THE ECONOMICS OF INFORMATION
(Mathematical Gnostics for Data Analysis)

Pavel Kovanic and Marcel B. Humber

September 6, 2015

Contents

Preface

Quantitative information ("how many", "how much", "at what cost" etc.) is one of the necessary elements for cognition[1]. Information of this nature results from a numeric data structure, which depicts (*quantifies*) the quantitative features of real objects and processes. Quantification is always perturbed by uncertainties, therefore methods that can suppress the imbedded uncertainty must be used to extract information from such data. The need for information has been growing at an increasing rate and both data and their processing are expensive.

Following the classical definition of an *economic good* as a good which is scarce relative to the total amount desired and *economic efficiency* as producing such goods at the lowest possible cost, then the idea of Economics of Information as the production of the maximum output of information given the cost (of data and of their treatment) is reasonable and leads to economic efficiency as well.

The first requirement of a methodology to treat uncertain data is that it must: **extract the maximum amount of information from a given collection of data.** It is the goal of this book to present

1. a mathematical theory of individual uncertain data,
2. an extension of this theory to small samples of uncertain data,
3. a development of data treatment methods based on this theory,
4. some demonstrations of results achieved by using these methods to analyzes of real data from several application fields, especially from economics.

These elements covering both theory and its usage form what will be called *mathematical gnostics*[2]. The book should demonstrate both theoretically and practically that the challenging task of getting maximum information from 'bad' data can be solved by the methodology of mathematical gnos-

[1] Cognition is used in the sense of the process used to obtain knowledge.

[2] There is only a philological relation of this notion to religious interpretation of words connected with the Greek word *gnosis* (knowledge).

tics.

The task is unusual and the approach to its accomplishment is to be unusual, too. The theory grew up on boundaries where several scientific domains contact each other: abstract algebra, measurement theory, Euclidean and non-Euclidean geometries, both classical and relativistic mechanics, thermodynamics, mathematical statistics both classical and robust. Reading of the first two parts of the book cannot be therefore easy for all readers. The real power of the new methodology can be vividly demonstrated by solving practical tasks. The Part III not only summarizes theoretical results important for applications, but contains many examples which can be understood even without a deep penetrating into the theory. Some readers can therefore start approaching mathematical gnostics by reading the third part.

The notion of economics of information can be considered only if the amount of information is measurable. Because the harvesting of information is directly related to decreasing uncertainty it is necessary to have at hand and to be able to use a scientific model of uncertainty.

There are several concepts of uncertainty and of its *paradigm*[3], the most popular of which is the statistical paradigm. However, the statistical paradigm—as are other related concepts—is tied to the uncertainty of mass events: statistical evaluation of a quantity of information is possible only for a sufficiently large "family" of events, not for a single event nor for a small number of events. Moreover, in order to successfully undertake such a task, an a priori model must be available and when the model does not fit the reality of the data, significant damage to the quality of the estimates can occur.

The wandering towards the information of individual data has to be started in the real world, with real objects, quantitative features of which data to be treated reflect.

[3] The notion of "paradigm" is understood to mean the prevailing opinion of the scientific community as to the arrangement and operation of things within a certain scientific field.

Part I

The Gnostic Theory of Individual Uncertain Data

Chapter 1

The Ideal Quantification

1.1 Quantitative Information

The term *information* has many meanings depending on the context. It is often related to such concepts as message, meaning, knowledge, text, image, communication, data, fact and many others. In all these examples the general definition of the http://en.wikipedia.org./wiki/Information is satisfied: *Information is the result of processing, manipulating and organizing data in a way, that adds to the knowledge of the receiver.* This book is oriented to a much narrower notion, to the *quantitative information* providing numerical description of real quantities and processes. Such descriptions are obtained by *numerical data*, that map the quantitative features of qualitatively specified real objects and processes. Increasing information decreases the receiver's uncertainty. This is why the Shannon's information theory measures the amount of signal's information by the Boltzmann's statistical entropy and why the Fisher's statistical measure of the information carried by an estimate of an unobservable parameter is a function of the variance. However, both these measures assume the availability of "mass data" to estimate a complete probabilistic model for the former and the variance for the latter evaluation. These measures of the amount of information cannot be therefore applied to individual data and to small data samples. Existence of the information contained in each data item is undoubted, it is justified by the fact, that information is carried by data sets. Usefulness of the ability to measure the information of the individual data and small samples results from the practical limitations of amount and quality of the data available to the treatment in practise. Unsuitability of the standard methods to solving this problem motivates the quest to find a suitable alternative method enabling the quantitative recognition of the reality.

1.2 Quantification

Quantification is *a procedure, which relates some real quantities to the numbers.* The numbers resulting from the quantification will be called *data*. There are two modes of the quantification, *counting* and *measuring.*

Every physical object has a given quality and exists in some quantity. Quality is an aggregation of traits, which define the nature of a thing. The first step in quantification is to establish the boundaries of a set of things in a qualitative sense, ie to describe the thing by listing its characteristics, so as to ensure the homogeneity and comparability of the underlying quantities. To quantify a herd of sheep, one must be able to distinguish between a sheep and a what is not a sheep and to accept as members of the set only the members of the herd, that fit the description. This set (flock) can be quantified either by counting or by measuring.

Counting the sheep means to assign to each animal the numeric unit (1) and to sum up all the units.

The objects of quantification, in this case the sheep, have their own properties, and these are separate from and independent of those of the persons, who actually do the counting. The properties of the sheep are fixed, and objectively given; the assignment of a unit of quantification, the number 'one' to an animal, is an intellectual (and subjective) activity performed by the "measurer", dependent on his or her abilities, and therefore subject to uncertainty: how well can he/she see? How carefully will he/she ensure, that all sheep in the flock have been located and counted? Is he/she sure, that some small animals were not blocked from view by the larger ones, etc.?

Measuring these same animals could consist eg of weighing the sheep. Two necessary elements are required for a measuring procedure: a (standard) unit of measure and a measuring instrument. To weigh a sheep, one needs a unit weight (pound, kilogram, ton) and a weighing machine to determine, how many times the weight of the sheep exceeds the standard unit, or how many times the unit exceeds the weight of the sheep. Measuring thus requires not only intellectual actions, but also some manipulations with physical tools. The results of a measuring process are the positive rational numbers: again a product of an intellectual activity.

Quantification in both of its modes is thus a mapping defined over a delimited part of the physical world, providing values, which are expressed in terms of the ideal world of mathematical objects: in numbers. This constitutes the fundamental difference between the art of the quantification and "pure" mathematics, which deals only with artificial (abstract) objects—products of the human brain. The quantification is a real technology, developed over thousands of years as a necessary element of the merchandizing. Just as with other technologies, it has been subjected to scientific analyzes, the outcome of which resulted in the establishment of a theory (the theory of measurement) ([17]). This theory provides us with the rules, which ensure, that the quantification procedures are consistent. However, there exists a serious limitation of this useful theory: it considers only an ideal quantification process, only the precise mapping and a disturbance-less measuring. Metrologists are aware of uncertainties, but they leave the treatment of the uncertain data "to statistics". A symptomatic characteristic can be found in http://en.wikipedia.org/wiki/Uncertainty/Measurement-Uncertainty:

> *Measurement uncertainties have to be estimated by means of declared procedures. These procedures, however, are intrinsically tied to the error model referred to. Currently, error models and consequently the procedures to assess measurement uncertainties are considered highly controversial. As a matter of fact, today the metrological community is deeply divided over the question as how to proceed. For the time being, all that can be done is to put the diverging positions side by side.*

A significant achievement of the measurement theory was the considering the quantification not only as a mapping of some real quantities onto numbers, but more specifically, as a mapping of some empirical relational structures onto the algebraic structures. This approach has been generally accepted as fruitful. An interesting requirement results from the Kuhn's ([65]) conception of the scientific revolution caused by a change of the paradigm: a new scientific paradigm should reveal some hidden assumptions, justification of which conditioned the validity of the old paradigm. This should make the old paradigm a special case of the new one. An example of this is the Newtonian mechanics still applicable to sufficiently slow movements. It could be therefore expected, that a more general theory of the quantification including the uncertainty would be able to establish conditions, under which the present measurement theory would hold up.

There are many assumptions about structures considered in the measurement theory ([87]), under which the relations of equivalence and preferences/ordering hold. A substantial, but plausible, simplification of these axioms was adopted in [56] about the nature of the underlying structures being isomorphic with the abstract commutative groups. A detailed mathematical justification of this simplified model was presented later in [100] and [99].

For our purposes, it is sufficient to limit the exposition of this topic to the presentation of two special cases of the commutative groups. These are closely connected to the structures of quantified real objects and their mathematical images.

1.3 Empirical Structure of Quantities

1.3.1 Mathematical Structures

The aim is to develop and justify a theory of individual uncertain data. The most reliable theories are created by the axiomatic method based on mathematics because of the power of mathematics to prove and disprove statements. Such a method essentially always produces a mathematical structure.

A mathematical structure is a set[1] (or several sets) of objects endowed with some relations and operations satisfying various assumptions (axioms). Identity of the elements of a set, as well as the rules to which the relations and operations are subjected, are uniquely and consistently established in mathematics by definitions. However, subjects of quantification are not abstract notions, but traits of the real world thought of as empirical structures. Moreover, as already mentioned, quantification is not a purely intellectual activity of men, but a technology including manipulation with real objects and instruments. To warrant consistency of the mapping of the empirical structures into the abstract world of mathematics, the measurement theory used the mathematical language to establish rules for empirical relations and operations. These correspond to the rules governing the relations and operations in mathematical structures, which represent the empirical structures.

To illustrate this thought, the notion of the *structure of cash flows* is introduced.

[1] A mathematical set is a collection of distinct objects considered as a whole. The extraordinary complex problem of deciding if an object is a member of a specific set (the "membership" problem) is to be postponed until particular sets and structures are considered.

1.3.2 Additive Group

Amount is one of basic quantitative characteristics of the elements of real sets. The number of a set's elements can be increased as well as decreased. Balance of incomes and expenses is a well-known worry in personal life, as well as, in the activity of enterprizes and institutions. An individual financial transaction realized in cash denoted as A_k can be thought of as an element of the cash flow. It has an empirical nature and a collection of all such elements forms an empirical set of cash flows (denoted $\mathcal{A}$). The adjective "empirical" is used to emphasize the material, and not only abstract, character of the objects. The fact of the membership in the set $\mathcal{A}$ will be denoted $A_k \in \mathcal{A}$.

The relation of the equivalence ($A_j \;=\; A_k$) of two elements A_j and A_k and the relation of preceding/preference ($A_m < A_n$) for some other elements of the cash flow can be accepted as natural as the binary operation of cumulation (the *additive aggregation rule*) written as $A_p \oplus A_r$ for each pair of the elements A_p and A_r of cash flows (denoted as $A_p,\ A_r \in \mathcal{A}$). Consider following requirements related not only to cash flows, but to all structures, which satisfy these conditions:

Closedness: Let A_m and A_n be two arbitrary elements of the structure ($A_m, A_n \in \mathcal{A}$). Then the aggregation of these elements is also an element of the structure:

$$A_m \oplus A_n \in \mathcal{A}. \tag{1.1}$$

Associativity: For each triple of elements $A_k, A_m, A_n, \in \mathcal{A}$, it holds, that

$$(A_k \oplus A_m) \oplus A_n = A_k \oplus (A_m \oplus A_n). \tag{1.2}$$

Commutativity: The order of operands does not change the aggregation:

$$A_m \oplus A_n = A_n \oplus A_m \tag{1.3}$$

for all pairs of elements A_m and A_n.

Neutral element: There exists in $\mathcal{A}$ an element (the "zero element" denoted by $O \in \mathcal{A}$), so that its aggregation with an arbitrary element ($A_m \in \mathcal{A}$) does not change the value:

$$A_m \oplus O = A_m. \tag{1.4}$$

Invertibility: There is also an element $\ominus A_m \in \mathcal{A}$ to an arbitrary element $A_m \in \mathcal{A}$ such that

$$A_m \oplus (\ominus A_m) = O. \tag{1.5}$$

The pair $\langle \mathcal{A}, \oplus \rangle$, of the empirical set $\mathcal{A}$ and of the aggregation operator $\oplus$ satisfying these requirements, can be called *the additive empirical structure.* Mathematicians like to deal with the abstract ("dematerialized") objects; they call these abstract structures obeying the rules of closedness, associativity and commutativity, having a neutral element, and inversions to all elements, *the commutative group.* A wellknown example of such a group is *Abel's group of real numbers*

$$\mathcal{G}_+ := \langle \mathcal{R}^1, + \rangle \tag{1.6}$$

where $\mathcal{R}^1$ denotes the set of real numbers and the $+$ is the operator of "ordinary" (numeric) addition of real numbers.

The numbers obtained by quantification of an additive group of empirical quantities will be called *additive data.*

This model is broadly applicable, but only to structures really obeying the formulated assumptions. Before the start of a data analysis, it is necessary to verify the additive nature of the structure quantified by the considered data, as they may represent an alternative, multiplicative, or even a more complex structure. Additivity of data, as well as of quantitative characteristics of real objects, is not a trivial problem. So, for example, velocities of real objects are not aggregated additively like in Newtonian mechanics, but in a nonlinear way respecting the limit of the speed of light. It will be shown below, that data uncertainty should be also aggregated by addition of certain nonlinear functions of uncertain data and not by additive aggregation of data like in statistics.

1.3.3 Multiplicative Groups

The task of measuring establishes how many times the measured quantity exceeds the quantity accepted as a unit, or how many times is the quantity smaller than the unit. The measurement's result is thus a ratio (multiplier), but this is a notion borrowed from the mathematics connected with numeric multiplication or division. A question may arise as to what is the real sense of the empirical operations of "multiplication" or "division". Their historical roots can be found in the distant past, in the development of barter markets. An example of a barter exchange can be the transaction "five arrows for one skin" written as

$$t_1 := (5\, arrows) \cong (1\, skin). \tag{1.7}$$

This expression establishes a relation (equivalence) between the values of two assets, ie their prices. Another transaction may follow:

$$t_2 := (1\, skin) \cong (20\, eggs). \tag{1.8}$$

The resulting "price" of arrows expressed in eggs is

$$t_1 \otimes t_2 := (5\, arrows) \cong (20\, eggs). \tag{1.9}$$

Ratios of exchanging things according to 1.7, 1.8 and 1.9 were different not only numerically (1/5, 1/20 and 5/20), but also with respect to the "measuring unit" or "dimension", the "price" being expressed in units of arrows, skins and eggs. Chaining the barter operations led to the changing of the units of measurement. A significant simplification of barter operation modes resulted from introducing of some kinds of currency as "units of the values of things". Worth of things could thus be expressed as a price determined by the ratio of the amount of a thing divided by by the amount of money. Necessity to determine the amount of things led to introduction of measuring units and to the development of instruments and measuring procedures providing the ratios of quantity/unit. Expressing the quantities in the same units allowed the quantitative features of things to be multiplicatively related by dimensionless ratios (multipliers). Pharmaceutical scales are real objects and when they determine, that two quantities' weights are W_1 and W_2 gram, then the relation $W_2 \ = W_2/W_1 * W_1$ becomes a mathematical model of the empirical relation existing between two quantities saying, that the empirical quantity W_2 is W_2/W_1-times larger or smaller than W_1.

The *Empirical Multiplication Factor (EMF)*, depicted by the multiplier W_2/W_1, can be defined by the relation

$$EMF_k = W_k \ \oslash \ W_{k-1} \tag{1.10}$$

where quantities W_k and W_{k-1} are 'empirical originals' mapped to the numbers and where $\oslash$ is the symbol of the empirical operation represented in mathematics by the operation of numeric division. This extremely simplifying representation of a relation is borrowed from cybernetics. Such an operation is presented as a black-box, the only interesting feature of which is its multiplication of the input values. It can serve as an important example of the multiplicative group. It is obvious that the output of a black-box can be used as the input for another black-box. The multiplicative factors defined by three quantities W_1, W_2 and W_3 can be thus chained multiplicatively:

$$EMF_3 \otimes EMF_1 == W_3 \ \oslash \ W_2 \otimes W_2 \ \oslash \ W_1, \tag{1.11}$$

where $\otimes$ is a symbol of empirical multiplication.

There are factors in economics, which have similar multiplicative nature: the examples of these characterize inflation/deflation, discounting, indexing, and interest factors among others. Prices viewed as multipliers can also be chained in a multiplicative manner: if price $P1$ is 3-times higher then $P0$, and price $P2$ is 4-times higher than $P1$, then $P2$ is 12-times greater than $P0$. Another important example relates to measuring by the application of a physical or chemical unit. Results of measurements viewed as multipliers (or ratios) can also be chained in the multiplicative manner: if a quantity Q_1 is exceeding the unit Q_u 3-times, and quantity Q_2 is 4-times larger than Q_1, then Q_2 is 12-times larger than Q_u.

We have considered in both 1.10 and 1.11 the *empirical* values W and empirical multiplication factors EMF. Quantifying mapping (counting or measuring) enabled the empirical inputs and outputs of operations to be numerically depicted and the multiplicative factors to be chained in a manner recalling numeric multiplication and division. The multiplication factors can be thus thought of as elements of a structure similar to the structure $\mathcal{G}_*$ formed by the set $\mathcal{R}_+$ of positive real numbers, over which the numeric multiplication $(*)$ is defined along with its inversion (division "/"):

$$\mathcal{G}_* := \langle \mathcal{R}_+, * \rangle. \tag{1.12}$$

This structure is the *multiplicative commutative group.* However, numeric values of quantities viewed as multipliers, and other numeric multiplication factors, are abstract images of quantitative features of real things. As such, they are results of quantification. The structure $\mathcal{G}_*$ is thus a mathematical image of an empirical structure of the real quantities. So as not to limit ourselves to a single specific type of subject, we shall denote a multiplication factor of a general type by M and a set of such elements of the same kind by $\mathcal{M}$. To ensure consistency in quantification, we assume that there is an empirical aggregation rule (operation) $\otimes$ defined over the set $\mathcal{M}$, for which the following relations hold:

Closedness: Let M_m and M_n be two arbitrary elements of $\mathcal{M}$. Then the result of the aggregation is also an element of this set:

$$M_m \otimes M_n \in \mathcal{M}. \tag{1.13}$$

Associativity: For each triple of elements $M_k, M_m, M_n \in \mathcal{M}$ it holds, that

$$(M_k \otimes M_m) \otimes M_n = M_k \otimes (M_m \otimes M_n). \tag{1.14}$$

Commutativity: The order of operands does not change the aggregation:

$$M_m \otimes M_n = M_n \otimes M_m \tag{1.15}$$

for all pairs of elements M_m and M_n of the set $\mathcal{M}$.

Neutral element: There exists in $\mathcal{M}$ an element (called "unit" and denoted $I \in \mathcal{M}$) such that its aggregation with an arbitrary element $M_m \in \mathcal{M}$ does not change its value:

$$M_m \otimes I = M_m. \tag{1.16}$$

Invertibility: There exists an element $\oslash M_m \in \mathcal{M}$ to an arbitrary element $M_m \in \mathcal{M}$ such that

$$M_m \otimes (\oslash M_m) = I. \tag{1.17}$$

The pair $\langle \mathcal{M}, \otimes \rangle$, of the empirical set $\mathcal{M}$ and of the aggregation operator $\otimes$, which obeys the above defined conditions can be called the *multiplicative group.* It is obvious that all five of the above conditions are satisfied for the mathematical structure $\mathcal{G}_*$, which is called *the multiplicative group of positive real numbers.* Note that not all multiplicative groups are necessarily commutative, but these particular groups are.

The concept of the multiplicative group is thus general enough to be applied to the measuring of quantities belonging to very different sets endowed with the operations $\otimes$ and $\oslash$ of various nature.

The numbers obtained by the quantification of a multiplicative group of empirical quantities will be called *multiplicative data.*

1.4 Isomorphism

From *iso* = same and *morph* = form. Let us consider two structures $\langle \mathcal{S}_1, \sigma_1 \rangle$ and $\langle \mathcal{S}_2, \sigma_2 \rangle$ where $\mathcal{S}_1$ and $\mathcal{S}_2$ are sets of elements $s_{1,m}$ (or $s_{2,n}$) for $m, n = 1, ...N$. Symbols σ_1 and σ_2 identify structural operations over the corresponding sets. A characteristic of these structures will be called an *isomorphism* if the following conditions are satisfied:

1. A one-to-one mapping $\tau_S : \mathcal{S}_1 \to \mathcal{S}_2$ exists, so that $s_{2,m} = \tau_S(s_{1,m})$ and $s_{1,m} = \tau_S^{-1}(s_{2,m})$ for all $m = 1, ..., N$.

2. For each pair of indices m and n ($m = 1, ..., N$ and $n = 1, ..., N$) the following implication holds:

$$s_{2,m}\sigma_2 s_{2,n} = \tau_S(s_{1,m}\sigma_1 s_{1,n}). \tag{1.18}$$

This definition can be illustrated by a simple example. We have already introduced the additive group of real numbers $\mathcal{G}_+$ in equation 1.6. Denote $s_{1,l}$ an arbitrary element of this group. Consider the relation

$$s_{2,m} = \exp(s_{1,m}). \tag{1.19}$$

The real number $s_{2,m}$ is strictly positive for all $s_{1,m}$. Moreover, it holds, that

$$s_{2,m} * s_{2,n} = \exp\left(s_{1,m} + s_{1,n}\right) \tag{1.20}$$

for all m and n. This means that the commutative group $\mathcal{G}_*$ is isomorphic with the additive group $\mathcal{G}_+$. The role of the mapping $\tau_S : \mathcal{G}_+ \to \mathcal{G}_*$ is played by the exponential function. The inverse mapping is the natural logarithm.

The power of the concept of isomorphism comes from focusing on the fundamental algebraic features of a large number of different structures and in omitting all their special characteristics. This thought can be illustrated by considering the group of cash flows. It is isomorphic with the abstract additive group $\mathcal{G}_+$: Each (material) cash flow has its (numeric) image (a numeric value). Each pair of cash flows produced by the aggregation operator $\oplus$ has as its (numeric) image the (numeric) sum of images of both cash flows obtained by the application of the numeric operator $+$. The numerical 0 corresponds to the neutral (zero) asset flow O. The inversions of cash flows are represented by negative images of cash flows.

Analogously, the group of measuring ratios (multipliers) is isomorphic with the abstract multiplicative group $\mathcal{G}_*$. Images of the multipliers are positive real numbers. The operator $\otimes$ is represented by the numeric operator of multiplication ($*$) and the inversion is represented by the reciprocal values of the ratios. We have already noted that the two abstract groups $\mathcal{G}_+$ and $\mathcal{G}_*$ are isomorphic. The group of measuring multipliers is thus isomorphic with the group of cash flows although their "material" natures substantially differ.

Note the differences between an additive an a multiplicative numeric group. Elements of the former can be positive, as well as negative or zero. Elements of the latter can only be strictly positive while zero element cannot exist.

The typical movement in the additive group is linear (repeated addition or subtraction of a constant). In contrast, the typical movement in the multiplicative group is exponential (repeated multiplication or division by a constant). Members of the additive group can be 'named' by having a measuring, such as eg physical, dimension: kilogram, meter. The group operation is not prevented by the (same) 'name': three meters plus two meters are five meters. However, the structure operation is not defined for elements having different dimensions. Unlike this, elements of a multiplicative group are dimensionless. This means, that results of measurements of quantities, which were expressed in the same measuring units, are to be viewed as measuring ratios (multipliers).

1.5 Ideal quantification

Using the notion of the isomorphism of groups, we can come to a more precise definition of quantification. This step is necessary in order to distinguish between ideal and practical quantification. The notion of *ideal quantification* is defined as an isomorphism between a group of empirical quantities and a group of real numbers. To explain:

It was already demonstrated by the example, that the historical origins of quantification (counting and measuring) were closely connected with the development of goods markets. The roots of mathematics are in these same practical needs of "ordinary life." Mathematics was primarily created to serve such plebeian, but necessary, activities. Due to its power and its universal generality, it developed as an abstract science completely isolated from everyday reality. Some of the "rich fruits" of mathematics were gathered by non-mathematicians: physicists, technicians, economists and other "practical" people. The idea of quantification is just another example of such an application. It clearly extends the borders of mathematics because of its reference to structures of a non-mathematical nature: the empirical quantities. On the other hand, it makes use of such a strictly mathematical notion as a group to describe the fundamental features of these non-mathematical structures. We shall see, that this mixture of the mathematical/non-mathematical approaches is very useful.

At this point we need to recall the one-to-one character of the mapping, which defined isomorphism and quantification. This was necessary to ensure the consistency of the ideal quantification. When the outcome of counting is five, we want to be sure, that there is a group of exactly five sheep in the flock, and that they correspond to number 5 and not to a greater or smaller number. This means, that the ideal quantification is a **precise** mapping. We now come to a difficult point: It is, at least at first sight, easy to imagine an accurate real number. It is much more difficult to do the same with real structures. As already stated, qualitative identification must precede the quantification. Perfect identification of objects subjected to quantification is difficult or even impossible. The same relates to quantification because of its necessity to involve real processes and objects. Imperfections play role in both cases making the results uncertain. The concept of ideal quantification can be seen as a simplified view of the way the quantitative features of the real world are observed and measured. What is urgently needed is a notion of *uncertainty.*

1.6 Summary

Real data can be considered as an outcome of a quantification process, as a collection of numeric images of real quantities. These quantities can be thought of under certain conditions as belonging to one of two (empirical) structures, the additive and/or multiplicative groups. An example of an empirical additive group is a set of cash flows over which the additive aggregation rule is defined. A mathematical model of empirical additive groups is the commutative Abel's group. An example of the empirical multiplicative group is a set of measuring ratios endowed with the multiplicative aggregation rule. Such groups are represented in mathematics by the abstract multiplicative group, which is isomorphic with the Abelian commutative group. Real data can thus be viewed from one of two points of view:

1. The additive group, that quantifies an empirical additive group of real quantities.
2. The multiplicative group, that quantifies tan empirical multiplicative group of real quantities.

Both these views are simplified. They are based on the assumption, that quantification is ideal, ie, that there is no uncertainty.

Chapter 2

Uncertainty

Unlike ideal quantification, the process of counting or measuring real quantities involves uncertain factors. To prepare mathematical modeling of the quantitative uncertainty as a component of real quantification, it is useful to start with consideration of some sources of the uncertainty.

2.1 Nature of Quantitative Uncertainty

G.W.F.Hegel in his Book One of "The Doctrine of Being" considers quality, quantity and measure as three grades of Being:

> *Quality is, in the first place, character identical with being: so identical, that a thing ceases to be what it is, if it loses its quality. Quantity, on the contrary, is the character external to being, and does not affect the being at all. Thus, a house remains what it is, whether it be greater or smaller; and red remains red, whether it be brighter or darker.*
>
> *Measure, the third grade of being, which is the unity of the first two, is a qualitative quantity. All things have their measure: ie the quantitative terms of their existence, their being so or so great, does not matter within certain limits; but when these limits are exceeded by an additional more or less, the things cease to be what they were.*

The last sentence points in the direction developed later in Marxist dialectic materialism in the form of the Law of Transformation allowing for the reverse with quality affecting quantity: "Continuous quantitative development results in qualitative 'leaps' in nature, whereby a completely new form or entity is produced."

This complex interdependence of the “three grades of Being” is reflected by complexity of the uncertainty: mistaken qualitative identification contributes to quantitative determination and distortion of the measure. Even the seemingly simple task of “counting sheep” can be difficult, when the qualitative recognition fails as demonstrated in Homer’s Odyssea by the blinded Cyclope Polyphemus incapable to distinguish between sheep and members of the Odysseus’ crew masked by sheep skins.

Data uncertainty can occur everywhere on their path from the observed quantity to the analyst’s computer. Its nature can be different in dependence on the essence of the particular field.

To make the notions to be considered below sufficiently specific, we narrowed the vast idea of information by focusing to its quantitative character numerically expressible. This information is contained in numeric data along with disturbances derived from uncertainty.

2.2 Uncertainty in Numeric Data

Numeric data resulting from quantification can be uncertain due to many different causes:

- Imperfect identification of the quantified object or process.
- Imperfect observation or design of experiment:
  - Instability of observed objects or volatility of the process.
  - Impacts of environment.
  - Insufficient number of repeated measurements.
  - Insufficient isolation of the observed quantity from impacts of other quantities.
  - Overlooked or neglected interactions between measured objects.
- Instable or faulty measuring instruments.
- Transformation errors.
- Communication errors.
- Incompleteness of measuring (data censoring).
- Geometric errors.
- Events or data aggregation errors.

Transformation errors can result from the necessity to exchange information-bearing media: many physical and chemical quantities are ultimately measured using electric instruments, but this requires conversion of the original quantities into electric currents, voltages, impulses or codes. These conversions can be imprecise and noisy. Further contributions to the uncertainty can result from the transmission of these signals. Chemical measuring can include dilution or concentration of liquids, distillation and other transformations, which distort the quality of the measurement.

All measuring instruments operate only within a limited scale. Values of some quantities can occur below the limit of detection, others can exceed the upper bound of the measuring range. Measurement can be then interpreted only in terms of an inequality instead of a discrete number. Neglecting such data would be a bad idea, because these (*censored*) data can be the best ones (when signalling, that there is only a small danger) as well as the vitally important ones (a danger exceeded a limit). Some censored data can have only an interval nature. All censored data contain some information in spite of their special uncertainty.

Geometric errors can result from application of an unsuitable geometry to measuring. A nearly trivial example is using plane maps to quantify distances and angles in the real three-dimensional word. Nontrivial examples of the necessary application of non-Euclidean geometries to measure some real processes and objects are known from physics, but relevance of the non-Euclidean geometries to measuring uncertain data also exists and will be considered in the sequel.

The simplest (additive) way used to aggregate events and/or data cannot always be adopted. Sometimes either weighted aggregation or an even more complex method is to be applied.

Examples of uncertainty in measurements can be seen in systems of the Quality Assessment Control broadly used in the technology for maintaining quality standards of industrial products. These standards are defined by sets of measurable parameters checked by the system. Quality can be disturbed by many factors, causing the uncertainty in the parameters' values: bad quality of the raw materials used for production, tools and machinery, deviations from the prescribed technology of production or storing, malfunction of instruments, human factors and others.

Fundamentally different types of uncertainty exist in economics and in economic data.

2.3 Uncertainty in Economic Data

An old proverb goes: "When two people do the same thing, it isn't the same thing at all." Applying this thought to "measurement" one could say: "When an economist measures something, it is not the same process as that used by a physicist." In the methodology of physics, chemistry, engineering and biology, measurement frequently produces a high level of precision, perhaps resulting in the convergence of the outcome to a **unique** value. This value exists objectively, it is the same for all observers and for their measuring tools. Trying to emulate this technique in accounting or in the collection of other economic data is a waste of time. The objective existence of worth or value of things is for an economist proven only in terms of philosophy: "Yes, a value exists always, if the thing can be sold or exchanged." Divergence of opinions inevitably starts with quantification of the value, because there is no objective economic value of a thing. The first problem stems from the use of the same word "measurement" and the manner, in which it may be understood by economists. Another problem is connected with uncertainty.

Paul A. Samuelson [97] introduces uncertainty in economic behavior in the following way:

> *. . . No study of the realities of economic life is complete without a thorough study of the fascinating interplay of uncertainty and strategy . . .*
> *. . . In reality, business life is teeming with risk and uncertainty. The demand for a firm's output will fluctuate from month to month; input prices of labor, land, machines, and fuel are often highly volatile; the behavior of competitors cannot be forecast in advance . . . Life is a risky business . . .*

A fundamental role in providing the economics with data is that of accountancy. The layman's view, that accounting data represents true values, is not correct. This is due to the fact, that accounting documents reflect *adjusted* historical costs. Prices (or current market values), on the other hand, reflect the analyst's or investor's *perception* of the *present value* of the *future,* and therefore unknown, *cash flows.* Problems can be seen already in the definition of the accounting ([23]):

> *Accounting is the process of identifying, measuring and communicating economic information to permit informed judgements and decisions by users of the information.*

There are two similar definitions in [23]. They leave some fundamental questions opened: what should be identified and measured to get and communicate the information about it? One of these definitions tries to be more specific by description of the object as "quantitative information, primarily financial in nature, about economic entities, that is intended to be useful in making economic decisions ...". However, this is a tautology. The missing notion is **worth** or - as reflected by accounting - **value**. This notion is an agelong worry of economists as can be seen in the book [102] first published in 1776:

> *The word VALUE, it is to be observed, has two different meanings, and sometimes expresses the utility of some particular object, and sometimes the power of purchasing other goods, which the possession of that object conveys. The one may be called 'value in use'; the other 'value in exchange'.*

It is worth mentioning, that this dual viewing was introduced ages ago by Aristotle (384-322 B.C.). It is not less noteworthy, that it is applied even recently, when the valuation problem became an object of international efforts directed to standardization of valuation ([37]). International Valuation Standards distinguish nonmarket and market approaches. *Nonmarket values* include:

- *Value in Use*: worth for a particular owner.
- *Value in Exchange*: a value acknowledged by the market, where could be a hypothetical exchange of the asset realized.
- *Investment Worth*: a value for an investor following his specific goals.
- *Going Concern Value*: worth of an enterprise as a whole.
- *Insurable Value*: value specified in an contract of insurance.
- *Assessed, Rateable or Taxable Value*: values defined by corresponding legal regulation.
- *Salvage Value*: the net realizable value.
- *Liquidation Value or Forced Sale Value*: value in a situation, when the application of the market value is impossible.
- *Special Value*: value regarding some specific factors distinguishing the transaction from the current market conditions.
- *Mortgage Lending Value*: evaluation for securing a mortgage.

The *market value* is specified as an estimate of the price obtainable in a hypothetical transaction on a specified date or a measure of the value, that will accrue from ownership by a particular party. This value depends on the valuation basis adopted and the required valuation premise.

It is obvious, that all these quantities are evaluated subjectively, being dependent on the needs of subjects, on the conditions of the market, on intentions of investors, on the policy of banks and insurance companies, on law making and especially on judgements of specialists performing the valuation.

Fundamental information used for economic valuation stems from accountancy.

2.4 Measuring in Accountancy

To measure, a physicist or engineer applies a measuring tool or apparatus. Such material instruments do not exist to be used in accountancy. Instead, "measuring tasks" are realized by "generally accepted" procedures formulated as national and international standards ([35], [36]). Such "measuring manuals" are not simple, the former has 1265 pages.

It is remarkable, that the International Accounting Standards do not consider the notion of measurement as a primitive[1]: a definition is provided ([35]), and the idea is further clarified by the inclusion of specific measurement bases:

> *Measurement* is the process of determining the monetary amounts, at which the elements of the financial statements are to be recognized and carried in the balance sheet and income statement. This involves the selection of the particular *basis of measurement.*

A number of measurement bases are employed and these are applied to various portions of financial statements in varying degrees. They include the following:

(a) *Historical cost.* The amount of cash or cash equivalents paid, or the fair value of the consideration given at the time of acquisition.

(b) *Current cost.* The amount of cash or cash equivalents, that would have to be paid, if the same or an equivalent asset was acquired currently. Liabilities are carried at the undiscounted amount of cash or cash equivalents, that would be required to settle the obligation currently.

(c) *Realizable (settlement) value.* The amount of cash or cash equivalents, that could currently be obtained by selling the asset in an orderly market. Liabilities are carried at their settlement values; that is, the undiscounted amount of cash or cash equivalents expected to be paid to satisfy the liabilities in the normal course of business.

(d) *Present value.* Assets are carried at the present discounted value of the future net cash inflows, that the item is expected to generate.

[1] In mathematics, 'primitive' is a notion, which does not require a definition because "everybody knows, what it means."

This latter idea attempts to inject the notion of market value. One should be willing to pay just enough for an asset to offset the future value of the cash flow, which it is expected to contribute. As satisfying as this thought may be, it brings along as extra baggage several additional sources of uncertainty.

First of all, the future cash flows are unknown, therefore risky, and they are represented by a best guess of their magnitude[2]. Then the present value of these flows must be estimated using an uncertain discount rate, which reflects that risk.

Another issue is the relative risk of certain elements in the cash flow stream. Operational flows are one thing, but once an asset has been acquired, how certain is the depreciation, that will be taken? What about the resale value of the asset at the end of its useful life?

Another source of uncertainty comes from the fact, that today's market rates, whether they are short, intermediate, or long term, all imbed current perceptions of *expected* future market conditions. However, for example, the second year's risky cash flow should be discounted at the second year's relevant rate, but this rate will not be known until the second year, when the cash flow is actually received. The rates, that were expected, will probably not be the ones, that actually occur, and the true value of the cash flow at the end of the second year will not be that same value, which was estimated, when the asset was purchased.

Most commonly, enterprises use historical cost when preparing their financial statements. However, this is usually combined with other bases (inventory is usually carried at the lower of cost or market value, marketable securities can be carried at market value, pension liabilities and capital lease obligations are carried at their present value, etc.). When inflation is a persistent and recurrent problem, and where it is permitted, some firms turn to the use of current cost basis in response to the inability of the historical cost accounting model to deal with the effect of the changing prices of nonmonetary assets. In some countries, accountancy law specifies the use of different measurement bases for different elements of financial statements.

There are further sources of accounting uncertainty:

Vagueness of definitions: Complexity of the problems and diversity of the processes do not enable creation of uniquely interpretable definitions. Neither the basic accounting notion of *recognition* cannot be

[2] There are many techniques, which can be used to refine these estimates.

described precisely. According to Standards, an item is recognized if:

- (a) it is probable, that any future economic benefit associated with the item will flow to or from the enterprise; and
- (b) the item has a cost or value, that can be measured with reliability.

Ambiguity in terminology: Expressions like ...fair value..., ...expected to be paid..., ...equivalent asset..., ...expected to generate..., the normal course of business..., ...expected to be required... are more or less fuzzy and are permanently discussed by specialists of the field.

A degree of tolerance in selection of the particular basis of measurement for a particular task must exist. However, different measurement bases → different results from measurement.

Subjectivity: Unlike 'technical' measurements tending to maximal independence of the results on the subject, who performs the measurement, the role of **professional judgments** is emphasized. However, different professionals can come to different conclusions.

Ethical issues and criminal actions can also contribute to uncertainty.

It can be concluded, that both measurement and recognition in accounting are inevitably connected with vague notions and actions, which include unknown and unpredictable factors, that increase the uncertainty of accounting data. These sources of uncertainty have a fundamental nature, which the intensive long-term effort of the international community of accountants has not yet been able to overcome.

2.5 Fighting against Uncertainty

Aims to minimize negative effects of uncertainty on quality of measurement were the driving forces of the development of measurement technology. A leap in evolution of the field was brought by the development of statistics. Decreasing uncertainty of the quantification by repeating measurements followed by statistical analysis enabled results' quality to be significantly improved. Moreover, a statistical way of thinking helped to scientists to effectively model complex natural "uncertain" processes of demography and public administration and to create new scientific fields like statistical thermodynamics, nuclear physics along with nuclear engineering and the theory of information. The development of computers allowed to apply statistical methods to satisfy rapidly increasing needs of improving the results of measurements. However, serious problems with applications of statistics arose in many fields of practice, where uncertainty was not adhering the necessary assumptions of statistical modeling. Difficulties manifested significantly around ensuring a “sufficient” amount of data. Modern measuring can be expensive. There is a risk of delay as results must be obtained as soon as possible. Losses caused by destructive measurements can be excessive. Monitored processes can be too fast to allow many measurements. Application of statistical methods in all such cases is no longer feasible.

The efforts to decrease uncertainty are motivated by the aim to maximize the opposite: the information. However, as already mentioned, statistical models of uncertainty and information are based on the idea of mass uncertainty.

2.6 Summary

Unlike ideal quantification, real quantification is neither precise nor consistent, although the existence of an unique true value of the quantity can be assumed in many application fields, apart from economics. However, there is always an unavoidable participant in quantification - uncertainty. The nature of uncertainty can be very different; it can originate from many sources. Its presence in data implies risks, errors and the degradation of results’ reliability and utility. The availability of the large amounts of data required to remove uncertainty using statistical methods cannot be always assumed. Examples of existing causes and sources of uncertainty show, that the application of statistical models of uncertainty can be therefore inadequate to many required tasks. To minimize effects of uncertainty on quantification process in such applications, a realistic theory of real quantification, which is applicable to small data samples, is needed.

Chapter 3

Geometric Paradigm

3.1 Paradigm

A theory is an abstraction conceived to explain or predict reality. It needs to include sufficient information, but also to suppress irrelevant facts. Therefore a model is proposed as a reproduction of what is observed as reality. Theories evolve over time and are improved or negated as empirical investigation either supports or refutes what had previously been postulated.

The notion of a *paradigm* is close to that of a *model* but Thomas Kuhn introduced [65] a special interpretation which pertains to scientific revolutions. He suggests that a paradigm represents a collection of generally accepted views which dominate the thinking of "experts" in a scientific field at some point in the development of a theory. As shown later by Joel Barker [5], the problem of the paradigm is much more universal in its nature and it is one of the most important questions in the development of our everyday life. A paradigm consists of two major parts. It:

1. delimits the boundaries of a class of problems, and
2. includes a collection of rules to solve the problems which exist within the given boundaries.

The acceptance of an existing paradigm results in several advantages:

The paradigm

- helps to distinguish between the important and the insignificant,
- offers advice and recommendations as to how to move successfully within the given boundaries,
- aids in communication between its adherents because they are all familiar with and use the same notions, terms and language,

- helps in the understanding of changes within its “valid” framework because it is understood as being “legal” and it does not give rise to suspicions of “heresy” within the domain and it does not lead to conflicts with the “pontiffs”, who dominate the field,
- assists in legitimizing activities within its boundaries, thus increasing the number of its adherents and sponsors.

There are also negative features. A paradigm is a “filter”, which selects and adapts incoming information to support itself and to eliminate inconsistencies with any new facts. Murphology has two observations on this issue [8]:

1. **Maier’s law:** “If facts do not correspond to your theory, get rid of them as fast as possible.”
2. **Finagl’s credo:** “Science is always right. Do not be confused by facts.”

These theses are (sadly) more than a joke; if these conclusions were not frequently real, the acceptance of a new paradigm would be much easier and faster. Blind and uncritical adherence to an existing paradigm often results in a closed mind and to an erroneous conviction that everything successful in the past must be successful in the future because the future is nothing more than a simple extrapolation of the past.

In defense of maintaining the old order, a change in the paradigm might encompass large risks:

- At the moment, when a revolutionary paradigm is accepted, a great deal of the built-up intellectual or spiritual “capital” of those, who supported, nurtured, and maintained the supplanted paradigm is lost, and nearly everyone starts from zero once again.
- New paradigms ordinarily appear at the boundaries of several scientific fields, which are not familiar to the “priests” of the old paradigm. Younger scholars with fresh new knowledge, and newcomers from the new “neighboring” fields are favored.
- As a potential revolution of the paradigm develops, it is not sure, who will win. Many will prefer to wait on the sidelines to see, which way the wind blows, before putting their necks on the chopping block or opening themselves to criticism.
- The old paradigm is rarely completely refuted; this permits the established ideas to continue to be harvested until the new ones are completely established. Moreover, some new paradigms are more general than the old ones and include the former as valid special cases. This,

of course, does not apply to conflicts between hostile paradigms such as between social systems or between paradigms, which exclude each other (such as the Ptolemaic versus the Galilean paradigms).

- Occasionally, a seemingly new and better paradigm appears to be more a fashion or a fad than a well justified innovation (eg many slimming cures, following the 'herding instinct' in jogging, etc.).
- "New" is not automatically the same as "progressive" or "better". History is replete with many new ideas or discoveries, which have lead to dead end roads or U-turns, eg DDT, cheap and safe nuclear energy, small-pox vaccinations for children, etc.
- Some paradigms, especially those related to the use of power may be highly dangerous, (eg Hitler's "Blitzkrieg" or religious fundamentalism).

Having considered the advantages of continuing to subscribe to an old paradigm with the risks of accepting new or revised ideas, one can develop a better understanding of conservatism in thought and general resistance to change.

It is logical to ask, why such a philosophical problem is dealt with in a book, which aims to contribute to the analysis of real data . The answer is that to analyze, one needs data and analytical methods. Real data contain strong uncertainty. Hence, the methodology, which will be applied, must be able to cope with the inherent uncertainties. There are different paradigms of uncertainty; to select the most suitable, it is necessary first to reflect on the choice of the proper geometrical paradigm, then to pick correct paradigm of uncertainty.

3.2 Why Geometry?

The purpose of this book is to present a new method for the solution of problems in the applied sciences; therefore, at first, it may seem puzzling to have geometry play a major role. However, applied sciences deal with data. Data result from measurement, and measurement is the main task of geometry.

Let us now explore this problem. It is a widely held point of view that geometry is the branch of mathematics that deals with the properties, measurement, and the relationship of points, lines, planes, and solids [112]. Three notions of fundamental importance are missing here: *space! transformations,* and *invariants.* These are necessary because all geometric objects must be placed somewhere and in some manner moved or changed. Felix Klein[1] stated (Klein, 1921) that these particular notions are the most important features of geometry:

> *Geometry is the science studying invariants of figures, ie properties, which do not change by movements.*

3.2.1 Space and its Geometry

One idea is that space is something like a box "containing" geometric objects. Such thoughts are promoted by the traditional approach to geometry, which focuses on only a single geometry (the Euclidean one) and ignores the existence of a large number of others, some of which are extremely important in our lives. Indeed, one box (eg a shoe box) is very similar to another box (a hotel room) in that the distance between two points is the length of a straight line connecting the points. We even know that this length is the minimum of lengths measured along all possible different paths (this feature is called *the variational principle*). The notions of the "right angle" and "parallelness" are identical in both boxes. The same relates to angles between lines, which can be measured using the same protractor. These notions are simple and natural, but only because we are accustomed to viewing the world from the standpoint of Euclidean geometry, ie because we were educated to uncritically accept the Euclidean geometric paradigm.

When evaluating distances in Euclidean geometry, the elements of the path are summed, giving each the same weight, but are they all of the same importance?

- The manager of a citrus grove monitoring a winter forecast would not care if the temperature was expected to fall to 60° from 65°. However, he might need to start thinking about how to react should this five degree change be between 40° to 35°, and he would certainly take desparate action should it range from 33° to 28°. The "weight" of a degree of temperature change depends on its level!
- The CEO of a firm would probably find justification for a 2% drop in revenue or net income by blaming a competitor's new product, inflation, or a blip on the GNP growth chart, but were this change

[1] (1849-1928) A well known mathematician and geometer at the University of Göttingen.

to amount to 10%, some very serious activity would take place in the boardroom.

This idea of imposing different weights on different segments of a distance scale depending on a "locally" determined value bears a close resemblance to the assessment of uncertainty (or the errors) in data by measuring the distances between an (unknown) true datum and the observed value. If there are several observations and an estimate of the true value (the "central" value), it is natural to weight the observations closer to the center more heavily and to give smaller weights to those more distant. This is a straightforward application of a non-Euclidean geometry. What is more complex is the determination the "weighting function,"—in other words—the choice of the particular non-Euclidean geometry to be employed.

The point is:

Different geometries ⇔ different measures.

The German mathematician Bernhard Riemann (1826-1866) developed his geometric paradigm in the middle of the 19th century. Riemannian space is a set of points (manifold) endowed with specific instructions defining the way lengths and angles are measured at all points in the space. (The measurement method—*metric*—may be different for different points). All the characteristics of the space depend on its specifications (curvature, variational principles, etc.). One of Riemann's hypotheses is of special importance to our purposes; it can be roughly stated in this way:

It is not for mathematicians to choose metrics of spaces to model real processes. Metrics are given objectively by laws of Nature.

No one could confirm Riemann's hypothesis in his time because the progress of science had not matured sufficiently and very few, if any, had any conception of what he was saying. It was more than half a century later before Albert Einstein proved through his special theory of relativity that the space we live in is not Euclidean but substantially different—Minkowskian. This conclusion resulted from experiments documenting the finite speed of light—a law of Nature. Einstein's gravitation theory (the general theory of relativity), a decade later, explained other physical experiments, which proved that the geometry of outer space is Riemannian. The (local) metric at each point is determined by the (local) gravitation forces—again by something objective.

It is now the right moment to ask a modified Riemannian question:

What is the proper metric to use to measure the uncertainty of real data obtained by quantification?

In other words:

What is the proper geometric paradigm and the proper paradigm of uncertainty, on which methods for analyzes of real data should be based?

Indeed, data uncertainty results in data errors. What is the "size" of an error? If the reply were that the error should be measured as the difference between the true (ideal) value and the observed data, this approach being derived from Euclidean geometry, the immediate follow-up question would be to explain the **reasons, which motivate the choice of Euclidean geometry.** Error is obviously distance but to measure distances, we should know the geometry of the space.

It should now be clear, why we are concerned with geometry, when dealing with data treatment. Our problem is to find the reasons that establish the "proper" geometry of spaces for real events and processes. One intuitively feels that this geometry is closely connected with uncertainty.

3.2.2 Transformation of a Space

Transformations play an important role in geometry. Three notions are essential for our purposes:

1. geometric movement,
2. invariants of transformations,
3. replacement of coordinates.

Anyone, who has used computer graphics, is familiar with various types of *geometric movement.* Moving a point by means of the "mouse," one produces straight lines or curves. Rotating a line having a constant length about a point draws a circle. Using the sequence of operations 'copy'—'paste'—'rotate' a square can be constructed from one of its sides. Not all users of these techniques are aware of the fact that they are actually applying geometric transformations. They slightly modify the values of the coordinates of the last point thus generating a new point, store the new point's coordinates and use the new point as the point of departure for the next movement. By such movements, geometry creates lines from points, figures from lines and solids from figures. Further, by geometric movement, a space with a higher dimension can be created from one of a lower dimension.

It is important not to confuse geometric (virtual) with physical (or—more generally—real) movement. A geometric movement can model a physical one, but it is a more general notion. Not all possible geometric movements can be realized in the real world. A plate lying on a table cannot move down vertically but a geometrical point on the plate can move down freely along a vertical line as well as up. Another significant difference between geometric and physical movement is that of the time aspect. Physical movement is strongly parametrized by the time coordinate and its speed cannot exceed the speed of light. Geometric—purely mathematical—movement (not represented by the physical movement of the cursor on the screen) has no relation to time. Geometric movement can of course model other kinds of movement, eg movements taking place in the financial world. The development of the individual price of shares can be thought of as a geometric movement (this is what "chartists" try to do in attempting to predict future prices by drawing straight lines, or other more complicated geometric shapes representing the path of past prices).

Examples of a "collective" movement in the financial world are inflation or the devaluation of currency. The affected group participates in a real macroscopic movement, yet its individual members interact with each other on the "micro" level. These movements also have a double nature, when they occur: they are real, but on the analysts' screens. They have a different character, since the real (financial) movement is modeled by the physical movement of electrons, which draw lines, thus representing a geometrical movement.

There are many kinds of transformations; a convenient way to classify them is by the use of *invariants.* Invariants are features of geometric objects, which permit them to remain unchanged through certain transformations. It is important to emphasize that there is a requirement for a triple of notions: **space—transformation—invariant**, all of which are mutually bounded. There are *affine transformations* within Euclidean space such as shifts and orthogonal rotations. Invariants with respect to these transformations are lengths and angles. Note that these words “lengths,” “angles” and “orthogonal” should be understood as “lengths”, “angles” and “orthogonal” in the sense of Euclidean geometry. In a different geometry they may mean something entirely different.

There are invariants of transformations also in the space, which represents the financial world: the rate of interest is a time invariant of the exponential curve depicting the change in value of a bank deposit in the case of a constant interest; the ratio of liabilities to receivables does not change with a currency’s devaluation.

Mathematicians have the freedom to choose the coordinate system of a space from a whole class of permissible systems and they ordinarily use the most convenient one. To solve problems of a rectangular nature, they use a Cartesian system of coordinates (eg $\langle x,\ y,\ z\rangle$). When working within a sphere, they choose spherical coordinates ($\langle \rho,\ \phi,\ \theta, \rangle$) where ρ is the diameter, ϕ the azimuth and θ is the altitude) and so on. This does not mean that another system, perhaps the Cartesian system, cannot be used to measure the position of stars, just that it might be more cumbersome. However, it is sometimes necessary to change from one coordinate system to another, *to replace coordinates.* One extremely important aspect of replacing coordinates is that of invariants. This is especially familiar to physicists because it allows objective features to be distinguished from subjective ones. A good example is a tension within a solid body. Such a stress together with resulting deformations is something objective, it exists independently of the observer trying to model it mathematically. However, the written appearance of equations, which describe the state of the body, is dependent on the choice of coordinate system, it is thus subjective. If the load and resulting stresses exceed a certain critical value, the body will break. This outcome is objective, it is independent of the chosen coordinates. The tensions therefore must be invariant to transformations that exchange coordinates. There is an ideal mathematical tool, *tensor*, () which can determine invariants as eigenvalues of tensors for a whole class of transformations, which replace coordinates.

3.3 Summary

To analyze real data, which, as a rule, are strongly disturbed by uncertain components of various origins, one needs good analytical methods. To develop good analytical methods, a well conceived theory is essential. Every theory is based on a paradigm, which is a collection of views which have been generally accepted during the historical development of a field and which dominate the collective thinking of experts in that field.

The choice of geometry to be used to represent real quantities (or their movement) is critical to the nature of the results, that will be obtained. Since there is no freedom in the selection of the geometric paradigm, it is necessary to find suitable laws of Nature, which can determine the appropriate geometry. Preliminary consideration of the problem reveals, that there is no confidence in the suitability of the Euclidean geometric paradigm and its corresponding analytical methodology for the treatment of real data. A difficult problem which remains is the justification of the specific geometry of a Riemannian type for this class of applications.

Chapter 4

Paradigms of Uncertainty

4.1 Statistical Paradigms

When using the word *statistics,* one must distinguish between two substantially different meanings:

1. numbers, that have been collected in order to provide information about something,
2. the science of collecting and analyzing these numbers.

The numbers cited in both definitions are *data*, ie outputs of the quantification process, which map real quantities. A quantitative depiction of reality would be impossible without data. The statistical activity described by the first definition is an absolutely necessary part of all methods used to obtain quantitative information. The second meaning also defines an objective of mathematical statistics, but it does not imply its uniqueness as a tool for data analysis.

The use of the plural in the heading above might be a shocking revelation for one, whose acquaintance with statistics is via the most popular paradigm, that of *relative-frequency* statistics. It is based on one of the oldest paradigms closely connected with games of chance such as dice, cards or roulette. The *relative frequency* (number of successes divided by the number of trials) characterizes the success of repeating a given (*random*) experiment under fixed conditions. The thrust of this paradigm is described by [22] as follows:

1. In many important cases relative frequencies appear to converge or stabilize, when the random experiment is repeated a sufficient number of times.
2. This apparent convergence is an empirical fact and a striking instance of order in chaos.

3. The apparent convergence imputes a hypothesis, that the relative frequency of outcomes in as yet unperformed trials of an experiment can be extrapolated from the observed relative frequency of trials already run.
4. Probability can be interpreted through the limit of relative frequency and assessed from relative frequency data.

This statistical paradigm is not unique. There are altogether seven classes of theories of probability based on different paradigms described in detail and analyzed in [22]. The conclusions drawn therein are far from optimistic:

> *... The many difficulties encountered in attempts to understand and apply present-day theories of probability suggest the need for a new perspective. Conceivably, probability is not possible. A careful sifting of our intuitive expectations and requirements for a theory of probability might reveal, that they are illusory or even logically inconsistent. Perhaps the Gordian knot, whose strands we have been examining, is best cut. However, where would such a drastic step leave the world of practice?*
> *... Clearly much remains to be understood about random phenomena before technology and science can be soundly and rapidly advanced. It is not only the "laws" of today that may be in error, but also our whole conception of the formation and meaning of laws.*

Perhaps this sad state of affairs can be interpreted as a call for a good nonstatistical paradigm to assess uncertainty.

4.2 Nonstatistical Paradigms of Uncertainty

Many problems of a theoretical nature have given rise to new attempts to reconsider various statistical paradigms. The rapid development of computers after World War II enabled existing statistical methods to be applied to real problems to an extent never thought possible before. However, results have been far from satisfactory. This may be explained by the fact, that statistical methods are products of mathematics; and as such, if they were developed from nonconflicting assumptions in a consistent manner, they cannot be wrong. If a mathematical statistics methodology fails, when it is applied to real data, it is necessary to look for the cause in the conflict between the theoretical assumptions and the real nature of the data. A statistician may warrant his methodology to one, who requests an analysis, only if he in turn provides the statistician with a warranted statistical model of the data. Statisticians ordinarily make the requester responsible for the choice of data model. **The stumbling block is, that one very rarely knows the statistical model of real data.**

Both theoretical and practical problems with statistical paradigms have lead to the fast development of methods based on alternative, nonstatistical paradigms. As outlined in [72], several of these methods are being applied in forecasting and decision making in the financial markets. The use of methods based on pattern recognition, neural networks, fractal geometry, deterministic chaos, fuzzy logic, genetic algorithms and nonlinear dynamic theory are discussed. However, this list of existing alternatives to statistics is short by at least one candidate. It seems, that there is an informal, but ferocious, "race" running for alternative paradigms of uncertainty. It is not the aim of this text to deal with the broad spectrum of nonstatistical paradigms of uncertainty. Instead, we shall concentrate on a single participant in the "race," on the *gnostic* paradigm, which was absent from the roll call given in [72].

4.3 Is there a Need for an Alternative to Statistics?

Economists rely on statistics to collect data, but the use of mathematical statistics as the unique technology of choice for extracting information from these data may be a questionable procedure[1].

The historical achievements of statistics, especially in physics, has justified consideration of this methodology for use in the analysis of real phenomena. Theories of statistical thermodynamics, chain fission reaction, and neutron slowdown and diffusion yield precise engineering calculations for nuclear reactors. These constitute some of the unchallenged successes of the statistical approach. However, is this sufficient reason to expect, that the application of the same principles will yield equally successful results when applied, for instance, to economics? Because economic processes are substantially different from physical ones, it is not likely. Benjamin Graham, the father of "fundamental" investment analysis stated [30]:

> ... *The art of investment has one characteristic that is not generally appreciated. A creditable, if unspecular, result can be*

[1] The basis for the material in this section is taken from a presentation made by the authors at the third international Artificial Intelligence in Economics and Management (AIEM) workshop in Portland, Oregon in August of 1993.

> *achieved by the lay investor with a minimum of effort and capability; but to improve this easily attainable standard requires much application and more than a trace of wisdom. If you merely try to bring* just a little *(emphasis added) extra knowledge and cleverness to bear upon your investment program, instead of realizing a little better than normal results, you may well find, that you have done worse.*
>
> *Since anyone—by just buying and holding a representative list—can equal the performance of the market averages, it would seem a comparatively simple matter to "beat the averages"; but as a matter of fact, the proportion of smart people, who try this and fail, is surprisingly large. Even the majority of investment funds, with all their experienced personnel, have not performed so well over the years as has the general market . . . there is strong evidence, that their calculated forecasts have been somewhat less reliable than the simple tossing of a coin.*

Although there is no reference to specific forecasting methods, it can be inferred, that the word "calculated" refers to the mathematical methodology of statistics, that has almost exclusively dominated econometrics for decades. Among other pertinent critiques of the statistical approach to economic problems the following remarks by Los ([70]) can be mentioned:

> . . . *It is clear to most people, that economic forecasting still amounts to little more than educated guessing, despite the aura of precision created by computerized models of the economy.*
>
> . . . *Scientific economic analysis, in the true sense of these words, still does not exist.*
>
> . . . *Since objective modeling has not been practiced, economics as a science has not progressed.*
>
> . . . *Recently, simple cost-benefit analysis has created strong financial incentives to obtain better and more accurate economic forecasts in the private sector. But, paradoxically, the main obstacle to this progress in economics is the conventional pseudoscientific methodology of econometrics adopted in the 1940's and 1950's. The conclusion is clear: first the problem of objective identification from noisy data has to be solved.*

Professor R. E. Kalman, who made a substantial contribution to cybernetics with his famous filters, expresses his view of the issue as follows

[42]:

> *...Statistics is not science, but a kind of prescience, a pseudoscience, a "gedankenscience."*[2] *Perhaps it's best called an "ersatzscience."*[3]
>
> *...Uncertainty in nature cannot be modeled (and therefore must not be modeled) by conventional, Kolmogorov*[4] *probability schemes, because no such scheme may be identified from real data.*
>
> *...The trouble is, that probabilities are not identifiable.*

Even the name of the scientific meeting, where Kalman's contribution was presented, could be interpreted as symptomatic—*Foundation Crisis in Econometrics within the Standard Statistical Paradigm.* As pointed out in [70], the criticism of current methodologies of data treatment has long ago left academia for the popular press, for instance in the Wall Street Journal [114]:

> *Fickle Forecasters. How Three Forecasters, After Crash, Revised Economic Predictions*

and [44]:

> *Into the Void: What Becomes of Data Sent Back From Space? Not a Lot as a Rule.*

We do not reject statistics because it is a "gedankenscience." The power of mathematics results from the fact, that it is a "gedankenscience," due to its independence from the facts of real life. However, the practical applicability of mathematical or statistical models goes outside the borders of a "gedankenscience." Many processes studied in physics are can be modeled by "gedanken-experiments" because useful models of their behavior are simple enough to be formulated by humans. We can come very close to describing the orbit of the earth relative to the sun using only Newton's gravitational principle and the masses and distances of the earth, sun, and moon. For most purposes, we can ignore the effects of other planets, other stars, and air disturbances due to (say) the flight of butterflies.

[2] *Der Gedanke...* the thought (in German). Appears frequently in natural sciences in the word *der Gedanken-experiment* —a thought experiment not really performed, but obeying an a priori given system of laws. In use without translation in many languages. The most popular application of such an approach was the A. Einstein's cosmic elevator used in the General Theory of Relativity.

[3] German word *der Ersatz* means a not quite perfect substitute, an artificial Christmas tree or a "hamburger" made from soy beans.

[4] A. N. Kolmogorov: Russian mathematician (1903-1997), who developed in the 1930's the most commonly accepted version of probability theory.

However, in economics, it is not simple to distinguish the perturbations of the data resulting from influences, that (if we knew, what they were) could be ignored, and the essential ones. It is impossible to discriminate from the flapping wings of a butterfly and the mass of the sun. Moreover, we have not yet identified anything that remotely corresponds to Newton's laws. Such principles, invariant for all time, may not even exist. Nothing is stationary and replicable in economics. One of the major issues is the independence of events; the collision of two gas particles at a specific point can be considered completely independent of a collision of particles at a distant point. Economic events not only are influenced by economic transactions, but also by seemingly unrelated activities across the globe, which may even cause a strong synchronous reaction throughout the world.

The impropriety of statistical applications to many propositions in real life is reflected by the manner in which many problems are stated. They begin with the assumption: *Let* $x_1, .., x_N$ *be the* $N-$ *tuple of i.i.d. random variables.* The idea of independence, as noted above, is probably unsuitable for all real events. *Identical distribution* refers to stationarity and repeatability, which is also a doubtful characteristic of many real data. However, the most discordant is the notion of *randomness.* This is pure agnosticism, a complete abdication of the notion, that the human mind has the ability to discern, confirm, and establish the cause of events.

Returning to economics: are the fluctuations of prices on the stock market random? Ask market experts this in a more specific way: "Was yesterday's change in company X's share price random?" The explanation, (or several explanations) received will suggest, that what occurred was a necessary consequence of having new public information about X's earnings or prospects, or a change in the discount rate by the Fed., etc. The change might seem random for those, who perceive the market only as a big roulette wheel. Often, no reason can be elicited, and the response is: "I have no idea;" (read, "I have no information,") rather than, "It was random." Corporate financial data in the form of various financial statements play an important role as 'raw materials' of economic and financial analysis. Market analysts and economists use data from multiple sources; even those, who are responsible for a single firm's planning and policies will examine information on the competition as well as the economy as a whole. To estimate the financial position of a company, an analyst uses data not only of the company under consideration, but also data of other companies. Using the language of statistics for a moment, we may say, that data from a *sample* of companies will be used.

Let us see how well the standard statistical assumptions are met in the case of financial statement analysis:

1. To evaluate the financial position of a company one needs to compare the company's parameters with those of "similar" enterprises. However, the number of really comparable companies is always very small. Finding companies quoted on the NYSE within a range (size, capitalization, etc.) in a specific industry will result not only in a strictly bounded set, but frequently in only a very small number of firms, often numbering not more than 10. Is it possible to accept the idea, that analysts are *randomly* choosing companies for their comparisons from an infinite population of mutually independent companies, for which the same statistical model describes their economic and other parameters? Such an idea is implausible for many reasons. To compare the comparable, the analyst's choice is **systematic**. Systematic choice is of course biased and the exact opposite of random.
2. The small size of a "sample" of comparable companies cannot be increased to raise the reliability of the analysis. It is easy to add to the number of trials, when throwing dice, but the idea of increasing the number of companies directly comparable to eg Coca-Cola is absurd.
3. To assume the independence of financial statement data for different companies, as is required by many statistical methods, would be even more unrealistic. Comparable companies may react in a similar way to changes in the economic environment (recession, taxes, custom duties, inflation, prices of raw materials and energy, technological innovations, etc.). This similarity of reactions forces the economic parameters of these companies to be mutually dependent.
4. Another problem is centered on the fact, that many statistical methods are based on the assumption, that the data fit a particular probability distribution. One of the most widely used statistical distributions is the *Gaussian* or *normal.* It is customarily chosen because many applications are based on mass random events, in which case its choice may be justified. This justification is based on the *Central Limit Theorem* (the Law of Large Numbers), which is the most effective weapon in the arsenal of mathematical statistics. However, there is no reason to expect, that what works for large data samples will also be applicable to small ones. Moreover, the fundamental differences between random events and the actual causes of real events infer, that the universal application of standard

statistical reasoning to real data is illegitimate. To demonstrate this, it is useful to recall the Central Limit Theorem ([110]):

Central Limit Theorem:

Given:

1. The random variable x has some distribution with mean μ_x and standard deviation σ_x.
2. Samples of size N are randomly selected from this population.

Conclusions:

1. The distribution of all possible sample means $\overline{x}$ will approach a *normal* distribution.
2. The *mean* of the sample means will be μ_x.
3. The *standard deviation* of the sample means will be $\sigma_x/\sqrt{(N)}$.

An (infinite) population is thus assumed, from which samples of size N are randomly selected. For samples of size N greater than 30, the sample means are approximated reasonably well by a normal distribution. However, if there are only 10 comparable enterprises, it will be impossible to obtain the necessary sample size, and adjustments will have to be made, each of which will degrade the outcome.

Moreover, some interpret the central limit theorem as being valid for any distribution of the population. This may be a crucial error because the theorem applies only to distributions, which have a **mean** and a **standard deviation.** Not all distributions fulfill this requirement, eg a Cauchy distribution has neither a mean nor a standard deviation, and it is sometimes used by statisticians to describe the effects of gross errors (*outliers*). Since outliers are rare in large samples, only a small portion of the data behave in this manner. However, if there are only 10 companies in a sample, a “small part” of these data may number one or two. Moreover, in various real data, outliers are generally present and they cannot be ignored. Is it possible to treat these unusual data points as if they have ‘normal’ statistical properties? On the other hand, if we accept a Cauchy distribution to characterize outliers, we are not able to apply the central limit theorem. How then can it be expected, that reliable results will be produced from such an analysis?

If data are not normally distributed, one may still compute any desired statistic (if it exists), but it is no longer possible to impute any *meaning* or significance (in the statistical sense) to these measures. This is applicable to several customary tests, for instance:

1. the Student's distribution (the t-distribution) for testing statistical hypotheses on population means,
2. the chi-square (or χ^2) distribution used for tests on population variances and for tests in curve fitting,
3. the Fisher's (or F-) distribution for testing hypotheses on ratios of variances and for variance analysis (ANOVA) in solving regression problems.

No longer being able to apply the concept of normality leads to the "illegal" application of these popular statistical tests. There are those, who try to escape these difficulties by using nonparametric statistical methods not relying on the normality assumption, however the danger posed by the mutual dependence of events may once again subvert their effort.

Another difficult problem is connected with the notion of "identically distributed." This idea refers to the *homogeneity* of the data sample: all data should be "of the same origin." This is rarely the case in practice. In the analysis of an industry, it is not unusual to find subgroups of companies, within which the members behave in a similar manner, and which manner is different from the behavior of members of other subgroups. This is noticeable in the probability density functions of economic parameters: instead of a single density maximum (typical for eg normal distributions) several local maxima appear—the density is *multimodal.* Such "mixtures" of differently distributed subsamples do not conform to the central limit theorem.

In fairness, it must be said, that a large number of tests for the above conditions have been developed and numerous data "treatments" have been unfolded in an attempt to mitigate these problems, and mathematical statistics has evolved as new approaches to problem solving have been developed: more robust statistical methods, Bayesian and recursive procedures, etc. These and similar innovations provide a little extra knowledge and sometimes a good bit of cleverness (as Benjamin Graham would have said), but taken as a whole, they do not go a very long way in overcoming the noted dilemmas. The use of these newer methodologies does not provide a complete solution of the problem.

Some new nonstatistical approaches may yield better results than statistical methods in a particular application, but they are not competitive with mathematical statistics as products of a scientific theory, which systematically covers a broad field of theoretical problems. The statistical paradigm is perfectly suited to explain the outcome of mathematical models for processes in many scientific fields, particularly those in the physical sciences (treating mass events in physics such as the movements of molecules of a gas or neutrons in the core of a nuclear reactor). As already discussed, the development of scientific thought, as new information is discovered, leads to modifications of theory, in which frequently the "old" processes remain as special cases. For instance, Einsteinian relativistic mechanics produces the same dynamic models as Newtonian mechanics, but only for an "asymptotic" case of extremely slow movements. However, most nonstatistical models of uncertainty do not fulfill the same function of "asymptotic consistency" with mathematical statistics.

It might seem, that fuzzy-set theory and the probability theory based on fuzzy sets provides a desirable "generalization" of the statistical paradigm, because a fuzzy set may be seen as a generalization of the classical notion of a set. However, there are other difficulties; a major one is, that the paradigm of fuzzy-set theory is based on the assumptions being the foundation stones of the theory:

1. The "membership function" is given, which determines the degree, to which each event belongs to the set.
2. The formulae of the fuzzy logic are given.

These necessary elements must be chosen a priori and subjectively. This could prove to be even more difficult than to establish an a priori statistical model of data.

Other nonstatistical methods suffer from the absence of a complete mathematical theory and have only a heuristic nature.

There exists a more radical solution: to depart entirely from the statistical environment and the existing paradigms of uncertainty, and to try something entirely new.

4.4 Paradigm of Gnostics

4.4.1 Gnostic System

The semantics of this unusual word is treated below; we use this notion extensively, but the word itself is not our creation, it has an extremely long and exciting history. It is no wonder, that this word has been redefined and applied to a really recent problem: a cybernetic recognition system, by modern philosophers [88]. In a general setting, the gnostic system can be characterized as the pairing of an object and of a subject (observer), whose task is to recognize the object. The observer gets his information by means of an *object* → *subject* channel. To compare his improving knowledge (through repeated observations) about the object's reality he uses another channel (feedback *subject* → *object*). The recognition process is thus active. A typical example is the evolution of experience as an iterative cycle *initiating an action—gaining experience from the action—evaluating and accumulating experience—using the updated experience to initiate a new action—etc.*

For a gnostic recognition system, a more specialized formulation is needed. It is sufficient to restrict our exposure to the *gnostic system of quantitative recognition.* The idea of such a system is presented in Fig. 4.1.

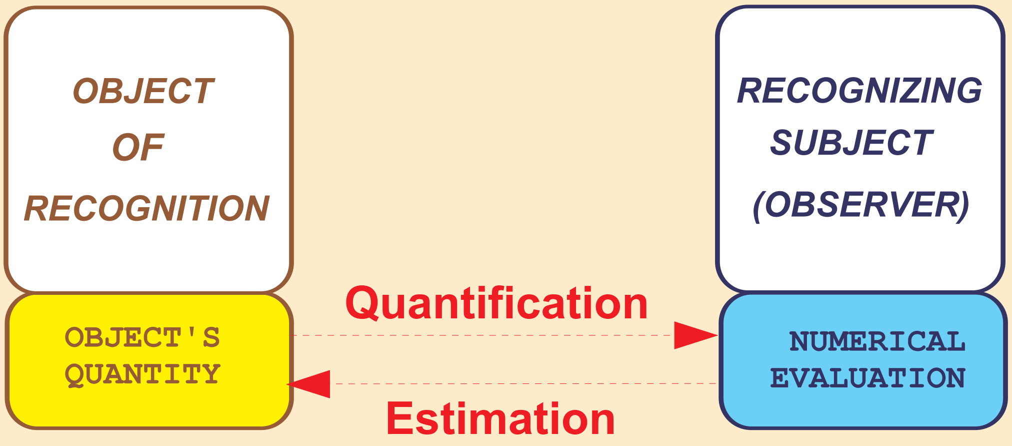
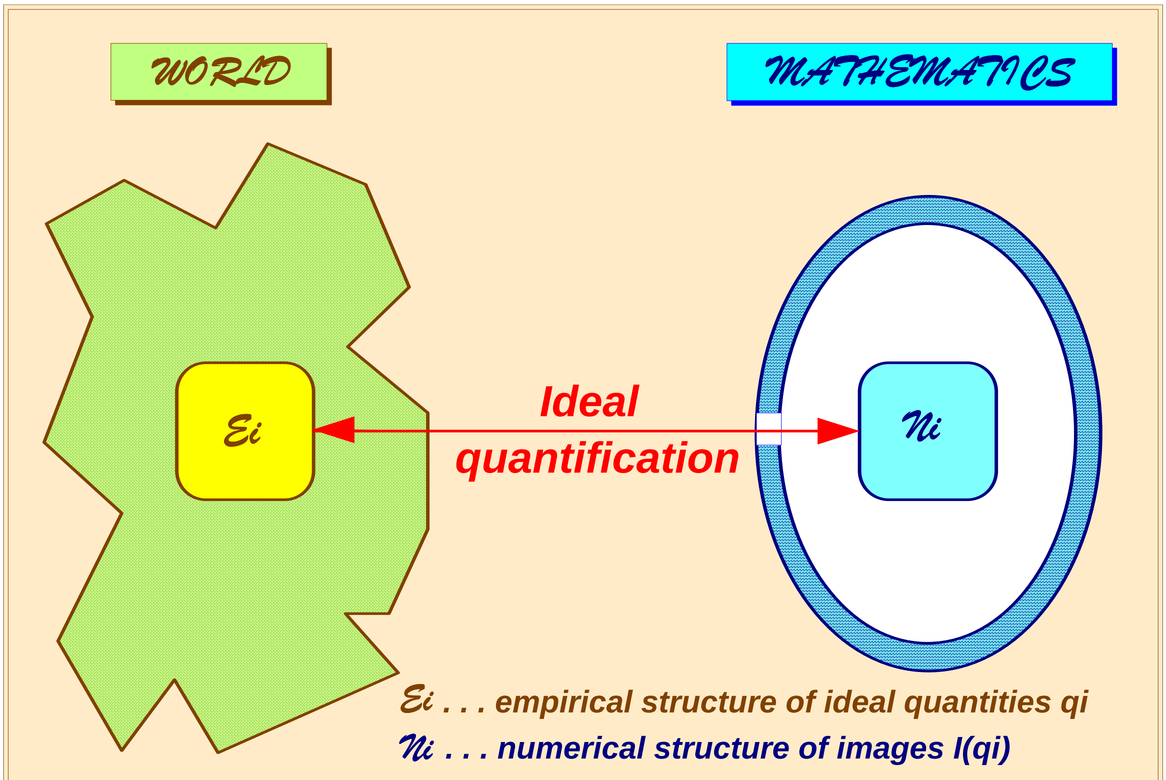


Fig. 4.1 Gnostic system for quantitative recognition

There exists an object with quantitative parameters, which should be recognized by the subject/observer. Recognition is reduced to the object's quantification. As seen previously, quantification is the mapping of a structure of quantities onto a structure of numbers (data). This is the feedforward link *quantity* → *number* in Fig. 4.1. The basic task of the observer is *to estimate* the true (*ideal*) value from the data. This is represented in Fig. 4.1 as the feedback link *number* → *quantity*. The estimating phase of the cognition cycle is required because of the inevitable disturbances to quantification caused by uncertainty. We shall examine the problems of estimation later. The present point of interest is still quantification.

We learned in Chapter 1, that quantification as a mapping is consistent if some simple rules hold. Using mathematical language, we characterized structures of both quantities and their numerical images as additive or multiplicative groups. This scheme of quantification corresponds to that of measurement theory. We can summarize our notion of ideal quantification by Fig. 4.2.



Ei . . . empirical structure of ideal quantities qi

Ni . . . numerical structure of images I(qi)

Fig. 4.2: Ideal quantification

Anywhere in our world, an empirical structure $\mathcal{E}_i$ of ideal quantities qi exists mapped onto the numerical structure $\mathcal{N}_i$ of numbers denoted by $I(qi)$—images of qi. We already know the mathematical nature of both structures: they are both commutative groups isomorphic with Abel's additive group.

This idea of ideal quantification is suitable for measurement theory, but it has only a limited use for application to real measurements. There is consensus among those, who develop and use measurement theory, to limit their studies to precise measurements and to let statisticians deal with imprecisions. We are not going to follow this scheme:

1. Such schemes are useful for applications, where data precision is—as a rule—high, and imprecise data are a rare exception. We saw in Chapter 2, that the opposite is frequently true in practise: data can contain gross errors and precise measurements, as a practical matter, can be impossible.
2. The unreliable results of data treatment in many applications lead to incorrect decisions, the outcomes of which are too costly to ignore.
3. It is a normal use for statistics to model and treat data having a statistical character; however, as noted earlier, it is unlikely that all real data are statistical in nature, and that they can be modeled using statistical methods. Statistics, as a strongly mathematical discipline, should not be used to work with nonstatistical data.
4. The main obstacle in the path of consistent successful use of statistics in many fields, especially in economics, is the specific nature of the uncertainty, which surrounds real events. They are not always some consequences of "random" factors, which is a necessary condition for "good" statistical data.
5. There is a natural and direct way to extend the useful axioms of measurement theory to the quantification of uncertain data. This is the primary focus of the theory that will be developed.

The approach we are going to use is based on the *gnostic theory of uncertain data* (on *mathematical gnostics*) or—in short—on *gnostics.* The ideas imbedded in the mathematical gnostics were first introduced in the literature in 1984 in three papers [56], [57] and [58]. A summary [59] of the new theory was presented at the IX-th World Congress IFAC (International Federation of Automatic Control). Scientific interest in this contribution lead to an invitation to publish an extended review of the theory [60] in the official journal of the IFAC. A more complete exposition of the development of gnostics through 1990 is in [61].

The word "gnostic", has already been used and it is therefore time to look at its semantic roots.

4.4.2 Gnostics, the Choice of a Name

Every meaningful concept should have a name. There has been a long and successful experience in using gnostic algorithms, showing their superiority over other approaches under practical non-laboratory (real world) conditions. This methodology, then, should be distinguishable from other analytical concepts and procedures. In addition, the uniqueness of this theory and its algorithmic applications place it outside of the framework of other known methodologies, and the name should distinguish this concept from all of the others.

A more interesting question is "Why just **this** name?" The birth of this idea comes from the very old Greek word *gnosis*, which can be translated as "knowledge" or "art of knowing". Its root is found in many newer words such as prognosis, diagnosis, physiognomy, gnome, gnomon, gnoseology and—via the Latin reflection of Greek (*gnoscere* = know)—cognition, cognizable, cognizance, cognizant, recognition, recognizance, recognize, etc. All these words are in a way connected with knowledge. However, three words from this family deserve special comment: *agnostic, Gnosticism, gnostic.*

The word Gnosticism is used ([112]) to define a system of mystical doctrines combining early Christian, Greek, and Oriental philosophies. *Gnostic* means 1) "of or having knowledge," 2) "believer in Gnosticism." These definitions are rather formal, however the first use of the word was Plato's (427–347 B.C.), while the second is connected with a religious movement and is about five centuries younger. Gnostic philosophers sought the truth outside of the official Gospels and were sternly rejected by the Church authorities and the Gnostic movement was considered a great danger and was repressed. Examples of recent words using this ancient kernel show, that there is no religious connotation in the modern usage of the notion of "gnosis"; the accent is on knowing and knowledge as in the time of Plato. This will also be true here.

The word *agnostic* is in a way closer to our purpose. According to [112], an agnostic is one, who thinks it is impossible to know or learn, whether there is a God, or if anything exists beyond material phenomena. Now applying this definition to the world of data and substituting the "true (ideal) value" of a datum[5] instead of God, someone could be called "data agnostic" if he or she thinks it is impossible to know if there *is* a "true value" beyond the observed data, and believes, that it is impossible to

[5] "Datum" is the singular of "data": a data item.

approach (to estimate) the true value from the data itself. In contrast, we shall not only profess the existence of "something like a true or fair value", but we shall also look for (and find) the best estimating methods, with which this value can be approximated. By doing this, we shall attempt at all times to be "non-agnostic," ie to be **gnostic**.

4.4.3 The Framework of Gnostics

The overall breadth of **gnostics** will encompass the following elements:

1. A specific mathematical theory called *The gnostic theory of uncertain data.*
2. Data treatment algorithms based on the gnostic theory—gnostic algorithms.
3. The applications of gnostic algorithms to solve different types of problems.

By including applications in the framework to be considered, it is intended to show, that gnostics has been developed to serve practical needs inspired by close contact with specific problems. The need to provide applications to solve these real world problems provides a rich source of inspiration and motivation for the theory and for the development of tools to test and verify its suitability as well. It will be seen below, that these gnostic applications have very specific properties.

Each mathematical theory defines the necessary notions, which permit the formulation of the assumptions or axioms of the theory. Applying consistent judgment, mathematicians develop and prove all the statements of the theory from the assumptions. This means, that all the cleverness and fruitfulness of a theory is hidden or rooted in its assumptions. A mathematician "only" makes them obvious by using the exact apparatus of mathematical reasoning. This is why it is important to understand the main ideas, from which the axioms are derived. With respect to gnostics, these principal ideas can be summarized as follows:

1. Data are not arbitrary numbers, but products (outputs) of a highly developed technology called quantification. As such, they obey strict regularities.
2. A real datum is the pairing of an *informative (ideal)* and an *uncertain* component, which disturbs the observation.
3. The uncertain component is a result of our lack of knowledge of the causes of discrepancy between the observed datum and its informative

component. There is no randomness involved.

4. The observed datum is a one-dimensional "projection" of the pair containing these two components; the direct determination of the ideal component is thus impossible.
5. The shape of the quantification process is a special path, along which the image of the datum moves under the influence of the uncertain data component. This path is thus a geometric model of quantification.
6. Neither the metric nor the geometry of the data space is an a priori assumption, everything is determined by the data following the principle, "Let the data speak for themselves."
7. By analyzing the path of the data pair, one can develop a complete theory of individual uncertain data.
8. The gnostic theory of data samples is developed from the theory of individual uncertain data by applying a aggregation axiom motivated by a close relationship to relativistic physics.

The thrust of gnostics is, that in contrast to statistics, gnostics constructs its theory of data samples based on its own theory of individual data; it is finite, not based on samples randomly drawn from an infinite population.

4.5 Randomness versus Data Uncertainty

The idea of *randomness* in discussing the disturbance of data pairs will not be used for the following reasons:

1. Random experiments, random events, random variables and random functions have precise mathematical definitions in mathematical statistics. We introduce a different notion, that of *uncertainty.* Different things cannot use the same name.
2. In spite of its importance in mathematics, the notion of randomness rarely has been given a clear interpretation in statistical literature. For example, [16]:

   > *"Obviously, it is impossible to define precisely, what is understood by the word 'random'. Its sense can be clarified best by means of examples."*

3. The statistical notion of randomness is inherently connected to the idea of the unlimited repeatability of a random experiment. Gnostics considers finite data samples under no assumption of unlimited

extensions to the number of data.

Due to the principal ideas of gnostics, which have been exposed above, there is no "magic" in the gnostic interpretation of the disturbance component of a datum:

Data uncertainty comes from the observer's lack of knowledge of the given quantification process on the observer's part.

There are three important aspects to this statement:

- **Objectivity:** The uncertainty of the quantification process is given, it is a part of the real world.
- **Knowledge:** If the observer knew perfectly all of the conditions of the quantification and the factors influencing each datum's value, he could completely explain all the details of the data values. Uncertainty is the unexplained portion of the imperfectly clarified data.
- **Subjectivity:** The ability to explain a data value depends on the subject (observer), and on his particular knowledge. What is uncertain for one observer, can be quite precise for another.

If this is true, then the same—objectively given—piece of data can be deaggregated into different ideal and uncertain components by different observers. This individuality in the points of view of the observers is yet again another reason to search for a law of Nature, that explains quantification, and which does not depend on the individuality and capability of the observer for its outcome.

4.6 Summary

In spite of the commendable praise, which statistics deserves for solving problems in both scientific and practical application fields, experience does not support the universal application of methods, which are based on the paradigm of mathematical statistics. Most of the important problems of economics attempts to predict the market's behavior, financial statement analysis across an industry, and other similar activities remain practically unsolvable by statistical methods. There are other fields of practice, where availability of data is limited because of difficulty and/or high costs of measurements like in environmental control, geology, medicine, reliability and endurance assessment. Other obstacles can be caused by non-stationarity and a high speed of processes.

There also exist serious theoretical objections to modeling such processes by statistical methods. Both theoretical and practical needs have given rise to a number of alternative, nonstatistical methods based on different paradigms of uncertainty. One of the new paradigms of uncertainty—the gnostic one—is represented by a set of simple assumptions, according to which uncertain data are the products of a quantification process. The uncertainty of a datum is the result of a lack of information and not due to a random event. The first goal of gnostic theory is to derive a mathematical model of uncertainty for individual data from this paradigm.

Chapter 5

Model of Uncertain Quantification

5.1 Model of Uncertain Data Component

In Chapter 1, where the process was defined, the material (real) nature of quantification was emphasized. Its mapping of a structure not only involves real elements, but also interactions between objects being quantified and elements and forces belonging to other structures. The results of these interactions manifest themselves—if they are not explained—as uncertain components in the observed data. This means, that the uncertain components have the same nature (and can be expressed using the same measuring unit) as the ideal component. If we quantify the value of an asset, then the nature of an observed datum is money and the numeraire is the dollar; the same units are valid for both the ideal and the uncertain components of the observed datum. All three quantities (ideal, uncertain, and quantified components) thus have the same financial nature. When the concentration of a dangerous pollutant in a river is to be quantified, the disturbations of the observed quantity are of the same nature, concentration. If the data structure is additive, then the influence of the uncertain components is also additive. In the case of multiplicative data the interaction of both components of the data pairs is also multiplicative.

We can now proceed one important step further in the analysis of the features of a structure of uncertainties. It has already been noted, that the interactions of uncertain/ideal components are the same as ideal/ideal interactions. To be consistent, we recognize the same rule for uncertain/uncertain interactions. Consider a particular case: the structure of cash flows. We have already accepted the idea, that this structure (when it is quantified without the influence of uncertainty) can be modeled by the multiplicative group. However, the current price of each asset (represented by the item of the cash flow) is a product of its ideal (historical, original, nominal) value and of the current value of a factor representing inflation (**inflator**). The current prices of assets are thus data pairs having the form ⟨ *ideal price, inflator* ⟩. Observations of the data are one-dimensional, only the product of the two factors can be observed. Each inflator, in itself, can be a product of factors characterizing different specific influences. Daily price changes produce an inflator for each day. A change occurring over two days can be then evaluated by the product of each daily inflator. This product is commutative. Products of three inflators are associative. Deflation can also occur, so a reciprocal ("inverse") value can exist. All inflators are finite and positive and the zero inflator does not exist. Thus it can be concluded, that the structure of inflators may be also modeled by the multiplicative group.

We now arrive at a reasonably general model:

The structure of uncertainties is a commutative group.

5.2 Algebraic Model of Real Quantification

Both the objects and the results of an ideal quantification have been represented by algebraic structures isomorphic with the Abelian group. The same model was then introduced to represent a structure of uncertainties. This structure of uncertainties can be quantified in the same manner. It is also true, that the respective positions of the "ideal" and the "uncertain" components of a data pair can be interchanged:

Imagine the case of a tourist visiting a foreign country. As he exchanges money, he receives a sum of local banknote and coins. He believes, that an "ideal" value for this amount of money exists, but not knowing the current value of the inflator, he is not able to quantify it. For him, the structure of inflators is a structure of uncertainties. A second observer's task is to analyze the current inflation factors. As he accumulates a set of prices for comparable products, he considers the inflator as an "ideal" value to be estimated. His data are "disturbed" by the variability of the historical (deflated) prices of these mutually comparable products. Hence, for him, the group of uncertainties is the structure of unknown historical prices[1]. This interchangeability between the roles of the data pair's components will be called *symmetry*.

We have arrived at a two-dimensional structure, which could be mathematically modeled as the Cartesian product of two groups. The interactions between elements of these two groups have the same character as those within the groups: all interactions are either multiplicative or additive. This notion of real quantification composed of the pair of ideal quantification processes can be illustrated by Fig. 5.1.

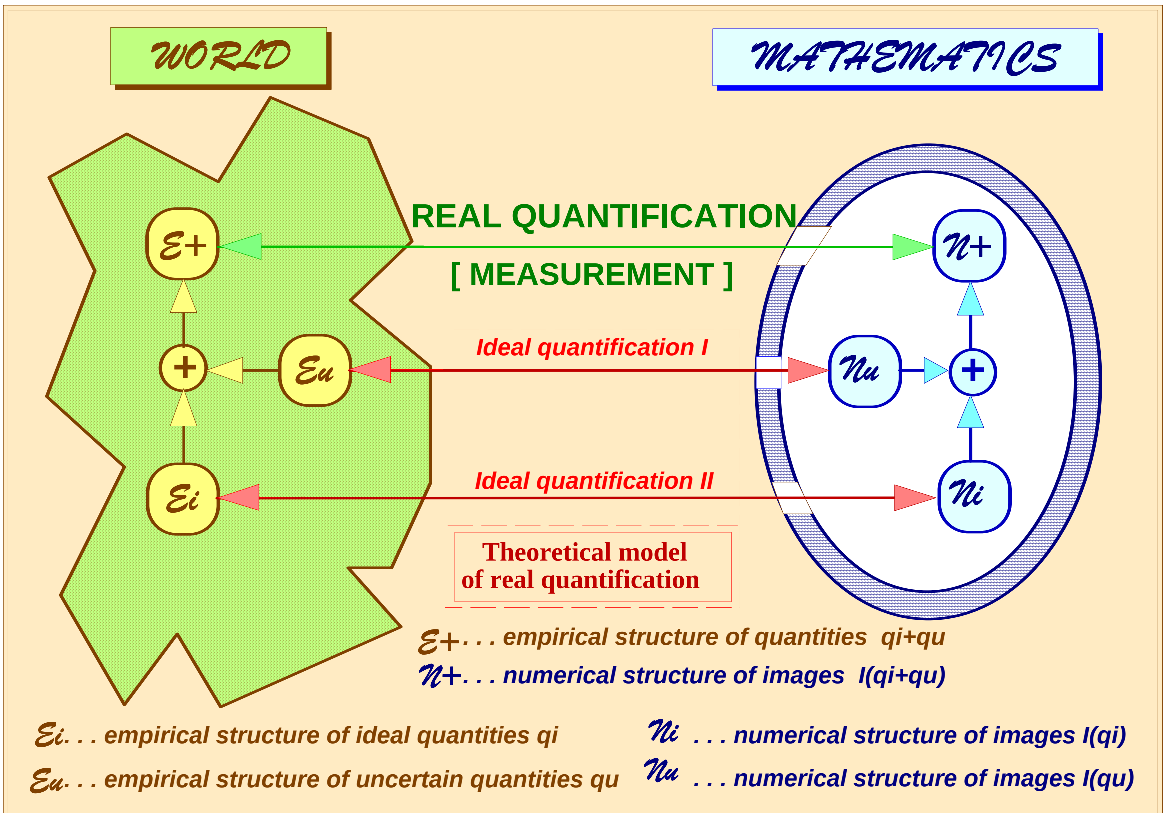

WORLD
MATHEMATICS
REAL QUANTIFICATION
[MEASUREMENT]
Ideal quantification I
Ideal quantification II
Theoretical model of real quantification
E+
Eu
Ei
+
N+
Nu
Ni
+
E+. . . empirical structure of quantities qi+qu
N+. . . numerical structure of images I(qi+qu)
Ei. . . empirical structure of ideal quantities qi
Eu. . . empirical structure of uncertain quantities qu
Ni . . . numerical structure of images I(qi)
Nu . . . numerical structure of images I(qu)

Fig. 5.1 Real quantification

Elements of a group Ei of ideal (true) empirical (real) quantities are denoted by qi and their numerical images, (which form the numeric group Ni) by $I(qi)$. (The operator $I(...)$ symbolizes the ideal quantification). There also is a group Eu of empirical (real) uncertainties qu and a group $E+$ of compositions of ideal and uncertain quantities. The result of the real quantification (the observed data) are neither group Ni of images of ideal quantities, nor group Nu of images of uncertainties, but the group $N+$

[1] To view prices as elements of the multiplicative group, one applies the same notion of multipliers as in the case of measuring ratios: a good's price says how many times the price exceeds the unit of currency.

of quantities $I(qi + qu)$ obtained by the sums of images of the composed elements of both empirical groups. (The composition operator is denoted by the symbol $+$, which can be interpreted here generally as $\oplus$ as well as $\otimes$, or as inverse operations). Each of the three empirical structures can be (at least theoretically) subjected individually to ideal quantification. In this way, three views of the data are obtained: theoretical images $I(qi)$ and $I(qu)$ and the actually observed numeric images $I(qi+qu) = N(qi)+N(qu)$ of the composition (sum) of numeric images of components qi and qu. The real quantification is thus modeled as an ideal quantification of the composition of two (unidimensional) quantifications. Problem to be solved by data treatment consists in estimation of the $N(qi)$ (and $N(qu)$) from the composed pair of images.

This concept can be now described more formally (as in [61]) using the following notation:

Definition 1: Let $\mathcal{A}_0$, $\mathcal{A}$, and $\mathcal{N}$ be nonempty sets. Elements of the set $\mathcal{A}_0$ are the *ideal values*, while elements of $\mathcal{A}$ are the *data;* and those of $\mathcal{N}$ the *uncertainties.* Let $\mathcal{R}^1$ be the set of real numbers and $\mathcal{R}_+$ the set of positive real numbers. Now introduce the following mappings:

$$\upsilon : \mathcal{A}_0 \longrightarrow \mathcal{R}^1 \tag{5.1}$$

$$\vartheta : \mathcal{A} \longrightarrow \mathcal{R}^1 \tag{5.2}$$

$$\nu : \mathcal{N} \longrightarrow \mathcal{R}^1 \tag{5.3}$$

$$\sigma : \mathcal{N} \times \mathcal{N} \longrightarrow \mathcal{N} \tag{5.4}$$

$$\pi : \mathcal{A}_0 \times \mathcal{N} \longrightarrow \mathcal{A}. \tag{5.5}$$

The mapping υ as well as the mapping ν are *ideal quantification*, while ϑ is *real quantification.* The domain of definition of a mapping ξ will be denoted $Dom(\xi)$ and its range of values $Ran(\xi)$.

Instead of using this model in its full generality, only a special case delimited by the following axiom will be considered:

Axiom 1 (axiom of the additive model of possible data):

A1.1: Mappings υ, ν and ϑ are one-to-one and the range of values $Ran(\upsilon)$ is identical to $\mathcal{R}^1$.

A1.2: Mapping ν is an isomorphism of the structure

$$\mathcal{G}_\sigma := \langle \mathcal{N}, \sigma \rangle \tag{5.6}$$

and of the additive group $\langle \mathcal{R}^1, \ +\rangle$.

A1.3: There exists $S \in \mathcal{R}_+$ such that for all $a_0 \in \mathcal{A}_0$ and for all $n \in \mathcal{N}$ the following relation holds:

$$\vartheta(\pi(a_0, n)) = \upsilon(a_0) + S\nu(n). \tag{5.7}$$

The structure $\mathcal{N}$ is isomorphic with the additive group according to A.1.2. It might seem, that the assumptions related to the group features of the set $\mathcal{A}_0$ have disappeared from our definitions because there is no composition operation defined over this set. However, this is not true because of the already noted interchangeability of the ideal and the uncertain components, which can be achieved by renaming the two components. On the other hand, we need only one (fixed) ideal quantity to model the effects of uncertainty on a datum. The ideal quantity to be quantified is certain, but for the observer, it is an unknown value. The uncertain component also has a fixed (but unknown) value; our goal is to explore the consequences as the uncertain value changes. More specifically, the objective (given a fixed, but unknown, ideal value) is to analyze the path of a geometric movement of the datum-pair driven by an uncertain component.

The main ideas of the axioms A1.1–A1.3 may be thus interpreted in the following way:

1. Both uncertainties and ideal values can be considered members of commutative groups.
2. The theoretical model of both uncertainties and ideal values is that of ideal quantification.
3. The *theoretical* model of a piece of uncertain data (of **real** quantification) consists of **two** theoretical models of **ideal** quantification. The *geometrical* model of a piece of real data is thus a point on a **plane**; it is two-dimensional.
4. The *actually observed* piece of uncertain data is a **single** number, it is one-dimensional.
5. The *actually observed* piece of data is formed by composition of the outputs of both ideal quantifications ($I(qi \oplus qu)$), the composition law between the two groups of outputs being the same as that between elements of each of the groups.

The fundamental role is thus played by the assumption, that the ideal quantification is an isomorphic mapping of a **commutative group** of real quantities (domain of the ideal quantification) onto the commutative group of real numbers (range of values of the ideal quantification). One can base such an assumption on practical experience—as was the case of the group of cash flows or of the group of measuring multipliers. However, there is a much more reliable reasoning, that derived from measurement theory, which condenses the experience of thousands years in the form of its axioms. Based on the axioms of measurement theory , the assumption that a suitable mathematical model of ideal quantification is a mapping of a commutative group of real quantities was established in [56] as the basic axiom of gnostic theory. As shown later in [99], this assumption can be supported by a strict mathematical reasoning, which shows that the commutative group is an acceptable model for an empirical structure of quantities within the framework of measurement theory, as set out in [87]. This is the sense of the statement, that the first gnostic axiom is supported by measurement theory.

From **A1.2** it is seen, that

$$(\forall n_1,\ n_2 \in \mathcal{N})(\nu(\sigma(n_1,\ n_2)) = \nu(n_1) + \nu(n_2)). \tag{5.8}$$

Using notation

$$A := \vartheta(\pi(a_0, n)), \qquad A_0 := \upsilon(a_0), \qquad \Phi := \nu(n). \tag{5.9}$$

we recast **A1.1** and **A1.3** in the form of a relation

$$A = A_0 + S\Phi \quad (A,\ A_0,\ \Phi \in \mathcal{R}^1,\ S \in \mathcal{R}_+), \tag{5.10}$$

called the *additive form of possible data.* Data A will be called *additive,* and the positive number S is the *scale parameter.* The scale parameter will be introduced, when data samples are analyzed to recognize, that successive data elements may have different volatilities. The scale parameter thus unifies the measuring units to evaluate the intensity of uncertainties.

Definition 2:
Multiplicative data will be additive data transformed using the following relation:

$$Z := \exp(A). \tag{5.11}$$

The choice of this transformation is natural. It is defined by the one-to-one function; the domain $Dom(\exp(*))$ coincides with that of the additive data ($\mathcal{R}^1$). The range of values $Ran(\exp(*))$ is the same as that of multiplicative data ($\mathcal{R}_+$). It is seen from (5.10) and (5.11), that

$$Z = Z_0 \exp(S\Phi), \tag{5.12}$$

where

$$Z_0 = \exp(A_0). \tag{5.13}$$

The expression (5.12) is called the *multiplicative form of possible data.*

This and the several following chapters develop the gnostic theory of an individual datum. The scale parameter S, introduced to take into account different volatility for different data, remains constant, when only one datum is being considered. To simplify the formulae, we shall therefore temporarily assume, that $S = 1$ using a unified scale, which provides a simplified equivalence

$$Z = Z_0 \exp(\Phi). \tag{5.14}$$

When the solution of problems requires non-unity scale parameters, the general multiplicative form (5.12) and its derivatives will be used.

5.3 Realism in Data Models

The customary methodology used to interpret (statistical) data relies on premises, which are extremely difficult to justify in the treatment of some real data. The primary thrust of the methodology being developed here is to overcome these problems.

5.3.1 Statistical Data Models

The first problem concerns the very nature of the data. A data form analogous to the additive form 5.10 is frequently used as a basis for statistical analysis, but only after assuming, that the data behave in a specific manner (defined by an a priori chosen statistical model).

As discussed in the previous chapter, such a model—as a rule—includes the following assumptions:

1. Data are obtained by sampling from a large population of random events.

2. The sampling is purely random and independent (ie the probability of being chosen is equal for all members of the population—sampling with replacement—and not influenced by the results of preceding choices).
3. All events (ie included members) of the population are identically distributed.
4. A "large" population means, that increasing the number of data in the sample is actually possible and theoretically reasonable. Moreover, a "large" population is assumed to be suitable for consideration in limiting cases, which have an infinite number of data.

The conclusion to be drawn from the above is, that (at least) for some applications, one can neither rely on the a priori assumptions of a statistical data model nor verify the realism of these assumptions if such a model is used.

5.3.2 Gnostic Data Models

It is worth restating here, that the additive data form described previously was motivated by measurement theory. The choice of such a point of departure for gnostics is advantageous:

1. Measurement theory summarizes and generalizes technology of precise quantification. There has been a much longer period of time to verify and refine the rules of quantification than has been possible in the case of statistics.
2. Axioms of measurement theory are much more elementary than those of statistics. They have a purely algebraic nature without the introduction of "magic" and complex ideas of randomness and statistical independence. This is why the range of applicability of these axioms is much broader and more universal than those of statistics.
3. The validity of the axioms of measurement theory (as they have been summarized in the form of the group model of ideal quantification) can be verified. It is a simple matter to establish experiments to test (finite) data structure features such as closedness, associativity, commutativity, the existence of the neutral element and invertibility (the validity of (1.1)–(1.5) or (1.13)–(1.17)). The slightly more complex assumptions of Axiom 1 could also be tested experimentally.
4. The gnostic paradigm does not contradict classical statistics. Gnostic theory simply extends the field to (small) finite samples of data, which

can not have a statistical model.

5. There are strong scientific arguments, which support the use of gnostic axioms: it will be shown, that this theory bears a close relationship to recent developments in several important branches of science: measurement theory, geometry, physics and information theory. Further, gnostics establishes special conditions, under which both gnostic and statistical methods provide identical results; this occurs if and only if the uncertainty is very weak.

Each theory, that is proposed, should include the means for its verification and/or rejection, and establish the limits of its applicability. All the implications of a mathematical theory are defined by its axioms and other assumptions. The more realistic these roots, the greater the usefulness of the theory in its application to real problems. From a practical point of view, the broader range of applicability, the more universal the theory becomes. Successful applications are then also important tools for validation.

In the context of the universality of mathematical models, a comment should be made regarding the perception of a limitation of the gnostic data model. Both data forms 5.10 and 5.12 using the two models of ideal quantification are basic in the gnostic theory. However, this does not imply unsuitability of the theory for other data forms. Numbers obtained from multiplicative or additive data using a suitable transformation can also be called data, but the transformation, which is applied, must always be specified. An important notion connected to data transformation is *data support*—the domain, over which data values are defined. We have so far considered only two basic types of data support: $\mathcal{R}^1$ and $\mathcal{R}_+$—the infinite ones. However, in practice, most data sets have finite limits $\longrightarrow$ bounded data supports. Bounded data can also be treated by gnostic methods after they have been properly transformed onto the infinite data support $\mathcal{R}_+$, for which the theory was originally developed.

The premise, which has been proposed in this section, infers, that the thrust of gnostics is to provide a reliable methodology for the treatment of small samples of "bad" real data. It should therefore not escape the reader's attention, that these axioms set out a sufficient foundation for the establishment of a unique theory of **individual** uncertain data.

5.4 Matrix Model of Real Quantification

Using the same symbols Z_0 and Φ as shown in (5.9)–(5.14) we introduce variables

$$x := Z_0 \cosh(\Phi), \qquad y := Z_0 \sinh(\Phi), \tag{5.15}$$

a matrix

$$\underline{M}(A_0, \Phi) := \begin{pmatrix} x & y \\ y & x \end{pmatrix}, \tag{5.16}$$

and a set of matrices using Definition 1:

$$\mathcal{M} := \{\underline{M}(A_0, \Phi) | A_0 = \upsilon(a_0),\ \Phi = \nu(n),\ a_0 \in \mathcal{A}_0,\ n \in \mathcal{N}\}. \tag{5.17}$$

Matrix products will be written in the usual manner, and the symbol of matrix multiplication ($\otimes$) will be shown only if necessary. Such a special case follows:

Theorem 1:

Let

$$\mathcal{S}_m := \langle \mathcal{M}, \otimes \rangle \tag{5.18}$$

be a structure. Denote sets $\{A_0\} := \{A_0 | A_0 = \upsilon(a_0),\ a_0 \in \mathcal{A}_0\}$ and $\{\Phi\} := \{\Phi | \Phi = \nu(n),\ n \in \mathcal{N}\}$.

It then holds, that structure (5.18) is isomorphic with the direct product of commutative groups $\langle \{A_0\}, + \rangle$ and $\langle \{\Phi\}, + \rangle$.

In order not to overburden the reader, only an outline of proofs is presented.

Outline of the proof:

- Structures $\langle \{A_0\}, + \rangle$ and $\langle \{\Phi\}, + \rangle$ are isomorphic with the additive group by the definition of ideal quantification.
- Matrix (5.16) can be written as a commutative matrix product

$$\underline{M}(A_0, \Phi) = \underline{M}(A_0, 0) \otimes \underline{M}(0, \Phi), \tag{5.19}$$

where

$$\underline{M}(A_0, 0) = \begin{pmatrix} Z_0 & 0 \\ 0 & Z_0 \end{pmatrix} \tag{5.20}$$

and

$$\underline{M}(0,\Phi) = \left(\begin{array}{cc} \cosh(\Phi) & \sinh(\Phi) \\ \sinh(\Phi) & \cosh(\Phi) \end{array} \right). \tag{5.21}$$

- Show, that both structures $\langle\{\underline{M}(A_0,0)\},\otimes\rangle$ and $\langle\{\underline{M}(0,\Phi)\},\otimes\rangle$ are commutative multiplicative groups.
- Show isomorphism of the structure $\langle\{\underline{M}(A_0,0)\},\otimes\rangle$ with $\langle\{A_0\},+\rangle$.
- Show isomorphism of the structure $\langle\{\underline{M}(0,\Phi)\},\otimes\rangle$ with $\langle\{\Phi\},+\rangle$.
- There exists only one common element to both groups of matrices $\underline{M}(A_0,0)$ and $\underline{M}(0,\Phi)$, namely the matrix unit $\underline{M}(0,0)$.

Therefore, the matrix structure $\mathcal{S}_m$ is also a model of uncertain data. The uniqueness of this matrix data model is examined below.

5.5 Data Uncertainty as an Operator

Consider the matrix structure (5.18) denoting elements of the matrices $\{\underline{M}(A_0,\Phi)\}$ placed in the $i-$th row and the $j-$th column as $m_{i,j}$. All matrices of this type satisfying the condition $|\Phi| < \infty$ have several special properties:

1. It holds for normalized elements of all matrices of this type, that
$$x/Z_0 + y/Z_0 = 1/(x/Z_0 - y/Z_0) = \exp(\Phi). \tag{5.22}$$
2. The determinant
$$Det\{\underline{M}(A_0,\Phi)\} = x^2 - y^2 = Z_0^2 \tag{5.23}$$
is constant for each fixed $Z_0 \in \mathcal{R}_+$ and for **all** real values of the uncertain parameter Φ.
3. Double symmetry:
$$m_{1,1} = m_{2,2} = x, \quad m_{1,2} = m_{2,1} = y, \tag{5.24}$$
where $x \in \mathcal{R}_+$ and $y \in \mathcal{R}^1$.
4. Introducing a special symmetric matrix
$$\underline{T} := \left(\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right), \tag{5.25}$$
one obtains
$$(\forall \underline{M} \in \mathcal{S}_m)(\underline{TM} = \underline{MT}). \tag{5.26}$$

The interpretation of these easily verifiable relations starts from (5.19). When considering a given (fixed) ideal value Z_0 under influence of different uncertainties (ie different values of the parameter Φ), the fixed matrix $\underline{M}(A_0, 0)$ is mapped onto the structure $\mathcal{S}_m$. Matrices $\underline{M}(0, \Phi)$ are playing the role of *linear operators* in these mappings. The relations (5.22)–(5.25) specify interesting properties of these operators, which (as we already know) are elements of the multiplicative group.

Relation (5.22) written as a pair of equations uniquely yields (5.15). Taking into account the double symmetry (5.24), one concludes, that the matrix operator is uniquely determined by these relations.

Relation (5.23) says exactly the same thing as (5.22), but its form leads to another interesting interpretation. The equality (5.19) defines a linear transformation with the matrix $\underline{M}(0, \Phi)$ as the operator. Imagine the parameter Φ continuously changing from zero to a nonzero value Φ', while Z_0 is constant. Points $\langle x, y\rangle$ (5.15) would then "move" along a continuous line within a real plane, describing a path, which is a geometric representation of the quantification process. (This is the *geometric—virtual—movement*). The point is, that (according to (5.23)) the square of the ideal value, Z_0^2, does not change as Φ varies. This means, that **the ideal value of a data pair is the invariant to the quantification process**. The equality (5.23) together with the initial and finite values of the parameter Φ thus defines the geometric path depicting the quantification.

The feature (5.26) of the matrix operator $\underline{M}(0, \Phi)$ can be called *commutativity with respect to transposition.* We ordinarily understand by the "replacement of coordinates" of a plane, that a transformation of variables $\langle x, y\rangle$ to some function $(\langle f(x, y), g(x, y)\rangle)$ has taken place. Introducing a different notion, the *exchange of coordinates* for the special replacement of the type $\langle x, y\rangle \longrightarrow \langle y, x\rangle$, it is possible to interpret (5.26) as invariance of the matrix $\underline{M}$ with respect to the exchange of the definition $Dom(\underline{M})$ and $Ran(\underline{M})$ (The matrix $(\underline{M})$ is considered here as an operator). In an explicit form, using the well-known formulae of hyperbolic functions

$$\cosh(\Phi_1 + \Phi_2) = \cosh(\Phi_1)\cosh(\Phi_2) + \sinh(\Phi_1)\sinh(\Phi_2) \tag{5.27}$$

and

$$\sinh(\Phi_1 + \Phi_2) = \cosh(\Phi_1)\sinh(\Phi_2) + \sinh(\Phi_1)\cosh(\Phi_2) \tag{5.28}$$

we can get from (5.21)

$$\underline{M}(0, \Phi_2 + \Phi_1) = \underline{M}(0, \Phi_2) \otimes \underline{M}(0, \Phi_1) \tag{5.29}$$

and from (5.16) and (5.19)

$$\begin{pmatrix} x_{21} \\ y_{21} \end{pmatrix} = \begin{pmatrix} \cosh(\Phi_2) & \sinh(\Phi_2) \\ \sinh(\Phi_2) & \cosh(\Phi_2) \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \tag{5.30}$$

which says exactly the same thing as

$$\begin{pmatrix} y_{21} \\ x_{21} \end{pmatrix} = \begin{pmatrix} \cosh(\Phi_2) & \sinh(\Phi_2) \\ \sinh(\Phi_2) & \cosh(\Phi_2) \end{pmatrix} \begin{pmatrix} y_1 \\ x_1 \end{pmatrix} \tag{5.31}$$

because the matrix satisfies (5.26). The transformation is thus unchanged although we have **renamed** (exchanged) the coordinates of x and of y. Our choice of coordinate names is subjective, while the transformation represents the **real**, objective contribution of uncertainty to the datum. It would be strange to find, that a real process is dependent on the names we gave to the coordinates! We can thus take (5.26) as natural and reasonable. We shall see below, that there are other important reasons supporting this seemingly trivial property of the operator, which represents the influence of uncertainty on the observed datum.

5.6 The Uniqueness of the Matrix Model

The structure $\mathcal{S}_m$ (5.18) formed by 2∗2 matrices $\underline{M}(A_0, \Phi)$ has been shown above to be a matrix model of uncertain data obtained by real quantification. A careful reader will note, that many interesting features of the matrix model resulted from the special choice of the coordinates x and y in (5.15). This substitution is legal because of the identity

$$\exp(\Phi) = \cosh(\Phi) + \sinh(\Phi), \tag{5.32}$$

which was applied to go from (5.14) to (5.15). This identity results from the definitions of hyperbolic and trigonometric functions and it is valid for all real and imaginary values of the argument Φ. However, it is logical to think, that other decompositions of the exponential function might lead to different pairs of coordinates, which would imply different models of uncertain data. A question then arises as to the uniqueness of coordinates (5.15), which is answered by the following statement:

Theorem 2:
The relation (5.32) is the only additive decomposition of the exponential function $\exp(\Phi)$ of a real and imaginary argument Φ into a pair of

different functions, for which relation 5.26 with $Det\{\underline{M}(0,\Phi)\} = 1$ holds identically for all values of the argument Φ.

Proof of the Theorem 2:

Let

$$(\forall\Phi \in \mathcal{R}^1)(\exp(\Phi) = f(\Phi) + g(\Phi)), \tag{5.33}$$

where $f(\Phi)$ and $g(\Phi)$ be some functions, which satisfy the assumptions of Theorem 2 so that the $M(0,\Phi)$'s elements are $M_{1,1} = M_{2,2} = f(\Phi)$ and $M_{1,2} = M_{2,1} = g(\Phi)$. The $M(0,\Phi)$'s determinant should be

$$f^2(\Phi) \ - \ g^2(\Phi)) \ = \ 1. \tag{5.34}$$

Substituting $f(\Phi) \ = \ \exp(\Phi) \ - \ g(\Phi)$ into 5.34, relation

$$g(\Phi) \ = \ (\exp(\Phi) \ - \ 1/\exp(\Phi))/2 \tag{5.35}$$

follows equal to $\sinh(\Phi)$ by definition. Hence

$$f(\Phi) \ = \ \cosh(\Phi). \tag{5.36}$$

This confirms, that the matrix model of uncertain data is unique.

5.7 Summary

Real quantification differs from ideal quantification due to participation of quantities, which include not only ideal, but also uncertain components. Gnostic theory considers structures composed of both ideal and uncertain quantities as commutative groups, which can be manipulated by the same kind of structural operations. The same operation also composes the ideal and uncertain quantities to form the quantity, which is observed. The theoretical model of a real (uncertain) quantification is thus formed by a pair of ideal quantification processes, which produce three images: the numerical image of the ideal quantity, the numerical image of the uncertain quantity and a (multiplicative or additive) numerical composition of both images, which *is equal* to the observed, but uncertain, datum. From the axiomatic setting of this model it has been shown, that the uncertain data component plays the role of a transformation, the invariant of which is the ideal value (the numerical image of the ideal quantity). The axioms of quantification have been used for the derivation of a unique matrix model of the quantification process.

Chapter 6

The Geometry of Real Quantification

6.1 Distance as a Problem

Let us consider two points a and b on a straight line with coordinates c_a and c_b, and ask once more the question, "What is the distance (L) between the points?" The answer depends on the level of the reader's mathematical skill:

Basic: It is simple,

$$L = |c_a - c_b|. \tag{6.1}$$

Thoughtful: It is not that elementary, because the path of integration, along which the distance should be measured, has not been defined. For the path denoted $\mathcal{P}(a, b)$ the distance would equal to the path integral

$$L = \int_{\mathcal{P}(a,b)} dp, \tag{6.2}$$

where dp is the length of an element of the path. Only in the case of the integration path coinciding with the straight line does the expression (6.2) reduces to the ordinary integral

$$L = |\int_{c_a}^{c_b} dx|, \tag{6.3}$$

which provides the same result as (6.1).

Advanced: It is complicated because neither the integration path nor the geometry is specified. Assume, that the integration path is $\mathcal{P}(a, b)$, and that such geometry is chosen, that the weight of an element dp of the path of the point z is $g(z)$. Then the distance is

$$L = \int_{\mathcal{P}(a,b)} g(z)dp. \tag{6.4}$$

> This expression reduces to (6.2) only if the weight $g(z)$ is a constant equal to 1, which is the case, when Euclidean geometry is employed.

The importance of knowing the path, when measuring distances, can be illustrated by the following: imagine a lady, whose home is at point a and has an office at point b. Even in the case, when both points are on the same straight street, the distance to be actually walked or driven between them would seldom be the same each time. It depends on stops made to shop, visit a hairstylist or a friend, etc., on the way. This problem of the path is important particularly, when measuring uncertainty, because uncertainty can move the geometrical image of data along a curve and not along a straight line.

The need to use a variable weight $g(z)$ in (6.4) can be also illustrated by an example. Consider the process of estimating the value of an asset by several differently qualified experts. Assume, that an "ideal" value for the property exists, eg as estimated by an omniscient Expert, but no such expert is on hand. It is felt, that all of these experts, together, should come close to the ideal value. The estimates are real numbers and the evaluation process is begun by calculating their arithmetic average. A second approximation is made to produce a more realistic result; it represents a **weighted** average using weights, which are dependent on the distance of each individual's estimate from the previous round's mean. Weights equal to 1 are given to estimates, which have the value of the previous average and these weights decrease with increasing distances from the mean. The new value of the weighted average is used for the next round and the iterative process ends, when the weighted average remains constant. The estimates of "bad" experts are thus suppressed and those of "good" ones are emphasized. The idea is simple—instead of assuming, that judgments of all experts are the same, and that each have the right to be taken equally into account, we evaluate the individual qualification of each expert by the quality of his work. This is accomplished by using a particular weighing function $p(z)$. The remaining problems are:

1. What is the form of the path, along which the errors (distances between the "ideal" and "estimated" values) should be measured?
2. What weighing function should be chosen (what kind of geometry) to get the "best" result?
3. How is "the best result" to be interpreted?

Gnostics can answer these questions for those readers who persevere.

From the point of view of the “man in the street,” the distance between two points, assuming that it is to be measured along the segment of the straight line connecting the points, is “obvious”: “everybody knows, that this is the shortest path among all those possible.” However, this statement is true for Euclidean geometry and may be false in another geometry. To show the changes in the distance, which result from path variations, it is necessary to extend the dimensionality of the problem—to go over from the geometry of a straight line to the geometry of (at least) a plane. Because distance is a special application of *the scalar product* of two vectors, it will be instructive to linger a moment on this notion.

6.2 A Bit of Geometry

No one is a complete geometric neophyte, since geometry is as basic as reading. Notions such as length or distance, angles and circles are part of the common body of knowledge. The problem—as it has already been mentioned—is, that many are not aware, that their basic understanding is tied unequivocally to a single geometry, the Euclidean one. A full course in non–Euclidean geometries is not going to be presented here, however, a brief and elemental introduction to this field is required in order to unveil the secrets of uncertain data.

6.2.1 Riemannian Scalar Product

Let us consider a plane $\mathcal{P}_2$, the points of which are presented as coordinate pairs $\langle x, y\rangle$. Let functions $\eta(x, y)$ and $\theta(x, y)$ be continuous, one-to-one and at least once differentiable functions. The values of functions x and y will be called *coordinates*, the mapping

$$x' = \eta(x, y), \quad y' = \theta(x, y) \tag{6.5}$$

is a sufficiently general form of the replacement of coordinates or transformation of the plane.

Subjecting the variables x and y to differential changes we obtain

$$\begin{pmatrix} dx' & dy' \\ dy' & dx' \end{pmatrix} = \begin{pmatrix} \frac{\partial\eta}{\partial x} & \frac{\partial\eta}{\partial y} \\ \frac{\partial\theta}{\partial x} & \frac{\partial\theta}{\partial y} \end{pmatrix} \begin{pmatrix} dx & dy \\ dy & dx \end{pmatrix}. \tag{6.6}$$

This is a general formulation; however, rewriting it as

$$\begin{pmatrix} dx' & dy' \\ dy' & dx' \end{pmatrix} = \begin{pmatrix} g_{11}(x,y) & g_{12}(x,y) \\ g_{21}(x,y) & g_{22}(x,y) \end{pmatrix} \begin{pmatrix} dx & dy \\ dy & dx \end{pmatrix} \tag{6.7}$$

and denoting the matrix

$$\underline{G}(x,y) = \begin{pmatrix} g_{11}(x,y) & g_{12}(x,y) \\ g_{21}(x,y) & g_{22}(x,y) \end{pmatrix}, \tag{6.8}$$

the Riemannian scalar product can be defined over the plane $\mathcal{P}_2$ in the following way:

Definition 3:

Let $\underline{v}$ and $\underline{v'}$ be column vectors defined by two points on the plane $\mathcal{P}_2$

$$\underline{v} = \begin{pmatrix} x \\ y \end{pmatrix} \qquad \underline{v'} = \begin{pmatrix} x' \\ y' \end{pmatrix}, \tag{6.9}$$

and $d\underline{v}$ and $d\underline{v'}$ their differentials. Let the matrix $\underline{G}(x,y)$ (6.8) satisfy the conditions of *regularity:*

$$g_{11}g_{22} - g_{12}g_{21} \neq 0, \tag{6.10}$$

and *symmetry:*

$$g_{12} = g_{21}. \tag{6.11}$$

Then the matrix $\underline{G}(x,y)$ is *the metric matrix* and the expression

$$[d\underline{v}, d\underline{v'}]_G = d\underline{v}^T \underline{G}(x,y) d\underline{v'} \tag{6.12}$$

(where the upper index T denotes the transposition operator) is the *scalar product* of the two vectors.

Within this framework, the square $(dL)^2$ of the differential of distance between two points $\langle x+dx, y+dy\rangle$ and $\langle x,y\rangle$ is calculated by

$$(dL)^2 = g_{11}(x,y)(dx)^2 + 2g_{12}(x,y)dxdy + g_{22}(x,y)(dy)^2. \tag{6.13}$$

This is the Riemannian *metric form* to measure lengths.

The differential form of (6.13) together with the dependence of the metric matrix on the point where the element of the distance is to be determined means, that—in general—

1. the method of measurement may be different at different points in the plane,
2. the distance between two points may be dependent on the form of the curve connecting the points, ie on the path,
3. the distance should be calculated using path integration,
4. each element of the path may receive a different weight, and therefore have a different influence on the value of a distance.

Making the weight of each segment of the distance dependent on "something" (ie on the discrepancy of a particular judgment with respect to the "central" meaning) was exactly, what we did in our practical example by intuitively treating differently the individual points of view of each member of an expert team.

Returning to abstract geometry, two important very special cases of Riemannian planes are considered: the Euclidean and Minkowskian ones. They both are obtained for some **constant** metric matrices.

6.2.2 Euclidean Plane

A point (U) in a two-dimensional real plane having coordinates U_1 and U_2 can be interpreted both as a couplet $\langle U_1, U_2 \rangle$ and as a (column) vector $\underline{U}$. The *transposed* vector is then the row vector $\underline{U}^T = (U_1, U_2)$. We all remember the (Euclidean) formula for calculating the length L of this vector

$$L = \sqrt{U_1^2 + U_2^2} \tag{6.14}$$

equal to the (Euclidean) distance of the point U from the origin $\langle 0, 0 \rangle$ of the coordinate system. Introducing the seemingly trivial (identity) matrix

$$\underline{G}_{2,E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{6.15}$$

one may rewrite (6.14) in the form of a special case of (6.13)

$$L^2 = [\underline{U}, \underline{U}]_{2,E}, \tag{6.16}$$

ie as a particular application of the Riemannian scalar product. It is possible to represent it in the integral form because the (Euclidean) metric matrix is a constant. Index 2 gives the dimension of the space while index E designates the type of the metric matrix as Euclidean.

The cosine of the angle α between two vectors $\underline{U}$ and $\underline{V}$ is then evaluated using the well-known (Euclidean) formula

$$\cos(\alpha) = \frac{U_1 V_1 + U_2 V_2}{\sqrt{U_1^2 + U_2^2}\sqrt{V_1^2 + V_2^2}}, \tag{6.17}$$

which can be also expressed using the scalar product:

$$\cos(\alpha) = \frac{[\underline{U}, \underline{V}]_{2,E}}{\sqrt{[\underline{U}, \underline{U}]_{2,E}}\sqrt{[\underline{V}, \underline{V}]_{2,E}.}} \tag{6.18}$$

It might appear, that there is no advantage in appealing to Riemannian geometry and to use more complicated formulae with such simple relations; however, as it is usually found in mathematics, a higher level of observation broadens the horizon. This can be shown by using Minkowskian plane geometry but, before considering the Minkowskian case, it is instructive to demonstrate, that the important notions of classical statistics are based on Euclidean geometry.

6.2.3 The Euclidean Scalar Product in Statistics

Imagine N pairs $\langle u_i, v_i \rangle$ of observations, which represent the results of a set of repeated experiments ($i = 1,\ ...,\ N$). Denote

$$\overline{d} = \frac{\sum_{i=1}^{i=N} d_i}{N} \tag{6.19}$$

the arithmetical mean of a sample of data d_i. Let the population variance, from which the sample was chosen be σ_d^2 and its point estimate $\widetilde{\sigma_d^2}$. Let $\underline{\mathrm{d}}$ be an $N-$dimensional column vector, the components of which are $d_i - \overline{d}$. This data vector is thus—as statisticians say—*centralized.* Introducing the $N * N$ identity matrix written as an $N-$dimensional Euclidean metric matrix $\underline{G}_{N,E}$, one may rewrite the well known formula for estimated variance by means of the arithmetic mean of scalar products

$$\widetilde{\sigma_d^2} = \overline{[(\underline{d} - \underline{\overline{d}}), (\underline{d} - \underline{\overline{d}})]}_{N,E}. \tag{6.20}$$

Using the same notation, the point estimate of the covariance between two $N-$dimensional data vectors $\underline{u}$ and $\underline{v}$ is

$$\widetilde{\mathrm{cov}}(\underline{u}, \underline{v}) = \overline{[(\underline{u} - \underline{\overline{u}}), (\underline{v} - \underline{\overline{v}})]}_{N,E}. \tag{6.21}$$

Since correlation coefficients are based on estimated covariances and variances and correlation functions and correlation matrices are built using correlation coefficients, all these popular and widely used statistical tools originate from the notion of the Euclidean scalar product.

It has been already pointed out, that Euclidean geometry is an important paradigm. Its axioms are nearly 23 centuries old, and its roots are even older. So eg the famous Pythagorean theorem—if it really was formulated personally by Pythagoras—could date back as much as 2,600 years. But, we know today, that it holds in what we now call Euclidean geometry, and that it may be false in other geometries.

What lesson can we learn from these reminiscences? The following: there are important notions used in every-day mathematical statistics, which are based on the Euclidean geometric paradigm and their roots go back to ancient times. There would be nothing strange in making use of “good old” knowhow if scientific progress had not shown “in between” the substantial limitations of the ancient science. It is a habit of small children to continually ask “why?” It is a pity, that some adults frequently hesitate to pose the same question. Many unquestioningly believe in the authority of their schools, their text-books, and the pronouncements of their professors and other learned persons. In other words, they conservatively accept the “ruling” paradigm because it is more comfortable and less dangerous then to independently question the problem.

6.2.4 Minkowskian plane

Let all U_1, U_2, V_1 and V_2 be again reals. Let $\underline{G}_{2,M}$ be the (*Minkowskian*) constant metric matrix

$$\underline{G}_{2,M} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{6.22}$$

The vector form of two points on the Minkowskian plane is

$$\underline{U} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \qquad \underline{V} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}. \tag{6.23}$$

The Riemannian differential form (6.12) can in this case be integrated to get

$$[\underline{U}, \underline{V}]_{2,G} = U_1 V_1 - U_2 V_2. \tag{6.24}$$

The squared length of a (column) radius vector $\underline{v}$ (having components x and y) denoted as

$$\rho_M^2(\underline{v}) = [\underline{v}, \underline{v}]_{2,G} \tag{6.25}$$

obviously equals

$$\rho_M^2(\underline{v}) = x^2 - y^2. \tag{6.26}$$

This is a substantially different notion to that, which we are accustomed in Euclidean geometry. The square of the length may be negative as well as positive or zero—it depends on the direction of the vector. Some paths in the Minkowskian plane leading from one point to another may thus have not only a real but also an imaginary or zero length. From this point of view, the Minkowskian plane can be split into several parts. Such a reasonable splitting will be considered in detail in Chapter 10.

6.2.5 Invariants, circles and rotations

Consider a matrix $\underline{R}_{2,*}(\phi)$, for which the relation

$$(\forall\phi_* \in \mathcal{R}^1)(\underline{R}_{2,*}^T(\phi_*)\ \underline{G}_{2,*}\ \underline{R}_{2,*}(\phi_*) = \underline{G}_{2,*}) \tag{6.27}$$

holds. The star designates 'E' or 'M' depending on the geometry. When both vectors $\underline{v}$ and $\underline{v'}$ are transformed (multiplied) by a matrix satisfying (6.27), the scalar product (6.12) does not change. In other words, the scalar product is *invariant* to such a transformation. It is obvious, that equation (6.27) and the metric matrix delimits the form of the matrix $\underline{R}_{2,*}(\phi_*)$. It can be verified by substitution, that all matrices having the form

$$\underline{R}_{2,E}(\phi_E) = \begin{pmatrix} \cos(\phi_E) & -\sin(\phi_E) \\ \sin(\phi_E) & \cos(\phi_E) \end{pmatrix} \tag{6.28}$$

satisfy (6.27) with the Euclidean metric matrix $\underline{G}_{2,E}$ while matrices

$$\underline{R}_{2,M}(\phi_M) = \begin{pmatrix} \cosh(\phi_M) & \sinh(\phi_M) \\ \sinh(\phi_M) & \cosh(\phi_M) \end{pmatrix} \tag{6.29}$$

solve the equation in the case of the Minkowskian metric. Multiplying a (column) radius vector $\underline{v}$ of a point on the Euclidean or Minkowskian plane (having components x and y) by a matrix $\underline{R}_{2,*}(\phi_*)$ of either class, we obtain

$$v^{'} = \underline{R}_{2,*}(\phi_*)v. \tag{6.30}$$

Recall, that a *circle* is a line, the points of which are equally distant from a fixed point, *center*. Distance is specified by geometry, therefore circles can have different forms. Denote ρ and ρ' the lengths of the vector $\underline{v}$ and of the transformed vector $\underline{v}'$, respectively. From (6.27) substituted into (6.12) for both transformed vectors, the identity

$$(\forall\phi_* \in \mathcal{R}^1)(\rho_*^{'} = \rho_*) \tag{6.31}$$

holds. It is worth rewriting this relation in both explicit forms (as in (6.16) and (6.25)):

$$(\forall\phi_E \in \mathcal{R}^1)(\rho_E^2 = x^2 + y^2) \tag{6.32}$$

and

$$(\forall\phi_M \in \mathcal{R}^1)(\rho_M^2 = x^2 - y^2). \tag{6.33}$$

Relation (6.32) may be interpreted as the equation of the (Euclidean) circle with its center at the point $\langle 0, 0\rangle$ and a radius of ρ_E. Analogously, (6.33) is the Minkowskian circle, the center of which is again $\langle 0, 0\rangle$ with the radius ρ_M. It should be emphasized, that all points $\langle x, y\rangle$ considered in (6.32) and (6.33) are points of the same real 2-dimensional linear space R^2, which is endowed with the Euclidean and alternatively with the Minkowskian metric. The matrix $\underline{\underline{R}}_{2,E}(\phi_E)$ defined by (6.28) (resp. $\underline{\underline{R}}_{2,M}(\phi_M)$ by (6.29)) **rotates** vectors, preserving their lengths and the angles between vectors measured by either Euclidean or Minkowskian) geometry.

6.3 Minkowskian Nature of Real Quantification

Let us identify variables x and y having the form of (5.15) with those appearing in (6.33). One obtains by substitution of (5.15) into (6.33)

$$\rho_M^2 = Z_0^2. \tag{6.34}$$

This result is crucial to the theory and is therefore recast as a theorem:

Theorem 3:

Let Z_0 be the numerical image of an ideal quantity subjected repeatedly to quantification under effects of different uncertainties represented by different values of the variable Φ. Let scale parameter S be fixed equal to 1. Let relations (5.15) define points in a Minkowskian plane.
It holds that:

1. All points representing observed data having an arbitrary uncertain component Φ lie on the Minkowskian circle.

2. The center of this circle coincides with the origin of the coordinate system ($\langle 0, 0\rangle$).
3. The diameter of this circle measured in Minkowskian geometry equals $\sqrt{Z_0^2}$.

These statements were derived directly and it is therefore not necessary to prove them. The message, which they convey, is important and is worth further analysis.

An objection to the model could be raised on the grounds, that an infinite number of data have been assumed because a continuous line (an arc of a circle) has been used as a model. However, this is not a correct interpretation of what has been done. What has been modeled is the path of a **single** datum as it is transported from the "ideal" point $\langle Z_0, 0\rangle$ of a plane to its "final" point $\langle Z_0, \exp(\Phi)\rangle$. This "transport" of the point was characterized by a geometric (virtual, not a real, or physical) movement. Although the model was designed for a **single** datum, it is valid for **all possible** values of its uncertain component thus making the arc continuous.

6.4 Summary

The notions of the metric matrix and scalar product are basic in determining the character of a geometry. Starting with the general Riemannian definition of these notions, one may obtain (among others) two important special cases related to constant metric matrices, the Euclidean and Minkowskian ones. An examination of the basic ideas of mathematical statistics leads to the conclusion, that they are deeply rooted in the Euclidean geometric paradigm. In contrast, the gnostic theory of real quantification develops a model, which is inherently connected to Minkowskian geometry. It is important to state, that the Minkowskian character of real quantification is not an assumption of the gnostic theory, but is only an interpretation of its results, which are derived from its first axiom. Within this geometric interpretation, a matrix operator parameterized by the uncertainty moves the 2-dimensional vector image of the observed datum on the Minkowskian plane along an arc of a Minkowskian circle. The ideal value of the observed datum is an invariant of this geometric movement, playing the role of the (Minkowskian) radius of the circular path.

Chapter 7

Quantification and Relativistic Physics

7.1 Lorentz's Transformation

A small excursion into physics will be useful for clarification of the gnostic paradigm[1]. Let v_* be the speed of light and let the general 4-dimensional time-space model of special relativity theory discussed in [113] be reduced to 2 dimensions. Define a time-space point in a coordinate system by a vector $\underline{u} = (v_* t, s)^T$, where t is the time and s the space coordinate. In another coordinate system moving with respect to the former one with a velocity v, the same vector will be observed as $\underline{u'} = (v_* t', s')^T$:

$$\underline{u'} = \underline{L}(v/v_*)\underline{u}, \tag{7.1}$$

where

$$\underline{L}(v/v_*) = \begin{pmatrix} \gamma(\frac{v}{v_*}) & (\frac{v}{v_*})\gamma(\frac{v}{v_*}) \\ (\frac{v}{v_*})\gamma(\frac{v}{v_*}) & \gamma(\frac{v}{v_*}) \end{pmatrix}, \tag{7.2}$$

and where

$$\gamma(\frac{v}{v_*}) = \frac{1}{\sqrt{(1 - (\frac{v}{v_*})^2)}}. \tag{7.3}$$

The matrix $\underline{L}(\frac{v}{v_*})$ is the matrix representation of the 2-dimensional *proper homogeneous Lorentz's transformation group*[2] [113]. Its significant

[1] The Dutch physicist, Hendrik A. Lorentz (1853-1928); developed the coordinate manipulations in the late 19th century to explain optical and electromagnetic phenomena.

[2] Newtonian mechanics is based on so-called Galilean geometry, which permits motion at unlimited speeds. With the advent of the special theory of relativity, it became necessary to take into account the limited speed of light. This can be done by using the Lorentz transformation, which makes both place and time measurements dependent on the speed of the observer.

feature is that it preserves the Minkowskian scalar product. The squared length of the vector $\underline{u}$ is thus invariant to the transformation (7.1):

$$(\forall v/v_* \in R^1)([\underline{u}', \underline{u}']_{2,M} = [\underline{u}, \underline{u}]_{2,M}). \tag{7.4}$$

As shown in [113], the transformations according (7.1)–(7.3) are the **only** nonsingular homogeneous coordinate transformations $\underline{u} \to \underline{u}'$ that leave the scalar product invariant. (Nonsingular means that both $\underline{u}'(\underline{u})$ and $\underline{u}(\underline{u}')$ are well-behaved differentiable functions. Relation (7.1) is called homogeneous because it does not include an additive constant.) A stronger version of this statement related to uniqueness was proved by purely algebraic means in [100] without requiring the differentiability of the two functions.

The uniqueness of the Lorentz's transformation and thus of the class of matrices $\underline{L}(v/v_*)$ (7.2) might perhaps lead to confusion because another (also unique) class of matrices $\underline{R}_{2,M}(\phi_M)$ (6.29), which leave the scalar product invariant has already been discussed. However, there is no conflict here because the difference between the two classes is only formal. Indeed, substituting

$$\tanh(\phi_M) = \frac{v}{v_*} \tag{7.5}$$

and using the formulae of hyperbolic functions, one identifies $\underline{R}_{2,M}(\phi_M)$ (6.29) with the Lorentz's matrix $\underline{L}(v/v_*)$ (7.2). The importance of this relation is that it gives the fundamental tie between the theory of uncertainty and well established physical phenomena.

The principal reason for appealing to the theory of special relativity is to apply the *Lorentz-invariance* principle, which requires that the form of equations should be the same in all *inertial* coordinate systems, i.e. in systems moving with a constant velocity each with respect to other, since the speed of light is the same in all inertial systems. The Lorentz's transformations explain the time dilation and size contraction of **images** of objects observed from a coordinate system moving relative to the object[3]. This transformation provides an *information channel* for the observed input, which is the real object, and the output, which is a distorted image of the object. This channel is parameterized by the relative velocity of the observer. One sees an immediate analogy with the idea behind quantification, which also "works" as an information channel. It has the ideal quantity

[3] These effects are sometimes misinterpreted as a "dependence of time and size of an object on its velocity". In reality, the object remains unchanged (having its *proper* space and time coordinates measured in a reference system, in which the object is at rest) and independent of the observer, whose frame of reference is in motion (relative to the object's reference system). What is velocity-dependent, is the (subjective) observation described by Lorentz's transformation.

as its input and the distorted two-dimensional image as its output, the distortion being parameterized by the amount of uncertainty. Equation (7.5) developed earlier establishes the relation between the parameters of both "information channels"; however, there is much more than a formal analogy between these seemingly unrelated processes[4].

7.2 Relations to Relativistic Mechanics

7.2.1 Isomorphism of Two Groups of Transformation

The first fundamental relation has already been shown: the Lorentz's transformation matrix $\underline{L}(v/v_*)$ (7.2) has been identified with the matrix operator $\underline{R}_{2,M}$ of Minkowskian rotation. However, the latter matrix is identical to the matrix $\underline{M}(0, \Phi)$ (5.19). Moreover, this identity exists for all values of the parameters v/v_* and Φ, for which (7.5) holds, as can be easily verified:

$$(\forall \Phi_k \in R^1)(\forall v_k/v_* | v_k/v_* = \tanh(\Phi_k), k = 1, 2)$$

$$(\underline{M}(0, \Phi_1)\underline{M}(0, \Phi_2) = \underline{L}(v_1/v_*)\underline{L}(v_2/v_*)). \tag{7.6}$$

This result can be formulated as a theorem:

<u>**Theorem 4:**</u>
The commutative group of matrix operators $\underline{M}(0, \Phi)$ representing the effect of uncertainties on data within the quantification process is isomorphic with the group of Lorentz's transformation of the Minkowskian plane formed by matrices $\underline{L}(v/v_*)$.

The importance of this theorem for the gnostic theory is closely connected with the relation of the quantification process to the conservation law of relativistic physics.

7.2.2 Quantification and the Relativistic Conservation Law

The idealized model for a system of a large number of freely moving particles (particles not subject to forces) is a continuously distributed mass

[4] Of course, relativity theory has a negligible impact on the perceptions of an observer in cases, where velocities are very small compared to v_*, nevertheless motions at high speed as well as strongly uncertain data exist and require the application of Lorentz transformations.

having a density μ. The *proper* value of this density observed from within the reference system moving with the mass will be denoted μ_0. From the special theory of relativity [113], the matrix representation of the energy-momentum tensor $\underline{E}(v/v_*)$ of the subject mass has the form

$$\underline{E}(v/v_*) = \begin{pmatrix} \mu_0 v_*^2 \gamma^2(\frac{v}{v_*}) & \mu_0 v v_* \gamma^2(\frac{v}{v_*}) \\ \mu_0 v v_* \gamma^2(\frac{v}{v_*}) & \mu_0 v_*^2 \gamma^2(\frac{v}{v_*}) \end{pmatrix}, \tag{7.7}$$

where $\gamma(\frac{v}{v_*})$ is the expression as that given in (7.5). Introducing the matrix

$$\underline{E}_0 = \begin{pmatrix} \frac{\mu_0 v_*^2}{2} & 0 \\ 0 & \frac{\mu_0 v_*^2}{2} \end{pmatrix} \tag{7.8}$$

and using relations

$$\gamma(\frac{v}{v_*}) = \cosh(\Phi) \qquad \frac{v}{v_*}\gamma(\frac{v}{v_*}) = \sinh(\Phi), \tag{7.9}$$

which result from (7.5) and identities $2\cosh^2(\Phi) = \cosh(2\Phi) + 1$ and $2\cosh(\Phi)\sinh(\Phi) = \sinh(2\Phi)$, one arrives at a relation, which is important enough to be stated as another theorem:

Theorem 5:

Let $\underline{M}(0, 2\Phi)$ be the matrix operator representing the effect of uncertainty on the observed datum and having the form of (5.16). Let $\underline{G}_{2,E}$ be the identity matrix (6.15). Let $\otimes$ denotes matrix multiplication.
Then following identity holds:

$$(\forall\Phi \in R^1)(\forall v/v_* | v/v_* = \tanh(\Phi))(\underline{E}(v/v_*) = \underline{E}_0 \otimes (\underline{M}(0, 2\Phi) + \underline{G}_{2,E})). \tag{7.10}$$

This formal statement deserves further comment. Energy and momentum (elements of the matrix $\underline{E}(v/v_*)$) are components of one of the most fundamental notions of physics in view of their role in the formulation of the Energy-Momentum Conservation Law. This generally accepted theoretical hypothesis, which has strong empirical support, enables many physical processes to be mathematically modeled. It will be shown in the following sections, that the matrix $\underline{M}(0, 2\Phi)$ plays an important role in gnostics because its elements evaluate and weigh observed data in a nonlinear and non-quadratic manner, which is interpretable as an application of Riemannian geometry. The importance of the 7.10 for gnostics lies in its ability to motivate the aggregation law of uncertain data. Further, it is Lorentz-invariant, for it holds for all possible velocities v/v_* and for all corresponding uncertainties (Φ).

Theorems 4 and 5 thus provide a link between events and processes, which exist in different scientific fields: the theory of uncertain data (mathematical gnostics) and relativistic mechanics. That such a relationship exists should not surprise those, who accept, that:

1. there is a mutual dependence between all real processes because of the unity and universality of the world,
2. the same real process can be analyzed from the points of view of different sciences,
3. the boundaries between different scientific fields are as 'fuzzy' and 'mixed' as those between the different features of real processes,
4. classical statistics and classical (Newtonian) mechanics have been linked for many years.

Most of the above statements will draw general support but statisticians, with a strong belief in the "purely mathematical roots" of their science may balk at the last one. We shall return to this important point later, when the aggregation problem of uncertain data is considered.

7.3 Uncertainty in Relativistic Observations

From the point of view of the gnostic paradigm, data uncertainty is caused by lack of information. This idea can be illustrated using relativistic observations. Imagine two space ships S1 and S2 moving along parallel straight paths with different but constant velocities. An observer in S2 wants to measure the proper length L_1 of S1 (i.e. the length as would be measured in S1's own coordinate system). He, of course, can only obtain L_2, the only measurement, which can be obtained from within S2. In order to obtain the *exact* result, which he is seeking, the observer must also know the relative velocity v/v_0 of S2 with respect to S1. The observer's insufficiently precise knowledge of the energy-momentum tensor "contaminates" the observed datum by the uncertainty Φ, which is connected to the velocity by means of (7.5). The example shows that to explain uncertainty, there is no reason to introduce randomness, mutual independence and unlimited repeatability of the observations, stationarity, and so on. An improvement in the quality of one single observation can be achieved by obtaining better information on the relative velocity and by making use of knowledge of

Lorentz's transformation. Analogously, an improvement in the results of a series of different observations contaminated by uncertainty can be obtained by using data treatment methods based on knowledge of the nature of quantification and of optimal estimation.

A question can be expected at this point, as to whether references to and an understanding of some nonmathematical scientific fields is really a necessary element of gnostics? The answer is decisively negative. Readers wishing to consider gnostics as a purely mathematical theory can be satisfied: the skeleton of the theory has been developed using the consistent 'definition—axiom—theorem—proof' mathematical procedures. However, there are others, who may wonder **why these and not other** definitions and axioms have been chosen: where are the roots? To explain (and then to understand) these roots, it is necessary to venture across the boundaries of mathematics because the roots of gnostics are in the real world.

7.4 Summary

There is a formal link between the mathematical model of uncertain data obtained by quantification and Lorentz's transformations of the Minkowskian plane, which is manifested by

1. the isomorphism of the group of gnostic matrix-operators representing the data uncertainty with the group of Lorentz's 2-dimensional time-space transformations of relativistic physics,
2. the linear relationship between the relativistic energy-momentum matrix and gnostic matrix, which is parameterized by the value of the uncertainty.

These mappings are Lorentz-invariant, i.e. their forms stay unchanged for all values of uncertainty. Their existence may be explained as a manifestation of the unity of Nature, where both real processes and information, which can be obtained about them, are inherently related. Therefore, information is a complementary dimension of space, time and any other dimension of the process being examined.

Chapter 8

A Bit of Algebra and Analysis

8.1 Pair Numbers

8.1.1 Complex and Double Numbers

The notion of *complex numbers* is well-known; and it is an important part of the mathematics used in several applied sciences. It would be difficult to solve the simple problem of finding the roots of a quadratic equation without using them. The basic idea is simple—to map the Euclidean plane onto the Gaussian one using the mapping

$$(\forall x, y \in R^1)(\langle x, y\rangle \leftrightarrow x + i\, y), \tag{8.1}$$

where the symbol i is an indeterminate satisfying the relation

$$i^2 = -1. \tag{8.2}$$

A parallel notion to that of complex numbers can be obtained, when mapping the Euclidean plane onto the Minkowskian plane:

$$(\forall x, y \in R^1)(\langle x, y\rangle \leftrightarrow x + j\, y), \tag{8.3}$$

where the symbol j is another indeterminate, which satisfies

$$j^2 = 1. \tag{8.4}$$

Elements of the range of this mapping introduced for the first time by the English mathematician William K. Clifford (1845–1879) will be called *double numbers*. Unlike the case of complex numbers, which exist for all real pairs x, y, the double numbers exist only for real pairs x, y, whose squares are **not** equal, ie for those satisfying condition

$$x^2 - y^2 \neq 0. \tag{8.5}$$

This is a natural constraint. It can be seen from (6.26), that condition (8.5) excludes from consideration all points on the Minkowskian plane $(\langle x, y\rangle)$, for which the distance from the origin $(0, 0)$ is zero. There are an infinite number of such points: those on both diagonals $(|x| = |y|)$.

The formal similarity of complex and double numbers enables them to be combined in a simple way by the introduction of another indeterminate c so, that it covers both cases:

$$c \in \{i, j\}. \tag{8.6}$$

An expression of the form

$$(x, y \in R^1)(x^2 \neq y^2)(c \in \{i, j\})(u_c = x + c\, y) \tag{8.7}$$

will be called a *pair number.* A pair number is thus a complex or a double number depending on the choice of $c = i$ or $c = j$. As will be seen below, using the concept of pair numbers will permit the number of formulae used to be decreased by 50 %.

To understand the idea of an indeterminate, it should be recognized, that it is not a real number. (Many use the name “imaginary unit” for the indeterminate i. However, the notion of an indeterminate is more general [71] and more exact.) An expression such as $a + c\, b$ for reals a, b and indeterminate c is not an “ordinary” sum but only the notation for a pair of two different objects, for which the operation of addition is not defined, ie, the pair $\langle a, b\, c\rangle$. On the other hand, multiplication of the indeterminate c by a real b denoted $b * c$ (or $c * b$ because of commutativity) makes good sense: the product is another indeterminate, for which $(b * c)^2 = b^2 * c^2$ is a real number. Because this operation is an analogue to numerical multiplication, we write these products simply as cb.

8.1.2 The 2-algebra of Pair Numbers

The set of pair numbers will be denoted by U_c, and its subsets (classes) of complex and double numbers are U_i or U_j, corresponding to their character. It can be shown [71], that the set U_c can be interpreted as the associative and commutative *2-algebra* (algebra of dimension 2) over the field of real numbers having a unit u_o

$$u_o = 1 + c\, 0. \tag{8.8}$$

The sum of two elements $u_1 = a_1 + c\,b_1$ and $u_2 = a_2 + c\,b_2$ of this algebra has the form of

$$u_1 + u_2 = (a_1 + a_2) + c\,(b_1 + b_2), \tag{8.9}$$

while their product is

$$u_1 u_2 = (a_1 a_2 + c^2\,b_1 b_2) + c\,(a_1 b_2 + a_2 b_1). \tag{8.10}$$

It is worth noting, that both addition and multiplication of pair numbers is defined only for numbers belonging to the **same class,** ie either to U_i or to U_j. Operations between elements of two different classes are not considered and will not be used.

Let $u = a + c\,b \in U_c$ be a pair number. The pair number

$$\bar{u} = Co(u) := a - c\,b \tag{8.11}$$

is the *conjugate* of u and

$$Tp(u) := b + c\,a \tag{8.12}$$

is the *transposed* pair number u: it is obtained by exchanging the *components* a and b of the pair number.

Expression

$$|u|_c := \sqrt{a^2 - c^2\,b^2} = \sqrt{u\,\bar{u}}, \tag{8.13}$$

is the *modulus* of the pair number u.

Division of two pair numbers $u_1,\ u_2 \in U_c$ belonging to the same class for $|u_2|_c \neq 0$ is given by the formula

$$u_1/u_2 = u_1\,\bar{u_2}/|u_2|_c^2. \tag{8.14}$$

As a result of these definitions, the following relations hold for both subsets $c = i, j$:

$$(\forall u \in U_c)(Co(Co(u)) = u), \tag{8.15}$$

$$(\forall u, u' \in U_c)(Co(u + u') = Co(u) + Co(u')), \tag{8.16}$$

$$(\forall u, u' \in U_c)(Co(u\,u') = Co(u')\,Co(u)), \tag{8.17}$$

$$(u \in U_c)((Co(u) = u) \iff ((\exists a \in R^1)(u = a\,u_o = a + c\,0)(u_o \in U_c)), \tag{8.18}$$

$$(\forall u \in U_c)(u\,Co(u) = (a^2 - c^2\,b^2)\,u_o = Q(a, b)\,u_o), \tag{8.19}$$

where $Q(a, b)$ is a quadratic form. Let us now cite a reformulation of the theorem proved in [96] as a generalization of Frobenius's well known theorem related to the algebra of complex numbers:

Theorem 6:
Structures of double and complex numbers are the only 2-algebras with the unit u_0, such that $u_0 = 1 + c\,0$, (where c is defined by (8.2), (8.4) and (8.6)) and with the unitary conjugate operation satisfying (8.15)–(8.19), where $Q(a, b)$ in (8.19) is a non-degenerate quadratic form.

Theorem 6 thus ensures the uniqueness of the two 2-algebras. Moreover, it also ensures the uniqueness of two different plane geometries, the Minkowskian and Gaussian. It has already been shown, that the former geometry is valid for the range of the unusual function called quantification. However, each point $\langle x, y \rangle$—the theoretical model of the real quantification of a real quantity—will be either a double number $x + j\,y \in U_j$ or a complex number $x + i\,y \in U_i$.

More formally: such a mapping $\eta : U_j \longrightarrow U_i$ can be introduced so, that

$$\eta : a + j\,b \longleftrightarrow a + i\,b. \tag{8.20}$$

In general, this mapping preserves addition but not multiplication.

The double interpretation of the results of real quantification plays a very important role in gnostic theory because it opens a path to the world of optimal estimation.

A comment to Theorem 6 is notable related to a degenerated case of the $Q(a, b)$ in (8.19), where $c^2 = 0$. There is a third geometry attached to this case called the Galilean geometry in [118]. This name was given to the oldest geometry which was ruling over a long time period till the XIX-th century. Its physical interpretation is obvious in the case of variable x representing time and y stating for a space coordinate of an object moving along the path $x = 0$ with an infinite velocity. This path can be called “Galilean circle”. The Galilean geometry can be helpful in showing the extreme lengths of paths in Minkowskian and complex plains interpreted as estimation errors.

8.1.3 Geometric Interpretation of Pair Numbers

The range of the real quantification process is a set of pairs $\langle x, y \rangle$ of real numbers x and y, ie the Cartesian product $R^1 \times R^1$. This *quantification event* can be interpreted from several perspectives:

1. a point (x, y) in the Euclidean plane,
2. a point $(x, i\,y)$ in the Gaussian (complex) plane,

3. a complex number $x + i\,y$,
4. a point $(x, j\,y)$ in the Minkowskian plane dual to the Gaussian one,
5. a point $(x, j\,y)$ in the plane (x, y) endowed by the Galilean geometry,
6. a 2∗2 twice symmetrical matrix $\underline{M}(A_0, \Phi)$ (5.16),
7. a double number $x + j\,y$.

It is easy imagine one-to-one mappings between each pair of these representations. There is obviously no fundamental difference between the interpretations of the first through the third from the geometric point of view because the same (Euclidean) metric is used to measure the distances (or the lengths of vectors). Nor is there any difference between cases No. 4 through No. 6 in the following sense: the determinant of the matrix $\underline{M}(A_0, \Phi)$ equals the squared length of the radius vector of the point $(x, j y)$ as well as the squared modulus of the double number $x + j\,y$. It can be concluded, that pair numbers are closely connected with the joint use of the Euclidean and Minkowskian plane geometries, and that there is a one-to-one mapping between double numbers and the matrix representations of the quantification events. A much deeper relation can be shown, if the isomorphism of the structures is considered.

8.1.4 Modeling Quantification with Double Numbers

The structure $\mathcal{S}_m$ (5.18) of matrices $\underline{M}(A_0, \Phi)$, (which models the quantification process) is isomorphic with the direct product of commutative groups $\langle\{A_0\}, +\rangle$ and $\langle\{\Phi\}, +\rangle$ by Theorem 1. As stated in the outline proof to Theorem 1, the structure $\mathcal{S}_m$ is a direct product of commutative groups $\langle\{\underline{M}(A_0, 0)\}, \otimes\rangle$ and $\langle\{\underline{M}(0, \Phi)\}, \otimes\rangle$ isomorphic with $\langle\{A_0\}, +\rangle$ and $\langle\{\Phi\}, +\rangle$. Let us introduce three structures of double numbers:

Definition 4:

Let $\mathcal{M}_d$, $\mathcal{M}_0$ and $\mathcal{M}_n$ be the following sets of double numbers:

$$\mathcal{M}_d := \{x + j\,y \,|\, x = Z_0 \cosh(\Phi),\; y = Z_0 \sinh(\Phi)\}, \tag{8.21}$$

$$\mathcal{M}_0 := \{Z_0 + j\,0\}, \tag{8.22}$$

$$\mathcal{M}_n := \{\cosh(\Phi) + j\;\sinh(\Phi)\}. \tag{8.23}$$

Let ∗ denotes multiplication of double numbers. Define following structures:

$$\mathcal{S}_d := \langle \mathcal{M}_d, * \rangle, \tag{8.24}$$

$$\mathcal{S}_0 := \langle \mathcal{M}_0, * \rangle \tag{8.25}$$

and

$$\mathcal{S}_n := \langle \mathcal{M}_n, * \rangle. \tag{8.26}$$

It can be easily verified by substitution, that the following implication holds:

$$(\forall \underline{M}(A_{0,k}, \Phi_k) \in \mathcal{M}_d \mid k = 1, 2, ...)(\forall \underline{M}(A_{0,l}, \Phi_l) \in \mathcal{M}_d \mid l = 1, 2, ...)$$

$$(\underline{M}(A_{0,k}, \Phi_k) \leftrightarrow \exp{(A_{0,k})}(\cosh{(\Phi_k)} + j\ \sinh{(\Phi_k)}) \Longleftrightarrow$$

$$(\underline{M}(A_{0,k}, \Phi_k) \otimes \underline{M}(A_{0,l}, \Phi_l) \leftrightarrow$$

$$\exp{(A_{0,l} + A_{0,k})}\ (\cosh{(\Phi_l + \Phi_k)} + j\ \sinh{(\Phi_l + \Phi_k)})). \tag{8.27}$$

The one-to-one mapping of the matrix models of quantification onto corresponding double numbers thus implies a one-to-one mapping of matrix products onto products of double numbers. Structure operations of structures $\mathcal{S}_m$ and $\mathcal{S}_d$ are therefore preserved by the mapping. Using relation (8.27), one can prove the following theorem:

Theorem 7:

1. Structures $\mathcal{S}_m$ (5.18) and $\mathcal{S}_d$ (8.24) are isomorphic.
2. Structure $\mathcal{S}_0$ is isomorphic with the structure $\langle \{A_0\}, + \rangle$ of ideal values.
3. Structure $\mathcal{S}_n$ is isomorphic with the structure $\langle \{\Phi\}, + \rangle$ of uncertainties.

The messages imparted by this Theorem are important:

1. Structure $\mathcal{S}_d$ (8.24) of double numbers is also a model of the quantification process. These double numbers are the (theoretical) models of the observed data.
2. The modulus of a double number belonging to the set $\mathcal{M}_d$ (8.21) is equal to the multiplicative image Z_0 of the ideal value.
3. Double numbers belonging to the set $\mathcal{M}_n$ model the effect of the uncertainty on the observed data.

Consider a double number $e_j(\Phi) \in \mathcal{M}_n$. Taking into account (8.23) and the definitions of hyperbolic functions, we can write

$$e_j(\Phi) = \exp{(j\ \Phi)}. \tag{8.28}$$

An easy way to verify this expression is to develop functions $\cosh(\beta)$ and $\sinh(\beta)$ into a power series, and to substitute $j\,\Phi$ for β, while making use of (8.4). The following identities may be shown by using the same method:

$$\cosh(j\,\Phi) = \cosh(\Phi), \qquad \sinh(j\,\Phi) = j\,\sinh(\Phi), \tag{8.29}$$

and

$$e_j(\Phi_1) * e_j(\Phi_2) = \exp(j\,(\Phi_1 + \Phi_2)). \tag{8.30}$$

A general result of quantification $x + j\,y \in \mathcal{M}_d$ may thus be presented in the form

$$x + j\,y = Z_0 \exp(j\,\Phi), \tag{8.31}$$

where the modulus Z_0 can be calculated by

$$Z_0 = \sqrt{x^2 - y^2}. \tag{8.32}$$

This is the same result as that obtained by (5.23), which was interpreted as the Minkowskian length of the radius vector. Other important relations, which result from (8.21)for the uncertainty 2Φ and from the formulae of hyperbolic functions, are:

$$y/x = \frac{\exp(2\Phi) - \exp(-2\Phi)}{\exp(2\Phi) + \exp(-2\Phi)} \tag{8.33}$$

with its inverse

$$2\Phi = \ln\sqrt{\frac{x+y}{x-y}}. \tag{8.34}$$

It is obvious from (8.32) and (8.34), that the moduli and angles of double numbers are independent of each other. In other words, the modulus is invariant to the transformation performed by multiplication by the double number $\exp(j\,\Phi) \in \mathcal{M}_n$. This number thus plays the role of the operator of rotation in the same manner as its matrix image $\underline{M}(0, \Phi)$. On the other hand, the angle Φ is invariant to the transformation realized by changing the multiplier Z_0. We have seen, that changes of Φ model *virtual movement* of the observed data along the circular path. Changes in Z_0 may correspond to an increasing or decreasing ideal value, ie they model *real movement* along a straight line (if the angle Φ stays unchanged).

There is a counterpart to (8.33), which will be needed: by interpreting the same real pair $\langle x, y\rangle$ as a point in the Gaussian plane, ie as the complex number

$$x + i\,y = \sqrt{x^2 + y^2}\,(\cos(\phi) + i\,\sin(\phi)), \tag{8.35}$$

there is an important relation, which links both representations:

$$\tanh(\Phi) = \sin(\phi). \tag{8.36}$$

Note, that the above equation defines the same point, $\langle x, y\rangle$, by using each of the two geometries (see 8.20).

8.2 Analyticity of Pair Numbers

The notion of *analyticity (holomorphy)* of functions of a complex variable is well-known. Let $u = x + i\,y$ be a complex variable. Then a complex function $u'(u) = x'(x + i\,y) + i\,y'(x + i\,y)$ of this variable is *analytical*, if the following relations (called the *Cauchy-Riemann conditions*) hold:

$$\frac{\partial x'}{\partial x} = \frac{\partial y'}{\partial y} \qquad \frac{\partial x'}{\partial y} = i^2\,\frac{\partial y'}{\partial x}. \tag{8.37}$$

Analytical functions are well-behaved in the sense of unlimited differentiability and predictability. According to [96], the notion of analyticity can also be used with double functions of double variables having the form $v'(v) = x'(x + j\,y) + j\,y'(x + j\,y)$, if the functions satisfy a modification of the Cauchy-Riemann conditions:

$$\frac{\partial x'}{\partial x} = \frac{\partial y'}{\partial y} \qquad \frac{\partial x'}{\partial y} = j^2\,\frac{\partial y'}{\partial x}. \tag{8.38}$$

Double functions, for which (8.38) hold, can also be called analytical. They are well-behaved in the same sense as complex analytical functions. It is therefore obvious, that we may rewrite (8.37) and (8.38) in the form valid for pair functions.

Definition 5:

A function of a pair variable $v'(v) = x'(x + c\,y) + c\,y'(x + c\,y)$ will be called c-analytical, if generalized Cauchy-Riemann conditions

$$\frac{\partial x'}{\partial x} = \frac{\partial y'}{\partial y} \qquad \frac{\partial x'}{\partial y} = c^2\,\frac{\partial y'}{\partial x} \tag{8.39}$$

hold.

All three versions (8.37) through (8.39) of the Cauchy-Riemann conditions of analyticity have interesting interpretations. They can be explained using the following Theorem.

Theorem 8:
Let $v'(v) = x'(x + c\,y) + c\,y'(x + c\,y)$ be a differentiable pair function of the pair variable v. Let

$$d\underline{M}_c := \begin{pmatrix} dx & c\,dy \\ c\,dy & dx \end{pmatrix} \tag{8.40}$$

and let $d\underline{M'}_c$ be an analogue of (8.40) aggregated of differentials dx' and $c\,dy'$. Let

$$\underline{J_c} := \begin{pmatrix} \frac{\partial x'}{\partial x} & \frac{\partial x'}{c\partial y} \\ \frac{\partial x'}{c\partial y} & \frac{\partial x'}{\partial x} \end{pmatrix}. \tag{8.41}$$

Then the relation

$$d\underline{M'}_c = \underline{J_c} d\underline{M}_c \tag{8.42}$$

holds, if and only if the function $v'(v)$ is c-analytical.

Proof of Theorem 8:
Consider total differentials of the functions $x'(x, y)$ and $y'(x, y)$:

$$dx' = \frac{\partial x'}{\partial x} dx + \frac{\partial x'}{\partial y} dy \tag{8.43}$$

$$dy' = \frac{\partial y'}{\partial x} dx + \frac{\partial y'}{\partial y} dy. \tag{8.44}$$

Multiply the right-hand matrices of (8.42). The resulting diagonal elements $(d\underline{M'}_c)_{1,1}$ and $(d\underline{M'}_c)_{2,2}$ have a form, which coincides with (8.43), and there is no need to consider them in the proof. The non-diagonal elements have different forms:

$$(d\underline{M'}_c)_{1,2} = c\,\frac{\partial x'}{\partial x} dy + \frac{\partial x'}{c\partial y} dx \tag{8.45}$$

and

$$(d\underline{M'}_c)_{2,1} = \frac{\partial x'}{c\partial y} dx + c\,\frac{\partial x'}{\partial x} dy. \tag{8.46}$$

A) Let function $v'(v)$ be c-analytical. Then (8.39) holds. Substitution of both generalized Cauchy-Riemann conditions into (8.45) and (8.46) results in their equivalence with (8.44) multiplied by c. Hence, (8.42) holds.
B) Let (8.42) hold. The function $v'(v)$ is differentiable, therefore the total differential exists and has the form (8.44). The identity of both (8.45) and (8.46) with the formula of the total differential can be achieved, only if (8.39) holds, ie only if the function $v'(v)$ is analytical.

The most striking feature of the equation (8.42) is the **double symmetry of all three matrices**, ie the fact, that all three satisfy condition (5.26) in the same manner as the quantification matrix $\underline{M}(A_0, \Phi)$. Theorem 8 says, that this takes place, if (and only if) the pair function $x'(x+cy)+cy'(x+cy)$ is analytical. This means, that the simple algebraic feature (5.26) is equivalent to a much deeper requirement of analyticity of functions representing the effects of uncertainty on data. Such a requirement is acceptable from the mathematical standpoint because it reduces the class of possible models to well-known and well-behaved functions. We shall see, that even such "simple" models are sufficient to yield rich and far-reaching results from the gnostic theory.

8.3 Summary

Double numbers are a natural and straight-forward extension of the algebra and analysis of complex numbers and variables. Each point on the Minkowskian plane can be interpreted as a double number and vice versa. To combine operations with both complex and double numbers, a more general notion of pair numbers is introduced. Both structures of double and complex numbers are unique in the sense, that they satisfy a certain set of conditions. The structure of double numbers (the 2-algebra) is isomorphic with the structure of matrix models of the quantification process, therefore the structure of double numbers can also be used as a mathematical model of quantification.

An important feature of double functions of double variables is their analyticity. The sufficient and necessary conditions for analyticity of a double function are an analogue of the Cauchy-Riemann conditions of analyticity of complex functions. It is therefore reasonable to introduce a generalized version of Cauchy-Riemann conditions valid for pair functions of a pair variable. An interesting connection exists between the double symmetry of matrix models of uncertain data and the analyticity of the pair function modeling data uncertainty.

Chapter 9

Estimation/Quantification Duality

9.1 Quantification and Estimation Characteristics

It was shown in previous chapters, that the real quantification process (quantification disturbed by an uncertain component) can be theoretically modeled as a bi-dimensional mapping which is isomorphic; hence, an inverse to quantification always exists theoretically. However, in practice, the datum is observed as a one-dimensional object, which is not sufficient to precisely determine the ideal quantity imbedded in the datum. Since a precise inverse to quantification is not possible from a practical point of view, one must look for the best possible *estimation* process. The notion of the "best" is obviously connected with the problem of "how to estimate." To find a useful solution, an estimation theory is needed. The highly developed statistical theory of estimation is based on the **random** conception of uncertainty. Gnostics has its own, entirely different notion of uncertainty, therefore another estimation theory independent of statistical concepts must be developed. By the introduction of the concept of pair numbers and pair functions, a suitable technique for a joint consideration of both quantification and estimation theory has already been prepared.

It was shown in Chapter 6, that real quantification is modeled in gnostics as the rotation (in the sense of Minkowskian geometry) of the vector, which represents the uncertain datum. The role of the matrix rotation operator is played by the matrix (6.29). This matrix operator is an analogue to the Euclidean matrix (6.28). In Chapter 8, it was demonstrated, that the algebra of double numbers enables quantification to be modeled using the Minkowskian rotation operator, which has the form of the special double number $\exp(j\,\Phi)$ ((8.28), (8.29) and (8.30)) (where Φ is the numerical image of uncertainty). This operator also has its Euclidean counterpart written as the complex number $\exp(i\,\phi)$ (see (8.35)). Using pair numbers and denoting

$$\Omega_j := \Phi \qquad \Omega_i := \phi \tag{9.1}$$

and recalling (8.29) and the identities $\cos(\phi) \equiv \cosh(i\,\phi)$ and $i\,\sin(\phi) \equiv \sinh(i\,\phi)$, both rotation operators can be written as a single expression which in the case of a rotation by $2\Omega_c$ have the following form:

$$\exp(2c\Omega_c) = \cosh(2c\Omega_c) + \sinh(2c\Omega_c). \tag{9.2}$$

Both components of this rotation operator are very important in gnostics.

Definition 5:

Let U_c be the 2-algebra of pair numbers ($c \in \{i, j\} \mid i^2 = -1,\ j^2 = 1$) representing gnostic events.

Let $\sqrt{x^2 - c^2\,y^2}\,\exp(c\,\Omega_c)$ be the exponential form of the pair number $x + c\,y$, which has particular forms $\sqrt{x^2 - j^2\,y^2}\,\exp(j\,\Phi)$ (8.31) and $\sqrt{x^2 - i^2\,y^2}\,\exp(i\,\phi)$ (8.35).

Let $\chi\ : U_c \to R^1$ be a function of the argument Ω_c. Then

1. the function $\chi(\Omega_c)$ will be called *the gnostic G-characteristic of the gnostic event $x + c\,y$*,
2. the function $\chi(\Phi)$ will be called *the quantification characteristic* or shortly *Q-characteristic*,
3. the function $\chi(\phi)$ will be called *the estimation characteristic* or shortly *E-characteristic*,
4. the gnostic characteristic obtained as the first component of the rotation operator (9.2) will be called *the G-weight*:

$$f_c := \cosh(2c\Omega_c) \tag{9.3}$$

and the second one *the G-irrelevance*:

$$h_c := \sinh(2c\Omega_c). \tag{9.4}$$

The G-weight and G-irrelevance are the basic gnostic characteristics of the uncertainty of the datum observed as the output of the quantification process $Z = x + c\,y$ (5.11).

9.2 Data Weight and Irrelevance

9.2.1 Formulae

In accordance with Definition 5, the angle Ω_c is also a gnostic characteristic. The particular form of the function χ in this most basic case is $\chi(\Omega_c) \equiv \Omega_c$. The Q-angle is Φ and the E-angle is ϕ. We have already seen, that the angle 2Φ is a function of the ratio y/x (8.34), and that relation (8.36) combines both versions of the angle Ω_c (the Minkowskian and Euclidean versions). Not only Ω_c, but all gnostic characteristics, are thus functions of the ratio y/x, which is equal to both $\tanh{(\Phi)}$ and $\tan{(\phi)}$. Using the general definitions of the weight and irrelevance and the formulae of trigonometric and hyperbolic functions, one arrives at following formulae for weights and irrelevances:

$$\textit{G-weight:} \quad f_c = \frac{1 + c^2 (y/x)^2}{1 - c^2 (y/x)^2} \tag{9.5}$$

and

$$\textit{G-irrelevance:} \quad h_c = \frac{2(y/x)}{1 - c^2 (y/x)^2}. \tag{9.6}$$

Substituting $c = j$ into this formulae, one obtains the Q-weight and Q-irrelevance and by $c = i$ the E-weight and E-irrelevance. However, it is important to proceed one step further and to develop formulae which can be used in algorithms. We now take leave of the simplified formula (5.14) (which assumes $S = 1$) and shift to the general case (5.12). In order to simplify computations, an auxiliary variable

$$q = (Z/Z_0)^{2/S} \tag{9.7}$$

is introduced. Recall that Z is the (known) multiplicative form of the observed datum (5.11), Z_0 is the (unknown) “ideal” or “true” value (5.13) of the quantity, which is to be estimated, and S is the (unknown) scale parameter. Now,—by (5.12) and (9.7)—

$$q = \exp{(2\Phi)}. \tag{9.8}$$

The expression (8.33) can be rewritten:

$$y/x = \frac{q - 1/q}{q + 1/q}. \tag{9.9}$$

The G-weight is now presented in a simplified form

$$f_c = \left(\frac{q+q^{-1}}{2}\right)^{c^2} \tag{9.10}$$

and the G-irrelevance is

$$h_c = \frac{2(q-q^{-1})}{(q+q^{-1})(1-c^2)+2(1+c^2)}. \tag{9.11}$$

9.2.2 Geometric Interpretation

The first geometric idea of G-weights and G-irrelevances follows from the role of the pair number $\exp(2c\Omega_c)$ as the rotation operator. Recall that the Q-angle Φ (equal to Ω_j) is the numerical value of the data uncertainty, the error measured in the most popular (additive) way

$$\Phi = \frac{A-A_0}{S} \tag{9.12}$$

given by 5.10. There is another "scale" to measure this same error which results from the identity (8.36); the Euclidean angle ϕ is used this time. A question might be asked as to why **twice** the value of the angles is taken as the argument of the characteristic being considered. The answer merits attention.

Denote $\arg_c(u)$ the angular parameter of a pair number u, which represents a gnostic event using the polar coordinates $|u|_c$ and Φ:

$$u := |u|_c \exp c\,\Omega_c. \tag{9.13}$$

This relation unifies (8.31) and (8.35). One may thus write $\arg_c(u) = \Omega_c$. This quantity is the G-angle of the rotation, which is necessary to go from the original, ideal value expressed as $u(0) = Z_0 + c\,0$, to the 'final' value u. The difference, $\arg_c(u) - \arg_c(u(0)) \equiv \Omega_c$, could be viewed as a universal characteristic of uncertainty, but this is not a good idea, because what is needed, is not only the 'error' of u, but an evaluation of its **relationship** to another gnostic event, say u'. Compare two possibilities: $\arg_c(u) - \arg_c(u')$ and $\arg_c(u) + \arg_c(u')$ and consider the case $u = u'$. The former expression would evaluate the relation between the events as zero, meaning that it is **independent of the uncertainty**, while the latter returns $2\Omega_c$, which preserves the information on uncertainty. A more generally applicable representation of the uncertainty existing "between" u and u' should be used: the G-angle $\arg_c(u) - \arg_c(\bar{u'})$, where $\bar{u'}$ is the conjugate of u' defined by (8.11). The relation of $\bar{u'}$ to u can be interpreted in the following way: imagine a mirror placed on the horizontal ("real") axis of the plane, where the pair numbers are plotted. Then the $\bar{u'}$ is the image of the u' seen in the mirror from point u'. The expression $\arg_c(u) - \arg_c(\bar{u'})$ is thus the angular measure of the circular path from the mirror image of u' to u. In the special case of $u = u'$ this measure is $2\Omega_c$.

The reason for preserving the value of each datum's uncertainty when evaluating mutual uncertainty between two data results from the need to give to each datum its individual weight dependent on its own respective uncertainty.

This manner of evaluation for both "own" and "mutual" uncertainty is general enough to even cover the case of gnostic (weighed) covariances.

There are other reasons for accepting the doubled angular arguments for these gnostic characteristics. They have a fundamental importance, because they are connected to the notions of entropy, information and the probability of individual data, which will be examined in the next several chapters.

The second geometric interpretation of the G-weights and G-irrelevances requires recalling the basic notions of Riemannian geometry. Consider a special (one-dimensional) case of the Riemannian metric form (6.13). The single coordinate will be α and the role of the metric matrix will be played by a positive scalar $g^2(\alpha)$. The metric form reduces to

$$dL = g(\alpha)d\alpha. \tag{9.14}$$

Let us compare this expression with two particular cases:

$$d(\sinh(2\Phi)) = \cosh(2\Phi)d(2\Phi) \qquad d(\sin(2\phi)) = \cos(2\phi)d(2\phi). \tag{9.15}$$

This shows that the G-weights play the role of the one-dimensional metric matrix. They really perform the function corresponding to their names: they **weigh** all the individual segments of the angular path with respect to their position within the whole path. The G-irrelevances quantify 'distances' (understand: uncertainties) measured using special Riemannian geometries, they **measure** the errors.

It is interesting to rewrite (9.15) to obtain the differential $d\Phi$ of the additive value of uncertainty (which equals to dA/S—see (9.12)). The relation between differentials $d\Phi$ and $d\phi$ follows from (8.36) by differentiation. Substitution of the result into (9.15) yields

$$d(\sin(2\phi)) = \cos(2\phi)\frac{(1+\cos(2\phi))^2}{4}d(2\Phi), \tag{9.16}$$

which shows an even more intensive suppression of large errors than characterized by Φ alone in (9.15).

We shall return to the formulae of G-weights and G-irrelevances repeatedly, when designing appropriate algorithms. The unique features of gnostic methods are derived from the special nature of these characteristics: among others, an inherent, natural robustness with respect to outliers/inliers and peripheral/internal data clusters (depending on the choice of $c = i$ versus $c = j$) and optimality. It is therefore worth visualizing these characteristics by graphing the formulae presented in foregoing paragraph.

9.2.3 Behavior of Data Weights

Data weights f_c are given by 9.10. Note that

$$f_j = \frac{(q \; + q^{(-1)})}{2} \equiv \cosh(2\Phi), \tag{9.17}$$

$$f_i = \frac{2}{(q \; + q^{(-1)})} \equiv \cos(2\phi) \tag{9.18}$$

so that

$$\cos 2\phi = \frac{1}{\cosh(2\Phi)}. \tag{9.19}$$

The red family of curves in the upper part of Fig. 9.1 shows the dependence of the Q-weights (f_j) on the ratio Z/Z_0 for different values of the scale parameter S. The green family (the lower parts of Fig. 9.1) belongs to the E-weights. The blue vertical/horizontal axis represents the limiting case of curves for $S \to 0$ and for $S \to \infty$, respectively.

The horizontal blue line demonstrates, that for a very large value of the scale parameter, the weight of an individual datum does not depend on the value of the Z/Z_0 (on the quantification error, ie on the "distance" of the observed datum from the "true" value); all observed data (both "good" and "bad") are taken and treated equally with weight 1. This is equivalent to using the ordinary (non-weighted) arithmetic mean as an estimate of the location parameter of a data sample.

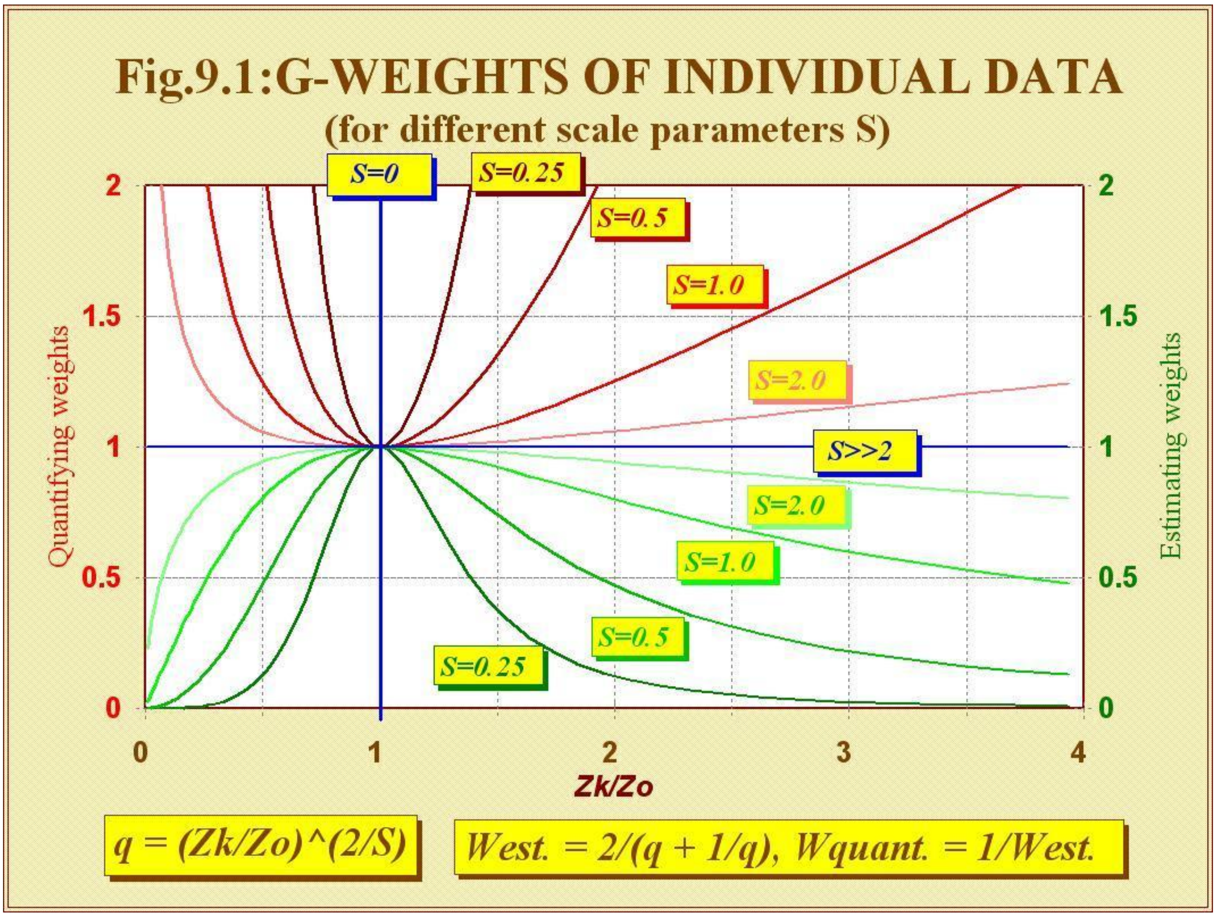


Fig.9.1:G-WEIGHTS OF INDIVIDUAL DATA

(for different scale parameters S)

q = (Zk/Zo)^(2/S)

West. = 2/(q + 1/q), Wquant. = 1/West.

In contrast, gnostic weights give different "preferences" to different individual data depending on their quality. The weighing provided by Q-weights is qualitatively different from that of the E-weights: Q-weights increase as the Z/Z_0 ratio declines from 1 while the E-weights decrease over the same range.

Starting with E-weights, consider some data spread about a central point (Z_0). The closer a datum (Z) to the central point, the larger its E-weight. The maximum weight (1) is given only to a datum positioned exactly at the central point ($Z = Z_0$). The 'central' data are thus preferred, while the 'peripheral' data are 'penalized'. The central data are thus considered as 'good', the peripheral ones are 'bad'. The 'bad' data are not cut-off completely (as if given a zero weight). It will be seen later that all data contribute to information, but the 'good' ones provide more, while the 'bad' ones give less. It will be proved, that the estimation transformation ensures, that the maximum possible information contained in each datum is retained.

The variability of the E-weight (called *robustness with respect to outliers*) is an important outcome, which can be demonstrated by consideration of the arithmetic mean of the E-weights of a collection of data spread about a central point. The mean data weight will be determined mainly by the central data while the peripheral (distant) data will be taken into account to a lesser extent, the further they lie from the central point. Such a data sample can be seen as resulting from good measurements (central data) disturbed by some strong unknown factors, which result in the larger uncertainty of the outliers. Example: To evaluate the mean weight of the ROA (Return on Assets) of a sample of firms and to compare it with another sample in the same or different industry, we are interested in the 'central' value of the group, while the 'peripheral' (extreme, atypical) values disturb our observation. We therefore welcome robustness, which suppresses the contribution of the outliers.

The case of Q-weights is quite different because these weights rise with increasing distance of the observed data value (Z) from the 'central' point Z_0. In this case, the 'peripheral' data of a data sample are preferred, being assigned larger weights; they are 'good', while the 'central' data are 'bad'. However, this result is not unnatural, it simply leads to a robustness which is opposite to the previous (E-) case, *robustness with respect to inliers.* Again an example: we observe a series of a volatile share price oscillating about a practically constant 'central' value. The 'normal' volatility is 'noise' in our observations. The objective is to signal, when the share price leaves the stationary path, and volatility exceeds the 'noise.' The disturbances in this case are the central data, which correspond to 'central' volatility (inliers), while the objects of interest (the 'good' external data) are the outliers. Therefore the definition of 'good' or 'bad' depends on the objective of the analysis, and the analytical technique used must be faithful to either type of application.

The curves in Fig. 9.1 are also useful in demonstrating the impact of the scale parameter S. When examining some data samples, this parameter will be used as a quantitative characteristic for the spread of data. The larger the data spread—the larger the S. The more precise the data—the smaller the S. It can be seen, that the curves in Fig. 9.1 are 'more tolerant' to deviations of the ratio Z/Z_0 from 1 in cases, when S is larger. In the case of precise data (small S), in contrast, deviations are unusual, therefore the stronger reaction of the curves to deviations.

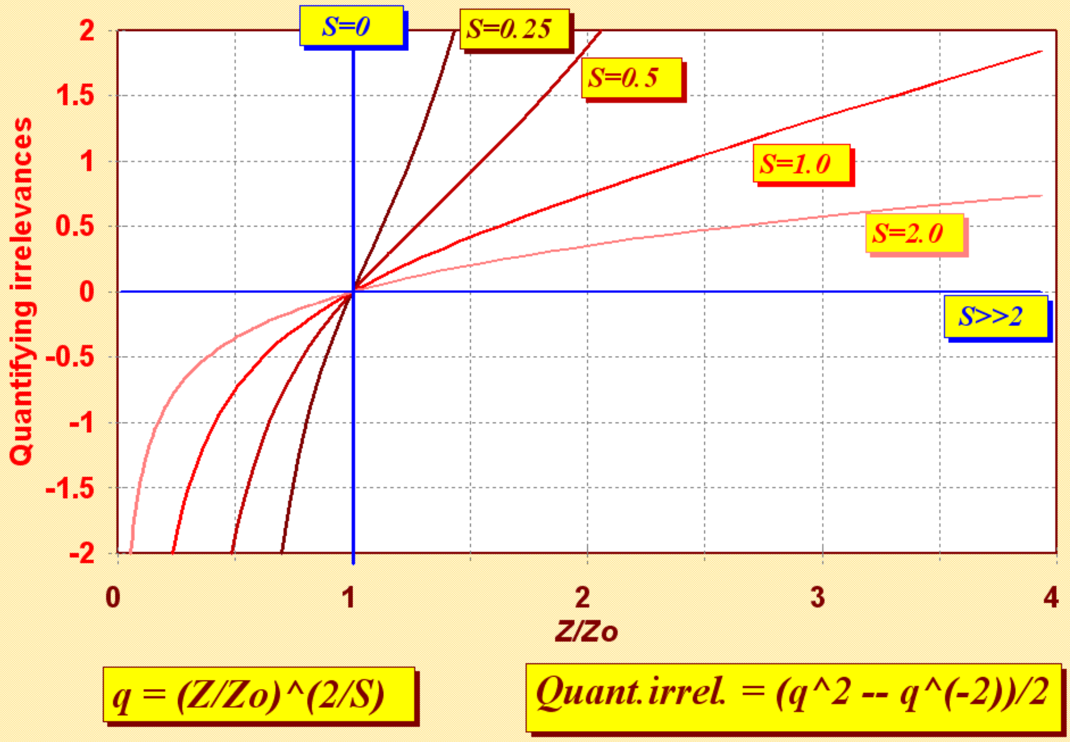
Both kinds of G-weights assigned to an individual datum assess the importance of the datum in a data treatment process. They measure the quality of the datum.

9.2.4 Behavior of Irrelevances

The red family of curves in Fig. 9.2 depicts the Q-irrelevances corresponding to the red family of Q-weights in Fig. 9.1, while the green family of E-irrelevances in Fig. 9.3 corresponds to E-weights in Fig. 9.1. Irrelevances measure data uncertainty (errors) using certain Riemannian geometries.

Fig.9.2 Q-IRRELEVANCES OF A DATUM

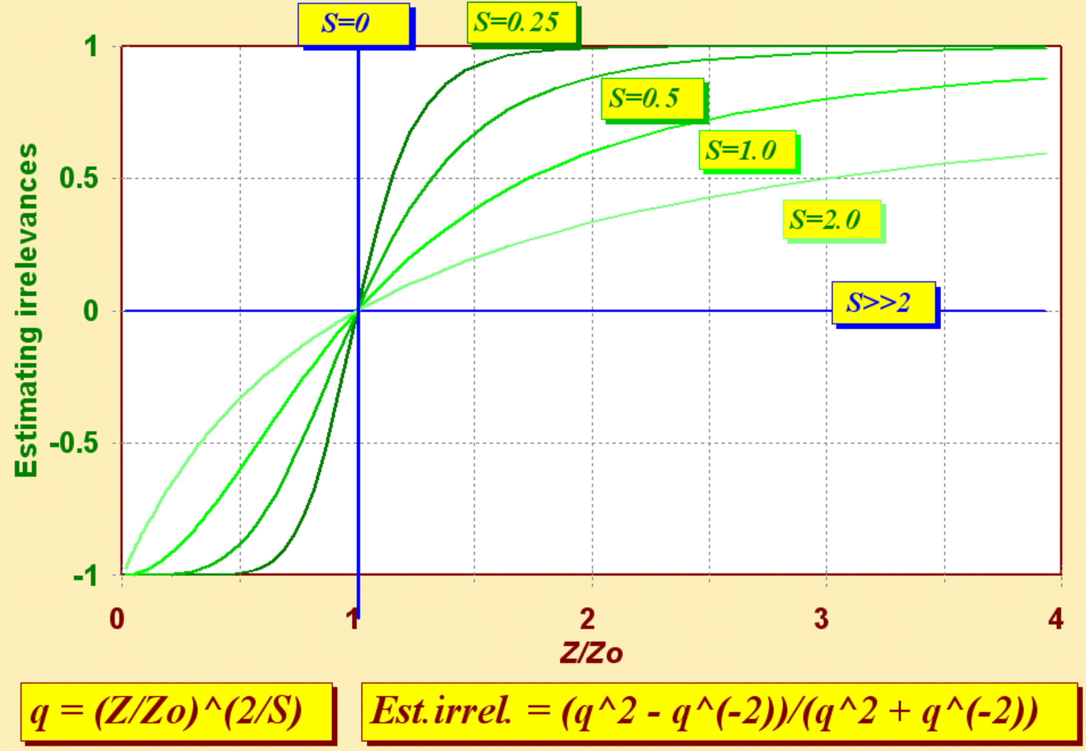
(for different scale parameters S)



Both Q– and E–irrelevances are zero for an exact datum ($Z = Z_0$), they are positive for $Z > Z_0$ and negative for $Z < Z_0$. They thus evaluate the "distance" between Z and Z_0 (error) respecting its sign (the direction of the path). The slope of the Q-irrelevances rises with the increasing deviations of the ratio Z/Z_0 from the 'central' value 1. The most interesting feature of the E-irrelevances (Fig.9.3) is, that—unlike the case of Q-irrelevances—the slope decreases to zero for large data errors. The E-irrelevances are thus limited, bounded by the limiting values -1 and $+1$. The maximum sensitivity of the E-irrelevance with respect to data error appears about the ideal value Z_0 and the larger the error, the less sensitivity to it.

Fig.9.3 E-IRRELEVANCES OF A DATUM

(for different scale parameters S)



These characteristics are simple to explain, because they correspond to instinctive human behavior:

- If one were to ask a group of people randomly chosen on the street to estimate the population of the Republic of Indonesia and the responses were 140, 210, 200, 190, 20, 250, 190, 210, 1000, 190 millions respectively (given a true value of 197.6 million), one would probably not treat these values linearly. It would generally be felt, that the extremes of 20 and 1000 are well off the mark, differing from the 'center of gravity' of the other (and larger number) of estimates leading to their deletion by giving them a very small weight. The perception is, that a response of 2000 would not be greatly different than that of 1000, nor 5 from 10, so that the asymptotic convergence of the E-irrelevance to ±1 corresponds to a natural interpretation.

By the way, this is the technique used by judges to evaluate participants in different sport competitions (gymnastic, skating, dancing, etc.): the best and the worst marks are deleted and the rest are evaluated by the arithmetic mean. Treating the results of the above “pool,” the ‘central’ value would be 197.5—very close to the true value—while the arithmetic mean of all 10 guesses would be 260 which demonstrates the non-robustness of the ordinary arithmetic mean. The adjective ‘ordinary’ is used in this context to distinguish this estimate from the *trimmed mean*, which is the name of the robust statistical method, according to which a part of the peripheral data is cut off before the arithmetic mean is evaluated.

- Nor is the increasing sensitivity of the Q-irrelevances with respect to increasing distance from the central data point unnatural. In monitoring the time series of share prices, the more the actual value exceeds the usually expected fluctuation, the greater the importance, that is placed on the observation, and the surer one can be, that something unusual is happening. The monitored share price has departed from its quasi-stationary state.

The behavior of irrelevances is thus acceptable.

9.2.5 Consistency with Statistics

The convergence of both G-weights to 1 with Z approaching Z_0 as well as the convergence of both G-irrelevances to zero in the same case have certain important consequences. Let ξ be a real number, N an integer and $O(\xi^N)$ be the so-called Landau's symbol characterizing the order of magnitude. This symbol is equivalent to the statement, that for ξ converging to zero, the variable $O(\xi^N)$ converges to zero as ξ^N. Expanding the formula $\phi(\Phi)$ resulting from (8.36) into the Taylor series, one can easily verify the relation

$$\phi = \Phi + O(\Phi^3). \tag{9.20}$$

Taking the same approach to (9.3) and (9.4) one obtains

$$f_c = 1 + c^2\frac{(2\Phi)^2}{2} + O(\Phi^4) \quad \text{and} \quad h_c = 2\Phi + O(\Phi^3). \tag{9.21}$$

Using the term *a sufficiently precise datum* to imply a datum, for which the (additive) error Φ (9.12) permits the terms $O(\Phi^3)$ and $O(\Phi^4)$ in 9.20 and 9.21 to be neglected. Then these relations prove, that the following statements are valid **for sufficiently precise data**:

1. The difference between the Euclidean and Minkowskian angles ϕ and Φ, which evaluate the data errors, tends to zero.
2. The gnostic evaluation of h_c (irrelevance) of the data error tends to the additive error 2Φ.
3. The gnostic weight f_c tends to a simple quadratic function of the additive error.
4. The sum of G-irrelevances of several data tends to zero simultaneously with the sum of the additive errors of data.
5. The sum of G-weights of several data is minimized simultaneously with the sum of squared additive errors.

It is obvious from the above general relations, that these statements hold, **if and only if** data are sufficiently precise.

The additive error Φ is a linear function of the (unknown) ideal value A_0 (9.12). The estimate of this quantity can be obtained easily by minimizing the sum of squares of additive data errors. This is the well-known *OLS* statistical estimating methodology (*ordinary least squares*) in which an estimate is *unbiased* (the sum of estimating errors equals zero). It is obvious from the foregoing statements, that the gnostic characteristic, which have been considered, approach this most popular and frequently used statistical technique (additive errors and their squares), if the data are sufficiently precise. Gnostics is thus consistent with statistics in the following sense:

> Gnostics is consistent with statistics in the sense, that for sufficiently precise data (and only for such data), the basic gnostic characteristics of data uncertainty (G-weights and G-irrelevances) converge respectively to a quadratic and to a linear function of the additively measured data error.

9.3 Virtual and Real Movements

The notion of estimation was introduced rather formally, making use of the duality of plane geometries (Minkowskian/Euclidean) and of the duality of pair (double/complex) numbers. It will be shown in what follows, that the estimation process derived from this duality is “the best” in a very acceptable sense. Before doing this, it must be demonstrated, that there is a third duality which is related to the previous ones.

Theorem 9:

Let $u = x + c\,y$ be a gnostic event, $c \in \{j, i\}$. Let $u'(u) = x'(x + c\,y) + c\,y'(x + c\,y)$ be a pair function of the pair variable u. Let $\underline{J_c}(u)$ be the matrix (8.41) composed of partial derivatives of the function u. Let K be the value of the determinant of the matrix $\underline{J_c}(u)$. Let K be a constant defined in the following way:

$$(\exists K \in R^1)(\forall (x + c\,y) \in U_c \mid 0 < |y|/x < 1)(Det(\underline{J_c}(u) = K). \qquad (9.22)$$

Then the following statements, **A** and **B,** are equivalent:

Statement A:

1. The matrix $\underline{J_c}(u)$ exists,
2. $K > 0$,
3. $u(0 + c\,0) = 0 + c\,0$ (the condition for homogeneity).

Statement B: The function u has the form

$$x' + c\,y' = (a + c\,b)(x + c\,y)K^{1/2}, \qquad (9.23)$$

where $(a + c\,b)$ is the rotation operator, for which

$$a^2 - c^2b^2 = 1 \qquad (9.24)$$

holds and where K is a positive constant.

The validity of this statement is not obvious, therefore a complete proof is shown:

Proof of Theorem 9:

Let **A** hold. Then the function u is analytical and (8.42) holds by Theorem 8. Substituting the generalized Cauchy-Riemann conditions of analyticity (8.39) into (9.22) one comes to a pair of partial differential equations:

$$(\frac{\partial x'}{\partial x})^2 - (\frac{\partial x'}{\partial y})^2 c^2 = K \qquad (\frac{\partial y'}{\partial y})^2 - (\frac{\partial y'}{\partial x})^2 c^{-2} = K. \qquad (9.25)$$

There exist exactly two solutions for each of these equations, a quadratic and a linear one. A quadratic solution does not satisfy (8.39), while a linear one does. Its constant term should be zero to satisfy the condition of homogeneity. The solutions have thus the homogeneous linear form

$$x' = (Ax + By)K^{1/2} \qquad y' = (ay + bx)K^{1/2}, \qquad (9.26)$$

where A, B, a and b are real numbers. It results from (8.39), that $A = a$ and $B = c^2 b$. Using these relations together with (9.26) one obtains (9.23). Substituting the solutions (9.26) into equations (9.25), one comes to (9.24). Thus it holds **A** ⇒ **B**.

Let **B** hold. It then follows from (9.23), that

$$x' = (ax + c^2 by)K^{1/2} \qquad y' = (bx + ay)K^{1/2}. \tag{9.27}$$

The function $x' + c\,y'$ is thus homogeneous (no constant term), analytical (conditions (8.39) satisfied) and the determinant of the matrix $\underline{J_c}$ is a constant equaling K (condition (9.22) satisfied). Implication **B** ⇒ **A** thus also holds.

The third duality of quantification and estimation models is in the special form of 9.23. It is known from Theorem 8, that both quantification and estimating transformations are analytical. It is also known, that condition 9.24 defines the pair number $a + c\,b$ as a rotation operator. According to Theorem 9 both of these transformations have a quite special form:

1. They are linear.
2. They include two possible phases:
   (a) Rotation by the operator $a + c\,b$ while the modulus remains unchanged.
   (b) The adjustment of the modulus by $K^{1/2}$.

The quantification rotation is already known, it was interpreted as a virtual movement of the point representing the evolution of an uncertain datum from its ideal value to the model of the observed value along an arc of a Minkowskian circle. The estimating rotation is another virtual movement formally dual to the quantification ones—the same point follows a Euclidean circle.

The adjective 'virtual' deserves a comment. It serves as a contrast to the adjective 'real', which would define changes in the path of an observed datum taking place in real time, if it were to actually follow the circular path. Something slightly weaker, however, has been derived from Axiom 1 for quantification: all outcomes from observing a fixed quantity under the influence of different uncertainties are to be modeled as points of a fixed circle in the Minkowskian plane. The observations are real, but the points of the circular path and the path itself are virtual, mathematical notions.

The uncertainty of quantification can be interpreted as a "draw" of Nature playing a game with the observer. Nature's objective is not to disclose its secrets too readily. Nature's draw maximizes (within some rules) the potential harm, which could result from the uncertainty imposed. The observer's response follows a strategy to minimize this impact, and it will be shown, that both the virtual movements do indeed realize their goals.

Both of these circular movements correspond to changing the 'quantity' of uncertainty (by changing angles Φ (Minkowskian) and ϕ (Euclidean)), while leaving the radius (modulus of the representative pair number, multiplier $K^{1/2}$ in (9.23)) fixed. A third special kind of a movement can also occur with (9.23): the adjustment of the modulus $K^{1/2}$, while the uncertainty is left unchanged. Such a movement is real (when the modulus increases), because it can be interpreted as an actual change in the ideal value Z_0; this is possible, because this number is a numerical image of a real quantity. Such a movement can be viewed as a *contraction* or *expansion* in dependence on the decreasing or increasing value of the modulus.

9.4 Summary

The idea of pair numbers can be used to introduce a gnostic notion of estimation, which is dual to quantification. The role, which estimation is expected to play, is to serve as a reverse transformation leading to more precise quantification by minimizing the uncertainty, with which a datum is contaminated. The formal difference between estimation and quantification lies in the use of complex numbers and variables in estimation instead of the double numbers used for quantification. A more material difference results from the duality of the Euclidean and Minkowskian geometries. Because of this difference, the basic quantification (Q-) and estimation (E-) characteristics of uncertainty (data weights and irrelevances) manifest fundamentally different features. Data weights quantify data quality, irrelevances measure data errors using certain Riemannian non-linear geometries. This non-linearity leads E-characteristics to robustness with respect to outliers, and Q-characteristics to robustness with respect to inliers. Depending on the application, either of these types of robustness will be useful.

The basic gnostic (G-) characteristics converge to the basic statistical ones (additive error and its square) in the case of sufficiently precise data. This means, that there is consistency between gnostics and statistics, when treating 'good' data. For insufficiently precise data the outcomes of both theories may be significantly different because of the non-linearity of the G-characteristics.

Both quantification and estimation operations can be thought of virtual processes represented by two special transformations—geometric rotation. The real changes of the data are modeled as contraction/expansion. This allows not only virtual but also real movements of the object of quantification and estimation to be considered.

Chapter 10

Entropy, Information, Probability

10.1 Strangeness of the Chapter

In the preface, readers were given notice, that the uniqueness of the approach to uncertainty being pursued would require an open mind, the ability to embrace new concepts, as well as the need to cast off personal biases, which could have been formed from traditional exposure to the ideas being discussed.

As the preceding chapters have demonstrated, gnostics develops its models of data uncertainty from new assumptions, distinct from those, on which other approaches to this subject are based. The need for such a radical departure is necessitated by the fundamentally different goal of gnostics: to be a complete mathematical theory, not of mass uncertain events, but of the uncertainty of individual data and of small data samples. To further this objective, it has been necessary to question a number of concepts, which have heretofore been accepted as given. It was shown, that a response to the question, "Which geometry is to be used to measure data errors," is not self-evident. The unexpected answer was found as a result of strict mathematical reasoning originating from more elemental assumptions. To derive and interpret models of uncertain data within the framework of gnostics, it became increasingly evident, that other concepts, most of which have had no application in other theories of uncertainty, needed to be employed. Notions such as path, real and virtual movement, and Lorentz invariance remind one more of mechanics than statistics. Other well known concepts, such as entropy, probability and information, which are used in this chapter, are derived in an entirely different way, and new formulae are given applicable to real data, the character of which is unknown.

The above may sound a bit bizarre to those, who are familiar with other theories, which deal with uncertainty. However, the list of significant differences between gnostics and traditional approaches also includes the following points related to entropy and information:

Reasons for inclusion: Despite the fact, that this book is intended to serve as an introductory text, it uses entropy and information as indispensable tools. These subjects are rarely touched in the basic statistical literature, nor is entropy and information treated in most statistical or econometric textbooks in use today. This is not meant to infer, that statistical theorists do not develop methods based on information theory; there is a large number of such contributions, but their practical impact has been small. Even one of the most recent and complete statistical software packages (S-PLUS [103]), which contains an enormous number of procedures, has only one program, which uses the notion of entropy, and none, which would use and treat data information.

Applicability to a single datum: Common definitions of data entropy and data information assume, that the probabilistic model is given a priori, before data is even gathered, and that it is related to random (collective) events. The application of these definitions to a specific single event is not defined. In contrast, gnostics is grounded in a highly developed theory of individual data.

Sequence of development: In gnostics, the notions of entropy → information → probability are developed in that order, while the paradigm of current statistics and information theory starts with the probabilistic model of collective events and only then introduces entropy and information. Gnostics begins with the entropy of individual uncertain events, derives information formulae from the entropy and then probability from the information.

Origin of notions: Standard approaches to uncertainty introduce probability through a series of axioms as the starting point of the theory. In gnostics, probability is one of products of the theory, which is derived from more elemental axioms.

Duality of notions: Gnostic concepts have a dual character: quantification/estimation, double/complex numbers, Minkowskian/Euclidean geometries, Q/E circles and paths, Q/E weights and irrelevances. These in themselves also generate other unexpected dualities: Q/E entropies, Q/E information, probability/improbability, Q/E robustness.

The above sets the stage for the text which follows, and serves notice, that the ensuing discussion will be neither commonplace nor simple, but perhaps of a nature to break new ground and so, exciting.

10.2 Gnostic Virtual Movements

10.2.1 Five Kinds of Double Numbers

Both diagonals of the Minkowskian plane have a special character. The radius vector $x + j\, y$ of an arbitrary point on a diagonal has modulus $(\sqrt{x^2 - y^2})$ which equals zero, because the relation $|x| = |y|$ holds on the diagonal. In other words: the distance between two arbitrary points on a diagonal is zero. It is therefore reasonable to consider the Minkowskian plane as a union of five sets of points:

Definition 6:

Let α and β be arbitrary real numbers and $\alpha + j\, \beta$ the radius vector of the corresponding point in the Minkowskian plane.
Let ${}_kU_j$ (for $k = 0,\ 1, ...,\ 4$) be following sets:

$$ {}_0U_j := \{\alpha + j\, \beta \mid |\alpha| = |\beta|\}, \tag{10.1}$$

$$ {}_1U_j := \{\alpha + j\, \beta \mid \alpha > |\beta|\}, \tag{10.2}$$

$$ {}_2U_j := \{\alpha + j\, \beta \mid |\alpha| < \beta\}, \tag{10.3}$$

$$ {}_3U_j := \{\alpha + j\, \beta \mid -\alpha > |\beta|\}, \tag{10.4}$$

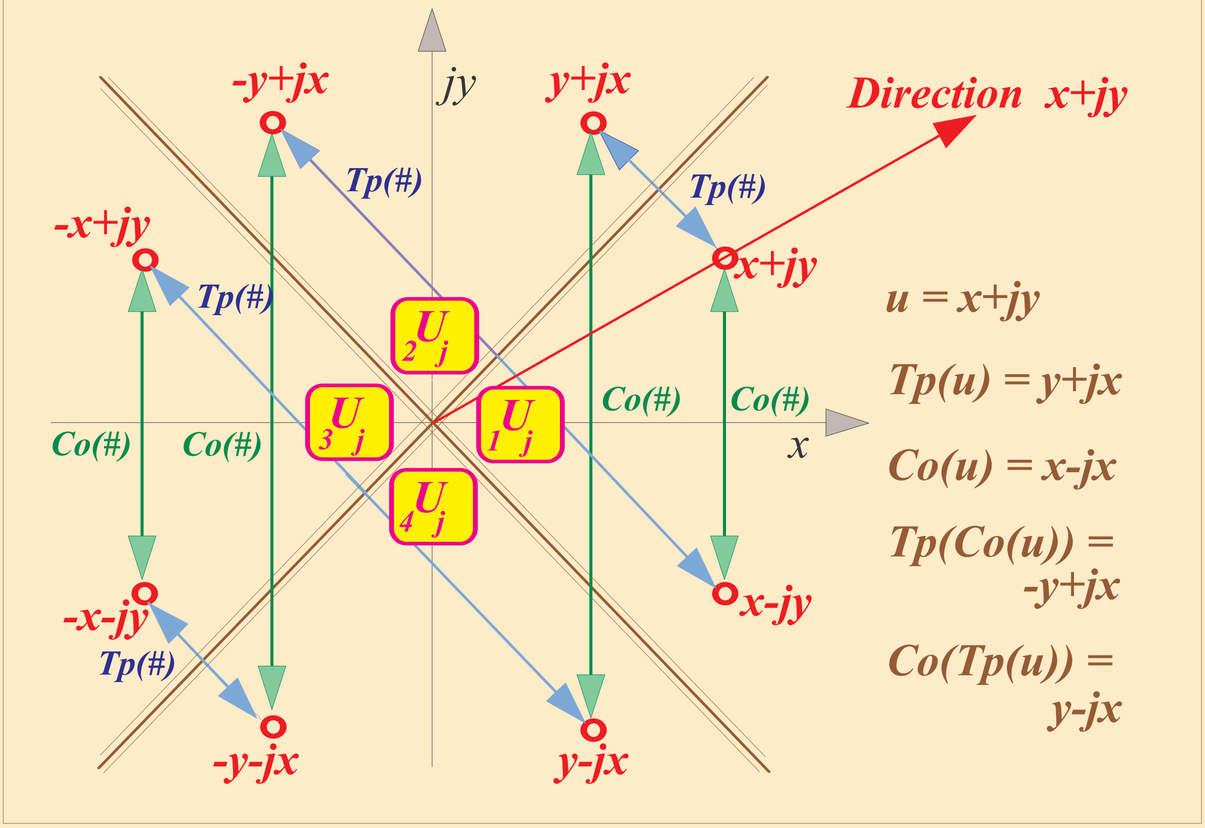
$$ {}_4U_j := \{\alpha + j\, \beta \mid |\alpha| < -\beta\}. \tag{10.5}$$

The double number modeling the point $u \in {}_kU_j$ will be called the *double number of the k-th kind.*

The Minkowskian plane is thus split by the diagonals into four open cones ${}_1U_j$, ${}_2U_j$, ${}_3U_j$ and ${}_4U_j$ as illustrated in Fig. 10.1.

A double number $x+jy$ may be viewed as a point or as the radius vector of this point showing the direction from the point $0 + j\, 0$ to $x + j\, y$ and having as its modulus the distance between the two points. It is a special feature of the Minkowskian plane that a continuous line consisting of finite points cannot cross a diagonal: therefore each finite continuous line lies in a single cone, and radius vectors of all points of such a line are of the

Fig. 10.1 FOUR KINDS OF DOUBLE NUMBERS



same kind. However, there are invertible transformations $_mU_j \leftrightarrow {}_nU_j$ for all $0 < m \leq 4$, $0 < n \leq 4$, $m \neq n$. The first of these transformations was defined by (8.12) as the *transposition* $Tp(*)$ of double numbers. Using notation introduced by Definition 6, one can rewrite (8.12) in the form

$$(k = 1,\ 2,\ 3,\ 4)(\alpha + j\,\beta \in {}_kU_j)(Tp(\alpha + j\,\beta) := \beta + j\,\alpha). \qquad (10.6)$$

This operation is complemented by the *conjugation* introduced by (8.11) which now has the form

$$(k = 1,\ 2,\ 3,\ 4)(\alpha + j\,\beta \in {}_kU_j)(Co(\alpha + j\,\beta) := \alpha - j\,\beta). \qquad (10.7)$$

There is an imaginative interpretation of both operations: mirror reflection (previously mentioned in section 9.2.2). Take a double number u. Then the relation of $Tp(u)$ to u is the same as the mirror image of u observed in a mirror placed along the South–West/North–East diagonal. The relation of $Co(u)$ to u is identical to the mirror image of u seen in the mirror placed along the horizontal axis.

It is clear from Fig. 10.1, that using pairs of operations $Tp(*)$ and $Co(*)$, one can produce double numbers, which belong to all four cones from an arbitrary double number, which is not on a diagonal.

There is a difference between both transformations, which have been considered. A transposition always maps one cone into another (${}_1U_j \leftrightarrow {}_2U_j$ and ${}_3U_j \leftrightarrow {}_4U_j$), while a conjugation maps two cones into themselves (${}_1U_j \leftrightarrow {}_1U_j$ and ${}_3U_j \leftrightarrow {}_3U_j$), or two cones each into another (${}_2U_j \leftrightarrow {}_4U_j$).

It was proved by Theorem 8 that the condition for the analyticity of a pair function of a pair variable is equivalent to the simple algebraic condition (5.26) of *commutativity with respect to transposition.*

Using the idea of mirror reflection, one can come to another interesting geometric interpretation of the analyticity of a double function of a double variable: such a function is analytical if and only if its transposed graph coincides with the graph's image observed in a mirror placed along the South–West/North–East diagonal. The small Alice would be disappointed by this analytical Wonderland: life behind the mirror would be the same as in front of it, only inverted left-to-right—no wonder at all! However, more interesting is, that the simple analytical functions—orthogonal rotations—lead to paths, which have wonderful features.

10.2.2 Paths of Gnostic Virtual Movements

The concept of gnostic virtual movement consists of the following:
the observed datum modeled by (10.8) is distorted by the uncertainty Φ, which is—at the moment of observation—an unknown constant. We imagine, that this final value resulted not from a discrete "jump" but from a continuous change in uncertainty starting at $\Phi = 0$ and ending at $\Phi = \Phi_k$.

The statements proved in Theorem 9 motivate the following definitions for three paths of gnostic virtual movement:

Definition 7A:

Let the observed datum be modeled by a fixed double number

$$u_k := Z_0 \exp\left(j\; S * \Phi_k\right) \tag{10.8}$$

where $Z_0,\; S \in R_+,\; \Phi_k \in R^1$ are given constants. Let

$$\phi_k := \arctan\left(\tanh\left(S\Phi_k\right)\right)/S. \tag{10.9}$$

Let $\mathcal{P}_Q$, $\mathcal{P}_{iE}$ and $\mathcal{P}_{jE}$ be the following sets called *quantification, i-estimation and j-estimation paths* (shortly *Q- and E-paths*), correspondingly:

$$\mathcal{P}_Q := u|u = Z_0 \exp\left(j\; S\Phi\right),\; \Phi \in [0, \Phi_k], \tag{10.10}$$

$$\mathcal{P}_{iE} := u|u = Z_k \exp\left(i\; S\phi\right),\; \phi \in [\phi_k, -\phi_k], \tag{10.11}$$

$$\mathcal{P}_{jE} := u|u = Z_0 \exp\left(j\; S\Phi\right),\; \Phi \in [-\Phi_k, 0], \tag{10.12}$$

The union of all three paths

$$\mathcal{IGC} := \mathcal{P}_Q \cup \mathcal{P}_{iE} \cup \mathcal{P}_{jE} \tag{10.13}$$

is defined as the *ideal gnostic cycle (IGC).*

The definitions show, that the ideal gnostic cycle consists of three branches, quantification and two sections of estimation paths. For the case of $\Phi_k > 0$, the virtual motion increases the angle from zero to Φ_k, while the "mirror image" $-\Phi$ decreases from zero to $-\Phi_k$. An analogous duality occurs in the case of the estimation path; the correspondence is established by the relation (10.9). Recall, that the important gnostic characteristics of uncertainty—data weights and irrelevances—introduced in Chapter 9 are functions of the angular distance between points on the paths and their conjugates (mirror images). These angular distances are $\Phi - (-\Phi) = 2\Phi$ and $\phi - (-\phi) = 2\phi$.

These notions are illustrated in Fig. 10.2 for the case of $\Phi > 0$.

The ideal value Z_0 plays here a role only in the graph's scale; hence, it is assumed that $Z_0 = 1$.[1] The image of the ideal value is thus $1 + j\,0$ (point $U0$ in Fig. 10.2). An increasing uncertainty moves the point $x + j\,y$ (modeling the datum) from the ideal value along the Minkowskian circle

[1] This simplification holds only for graphs. The formulae are written for the general case of an arbitrary positive S.

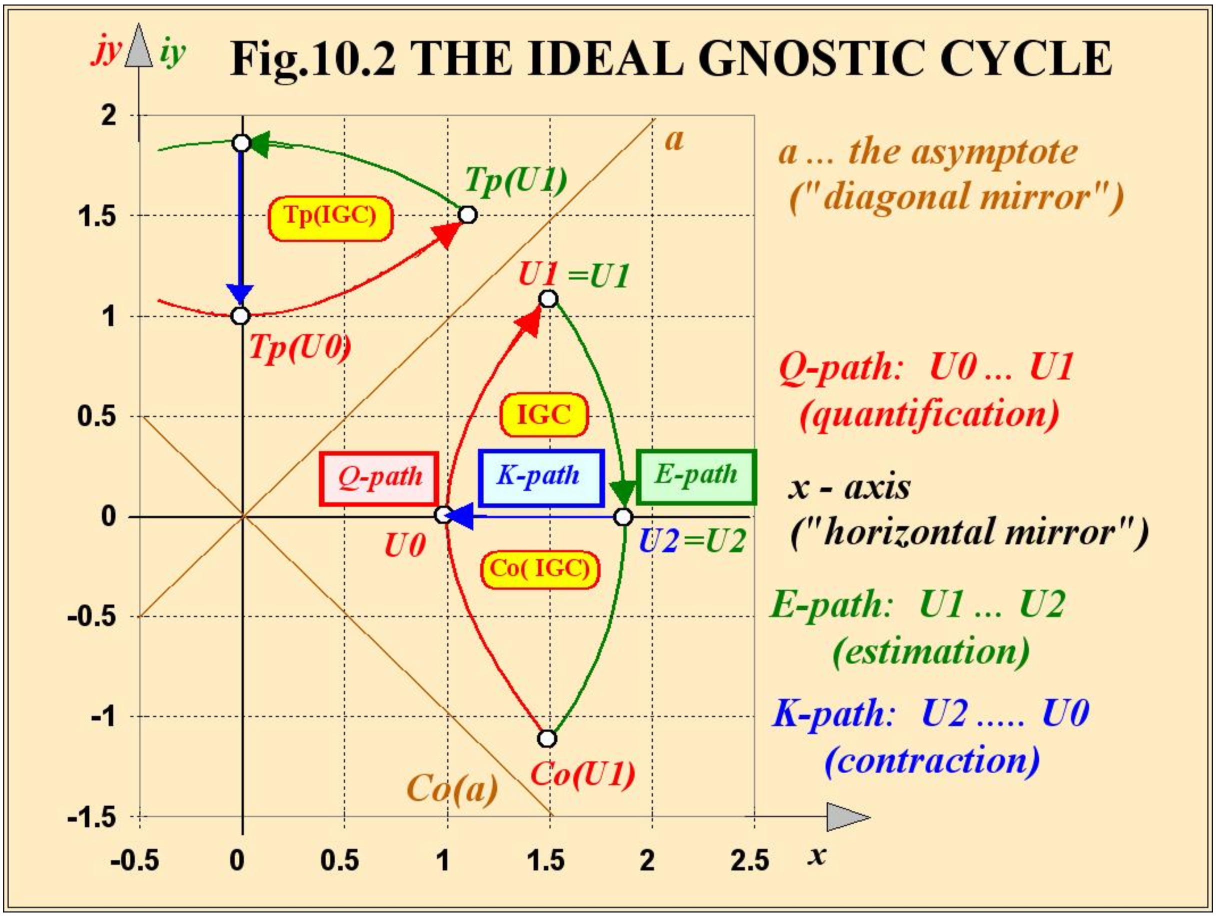


Fig.10.2 THE IDEAL GNOSTIC CYCLE

(the red curve, Q-path) to the point $U1$. The equation of the Q-path is

$$x^2 - y^2 = 1, \qquad (10.14)$$

because during quantification relations $x = \cosh(S\Phi)$ and $y = \sinh(S\Phi)$ hold. The quantification (Minkowskian) circle has the form of a hyperbola in Fig. 10.2, because the graph is drawn using Euclidean geometry. The end point of the Q-path ($U1$) is determined by the argument of the observed datum $S\Phi_k$. The radius vector of this point has the same length as all the points on the Q-path, for which the Minkowskian formula 10.14 holds. A different length is obtained, when the vector's length is measured using Euclidean geometry. Assuming, that at the point $U1$ relations 10.8 and $Z_0 = 1$ hold, then the Euclidean length of the radius vector is obtained as $\sqrt{\cosh^2(S\Phi_k) + \sinh^2(S\Phi_k)} = \sqrt{\cosh(2\Phi_k)}$. The Euclidean angle of this vector is ϕ_k determined by 10.9. These relations result from the characterization of the same point $U1$ using both geometries (see 8.36).

The E-path consists of estimating part $\mathcal{P}_{iE}$ and quantifying part $\mathcal{P}_{jE}$. The former part is represented in Fig. 10.2 by the green arc leading from the point $U1$ to $Co(U1)$. Its (Euclidean) radius is constant ($\sqrt{\cosh(2\Phi_k)}$) while the moving radius vector changes its (Euclidean) angle from $S\phi_k$ to -$S\phi_k$. Thus for all points of this path, equation

$$x^2 + y^2 = \cosh(2\Phi_k) = Z_k^2 \tag{10.15}$$

holds and the $\mathcal{P}_{iE}$-path is an arc of the Euclidean circle. The green $\mathcal{P}_{jE}$-path closing the Ideal Gnostic Cycle leads from the point $Co(U1)$ to the ideal value $U0$.

It was proved by Theorem 9 that the virtual movement along the Q-path (Q-rotation) is analytical in the sense of the analysis of double functions of double variables, while the virtual movement along the E-path (E-rotation) is analytical in the sense of complex analysis. All three branches of the ideal gnostic cycle are thus analytical and the transposition of all their points must coincide with the reflection observed in a mirror held along the diagonal. All points of the IGC shown in Fig. 10.2 are in the same cone, shown as ${}_1U_j$ in Fig. 10.1—the corresponding double numbers are all of the first kind (see Definition 6). However, this IGC has its "phantom" (reflected by the diagonal mirror) in the cone ${}_2U_j$ composed of double numbers of the second kind. Fig. 10.2 shows the IGC for $\Phi_k > 0$. In the case of a negative Φ_k the IGC and its conjugate exchange their positions.

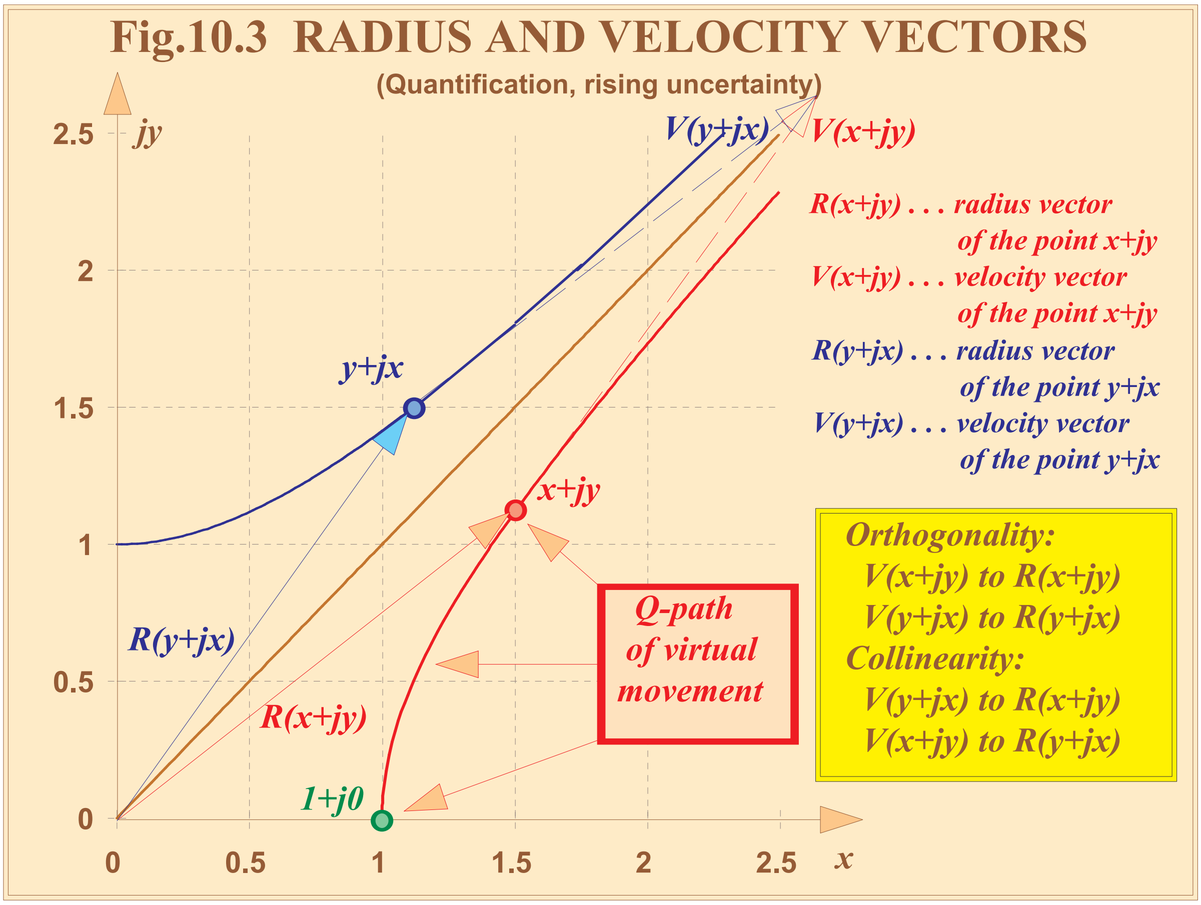
10.2.3 Velocity Vectors of Virtual Movements

Interesting relations between events in ${}_1U_j$ and in ${}_2U_j$ can be shown by considering the kinetics of the virtual movement. The velocity[2] of a point $u_j = x + j\,y$ on the Q-path may be evaluated as

$$V(x + j\,y) := \frac{du_j}{d\Phi} = S * (y + j\,x) = S * Tp(x + j\,y), \tag{10.16}$$

because relation $u_j = Z_0 * (\cosh(S\Phi) + j\sinh(S\Phi))$ holds for all points of the Q-path. The double number $x + j\,y$ may be interpreted both as a point on the Minkowskian plane and as the radius vector $R(x+j\,y)$ of this point. One can see in Fig. 10.3, that the velocity vector $V(x + j\,y)$ of the quantification movement in the point $x + j\,y$ is collinear with the radius vector of this point's mirror image.

[2] Recall that in mechanics the velocity vector of a moving point is the first time derivative of the radius vector. In gnostics, the role of time is played by the uncertainty, Φ, as the "driving force"; therefore the velocity of the virtual movement is given by the first derivative of the double number by its angle.



The dual collinearity exists between the velocity vector $V(y + j\,x)$ of the movement along the mirrored Q-path and the radius vector $R(x + j\,y)$. These relations are shown in Fig. 10.3 for the case $Z_0 = 1$ and $S = 1$.

Just as in mechanics, the velocity vectors of a point take the direction of the tangential line of the curve at this point. It would be customary to see the velocity vector of circular movement orthogonal to the radius vector of the tangential point. This, the Euclidean notion of orthogonality, is derived from the "right" angle (90°) between the direction of the vectors, which means, that to reach the orthogonality of one vector with a collinear one, one should rotate it by this angle. However this notion of orthogonality is not suitable in Minkowskian geometry. The Minkowskian angle between the radius vector of an arbitrary point on the Q-path and the direction of a diagonal is infinite: one cannot reach a point in ${}_2U_j$ by moving it along the circular Q-path (by rotating its radius vector). However, a more generally applicable notion of orthogonality of a pair of vectors may be introduced:

Definition 7B:
Let $u_j \in {}_kU_j$ and $v_j \in {}_mU_j$ be vectors, for which $k, m = 1, ..., 4$.
Let relation

$$[u_j, v_j]_{2,j} = 0 \tag{10.17}$$

hold.
Then the vectors will be called *orthogonal.*

Note that a zero value of the scalar product is also equivalent to the condition of orthogonality of a pair of vectors in Euclidean geometry. Returning to Fig. 10.3, we find, that the velocity vectors of all points along the Q-path are orthogonal to radius vectors of these points in the sense of (10.17). This results from the obvious orthogonality of all double numbers with their transpositions. Velocity vectors of a quantification movement are thus orthogonal at all points of the Q-path with radius vectors of these points as in Newtonian/Euclidean kinetics.

The kinetics of estimation shown in Fig. 10.4 is closer to the common view.

Let us consider the end point of the Q-path ($U1$ in Fig. 10.2) interpreted this time as a complex number, ie as $u_{i,k} = Z_k * \sqrt{x^2 + y^2} * (\cos(S\phi_k) + i \sin(S\phi_k)))$. The estimation movement starts at this point. Its velocity vector is

$$V(x_k + i\, y_k) := \left(\frac{du_i}{d\phi}\right)_k = S * (-y_k + j\, x_k). \tag{10.18}$$

This velocity vector is orthogonal to the radius vector $R(x_k + i\, y_k)$ of the tangential point in the Euclidean sense. Analogous relations take place at all points along the Q-path, from which the estimation starts. It is therefore sufficient to show the case $Z_0 = 1$ and $S = 1$ and a general point $x + i\, y$ in Fig. 10.4.

10.3 Energy and Entropy of a Datum

10.3.1 Energy of an Individual Datum

The outcomes of a “pure” mathematical theory are hidden in its axioms and its definitions. Whether a theory will bear fruit depends on its assumptions. The assumptions, that define the bounds of the theory, can be chosen in several ways. Most talented researchers need no particular method, they

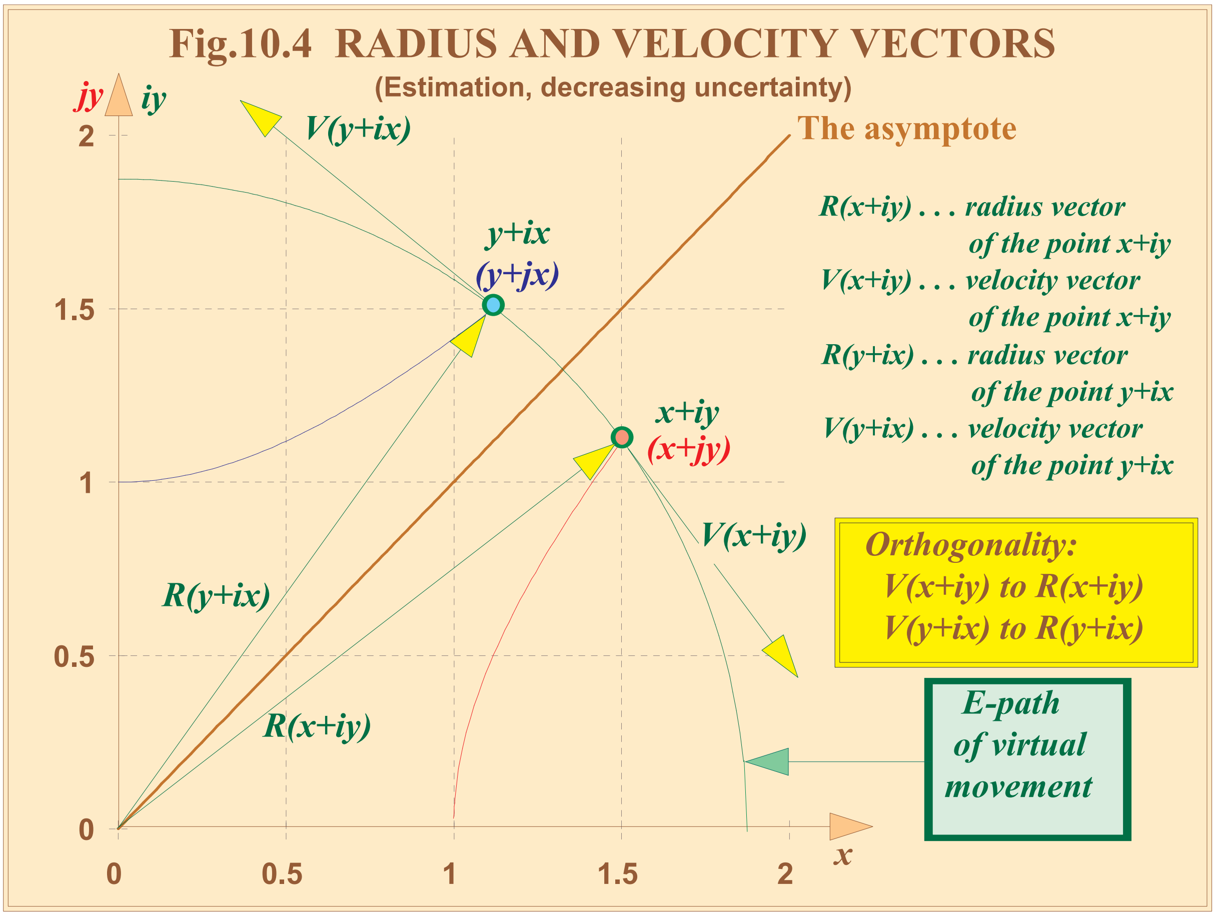


Fig.10.4 RADIUS AND VELOCITY VECTORS

(Estimation, decreasing uncertainty)

jy
iy
V(y+ix)
The asymptote
R(x+iy) . . . radius vector of the point x+iy
V(x+iy) . . . velocity vector of the point x+iy
R(y+ix) . . . radius vector of the point y+ix
V(y+ix) . . . velocity vector of the point y+ix
y+ix
(y+jx)
x+iy
(x+jy)
V(x+iy)
Orthogonality:
V(x+iy) to R(x+iy)
V(y+ix) to R(y+ix)
R(y+ix)
R(x+iy)
E-path of virtual movement
2
1.5
1
0.5
0
0
0.5
1
1.5
2
x

feel their way by intuition using their “sixth sense”. Some having learned by experience gained in previous efforts proceed step by step making small adjustments to axiomatic systems, that have already been successful, while others use a “Monte Carlo” approach, randomly varying assumptions along the way, however this seldom provides very productive results.

When a theory is intended to be used in practical applications for the solution of real as well as theoretical problems, its development is more complicated. Mathematics does not have tools to determine the realism of any a particular statement. It is then necessary to cross the boundaries of mathematics and look for assistance in the natural sciences: a realistic statement must respect the Laws of Nature. A methodology using the *Gedanken-experiment*, already noted in Chapter 4, has acquitted itself very well in the history of science as a source of assumptions for new theories, and we are now going to make use of it to logically interconnect the idea of a very general kind of data (such as eg from economics) to such seemingly foreign notions as energy, temperature, heat flow and entropy.

An initial example to discover the link between a datum and energy is to consider modeling and computing technology, where data values are represented by physical variables. Imagine an analog computer, within which a data value x is represented by a voltage V. A condenser with capacity C could be charged by the voltage V to accumulate energy $CV^2/2$ showing, that the physical mapping $x \leftrightarrow V$ implies the mapping $x^2 \leftrightarrow CV^2/2$, which results in a practical mapping of *data* $\leftrightarrow$ *energy*.

Another illustration is to imagine the screen of a monitor graphically displaying either a Minkowskian or a Euclidean plane. Applying Cartesian coordinates $\langle x, y\rangle$, the energy at the coordinates (necessary to deflect the impact point of streams of electrons of cathode rays) would be proportional to x^2 and y^2 (with the same coefficient of proportionality); and the energy of a datum represented by a pair number $x + c\,y$ would be proportional either to $x^2 - y^2$ (for $c = j$) or to $x^2 + y^2$ (if $c = i$).

10.3.2 Entropy of an Individual Datum

Energy can be converted into temperature, eg by discharging the condenser into a resistance placed in a calorimeter, the inner temperature of which will increase. The amount of heat, created by the amount of discharged electrical energy, can be measured and the corresponding absolute temperature corresponding to the energy can be calculated. Our *Gedanken-experiment* thus shows, that we can think of three <u>proportional</u> mappings:

squared datum $\leftrightarrow$ *energy*,
squared datum $\leftrightarrow$ *heat flow*,
squared datum $\leftrightarrow$ *absolute temperature.*

The next step is to recall the conditions, under which the famous Clausius[3] inequality for <u>non-statistical</u> *entropy* holds

$$\oint \frac{dQ}{T} \leq 0 \tag{10.19}$$

where Q is heat, T is absolute temperature and the integral is taken over either a complete thermodynamic cycle. This relation leads to one of the possible formulations of the Second Law of Thermodynamics: in a closed thermodynamic cycle the entropy increases. The adjective ‘non-statistical’ is emphasized to prevent a misunderstanding, which could result from the frequent habit (in both information theory and statistics) to define entropy

[3] Rudolf Clausius (1812-1888).

for a thermodynamic system, which can exist in N states with probabilities $p_1, ..., p_N$, in the statistical form introduced by Boltzmann[4]. Boltzmann's entropy is

$$S_B = \sum_{n=1}^{N} p_n \ln(p_n). \tag{10.20}$$

The most popular definition of information (Shannon's[5]) is $-S_B$, according to which information differs from (Boltzmann's) entropy only by its sign. It is for this reason, that it was emphasized at the beginning of the chapter, that in order to employ these notions (Boltzmann's entropy or Shannon's information), one must already have a priori knowledge of the probabilistic model. Even so, Boltzmann's point, that entropy is disorder, is of universal importance for it suggests that 'Nature prefers disorder'.

Returning to the screen of our monitor, it is seen, that at the point, that models the virtual movement of the datum within quantification, the energy of the double number is constant along the Q-path, proportional to $x^2 - y^2$, while the energy of the complex number $(x^2 + y^2)$ increases. The difference between these energies equals $2y^2$. This energy difference can be converted into a heat flow which may be positive or negative depending on the sign of c^2. Hence, the heat change relation

$$dQ_c = K_q d(2c^2y^2) \tag{10.21}$$

with a proportionality coefficient K_q holds. Converting the energy $x^2 - c^2y^2$ into absolute temperature leads to

$$T_c = K_t(x^2 - c^2y^2) \tag{10.22}$$

where K_t is a constant. When describing the Q-path of the process ($c^2 = 1$), the energy is constant, proportional to the squared radius of the Minkowskian circle; and the case of the E-path ($c^2 = -1$) is analogous, the energy is proportional to the squared radius of the Euclidean circle. In both situations, the temperature is constant and equal respectively to the radius of the Euclidean or the Minkowskian circles. Now, it will be shown, that the change in entropy between the "same" point on either circle is a simple function of the gnostic weight, f_c: the entropy change from the state Q_1 to Q_2 is

$$E_c^{'} = \int_{Q_1}^{Q_2} \frac{dQ_c}{T_c}, \tag{10.23}$$

[4] Ludwig E. Boltzmann (1844-1906).

[5] Claude Shannon, an engineer at Bell Labs, first published these ideas in the Bell Technical Journal in 1948.

where the temperature T_c is constant for both values $c = j$ and $c = i$, because the two cases are evaluated separately, either for the Q-path or for the E-path.

Relation (10.23) may be thus rewritten as

$$E_c^{'} = T_c^{-1} \int_{Q_1}^{Q_2} dQ_c = \frac{(Q_2 - Q_1)}{T_c}, \tag{10.24}$$

from which formula

$$E_c^{'} = \frac{K_q}{K_t} \frac{2c^2y^2}{x^2 - c^2y^2} \tag{10.25}$$

results. Adjusting the scale for measuring the entropy change, E_c, by choosing $\frac{K_q}{K_t} = 1$ and substituting the G-weight f_c from (9.5), one arrives at

$$E_c = f_c - 1. \tag{10.26}$$

All the foregoing was necessary to support the statement, that the following definition is reasonable:

Definition 8:
Let $f_c(x + c\,y)$ be the G-weight (9.5) at the point $x + c\,y \in {}_1U_j$, $c \in \{j, i\}$. Then the quantity E_c evaluated by (10.26) is *the change of entropy* of quantification (when c=j) or correspondingly of estimation (for c=i).

It was shown in Chapter 9 that G-weights play a significant geometric role as (one-dimensional) metric matrices in the determination of the specific Riemannian geometry to be employed (9.15). The *Gedanken-experiment* which led to definition 8 has demonstrated an unexpected connection between the choice of a metric for a geometric space and "something physical"—entropy. As Riemann had stated over a century ago (see Chapter 3), "Metrics are given objectively by laws of Nature." It can be concluded that the metric of the space of an uncertain datum is determined by entropy changes, which themselves are determined by the value of the observed datum.

It therefore holds:

The metric for measuring an individual data's uncertainty is determined by the uncertainty of the datum being considered.

We can now investigate the consequences of this interesting finding.

10.4 The Entropy Field

10.4.1 Data Entropy as a Scalar Field

Expression (10.26) defines a specific scalar field of entropy change over the Minkowskian/Gaussian plane. This allows an evaluation of the change in entropy caused by uncertainty for all points $x + cy$ for which $x > |y|$ (ie for the open cone ${}_1U_j$ of the Minkowskian plane as well as for the corresponding cone of the complex plane).

Entropy is one of the important gnostic characteristics of uncertainty and—as such—it depends only on the ratio y/x. Figure 10.5 illustrates *isoentropic lines* (points within the entropy definition's range for which the entropy values are constant), which are straight lines passing through the origin $(0 + c\,0)$.

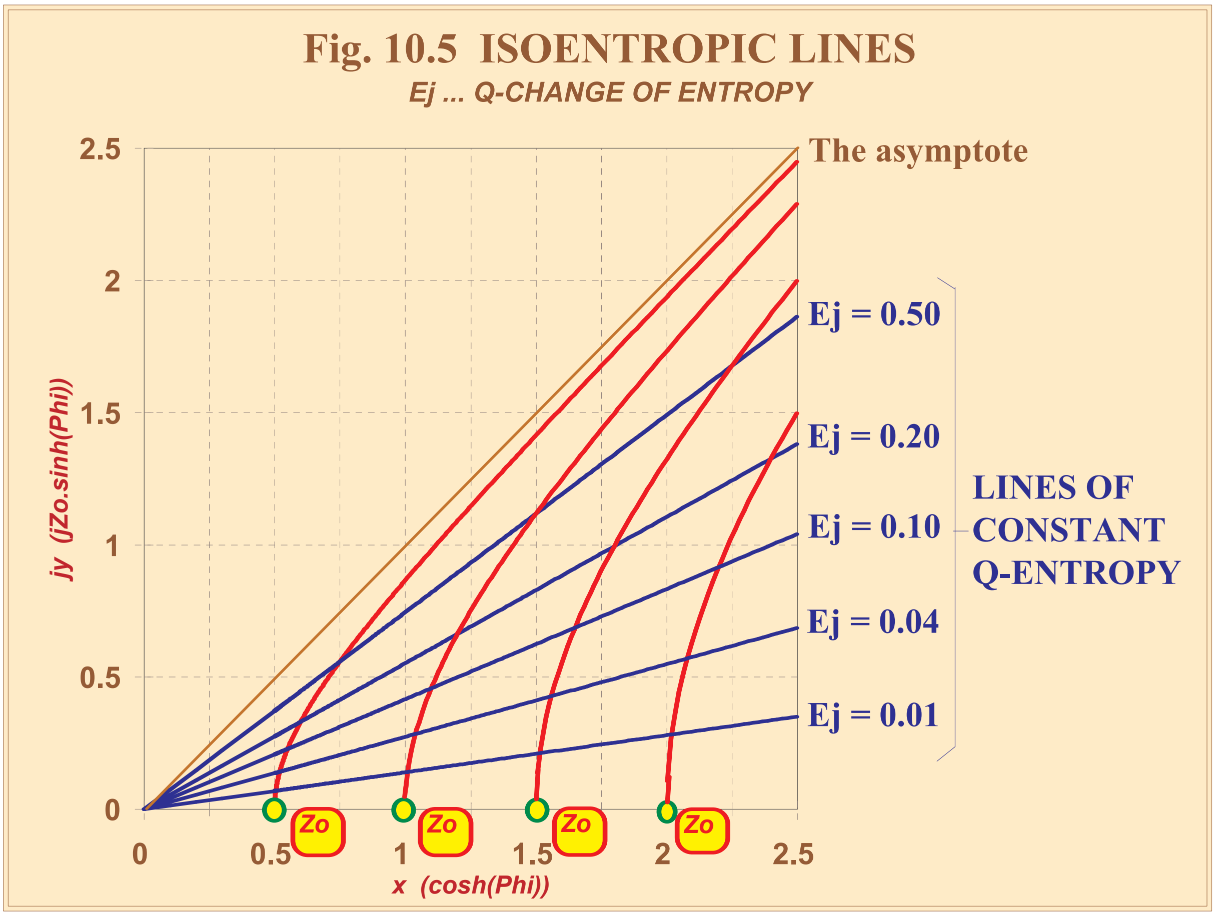


Fig. 10.5 ISOENTROPIC LINES

Ej ... Q-CHANGE OF ENTROPY

Four Q-paths are shown for ideal values ($Z_0 = 0.5,\ 1.0,\ 1.5,\ 2$) along with isoentropic lines for $E_j = 0.01,\ 0.04,\ 0.10,\ 0.25,\ 0.50$. The figure demonstrates how entropy rises as the point representing an observed datum driven by uncertainty moves along a Q-path. The case of estimation is treated by Fig. 10.6.

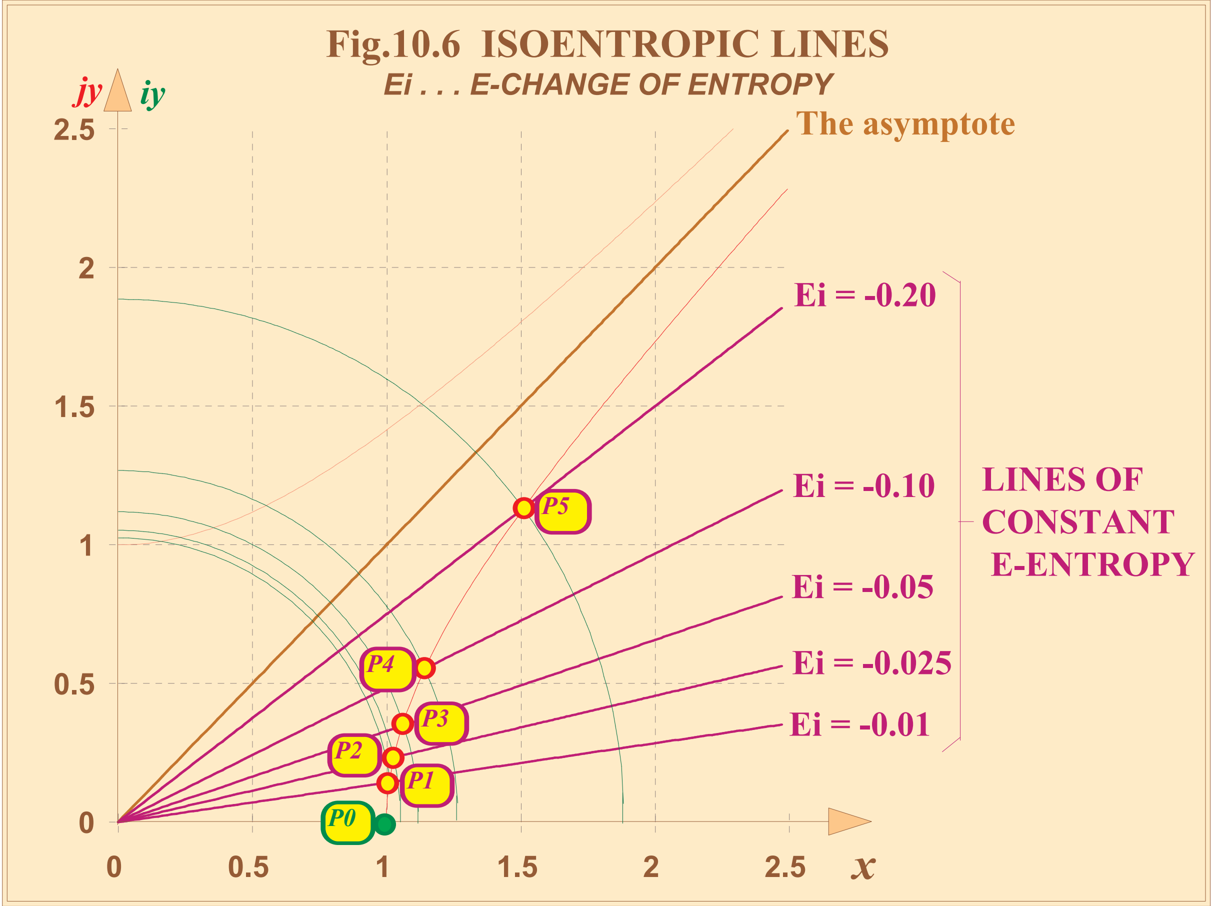


Fig.10.6 ISOENTROPIC LINES

Ei . . . E-CHANGE OF ENTROPY

Only one Q-path is shown (for $Z_0 = 1$), but there are several end-points for the Q-path ($P0, ..., P5$), which correspond to different observed values for the same datum. All entropy changes are negative in Fig. 10.6, they quantify the fall of entropy resulting from the virtual movement of the representative point along the E-path from the end point of the Q-path to the horizontal axis. Note that the entropy increase (which corresponds to an increase in disorder) along the Q-path for a given observed datum cannot be completely compensated by the corresponding fall in entropy during estimation. This observation, which is of fundamental importance, will be considered in more detail in the following section.

Illustrating with a Q-path for $Z_0 = 1$, and taking figures 10.5 and 10.6 together, the former shows that the isoentropic line representing Q-entropy of 0.2 is reached at a jy value of about .65 while the same Q-path plotted on figure 10.6 results in an entropy fall of 0.2 only if $i\,y = 1.12$ (corresponding to $P5$). The conclusion is, that for any level of uncertainty, the Q-entropy increase is larger than the E-entropy's fall. Hence, the increase in information resulting from data treatment cannot fully compensate for the increase in entropy resulting from the uncertainty inherent in the system. The IGC is not reversible, which is consistent with the Laws of Nature.

10.4.2 Entropy Field Gradients

It is useful to examine the *gradients of the entropy field*, ie vectors which delineate the direction of the steepest descents or ascents of the field. These quantities will be indexed by c, because they have different forms for each of the two geometries considered. Let α be a differentiable scalar function of a pair variable $x + c\,y$ ($c \in \{j, i\}$) interpreted as a scalar field over the Minkowskian/Euclidean plane. For these geometries, there are two kinds of gradients [45]: the *covariant gradient* ($\nabla_{c,1}$) defined for a scalar field α by

$$\nabla_{c,1}(\alpha) := \frac{\partial \alpha}{\partial x} + c\,\frac{\partial \alpha}{\partial y} \tag{10.27}$$

and the *contra-variant gradient* ($\nabla_{c,2}$)

$$\nabla_{c,2}(\alpha) := \frac{\partial \alpha}{\partial x} - c^3\,\frac{\partial \alpha}{\partial y}. \tag{10.28}$$

Using these formulae and (10.26), one obtains:

$$\nabla_{c,1}(E_c) = \frac{2c^2 h_c}{x^2 - c^2 y^2} * (-y + c\,x) \tag{10.29}$$

$$\nabla_{c,2}(E_c) = \frac{2c^2 h_c}{x^2 - c^2 y^2} * (-y - c^3\,x). \tag{10.30}$$

Examining the more ordinary case of the gradient of E-entropy, it is obvious that $\nabla_{i,1} \equiv \nabla_{i,2}$, because $i = -i^3$ for the imaginary unit i: there is only one gradient in this case. It is shown at the point $x + i\,y$ in Fig. 10.7 orthogonal to the radius vector at that point and therefore tangent to the E-path.

Using expression (10.26), $E_i = f_i - 1$, and the fact, that the gradient of a constant (1) is zero, it is seen, that the gradient (denoted Ge in Fig. 10.7) points in the direction of the steepest increase in the estimation weight f_i, which is equal to $\cos(2S\phi)$. The E-entropy change is negative

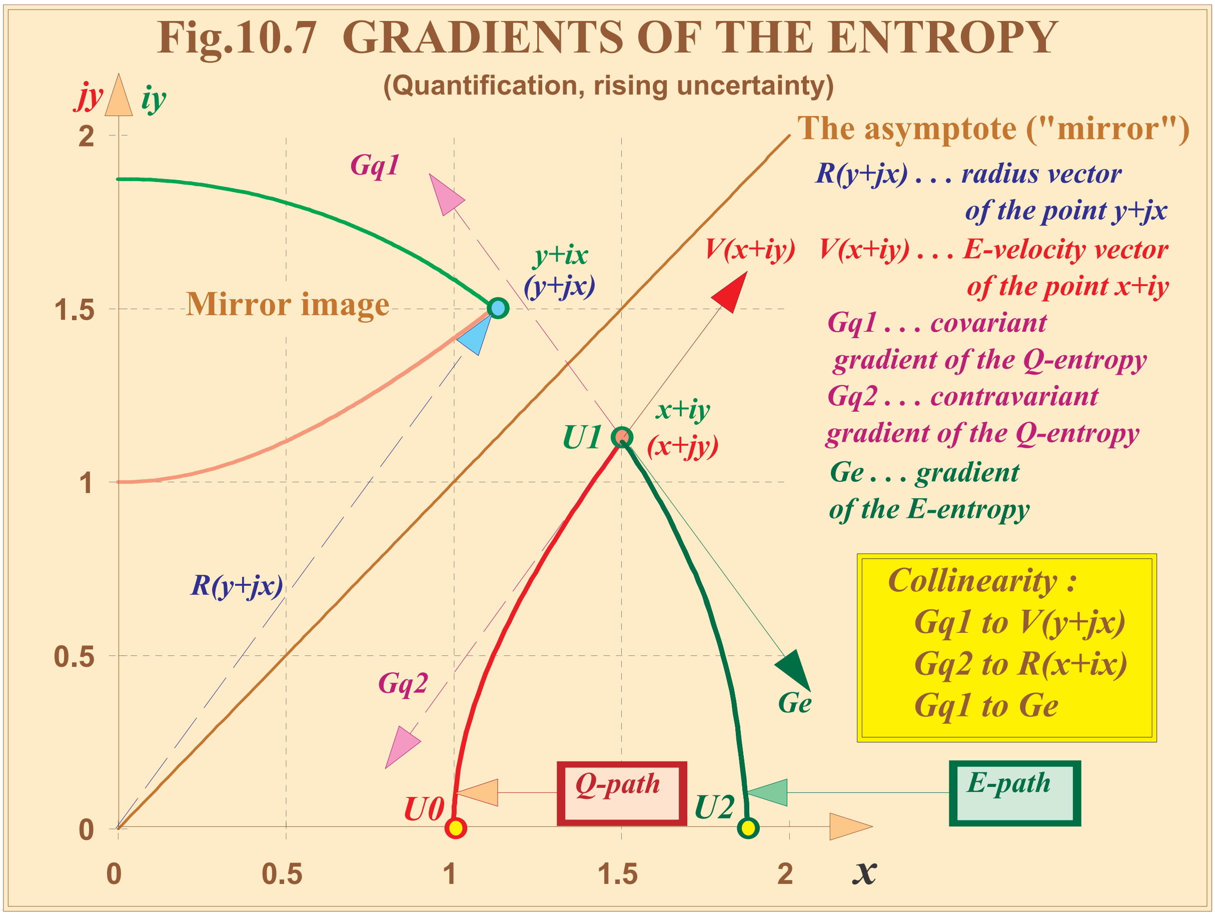


Fig.10.7 GRADIENTS OF THE ENTROPY

(Quantification, rising uncertainty)

for a non-zero uncertainty, because $\cos(2S\phi) < 1$ for $\phi \neq 0$. The E-gradient thus points in the direction of the steepest decrease of the entropy E_i for all points along the Q-path. We have arrived at an important result:

The local path from each point of the Q-path leading to minimization of the increase of the entropy coincides with the local E-path.

This is why the E-path is the first candidate offered as the best mode of estimation.

Unexpected results are obtained for the quantification case: The first—and from the point of view of Euclidean geometry, unusual,—feature is, that there is not only one gradient, but there are two. The second feature concerns the direction of the two gradients at the point $U1$. The covariant gradient $(Gq1)$ is collinear to the gradient of the E-entropy but its direction is opposite. This can be explained by referring to Fig. 10.5: All isoentropic lines intersect at the origin. To go from a point on one of these straight lines to a point that corresponds to a slightly higher level of entropy, one should move in a direction orthogonal to the former line, ie in the direction tangential to the E-path (opposite to the gradient Ge). This means, that the covariant gradient $Gq1$ points in the direction of the steepest increase in entropy E_j.

The contra-variant gradient ($Gq2$) is collinear with the velocity vector $V(x + j\ y)$ of the virtual movement along the Q-path and points in the direction opposite to the virtual movement. But the virtual movement from the point $U0$ in Fig. 10.7 to the point $U1$ increases the entropy. The direction of the contra-variant gradient $Gq2$ is thus the second best candidate as a mode of estimation. There are then two directions, which can be used as alternate paths to return to the point $U0$ representing the real and unknown ideal value:

1. Along the E-path opposite to the direction of $GQ1$ to the point $Co(U1)$ followed by the movement along the Q-path (*estimation path*).
2. Returning from $U1$ back along the Q-path following the direction of $Gq2$ (*anti-quantification*).

We will now examine the choice between these alternatives in more detail.

10.4.3 Which Estimation Path?

A choice between anti-quantification and the IGC-path can be based on a preliminary consideration of the errors which would result. The errors result from the fact that one can never know the actual value of uncertainty. The observed datum is given in the form of (5.12). To compare the effect of the choice of paths, the simplifying assumption $S = 1$, can be made (for this section), because the scale will be the same for either path. Let us consider the angular distance D_0 between the points $U1$ and its mirror image $Co(U1)$ in (Fig. 10.2). Relation

$$D_E = 2\Phi \tag{10.31}$$

may seem to be "natural", but only till its more general form

$$D_0 = |\int_0^{2\Phi} C_0\ d\psi| = 0 \tag{10.32}$$

with $C_0 = 0$ is presented demonstrating that D_0 has been obtained using the Galilean geometry to evaluate the path integral of angels' differentials along the vertical straight line connecting the points $Co(U1)$ and $U1$ in the two-dimensional plane endowed with the Galilean metric. (This line segment is a Galilean "circle"). However, there are two alternative circular paths in this plane, the Minkowskian (quantifying) and Gaussian (estimating)ones, also connecting these points, for which path integrals

$$D_j = |\int_0^{2\Phi} \cosh\psi d\psi| \tag{10.33}$$

and

$$D_i = |\int_{2\Phi}^{0} \cos\psi d\psi| \tag{10.34}$$

result by substitution of relations 9.15 attached to the quantification and estimation paths, correspondingly. These relations enable an important statement to be formulated:

<u>Theorem 10:</u>
Let Φ be the true numerical value of uncertainty of an observed datum 5.12. Then relation

$$D_j \geq -D_i \tag{10.35}$$

holds. The equality occurs, if and only if $\Phi =\ 0$.
Moreover:
The path integral D_j represents the maximum value among all integrals taken along the alternative paths connecting the points $Co(U1)$ and $U1$ obtained by limited variations of the quantifying circular path.
The path integral $|D_i|$ represents the minimum value among all integrals taken along the alternative paths connecting the points $Co(U1)$ and $U1$ obtained by arbitrary variations of the estimated circular path.

Theorem 10 says not only why the estimation path is to be chosen and its uniqueness, but also clarifies the extreme features of the Ideal Gnostic Cycle. It is a special case of a theorem considered in detail and proved in [61], paragraph 3.7.1, Theorem 8. It compares angles (interpreted as relative lengths, i.e. lengths divided by the circles' radiuses) of segments of two circular paths over the two-dimensional planes endowed with different Riemannian metrics. The extremity of the Q-path is an objective fact recognized by the analysis. Unlike this, the choice of estimating path is a matter of an analyst's subjective decision. As shown below in Chapter 12, choosing the unique (the shortest) circular E-path ensures the optimality of estimation.

An evaluation of the Theorem 10 requires the problem of robustness to be taken in account. Relation 10.33 results from quantification version of 9.15 increasing the weights of large errors contrary to the estimating version which prefers the weak errors. The estimating versions of formulae will thus provide results robust with respect to outliers opposite to quantification formulae preferring the large errors. Both concepts of robustness have applications in dependence on tasks to be solved. Suppression of outliers protects the results from bad observations while preference of extreme "errors" enables unusual events (eg a bad product, a rare signal) among many "normal" ones to be reliably recognized.

A natural question is,"How to go about optimal algorithmic applications of these extremities?" This is not a problem for the theory of individual data. A complete explanation of gnostic procedures, and answers to this and other questions will be given after the gnostic theory of data samples is provided.

10.4.4 Sources of the Entropy Field

The sources of a field determine its character, its spatial distribution and explain its origin. For an electrostatic field, the role of the source is played by an electric charge, for electromagnetic fields it is electric current and for a gravitational field, gravitation masses. Some field sources may be negative as well as positive (outflows or inflows). Sources may create fields and/or let fields vanish. It is for this reason, that sources of entropy fields have to be investigated.

Mathematical analysis has derived formulae for the calculation of a point source of a scalar field (eg α) for different geometries. In our case the formula has the following form ([45]):

$$\nabla^2(\alpha) = \frac{\partial^2 \alpha}{\partial x^2} - c^2 \frac{\partial^2 \alpha}{\partial y^2}, \tag{10.36}$$

where ∇^2 is the Laplace's operator. Denoting

$$r_c^2 := x^2 - c^2 y^2 \tag{10.37}$$

the squared radius of a circular path, and twice differentiating the entropy, one gets

$$\nabla^2(E_c) = -4 r_c^{-2} f_c. \tag{10.38}$$

Let us introduce the complementary indeterminate $\hat{c}$ such that $\hat{\imath} \equiv j$ and $\hat{\jmath} \equiv i$. The complementary radius of a circle may be thus written as

$$r_{\hat{c}}^2 := x^2 + c^2 y^2. \tag{10.39}$$

The relation

$$r_{\hat{c}}^2 \nabla^2(E_c) = \frac{\partial^2 E_c}{\partial(x/r_{\hat{c}})^2} - c^2 \frac{\partial^2 E_c}{\partial(y/r_{\hat{c}})^2} \tag{10.40}$$

can be interpreted as

1. the source of Q-entropy (eg of E_j) at a point on the E-path, which has a radius equal to 1 (when $c = j$),
2. the source of E-entropy (eg of E_i) at a point on the Q-path, which has a radius equal to 1 (when $c = i$).

Division of both coordinates x and y by $r_{\hat{c}}$ leads thus to the unification of all circular paths corresponding to different Z_0's—all have been mapped onto a single value with $Z_0 = 1$.

Multiplying (10.38) by $r_{\hat{c}}^2$ and taking into account that $r_{\hat{c}}^2/r_c^2 \equiv f_c$, one comes to

$$r_{\hat{c}}^2 \nabla^2(E_c) = -4 \frac{1}{(1 - 4c^2x^2y^2/(x^2 + c^2y^2)^2)}. \tag{10.41}$$

To prevent a misunderstanding, it is useful to analyze three interpretations of this equivalence:

<u>Definition 9:</u>

Let h_j and h_i be the Q- and E-irrelevance, correspondingly, expressed as in (9.6). Let $|y| < x$. Then

$$p := (1 - h_i)/2 \quad p_j \ = (1 - j\, h_i)/2 \quad p_i := (1 - i\, h_j)/2. \tag{10.42}$$

The first of these variables is a real number $p \in (0,\, 1))$ while the second and third are double numbers. There are thus three versions of 10.41:

1. As follows for the scalar case from (10.41),

$$r_i^2 \nabla^2(E_j) = -\frac{1}{p * (1 - p)}. \tag{10.43}$$

2. In the case $c = j$,

$$r_i^2 \nabla^2(E_j) = -\frac{1}{p_j * (1 - p_j)} \tag{10.44}$$

results.

3. Equivalently, in the case $c = i$,

$$r_j^2 \nabla^2 (E_i) = -\frac{1}{p_i * (1 - p_i)} \tag{10.45}$$

is obtained.

To show the equivalence of double interpretation of the source of quantifying field, following relations are to be taken in account:

$$p * (1 - p) \equiv (1 - h_i^2)/4 \tag{10.46}$$

and

$$p_j * (1 - p_j) \equiv \overline{p_j} * p_j \equiv |p|_j^2 \equiv (1 - h_i^2)/4 \equiv \cos^2 2\phi/4 \tag{10.47}$$

There also is a dual relation of this type for the estimation field:

$$p_i * (1 - p_i) \equiv \overline{p_i} * p_i \equiv |p|_i^2 \equiv (1 + h_j^2)/4 \equiv \cosh^2 2\Phi/4. \tag{10.48}$$

The expression $\overline{p_c} * p_c$ is a square of the pair number's p_c modul because $\overline{p_c}$ is its conjugate. Expressions 10.43 and 10.44 are interchangeable due to double interpretability of the variable $(1 - j\, h_i)/2$ as a real number and/or as a double number resulting from the equivalence

$$\exp(\Phi) \equiv \cosh(\Phi) + \sinh(\Phi) \equiv \cosh(\Phi) + j\ \sinh(\Phi) \tag{10.49}$$

The interpretation of the equations 10.44 and 10.45 is important: there are two fields, the sources of which balance the quantifying and estimating sources of entropy.

Let us investigate these interesting fields.

10.5 Four Integrals of Gnostic Movements

10.5.1 The Birth of E-Information and Probability

This subsection deals only with changes of entropy E_j, i.e with Q-entropy. This entropy has been defined over the cone ${}_1U_j$, as a function of two coordinates x and y. It is now reasonable to restrict the definition of the range of this entropy to points of the circular E-path, which have the unitary radius ($r_i = 1$). This restriction enables the left hand side of (10.43) to be interpreted as an evaluation of the strength of the Q-entropy sources applicable to all the data, independent of the value of their Z_0, which implies, that the normalized Q-entropy will be a function of a single variable, eg of y/x or of $S\Phi$. In other words, Q-entropy is a gnostic characteristic of the uncertainty of an individual datum. The right hand side of (10.43) is also a real function of the single real variable (p). Introducing an auxiliary function

$$(0 < p < 1)(H(p) := -p * \ln(p) - (1-p) * \ln(1-p)) \tag{10.50}$$

one can easily verify the relation

$$-\frac{1}{p*(1-p)} = \frac{d^2(H(1/2) - H(p))}{dp^2}. \tag{10.51}$$

We thus know a scalar field and its source (second derivative) which balances the source of Q-entropy. It only remains to give suitable names to this function and to its argument, and to interpret their characteristics.

Definition 10:

Let $Z = Z_0 \exp(S\Phi)$ be the model (5.12) of a given observed datum.

Let p be the gnostic characteristic of this datum defined by (10.42) and (9.4).

Then p will be called the *gnostic probability of an individual datum*, shortly *probability.*

The function

$$(0 < p < 1)(I_j := H(1/2) - H(p)) \tag{10.52}$$

will be called the *estimation change of information* (shortly *E-information*) of the given individual datum.

The interchangeability of the variables p and p_j enabled the scalar form of H's argument to be preferred to keep the tradition of measuring the probability by a real number. On the other hand, its double form maintains the duality of quantification and estimation processes.

The adjective "estimation" stresses the fact that the information I_j has been obtained by (double) integration of the sources of entropy E_j along the E-path. The integration constants in (10.52) are chosen so as to satisfy the natural requirement, that the information change of a precise datum ($\Phi = 0, h_i = 0$) is zero. The character and features of probability and information are analyzed below.

It is interesting to present the information I_j (10.52) as a function of the uncertainty Φ. The E-irrelevance h_i is defined by (9.4) for a non-unitary

S and $\Omega_i = S\phi$ as $\sin(2S\phi)$. It can be therefore rewritten as $2\frac{\tan(S\phi)}{\sqrt{1+\tan^2(S\phi)}}$. The basic relation (8.36) (binding the Q- and E-angles) enables the substitution of $\tanh(S\Phi)$ for $\tan(S\phi)$ so, as to arrive at the equivalence

$$\sin(2S\phi) = \tanh(2S\Phi). \qquad (10.53)$$

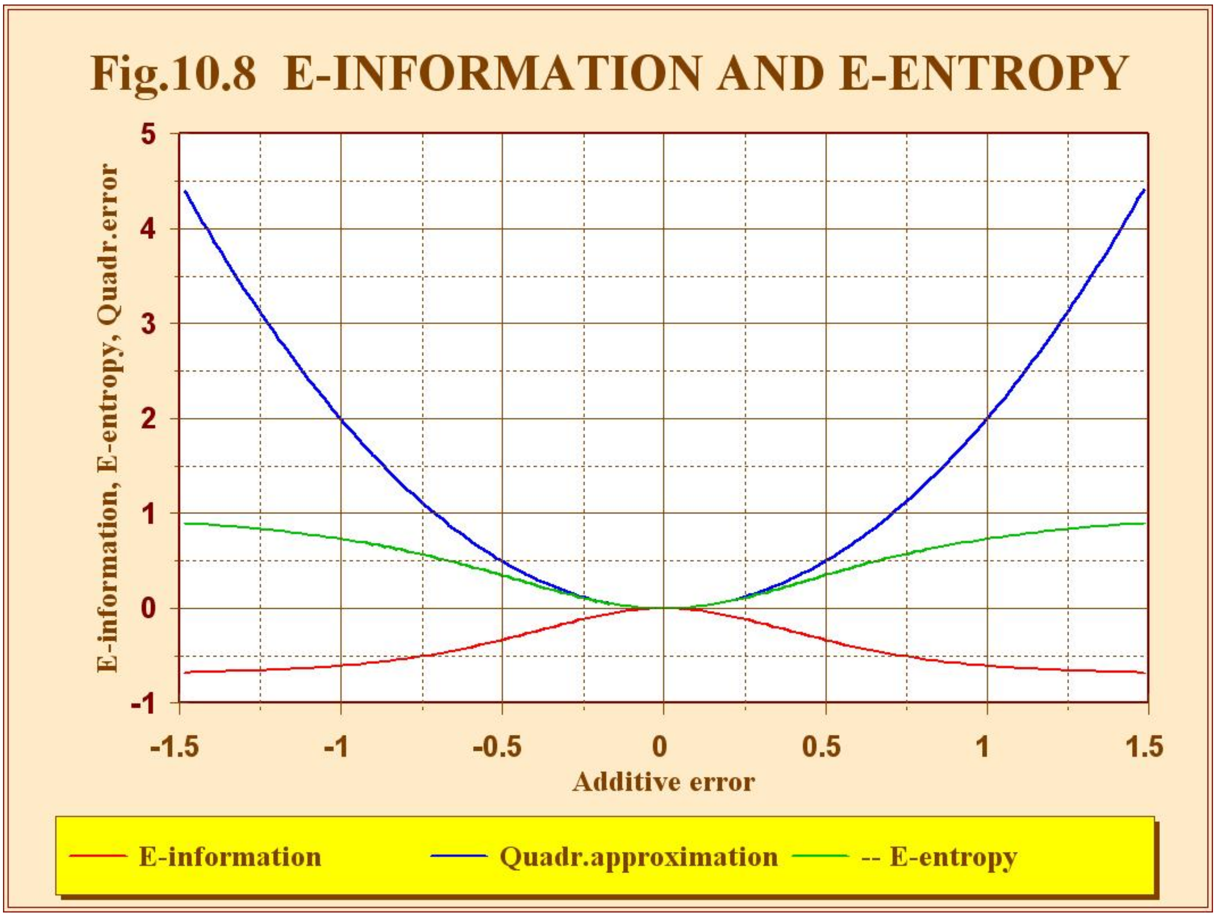
Taking into account that $|h_i| < 1$ and applying the usual formula of hyperbolic functions one obtains

$$2S\Phi = \ln\left(\sqrt{\frac{1+h_i}{1-h_i}}\right) \qquad (10.54)$$

and

$$I_j = 2S\Phi * \sinh(2S\Phi) - \ln(\cosh(2S\Phi)). \qquad (10.55)$$

This function together with the E-entropy (10.26) and with a quadratic approximation are depicted in Fig. 10.8.



(The "additive error" is $2S\Phi$.) All three functions can be used to measure the amount of uncertainty. The forms of both gnostic functions substantially differ from the quadratic function: they both are bounded for gross errors. This feature is again one of the vital characteristics of gnostics, because it leads to robustness with respect to outliers.

10.5.2 The Birth of Q-information and Improbability

Equivalence (10.41) is valid for both $c = j$ and $c = i$. In the latter case, (10.45) holds. For the case of $c = j$, a formal analogy to (10.50) is introduced:

$$(p_i := (1 - i\,h_j)/2)(h_j \in R^1)(H_i(p_i) := -p_i * \ln(p_i) - (1 - p_i) * \ln(1 - p_i)). \tag{10.56}$$

One can then verify the relations

$$-\frac{1}{p_i * (1 - p_i)} = \frac{d^2 I_i}{dp_i^2}. \tag{10.57}$$

$$I_i := H_i(1/2) - H_i(p_i). \tag{10.58}$$

Functions p_i and I_i also deserve names which characterize their substance.

Definition 11:
Let $Z = Z_0 \exp(S\Phi)$ be the model (5.12) of a given observed datum. Let p_i be the function of the E-irrelevance defined for this datum by (10.42) and (9.4).
Then p_i will be called the *gnostic improbability* of the given individual datum, shortly *improbability*.
The function I_i defined by (10.58) will be called the *quantification change of information* (shortly *Q-information*) of the individual datum being considered.

Analogously, the adjective "quantification" expresses the origin of the information I_i as the result of (double) integration of the sources of entropy E_i along the Q-path. And in the same manner as above, one arrives at the explicit dependence of (10.58) on uncertainty:

$$I_i = -2S\Phi * \sinh(2S\Phi) + \ln(\cosh(2S\Phi)). \tag{10.59}$$

Fig. 10.9 shows, that both this function and the Q-entropy E_j increase faster than the quadratic approximation.

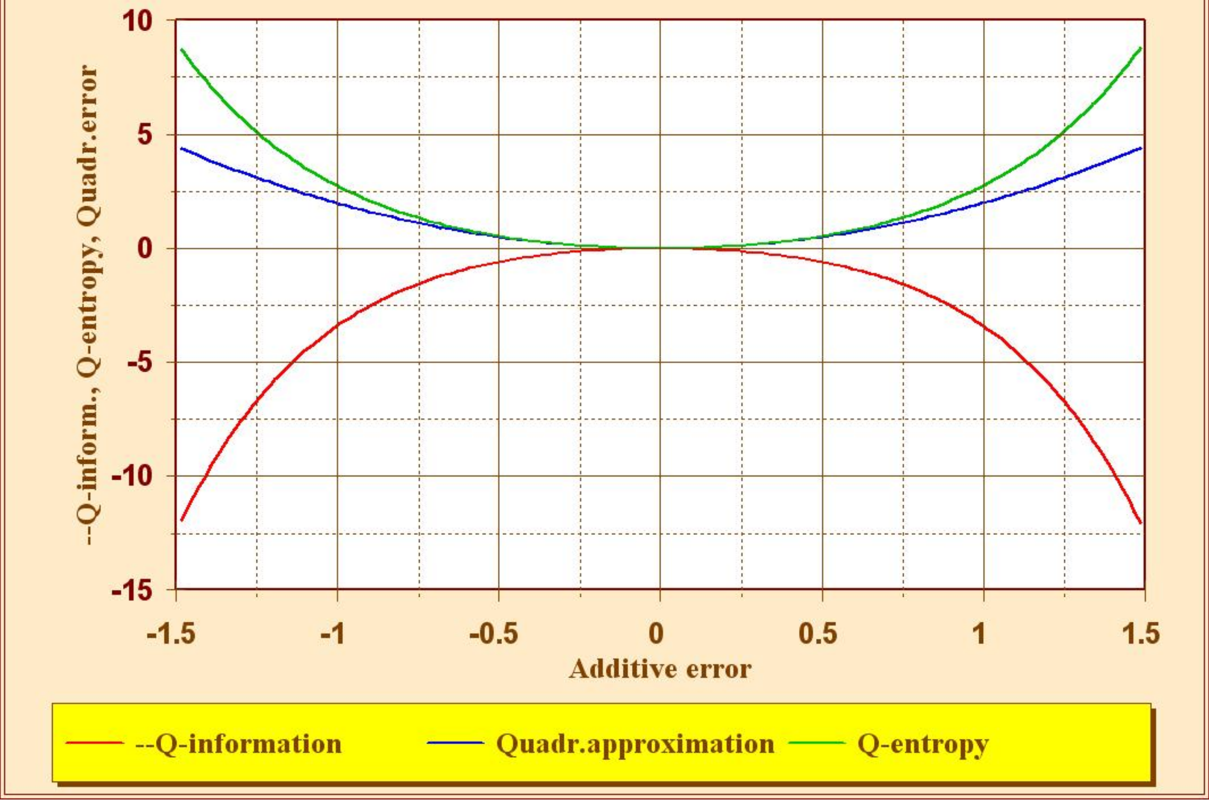


Fig.10.9 Q-INFORMATION AND Q-ENTROPY

10.5.3 Probability and Improbability

Features in common:

Domain: Just as the irrelevances, h_j and h_i are both defined over the open infinite interval $(-\infty, +\infty)$, so are the values of uncertainty $S\Phi$ as functions $p(S\Phi)$ and $p_i(S\Phi)$.

One common value: They have the same value (1/2) for zero uncertainty.

Monotonicity: They both rise monotonically as uncertainty Φ ranges from $-\infty$ to $+\infty$. (For p_i, this relates to the imaginary part of the improbability).

There also are substantial differences between probability p and improbability p_i (10.42):

Range of values: Probability p is a real function which has values in the closed interval $[0, 1]$. Improbability, on the other hand, is a complex function which takes on arbitrary values represented by points on the unbounded vertical straight line $x + i\,y = 1/2 - i\,h_j/2$ where h_j can have an arbitrary (finite) real value.

Slopes: For $|\Phi| \to \infty$ the values of the slope of the function p converge to zero, while for p_i, the values of the slopes of the p_i diverge under the same conditions.

The problem of the range of the probability forced a departure from the language of pair numbers, when proceeding from the pair equivalence (10.41) to a separate analysis of the quantification and estimation cases. This was not absolutely necessary, because the expression $(1 - j^2 h_i^2)/4$ can be decomposed not only as $p * (1 - p)$ (as was done with $p := (1 - h_i)/2$) but also as $p_j * (1 - p_j)$, where

$$p_j := (1 - j\,h_i)/2. \tag{10.60}$$

When the latter decomposition is chosen, one can continue writing both versions jointly using the indeterminate c. However, there are good reasons to prefer the scalar version:

1. The function p (probability) is used as an estimate of an element of probability; that is, as a measure of the expectation of the occurrence of some events. It is customary to express such expectations with real numbers and not by using a double number.
2. To derive the relationship between entropy sources and information using the variable p_j, one would need to use differential and integral calculus of double variables, which might give rise to additional questions. By using p, it was possible to apply the "ordinary" calculus instead.

Nevertheless, the possibility of introducing probability, expressed as a double number is useful to maintain the theoretical uniformity and duality of quantification and estimation analysis.

There are also common features as well as substantial differences between information I_j (10.52) and information I_i (10.58):

Domain: Both can be represented as functions of the uncertainty Φ.

Zero value: Both are zero for zero uncertainty.

Convergence: If uncertainty is sufficiently weak, both converge to a quadratic function of the uncertainty Φ.

Divergence: Unlike E-information converging to a constant for $|\Phi| \to \infty$, Q-information diverges under this condition.

In the case of a strong uncertainty, the different behavior of the four functions, which were considered, leads to greater robustness/sensitivity to outliers/inliers. These issues will be dealt with in the theory of data samples.

10.5.4 Conversion of Entropy to Information

Integrals of real movements (energies) play an important role in mathematical physics. Together with their first derivatives (moments), they are objects of the most powerful Laws of Nature—Conservation Laws. Applying variation principles to integrals of movement, one can derive differential equations which model the movements. Typical examples of integrals of movement include kinetic and potential energy (with respect to the Newtonian differential formulation of the laws of classical mechanics derived from movement integrals by differentiation), and the energy of electric and magnetic fields (with relation to Maxwell's partial differential equations). Typical features of different integrals of movement are, that through the process of movement, they are mutually converted from one to the other. A practical example from mechanics is a swinging pendulum periodically converting its potential energy into kinetic and vice versa. An oscillator consisting of a condenser and an inductance converts electric energy charged in the condenser into the magnetic energy of the coil and vice versa.

This is why we are interested in integrals of gnostic movement—of changes of Φ caused by uncertainty.

Relation (**??**) was derived by twice differentiating the entropy E_j. It follows, that one can obtain this entropy by twice integrating its sources. It can be easily shown that (10.41) is equivalent to the relation

$$r_{\hat{c}}^2 \nabla^2 (E_c) = -4 f_c^2, \tag{10.61}$$

which says, that the (normalized) sources of the entropy field are completely determined by the data uncertainty $S\Phi$, because $f_j = \cosh(S\Phi)$ and $f_i = \cos(S\phi)$, where $S\phi$ is also dependent only on $S\Phi$. This uncertainty arises through the virtual quantification movement. One can therefore take the entropies E_j and E_i as a pair of integrals of gnostic virtual movements.

E-information I_j (10.52) was derived as a field, the sources of which (obtained by twice differentiating) balance normalized entropy sources. The E-information is determined by p (10.52), which measures the gnostic virtual movement. This means, that information I_j is also an integral of this movement. Q-information, obtained by analogous operations, is also an integral of this movement.

In summary, it can be stated, that there are two pairs of integrals of gnostic virtual movement: $\langle E_j, I_j \rangle$ and $\langle E_i, I_i \rangle$, that are bound each to the other by equation (10.41), which can be now reformulated in the following way:

$$\boxed{\begin{gathered} Normalized\ source\ of\ entropy\ E_j\ + source\ of\ information\ I_j = 0. \\ Normalized\ source\ of\ entropy\ E_i\ + source\ of\ information\ I_i = 0. \end{gathered}}$$

Both these equations can be unified using the probability interpreted as the double number 10.60. The equivalence 10.41 thus obtained forms

$$\boxed{r_{\hat{c}}^2 \nabla^2 (E_c) + \frac{d^2 I_c}{dp_c^2} = 0. \qquad (10.62)}$$

This equation may be interpreted as the formula for the **mutual conversion of entropy to information and vice versa**.

A number of scientists have investigated the relationship between entropy and information. Even Maxwell had some thoughts on the subject in the late nineteenth century. A review of the development of these efforts published in 1956 ([7]) included contributions by L. Szilard (1929, [108]), N. Wiener (1949, [117]), R. C. Raymond (1950, [90] and 1951, [91]), L. Brillouin (1951, [12]) and D. A. Bell (1952, [6]). Such typical ideas were illustrated in [7] by two *Gedanken-experiments*:

1. The first example is based on Maxwell's idea of a demon controlling a door (with no mass) between two identical boxes filled with gas. He would permit fast molecules to pass in one direction and only slow ones in the other (to keep the number of molecules on each side constant). Temperature would rise on the 'fast' side without the addition of energy thus violating the second law of thermodynamics. Wiener [117] pointed out, that the demon would need information to distinguish between fast and slow molecules; he would consequently be converting information to entropy.
2. If a large number of trained monkeys were to sit at typewriters for a sufficiently long time, their output could even include Shakespeare's Hamlet. Such a result would correspond to a much lower entropy than the more probable chaotic expected outcome. On the other hand, were the letters corresponding to the text scrambled, all the information

would be lost and entropy would increase. The conclusion is, that the more information, the lower the entropy and vice versa.

These examples illustrate the traditional paradigm of entropy ↔ information conversion: (1) collective random events and their probabilistic model as primary notions, (2) entropy/information as secondary notions introduced to measure the degree of randomness and (3) entropy ↔ information conversion as an exchange of these (integral) measures. History of the ideas related to the Maxwell's demon were summarized in [13].

In contrast to these concepts, equation (10.62) establishes the entropy ↔ information conversion not on the integral level but on the more basic level of sources of fields (second derivatives). Furthermore, it is applied not for collective (data sets) but for an individual (datum's) uncertainty.

What follows is an interpretation of the "mechanism" of the "information machine" using the Ideal Gnostic Cycle based on an observed datum $Z_k = Z_0 \exp(S\Phi_k)$:

Quantification: Due to the contribution of uncertainty Φ, which increases from 0 to Φ_k, entropy E_j rises from 0 to $\cosh(S\Phi_k) - 1$. Simultaneously with Φ, the E-angle ϕ increases to ϕ_k ((10.9) causing the Q-information to fall from its initial value (0) to $I_i(S\Phi_k)$ (10.59). These changes set the stage for **future, potential** changes of information I_j (by (**??**)) and entropy E_i (by (**??**)). The Q-modulus of the radius vector stays constant (Z_0), but the E-modulus increases from Z_0 to Z_k (10.8).

Estimation: The E-modulus is constant (Z_k), the Q-modulus increases from Z_0 to Z_k. The change in the E-angle ϕ from ϕ_k to zero realizes the potential change in information I_j and entropy E_i.

Contraction: Both Q- and E-angles are zero, the modulus of the radius vector of the representative point decreases from Z_k to Z_0.

Because neither the overall (residual) changes of entropy within the closed IGC, nor the overall changes of information are zero, the IGC is **irreversible**.

10.5.5 Residuals of Entropy and Information

The overall change of entropy, which results from passing through the full Ideal Gnostic Cycle (*the residual of entropy* denoted $\varrho_{E,IGC}$), can also be calculated. Because both quantification and estimation changes of entropy are already known (10.26), relation

$$\varrho_{E,IGC} = \cosh(2S\Phi) + 1/\cosh(2S\Phi) - 2 > 0 \qquad (10.63)$$

holds for all $\Phi \neq 0$. The conclusion drawn from this formula is significant: it is impossible to return a datum which contains a non-zero uncertainty back to its original state (with uncertainty removed): its residual entropy will always be larger, than that in the "clean" state. This fact is proved for estimation by (10.63) using the Ideal Gnostic Cycle. However, it will be shown in what follows, that all other closed estimation cycles would lead to even worse results. One can thus conclude, that a part of the damage caused by uncertainty can never be removed: the gnostic cycle is *irreversible.*

A look at the formulae (10.55) and (10.59) shows that they both include the same term, $\ln(\cosh(2S\Phi))$, but with opposite signs; these terms therefore balance each other—they represent the reversible part of the information. However, since there are also irreversible parts to both information changes, the residual of information change resulting from passing through the closed IGC is

$$\varrho_{I,IGC} = 2S\Phi * \tanh(2S\Phi) - \arctan(\sinh(2S\Phi)) * \sinh(2S\Phi) < 0 \quad (10.64)$$

for all $\Phi \neq 0$. Since it is impossible to recover all the information lost due to a datum's uncertainty, even by using the (best) estimation process, ie by following the ideal gnostic cycle, the closed cycle of information changes is also *irreversible.*

Residuals of entropy together with residuals of information along with the quadratic error are depicted in Fig. 10.10 as functions of the additive error $2S\Phi$.

It is obvious from the graph Fig. 10.8, that the quadratic function (which is frequently used in statistics as the criterion function) may be interpreted as a rough approximation of entropy and/or information, but only for weak data uncertainty. However, this does not hold for the residuals. Indeed, using Taylor's expansion and Landau's symbol, $O(*)$, one obtains from (10.26), (9.3), (10.55), (10.59), (10.63) and (10.64) the following approximations valid for sufficiently small errors $S|\Phi|$:

$$E_j = 2 * (S\Phi)^2 + O((S\Phi)^4) \qquad E_i = -2 * (S\Phi)^2 + O((S\Phi)^4), \qquad (10.65)$$

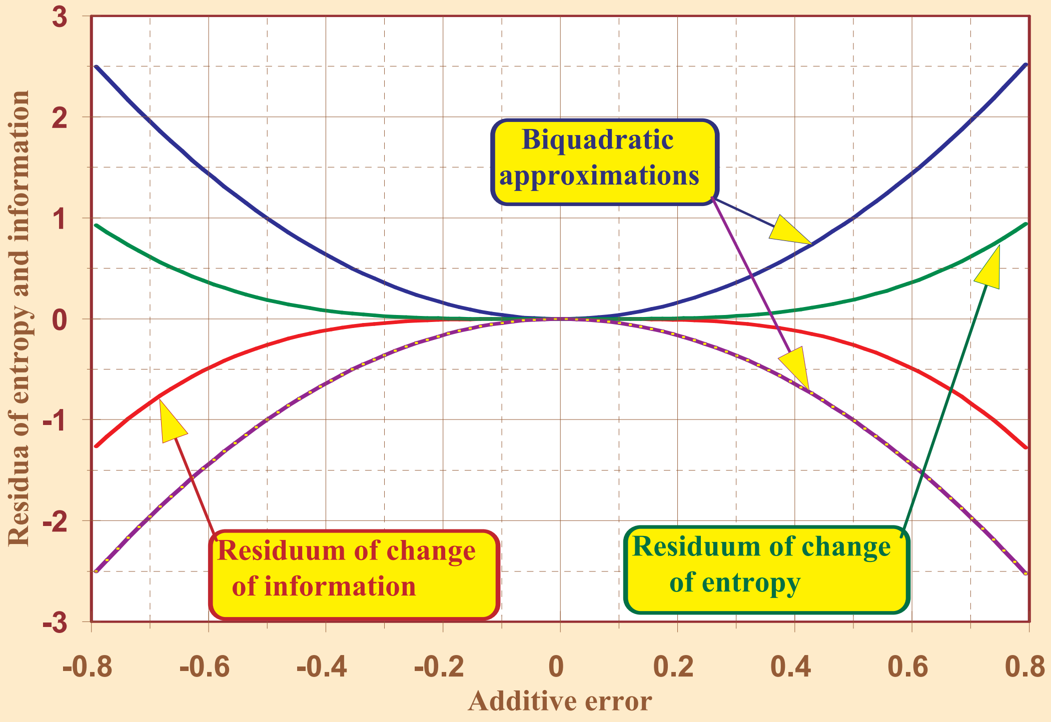
$$I_j = 2 * (S\Phi)^2 + O((S\Phi)^4) \qquad I_i = -2 * (S\Phi)^2 + O((S\Phi)^4), \qquad (10.66)$$

$$\varrho_{E,IGC} = 4 * (S\Phi)^4 + O((S\Phi)^6) \qquad \varrho_{I,IGC} = -16/3 * (S\Phi)^4 + O((S\Phi)^6). \qquad (10.67)$$

All these approximations are also shown in Figs. 10.8–10.10.

Fig.10.10 RESIDUA OF THE IGC

(IGC ... Ideal Gnostic Cycle)



The non-zero values of entropy and information residuals (10.67) (interpreted as the irreversibility of the Ideal Gnostic Cycle) prove a statement which has a fundamental importance comparable with that of the Second Law of Thermodynamics:

> It is impossible to create a machine to treat uncertain data so, that the output yields a greater amount of information, than that, which was contained in input data.

In other words: it is impossible to create an informational perpetual motion machine.

10.6 Summary

Gnostic virtual movement is a mathematical model of uncertainty's effect on data. In this movement, the role of time is taken over by the additive measure of uncertainty, the quantification angle Φ. This allows both kinetics (paths and velocities) and dynamics (behavior of integrals) of the virtual movement to be considered. A triple of special paths depicts the virtual movement, which is called the Ideal Gnostic Cycle.

Using the method of the Gedanken-experiment, each datum is endowed with a portion of energy, some of which is converted into heat flow and temperature. An analysis of the behavior of these energy-like characteristics of the datum over its virtual movement along the Ideal Gnostic Cycle permits the Q- and E-entropy changes caused by uncertainty to be evaluated. Analysis of the gradients of the entropies' fields confirms the favorable features of the circular E-path. Performing the estimation by following this path leads to an integral estimation error which is smaller, than that of the trivial “antiquantification” path.

An analysis of the sources of the entropies' fields reveals, that there are two scalar fields, the sources of which balance the entropies' sources. The formal appearance of these fields is reminiscent of probabilistic measures of information, although only an individual datum is being considered here. The use of these functions for Q- and E-information changes motivates the acceptance of their parameters for the determination of a measure of probability/improbability for an individual datum.

Q- and E-entropies together with Q- and E-information changes manifest interesting features, which are remindful of the integrals of movement of physics: equations of entropy ↔ information conversion for the second derivatives of these integrals describe the virtual movement of uncertainty. Using the integrals, one can prove, that the Ideal Gnostic Cycle is irreversible, ie that the damage caused to a datum by uncertainty cannot be completely eliminated, even by using estimation procedures based on this cycle. The interpretation of this irreversibility is, that the output can never provide more information, than was contained in the initial message.

Chapter 11

More on the New Notions

The Gedanken-experiment performed in the foregoing chapter used the logical link *datum* → *energy and/or temperature* → *entropy*. Assume at this point, that this reasoning provides sufficient justification for the idea of entropy E_c and its formula, 10.26. In contrast, the representations of gnostic probability, improbability, E-information and Q-information were derived from the entropy through mathematical manipulation. The process used to develop these latter ideas may have appeared complex; the objective of this chapter is therefore to show, that these newly introduced concepts are meaningful. In order to solidify the gnostic approach, it will be helpful to review several ideas from statistics and information theory.

11.1 Parzen's Estimators and Gnostic Kernels

In [81], Emmanuel Parzen, at Stanford University, expanded the ideas and results of [11], [95] and [115] by developing a series of theorems, which lead to a solution of the non-parametric estimation of a probability density function and mode:

> *Given a sequence of independent identically distributed random variables $X_1, X_2, \ldots, X_N$ with common probability density function $g(x)$, how can $g(x)$ be estimated?*

To solve the problem, a Borelian function $K(a) : R_1 \to R_1$, which satisfies the following conditions is introduced:

$$\sup_{-\infty<a<\infty} |K(a)| < \infty, \tag{11.1}$$

$$\int_{-\infty}^{\infty} |K(a)| da < \infty, \tag{11.2}$$

$$\lim_{a\to\infty} |aK(a)| = 0, \tag{11.3}$$

$$\int_{-\infty}^{\infty} K(a)da = 1, \tag{11.4}$$

$$(\forall a \in R^1)(K(a) = K(-a)). \tag{11.5}$$

Let $S > 0$. Then the required estimate may be constructed as

$$g_N(x) = \frac{1}{NS} \sum_{m=1}^{N} K(\frac{x - X_m}{S}). \tag{11.6}$$

The weighting function $K(a)$ has come to be called the *kernel* and the estimate (11.6) is the *kernel estimate* of the density function. The conditions required for unbiasedness, consistency and asymptotic normality of this estimate and for the density's mode (location of its maximum) can be found in [81]. Seven specific examples of kernels are given there, but it is obvious, that there are an infinite number of functions, which satisfy conditions 11.1–11.5. However, all possible kernels will not generate nicely smooth estimates of a density from a small data sample, although these estimates will still be acceptable from the asymptotic point of view.

The main idea behind the kernel estimation of distribution functions is simple: Suppose, that the task is to measure X_m. Any measurement taken will be imprecise. But if instead of accepting X_m as the real value, an (a priori assumed) "local" distribution $K((x - X_m)/S)$ of each possible measurement is used, the kernel $K(*)$ becomes an estimate of the probability density of the particular datum X_m. From 11.3 and 11.4, the integral of this kernel is the probability distribution function of the individual datum X_m, which can be written as $Pr\{X_m \leq x\}$. This distribution is conditional on the fact, that what was observed "really was X_m." The normalized additive aggregation of the density distributions of individual data (11.6) is then an estimate of the probability density of the data sample $X_1, X_2, \dots, X_N$. Its integral can be used as the kernel estimate of the common distribution function of the data sample.

Returning to gnostic theory, the substitution of 9.11 for $c^2 = -1$ into 10.42 and using 9.8 with 9.12, the gnostic probability can be recast in the form

$$p = \left(1 + \exp\left(\frac{4(A - A_0)}{S}\right)\right)^{-1}, \tag{11.7}$$

from which an important statement immediately results:

Theorem 11:
Let p be as defined in expression 11.7.
Then the *(E-kernel)*

$$\frac{dp}{dA_0} = \frac{4}{S}((\exp(2(A - A_0)/S) + \exp(-2(A - A_0)/S))^{-2} \qquad (11.8)$$

satisfies all the conditions of 11.1–11.5.

This easily verifiable theorem leads to an interesting interpretation: expression 11.8 is a kernel estimate of the probability density distribution of the unknown ideal value A_0 conditioned by the quantifying result A. Calling p in 11.7 the "gnostic probability" is justified, because it is a kernel estimate of the probability distribution of the unknown A_0, which has been quantified by observation of the individual datum A.

Using the definition of the G-weight, 9.10, one may rewrite 11.8 as

$$dp(A_0, A, S) = ((f_i)^2) * d(A - A_0)/S, \qquad (11.9)$$

ie as a metric formula of a certain Riemannian geometry applicable for measuring the distance between points A_0 and A, normalized by the scale parameter S (by integration). Formulae of this type have already been seen (the general one, 9.14, with a pair of its special cases, 9.15, valid for measuring errors in the Q- or E- irrelevance). The point is, that the choice of a kernel for the kernel estimation of a probability distribution/density can be understood as being the choice of a suitable Riemannian geometry. The statistical (say: Parzen's) approach is to require the observer to select and use one from an infinite set of possible kernels (geometries), which forces the choice of the geometry to be necessarily subjective. In contrast, the gnostic kernel 11.8 is unique for a given datum A and scale parameter S. The form of the kernel is obtained theoretically by strict mathematical reasoning from very elemental assumptions, ie in a way, which is more objective. Data are given objectively and the scale parameter (as will be shown later) will be determined from the data. It can be concluded, that the use of the gnostic kernel for the estimation of probability distributions is more objective, than that of the ordinary Parzen's approach.

Experience gained from application of this methodology has also shown, that the gnostic kernel generates surprisingly smooth estimates of probability densities even in the case of small data samples. Theorem 11 taken together with Parzen's theory legitimizes the application of the gnostic kernel even from the statistical point of view. The use of gnostic kernels for probability estimation is therefore justified on both theoretical and practical grounds.

The role of p in expression 10.42 is further support for the interpretation of 10.50 as probability. If one recognizes equation 10.50 as—at least formally—coinciding with Boltzmann's statistical entropy of a binary probabilistic system (or as Shannon's measure of the information of a binary message appearing on the output of an information channel with the probability p), the quantity p is confirmed as playing the role of probability.

11.1.1 A Gnostic Version of Parzen's Kernel?

As just seen, expression 11.8 can be interpreted as one of the particular versions of Parzen's kernels, suitable for kernel estimation of probability distribution functions and their densities. There are several features of this expression common to all kernels of Parzen's type: they all satisfy a set of conditions (11.1 through 11.5). These conditions are necessary from the statistical point of view—their fulfillment warrants the desirable asymptotic behavior of kernel estimates as the quantity of the sample's data increases without limit. There is no significant advantage to using a gnostic approach to treat all data samples regardless of the sample's size and its uncertainty. The power of gnostics is demonstrated by the application of the theory to cases, where data are scarce or their number and quality are limited due to the very nature of the data's source. The satisfaction of conditions resulting in "good statistical asymptotic behavior" is therefore not of vital importance for gnostic kernels. To show, that gnostic kernels may be viewed as estimators of Parzen's type at least in some special cases, is important mainly from a fundamental point of view. It has been already demonstrated, that the new theory has an important interface with the well-known results of statistics: the gnostic characteristics of uncertainty converge to the statistical ones, when there are weak uncertainties in the data. The estimation of distribution functions directly from data without appealing to an a priori data model is very important in practice. It is therefore a comfortable feeling, that the gnostic instruments applied to these tasks are supported not only by the gnostic, but also by statistical theory under circumstances, when statistical methodology can be properly applied.

There are some differences in the application of gnostic kernels 11.8 from the statistical ones, which include:

1. Aggregation of kernels, which result in the two kinds of robustness of

the estimated distribution functions (Chapter 15),

2. applicability to bounded data supports (see 11.1.2 and 15.2.2),
3. estimation of bounds of data support (15.2.2),
4. new methods of scale parameter estimation (16.2),
5. modifications, which allow censored data to be used (Chapter 19).

On Forms of E-Kernels

There are an infinite number of statistically valid kernels satisfying Parzen's conditions, but the ultimate choice is left to the users' subjective judgment; however the quality of the resulting estimate—especially of the probability density—critically depends on the form of kernel chosen.

In contrast, the gnostic kernel is not chosen in an arbitrary manner, but it is a product of the theory. Indeed, the kernel 11.8 (called *E-kernel*) was obtained by differentiating the distribution function 11.7 of an individual datum. This elemental distribution appeared as a parameter 10.42 of the field's source 10.43 of information. In its primary form (defined over infinite data support), this distribution is uniquely determined for each of the pairs of its parameters A (the observed datum) and S (the scale parameter). This distribution may be interpreted as conditional: $P\{A_0 \leq A|A,\ S\}$[1]. The datum is given and the scale parameter may be estimated from data using several methods, which will be discussed in the following sections. There is therefore no subjective element in the preparation of the gnostic kernel for the estimation of a sample's distribution or density.

The role of the observed value A as a kernel's parameter is obvious: it locates the kernel on the A-axis; it is thus the *location parameter* of the kernel. The scale parameter S determines the kernel's "width:" the greater S, the stronger the uncertainty, the more difficult to recognize the true (ideal) value A_0 hidden in the observed datum. A statistician would say, that the scale parameter is determined by the variance. Using the language of fuzzy set theory, one could say, that the scale parameter is a measure of the datum's fuzziness.

All Parzen's kernels also have parameters, which play the role of location and scale parameters. However, they also determine the analytical form of the kernel. The special form of the gnostic kernel 11.8 is unique because its origin is connected to the Ideal Gnostic Cycle. We have seen, that by

[1] This formulation reads: "the probability distribution of the ideal value A_0 given the observed value A and the scale parameter S".

following this cycle, one maximizes/minimizes the effects of uncertainty. The gnostic kernels thus promise informational optimality.

Measuring Scale Invariance and Equivariance

One of important features of (scientific) information is, that it is independent of the physical nature and other features of the carrier. Indeed, a message written (in a known language) on a piece of paper bears the same information as a message transmitted by a (noise-free and non-distorting) radio channel. The information content of speech is invariant with respect to strictly linear amplification. It is obvious, that information resulting from measurement does not depend on the choice of using centimeters or inches as a measurement unit. It is for reasons such as these, that notions such as the probability of events and the corresponding information must be independent of the scale of measurement, which is applied.

Examining the scale invariance of the probability distribution kernel 11.7, which equals the integral of the gnostic density kernel (11.8), it is seen, that with an increase in the data scale unit of eg k-times, all values A, A_0 and S change in proportion to $1/k$ and the value of the probability (11.7) remains unchanged. This invariance can also be observed directly in the definition of the irrelevance (9.6), which leads to the definition of probability (10.42). In the same manner, the other most important gnostic characteristic—the data weight (9.5)—is also seen to be scale invariant.

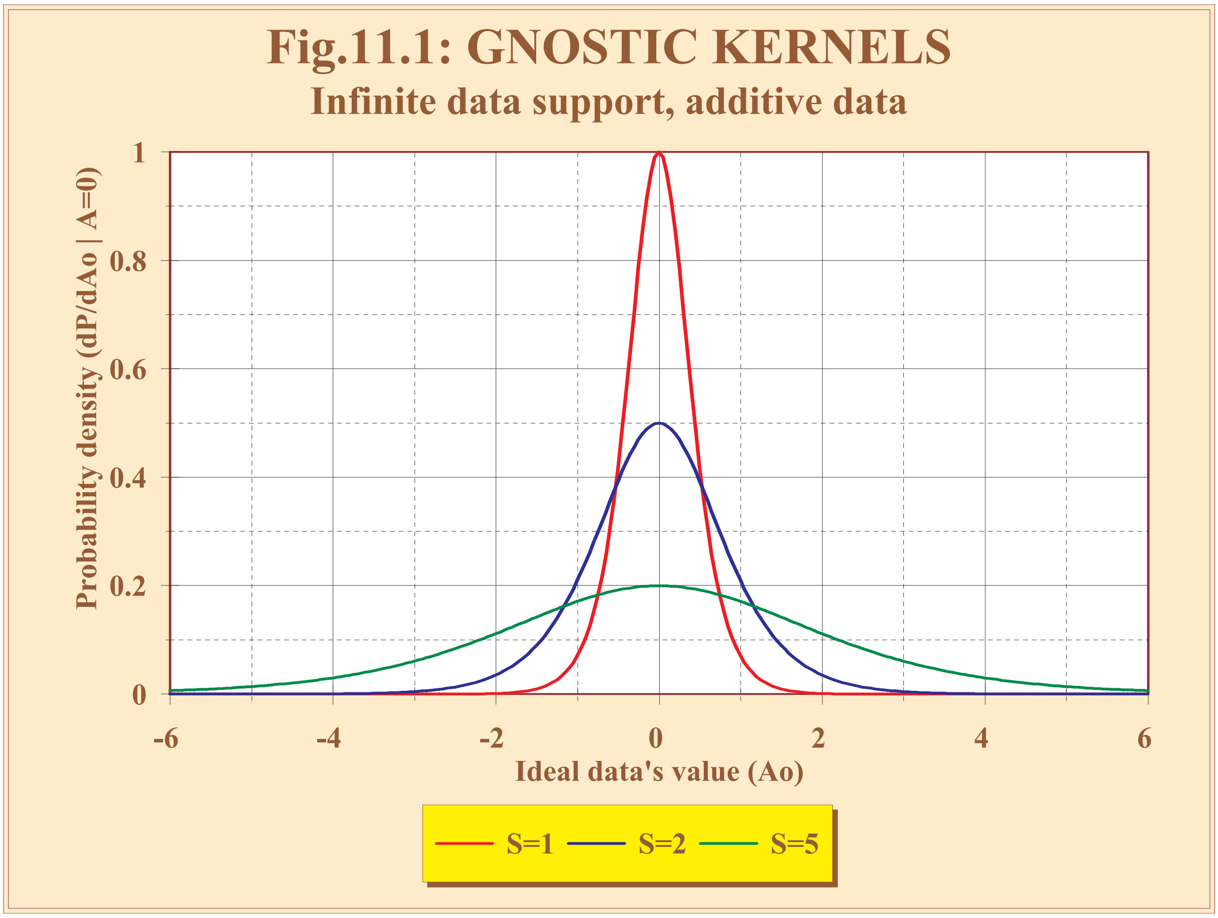
The notions of invariance and equivariance originated in statistical estimation theory. A typical example is the behavior of an estimate of the location and scale parameters in the case of a change of origin of the data scale ("the data shift"). Under such a change, the scale estimate found by the least squares method would be invariant, while the location parameter estimate would be equivariant ("changing in the same way as the origin"). The same scale parameter estimate would be equivariant with respect to the measuring scale. For the gnostic kernel of probability density 11.8, it is seen, that its relative form is "data location invariant" (because the difference $A - A_0$ is not dependent on the location of the origin of the data axis), but equivariant under changes of the measurement scale (because of the division of the invariant square of the data weight in 11.9 by the scale parameter S, which is equivariant with respect to the measurement scale).

Infinite Data Support

The set of possible data values can also be called *the data support.* This role was played by the set R^1 of real numbers (in the case of data forming the additive group) or by the set R_+ of the positive reals (in the case of data as elements of the multiplicative group) (see Chapter 1). These sets thus have the property of *infinite data support.* Both probability distribution (11.7) of the gnostic kernel and its density (11.8) are defined over the domain R^1, ie over infinite data support, which feature is not unusual for Parzen's kernels. So, eg, the Gaussian curve (the density of the normal distribution function) may also be used as one member of the family of Parzen's kernels. However, it should not escape the reader's attention, that the gnostic kernels vanish significantly faster than the Gaussian ones as the distance of the observed value A from the location parameter A_0 increases. If they have the same scale parameter (S), the Gaussian increases or decreases with increasing $((A - A_0)/S)^2$ proportionally to $1/\exp(((A - A_0)/S)^2)$, while the gnostic kernel (11.8) approaches zero as $4/\exp(((A - A_0)/S)^4)$. This higher rate provides a favorable flexibility for the gnostic kernels.

The dependence of the kernel's form on the scale parameter is demonstrated in Fig. 11.1 for the case of the infinite data support R^1. The data are thus considered to behave in accordance with the additive property. The observed value is zero ($A = 0$) for all three curves, while the scale parameters (S) take on different values (1, 2 and 5). The vertical axis is denoted "Probability Density." In using this term, we must remember, that the word "probability" in gnostics means something quite different than in its statistical definition. Here it is "our expectation based on observed data" and not a parameter from an a priori assumed statistical model.

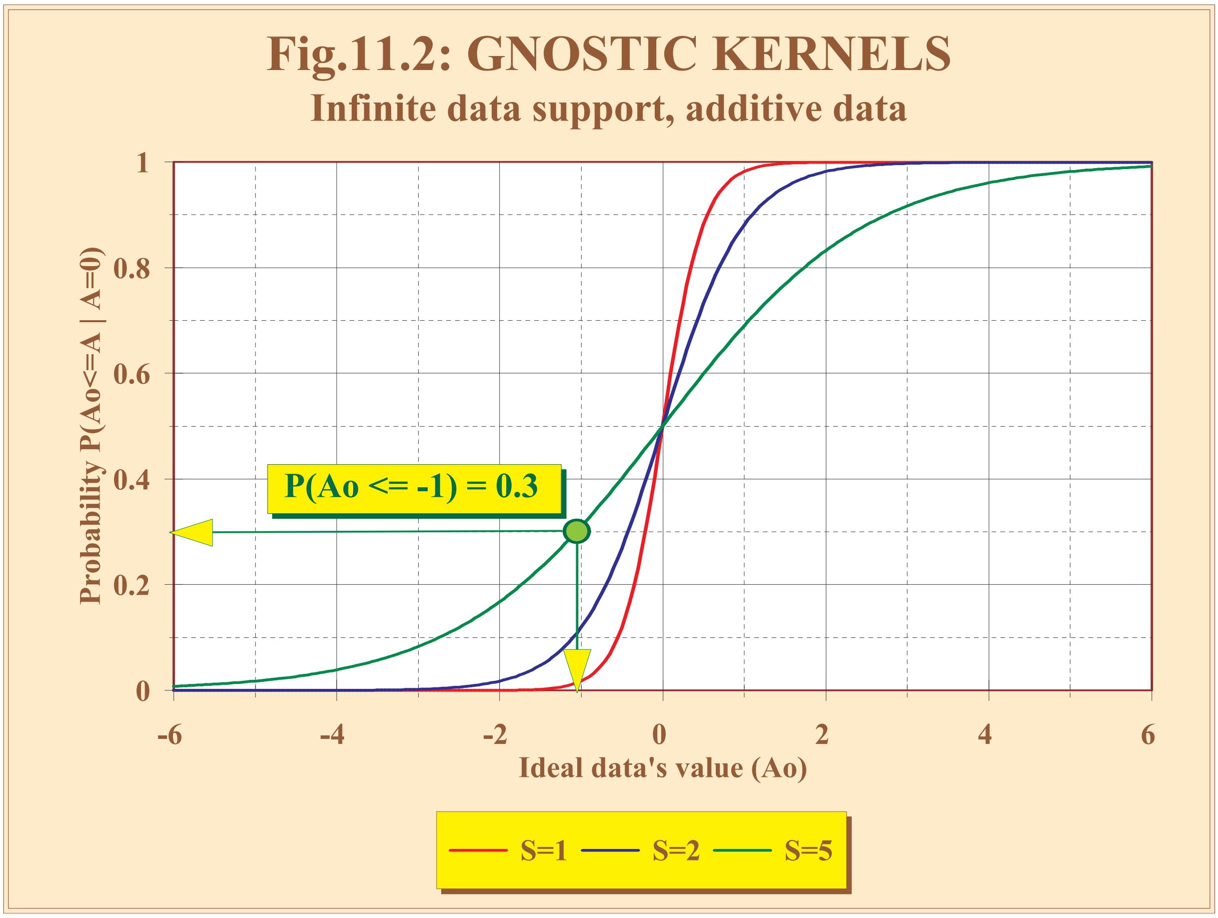
The curves in Fig. 11.1 depict the dependence of the probability density on the (unknown) ideal value (A_0) for given values of A and S. They answer the question: "The observed datum was zero: to what extent could one expect, that the (unknown) ideal value was close to the number A_0, if the (known or estimated) scale parameter equaled S?" In the case of a small scale parameter (eg $S = 1$), the expected values of A_0 are concentrated closely around the observed data's value (zero). Increasing S decreases the maximum density and flattens the density curve. (The areas under the curves are the same in all three cases because the overall density's integral equals 1). Using mathematical language, one denotes these probability densities as $\frac{dP(A_0,S)}{dA_0}$, where P is the probability $P\{A_0 \leq A|A,\ S\}$. Note, that (in this simple case) the location of the densities' maxima coincide



with the location of the observed value A. The probability distributions corresponding to these densities (their integrals) are shown in Fig. 11.2.

It is important to read these graphs properly. So in the case of $A = 0$, $S = 5$ (green lines) one reads: "the probability of A_0 **not exceeding** the value -1 is 0.3" or—equivalently—"the probability of **exceeding** the value -1 is 1 minus 0.3, ie 0.7." The distribution function thus attaches the probability value 0.3 (and its complementary value 0.7) to the ideal data value $A_0 = -1$. This relationship along with its inverse is shown in Fig. 11.2 by the green arrows.

Varying the observed value A, while maintaining the scale parameter unchanged, merely shifts the kernel along the horizontal axis. This is seen in Fig. 11.3 for the case of $S = 1$ and three observed values A (-3, 0 and 3). The kernel's relative form in this case is invariant. The location of the kernels is equivariant with changes in the origin of the horizontal axis. However, such a simple picture is only valid for infinite data support and additive data.



It was shown in Chapter 1, that there exist both additive and multiplicative data. The typical feature of multiplicative data results from the character of the structural operation, which creates from a pair of data a third datum—the multiplicative value. While the additive data tend to linear behavior, the "natural" behavior of multiplicative data is exponential. In other words, the multiplicative data, as a rule, cover a broad interval. One frequently sees samples of multiplicative data, within which data differ by many orders of magnitude. (Example: the total assets of companies in a given industry may span an interval ranging from several millions of US$ through many billions. Another example can be taken from the environmental control, where concentrations of pollutants in rivers can differ by orders of magnitude.) To work efficiently with such data, logarithms are used instead of the data themselves. (The focus of this discussion is not on accountancy, which uses the linear scale, but on data analysis.) A data distribution graph realized as a probability distribution function of logarithmic data also has its density. However, to maintain the com-

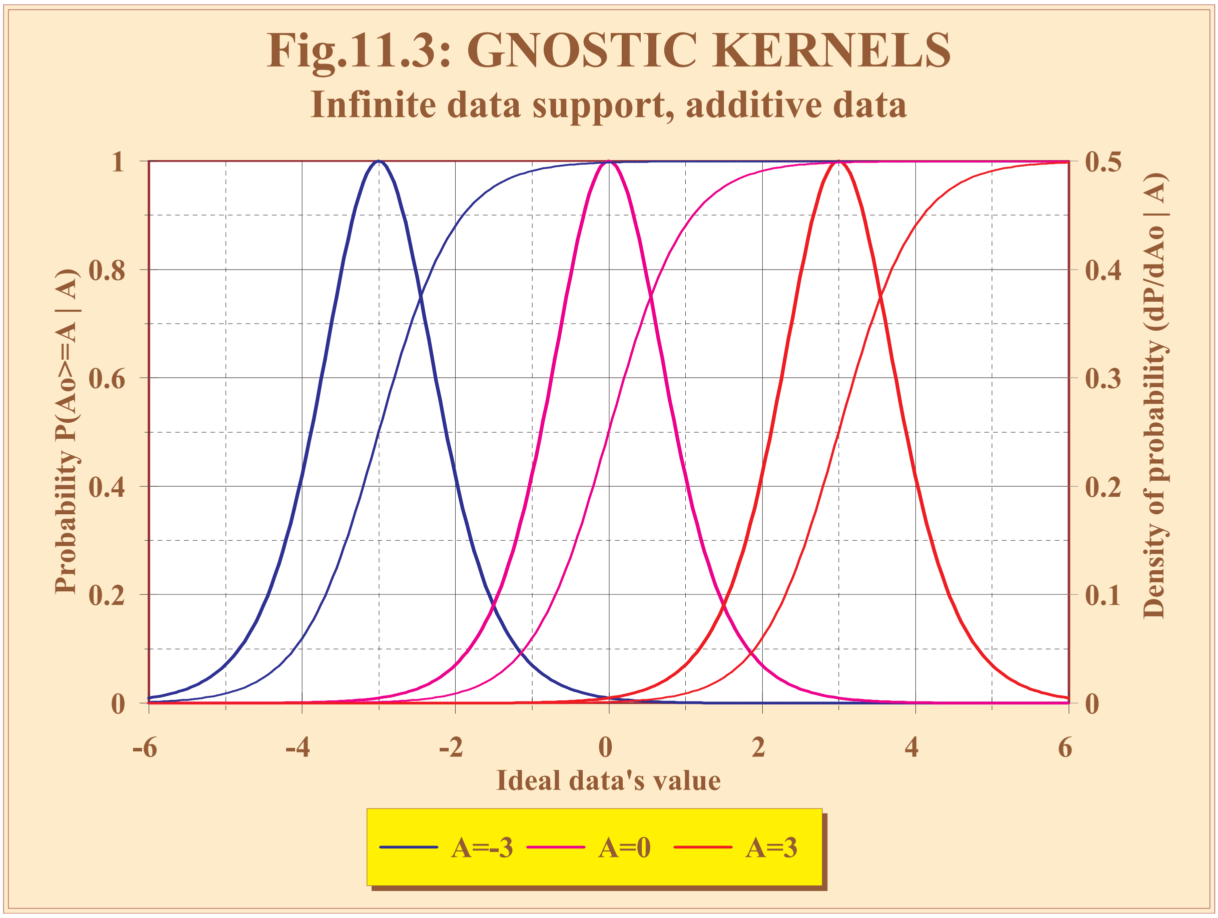


Fig.11.3: GNOSTIC KERNELS

Infinite data support, additive data

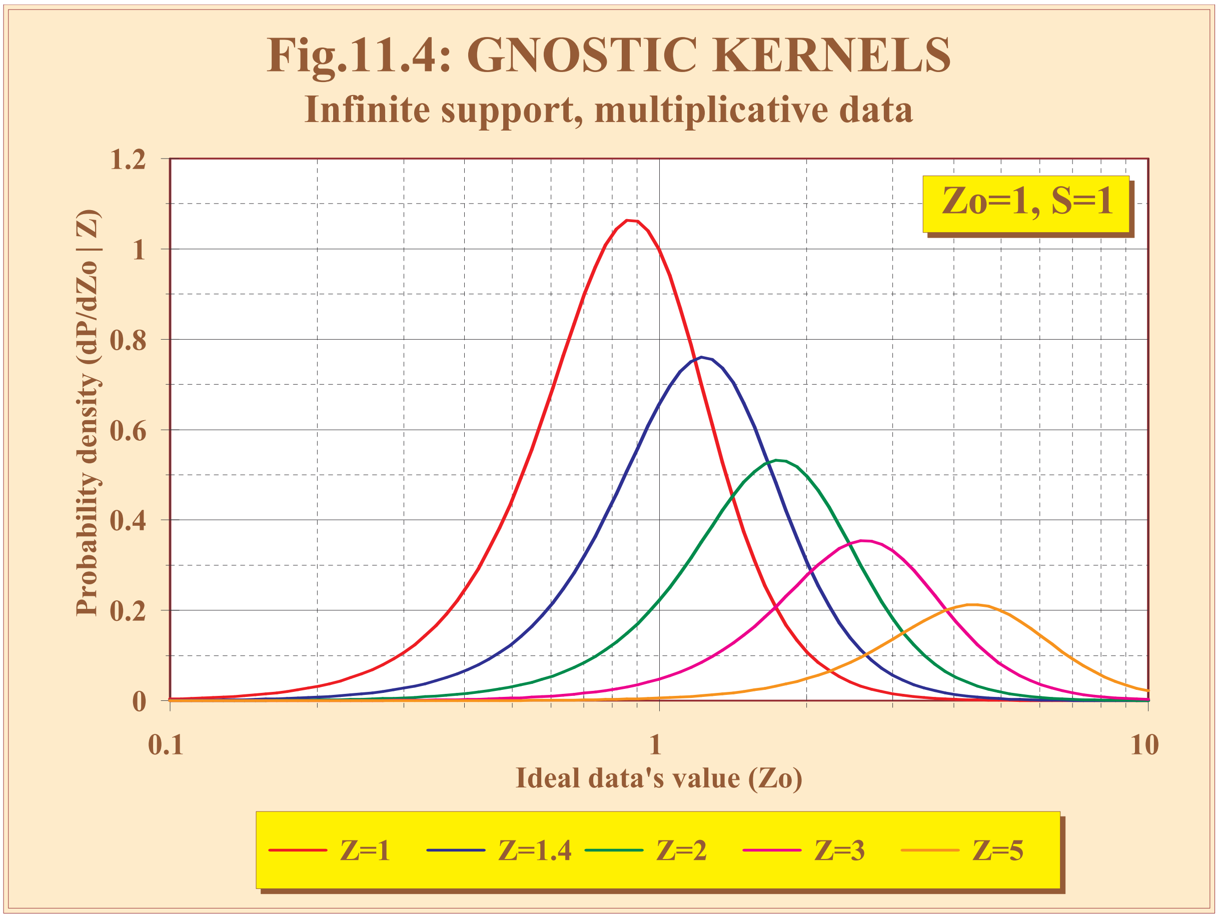
patibility of graphic representations of probability distributions and their densities, when a logarithmic datum is employed, the probability density is represented as

$$\frac{dP}{dZ_0} = \frac{dP}{dA_0}\,\frac{dA_0}{dZ_0}. \tag{11.10}$$

Substituting 11.9, 9.5, 5.12 and 5.15 into this relation one obtains

$$\frac{dP}{dZ_0} = \frac{1}{(S * Z_0)}\,\frac{4}{((Z/Z_0)^{2/S} + (Z_0/Z)^{2/S})^2}. \tag{11.11}$$

The division by Z_0 in this formula has the immediate consequence of shifting the density maximum with respect to the observed datum and changing the maximal density. This result is obvious as seen in Fig. 11.4, where the same scale parameter ($S = 1$) is used for all five curves, which depict different observed values of Z (1, 1.4, 2, 3 and 5) for an unknown and unobservable Z_0 of 1.0. Now in the case of multiplicative data, the densities' maxima and the corresponding values of the observed datum are no longer coincident.



It is worth noting, that in all the cases considered above, the density curves were such, that the probability distribution functions had the S-form. As has been noted, they differ from the most popular normal (Gauss') distribution function (which also are of the S-form) by their much faster decay. There are many other similar distributions in statistics, however, probability distributions, which have an “anti-S form” also exist both in nature and science. They are seen primarily in cases of finite data support and inner positive feed-back of the object under consideration. The typical density curve of an anti-S distribution is of the U-form. To estimate such distributions and densities, one cannot limit oneself to the kernels, which have a fixed form. To demonstrate the suitability of gnostic kernels to tasks of this type, it is useful to consider examples of their behavior over finite data support.

11.1.2 Finite Data Support

The gnostic model of quantification/estimation was developed for infinite data support, because this was the immediate consequence of the simple nature of the basic data structures, which were considered: the additive or multiplicative groups. However, real data more often exist as structures defined over finite data supports.

An example may be the structure of the depreciation of the assets of a group of enterprizes. Depreciation (expressed in percent) is a non-negative real number not exceeding 100. It is thus defined only over the range of $[0, 100]$. Another example is the set of market prices for a good. Intuitively, one feels, that a price should not exceed a certain value, but it also would not be expected, that the price would fall below some minimum. The maximum value is limited primarily by competitive forces, while the minimum is determined mainly by costs. To survive in such an industry, all comparable enterprizes must keep their economic parameters bounded within limits, which are typical for the given industry. Estimating these bounds from data may be one of the most interesting goals of the analysis. It is therefore useful to have analytical instruments, which can take into account these finite bounds for data supports.

A further example is concerned with the existence of a positive feed-back within the object under consideration, say, an industrial enterprize. Consider a prosperous company, which suffers through a very bad or even catastrophic period. The probability of failure for such an institution had been practically zero until the moment of the event but rises rapidly afterwards because of the "positive" feed-back: information about the threat of failure, which may cause the loss of credits, the fall of share prices, the loss of clients, all of this could accelerate the firm's downfall. (A feed-back amplifying the input changes of an object is "positive" only from the point of view of cybernetics.) If the initial impulse was sufficiently strong and if an "economic miracle" does not take place, the enterprize's failure is only a question of time. In other words, the probability of failure was zero until a certain moment, but it could increase to nearly 1 beyond some maximal survival time. The probability distribution may have the anti-S form in this and other similar cases.

In principle, the generalization of the theory to finite support is not difficult. Consider a strictly positive unbounded real number $Z_\infty \in R_+$. Let $0 < L < U < \infty$ and inequalities $L < Z_0 < U$ and $L < Z < U$ hold. Both the ideal datum Z_0 and its observed value Z are thus bounded. To prevent a misunderstanding, the bounded values will be denoted $Z_{0,fin}$ and Z_{fin}. Let a transformation $\mathcal{T} : (L, U) \leftrightarrow R_+$ be of the form

$$Z_\infty = \frac{Z_{fin} - L}{1 - Z_{fin}/U}. \tag{11.12}$$

This transformation is regular because its inverse

$$Z_{fin} = \frac{Z_\infty + L}{1 + Z_\infty/U} \tag{11.13}$$

exists. Moreover, let us consider a pair $Z_{1,fin}, Z_{2,fin} \in (L, U)$ of bounded data such that $Z_{1,fin} = \mathcal{T}^{-1}(Z_{1,\infty})$ and $Z_{2,fin} = \mathcal{T}^{-1}(Z_{2,\infty})$, where both $Z_{1,\infty}$ and $Z_{2,\infty}$ are unbounded multiplicative data. Let the structure operation (the "multiplication" $\otimes$) be defined as

$$Z_{1,fin} \otimes Z_{2,fin} := \frac{Z_{1,\infty}\, Z_{2,\infty} + L}{1 + Z_{1,\infty}\, Z_{2,\infty}/U}. \tag{11.14}$$

It can be easily verified, that the isomorphism between the additive group $\langle R^1,\ +\rangle$ and the structure $\langle \mathcal{S}_{Zf},\ \otimes\rangle$ exists, where $\mathcal{S}_{Zf}$ is the set of data defined over the finite data support. The latter structure is thus also the multiplicative group. This justifies the application of "transformed" results of the theory of individual unbounded data to transformed bounded data. Before proceeding to the consequences of the transformation, let us note several of its aspects.

1. The chosen transformation is not unique, others could be used. The advantage of making this choice is primarily its analytical simplicity.
2. The rate of convergence of Z_∞ to zero (for $Z_{fin} \to L$) and to infinity (for $Z_{fin} \to U$) is sufficient for applications.
3. Formula 11.12 has a natural motivation: $Z_{fin} - L$ is the Euclidean distance of the observed value from the lower bound of the data support, while $U - Z_{fin}$ measures the datum's distance from the upper bound. The ratio of these distances thus evaluates the "unbalance" caused by deviation of the Z_{fin} from the center of the finite data support. The ratio is computed by dividing 11.12 by U. This modification does not change anything in the case of a finite U, because the variable Z_∞ always appears in gnostic formulae divided by $Z_{0,\infty}$ (which is the ideal value transformed to infinite support in the same way). There are two advantages to this modification:
   (a) Formula 11.12 allows both limit cases $L \to 0$ and $U \to \infty$.

(b) This form of the formula preserves the symmetrical behavior of the density for both $Z \to L$ and $Z \to U$:

$$\frac{dA_\infty}{dZ_{fin}} = \frac{1}{S}\frac{d\log(Z_\infty)}{dZ_{fin}} = \frac{1}{S}\left(\frac{1}{Z_{fin} - L} + \frac{1}{U - Z_{fin}}\right). \tag{11.15}$$

4. Going from the simple case of infinite data support to finite ones, one is resigned to relying on Parzen's theory, which gnostics only uses to illustrate, that its approach is neither unnatural nor unexpected for a statistician, at least for a special case. However, the theoretical line of development of gnostics is independent of statistical ideas and is applicable to the more general problem of small data samples.

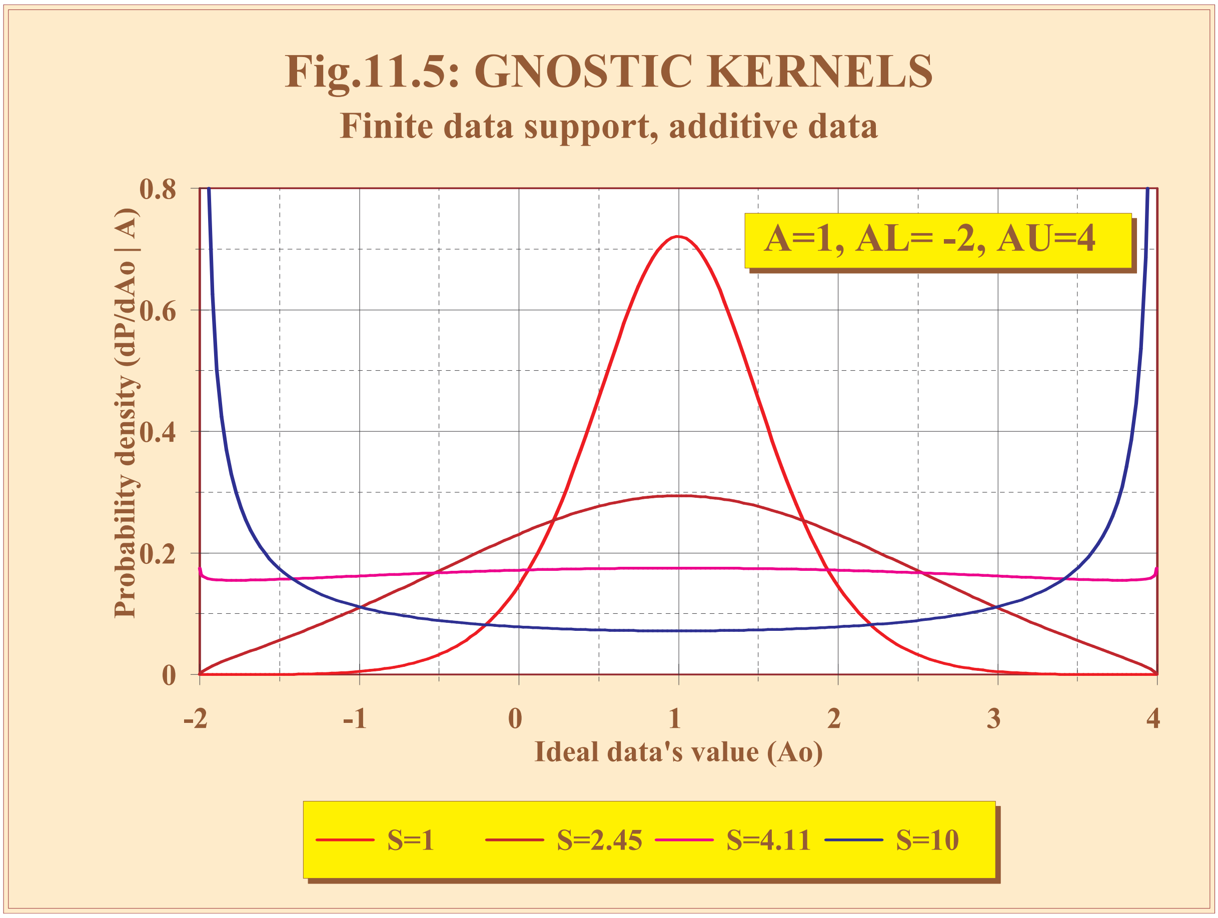
The importance of finite data support can be demonstrated by considering the gnostic kernels, that are generated by the transformation 11.12.

Four gnostic kernels shown in Fig. 11.5 are defined over the finite interval of additive data $(-2,\ 4)$ for the same observed datum $A = 1$ and for four values of the scale parameter (S =1, 2.45, 4.11 and 10).

The kernels are symmetric with respect to the observed datum, which is located at the center of the data support. The form of kernels is, at first sight, unexpected: only two kernels (S =1 and 2.45) have the concave forms, which are reminiscent of Parzen's kernels. The kernel for S =4.11 is practically flat over the whole data support, while the kernel for S =10 has the U-form. This variability of form is easy to see through reference to the formulae. The behavior of the kernels is determined by the product of two derivatives in the expression

$$\frac{dP}{dA_{0,fin}} = \frac{dP}{dA_{0,\text{inf}}}\,\frac{dA_{0,\text{inf}}}{dA_{0,fin}}, \tag{11.16}$$

where the first term of the product is 11.8 and the second 11.15. Both A and A_0 in the first derivative should now be interpreted as A_∞ and $A_{0,\infty}$ because they are transformed onto the infinite support by the formula 11.12) and $A_\infty = \log(Z_\infty)$. For $Z_{0,fin} \to U$ the transformed value $A_{0,\infty}$ approaches infinity and the derivative 11.16 approaches $\frac{4}{S\exp(4A_{0,\infty}/S)} * \frac{1}{U - Z_{fin}}$. This is an expression of the type "zero∗infinity", the value of which may be zero as well as infinity—depending on the scale parameter S. The case of $Z_{0,fin} \to L$ is symmetric as has already been mentioned. The rate of change of the density is thus determined not only by the difference $A_{fin} - A_{0,fin}$, but also by the scale parameter S, which can even change the character of the function from the concave form (small S) to the U-form (large S).

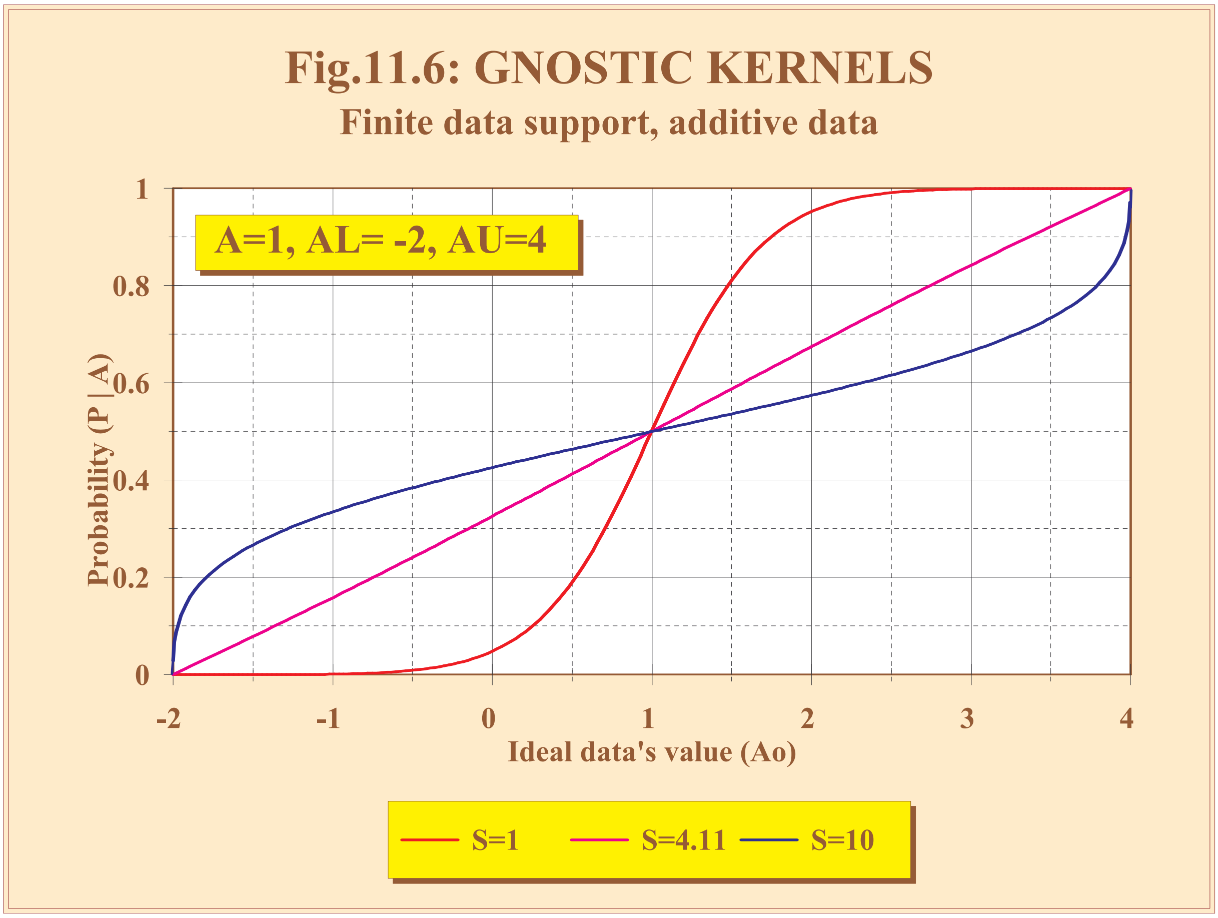


It is important to note, how the differences in the kernel's form are reflected by the forms of their integrals, the distributions of probability (Fig. 11.6): the S-form (eg with $S = 1$) can become a linear function ($S = 4.11$) or even an anti-S form with eg $S = 10$.

The previous examples were symmetrical, because the observed datum was at the central location, the point $A_{0,fin} = 1$. The influence of the observed datum on the kernel's form is illustrated in Fig. 11.7, where $A_{0,fin} = -1$.

The peak of the density is close to the observed datum only in the case of a small value of the scale parameter ($S = .5$ in the graph). Increasing S flattens the peak and its maximum approaches the lower bound. Large values of S change the form of the kernel entirely to a non-symmetric U-form. The effect of changing the observed datum's value, while keeping the scale parameter constant ($S = 3$), is shown in Fig. 11.8.

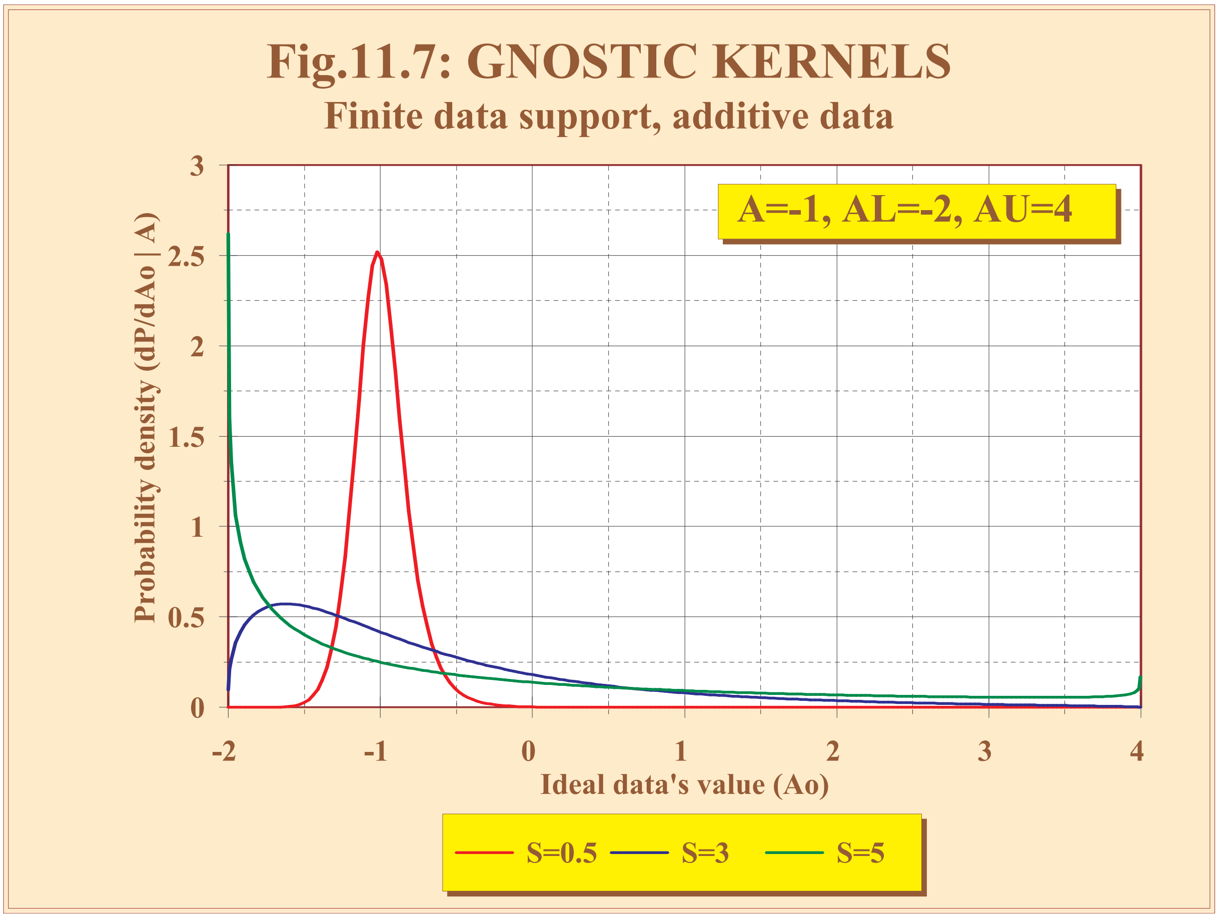
The gnostic kernels are parameterized in the general case by four pa-



rameters (Z, S, L and U) or by their transforms. The observed datum Z (or A) is given. It can be shown, that the parameters S, L and U are easily estimated from the data. This is the real sense of the gnostic motto "Let data speak for themselves". We have seen, that the choice of these parameters provides an extraordinarily rich palette of forms for the gnostic kernel. All this results in the expectation, that gnostic kernels will be useful to estimate distribution functions and densities of many forms, and that the process of finding the necessary parameters will be not only objective, but also suitable to a high degree of automation.

11.1.3 What About Q-kernels?

The foregoing analysis of gnostic kernels only developed the concept of the gnostic probability as defined in 10.42 as $p = (1 - h_i)/2$, where h_i is the estimating irrelevance 9.6. However, as is always the case in gnostics, there is a dual variable to the probability p, $p_i = (1 - i\ h_j)/2$ in 10.42.

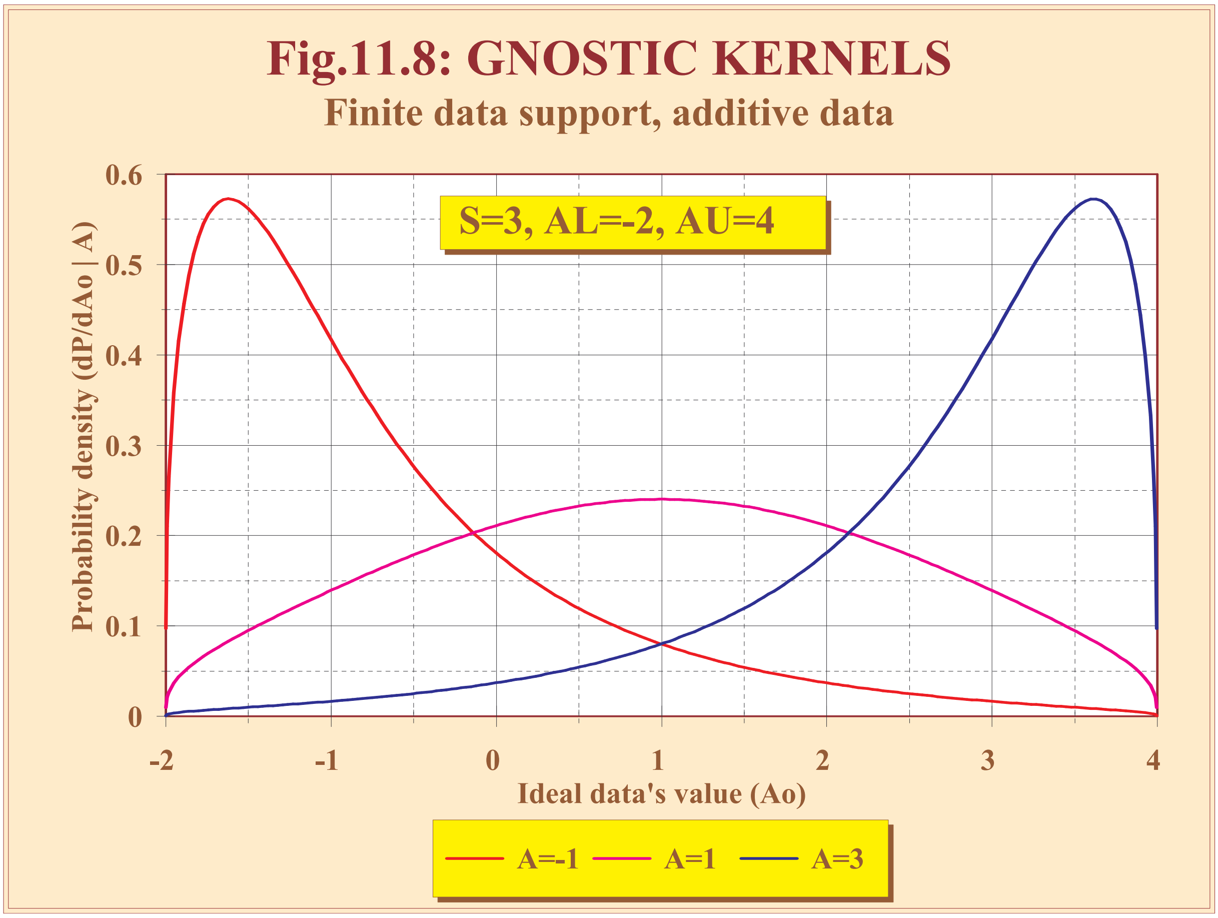


This duality becomes obvious, when the alternative interpretation of the probability, p_j 10.60 is used. Consequently, it could be asked, whether there exists a Q-kernel (a kernel based on the quantifying irrelevance h_j in a manner analogous to $\frac{dp_j}{dA}$). The problem is, that the improbability p_i is a complex number, the modulus of which is between 1 and infinity and we are accustomed to measure expectations with numbers lying in the interval $[0, 1]$. A transformation of the improbability to the interval $[0, 1]$ is straightforward, making use of formulae 9.17 and 9.18, from which the E-kernel 11.8 obtains the forms of

$$\frac{dp}{dA_0} = \cos^2(2\phi)/S \tag{11.17}$$

$$\frac{dp}{dA_0} = \cosh^{-2}(2\Phi)/S) \tag{11.18}$$

because of the equivalence $\cos(2\phi) \equiv 1/\cosh(2\Phi)$. This permits the consideration of the probability p 11.7 as the function $p(h_j)$ of the



quantifying irrelevance h_j and of its derivative as a Q-kernel 11.18. Such a transformation appears on the surface to bring nothing new, because this Q-kernel will be equal to the E-kernel considered above. However, this is true only for the *single* kernel produced by an individual datum. Significant differences will appear in the more general case of distributions obtained by the aggregation of several kernels. These differences are caused by the aggregation law, which gives different results, when it is applied to quantifying irrelevances rather than to estimating irrelevances. The final effect is, that the distribution functions obtained by E-kernels will manifest a different robustness than those produced by Q-kernels. This topic will be considered in more detail later on.

The results obtained, when considering Q-kernels, are thus of a sophisticated nature: although the kernels can be introduced in a manner, which identifies them with their E-kernel counterpart, (from the point of view of a single datum), they may also provide different (and useful) results, when they are applied to data samples. However, the implementation of gnostic kernels to data samples is not trivial:

1. The form of the kernels depends on the bounds of the data support as shown e.g. in Fig.11.8. To get rid of this, it is necessary to transform the finite support onto infinite one.
2. The weight of a data item applied within aggregating may depend on the location of this item.
3. The kernels' scale parameters may be location-dependent as well.

All this is to be reflected in estimating algorithms.

11.2 Albert Perez's Notion of Information

To show, that expression 10.50 can be used as a gnostic measure of the information brought by an individual datum, it is useful to recall, what the word "information" means. In general usage it is interpreted as "knowledge or facts" ([80]) or "knowledge acquired in any manner; data; facts; news; tidings" ([112]). It is the latter interpretations, that are preferred here, because they are more closely related to the scientific meaning of information. An example will serve to demonstrate the point. In the simple sequence, "You have a daughter," the knowledge, which is imparted, has a vastly different meaning (value) for a man anxiously waiting on a maternity ward, than for parents registering a girl for her first day at school. The message represents knowledge in both cases, but only in the former case does it impart **new** knowledge, which decreases the uncertainty of the person receiving the news.

When dealing with information imbedded in data, a narrower, more exact notion is needed. What is required is not merely a mathematical definition, but one, which also includes as many aspects of the idea to be defined as possible. The following points summarize this approach and are taken from [84]:[2]

(1) Information theory is a scientific discipline, the objective of which is to characterize the *abstract* notion of a *message*, without taking into consideration the various forms, that the message (the signal) may take, while attempting to remove as many *degrading* factors (noise) from the message as possible.

(2) The notion of a message assumes

[2] Any errors in translation are our responsibility.

   1. the existence of a pair of systems, eg one system representing the object to be observed and a second one representing the observer and/or his apparatus[3]
   2. an interaction between the two systems.

(3) This interaction manifests itself

   1. by mutual exchange of energy,
   2. by changes in the states of the systems.

(4) Changes in the state of these systems can be explained by changes in their energy, more specifically by the balance between changes in entropy and its "counterbalance", information.

(5) A basic characteristic of the notion of a message is the fidelity of the image, which is reflected or registered by the receiving system from the emitting system.

(6) The entropy of the whole system (of both the observed and observing sides)—in general—increases during transmission, however, the received information may be reused in partial compensation for this entropy increase.

(7) The random nature of the disturbance factors prevents the precise characterization of the interaction between the two systems as a fixed, unique transformation. The suitable tool is the probability of a real input value conditioned on the observed value.

(8) To measure any quantity of information, a function satisfying the following conditions is necessary:

   1. The function is non-negative.
   2. When the value of the function increases
      (a) the easier it is to recognize the message,
      (b) a more refined (a more detailed) description of the observed system's states is available.
   3. The function reaches its minimum value, when all of the observed system's states are equally probable.
   4. The function is zero if and only if the received signal is entirely independent of the emitted signal.
   5. The function's value cannot be increased by any measurable transformation.
   6. The function is additive.

(9) In order to serve a useful purpose in cybernetics, information theory must demonstrate in a practical fashion, that:

[3] The "object" is what we call "the ideal quantity". It plays the role of the **input** of the information channel. "Data" are the **output** of this channel, which add the noise (disturbances and uncertainty) to the input's clean (certain) value.

1. by efficiently suppressing the disturbance factors, it provides a maximum of information, on which to base a decision,
2. it provides a permanent (adaptive) improvement to the control mechanism by analyzing the messages, that are received, using the characteristic features of both the observed system as well as the disturbances.

This paper ([84]) is an example of the work, to which its author devoted his professional life in an endeavor to implement a notion of information theory based on classical statistical principles. It is remarkable, that his formulations are so comprehensive, that they embrace not only the ideas stemming from the classical approach, but also are applicable to the gnostic notions, which are being developed here. A brief comparison of the principal points set out above with respect to information theory (IT) with the requirements of gnostic theory (GT) follow:

(1): In GT, the IT's "message" is a datum, the "signal" is the datum's ideal value and the "noise," the numerical image of the uncertainty. These notions take no account of the physical and/or technical nature of the vehicle, which carries the message. This is demonstrated among others by equation 11.8, which shows, that the gnostic probability p depends only on the ratio $(A - A_0)/S$. The scale parameter has the same physical dimension as A and A_0. The vehicle, which carries the message (the data values) may have different physical dimensions (eg US$, Volt, ton, year), but they cancel out, when the ratio is evaluated. Information change I_j (10.52 depends only on the probability p. It is thus independent of the nature of the carrier of the message.

The gnostic concept is an abstract idea, which results from Axiom 1 (using abstract algebra), and the information formula (10.52) is developed from this axiom by pure mathematical reasoning.

The goal of GT is identical to that of IT: to minimize the effect of the uncertainty, which degrades the message (data).

(2): The basic structure of GT is analogous to that of IT: The observed system ("emitter" of the message) is the commutative group of real quantities, the observing system ("receiver") is given by the commutative group of observed data. The interaction between the two systems ("information channel") is mathematically modeled as the quantification process.

(3): The receiver/emitter interactions in the GT's systems (described in the previous chapter) cause changes in energy and in (thermodynamic) entropy, that are evaluated for the different states, that occur. GT's realization of the Perez's third point is preferred over the IT approach, because the entropy in GT is a measure of the quality of the signal's (the data's) energy, while in the IT methodology, Shannon's information and Boltzmann's negative statistical entropy have no direct relation to the signal's energy changes.

(4): In GT, the complete balance of (thermodynamic) entropy and information results from the analysis of the sources of the entropy field, which leads to formula 10.62, the mutual conversion of entropy to information.

(5): The fidelity of the "reflection" (to be understood as the quantification and/or estimation transformation) of the true (ideal) value is measurable using the G-weight (9.3) and/or G-irrelevance (9.4). These G-characteristics are closely interrelated: they can be interpreted respectively as the derivative and the integral of the latter and vice versa, because they can be expressed as trigonometric or hyperbolic functions (see 9.15). These characteristics really play a basic and crucial role in gnostics: they are components of the rotation operator (9.2), which represents the uncertainty, and they are closely connected with the metric used for measuring uncertainty (9.15). These features also determine the amount of thermodynamic entropy (10.26), the probability distribution (10.42) and density (11.8) of an individual datum, the sources of its entropy field (10.61), the sources of the datum's information field (10.62) and the information change (10.58) caused by uncertainty.

(6): An entropy increase over the quantifying process, due to observing interactions (stemming from the uncertainty) of the two systems, is obvious from 10.26 with f_i (by 9.10) and q (by 9.8). The fact, that the entropy increase can never be fully compensated by the results of observation (by using and treating the observed data), is demonstrated by the positive value of the entropy residua at the termination of the Ideal Gnostic Cycle, as computed by equation 10.63.

(7): Although there is no notion of randomness in GT, the emitting to receiving systems' interactions (the three stages of the Ideal Gnostic Cycle) are not described by a deterministic mathematical model because it is not possible to find the two unknown quantities (Z_0 and Φ) from the single observed quantity (Z). It is for this reason, that the conditional probability distribution 11.7 and its density function 11.8 must be used instead of a one-to-one mapping $Z_0 \leftrightarrow Z$.

(8): The function $H(p)$ 10.50, which is used to evaluate changes of information due to estimation in GT, satisfies the requirements listed above in items 8.1 through 8.6. This is a direct result of (Shannon's) information theory since the formulae used in GT are (formally) the same functions as in IT. Moreover, information theory shows, that this function is the unique real continuous function of a set of probabilities (eg $P_1, P_2, \ldots, P_M$), which satisfies the condition of invariance of its form, when the probability model is refined by increasing M, ie when the model is made more detailed, by describing more states.

(9): It will be shown through examples in succeeding sections, that the gnostic notions of information have practical application in a number of areas including decision and control as well as adaptive systems.

While Perez's vision of an ideal information theory (as we interpret it) as described in the foregoing summary was addressed to statisticians and to information theorists developing Shannon's ideas based on statistical notions of probability, it is also completely replicable by the application of the gnostic principles, which have been presented. However, the gnostic methods have additional features, which deserve attention, and which represent useful tools, that are not available in the standard information theory approach:

- All of the important notions of IT are represented in GT by mathematical formulae, which were obtained by mathematical reasoning, and which immediately connect them to the data.
- The notion of Boltzmann's statistical entropy as well as of Shannon's information formula are based on an a priori given probabilistic model (on a given system of probabilities $P_1, P_2, \ldots, P_M$), which describe the possible states of the system. However, the problem of estimating these probabilities from the data is external to IT, and has no connection to the data entropy or data information. In contrast, gnostics presents a joint mathematical model of data uncertainty, data entropy, data information, data probability and of the Ideal Gnostic Cycle as an instrument capable of estimating all these characteristics of data uncertainty.
- Both Boltzmann's and Shannon's approaches to uncertainty are based on the standard statistical concept of collective (mass) random events. Here again, the gnostic characteristics of uncertainty including probability, entropy and information have been derived to treat an individual event (neither a random nor a deterministic occurrence), but an uncertain one (due to a lack of knowledge about the event's characteristics).

- Boltzmann's formula for a system's statistical entropy is introduced by a direct definition of the statistical mean of the logarithmic probabilities of the system's states. The gnostic notion of entropy, however, has been derived (using a Gedanken-experiment) beginning with the more elemental notion of thermodynamic entropy of the Clausius's type, which is not based on any probabilistic concept.
- Shannon's information formula appeared heuristic as the negative of Boltzmann's probabilistic entropy. This step was motivated by the simple idea, that information is something directly opposite to entropy. Once again, in contrast, the gnostic notion of information is derived by consistent mathematical reasoning beginning with the elemental Axiom 1 all the way to the entropy $\leftrightarrow$ information conversion equation 10.62. The process has shown, that the relation between information and (thermodynamic) entropy is much more complex. A conversion law of this nature is not available within the framework of standard information theory.
- Probability in gnostics was determined as a by-product of the derivation of information and not as an a priori known notion. This derivation shows, that probability and information as used in GT are inseparably interdependent. This statement is also supported by another result of GT: Consider product $p * (1 - p)$, which is used in IT as an alternative measure of information entropy ([111]). In GT it has the following form:

$$p * (1 - p) = \frac{f_i^2}{4}, \tag{11.19}$$

  which results from 10.42, 9.4 and 9.3. There is a direct link between this product and information in GT—formula 10.51, the reciprocal of the product 11.19 is the source of the field of information (10.50). This is another useful linkage, which does not exist in IT.
- Improbability (p_j, 10.56) and Q-information (I_i, 10.58) have come about as by-products of this derivation and these notions extend the set of usable G-characteristics of uncertainty. All these gnostic characteristics are inherently robust, which permits them to suppress various kinds of data disturbances with a different intensity[4].
- The next chapter will demonstrate how the four gnostic integrals of virtual movement (Q- and E-entropy, Q- and E- information) are subjected to variation theorems, from which the optimality of the Ideal

[4] The uncertainty of a datum (from whatever sources), which would be different for each datum, results in the assignment of a specific weight to the datum.

Gnostic Cycle is derived. This manner of justifying an estimate's optimality, although widely used in physics, is seldom employed in theoretical statistics.

The unusual notion of improbability deserves a further comment. Unlike probability, it is a complex variable. A complex characteristic of the uncertainty of events is known from quantum mechanics—the wave function of a particle. This quantity is not directly measurable, but its modulus can be determined experimentally to characterize the expected spatial distribution of the particle (the spread of its location). This means, that the exclusion of a complex characteristic of uncertainty from consideration cannot be based on arguments of a merely formal nature. On the other hand, Q-information is a real function, which can be useful due to its special kind of robustness.

It can thus be concluded, that both the notions of gnostic probability and gnostic information changes are well justified.

11.3 Why the Least Squares Method (Sometimes) Works

The notion of "the best estimate" of an unknown quantity from observed data is closely connected to a definition of "the best". In mathematics, "the best" is ordinarily identified with the solution of an extremity problem. A criterion function evaluates the quality of the desired unknown quantity's estimate, and the estimate, that minimizes or maximizes the function, is accepted as the best estimate. The most popular criterion function is doubtless the second statistical moment of the estimate or the mean of the sum of quadratic estimating errors. This relates not only to one-dimensional estimation but also to multidimensional problems (eg to regression modeling).

There are several important reasons for the least squares method's popularity:

1. It can be well justified theoretically in statistics for a number of statistical data models:
   (a) The least squares estimate may be shown to be a special case of the broadly accepted estimation method of *maximum likelihood.*
   (b) The method may yield *sufficient* estimates, ie estimates, which make use of all knowledge about the estimated parameter available in the given data sample.

   (c) The estimate may be *efficient*, ie its variance may reach the lower bound of variance among all possible estimates of the given class.
   (d) The method yields an *unbiased* estimate.
2. It is simple and easily understood.
3. It is a familiar concept because it is taught in basic statistical courses and explained in statistical textbooks.
4. Its numerical solution is frequently simple because of linearity with respect to data.
5. Its numerical procedures are available not only in statistical software packages, but also in spreadsheet programs and even on pocket calculators.
6. Its results (when they consist of the solution of a system of linear equations) are unique. No interpretation problems arise in such cases.

As always in mathematics, if the theoretical assumptions of the method are warranted, then the method "works" in the sense, that its results have the theoretically predicted qualities. Conversely, if the assumptions cannot be justified, then not very much can be said about the quality of the results.

Interestingly enough, many practicing statisticians can confirm, that the least squares estimating method sometimes "works" in a practical manner not only for the "proper" uses, but also for unknown data models. Moreover, it sometimes works, even when applied to data, which evidently do not satisfy the theoretical assumptions. This outcome can be explained theoretically in the framework of gnostics by reference to the results of Chapter 10.

Indeed, when the absolute value of an (additive) data observation error $S\Phi$ is sufficiently small with respect to the first terms to neglect the second terms of 10.65 and 10.66, then all the characteristics E_j, E_i, I_j and I_i approach quadratic errors. This means, that—in such special cases of relatively precise data—minimization of quadratic errors also minimizes both entropy and information changes caused by data uncertainties.

The term "sometimes" used in the heading of this section in connection with a proper outcome of the least squares method can be now made more specific; the method can be expected to give good results if at least one of following assumptions is true:

Statistical: The data model satisfies all the requirements of statistical theory, under which the applicability of the method is warranted.

Gnostic: Data errors are sufficiently small to neglect the deviation of entropy and information changes from the quadratic functions in 10.65 and 10.66.

Another interesting conclusion can be drawn from 10.65 and 10.66 and is also reflected by 10.67: if the absolute data error is relatively small, then the Ideal Gnostic Cycle (approximately, up to the fourth power of the data error) is *reversible*. The *real* meaning of this conclusion is, that "ideal" estimators can exist, which completely remove the entropy increase and the decrease in information caused by data uncertainty. For other situations, gnostics is more realistic, showing that a process working in accordance to the Ideal Gnostic Cycle cannot completely recover the damage done to data by uncertainty.

From this point of view, the least squares method can be considered an approximation to the special case of gnostic estimating procedures, justified only in cases of sufficiently precise data.

The theoretical importance of this section is in that a theory (GT), valid generally even for gross data errors, has been shown under some special constraints (small data errors) to provide not only the same results as a broadly accepted (statistical) theory, but also to explain and quantify the limits of suitability of the particular statistical method. In other words, an important non-empty intersection of the two theories based on substantially different paradigms has been set out.

11.4 The Estimating Characteristics of Uncertainty

It will be useful to summarize the results of the previous sections, at least for the estimating phase of the Ideal Gnostic Cycle, in a vivid form. The practical importance of the estimating characteristics of data uncertainty is in their robustness with respect to **outliers**. It is this kind of robustness, that is needed more frequently in applications. This does not mean, that the opposite kind of robustness—with respect to inliers—is unimportant. Such robustness (achievable by using quantifying characteristics) is especially useful for problems, where values of rare large signals are observed over the noise created by many "false" impulses, which have smaller amplitudes. This robustness of quantifying characteristics results from the analytic form of quantifying weight and irrelevance (9.10 and 9.11 with $c^2 = 1$) as was explained in Chapter 9.

We limit ourselves at this juncture to the estimating characteristics because they give a clear insight into the nature of robustness in estimation.

A study, which includes both kinds of robustness will be examined later on.

To classify uncertainty by its size, it is useful to introduce the relative estimating error

$$\delta_k = \frac{A_0 - A_k}{S}, \tag{11.20}$$

where A_0 is again the ideal and A_k the observed value of the k-th datum, and where S is the same scale parameter as before. Four classes of data errors by size are defined in Tab. 11.1:

Error's size	**Symbol of the class**	**Approx. bounds**
Very small errors	VS	$\|\delta_i\| \leq 0.005$
Small errors	SE	$0.005 \leq \|\delta_i\| \leq 0.015$
General case	GC	$\|\delta_i\| < \infty$
Extreme	EX	$\|\delta_i\| \to \infty$

Tab. 11.1 Four classes of data errors

Tab. 11.2 clarifies the behavior of estimating characteristics with respect to the size of data error. The bounds were estimated by numerical methods.

Estimation characteristics	**Class of the error**			
	VS	SE	GC	EX
Data error	$2 * \delta_i$	$2 * \delta_i$	h_i (9.4)	-1_+ or $+1_-$
Data weight	1	$1 - 2 * \delta_i^2$	f_i (9.3)	0_+
Entropy fall	0	$-2 * \delta_i^2$	$f_i - 1$ (10.26)	-1_+
Probability	1/2	$1/2 - \delta_i$	p (10.42)	0_+ or 1_-
E-information	0	$2 * \delta_i^2$	I_j (10.52)	log(2)

Tab. 11.2 Estimation characteristics of data uncertainty for different classes of data errors

This small table is worth a careful examination, as it summarizes the results of the foregoing section. The fourth column (the general case) identifies each variable with the applicable general formula, which is valid for an error of any arbitrary size. The other columns contain approximations obtained for the special cases of data size, which were defined in Table 11.1:

Very small data errors (VS): The estimating error evaluated by the irrelevance approaches the value of the relative error, which is one of the traditional evaluations of the error. (The constant factor 2 does not play a significant role.) Data weight is a constant, independent of the error value. All data have the full weight of 1. Probability equals 0.5: all we can get from the datum's value is, that the unknown ideal value may be either less or more than the observed data value. This conclusion does not depend on the data error. Entropy's changes as well as information changes are completely ignored.

Small errors (SE): Error evaluation is the same as for very small errors, but the data weight decreases with increasing squared error—worse data are getting a smaller weight than the better ones. The probability of the ideal value is a linear function of the relative error. This permits a rough characterization of the probability's dependence to the error for data values close to the ideal value. The entropy and information changes are evaluated by the same formula, but with opposite signs. Entropy changes are thus completely balanced by the information changes—the quantification/estimation cycle is (approximately) reversible in this special case. The quadratic character of both functions of uncertainty supports use of the least squares method for data contaminated with small errors.

Extreme errors (EX): The bounds for errors obtainable for gross data errors and outliers are obtained as limits of the general case (GC) and are shown in the last column of Tab. 11.2. The most important fact is, that all estimating characteristics are bounded—an unlimitedly increasing or decreasing data error (outlier) cannot force an estimating characteristic beyond its finite range. This feature will be shown later to be the source of robustness with respect to outliers, which characterizes the gnostic estimating procedures.

11.5 Summary

The method of kernel estimation of probability density and distribution plays an important role in statistics and it is supported by a well developed theory, that establishes the conditions, under which it is unbiased and consistent. Its value is rooted in the broad range of distribution functions and data, to which it is applicable. It has been shown, that the derivative of gnostic probability functions satisfies all the conditions for use as a kernel estimate, and therefore the methodology can be used to estimate gnostic probability functions. It was also demonstrated, that the gnostic kernel plays a geometric role and is suitable for use as a Riemannian metric function for measuring distances (in probabilistic terms) between the observed datum and its estimated value. The concept of gnostic kernels can be generalized in a natural way far beyond the borders of the statistical concept of kernel estimation. The great flexibility of the forms of gnostic kernels extends their applicability to a very broad choice of distribution functions and densities. All the foregoing justifies the acceptance of the gnostic probability function.

A general concept of information theory as presented by Albert Perez (which can be found in the literature) is sufficiently broad, that it can support not only concepts of information using the usual familiar statistical probability theory, but it also can accommodate other approaches. When gnostics is coupled with information theory, not only are the necessary requirements fully satisfied, but a further advantage is gained by the addition of several features, which are not available, when the usual classical form is used.

The results of the gnostic theory explain, why the very popular least squares sometimes works, not only when its application is justified by statistical theory, but also in more general situations. Gnostics shows, that least squares methods are nearly optimal, if the relative data errors are sufficiently small. This—both theoretically and practically—important result is obtained by a detailed consideration of the behavior of gnostic characteristics. This analysis also documents the robustness of gnostic estimating characteristics of data uncertainty with respect to outliers or inliers.

Chapter 12

Optimality of Gnostic Characteristics

12.1 Pragmatism and Theory

A conflicting definition of the notion “pragmatic” is given in [80]:

> **Pragmatic** ... *dealing with problems in a practical way rather than by following theory or principles.*

If we are to accept this premise, then all science would be nothing, but a bare collection of ornaments, which embellish life as we know it. From this point of view it would be difficult to understand, why the world’s most successful industrial companies develop in their laboratories not only applied, but also basic research on a level frequently honored by the Nobel prize. Such activities are motivated by more than mere philanthropy. Nuclear power stations, antibiotics, laser, electronics, communication satellites and many other examples of recent technologies are all very practical, but they would not exist without a highly developed scientific background. From this standpoint, a more suitable interpretation of pragmatism can be found in [112]:

> **Pragmatic** ... *testing the validity of all concepts by their practical results.*

Now there is no contradiction in the notion of a “pragmatic theory”, which can be defined as being **a theory oriented to producing practical results in the best possible way**. This definition is in good agreement with the popular statement *the most practical thing is a good theory.*

Gnostics aspires to be such a practical theory; in this context:

Producing practical results is the outcome of data processing, which serves the needs of praxis by applying the principles of the gnostic theory of uncertain data.

The best way is determined comparing the results obtained, when a particular process is examined using different analytical methodologies. The goal of gnostic procedures is to minimize the uncertainty of the outcome by optimizing a suitable gnostic characteristic of the data's uncertainty. The most important among these characteristics are information loss and entropy increase. However, depending on the requirements of robustness, other gnostic characteristics, which are inherently connected with data information and entropy, can be used.

The notion of **the best possible** way for data treatment is based on the theoretically justified limits for the residua of information and entropy changes within the Ideal Gnostic Cycle. The information loss or the entropy increase caused by uncertainty cannot be completely eliminated (10.63 and 10.64); however, using the theory, one can only endeavor to obtain results, which are close to the theoretical limits.

Several gnostic characteristics as well as the notions of the Ideal Gnostic Cycle as a model of quantification and estimation have been developed. The goal of this chapter is to show, that this theory can be employed as the starting point for "producing practical results in the best possible way."

12.2 Gnostic Paths as Extremals

An *extremal* is a path of integration (one of a set of paths connecting two fixed points), for which the path integral's value reaches an extreme value. The popular belief, that the shortest path between two points is a straight line connecting the points is based on the hidden assumption, that one applies Euclidean geometry. However, the use of a different geometry may lead to a different result.

Indeed, consider a movement between two points $\langle x, y\rangle$ and $\langle x+dx, y+dy\rangle$ of the real plane R^2 written as pair numbers $x+cy$ and $(x+dx)+c(y+dy)$. Applying two simple geometries (the Euclidean and Minkowskian), the differential of the distance between the points is

$$d\Lambda_c = |\sqrt{(dx)^2 - (cdy)^2}|. \tag{12.1}$$

Introducing polar coordinates, combining 8.31, 8.32 and 8.35, assuming a constant positive scale parameter, and using $c \in \{j, i\}$, assuming $x > 0, x \geq |y|$ expressions

$$x = \varrho_c(cS\Omega_c)\cosh(cS\Omega_c) \qquad y = \varrho_c(cS\Omega_c)\sinh(cS\Omega_c) \tag{12.2}$$

may be written. (To consider both gnostic paths and alternative paths, the radius variable ϱ_c is taken as a function of the angular variable $cS\Omega_c$.) Differentiating 12.2, substituting into 12.1, and taking into account the equivalences

$$(c \in \{j, i\})(d\cosh(cS\Omega_c) = c^3 S \sinh(cS\Omega_c) d(\Omega_c)) \tag{12.3}$$

and

$$(c \in \{j, i\})(\cosh^2(cS\Omega_c) - c^2 \sinh^2(cS\Omega_c) = 1), \tag{12.4}$$

$$d\Lambda_c = |\sqrt{d\varrho_c^2(cS\Omega_c) - \varrho_c^2 d(cS\Omega_c)^2}|. \tag{12.5}$$

is obtained. To calculate the length of a path $\mathcal{P}$ from a point $u_a = x_a + c y_a$ to a point u_b, either the path integral

$$\Lambda_c(\mathcal{P}) = |\int_{\varrho_c(u_a)}^{\varrho_c(u_b)} \sqrt{1 - \varrho_c^2 \left(\frac{d(cS\Omega_c)}{d\varrho_c}\right)^2} \, d\varrho_c| \tag{12.6}$$

or

$$\Lambda_c(\mathcal{P}) = |\int_{\Omega_c(u_a)}^{\Omega_c(u_b)} \sqrt{\left(\frac{d\varrho_c}{d(cS\Omega_c)}\right)^2 - \varrho_c^2} \, d(cS\Omega_c)| \tag{12.7}$$

can be used, where $\varrho_c(u)$ and $S\Omega_c(u)$ are respectively the radius and angle of the pair number u. In a general case, the path's length may be a real or a complex number.

Due to the nature of the Minkowskian plane, two points can be connected by a continuous line only if both of them are in the same cone of the plane (${}_1U_j$, ${}_2U_j$, ${}_3U_j$ or ${}_4U_j$, see Definition 6, Chapter 10). Hence, we shall assume in what follows, that all points on any path being considered are restricted to the same cone. The same arbitrary restriction will apply in the case of the complex plane because of the one-to-one mapping of complex numbers onto the double numbers introduced in gnostics to represent the duality of quantification and estimation.

The goal is to show, that three special paths (quantification and estimation paths $\mathcal{P}_Q$ and $\mathcal{P}_E$, introduced in Definition 7 of Chapter 10, are extremals. For this purpose, alternative paths $\mathcal{P}^*$, which satisfy the following conditions are needed:

1. they are continuous and at least once differentiable,
2. they connect the same (initial and end) points u_a and u_b as the paths are subjected to variations,
3. the length of the radius vector (ϱ) is not necessarily constant,

4. the radius vector's angles $cS\Omega_c$ of the paths' points are not necessarily constant,
5. the inequality $d\varrho_c^2 - \varrho_c^2 d(cS\Omega_c)^2 > 0$ holds for all points on the path.

Those paths, which satisfy conditions 1 through 4 will be called *variated* paths, while those satisfying all five conditions will be *boundedly variated* paths.

Applying 12.6 to the simple special case of straight lines connecting points u_a and u_b, the angle $cS\Omega_c$ is constant for all points on the paths including the end points, and the second term in 12.6 vanishes. The lengths under the geometries considered are

$$\Lambda_c(\mathcal{P}_L) = |\varrho_c(u_b) - \varrho_c(u_a)|, \tag{12.8}$$

where $\mathcal{P}_L$ stands for the non-variated, straight-line connection of the two points.

For variated paths and $c = i$ the inequality

$$|\varrho_i(u_b) - \varrho_i(u_a)| = \Lambda_i(\mathcal{P}_L) \le \Lambda_i(\mathcal{P}_L^*) \tag{12.9}$$

obviously holds, because the second term added in 12.6 is nonnegative and the case of zero variation is not excluded. This relation becomes the well-known Euclidean *variational theorem* for straight lines:

The straight line connecting two points on the Euclidean plane
is the **shortest** of all variated paths.

However, the Minkowskian case ($c = j,\ c^2 = 1$) leads to more surprising results. The positive second term in 12.6 is subtracted leading to

$$\Lambda_j(\mathcal{P}_L^*) \le \Lambda_i(\mathcal{P}_L) = |\varrho_j(u_b) - \varrho_j(u_a)|, \tag{12.10}$$

where $\mathcal{P}_L^*$ denotes an alternative boundedly variated path. The corresponding Minkowskian variational theorem for straight lines thus states:

The straight line connecting two points on the Minkowskian plane
is the **longest** of all boundedly variated paths.

It has been shown, that linear paths can model a real (eg inertial) movement, while the circular gnostic paths (orthogonal to paths of real movements) model the quantification or the estimation process, ie the virtual movement under the action of uncertainty.

Let us now use the integral 12.7 to consider variation theorems for gnostic circular paths. Both the Q- and E-paths are completely determined by the theoretical model of an observed datum. Let a particular datum be modeled by the double number $u_m = Z_{0,m} \exp(S\Phi_m)$. The radius of the quantification circular path is equal to the ideal data value 8.32

$$\varrho_{j,m} = Z_{0,m}, \tag{12.11}$$

while the radius of the estimation path is (by 10.8)

$$\varrho_{i,m} = Z_{0,m}\sqrt{\cosh(2S\Phi_m)}. \tag{12.12}$$

It is important to note, that $Z_{0,m}$, S and Φ_m are constants for any given datum, therefore both radii $\varrho_{c,m}$ $(c \in \{j, i\})$ are constants also. The polar coordinates of points on the circular paths will be denoted by ϱ_c and $cS\Omega_c$. Circular paths having different radii are geometrically similar. To analyze a general case of Q- and E-paths, which have different radii, it is therefore useful to introduce *the relative length* of these paths (and of paths obtained by their variations); this is defined as

$$\lambda_{c,m} = \frac{\Lambda_{c,m}}{\varrho_{c,m}}, \tag{12.13}$$

where $\Lambda_{c,m}$ is integral 12.7 taken for $cS\Omega_c(u_a) = 0$ and $cS\Omega_c(u_b) = cS\Omega_{c,m}$. By substitution of 12.7 into 12.13

$$\lambda_{c,m} = |\int_0^{cS\Omega_{c,m}} \sqrt{(\frac{1}{\varrho_c}\,\frac{d\varrho_c}{d(cS\Omega_c})^2 - 1}d(cS\Omega_c)| \tag{12.14}$$

is obtained.

For a constant radius the first term vanishes and expression 12.14 reduces after trivial integration to

$$\lambda_{c,m}(\mathcal{P}_c) = \sqrt{-(c\,S\Omega_{c,m})^2}, \tag{12.15}$$

where $\mathcal{P}_j$ refers to $\mathcal{P}_Q$ and $\mathcal{P}_i$ refers to $\mathcal{P}_E$.

The relative length of the estimation path $(c^2 = -1)$ is thus real, while that of the quantification path is purely imaginary. For a variated E-path $\mathcal{P}_E^*$ equation 12.14 leads to the relation

$$S|\phi_m| = \lambda_{i,m}(\mathcal{P}_E) \leq \lambda_{i,m}(\mathcal{P}_E^*), \tag{12.16}$$

because the first term in 12.14 is non-negative. In the case of the Q-path a bounded variation of the path does not change the imaginary character of the relative length. For the moduli of imaginary relative lengths of Q-paths the relation

$$|\lambda_{j,m}(\mathcal{P}_Q^*)|_i \leq |\lambda_{j,m}(\mathcal{P}_Q)|_i = S|\Phi_m| \tag{12.17}$$

results from 12.14.

This leads to the following variational theorems for gnostic Q- and E-paths:

Theorem 12: Let $\mathcal{P}_c$ be a Q-path $(c = j)$ or an E-path $(c = i)$ of an Ideal Gnostic Cycle defined in accordance with Definition 7 for an observed datum, the dual model of which is $u_m = Z_0 \exp(j\ S\Phi_m)$.
Let $\mathcal{P}_c^*$ be a variated (for $c = i$) or boundedly variated (for $c = j$) path defined as above.
Let ϕ_m be the angular coordinate, for which relation $\tan(S\phi_m) = \tanh(S\Phi_m)$ holds.
Then

A. The relative length of the circular E-path $\mathcal{P}_E$ equal to $S|\phi_m|$ is the **shortest** of all variated E-paths $\mathcal{P}_E^*$.

B. The modulus of the relative length of the Q-path $\mathcal{P}_Q$ equal to $S|\Phi_m|$ represents the **longest** of all boundedly variated paths $\mathcal{P}_Q^*$.

These variational theorems, which can be proved even using a more strict and complete way (see [61] and [101]), have important consequences related to the interpretation and optimality of the Ideal Gnostic Cycle.

12.3 Extremality of Entropy and Information

12.3.1 Extremality of a Path Integral

Calculus of variations as a method for solving problems of the extremality of path integrals was motivated by both scientific and practical needs, especially in the development of physics. It became apparent, that important laws of physics could be formulated in extraordinarily economical and elegant forms, which are called *variational principles.* So, eg the basic (Newton's) equations of mechanics (primarily accepted as axioms) can now be obtained by assuming the validity of Lagrange's or Hamilton's variational principles [67].

Variational techniques not only serve as an alternative mathematical language to reformulate known regularities of Nature, but also as an inspirational tool for discovering new relationships.

Variational principles are deeply interconnected with the conservation laws (such as eg the Energy and Momentum Conservation Law) and with a special category of path integrals called *stationary* [1]. As stated above, if the value of these integrals depends on the integration path, then the path, for which the integral's value reaches an extreme value is called *the extremal.*

A simple example will illustrate these notions. An artificial satellite above the Earth can maintain (by inertia, without activating its rocket drives) an elliptical orbit. The parameters of this orbit are determined by the initial conditions, which result from the launch process. The orbit is the extremal, which satisfies the variational principle—the minimization of the total (kinetic and potential) energy of the satellite (which may be evaluated by a path integral). No additional energy and no additional momentum is needed for it to sustain this periodic movement. A deviation from the orbit (a variation of the path) is possible only by changing energy and momentum, which results in a deviation from the original inertial orbit (the previous extremal path).

There is a significant difference between the roles of path variation in physics and in gnostics. Physics uses **variational principles**, which represent formal alternatives to Laws of Nature. These laws were discovered and then generally accepted, because they did not conflict with experience and no scientific evidence was proposed falsifying them. However, their logical status are **hypotheses, which have not yet been disproved**; this gives them the same status as basic axioms. In contrast, gnostics proves such features as *variational theorems*, which result from much more elemental and directly verifiable gnostic axioms. However, the mathematical technique, that applies to the variation theorems of gnostics, is identical to that, which is used in the case of variation principles of physics.

There are several approaches to the formulation and solution of variational problems. In Riemannian geometry, the extremal line is called *the geodesic.* It can be calculated if the (Riemannian) metric of the space has

[1] A path integral is called stationary if its value does not change under small variations of the integration path.

been determined. Variational features of functionals may also be studied by direct analysis of integrals and integration paths. This technique was used in [61] and [101] to show the extremity of gnostic characteristics. The advantage of this approach lies in the precise and complete characterization of the neighborhood of the extremal, within which the variational theorem holds.

It is sufficient for the purpose of this book to make use of only one simple method based on a well-known classical lemma [67].

Lemma 1:
Let $u(t)$ be a differentiable and initially unknown function $R^1 \rightarrow R^1$. Denote $\dot{u} = \frac{du}{dt}$.

Let $F(u, \dot{u}, t)$ be a given differentiable function $R^3 \rightarrow R^1$ and t_1 and t_2 given fixed numbers.

Let I be the integral

$$I = \int_{t_1}^{t_2} F(u, \dot{u}, t) dt. \tag{12.18}$$

Then the integral I is stationary if and only if the following condition is satisfied:

$$\frac{\partial F}{\partial u} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{u}}\right) = 0. \tag{12.19}$$

The proof of this lemma can be found in the appendix to this chapter.

If a path is extremal, then the path integral must be stationary, but the opposite is not automatically true. The stationarity of a path integral does not directly imply extremity of the path, because there exist path integrals, which do not depend on the path at all. This is why a determination of stationarity has to include a demonstration, that the integral really depends on the path.

12.3.2 Variational Theorems for the Q- and E- Entropy Change

Lemma 1 can be used to show the extremity of the changes in entropy. Substituting 9.3 into the formula of the entropy change 10.26,

$$E_{c,m} = \int_0^{2c\, S\Omega_{c,m}} \sinh(2c\, S\Omega_c) d(2c\, S\Omega_c) \tag{12.20}$$

is obtained, where $\Omega_{c,m}$ is a fixed pair number and Ω_c a pair variable. This relation can be rewritten as

$$E_{c,m} = 4\int_0^{c\,S\Omega_{c,m}} \cosh(c\,S\Omega_c)\sinh(c\,S\Omega_c)d(c\,S\Omega_c). \tag{12.21}$$

Denoting $t = c\,S\Omega_c$, $u(t) = 2\sinh(c\,S\Omega_c)$ and $\dot{u}(t) = 2\cosh(c\,S\Omega_c)$ one has $F(t) = u\dot{u}$ and $E_c = \int_0^t F(t)dt$, whereby

$$\frac{\partial F}{\partial u} = \dot{u} = \frac{d}{dt}\frac{\partial F}{\partial \dot{u}}. \tag{12.22}$$

The integral $E_{c,m}$ 12.21 is thus stationary. Consider the integral 12.20 for a variated path $\mathcal{P}_c^*$:

$$E_{c,m}^* = \int_0^{2c\,S\Omega_{c,m}} \sinh(\lambda_c(\mathcal{P}_c^*))d(\lambda_c(\mathcal{P}_c^*)), \tag{12.23}$$

where $\mathcal{P}_c^*$ is again $\mathcal{P}_Q^*$ for $c = j$ and $\mathcal{P}_E^*$ for $c = i$. By applying Theorem 12 (inequalities 12.16 and 12.17) to integral 12.23 and assuming a non-zero uncertainty ($S\Omega_{c,m} \neq 0$), it can be seen, that for a non-zero boundedly variated quantification path, the inequality

$$0 < E_{j,m}^* < E_{j,m} \tag{12.24}$$

holds, while for non-zero variations of the estimation path

$$E_{i,m} < E_{i,m}^* < 0. \tag{12.25}$$

We have thus arrived at the variation theorem for the entropy changes.

Theorem 13: Let $E_{c,m}$ ($c \in \{j, i\}$) be the entropy change (10.26) taking place within the quantification or estimation phase of the Ideal Gnostic Cycle applied to an observed datum, the dual model of which is $u_m = Z_0\exp(j\,S\Phi_m)$.

Let $E_{c,m}^*$ be the entropy change corresponding to the variated ($c = i$) or boundedly variated ($c = j$) path of the integral 12.23 for $j\Omega_{j,m} = \Phi_m$ and $i\Omega_{i,m} = \phi_m$.

Let $S\Omega_{c,m} \neq 0$ and let the trivial case of variations identically equal to zero be excluded.

Then

$$0 < E_{j,m}^* < E_{j,m} \tag{12.26}$$

and

$$E_{i,m} < E_{i,m}^* < 0. \tag{12.27}$$

The increase in entropy from quantification is thus maximized, when the integration path $\mathcal{P}_Q$ is followed. Conversely, if the estimation path is used, $\mathcal{P}_E$, the entropy decrease is as large as possible.

12.3.3 Variational Theorems for E- and Q-information

To show in a simple way how both Q- and E-information are subjected to variational theorems, it is useful to recall equations 10.43 and 10.44, which describe the relationship between sources of entropy and information fields.

Q-information is obtained by integrating the **Q-entropy** along the **estimation path** with a radius of r_i. If a variated path $r_i^{'}$ is used for the same entropy field, the absolute value of the variated sources of information increases, because $r_i < r_i^{'}$ by 12.9. Hence, relation

$$-\frac{1}{(1-p)p} < -\frac{1}{(1-p^{'})p^{'}} < 0 \qquad (12.28)$$

(where $-\frac{1}{(1-p^{'})p^{'}}$ is given for a variated source of information) follows from 10.43.

E-information is obtained by integrating the **E-entropy** along the **quantification path** with a radius of r_j. For this radius relation 12.10 holds.

Relation

$$0 < \frac{1}{(1-p_i^{'})p_i^{'}} < \frac{1}{(1-p_i)p_i} \qquad (12.29)$$

based on 10.45 therefore exists between sources of E-information in the case of the path $\mathcal{P}_Q$ and of a boundedly variated path $\mathcal{P}_Q^{'}$. All the sources preserve their signs for all values of their parameters (p, $p^{'}$, p_i, $p_i^{'}$). The stronger the field's source, the stronger the field. Relations, which are valid for the information fields analogous to 12.28 and 12.29 also exist and have been developed for a given (fixed) entropy field. However, due to the extremality of the entropy fields as shown by Theorem 13, both the Q-entropy increase and the E-entropy decline are maximal. These features together with 12.28 and 12.29 justify the variational theorem for information.

Theorem 14: Let $I_{c,m}$ ($c \in \{j, i\}$) be the information change (12.26) taking place within the Ideal Gnostic Cycle of an observed datum, the dual model of which is $u_m = Z_0 \exp(j\ S\Phi_m)$.

Let $I_{c,m}^{'}$ be the information change obtained for the variated ($c = i$) or boundedly variated ($c = j$) integration path of information's sources 10.43 or 10.45.

Let $S\Omega_{c,m} \neq 0$ and let the trivial case of variations identically equal to zero be excluded.

Then

$$0 < I'_{j,m} < I_{j,m} \tag{12.30}$$

and

$$I'_{i,m} < I_{i,m} < 0. \tag{12.31}$$

Variational theorems for entropy and information of individual data represent important results of gnostic theory. They were originally proved using methods, which differ from those applied in this chapter (see [61] and [101]). Unfortunately, the technical details of these methods may complicate understanding of the main ideas, which were offered. The objective of this chapter has been to provide an insight to these thoughts, which is "as simple as possible, but not simpler."

12.4 Optimality as a Game with Nature

To reach a desired destination from a given starting point an aircraft needs a time interval and consumes an amount of fuel, which is dependent on the chosen path. A detailed knowledge of physics is necessary for a theoretical determination of the optimum path to be taken. The role of the aircraft's crew in trying to maintain the best possible path may be interpreted as that of men playing a game with Nature. By using its own laws (eg gravitation, air dynamics), Nature not only impedes the solution of the air transport problem, but also makes it difficult to follow the optimum path by introducing pressure disturbances and air turbulence as it makes its moves in a game played with the crew. A theorist computes the optimum path, a navigator checks for deviations from this optimum path, and the pilot tries to eliminate the deviations. Moves "against Nature" are more successful with better knowledge of both Laws of Nature and the true position of the aircraft.

A similar "game model" may be applied to data processing by representing Nature's move as the introduction of disturbances into the quantification process. A data analyst's "counter-move" is the application of an estimation method, which minimizes the "costs" imposed by the uncertainty contaminating the data. Gnostics provides data analysts with a detailed theoretical description of this game, the Ideal Gnostic Cycle. Theorems 13 and 14 describe Nature's "strategy"—to maximize data entropy and to minimize data information by following the Q-path of the

IGC. Nature's game is thus crafty and refined, but it is always played according to honest fixed rules. The same theorems show, that the entropy increase and the consequent reduction in information can be minimized by using the E-path. This is thus the best "counter-move." We already know from Chapter 10, that Nature always wins in its play with men: residua of entropy and information changes within the Ideal Gnostic Cycle cannot be completely eliminated. However, the amount of uncertainty can be decreased by using better measuring techniques and by identification and elimination of factors, which contribute to data uncertainty. There are therefore two ways of improving the results of the information game with Nature:

1. To maximize data quality by improving observation, identification and measuring techniques.
2. To maximize information obtained from given data by improving data processing techniques.

While the first task is self-evident, the second one is far from trivial. There exist many data processing methods; and all of them are based on a model of uncertainty. Statistical ideas have dominated this field for centuries, but recent doubts about the universal applicability of statistics (eg as discussed by [22], [42], [44], [70], [72], [75] and [44]) have brought about the present state of the art, which can be characterized as a highly structured competition between models of uncertainty and methods based on these models. The paradigm based on the gnostic theory of uncertain data is also one of those competing ideas. It is not likely, that there will be a clear winner; it is more probable, that suitable methods based on different approaches will be found to solve specific tasks. But, among these, the challenge represented by the gnostic model of uncertainty, which explains the "mechanics" and "physics" of uncertainty and culminates in unique variational theorems should not be left out.

12.5 Summary

Unlike physics, which uses variational principles to reformulate its axioms, gnostics develops variational theorems from its much more elemental and, in principle, experimentally verifiable axioms for the paths of gnostic virtual movement and for entropy and information changes caused by data uncertainty. Gnostic variation theorems relate to the phases of the Ideal Gnostic Cycle and state, that

1. The quantification and estimation paths are extremals:

   (a) the relative length of the quantification path represents the maximum from all the boundedly variated paths,
   (b) the relative length of the estimation path is the minimum of all variated paths.
2. For a given individual datum
   (a) the entropy increase caused by data uncertainty reaches its maximum, when the quantification process follows the extremal Q-path,
   (b) the datum's entropy falls to a minimum, when the estimation process follows the extremal E-path.
3. For a given individual datum
   (a) data information is minimized under the effect of uncertainty, when the quantification process follows the extremal Q-path, and
   (b) data information rises to a maximum, when the estimation process follows the extremal E-path.

The variational theorems have shown, that the Ideal Gnostic Cycle can be used as the optimal path to be followed by the data processing algorithms to minimize the effects of data uncertainty.

12.6 Appendix to Chapter 12, Proof of Lemma 1

Since knowledge of calculus of variations is not in everyone's tool-box, a sketch of the proof is included here to facilitate understanding of the principal steps and the results of the chapter, [67]:

Proof of Lemma 1:
Let I be the stationary value of integral 12.18 corresponding to the path $\mathcal{P}$, and $I^{'}$ the integral's value obtained for a path $\mathcal{P}^{'}$ modified by $\delta - variation$, in the following way: assume, that a point $(t, u+\delta u)$ of the variated path $\mathcal{P}^{'}$ is attached to each point (t, u) of the path $\mathcal{P}$. The variation is arbitrary, but sufficiently small and subjected to boundary conditions

$$\delta u(t_1) = \delta u(t_2) = 0. \tag{12.32}$$

This variation may be expressed as

$$\delta u = \vartheta \delta \alpha, \tag{12.33}$$

where α is a parameter defined over the path and ϑ is an arbitrary function of t, for which

$$\vartheta(t_1) = \vartheta(t_2) = 0. \tag{12.34}$$

The corresponding variation of the derivative $\dot{u}$ has the form

$$\vartheta \dot{u} = \dot{\vartheta} \delta \alpha, \tag{12.35}$$

where $\dot{\vartheta}$ denotes the derivative $\frac{d\vartheta}{dt}$. Since the variations are small, the integrand $F(u, \dot{u}, t)$ can be expanded into a Taylor series using only the first terms to obtain the integral $I^{'} = I + \delta I$ for the variated path as

$$I^{'} = \int_{t_1}^{t_2} [F(u, \dot{u}, t) + \frac{\partial F}{\partial u} \vartheta \delta \alpha + \frac{\partial F}{\partial \dot{u}} \dot{\vartheta} \delta \alpha] dt. \tag{12.36}$$

After integration by parts applying 12.34 one obtains

$$\int_{t_1}^{t_2} \frac{\partial F}{\partial \dot{u}} \dot{\vartheta} \delta \alpha dt = - \int_{t_1}^{t_2} \frac{d}{dt} \frac{\partial F}{\partial \dot{u}} \vartheta dt. \tag{12.37}$$

The integral's variation is thus

$$\delta I = \delta \alpha \int_{t_1}^{t_2} \left(\frac{\partial F}{\partial u} - \frac{d}{dt} (\frac{\partial F}{\partial \dot{u}}) \right) \vartheta dt, \tag{12.38}$$

which will be identically zero for an arbitrary function ϑ if and only if 12.19 holds.

Part II

The Gnostic Theory of Data Samples

Chapter 13

Aggregation of Uncertain Data

To this point the (gnostic) answer to the question, “How should the uncertainty inherent in the observed value of an individual datum be measured?” has been derived and justified. It has been shown, that the task of measuring uncertainty is far from a trivial effort, and that it is a long way from what might be considered a common sense approach. Now with the background of the gnostic theory of individual uncertain data, a second significant question can be posed:

Data aggregation problem: Given a sample of data, which are the results of (a real) quantification of a fixed (“ideal”) quantity: how should be the individual uncertain data and/or their characteristics aggregated to obtain those quantitative characteristics of the data sample, which will be useful in estimating the ideal quantity?

This is another seemingly trivial question, which is likely to motivate a non-trivial exploration of a range of issues.

13.1 Data Aggregation in Statistics

13.1.1 Linear and Nonlinear Weighing

It is useful to distinguish between classical and robust statistics: the latter notion is used extensively to describe statistical methodologies, which do not require the analyst to use a specific data model, but permit a broader class of models to be employed instead. The results of robust methods are therefore less sensitive to deviations of real data behavior than those, which rely on classical models. The practical application of statistical methods is thus enhanced at the price of an acceptable loss of efficiency [33]. We will use the notion of “classical” statistics to denote “pre-robust” methods.

A well known data aggregation problem in classical statistics concerns estimates based on data taken from several samples, which have different variances. It is an easy exercise to show, that eg an estimate of the mean of all the data (obtained by a weighed data sum) has the smallest variance, when the weights are the reciprocal of the relative variances of the respective data sources. The same weights must also be used to estimate the weighed mean’s variance as an additive aggregation of the variances of each of the individual data sources. These and similar procedures are worth further comment:

1. There is no connection between the uncertainty of a particular datum and its weight. The source of the weight, which is applied to the datum is the variance, a “collective” characteristic of all the data from the same source. The particular datum’s uncertainty has only an partial—through its contribution to ’collective’ variance—and limited role in the determination of its own weight.
2. The use of constant weights for the additive aggregation of data and of the squared deviations from the arithmetic mean (used in the estimation of the ordinary mean and standard deviation) is a special case, that is justified only when the variance of the data is assumed to be the same for all subsamples of the treated sample.
3. The additive aggregation of data and of the data squares is based on a Euclidean measure of data errors.
4. The data aggregation methods for robust statistics are derived in another way. The data weights, which are applied, depend on the errors of individual data, which in turn are dependent on the statistical models of the data, that are believed to apply. These are based on a priori assumptions about the data, which differ from the assumptions of classical statistics.

There is an alternative (geometric) interpretation of the nonlinear weighing of errors used in robust statistics. Indeed: errors are distances. A distance can be thought of as an interval additively aggregated of subintervals. The sum of the lengths of each subinterval measured along a straight line using Euclidean geometry is equal to the difference between the interval’s end points, independent of the location of the subinterval within the interval. The total length, that represents several Euclidean errors, is obtained by the addition of the length of each individual error; they enter the sum with the same (unitary) weight. The result of such a aggregation is therefore a linear function of the end points of each interval. In contrast, when using Riemannian geometry, the lengths of the subintervals in an interval depend on the location of the subinterval. This causes the aggregation law of errors to be a nonlinear function of the end points of the individual intervals.

Therefore, the data-dependent weights typical in robust statistics can be viewed as an application of a non-Euclidean geometry. A mathematical purist could become concerned about the mathematical consistency of robust statistics, as a structure built over classical statistics. Both approaches have been presented as being based on probability theory, however, probability is defined as an additive measure. It can be said, that statistics evolved from Euclidean geometry. Is it then legitimate to ask, whether it is (mathematically and philosophically) correct to apply non-Euclidean concepts within the framework of a theory, which has been grown from a Euclidean seed? This contradiction is fortunately not a problem with gnostic theory, which has not hidden its close connection with Riemannian geometry.

The data aggregation method used in classical statistics is also interesting, but from another point of view. What might be the origin of the additive aggregation of errors and their squares?

13.1.2 Newtonian Aggregation

Consider a group of N "small" material objects from the point of view of classical (Newtonian) mechanics. The mass of the k-th object will be m_k, and the projection of its velocity on a coordinate axis is v_k. The momentum of this object is thus $m_k v_k$ and its kinetic energy is $m_k v_k^2/2$. Because of the linearity of Newton's equations, the moments may be added so that it may be concluded, that if there are no forces acting on the objects, the sum of the moments (as well as the sum of kinetic energies) of all the objects is constant. Newton's equations are thus equivalents of the (Newtonian) Conservation Law. It is important, that the validity of this Law of Nature (for sufficiently small velocities) had been supported by experience over several centuries. If it is desired to find an object with a mass

$$m_e = \sum_{k=1}^{N} m_k, \tag{13.1}$$

which is equivalent to the mass of a series of smaller objects in the sense of classical mechanics, then the velocity of this equivalent is equal to

$$v_e = \frac{1}{m_e} \sum_{k=1}^{N} m_k v_k. \tag{13.2}$$

and the kinetic energy of the equivalent object is given by

$$m_e v_e^2/2 = \sum_{k=1}^{N} m_k v_k^2/2. \tag{13.3}$$

The aggregation preserves the sum of masses 13.1, moments 13.2, and kinetic energies 13.3.

Suppose, that a smooth curve, $X(t)$, is to be fit to a series of observations, $x(t_1), ..., x(t_N)$, which represent the movement of an object. The observation errors, e_k, evaluated in the Euclidean way are $X(t_k) - x(t_k)$ and their impact on the result is w_k (the manner, in which the weights are determined, is not important).

Inspired by the idea of Newtonian aggregation, the objective is to create a single, aggregated error equivalent to the sum of all N errors. Therefore the linear mappings

$$w_k = K_1 m_k \tag{13.4}$$

and

$$e_k = K_2 v_k \tag{13.5}$$

are introduced. Each weighed error $w_k e_k$ is thus attached to the momentum $m_k v k$ and the weighed quadratic error $w_k e_k^2$ to twice the value of the kinetic energy $m_k v_k^2$. It is then logical to attach the equivalent momentum

$$m_e v_e = \frac{1}{K_1 K_2} \sum_{k=1}^{N} w_k e_k \tag{13.6}$$

to the sum of weighed errors, and the equivalent energy

$$m_e v_e^2/2 = \frac{1}{2K_1 K_2^2} \sum_{k=1}^{N} w_k e_k^2 \tag{13.7}$$

to the sum of weighed error squares. For a scientist accustomed to seeing the world through Newtonian glasses, the best fit of the smooth curve will **minimize the equivalent energy of the fitting errors** under the constraint, that **the equivalent momentum of fitting errors is zero.** In other words, the requirement is, that the sum of the weighed squared errors be minimized, while keeping the sum of weighed errors equal to zero.

From the initial idea of the Conservation Laws of classical mechanics, we have followed a path, that has lead to the family of unbiased least squares estimates of classical statistics. This was not a particularly roundabout trip, for before the specialization and compartmentalization of the various scientific disciplines, researchers pursued interests in many branches: mechanics, astronomy, mathematics, etc., so that these ideas were not followed by a single individual, but very likely by several working independently, and over an extended period of time (Mayer, Euler, Laplace, Legendre, Gauss and others). The point is, however, that the methodology was developed to explain natural phenomena, ie, some perceived Laws of Nature. Among the objects of interest of these scientists an important role belongs to fitting smooth parameterized curves (eg ellipsis) to observed astronomic data. The idea of measuring the fitting errors by moments and energies was thus under then existing conditions natural. And these scientist were among founders of statistics ....

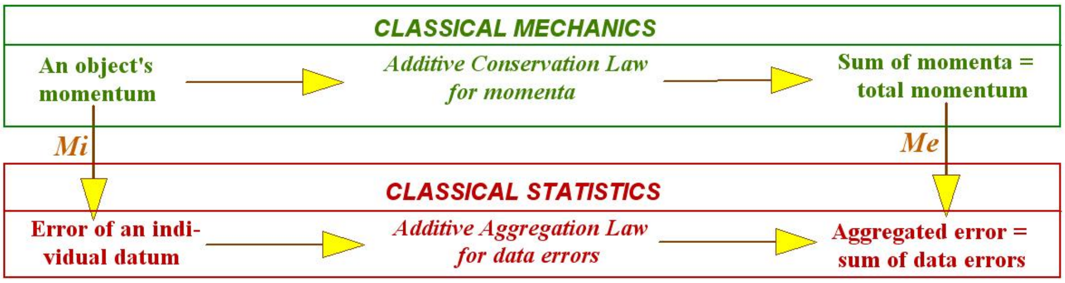

CLASSICAL MECHANICS
An object's momentum
Additive Conservation Law for momenta
Sum of momenta = total momentum
Mi
Me
CLASSICAL STATISTICS
Error of an individual datum
Additive Aggregation Law for data errors
Aggregated error = sum of data errors

Fig. 13.1 Classical aggregation of momenta and errors

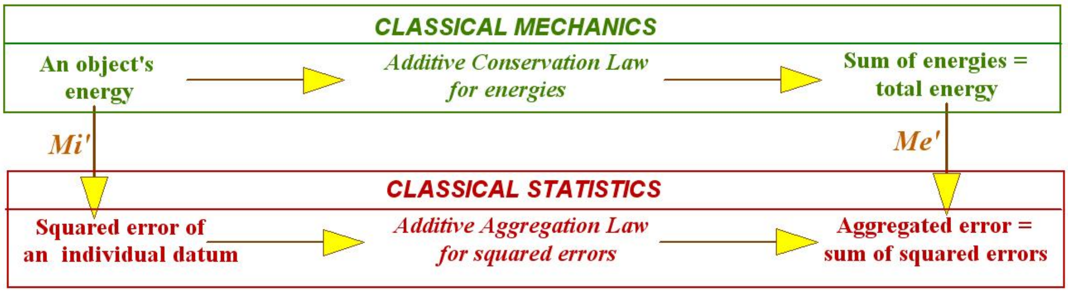

CLASSICAL MECHANICS
An object's energy
Additive Conservation Law for energies
Sum of energies = total energy
Mi'
Me'
CLASSICAL STATISTICS
Squared error of an individual datum
Additive Aggregation Law for squared errors
Aggregated error = sum of squared errors

Fig. 13.2 Classical aggregation of energies and squared errors

The thought behind this 'mechanical' explanation of statistical aggregation can be summarized using the commutative diagrams (Figs. 13.1 and 13.2). The simple mapping Mi in Fig. 13.1 defined by 13.4 and 13.5 depicts the mechanical momentum of an object onto the weighed error of an individual datum. Mapping Me illustrates the mechanical equivalent of the moments of all the objects—the sum of all moments—on the equivalent of all the errors. To ensure the commutativity of the diagram (the equivalence of the result of the transformation of *individual momentum* → *total momentum* → *aggregated error* with that of *individual momentum* → *individual error* → *aggregated error*), one also has to choose the additive aggregation law for data errors. It is obvious from 13.4 through 13.6, that the mappings Me and Mi cannot be chosen independently of each other).

Fig. 13.2 is an analogous commutative diagram for energies and squared errors. An important feature of the mapping of mechanics into statistics is its invariance with respect to the group of linear transformations of the coordinate system moving with a constant velocity.

Another possible inspiration of statistics from mechanics is covariance. In statistics, it is the mean of the product of two centralized random variables. In mechanics, a similar expression exists, and it evaluates the inertial moment of a rotating body. A popular aid for the visualization of multivariate correlation—the correlation ellipsoid—also has as a possible mechanical predecessor: the ellipsoid of inertia.

The link between classical mechanics and classical statistics demonstrated above undoubtedly exists, although it does not exclude other motivations for the additive data aggregation law, ie the aggregation of adding data, linear data errors, and squared data errors. So, eg, mathematicians, who might prefer the use of only a pure mathematical explanation can recognize the source of the mapping in the linear and quadratic character of the statistical and mechanical variables. The above discussion is not meant to imply, that the pioneers of statistics simply copied the mechanics of the time to create the basic notions of statistics. The only real connection is, that these similarities really exist and they provide solid support for the additive aggregation of data, errors and squared errors; however the common roots of this link lie with its geometry. This comes from not only that the notion of linear errors is based on Euclidean geometry just as the notions of classical mechanics, but also from the fact that Galileo made an important contribution to both mechanics and also to Euclidean geometry, which is not as well known as his other contributions to scientific knowledge. The notion of time was not considered in geometry until Galileo explicitly formulated the idea of "time-homogeneous" space, which is equivalent to the assumption of an unlimited speed of light ([118]). This assumption, of course, existed before Euclid's formulation of geometric axioms[1], but it was hidden and accepted unaware of its origin. History has shown, that some of the greatest scientific revolutions were preceded by the explicit statement of assumptions, which had heretofore been implicitly accepted. Later developments in physics rejected the implicit idea of the time-homogeneous space and lead to relativistic mechanics with aggregation laws, which substantially differed from those of classical mechanics. Given all of these parallels, it is not unnatural to expect, that the aggregation of uncertain data also requires substantial parallel revision.

13.2 Aggregation Axiom of Gnostics

13.2.1 Motivations

The decisive impetus for a substantial revision of both the mechanics and geometry of real space resulted from experiments, which proved the finite speed of light. Instead of a time-homogeneous purely geometric space it became necessary to consider space-time along with its non-Euclidean geometry. Relativistic relations respecting the finite speed of light appeared to be invariant with respect to the Lorentz's group of transformations of all inertial (moving with a constant velocity) coordinate systems. Thus this single physical fact determined both the geometry to be used as well as the class of transformations, which describe the underlying processes.

Gnostic theory, which models entirely different processes of quantification began from yet another experience resulting from quantitative observation and the measurement of natural processes: real data form a structure manifesting the features of a commutative group. In developing this idea, we arrived at data uncertainty models, which are in a close (linear) relationship (7.10, Theorem 5) with the energy-momentum tensor of relativistic mechanics. Major features of this relation need to be emphasized:

1. it has been derived for **individual** free relativistic particles and **individual** uncertain data,
2. it holds for **all** pairs joined by the mapping condition $v/v_* = \tanh(\Phi)$ and it is therefore **Lorentz-invariant**,
3. it is **linear**.

The relativistic Conservation Law states, that the energy-momentum ten-

[1] About 300 B.C.

sor of a group of free particles is equal to the sum of the tensors of the individual particles; moments as well as energies are thus aggregated additively. The invariance of the relation 7.10 with respect to Lorentz's transformations warrants its universality. Addition of the left-hand side of this relation results in the addition of matrices $\underline{M}(0, 2\Phi)$. These matrices contain the weights and the irrelevances of individual data. The sum of moments and energies of particles correspond to the sum of quantification weights and irrelevances. Both the quantification weights and the quantification irrelevances of individual data are thus aggregated additively. If one accepts the Conservation Law of relativistic physics, then—so as not to give rise to a mathematical contradiction—one **have** to accept the additive aggregation law for both weights and irrelevances of uncertain data, at least for quantification.

Commutative diagrams in Figs. 13.3 and 13.4 illustrate the idea in the same manner as for classical mechanics and statistics (as was shown in Figs. 13.1 and 13.2). By appealing to well known principles of physics, there is no need to introduce the aggregation law for the quantification phase of the Ideal Gnostic Cycle as an axiom. However, the gnostic the-

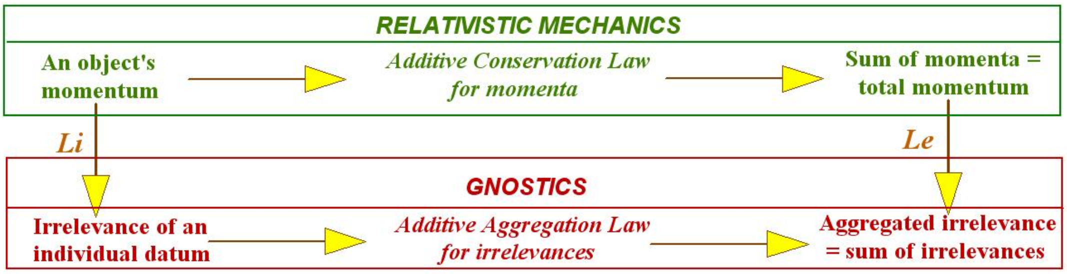

RELATIVISTIC MECHANICS
An object's momentum
Additive Conservation Law for momenta
Sum of momenta = total momentum
Li
Le
GNOSTICS
Irrelevance of an individual datum
Additive Aggregation Law for irrelevances
Aggregated irrelevance = sum of irrelevances

Fig. 13.3 Aggregation of relativistic momenta and of gnostic irrelevances

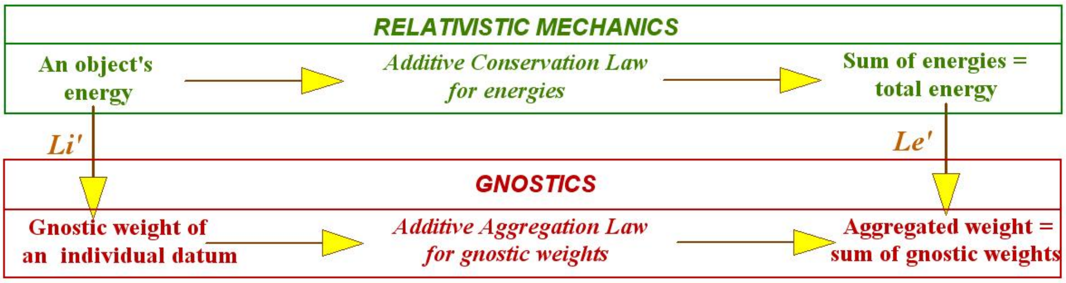

RELATIVISTIC MECHANICS
An object's energy
Additive Conservation Law for energies
Sum of energies = total energy
Li'
Le'
GNOSTICS
Gnostic weight of an individual datum
Additive Aggregation Law for gnostic weights
Aggregated weight = sum of gnostic weights

Fig. 13.4 Aggregation of relativistic energies and of gnostic weights of data

ory of individual data has been developed as a mathematically consistent theory from Axiom 1. Knowledge from other sciences (measurement theory, thermodynamics, mechanics) was used only as motivations to support the mathematical definitions. To maintain the mathematical autonomy of gnostics, it is desirable to accept the relation 7.10 again as a motivation and to introduce the aggregation law of uncertainty as another axiom. The case of estimation also gives reason to proceed in this manner.

There is a one-to-one mapping between quantification and estimation models of each uncertain datum induced by the relation of trigonometric and hyperbolic tangents 8.36. However, this relation **does not** map the group of Minkowskian rotations onto the group of Euclidean rotations. It can be easily seen that the mapping of individual angles considered does not result in the same mapping for their sum. The aggregation laws for estimation irrelevances and weights do not result automatically from those accepted for quantification. One must therefore look for other reasons to choose the estimation version of the aggregation law. The simplest idea is: to take over the additive aggregation law from quantification and to apply it to estimation. There are conditions, which support this idea: both quantification and estimation weights were interpreted as linear functions of thermodynamic entropy, which is an additive quantity. Both quantification and estimation irrelevances are derivatives of the corresponding weights. Hence, additive aggregation of weights results in the additive aggregation of irrelevances.

Other factors, which support the idea of the same aggregation law for quantification and estimation are of a formal mathematical nature. Both types of data models form special structures, 2-algebras of double and complex numbers. Addition is a defined operation within these algebras. Moreover, the similarity between these structures allowed pair numbers to be used and this lead to the formally identical appearance of both the quantification and the estimation formulae. To preserve this advantage and the formal unity, it is natural to accept the same—additive—aggregation law for both kinds of weights and irrelevances.

In order to formulate the aggregation axiom, it is necessary to describe precisely, what shall be understood to be a data sample.

13.2.2 Data Sample

In statistics, the notion of a sample is used together with the notions of population, parameter and statistic ([110]):

- *A population* is the complete and entire collection of elements (scores, people, measurements, and so on) to be studied.
- *A parameter* is a numerical measurement describing some characteristic of a population.
- *A sample* is a subset of a population.
- *A statistic* is a numerical measurement describing some characteristic of a sample.

Because of the substantial difference between the mathematical models, the role played by population in statistics can be only approximated by the set of possible data in gnostics. As we have already seen, the gnostic model of a set of possible data is the commutative group, while the most frequently used model of a statistical population is a much more complex structure, the sigma-algebra. Notions of samples are therefore also different. Statisticians require, that a sample be more than an arbitrary subset of a population; they assume, that the sample has been obtained from the population by a purely random selection. There is no randomness in gnostics and no random selection. Instead, a sample is a finite collection of uncertain data obtained by the quantification processes. They all are assumed to satisfy gnostic Axiom 1. Each of the data has a theoretical model of the type shown in Chapter 5. A sample may include data obtained by the quantification of several (say, L) ideal quantities. To characterize the effect of uncertainty on all the data aggregated into the sample, the characteristics of the sample's uncertainty must be obtained. The treatment of data samples will be the main task of the remainder of Part II of this book. A detailed definition of the important notion of the data sample in gnostics follows:

Definition 12: Given an integer number L ($L \geq 1$) of $N(l)$-tuples of multiplicative data $Z_1(l), ..., Z_{N(l)}(l)$ having models $Z_k(l) = Z_0(l) \exp(S(l)\Phi_k(l))$ ($k = 1, ..., N(l),\ l = 1, ..., L$), where $\Phi_k \in R^1$, and where $Z_0(l)$ and $S(l)$ are positive reals, constant for each fixed l.

Denote

$$\mathcal{Z}_l(N(l), Z_0(l), S(l)) := \langle Z_1(l), ..., Z_{N(l)}(l)\rangle \quad (l = 1, ..., L). \tag{13.8}$$

This $N(l)$-tuple is the *sub-sample* or *cluster*.

The L-tuple

$$\mathcal{Z}(L) := \langle \mathcal{Z}_1(N(1), Z_0(1), S(1)), ..., \mathcal{Z}_L(N(L), Z_0(L), S(L))\rangle \tag{13.9}$$

is the *data sample*. A data sample with an unknown number L of sub-samples will be denoted $\mathcal{Z}$.

A data sample will be called *homogeneous* iff $L = 1$. A data sample with $L > 1$ is *heterogeneous.*

Let $N = \sum_{l=1}^{N} N(l)$. Let $\langle \Xi_1, \Xi_2, ...\rangle$ be such a sequence of mappings, that for all integers N relation $\Xi_N : R_+^N \to R^1$ holds. Denote

$$q_k(l) := (Z_k(l)/Z_0(l))^{1/S(l)} \quad (k = 1, ..., N(l),\ l = 1, ..., L). \tag{13.10}$$

The value of a function

$$\Xi_N(\mathcal{Z}(L)) := \Xi_N(q_1(1), ..., q_{N(1)}(1), ..., q_1(L), ..., q_{N(L)}(L)) \tag{13.11}$$

is the *gnostic characteristic* of the data sample $\mathcal{Z}(L)$. Such a characteristic will be *additive*, if the relation

$$\Xi_{N+1}(\mathcal{Z}(L)) := \Xi_N(\mathcal{Z}(L)) + \varsigma(q_{N(l)+1}(l)) \tag{13.12}$$

(where l is an integer from the sequence $1, ..., L$) holds for a function $\varsigma : R_+ \to R^1$.

Let $f_{c,k}$ and $h_{c,k}$ ($c \in \{j, i\},\ k = 1, ..., N$) be the weights and irrelevances of all N data forming a data sample $\mathcal{Z}(L)$. The gnostic characteristic $F_c(\mathcal{Z}(L))$ is the *weight of the data sample* $\mathcal{Z}(L)$, if it is a function of the weights of the data in the sample. Similarly, the characteristic $H_c(\mathcal{Z}(L))$ is the *irrelevance of the data sample* $\mathcal{Z}(L)$, if it is a function of the irrelevances of data in the sample.

13.2.3 Axiom 2

We are thus prepared to accept following aggregation axiom:

Axiom 2 (axiom of the additive aggregation law): Let $\mathcal{Z}(L)$ be a data sample 13.9 aggregated of N data weights and irrelevances $f_{c,k}$ and $h_{c,k}$ ($c \in \{j, i\},\ k = 1, ..., N$). Then the weight $F_c(\mathcal{Z}(L))$ and irrelevance $H_c(\mathcal{Z}(L))$ of the data sample $\mathcal{Z}(L)$ are

$$F_c(\mathcal{Z}(L)) = \frac{1}{N}\sum_{k=1}^{N} f_{c,k} \qquad H_c(\mathcal{Z}(L)) = \frac{1}{N}\sum_{k=1}^{N} h_{c,k}. \tag{13.13}$$

Both the weight and the irrelevance of a data sample are additive gnostic characteristics of the data sample. In consonance with the motivations discussed above, these characteristics have the same form for both quantification and estimation.

The definition of gnostic characteristics 13.11 is more general than that of the weight and irrelevance set out by Axiom 2. While other characteristics will be useful in the development of what follows, the aggregation law 13.13 is of fundamental importance in defining the gnostic theory of the data sample.

It is emphasized, that the number of data N in a data sample is finite.

13.3 Summary

The additive aggregation law for data, errors, and squared errors commonly used in classical statistics can be justified by the formal coincidence of the basic statistical notions with those of classical mechanics. A weighed linear error is analogous to the momentum of a freely moving mass particle, and the square of this error corresponds to the particle's kinetic energy. The aggregation law for moments and energies is known, and it results from the Energy and Moment Conservation Law of classical mechanics. To complete the analogy and to show the strong support from mechanics to statistics, the additive aggregation law for statistics also must be accepted. The similarity of classical statistics to classical mechanics has deep roots, which are derived from the fact, that both theories are based on the same—Euclidean—geometry.

Just as it was seen in the gnostic theory of individual uncertain data, there is also a close formal relationship between the weights of data and their irrelevances on one hand and moments and energies of freely moving relativistic particles on the other hand. These relationships have a universal validity in the sense of Lorentz's invariance: they exist independently in a broad class of coordinate transformations. To preserve the validity of these relationships for data samples, one has to accept the additive aggregation law for both the gnostic weights of data and the data irrelevances. This choice provides support for a aggregation law of uncertainty from the Conservation Law of relativistic physics. The additive aggregation law for both the quantification and the estimation of data weights and irrelevances is accepted as Axiom 2 of gnostic theory. The similarities between gnostics and (special) relativistic mechanics comes from the fact, that both theories have common origins in non-Euclidean, Minkowskian geometry.

Chapter 14

Gnostic Characteristics of a Sample

14.1 The Modulus of a Data Sample

The data composition axiom defined the weight and irrelevance of a data sample. From these basic gnostic characteristics, a number of other important features of a sample can be derived so as to describe its uncertainty. Definition 12 introduced the notion of a data sample, which included data with different scale parameters, so that the k-th datum's uncertainty was $S_k\Omega_{k,c}$. So as to preclude the necessity of specifying particular scale parameters, unless it is otherwise noted, $\Omega'_{k,c} := S_k\Omega_{k,c}$ will be used to describe the uncertainty of individual data in this chapter.

Definition 13: Let $c \in \{j, i\}$. Let $\mathcal{Z}$ be a data sample, the weight and irrelevance of which are F_c and H_c. Denote

$$\Omega_{Z,c} := \frac{1}{2c} \arg\tanh(c\, H_c/F_c), \tag{14.1}$$

where the symbol argtanh($*$) represents the inverse of the hyperbolic tangent, so that $c\, H_c/F_c = \tanh(2c\, \Omega_{Z,c})$.

The pair number

$$e_{Z,c} := \exp(2c\, \Omega_{Z,c}), \tag{14.2}$$

the components of which are denoted by

$$f_{Z,c} := \cosh(2c\, \Omega_{Z,c}) \quad \text{and} \quad h_{Z,c} := 1/c\, \sinh(2c\, \Omega_{Z,c}), \tag{14.3}$$

is the *equivalent* of the data sample $\mathcal{Z}$.

Then the ratio

$$M_{Z,c} := F_c/f_{Z,c} \tag{14.4}$$

will be called the *modulus of the data sample* $\mathcal{Z}$.

It is obvious, that both pair numbers

$$E_{Z,c} := F_c + c\, H_c \tag{14.5}$$

and $e_{Z,c}$ characterize the uncertainty of the data in the sample. They each represent the overall uncertainty of the sample, but in a different manner. Unlike the pair number $E_{Z,c}$, the data sample's equivalent is normalized in the sense, that $|e_{Z,c}|_c = 1$. It is therefore an operator rotating a vector by the angle $\Omega_{Z,c}$, which is composed of all angles $\Omega^{'}_{c,k}$, and which represents the data of the sample ($k = 1, ..., N$). The composition rule for angles results from the composition axiom 13.13. The argument of the pair number $E_{Z,c}$ 14.5 is the same as that of $e_{Z,c}$, but the modulus $|E_{Z,c}|_c$ generally does not equal 1, because the arithmetic mean of sines (cosines) is not necessarily a sine (cosine) function.

14.2 Some Forms of the Data Sample's Modulus

The modulus of a data sample can be presented in several forms.

Theorem 15: Let $c \in \{j, i\}$.
Let $\mathcal{Z}$ be a data sample of data $Z_k \quad (k = 1, ..., N)$.
Let the pair models of the data be

$$u_k = |u_k|_c(\cosh(c\,\Omega^{'}_{c,k}) + c\, \sinh(c\,\Omega^{'}_{c,k})), \tag{14.6}$$

where

$$|u_k|_c = Z_{0,k} \quad (k = 1, ..., N), \tag{14.7}$$

and where the parameters S_k in $\Omega^{'}_{k,c} := S_k \Omega_{k,c}$ are not necessarily the same for all data.

Then the following statements hold:

a) The modulus of the data sample $\mathcal{Z}$ can be calculated using the relation

$$M_{Z,c} = \sqrt{F_c^2 - c^2 H_c^2}. \tag{14.8}$$

b) This can be rewritten as the modulus of the arithmetical mean of pair operators, which rotate vectors by twice the value of the angles of the sample's data:

$$M_{Z,c} = |\frac{1}{N} \sum_{k=1}^{N} \exp(2c\,\Omega^{'}_{c,k})|_c, \tag{14.9}$$

c) which can also be expressed as the geometrical mean of two special arithmetical means:

$$M_{Z,c} = \sqrt{\left(\frac{1}{N}\sum_{k=1}^{N}\frac{u_k}{Co(u_k)}\right)\left(\frac{1}{N}\sum_{k=1}^{N}\frac{Co(u_k)}{u_k}\right)}, \tag{14.10}$$

where $Co(u_k)$ denotes the conjugate 8.11 of the pair number u_k.

d) An alternative form is

$$M_{Z,c} = \frac{1}{N}\sqrt{\sum_{k,l=1}^{N}\cosh\left(2c(\Omega'_{c,k} - \Omega'_{c,l})\right)}. \tag{14.11}$$

e) If for all data in the sample $\mathcal{Z}$ relations $Z_{0,k} = Z_0$ and $S_k = S$ hold $(k = 1, ..., N)$, then

$$M_{Z,c} = \sqrt{1 + \frac{c^2}{N^2}\sum_{k>l}^{N}(f_{i,k}f_{i,l})^{(1-c^2)/2}((Z_k/Z_l)^{1/S} - (Z_l/Z_k)^{1/S})}, \tag{14.12}$$

where $i = \sqrt{-1}$, and where $f_{i,k} = 2/((Z_k/Z_0)^{2/S} - (Z_0/Z_k)^{2/S})$ is the estimation weight of the $k-$th datum.

Proof of Theorem 15: The result from 14.1 and 14.3 is

$$\frac{h_{Z,c}}{f_{Z,c}} = \frac{H_c}{F_c}. \tag{14.13}$$

Therefore, by 14.4:

$$f_{Z,c} = F_c/M_{Z,c} \qquad h_{Z,c} = H_c/M_{Z,c}. \tag{14.14}$$

Statement a) (above) results from 14.3, which states, that the modulus of the data sample equals the modulus of the pair number $E_{Z,c}$, which by 13.13 represents the arithmetic mean of rotation operators $f_{c,k} + c\,h_{c,k}$, the exponential form of which is $\exp(2c\Omega'_{c,k})$ (9.2). Hence we come to b).

The same form can also be applied to restate 14.8 written as

$$\sqrt{(F_c + c\,H_c)(F_c - c\,H_c)}$$

in the form

$$\sqrt{\frac{1}{N}\left(\sum_{k=1}^{N}\exp(2c\Omega'_{c,k})\right)\left(\sum_{k=1}^{N}\exp(-2c\Omega'_{c,k})\right)}$$

, which is identical to c). Now using the well-known decomposition $\exp(*) = \cosh(*) + \sinh(*)$ and calculating the product of sums, we arrive at d).

The quantification version of e) (14.12) results from d) (14.11) by elementary substitutions:

$$\cosh(2(\Phi_k - \Phi_l)) =$$
$$((\exp(\Phi_k - \Phi_l))^2 + (\exp(\Phi_l - \Phi_k))^2)/2 =$$
$$1 + (\exp(\Phi_k - \Phi_l) - 1/\exp(\Phi_k - \Phi_l))^2 =$$
$$1 + ((Z_k/Z_l)^{1/S} - (Z_l/Z_k)^{1/S})^2.$$

Recalling 10.53 and 9.5, relations

$$\cos(2(\phi_k - \phi_l)) =$$
$$\cos(2\phi_k)\cos(2\phi_l) - \sin(2\phi_k)\sin(2\phi_l) =$$
$$f_{i,k}f_{i,l}(1 + \sinh(2\Phi_k)\sinh(2\Phi_l)) =$$
$$1 + f_{i,k}f_{i,l}(1 - \cosh(2(\Phi_k - \Phi_l)))$$

can be derived, from which the estimation version of the statement e) (14.12) results by 14.11 and by the foregoing formulae.

Formula 14.12 unifies both the quantification and the estimation cases.

It is seen from this proof, that the modulus of a data sample is one of the gnostic characteristic of the data sample $\mathcal{Z}$.

14.3 Some Important Features of Moduli

14.3.1 Ordering of Moduli

Theorem 15 permits the data samples' moduli to be ordered.

Corollary 15.1: Let $M_{Z,j}$ and $M_{Z,i}$ be the quantification and estimation version of the modulus 14.4 of the sample $\mathcal{Z}$. Then following relation holds:

$$0 < M_{Z,i} \leq 1 \leq M_{Z,j} < \infty, \tag{14.15}$$

where the cases $M_{Z,j} = 1$ and $M_{Z,i} = 1$ take place simultaneously.

Proof: Let all data in the sample be precise. Then all arguments $\Omega^{'}_{c,k}$ in 14.9 are zero and both versions of the modulus are equal to 1. If one of either modulus is 1, then all arguments $\Omega^{'}_{c,k}$ in 14.9 must be zero. The quantification and estimation angles $\Omega^{'}_{c,k}$ are bound by the equivalence of their tangents 8.36, they must therefore reach zero simultaneously. The modulus $M_{Z,i}$ is positive, because the moduli of all non-zero complex numbers must be so. The modulus $M_{Z,j}$ is bounded, because all the uncertainties of the real data measured by the angles $\Omega^{'}_{j,k}$ are bounded. The inequalities in 14.15 reflect the fact, that the hyperbolic cosines in (14.11) are for non-zero arguments, and these are always greater than the corresponding trigonometric cosines.

It can be seen from 14.11, that the values of the data sample's modulus are determined by the data spread, which increases as the absolute differences between the data uncertainty and the data's true value become larger. Starting with values of 1 for the moduli of precise data, the quantification modulus increases and the estimation modulus decreases as the uncertainties increase. In order to use 14.11 to evaluate a modulus, the ideal value Z_0 must be estimated so as to obtain the angles $\Omega_{c,k}$. It is remarkable, that this requirement does not involve the quantification modulus, which is completely determined by the data values and by the scale parameter (see 14.12).

14.3.2 The Case of Concatenated Samples

It is instructive to analyze relations between the moduli of concatenated samples.

Corollary 15.2: Let $\mathcal{Z}'$ and $\mathcal{Z}''$ denote two homogeneous data samples, which have the same ideal value Z_0 and the same scale parameter S. Let N' and N'' be the number of data in the samples.

Let $\mathcal{Z}$ be the data sample created by concatenation of the samples $\mathcal{Z}'$ and $\mathcal{Z}''$. Let $c \in \{j, i\}$. Let $M_{Z,c}$, $M_{Z',c}$ and $M_{Z'',c}$ be the moduli of the data samples $\mathcal{Z}$, $\mathcal{Z}'$ and $\mathcal{Z}''$. Then the equivalence

$$M^2_{Z,c} = \left(\frac{N'}{N'+N''}\right)^2 M^2_{Z',c} + \left(\frac{N''}{N'+N''}\right)^2 M^2_{Z'',c} +$$

$$\frac{2}{(N'+N'')^2} \sum_{k=N'+1}^{N'+N''} \sum_{l=1}^{N'} \cosh\left(2c\left(\Omega^{'}_{c,k} - \Omega^{'}_{c,l}\right)\right) \tag{14.16}$$

together with the inequalities

$$(N'^2 M^2_{Z',i} + N''^2 M^2_{Z'',i})^{1/2} \leq$$

$$(N' + N'')M_{Z,i} \le (N'M_{Z',i} + N''M_{Z'',i}) \le (N' + N'') \le$$
$$(N'M_{Z',j} + N''M_{Z'',j}) \le (N' + N'')M_{Z,j} \qquad (14.17)$$

hold.

Proof: By substituting the concatenated data sample into 14.11 and by application of the features of the functions cos(∗) and cosh(∗).

14.3.3 Gnostic Covariance

To simplify notation, the following symbol for the arithmetic mean of an $N-$tuple of real numbers $Q_1, ..., Q_k$ is introduced:

$$\bar{Q} := \frac{1}{N} \sum_{k=1}^{N} Q_k. \qquad (14.18)$$

Definition 14: Let $c \in \{j, i\}$. Let $\mathcal{Z}$ be a data sample composed of data $Z_1, ..., Z_N$. Let $\Omega^{'}_{c,1}, ..., \Omega^{'}_{c,N}$ represent the same quantities as in Theorem 15. Then for all $K = 1, ..., N - 1$ define the *gnostic autocovariance*:

$$\mathrm{acov}_c := \frac{1}{N-K} \sum_{l=1}^{N-K} h_c(2c\Omega^{'}_{c,l}) h_c(2c\Omega^{'}_{c,l+K}). \qquad (14.19)$$

$\mathcal{Z}_A$ and $\mathcal{Z}_B$ are data samples composed of the same number N of data. Let $h_c(2c\Omega^{'}_{c,n,A})$ and $h_c(2c\Omega^{'}_{c,n,B})$ be the irrelevances of these data for $n = 1, \ldots, N$.

Then the *gnostic crosscovariance* is

$$\mathrm{ccov}_c := \frac{1}{N} \sum_{n=1}^{N} h_c(2c\Omega^{'}_{c,n,A}) h_c(2c\Omega^{'}_{c,n,B}). \qquad (14.20)$$

When the context clearly identifies the choice between auto- and crosscovariance, the composite label *G-covariance* can be used.

Gnostic autocovariances play a role, which can be clarified by means of another form of the data sample's modulus:

Corollary 15.3: Let $M_{Z,c}$ be the modulus of the data sample $\mathcal{Z}$ in accordance with Definition 13. It can be written alternatively as

$$M_{Z,c} = \sqrt{(\overline{f_c})^2 - \frac{c^2}{N}(\overline{h_c^2} + 2 \sum_{k=1}^{N-1} (1 - k/N)\, \mathrm{acov}_c(N, k))}. \qquad (14.21)$$

Proof: By expanding the square of the mean value of the irrelevances (H_c) in 14.8, reordering and summing their products, by using definition 14.21, and notation 14.18.

Unlike the values of $(\overline{f_c})^2$ and $\overline{h_c^2}$, which are always non-negative, the covariance terms in 14.21 may be either negative or positive. Their sign depends on the combination of the individual uncertainties. A general impression is, that if there is no systematic interdependence between angles $\Omega^{'}_{c,k}$ of individual data, the gnostic covariances might tend to zero in a manner similar to ordinary (statistical) covariances. Such a conclusion has support, as is shown below for the case of sufficiently precise data. Equation 14.21 shows, that the influence of covariances on the sample's modulus can change its value only if there is a systematic pattern in the data, which produces non-zero covariances. It can therefore be expected, that gnostic covariances (as in classical statistics) can serve as one of the tools for measuring the mutual dependence of data within a sample, but in a **robust manner**.

14.3.4 Gnostic Median

In statistics, the notion of the *median* is related to two situations:

1. The median of a data sample is the middle value, when the data are ordered.
2. The median of a distribution function is the fractile[1] related to the probability 0.5.

In statistics, the median may be (but is not necessarily) connected to another important notion, that of the *unbiasedness* of an estimate (a zero mean of the estimate's error). This idea also has two interpretations depending on the definition of "mean:"

1. The estimate of a data sample's *location* (also called *position*) is *unbiased*, if the arithmetic mean of its errors is zero.
2. In relation to a probability distribution function, an estimate is called unbiased, if the integral of its values weighted by the distribution's density equals zero.

In gnostics, errors are measured by the (quantifying or estimating) irrelevances (9.4). In accordance with Axiom II, the irrelevances are composed additively. The irrelevance H_c of a data sample $\mathcal{Z}$ is the arithmetic mean

[1] The *fractile* (also called *quantile*) is a number (point) x_p, at which the distribution function equals $p(x_p)$.

(13.13) of irrelevances. It is therefore natural to define the sample's *G-median* as the number Z_{med}, for which

$$\frac{1}{N}\sum_{k=1}^{N} h_{c,k}(Z_{med}) = 0 \tag{14.22}$$

holds (where $h_{c,k}$ is the same expression for irrelevances as in 13.13). This equation can be rewritten using 9.11 and 9.7 for the quantifying case as

$$\sum_{k=1}^{N} \frac{q_k^2 - 1/q_k^2}{2} = 0 \tag{14.23}$$

and for the estimating case as

$$\sum_{k=1}^{N} \frac{q_k^2 - 1/q_k^2}{q_k^2 + 1/q_k^2} = 0, \tag{14.24}$$

where

$$q_k = (Z_k/Z_{med})^{1/S} \quad k = 1, \ldots, N. \tag{14.25}$$

These relations show, that

$$\tilde{Z}_0 = Z_{med}, \tag{14.26}$$

ie that the role of the estimate of the (unknown) true value Z_0 is played by the G-median, Z_{med}, in this case. The G-median is thus an estimate of the sample's location parameter. (Other estimates of Z_0 and other location parameters are discussed below.)

Corollary 15.4: Let $Z_{Q,med}$ and $Z_{E,med}$ be quantifying and estimating medians, which satisfy equations 14.23 or 14.24, respectively.
Then

1. The quantile $Z_{Q,med}$ sets the Q-irrelevance $h_{Z,j}$ 14.3 of the quantifying equivalent of the data sample $\mathcal{Z}$ to zero, and maximizes both its Q-weight $f_{Z,j}$ and its modulus $M_{Z,j}$ 14.4.
2. The quantile $Z_{E,med}$ sets the irrelevance $h_{Z,i}$ 14.3 of the estimating equivalent of the data sample $\mathcal{Z}$ to zero, and maximizes both its E-weight $f_{Z,i}$ and its modulus $M_{Z,i}$ 14.4.
3. The sample's Q-median $Z_{Q,med}$ is identical to the quantile of improbability of 0.5.
4. The sample's E-median $Z_{E,med}$ is identical to the quantile of probability of 0.5.

Proof: Both statements 1) and 2) of the Corollary directly result from definitions 14.1 and 14.3.

Let $\bar{p}$ and $\bar{p_i}$ be the arithmetical means of probabilities and the improbabilities defined by 10.42. Writing equation 14.22 as $\overline{h_c(Z_{med})} = 0$, relations

$$\overline{p(Z_{med})} = 1 - \overline{p(Z_{med})} \qquad \overline{p_i(Z_{med})} = 1 - \overline{p_i(Z_{med})} \tag{14.27}$$

are obtained, which prove statements 3) and 4).

Hence, the gnostic notion of the two medians is always closely connected with the specific value (1/2) of probability or improbability.

14.3.5 Gnostic Variance

Variance is a measure of the volatility of data. Its estimate can be obtained in statistics as the arithmetic mean of the squared deviations from the mean. A similar role, in a way, is played in gnostics by the difference $1 - f_c$, where $c \in i, j$, and where f_c is the G-weight (9.10). As shown in 9.21, this difference is a function of the squared data error (Φ^2). The arithmetic mean of this difference is thus one candidate for measuring the volatility or spread of the data. However, there are other functions of Φ^2 available in gnostics, both Q- and E-information (10.66), and the square of Q- and E-irrelevances (9.21). All these functions have their own important special meanings in gnostics. The difference $1 - f_c$ evaluates by 10.26 the change of entropy caused by the uncertainty, the Q- and E-information evaluates the change of information caused by uncertainty, and squared Q- and E-irrelevances determine the intensity of the normalized sources of the entropy field (shown by substituting $f_c^2 = 1 + c^2 h_c^2$ into 10.61). These functions can therefore characterize the volatility of data in a precisely defined way, each deserving of its proper interpretation. However, when referring to the *G-variance* (*Gvar*) of a data sample below, it will be understood to be the arithmetic mean of the squares of Q- or E-irrelevances. There is a sound reason to prefer this version of the variance: it is a special case of covariance, which enables the *gnostic correlation* (*Gcor*) of two samples $\mathcal{Z}_\mathcal{A}$ and $\mathcal{Z}_\mathcal{B}$ to be defined in a way similar to correlation in statistics:

$$var_c(\mathcal{Z}) := \overline{h_c^2} \quad (c \in \{j, i\}) \tag{14.28}$$

and

$$cor_{c,A,B} := \frac{ccov_{c,A,B}}{\sqrt{var_c(\mathcal{Z}_A) var_c(\mathcal{Z}_B)}} \tag{14.29}$$

Recalling 9.7 and 9.11, the dependence of the new characteristics on data can be shown: for an n-th datum Z_n, and ideal (true) value Z_0, the irrelevances are

$$h_{j,n} = \frac{q_n^2 - 1/q_n^2}{2} \qquad h_{i,n} = \frac{q_n^2 - 1/q_n^2}{q_n^2 + 1/q_n^2}, \tag{14.30}$$

where

$$q_n := \left(\frac{Z_n}{Z_0}\right)^{1/S_n} \qquad (n = 1, \ldots, N). \tag{14.31}$$

Then the crosscovariances $Q - ccov$ and $E - ccov$ are found by substituting 14.30 into 14.20 and the variances are obtained by using formulae

$$var_j = \overline{h_j^2} \qquad var_i = \overline{h_i^2}. \tag{14.32}$$

Both variances and covariances are dependent on the unknown ideal value Z_0. There are two ways of overcoming this difficulty: to

1. choose the value to serve to a particular purpose, or to
2. substitute an estimate $\tilde{Z}_0$ for Z_0 instead of the true value.

The former is especially useful in exploring the behavior of the variance (or covariance) as the value of Z_0 changes; the latter allows a particular value for some of these functions to be set. This is illustrated by corollary 15.5:

Corollary 15.5: Let $h_{c,n}(Z_0)$ $(c \in \{j, i\})$ be the G-irrelevances (14.30) and $f_{c,n}(Z_0)$ the corresponding G-weights.

Let $var_{c,Z}(Z_0) = \overline{h_{c,Z}(Z_0)^2}$ $(c \in \{j, i\})$ be the G-variance of data sample $\mathcal{Z}$.

Let $\tilde{Z}_{0,G}$ $(G \in \{Q, E\})$ be estimates of Z_0, such that

$$\overline{(h_{j,Z} f_{j,n})}_{Z_0 = \tilde{Z}_{0,Q}} = 0 \tag{14.33}$$

$$\overline{(h_{i,Z} f_{i,n}^2)}_{Z_0 = \tilde{Z}_{0,E}} = 0. \tag{14.34}$$

Then

1. the root $\tilde{Z}_{0,Q}$ of equation 14.33 minimizes the Q-variance of the sample,
2. the root $\tilde{Z}_{0,E}$ of equation 14.34 locally minimizes the E-variance of the sample.

Proof of Corollary 15.5: By differentiation relations

$$\frac{\mathrm{d}(h_{j,Z}^2)}{\mathrm{d}Z_0} = -\frac{4f_j h_j}{SZ_0} \tag{14.35}$$

and

$$\frac{\mathrm{d}^2(h_{j,Z}^2)}{\mathrm{d}(Z_0)^2} = \frac{8(f_j^2 + h_j^2)}{(SZ_0)^2} \tag{14.36}$$

hold. The first shows, that the necessary condition for extremization is satisfied and the latter says, that the extremum reached under the condition 14.33 is a minimum. The second derivative (14.36) is positive for all Z warranting the uniqueness of the Q-variance's minimum.

Analogously,

$$\frac{\mathrm{d}(h_{i,Z}^2)}{\mathrm{d}Z_0} = \frac{4f_i^2 h_j}{SZ_0} \tag{14.37}$$

and

$$\frac{\mathrm{d}^2(h_{i,Z}^2)}{\mathrm{d}(Z_0)^2} = \frac{8(f_i^4 - 2f_i^2 h_i^2)}{(SZ_0)^2}. \tag{14.38}$$

The root $\tilde{Z}_{0,E}$ of equation 14.34 actually extremizes the E-variance. However, the sign of the second derivative of the (averaged) identity 14.38 depends on how the terms $\overline{f_i^4}$ and $\overline{2f_i^2 h_i^2}$ are related. The first term is strictly positive, while the latter can change its sign. For a sufficiently small Z_0 all irrelevances $h_i(Z_0)$ approach 1, and for a sufficiently large Z_0 they reach -1. In both extremal cases, the E-variance is bounded by 1. However, non-extreme values of Z_0 decrease the variance's addends. There exists, therefore, at least one (local) minimum.

Additional location parameters of a data sample have thus been developed: estimates $\tilde{Z}_{0,G}$, which minimize G-variance. These estimates can be called *G-centers* (or *center*$_c$) of a data sample. The possible existence of several E-centers of a data sample motivates an analysis of the inner structure of some samples by observing the behavior of the E-variance, while the parameter Z_0 varies.

14.3.6 Similarity and Correlations

The main goal of the gnostic theory is to provide mathematical models to depict the uncertain quantitative images of reality. However, in order to be useful, models must embody the important characteristics of the object or processes, which they are to represent. Using such similarities permits objects or events to be categorized, classified and sometimes even evaluated. An example of such an idea might be taken from financial statement analysis and entail a judgement as to the financial health of a firm by comparing its financial parameters with those of "similar" firms. Practical importance of applications of this nature necessitates the development of a more detailed insight into the notion of similarity. In this manner the following important statements can be justified:

1. Similarity is inherently connected with geometry.
2. Covariance and correlation enables dissimilarity to be measured.
3. The classical (statistical) definitions of variance, covariance and correlation are based on Euclidean geometry.
4. The gnostic generalization of these same measures is based on Riemannian geometry. This enables the results to be robust.

Consider a set of points in a plane connected by straight lines to form a figure. Strictly similar figures have the same relations (proportions) for the lengths of corresponding lines. Lengths are distances measurable in accordance with the 'accepted' geometry. This holds even in a more general case, when the requirement 'approximate' (instead of strict) similarity applies: an analysis of similarity/dissimilarity leads to the measurement of lengths (sometimes angles), ie to apply geometry. Let $\mathcal{X}$ and $\mathcal{Y}$ be samples of data X_m and Y_m (where in both cases $m = 1, ..., M$). A strict similarity between the samples could be defined by the linear relation

$$y_m = C_0 + K_{y,x} * x_m \quad (m = 1, ..., M), \tag{14.39}$$

the coefficients of which can be easily found. In the case of real data contaminated by uncertainties the relation will hold only approximately and the determination of the coefficients requires, that the notion of approximation be defined explicitly. A 'natural' condition is the validity of 14.39 for the arithmetical mean values of the data (the *unbiasedness*), that leads to the elimination of the constant C_0,

$$y_m - \bar{y} = K_{y,x} * (x_m - \bar{x}) \quad (m = 1, ..., M). \tag{14.40}$$

These equations cannot hold exactly due to uncertainties. Another 'natural' requirement is minimization of the sum of the equations' errors, from which the 'best' estimate of

$$\tilde{K}_{y,x} = \frac{\overline{(y_m - \bar{y}) * (x_m - \bar{x})}}{\overline{(x_m - \bar{x})^2}} \tag{14.41}$$

or—using statistical notions—

$$\tilde{K}_{y,x} = \frac{\mathrm{covariance(y, x)}}{\mathrm{variance(x)}} \tag{14.42}$$

results.

However, it is easy to see, that both of these 'natural' conditions are based on Euclidean geometry:

Take the m-th error of 14.41 in the form of the difference $e_m = L_m - R_m$ between the left hand equation's side L_m and its right hand side R_m. The absolute value $|e_m|$ is the distance between two points in the uni-dimensional (Euclidean) space 1D. Denote e_m^+ and e_m^- the positive and negative error. The condition of unbiasedness then requires relation $\tilde{e}_m^+ + \tilde{e}_m^- = 0$ to hold. In other words, the sum of the length of positive errors should equal the sum of the length of the negative errors.

Now consider M-dimensional Euclidean space MD with a rectangular coordinate system. Attach to each value e_m the point on the corresponding (m-th) coordinate axis to create M mutually orthogonal vectors, the vector sum of which is E. The second 'natural' condition defining the 'best' estimate $\tilde{K}_{x,y}$ requires, that the (Euclidean) length of the vector E be minimized. Note, that the Euclidean scalar product as in 6.20 (Chapter 6) is applied here.

The thusly obtained estimate is well-known as the coefficient of the uni-variate regression, which 'explains' $\mathcal{Y}$ by $\mathcal{X}$ in the 'best' (unbiased least-squares) way. A popular application is in **Beta analysis** using the *Capital Asset Pricing Model* (CAPM, [83])[2]. This example can be used to emphasize an important aspect, which in logic is called *symmetry*; the relation measured by Beta is asymmetric: it cannot be expected, that reaction of the market return to changes in the return of a stock would be the same as that measured by Beta. On the other hand, *similarity* is a typical example of a symmetric relation: if object A is similar to object B then B is similar to A. Therefore, the coefficient $\tilde{K}_{x,y}$ of the regression taken in the opposite direction should equal $\tilde{K}_{y,x}$. However, this will occur only in the exceptional case, eg when $\mathcal{X} \equiv \mathcal{Y}$ (then "$\mathcal{X}$ is similar to itself"). A suitable unique dissimilarity measure satisfying the condition of symmetry can be obtained by taking

$$\mathrm{cor(x, y)} = \mathrm{sign}(\overline{(y_m - \bar{y}) * (x_m - \bar{x})}) * \sqrt{|K_{x,y} * K_{y,x}|}, \tag{14.43}$$

[2] The Beta of an investment measures the expected sensitivity of its return (y) to the return of the market (x). The measure, normalized by the variance of the market, also has the form of 14.42.

which coincides with the ordinary statistical definition of the correlation coefficient 14.42.

Let us follow these same steps, but now using the Riemannian metric form 6.13, that applied to measuring an element of the length of a path in the uni-dimensional space of uncertain data. This measurement reduces to 9.15, which after substitution of 9.3 and 9.4 leads to

$$dh_c = c^{-1} f_c(2c\Omega_c) d(2c\Omega_c) \quad (c \in \{j, i\}), \tag{14.44}$$

where $c := j = \sqrt{1}$ defines the quantifying and $c := i = \sqrt{-1}$ the estimating metric. Recall the geometric interpretation of (9.12), which shows that when, because of uncertainty, the value of the observed m-th data item is A_m instead of the true A_0, then the quantifying version of the angle Ω_c denoted Φ_m is

$$\Phi_m = \frac{A_m - A_0}{S}, \tag{14.45}$$

where S is the scale parameter dependent on the spread of the sample's data. The error, which is caused by uncertainty is thus not directly measured by the difference $A_m - A_0$, but by its nonlinear function, the irrelevance h_c, which has the form of 9.11. Equation 9.10 allows the role of the metric weight F_c in this measurement to be evaluated: it decreases/increases in the estimating/quantifying case as A_m moves away from A_0. This is the feature, which makes the measurement robust with respect to the outliers/inliers of the sample.

The similarity requirement (14.40) thus takes the form

$$h_c(y_m) = C_{y,x} * h_c(x_m) \quad (m = 1, ..., M). \tag{14.46}$$

The requirement for the optimality of this approximation can use several gnostic functions depending on the criterial function, which is chosen. The sum of squares of irrelevances can also be used, because it extremizes the gnostic variance (14.28). Moreover, the squared Q- and E-irrelevances determine the intensity of the sources of the entropy field (as can be shown by substituting $f_c^2 = 1 + c^2 h_c^2$ into 10.61). As demonstrated in Chapter 10, these sources are offset by the sources of the information field. This means, that the extremization of squares of irrelevances affects changes in both entropy and information. The errors of equations 14.46 can be computed as differences of both sides, ie additively, because irrelevances are to be combined additively as prescribed by Axiom 2. The minimization of the sum of squared errors is ensured by the analogy/generalization of

14.41

$$\tilde{C}_{y,x} = \frac{\overline{(h_c(y_m)) * (h_c(x_m))}}{\overline{(h_c(x_m))^2}}, \tag{14.47}$$

ie

$$\tilde{C}_{y,x} = \frac{Gccov(y,x)}{Gcov(x)}. \tag{14.48}$$

When the nature of the object/process described by the data allows the similarity of the relation $\mathcal{Y}(\mathcal{X})$ to be assumed, the gnostic correlation coefficient is

$$Gcor(x,y) = \frac{Gccov(x,y)}{\sqrt{(Gcov(x) * Gcov(y)}} \tag{14.49}$$

in accordance with the equation 14.29.

These relatively simple remarks on the nature of correlation lead to the following conclusions:

1. Covariance and correlation can be interpreted as measures of the similarity of data samples based on the
   (a) choice of geometry defining the measurement errors and the optimality criterion,
   (b) characterization of similarity by a linear relation,
   (c) assumption of symmetry in the similarity relation for the case of correlation.
2. Gnostic notions of covariance and correlation are consistent generalizations of the classical statistical definitions.
3. The Riemannian character of the geometric background of these notions results in their robustness, the type of which can be chosen (inner or outer).
4. These notions have optimal features, because they are based on irrelevances, the optimality of which results from the optimality of the gnostic cycle as proved in Chapter 12.
5. The similarity of data irrelevances is linked directly to the similarity of data probabilities and improbabilities.

The last statement is justified by recalling relations 10.42: estimating irrelevances are simple linear functions of probabilities. The same relation holds between quantifying irrelevances and improbabilities. These relations are important, because they justify using regression models of probabilities or improbabilities (See Chapter 17).

14.3.7 Standard Equivalent of a Data Sample

The data sample's equivalent was defined in definition 13 by using mean values of G-irrelevances and G-weights. The idea was to represent the mean of pair numbers of a data sample by a single pair number. As shown in Corollary 15.4, the angular argument of this equivalent is zero at the point $Z_0 = Z_{med}$. Another equivalent characterizing the variance of the sample can be introduced. The statistical notion of the *standard deviation* is based on deviations from the mean (error), the mean square of which is the variance. An "analogous" definition of the *gnostic standard deviation* ($Z_{c,std}$) is the square root of the mean squared irrelevance equaling the minimal G-variance of the data. Corollary 15.6 expresses this characteristic and the formulae for its calculation.

Corollary 15.6: Let $c \in \{j, i\}$. Let $\mathcal{Z}$ be a data sample composed of data $Z_1, ..., Z_N$. Let $Z_{c,cen}$ be G-center of the sample, which minimizes its G-variance. Let $Z_{c,std}$ be the gnostic standard deviation, such that relation

$$\left(h_c(\frac{Z_{c,std}}{Z_{c,cen}})\right)^2 = var_{c,Z} \tag{14.50}$$

holds.

Then

$$Z_{j,std} = Z_{j,cen} * (\sqrt{var_{j,Z}} + \sqrt{1 + var_{j,Z}})^{S/2} \tag{14.51}$$

and

$$Z_{i,std} = Z_{i,cen} * \left(\frac{1 \pm \sqrt{var_{i,Z}}}{1 \mp \sqrt{var_{i,Z}}}\right)^{S/2} . \tag{14.52}$$

Proof: By substituting 14.30 into 14.50 and writing $Z_{j,std}$ (or $Z_{i,std}$) instead of $Z_{j,n}$ and $Z_{j,cen}$ (or $Z_{i,cen}$) instead of Z_0.

Before $Z_{i,cen}$ is used, it must be checked for its uniqueness.

14.4 Important Features of Gnostic Characteristics

14.4.1 Limit Values for Precise Data

It was shown in previous sections, that the data sample's modulus gives rise to several new properties such as $\overline{f_c}$, $\overline{h_c}$, $\overline{h_c^2}$ and $\text{cov}_c(N, K)$. Other useful features can be derived from data, when gnostic procedures are used. In Chapter 9, the notion of sufficiently precise data was introduced to identify data with a small enough uncertain component, that the higher members of the power series expansion of individual data characteristics could be neglected. Now the same concept can be applied to the gnostic characteristics of data samples. (See section 13.2.2 for the gnostic definition of a sample).

Let the scale parameter of all data be constant and be equal to S. Define

$$\Delta := \max_k(|S\Phi_k|) \quad (k = 1, ..., N) \tag{14.53}$$

to characterize the bound of the uncertainties of a data sample. Expanding the formulae of basic gnostic characteristics into a power series and using Landau's symbol $O(*)$ to represent infinitely small quantities, one can obtain the following approximations, which are valid for sufficiently precise data:

$$\overline{f_c} = 1 + 2c^2\overline{(S\Phi)^2} + O(\Delta^4), \tag{14.54}$$

$$\overline{h_c} = 2\overline{S\Phi} + O(\Delta^3), \tag{14.55}$$

$$\overline{E_c} = 2c^2\overline{(S\Phi)^2} + O(\Delta^4), \tag{14.56}$$

$$\overline{I_c} = 2c^2\overline{(S\Phi)^2} + O(\Delta^4), \tag{14.57}$$

$$\overline{h_c^2} = 4c^2\overline{(S\Phi)^2} + O(\Delta^4), \tag{14.58}$$

$$M_{Z,c} = 1 + 2c^2(\overline{(S\Phi)^2} - \overline{(S\Phi)}^2) + O(\Delta^4), \tag{14.59}$$

and

$$\mathrm{cov}_c(N, K) = \frac{1}{N-K}\sum_{l=1}^{N-K} S^2\Phi_l\Phi_{l+K} + O(\Delta^4). \tag{14.60}$$

In *this specific case*, there then exists a close relationship between basic gnostic characteristics and between the first and second statistical moments of errors. Under these same conditions, the gnostic covariance converges to that of classical statistics. However, it must be emphasized, that *no correspondence exists*, when there is a sizable data uncertainty. The salient feature of the gnostic characteristics of strongly dispersed data is their sensitivity/robustness with respect to outlying data. Let us consider this problem in more detail.

14.4.2 Sensitivity/robustness of Gnostic Characteristics

Chapter 11 examined the robust characteristics of individual data. These same ideas will now be explored for the broader spectrum of characteristics of data samples (see 13.2.2, Definition 12).

Definition 15: Let $\chi_N(\mathcal{Z}(L))$ be a gnostic characteristic of the data sample $\mathcal{Z}$ satisfying Definition 13 (13.11). This characteristic will be called *additive*, if for an arbitrary data sample $\mathcal{Z}$ composed of data $Z_1, ..., Z_N$ and for a data sample $\mathcal{Z}'$ composed of data $Z_1, ..., Z_N, Z$, the difference $\chi_{N+1}(\mathcal{Z}') - \chi_N(\mathcal{Z})$ depends only on the datum Z, its ideal value Z_0 and its scale parameter S.

The largest positive or the smallest negative real number γ, for which the inequality

$$0 < lim_{Z\to\infty}(|\chi_{N+1} - \chi_N|)Z^{\gamma} < \infty \tag{14.61}$$

holds for an arbitrary pair of data samples $\mathcal{Z}'$ and $\mathcal{Z}$, will be called the ***degree of robustness*** of the gnostic characteristic χ. The negative of the value of γ is the ***sensitivity*** of the gnostic characteristic.

Gnostic theory needs both additive (eg $\overline{f_c}$, $\overline{h_c}$, $\overline{f_c^2}$, $\overline{h_c^2}$, ...) and non-additive characteristics (eg $M_{Z,c}$ and cov_c). Using the notion of the degree of robustness/sensitivity, Tab. 14.1 summarizes the classification of the gnostic characteristics of data samples.

χ	$\overline{f_j^2}$	$\overline{h_j^2}$	$\overline{f_j}$	$\overline{h_j}$	cov_j	cov_i	$\overline{h_i}$	$\overline{h_i^2}$	$\overline{f_i}$	$\overline{f_i^2}$
γ	$-4/S$	$-4/S$	$-2/S$	$-2/S$	$-2/S$	0	0	0	$2/S$	$4/S$

Tab. 14.1 Sensitivity and robustness of several gnostic characteristics of a data sample

To verify the values of the degree of robustness (γ) in Tab. 14.1, $q := (Z/Z_0)^{1/S}$ (9.7) is substituted (Z_0 and S are the ideal value and the scale parameter of the datum Z). The results from formulae 9.10, 9.11 and 14.60 then show, that the differences $\chi_{N+1} - \chi_N$ have following forms:

- $((q^2 + q^{-2})/2)^{c^2}/N$ (for the mean of weights),
- $(q^2 - q^{-2})/(2N)$ (for the mean of Q-irrelevances),
- $((q^2 - q^{-2})/(q^2 + q^{-2}))/N$ (for the mean of E-irrelevances),
- $((N-K)/(N+1-K)-1)\mathrm{cov}_c(N,K) + h_{c,N+1-K}h_{c,N-K}/(N+1-K)$ (for covariances).

The values of γ can then be determined by substitution of these relations into the inequality 14.61.

Definition 15 introduces the notions of sensitivity and robustness from the standpoint of the response of a specific characteristic to the outlying data. Such a characteristic, which is sensitive to outliers is robust with respect to inlying data. The converse is also true. To prevent a misleading classification, we prefer to speak of robustness with respect to outliers and robustness with respect to inliers before using the notion of sensitivity.

The question may arise as to whether there is any real utility in having both kinds of robustness. The answer is decisively positive. If the return on equity for a sample of firms was being examined, for instance, then either extreme positive or extreme negative returns would be interpreted as "abnormal," distorting the picture of the dominant part consisting of the "normal" enterprises. These "peripheral" data (outliers) represent "noise" in the observations. This is analogous to some repeated physical or technical measurements, where the (rare) peripheral results are caused by gross measuring errors. An opposite situation occurs, when the objective of the analysis is to uncover the dynamics of a market's turning point, the goal of which is to discover the start of a sudden process of rising or falling share prices. In this case the "noise" is the "normal" volatility of the prices (inlying data) and the required "signal" is represented by (rare) peripheral data (outliers), which reach beyond the boundary of the frequent "normal" price movements. This becomes the signal to take some sell/buy action.

14.5 Summary

The sample's modulus is introduced to normalize both the sample's weight and irrelevance. This allows a single gnostic event to represent the whole of the sample. This is only a limited representation, similar to the case of a relativistic particle representing a group of particles in the sense of its having the momentum and energy of all of the particles. Many aspects of the individual data in the sample are not reflected by the sample's equivalent, the existence of which is warranted only when a sample is homogeneous. This draw-back in the utility of the data sample's equivalent (obtained by means of the modulus) is more than counterbalanced by the possibility of making use of this feature for testing the sample's homogeneity.

The data sample's modulus can be used to demonstrate the role of useful additive characteristics such as the arithmetic mean of weights as well as irrelevances and their squares and to introduce a new non-additive characteristic, the gnostic covariance. It was shown, that in the case of sufficiently precise data, there is a simple relation between the basic gnostic characteristics and the statistical first and second moments. In the case of gross errors, however, the behavior of gnostic characteristics is substantially different from those of statistical origin. This difference stems from the nonlinearity of the gnostic characteristics with respect to data and their squares. The result is a desirable robustness in the characteristics with respect to both outliers or inliers. These two mutually complementary kinds of robustness allows suitable characteristics for each given task to be selected. Robustness is a natural product of the theory, an inherent feature of gnostic characteristics of uncertainty, and not something “imported” from the outside to satisfy some additional requirements. The degree of robustness of the gnostic characteristics can also be chosen to suit the requirements of the problem to be solved.

Chapter 15

Distribution Functions of a Data Sample

15.1 Goodness of Fit

15.1.1 The Problem

The problem under consideration is to design a smooth *distribution function*, which characterizes the pattern formed by different values of a given data set (sample). In the previous description of the Parzen's and gnostic kernels (Chapter 11), it was made clear, that the objective of the discussion concerned distributions of probability and/or probability density. The outcome of the composition of gnostic kernels provides results, which are interpreted as probability distributions and/or as distribution functions of probability density (shortly *density*) of the data sample. It is to be recalled, that the notion of probability in gnostics differs from the statistical definition. As it has been pointed out, in gnostics, probability can be interpreted as the expectation based on the data in the sample.

The model should be a continuous distribution function of the actual data distribution. It can be obtained in three steps by choosing:

1. A *discrete distribution function* (DDF) as a set of primary estimates of the data's probability defined by the data values.
2. a family of smooth distribution functions suitable to model the DDF. (Only gnostic distribution functions GDF will be considered in this connection).
3. A *criterion function*, the extremity of which ensures the best *goodness-of-fit* of the specific GDF to the DDF.

Algorithms used to estimate the distribution functions can combine different choices of these elements. It is useful to examine the steps in a more detail.

The choice of the modeling function will be considered in detail below. The elements of the distribution functions to be constructed—the gnostic kernels, the unique form of which have been derived theoretically—are already known. It "only" remains to establish the composition rules to be applied to the kernels to satisfy the gnostic composition axioms. This question has a simple answer in the case of Parzen's method: the kernel estimate of a probability density function is obtained as the arithmetic average of kernels associated with all the data.

To obtain the discrete distribution function, DDF, a rule determining the probability to be assigned to each element of the data sample must be established. Three versions of the DDF will be considered:

1. The Empirical Distribution Function (EDF),
2. the Kolmogorov-Smirnov DDF (KSDDF),
3. the Maximum Entropy DDF (MEDDF).

An initial impression, that the problem of the best fit becomes a trivial one once the DDF and its model has been determined (since there are well-known solutions to the curve fitting problem such as least-squares, chi-square, etc.) would be incorrect, because each of these methods assumes a specific distribution for the fitting errors (most frequently the normal distribution). It should be obvious, that when the objective is to find an *unknown* distribution model, a fitting procedure based on having prior knowledge of the distribution cannot be used. Instead, fitting procedures based on more general principles are to be applied.

15.1.2 The Empirical Distribution Function

When treating sufficiently large data samples, one can approximate their density distribution by constructing a *histogram*. The data range is split into a series of intervals and the number of data falling within the range of each interval (frequencies) are counted and transformed into a bar graph, the horizontal axis of which delineates boundaries of the data value classes (intervals), while the vertical scale describes the frequencies of events for all the classes. A *frequency polygon* can be drawn over the set of average points representing the classes by connecting these points with straight lines. The *cumulative frequency polygon* or *ogive* is then obtained by connecting the cumulative frequency points by straight lines. These polygons can be used respectively as approximations of the data probability density and the probability distribution function. To estimate the frequency, there must be a sufficient number of data in each class, the more there are, the lower the estimation error. A well-known (Sturges') recommendation ([110]) as to the necessary number of classes is $1 + (\log(N)/\log(2))$, where N is the total number of scores (size of the data sample). It is obvious, that this method cannot be used for small data samples. Instead, methods for providing the probability estimates for each element of the sample's data are needed.

In statistics, the empirical distribution function (EDF) is defined for a random variable X as a function $F(x) = Pr\{X \leq x\}$[1], which assigns a measure (between 0 and 1) to data $x_1, x_2, ..., x_N$, so that $F(x_n)$ is the proportion of observed values not exceeding x. Formally,

$$F(x_n) = Pr\{X \leq x_n\} = \frac{n}{N} \qquad (n = 1, ..., N). \tag{15.1}$$

This relation defines the EDF for N points, but it is usually shown as a step function similar to an irregular staircase[2]. The lower level of the bottom step is zero, and its top is $1/N$. The level at the top of the last step is 1. It is proved in statistics, that under some generally acceptable assumptions, this EDF converges to the actual probability distribution function of the data population, when the sample size (N) tends to infinity.

The EDF is thus useful in statistics in the sense of asymptotic behavior. But this does not automatically mean, that it is acceptable for use with finite sample sizes. To demonstrate, recall that according to the axiomatic features of probability, the following identity holds for all x:

$$Pr\{X \leq x\} = 1 - Pr\{X > x\}. \tag{15.2}$$

Let the lower bound of the data support be LB and the upper bound UB. Then 15.1 gives $F_1 = Pr\{LB < X \leq x_1\} = 1/N$ and $F_N = Pr\{X \leq x_N\} = 1$. Using 15.2, the result is $Pr\{x_N < X < UB\} = 0$. The probability, that some X could lie "below the bottom step" (to not exceed x_1) is thus non-zero, while the probability of exceeding x_N ("over the top step") is zero. This asymmetry is of no importance in the asymptotic case, because the step's height tends to zero in this case. However, for a limited data sample of size N, this asymmetry leads to a contradiction. Indeed, the roles of probability and of its complement defined by 15.2 are completely

[1] The symbol $Pr\{X \leq x\}$ is to be read as "the probability that the variable X is less or equal to x".

[2] The irregularity of a step's height results from repeated data values.

symmetric. The choice between estimating either $F(x)$ or its complement is the matter of a subjective decision and both should lead to identical results. But this contradicts the fact shown above, that the first decision could assign a non-zero measure to one of the potentially infinite intervals, and a zero value to the other, while the latter choice assigns measures in the opposite manner.

In spite of the broad usage of the EDF in statistics, eg to depict an "ideal" distribution for a finite sample, the application of this step function to arbitrarily evaluate the goodness-of-fit cannot be recommended.

15.1.3 The Kolmogorov-Smirnov Points

The Kolmogorov-Smirnov test of goodness-of-fit represents an indirect application of the empirical distribution function. It is based on the *Kolmogorov-Smirnov statistic*, which is defined as the maximum absolute difference between values of two given distribution functions defined over the same data support. For a one-sample situation, the hypothesis to be tested is that the sample was taken randomly from a population, which has a known distribution function. This is not useful for gnostics, because no a priori data distribution is assumed. The two-sample Kolmogorov-Smirnov (KS) test is more relevant for our purposes, especially when one of the two distributions is the EDF. The KS statistic is then evaluated as the maximum absolute difference between the value of the distribution function being tested and the step's level at each of the data points. The two-sample KS tests are based on the following rather general assumptions:

1. the samples are random samples,
2. the two samples are mutually independent,
3. the data are measured on at least an ordinal scale,
4. the underlying distributions are continuous.

It should be emphasized, that there is no assumption made as to a particular form for either of the distribution functions. The application of the KS two-sample test includes the following steps:

a) determination of the absolute difference between the values of the two distribution functions at a given set of points (often at the data/points of the sample),

b) finding the maximum value of the absolute differences,

c) comparison of the maximum value with the critical value of the KS statistic (given by the sample size and the desired significance of the test),

d) decision on whether to reject the hypothesis of the goodness-of-fit.

Imagine a distribution model defined as a function (of a known analytical form) of not only the observed data, but also of an (unknown) parameter (eg the scale parameter) or even of several unknown parameters (eg scale parameter as well as the lower and upper bounds of a data support). In the sense of this test, the best possible choice for each of the unknown parameters (to ensure the best goodness-of-fit) is the one minimizing the KS statistic. In the case of a perfectly flexible distribution model[3], the model values at the data points would cross the points of the EDF's steps at exactly the center of the step's vertical edges. These "ideal" points will be called the *KS-points* and their height can be determined by the formula[4]

$$Pr_{n,KS} = \frac{2n-1}{2N} \quad (n = 1, ..., N). \tag{15.3}$$

The collection of these "ideal" points is symmetric, because the probabilities assigned to the intervals (LB, x_1) and (x_N, UB) are the same and are equal to $1/(2N)$. This is logical, because if there is no information as to the size of either interval, they can be assigned the same probability. The height of each of the steps is constant $(1/N)$. Again: there is no information on expectations, which would suggest, that different probabilities should be assigned.

The distribution of the KS-points is in consonance with the concept of Parzen's kernel estimation: consider a symmetric kernel, the maximum of which is located over the data point (say, x_n). Let the positive kernel's values exist only over the interval $[x_n - \Delta/2, x_n + \Delta/2]$ and let them be zero outside of this interval. The kernel is normalized so that its integral equals 1. Let us assume, that all data in the sample are different, and that the kernel's width Δ is small enough to ensure no overlapping of the positive parts of the kernels. The kernel estimate of the density function in this case is a collection of N separate kernels shifted along the data axis in correspondence with the location of the data. The weight of each kernel is $1/N$, because of the accepted additive composition law. The estimated probability distribution is given by integration of this density distribution

[3] One, in which the curvature can change depending on the value chosen for the scale parameter. The notion of flexibility relates to the ability to "bend" the DF by decreasing S so as to make the local radius of curvature unlimitedly small.

[4] A simplified case of data is considered here: the data values are not repeated and the a priori weights of all the data are equal. For a more general case see below.

and the estimated probabilities are then obtained by:

$$\tilde{Pr}(LB < X \leq x_1) = \tilde{Pr}(x_N < X < UB) = \frac{1}{2N} \tag{15.4}$$

and

$$\tilde{Pr}(x_{n-1} < X \leq x_n) = \frac{1}{N} \quad (n = 2, ..., N) \tag{15.5}$$

ie the same system of probabilities, which was formed by the KS-points $Pr_{n,KS}$ for $n = 1, ..., N$ (15.3).

15.1.4 The Maximum Entropy Goodness of Fit

Consider a k-th individual datum with E-irrelevance $h_{i,k}$ (where $i = \sqrt{-1}$) as in 9.6. As shown in the previous chapters, a value, p_k, (10.42) may be assigned to this datum and it can play three important roles:

1. Its value is the (gnostic) probability (expectation) of the observed datum's ideal value.
2. It is the function of an (unknown) ideal datum's value A_0 (or Z_0): the probability distribution function of A_0 or Z_0 (given observed value A_k or Z_k) and/or—after differentiation—the density distribution function (or a kernel for kernel estimation).
3. It is the parameter of the function $H(p_k)$ (10.50), which can be used for evaluation of E-information.

Let us assume for the while, that the sample's data are ordered, so that relations $z_1 \leq z_2 \leq ... \leq z_{N-1} \leq z_N$ hold. (If these relations are not satisfied, then the data are to be ordered and renumbered). The bounds of data support (z_L and z_U) always exist, but whether these values are known is not important at this juncture. Probabilities are also ordered as non- decreasing functions of the data. They define $N + 1$ non-negative differences $P_k = p(z_k) - p(z_{k-1})$ $(k = 1, ..., N + 1)$, where $p(z_0) = p(z_L) = 0$ and $p(z_{N+1}) = p(z_U) = 1$. Consider the expression

$$RE = \frac{-\sum_{k=1}^{k=N+1} P_k * \ln(P_k)}{\ln(N+1)}, \tag{15.6}$$

which will be called the *residual entropy* of a data sample. To justify this name, let us analyze four important features of the function:

1. Distribution: The sum of arguments $\sum_{k=1}^{N+1} P_k$ equals 1. This means, that the probabilities $p(z_k)$ $(k = 1, ..., N)$ together with the bounds

$p(z_L)$ and $p(z_U)$ define a discrete distribution of $N+1$ subintervals covering the whole interval of probability from 0 through 1.

2. Collapsed data support ($z_L = z_U = z_k$ for all k): All P_k are zero and so is RE, because $\lim_{p \to 0} p \log(p) = 0$. This is the case of precise data. There is no uncertainty in these data.
3. Convexity: It is well-known, that the numerator of RE is a convex function of its arguments P_k.
4. Maximum: It is also well-known from statistical physics, that the numerator of RE (formally identical with Boltzmann's statistical entropy of a dynamic system) reaches a maximum (equal to $-\log(N+1)$) if and only if all probabilities P_k have the same value and if their sum is 1. Data are uniformly distributed in this case, there is no data cluster, no "preference" to any part of the sample. The residual entropy RE reaches a maximum of 1.

An important note as to the nature of the numerator of expression 15.6 can be added by recalling the way it was derived: it represents the amount of information, which was provided by the estimation phase of the Ideal Gnostic Cycle. The greater this information, the greater the portion of the entropy increase offset by the estimation. The result is, that the system of N probabilities

$$P_k = \frac{k}{N+1} \quad (k = 1, ..., N) \tag{15.7}$$

forms a remarkable discrete probability distribution assigned to an ordered data sample $\langle z_1, ..., z_N \rangle$. Points calculated according to 15.7 will be called the ***ME-points*** (points of maximum residual entropy). Compare this system of ME-points with that of the KS-points considered above: a legitimate reason for taking the first and last value of the KS-points (which equal only half of the "inner" values) would be to have a priori knowledge of the data's behavior, information existing "outside" of the data sample. Example: We know, that the interval between z_L and z_1 is shorter than the other intervals, and that it therefore "deserves" only half of the measure of the other intervals. (But one can ask, why just a half, and why always a half, if such special—obviously not general—information is available). In contrast, the system of ME-points assumes no additional information on the probability distribution, only knowledge of the data values themselves.

The form of the discrete distribution of the ME-points 15.7 is especially suitable, when all the data in the sample are different. The a priori weight of each datum is the same and equals $W_k = \frac{1}{N+1}$. If there are repetitive data points, it is more practical to go over to a "compressed" sample, where all the data will be different, but with weights correspondingly multiplied. Summing the "accumulated" weights W_k creates a system of ME-points.

15.1.5 The Weighted Empirical Distribution Function

All three approaches to the construction of a discrete distribution function DDF defined directly by the data considered above (EDF,KS- and ME-) were based on the a priori assumption, that all data have equal importance: points on the EDF as well as the KS points assumed a priori weight $1/N$, while in the case of the ME-approach the weights were taken as $1/(N+1)$ for each element of the data sample. Such an assumption represents a limitation, which cannot be generally accepted. In practice, one cannot exclude cases of repeated data values. A simple example consists of data given as integers (eg counts of events) or as real numbers with limited precision. Under these conditions, the probability that a data sample will include two or more equal data values is not negligible. Another example of a data sample with repeated data values is for data provided in the form of a histogram. In this case there is not only a data vector, but also a vector of "a priori data weights", which represents the number of times each data value was observed.

It is important to distinguish between two kinds of data weights:

1. the prior weight,
2. the posterior weight, ie the G-weight defined by 9.5.

A priori in this case means "known before data analysis is begun" or "based on information available and obtained along with the data." Indeed, the information, that some of the data have the same value is available at the same moment as the data. In the case of "histogram data" the repetition is already made explicit by a weighting vector; for repeated observations or measurements, such a weighting vector may be obtained by simple calculations, which precede the data analysis. The fact, that some data values may appear more than once provides information about the data, which must not be neglected in the analysis. There are examples supporting this position: a simple case is that of the arithmetic mean of a finite number of data; it is strongly influenced by repeated data values. Among many other examples is the principle of making decisions based on a majority vote. The results of voting can be then given the weights proportional to number of votes. Another example is the treatment of data measured by methods, which differ in their accuracy.

In contrast to using a weighting scheme established before the analysis, the gnostic weight of each individual datum **results** from the gnostic analysis. It is thus a typical *posterior* weight available only after the data analysis has been concluded.

It can now be shown, that if a priori weights exist, then they can also be used to improve the results of the gnostic analysis. Assume, that a sample of multiplicative data Z_k and a collection of corresponding a priori data weights W_k, where $k = 1, ..., N$ is given. Each of the variants of the gnostic distribution functions considered in the following discussion can be adapted to use the data weights, W, that have been provided. What is being sought here is the discrete distribution function, which would make use of the a priori weights. It can be approximated by using the gnostic distribution function and a criterion function derived by testing the goodness-of-fit.

The form of this function, denoted $E_{W,k}$ is chosen so as to satisfy the following requirements. It:

- is applicable to an arbitrary distribution of weights W,
- respects the given relations between weights,
- is consistent with the system of ME-points (the maximum entropy ideal distribution) in the special case of all weights equal,
- is applicable to both finite and infinite data supports,
- has symmetrical behavior independent of the choice of either E_W or its complement, $1 - E_W$.

Consider the following version of this function using normalized values w_k of the a priori weights W_k:

$$w_k = \frac{W_k}{\sum_{n=1}^{N} W_n} \quad (k = 1, ..., N) \tag{15.8}$$

$$E_{W,1} = w_1/2 \tag{15.9}$$

$$E_{W,k} = E_{W,k-1} + (w_{k-1} + w_k)/2 \quad (k = 2, ..., N) \tag{15.10}$$

These relations imply, that

$$E_{W,N} = 1 - w_N/2. \tag{15.11}$$

Consider either a finite or infinite data support, an open interval $I_{L,U} := (Z_L, Z_U)$ split by the data into $N + 1$ semiclosed subintervals $I_{k-1,k} = (Z(k-1), Z(k)]$, where $k = 1, ..., N + 1$, $Z(0) = Z_L$ and $Z(N + 1) = Z_U$. The probability is again the measure of each interval's length. The value of the probability distribution function $E_{W,L}$ in a (zero or positive) point Z_L

is zero, and in the (finite or infinite) point Z_U, it is 1. The first value $E_{W,1}$ is a probabilistic measure of the first interval $I_{0,1}$ equaling the normalized weight w_1 (15.8). Similarly, the last value $E_{W,N}$ has the measure $1 - w_N$. Both of these values are related to the idea, that the measures are integrals over the intervals. The measure of both intervals is independent of the direction of the integration (from left to right or from right to left), which corresponds to the choice between the E_W and its complement $1-E_W$. Such a symmetry is ensured for the differences $E_{W,k} - E_{W,k-1}$, when they satisfy 15.10. The normalization of weights (15.8) is properly chosen, because the sum of the measures of all subintervals equals 1. This normalization does not change the proportions of the weights. It is easy to verify, that if all weights are equal, the points on the E_W coincide with the ME-points.

The discrete distribution functions E designated the *weighted empirical distribution functions* will be called WEDF.

15.1.6 Criterion Functions

Once a DDF is selected and a family of the GDF is chosen, the specific form of the gnostic distribution function can be optimized to reach the best possible godness-of-fit. To do this, a criterion function of the fitting errors is to be extremized by finding the best estimate of GDF's parameters[5]. Gnostic theory makes available several functions reasonably applicable to optimum estimation. They all are defined by the already introduced a posteriori weights and irrelevances of data.

The G-weight (9.10) and G-irrelevance (9.11), in the estimating case ($c^2 = -1$), have the form

$$f_{E,k} = \frac{2}{q_k^2 + 1/q_k^2} \qquad h_{E,k} = \frac{q_k^2 - 1/q_k^2}{q_k^2 + 1/q_k^2}, \tag{15.12}$$

where

$$q_k = (Z_k/Z_0)^{(1/S)} \tag{15.13}$$

is an auxiliary variable in similar form to that of 9.7, where Z_k is the k-th observed data value and Z_0 is the (unknown) ideal data value. To measure fitting errors, the role of Z_k takes over the probability estimated by the GDF in the $k-$th data point Z_k, while the ideal value Z_0 is determined as

[5] As shown in Chapter 16, there is no contradiction in the notion of 'parameters of non-parametric estimates of distribution functions.

$DDR(Z_k)$. Instead of 15.13, expression

$$q_k = \left(\frac{GDF(Z_k)}{DDR(Z_k)}\right)^{1/S} \tag{15.14}$$

is used.

The variable f_E (the estimating posterior weight of a datum) will be called *the fidelity*, because it measures the weight of the relationship between two numbers: the actual value and its required value. Both fidelity and irrelevance are robust with respect to outliers as has already been mentioned.

A formal note is in order here with respect to symbology. The bar or overline, ($\bar{X} = \frac{\sum_{k=1}^{N} X_k}{N}$) is a commonly accepted way of denoting the arithmetical mean of several values of a variable (eg X_k). When gnostic distribution functions are defined or used, a more general averaging method is required and the mean is weighted not only by constant weights, but also—in the case of different a priori weights—by normalized a priori weights w_k (15.8). Therefore the notation using a bar or overline will be retained in such instances as the estimating version of 13.13,

$$\overline{h_E} = \sum_{k=1}^{N} w_k * h_{E,k},, \tag{15.15}$$

which is applicable for both equal and different a priori weights. The index '$_E$' denotes "estimating" and it is used instead of the complex unit $i = \sqrt{(}-1)$, which might lead to misinterpretation.

Everything has now been prepared to introduce gnostic functions suitable as criterion functions for the goodness-of-fit:

$$CF(f_E) := \overline{f_E} \tag{15.16}$$

$$CF(h_E^2) := \overline{h_E^2} \tag{15.17}$$

$$CF(I) := -\overline{p\ln(p)}, \tag{15.18}$$

where $p_k = (1 - h_{E,k})/2$.

The additive composition applied in these formulae is justified by the theory: From Axiom 2, it is recalled, that the fidelities (data weights, entropies) are to be composed additively. The additive composition in 15.17 is justified by the fact, that the minimum of $CF(h_E^2)$ is reached, when $\bar{(}h_E) = 0$, ie when the weighted mean error vanishes. In this condition irrelevances are added in accordance to Axiom 2. The expression 15.18 was derived by application of the linear operation of integration.

The value of the scale parameter S in 15.14 deserves further comment. The measurement of uncertainty by gnostic weights and irrelevances (and by their functions such as entropy, information, probability etc.) is equivalent to the application of a Riemannian metric. The main characteristic of a particular metric is its curvature. The curvature in all gnostic formulae is determined by the scale parameter. As will be shown, the scale parameter is always estimated from the data and not assumed a priori. This is the point of the statement “Let data speak for themselves”: data determine the scale parameter and the scale parameter determines the curvature of the geometry applied to measure uncertainty in a particular set of data. The curvature in turn determines the degree of robustness of the gnostic characteristics. An illustration is given by formulae 15.12 and 15.13, which show how the value of S controls the form of the fidelity, the rate, at which it decreases with the increasing size of the interval between the observed Z_k and the ideal value Z_0. However, these formulae signal a problem: the limiting value of the fidelity for $S \to \infty$ equals 1 independently of the ratio Z_k/Z_0. In other words, a trivial “optimum” fit exists (independent of the data) by using a sufficiently large S, therefore the structure of specific algorithms must take this into account and provide a safeguard against this possibility.

The suitability of gnostic functions to goodness-of-fit problem does not exclude applications of other criteria. All three functions 15.16 through 15.18 are robust with respect to outliers (see Tab. 14.1). This robustness can be useful, when individual fitting errors reach extreme values. In some applications this robustness can cause discrepancies between model and discrete function, which represents data. Another criterion functions can provide an unrobust fit. Examples:

- Criterion function obtained by summation of weighted absolute fitting errors.
- Criterion function minimizing the Kolmogorov-Smirnov statistic.
- Application of weighted squares of fitting errors.

The problem of criterion functions will be examined further in Chapter 17 in connection with the optimality of multidimensional models.

15.2 Four Versions of the Gnostic Distributions

15.2.1 Introduction

According to the statistical (Parzen's) approach to kernel estimation, the question "How to compose kernels" has a seemingly trivial answer: "Take the arithmetical mean." This manner of composition does not present a problem in statistics, because it corresponds to the axiom of the additivity of the probabilistic measure. The initial motivation to proceed in this manner, as discussed above, probably originated from the possibility of formally mapping basic statistical variables onto variables of Newtonian mechanics along with the hidden assumption of the implied applicability of Euclidean geometry. Moreover, additive composition ensures the asymptotic behavior required by statisticians.

For gnostics, in the general case, additive composition of kernels cannot be taken for self-evident, because the notion of probability does not exist as a fundamental axiom. Probability, determined as a secondary product of the theory, cannot be manipulated arbitrarily, but only in accordance with the axioms. From this point of view, the irrelevance has a more fundamental meaning than probability, because Axiom 2 prescribes additive composition for irrelevances, but not directly for probability. An initial impression, that these two ideas are the same, because the integral of the probability kernel (10.42) is a linear function of the irrelevance, would be misleading:

1. There are **two** ways to compose irrelevances based on Axiom 2. The first is the arithmetical mean (13.13), but the second, 14.14, is neither additive nor linear. As previously mentioned, the composition law applied to the former does not imply homogeneity of the sample, while the latter does.
2. It was pointed out in Chapter 10, that in using the concept of double numbers one should introduce probability not only as a real number as in 10.42, but also as a double number (10.60). The real version of probability does not contradict the gnostic theory, because $p*(1-p)$ is equal to $p_j*(1-p_j)$ in the key expressions 10.43 and 10.44. It is customary to evaluate probability using a real, not a double number, and gnostics accepts the use of real probability for calculations, too. However, the double number version is better suited to the interpretation of the improbability p_i, which is complex.

The point is, that when gnostic kernels are aggregated, primary attention should be directed to the aggregation of irrelevances. There are two versions of irrelevances, and when addressing gnostic distribution functions, it will be necessary to distinguish between the two ***types:*** Q- (*quantifying*) and E- (*estimating*). As discussed above, there are also two composition rules for irrelevances, the universally applicable arithmetical mean and the normalized arithmetical mean, which is suitable only for a homogeneous data sample. To distinguish between the two ***kinds*** of gnostic distribution functions: the *local distribution function* will be denoted by L (which is based on the former composition rule), while the *global distribution function* will be identified by G (based on the latter rule). The result is four versions of gnostic distribution functions: the ELDF, EGDF, QLDF and QGDF. They all have a common root—they represent four applications of the same formula, and they are all related to the probability distribution of an individual datum (10.42) derived in Chapter 10:

$$**DF = (1 - h_{**})/2, \tag{15.19}$$

where h_{**} is the specific version of the data sample's irrelevance, either h_{EL}, h_{EG}, h_{QL} or h_{QG}. Each of these functions will now be examined in greater detail for samples, for which the scale parameter (S) is constant throughout the sample.

15.2.2 Transformations of Data Supports

The data weights and irrelevances as well as all their functions introduced in the previous chapters were taken to be defined over infinite data supports. There were two kinds of such data: additive data seen as arbitrary real numbers from the interval $R^1 := (-\infty, +\infty)$ and multiplicative data, strictly positive numbers from $R_+ := (0, \infty)$. However, real data are always bounded, and lie within a finite interval (LB, UB). The theoretical domain of gnostic distribution functions is R_+. In application of these functions, data must be transformed from their "natural" data support to the theoretical infinite domain of distribution functions. The converse is also true: when a quantile of a distribution function is estimated, it is set in the domain R_+; to obtain its "natural" form, it must be transformed backwards onto the finite data support. To unite the manipulations of additive and multiplicative data and to simplify the numerical calculation process, a unified finite closed interval is introduced,

$$\mathscr{Z}_e := [1/\exp(1), \exp(1)]. \tag{15.20}$$

To transform an element A_k from a sample of additive data onto the unified interval, the transformation

$$Tr_{az} := z_{fin} = \exp\left(\frac{2 * A_k - A_{max} - A_{min}}{A_{max} - A_{min}}\right) \tag{15.21}$$

can be applied; while in the case of multiplicative datum M_k formula

$$Tr_{mz} := z_{fin} = \frac{(M_k/M_{min})^{(2/\log(M_{max}/M_{min}))}}{\exp(1)} \tag{15.22}$$

can be used, where A_{max} and M_{max} are the largest and A_{min} with M_{min} the smallest data in the sample. The transformation $\mathcal{Z}_e \leftrightarrow (0, \infty)$ then can follow using the formula

$$Tr_{fininf} := z_{inf} = \frac{z_{fin} - LB}{1 - z_{fin}/UB}, \tag{15.23}$$

where LB is the lower and UB is the upper bound of the finite support. The backward transformations can be obtained by solving these equations with respect to their arguments.

Using 15.23 presents a problem, because it implies, that the data support is assumed to be an **open** interval, where the strict relations $z_{fin} > LB$ and $z_{fin} < UB$ must hold, but the solution of practical problems frequently requires closed data supports. Overcoming this difficulty is taken up in the next subsection.

15.2.3 Soft and Hard Data Bounds

The assumption, that real data are finite is based on the fact, that they are real objects in the real world, and that the world itself is a finite entity. While this fact is necessarily accepted, for data analysis, something more is needed: information as to the value of the data bounds.

In some special cases, the data support bounds LB and UB are known a priori, otherwise they must be estimated by using the EGDF. There are two different cases of a priori known bounds of data support:

1. The open interval of possible data (LB, UB): it is not expected, that data of values LB or UB will be seen in practice. Examples: zero weights for real objects (excluding balloons), reaching the speed of light by a non-zero mass particle.

2. The semi-closed or closed interval of possible data values, ie $(LB, UB]$, $[LB, UB)$ or $[LB, UB]$ data can reach the value of the closed end of the interval. So eg establishing the proportionality of a slice of pie always results in the closed interval $[0, 1]$, while the values of both bounds are possible: 0 means 'zero size for the slice' and 1 is equivalent to 'the whole pie.'

Data sets fitting the former case will be thought of as having *soft bounds*, while the latter one will have *hard bounds.*

Acceptance of the idea of finiteness of real data implies, that both the normal and lognormal distributions must be rejected, because their data support is infinite. While they may approximate the occurrence eg of people, who are close to average size, these distributions will not be very useful in predicting the presence of extremal sizes: dwarfs or those taller than NBA giants can serve as an example. Small but non-zero probabilities would be attached to heights exceeding several times the average. On the other hand, experience shows, that a maximum "possible" (understand "not improbable") height surely exists, but its value is different for each population. This value is of course uncertain, "fuzzy", but the form of a distribution function strongly depends on it. The motive for estimating these soft bounds is therefore obvious. The robustness of the EGDF can be used to estimate these bounds as shown in the applied portion of this book (Part III). The required (uncertain) information on the boundary values is thus 'mined' from the data. The value of the distribution function at the lower estimated soft bound (LB) is zero, while that of the UB is 1.

On the other hand, hard bounds for data come about from the **impossibility** of reaching data values beyond some point, but also having the **possibility**, that the boundary values could be attained: The total investment needed for a particular project cannot be negative nor can it exceed 100%, but a prospective investor could either fund its totality or refuse to participate. A holder of common stock cannot lose more than his initial stake, expenditures for R & D or capital goods cannot be less than zero, but then the firm need not make them at all, nor would dividends paid generally exceed 100% of earnings[6], but firms paying no dividends are relatively common in some industries. Such limitations are implied by the nature of things, the existence of hard bounds is sure. The frequency of occurrence of the boundary values can be directly estimated as a **discrete** probability. It is also true, that obtaining values "in between"

[6] In the relatively rare case, when dividends exceed EPS, the difference is a nontaxable 'return of capital' and deducted from retained earnings.

the two extremes is not excluded. The probability of nonextreme values is characterized by a (continuous or discrete) distribution function. In the continuous case, the application of a gnostic distribution function to data subjected to hard bounds easily combines the discrete estimate of the boundary probability ($Pr(LB)$ and/or $Pr(UB)$) of hard LB or UB with the continuous modeling of the nonextreme data. Its distribution function satisfies the constraints $Pr(LB)$ and $Pr(UB)$.

Distinguishing between the two types of bounds is important from both the theoretical and the practical (algorithmic) point of view.

15.2.4 The Estimating Local Distribution Function (ELDF)

To obtain this distribution function for a data sample $\mathcal{Z}$, the simplest version of the data sample's estimating irrelevance $H_i(\mathcal{Z})$ (13.13) as defined by Axiom 2, is applied. Averaging may be either with equal or unequal weights, therefore the sample's estimating irrelevance h_{EL} should be denoted by $\overline{h_E}$ in the same sense as in 15.15. By using the general expression 15.19, the *estimating local distribution function* (ELDF or simply EL) can be written in the form

$$ELDF \equiv EL(\mathcal{Z}, Z_0, S) := (1 - \overline{h_E})/2. \tag{15.24}$$

This same expression could also be obtained as the mean of probabilities 10.42 of individual data in the sample, which would result in a kernel estimation using the gnostic kernels.

Rewriting 15.24 in a more explicit form, again denoting the data in the sample as Z_k $(k = 1, ..., N)$, the ideal value as Z_0 and, using the auxiliary variables 15.13 and 9.11

$$EL(\mathcal{Z}, Z_0, S) = \overline{\left(\frac{1}{1 + q_k^4}\right)} \tag{15.25}$$

is obtained. The EL's density can be found easily by differentiation[7]

$$\frac{dEL}{dZ_0} = \frac{1}{SZ_0}\overline{\frac{4}{(q_k^2 + 1/q_k^2)^2}}. \tag{15.26}$$

[7] This formula—as well as all other formulae of probability densities in this chapter is based on the assumption of a constant scale parameter S.

It is useful to recall the interpretation of the probability distribution of an individual datum as the probability of the ideal value Z_0 given the observed value Z_k. This means, that this probability as well as its density are functions of Z_k, and that the ELDF (as well as other gnostic distributions) are functions of data and of the 'free' quantile Z_0.

The relations 15.25 and 15.26 reinforce the universal existence and applicability of this distribution function. Indeed, both the irrelevance and the gnostic kernel can be attached to an arbitrary datum. Gnostic kernels defined over the infinite data support are positive and finite and their arithmetic mean is positive and finite as well. Moreover, the range of the function EL (15.25) is $[0, 1]$. This function is non-decreasing, because it represents the integral of positive kernels, and it can therefore serve as a distribution function. Its interpretation as the probability distribution function is thus justified. The adjective "local" emphasizes the fact, that no "global" feature of the data sample such as its homogeneity was assumed. It can be seen, that the ELDF is really "local" in the sense, that it characterizes the data distribution even over a small subinterval of the data support if the scale parameter S is sufficiently small.

15.2.5 The Estimating Global Distribution Function (EGDF)

This distribution function is based on the idea, that if a data sample is homogeneous, then it is possible to represent it by a single gnostic event. The gnostic weight and irrelevance of this equivalent event can be calculated by formulae 14.14. In the estimating case ($c := i = \sqrt{-1}$), using 14.8 (the overline is the symbol of weighted averaging as above), one can rewrite the equivalent irrelevance from 14.14 as

$$h_{EG} := \frac{\overline{h_E}}{M_{Z,i}}, \tag{15.27}$$

where $\overline{f_E}$ and $\overline{h_E}$ are (weighted) means of fidelities and irrelevances of all the data in the sample $\mathcal{Z}$, and where $M_{Z,i}$ is the sample's estimating modulus 14.8,

$$M_{Z,i} := \sqrt{(\overline{f_E})^2 + (\overline{h_E})^2}. \tag{15.28}$$

In this case, the distribution function from 15.19 takes on the form

$$EGDF \equiv EG(\mathcal{Z}, Z_0, S) = \left(1 - \frac{\overline{h_E}}{M_{Z,i}}\right)/2. \tag{15.29}$$

The EG's density can be derived by differentiating (15.28) to get

$$\frac{dEG}{dZ_0} = \frac{1}{SZ_0} \frac{(\overline{f_E})^2 * F2 + \overline{f_E} * \overline{h_E} * FH}{M_{Z,i}^3}, \tag{15.30}$$

where

$$F2 = \sum_{k=1}^{N} w_k * f_{E,k}^2 \qquad FH = \sum_{k=1}^{N} w_k * f_{E,k} * h_{E,k} \tag{15.31}$$

represent the (weighted) means of squared data fidelities and the products of the fidelities and irrelevances, respectively. It is worth noting, that the second addend in the numerator of 15.30 may be negative. When all data are so precise, that all fidelities $f_{E,k}$ tend to 1, the term FH approaches $\overline{h_E}$ and its product with $\overline{f_E}$ is positive. However, in the case of strong uncertainties the variables FH and $\overline{h_E}$ reach negative values, not necessarily simultaneously, thus making the product negative. This effect may be so strong, that expression 15.30 is negative. In such cases the EGDF (15.29) does not exist, because it does not possess the important feature of a distribution function—that it not decrease. There are two interpretations of this characteristic of the EGDF's:

Bad: The applicability of the EGDF is not universal. This distribution function is suitable only for data samples, for which the EGDF's density is non-negative over its full range, and which have only one maximum.

Good: The EGDF may be used to test the homogeneity of a data sample. It will be shown shortly, that such tests are extraordinarily efficient and reliable, and that they reveal important features of the data.

For data with large uncertainties (which are far from the ideal value), the auxiliary variable q (15.13) tends to either zero or to infinity. This causes the estimating fidelities (15.12) to be negligible and the corresponding irrelevances to approach 1 or -1. The EGDF therefore suppresses the influence of the "peripheral" data and focuses on the "central" or "inner" data, for which the fidelities are close to 1 and the irrelevances tend toward zero.

Other interesting features of this distribution function will be discussed in the next chapter.

15.2.6 The Quantifying Local Distribution Function (QLDF)

There is an obstacle, which must be overcome in defining the quantifying versions of distribution functions: both the quantifying irrelevance h_i

(9.11) and the improbability 10.58 can reach infinite values, so that the general formula, 15.19, cannot be applied directly. However, as it was shown in connection with the problem of quantifying kernels in Chapter 11, it is possible "to see" quantifying irrelevance as if it were the estimating irrelevance by using formula 11.17. It is therefore possible to compose the quantifying irrelevances of all a sample's data and to "observe" the results by means of "estimating eyes" via the transformation 11.17[8]

More particularly:

$$h_{QL} := \frac{\overline{h_Q}}{\sqrt{1+(\overline{h_Q})^2}}, \tag{15.32}$$

where $\overline{h_Q}$ is the (weighted) mean of quantifying irrelevances of all of a sample's data. (Recall, that by 9.11 the quantifying irrelevance of the k–th datum is $(q^2-1/q^2)/2$, while the quantifying weight equals $(q^2+1/q^2)/2$ by 9.10.) In this ("local") case the non-normalized composition of irrelevances is applied as in the case of h_{EL}. Then the ***quantifying local distribution function*** takes on the simple form of 15.19

$$QLDF \equiv QL(\mathcal{Z}, Z_0, S) := (1-h_{QL})/2. \tag{15.33}$$

Denoting $\overline{f_Q}$ the (weighted) mean of the quantifying irrelevances of the sample's data and differentiating 15.33 one arrives at the density

$$\frac{dQL}{dZ_0} = \frac{1}{SZ_0}\frac{\overline{f_Q}}{(1+(\overline{h_Q})^2)^{3/2}}. \tag{15.34}$$

It is easy to see, that for a single datum ($N=1$) this expression provides the same results as is obtained for the densities of the ELDF's or the EGDF's—the gnostic kernel 11.9—because $\frac{dA_0}{dZ_0} = \frac{1}{Z_0}$. This does not mean, that the distribution functions will be similar for $N > 1$; the opposite is true, because the mean irrelevance $\overline{h_Q}$ and the mean quantifying weight $\overline{f_Q}$ may include unbounded terms caused by the observed data's significant divergence from the ideal value. Taking this into account, one can expect, that the QLDF will emphasize the "peripheral" data of the sample (the outliers).

[8] It is important to see, that this artificial step was taken to obtain quantifying distribution functions. These have values in the interval $(0,1)$, which are interpretable as probabilities. This is how distributions complementary to the EGDF, with their opposite robustness can be obtained. However, improbability as introduced in 10.42 is a complex function, which can give rise to a (complex) distribution function of data improbability, the modulus of which rises to infinity with increasing uncertainty. Such a function can also have a realistic interpretation: eg the wave function of quantum mechanics behaves in a similar way.

15.2.7 The Quantifying Global Distribution Function (QGDF)

The same problem exists here as with the QLDF and the same trick is employed to solve it, but using a different composition rule. Instead of (weighted) averaging in accordance with 13.13, the normalized (weighted) average 14.14 is used:

$$h_{Z,j} := \frac{\overline{h_Q}}{\sqrt{(\overline{f_Q})^2 - (\overline{h_Q})^2}}, \tag{15.35}$$

where again $\overline{f_Q}$ and $\overline{h_Q}$ represent (weighted) means of quantifying weights and irrelevances respectively. Using transformation 11.17,

$$h_{GQ} := \frac{h_{Z,j}}{\sqrt{(1 + (h_{Z,j})^2)}} \tag{15.36}$$

is obtained. The *quantifying global distribution function* is therefore simply

$$QGDF \equiv QG(\mathcal{Z}, Z_0, S) := \left(1 - \frac{\overline{h_Q}}{\overline{f_Q}}\right)/2 \tag{15.37}$$

and its density is

$$\frac{dQG}{dZ_0} = \frac{1}{SZ_0} * \left(1 - \frac{(\overline{h_Q})^2}{(\overline{f_Q})^2}\right). \tag{15.38}$$

It is again easy to show, that in the case of a single datum, this density reduces to the form of one gnostic kernel as in all the other cases. Moreover, for sufficiently precise data the mean weight $\overline{f_Q}$ will approach the data sample's quantifying modulus $\sqrt{((\overline{f_Q})^2 - (\overline{h_Q})^2)}$. This remainder will vanish and both the QLDF and the QGDF will return the same values except in the case of strongly dispersed data.

15.3 Main Features of the Distributions

15.3.1 The Unlimited Flexibility of the ELDF

Given data and a scale parameter S, the EL is a non-decreasing function of the unknown ideal data's value Z_0. Its character is determined by the scale parameter. The EL function converges to the step function (DDF), when the scale parameter S tends to zero and its derivative 15.24 converges to a collection of N δ-functions each placed at a different data point. Fig. 15.1

demonstrates the behavior of the EL-distribution function as it approaches the step function for a small data sample $\mathcal{Z}_4 \equiv \langle 1, 3, 4, 4.5 \rangle$ for $S = 0.5, 0.1$ and 0.05. Systems of both KS-points and ME-points are shown as well. Note in the figure, that the values of the EL-functions at each of the data point are close to the KS-points, but not necessarily near the ME points, especially for a large S.

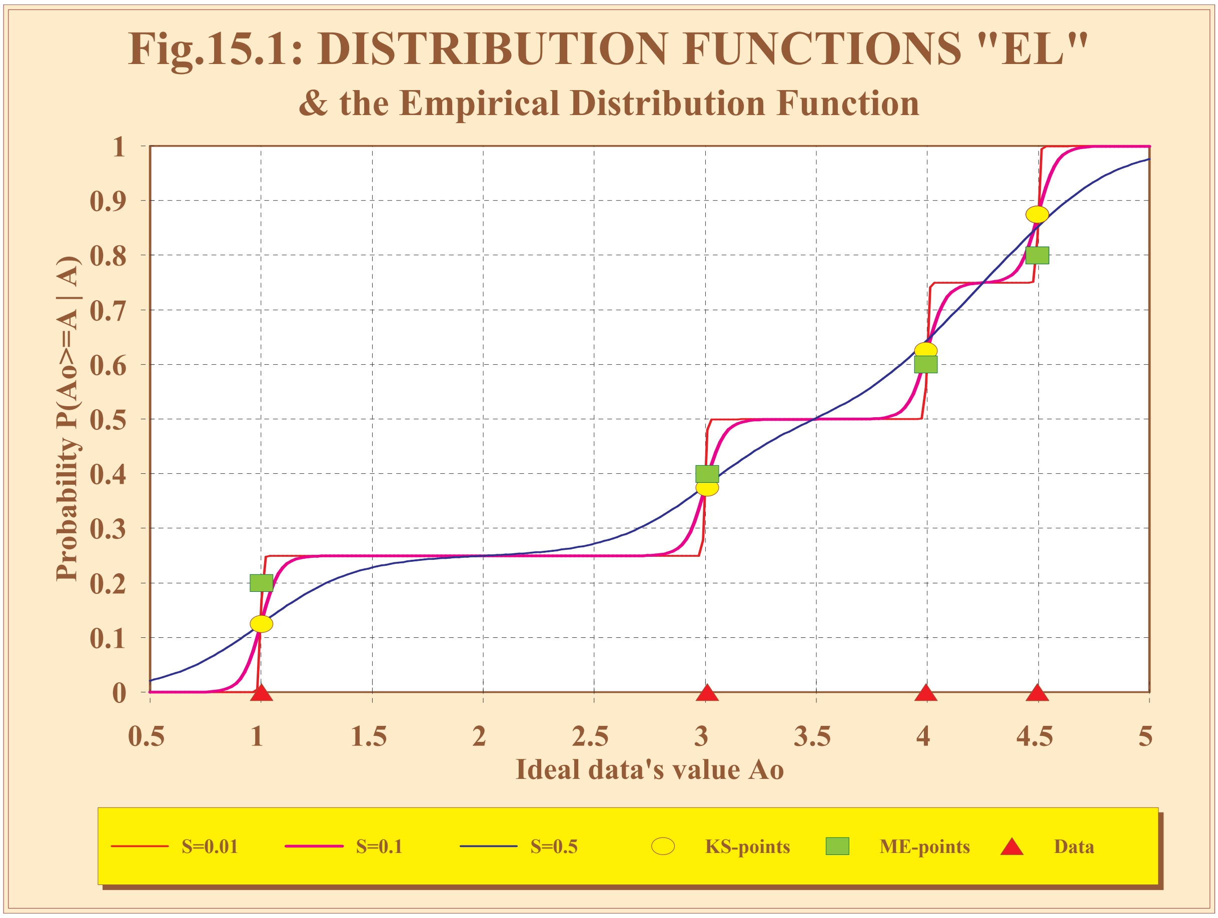
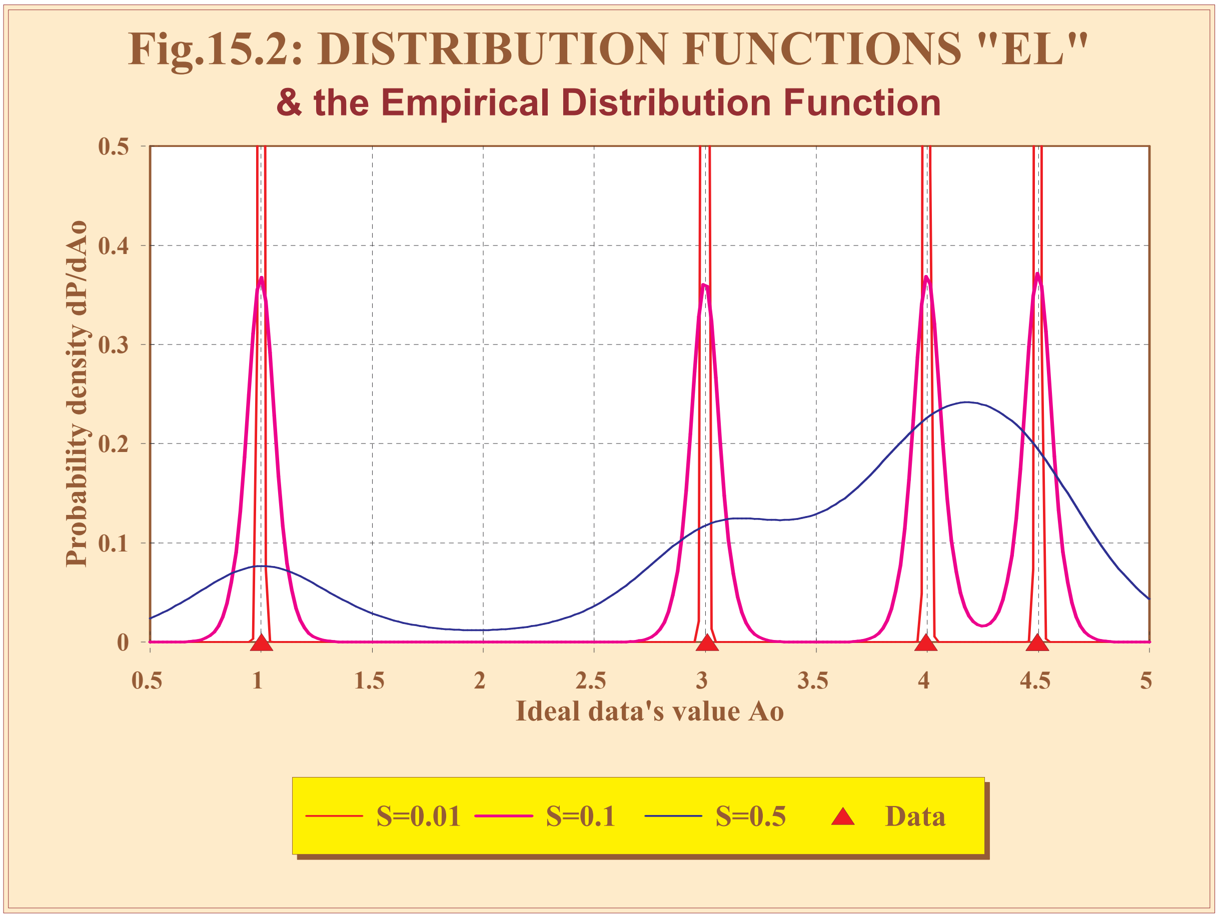


Fig.15.1: DISTRIBUTION FUNCTIONS "EL"

& the Empirical Distribution Function

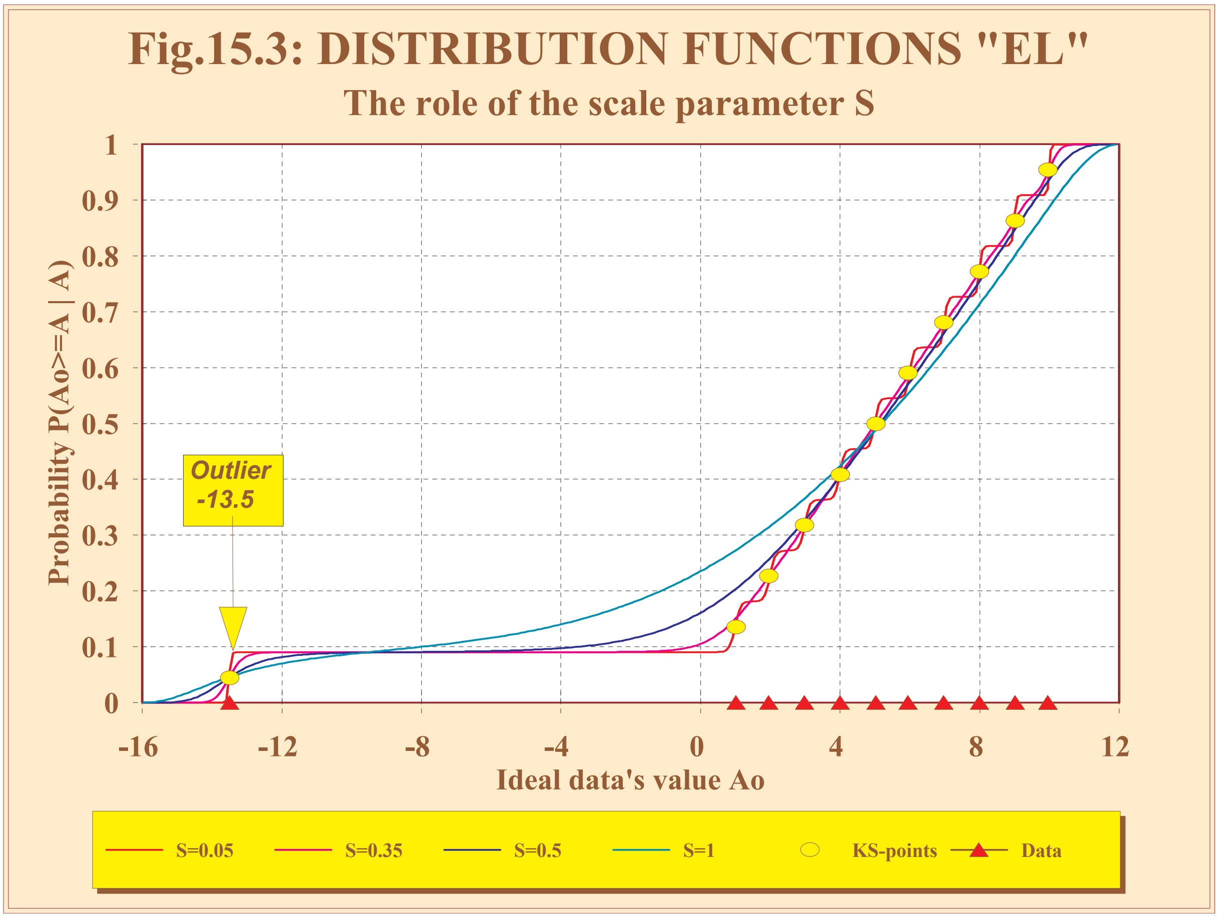
The dependence of the EL-density on the scale parameter is illustrated by Fig. 15.2. Observe, that the two kernels (*peaks*) belonging to data values 4 and 4.5 are entirely separate, when $S \leq 0.1$, but integrate into a perfectly smooth hill form, when S increases to 0.5. (Such a group of integrated peaks will be called a *cluster*). A scale parameter of 0.5 is large enough, in this case, to create an inflection in the density curve at the point 3.0. This causes the three peaks representing data values 3, 4, and 4.5 to be assimilated into one cluster. The appearance of such a plateau on a density curve always signals, that the density peak has screened out some other data. The largest cluster of data will be called the *main cluster*. We have now arrived at the idea of *marginal cluster analysis*.



15.3.2 Marginal Cluster Analysis

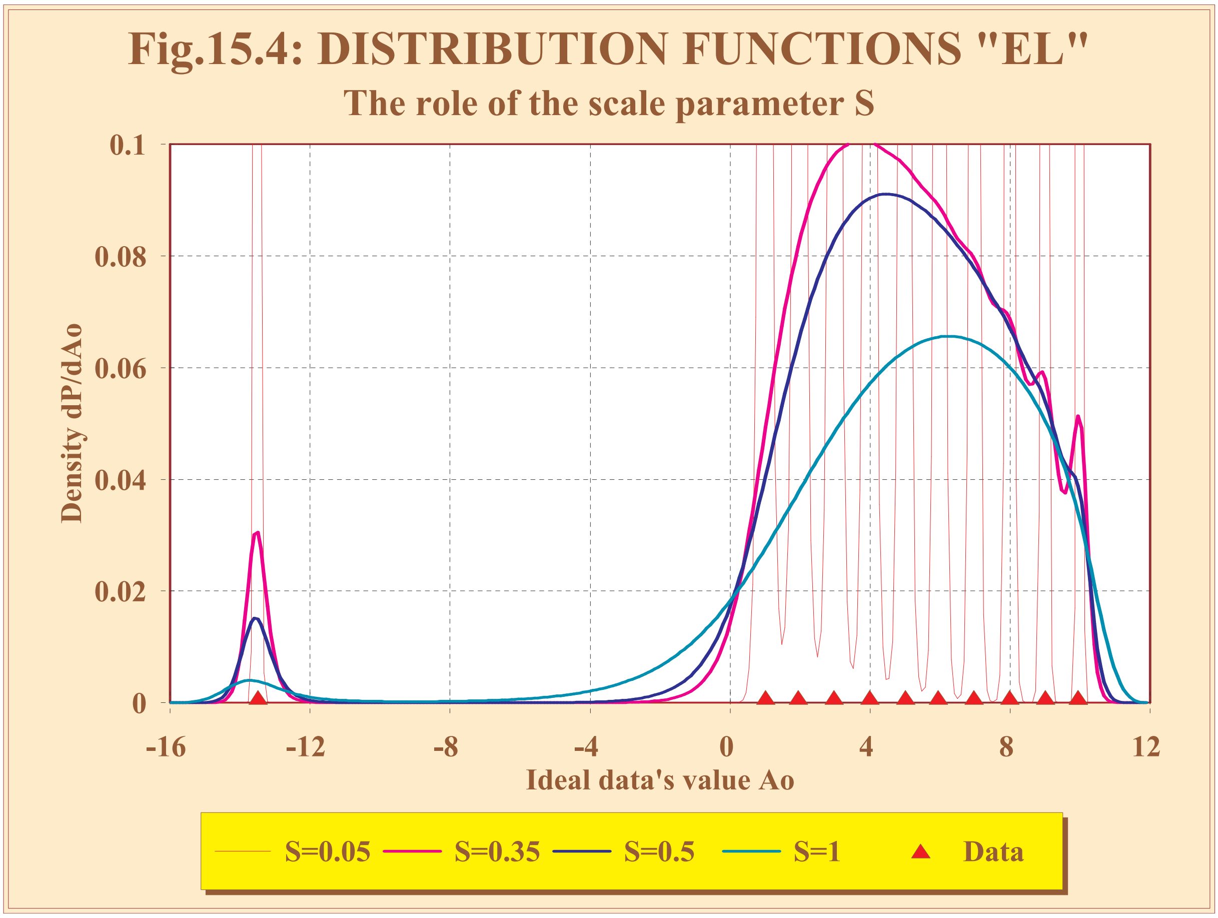
In statistics, the notion of “marginal distribution” describes the unconditional distribution of a single variable from a multidimensional model. When addressing *marginal cluster analysis* in gnostics, we shall keep in view the decomposition of a one-dimensional data sample into two or more subgroups, which manifest themselves in the sample’s density graph as separated clusters. A simple way of doing this is by using the EL-distributions.

A data peak such as the one belonging to the smallest datum (1.0) in Fig. 15.2 is sufficiently removed from other data to be considered as separate from the main cluster. Similarly, Figs. 15.3 and 15.4 demonstrate, that a data sample $\mathcal{Z}_{11} \equiv \langle -13.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \rangle$ is found to be composed of 10 regularly distributed data and an “outlier” (−13.5).



The EL-distribution is close to the step function (DDF) if the scale parameter is as small as 0.05. For the same value of scale parameter, the corresponding density (Fig. 15.4) reveals separate peaks for each datum, which subsequently dissolve into a principal cluster and an outlier as the scale parameter is increased. If too small scale parameter is used, the result is no better than just plotting the data values (the red triangles) on the horizontal axis. Increasing S smoothes the density curve and leads to the conclusion, that there is a “homogeneous” cluster of 10 data and an outlier (-13.5). The designation of this data value as an outlier is supported by the fact, that it has a separate peak even with as large a value of S as 1; this also puts it at the lower extremity of the probability curve of the series of KS-points in Fig. 15.3.

Another example demonstrating the flexibility of the EL-distributions and of the integration of groups of peaks into clusters is shown in Figs. 15.5 and 15.6 by a symmetric sample $\mathcal{Z}_{11,s} \equiv \langle -10, -9, -8, -0.2, -0.1, 0, 0.1, 0.2, 8, 9, 10 \rangle$. This artificial



sample contains one narrow central group of 5 data and two symmetrically placed peripheral clusters each formed by three data.

Marginal cluster analysis has to answer two categories of questions:

1. How many clusters are there in the data sample?
2. Which data form each of the clusters?

Even such a simple case as that of $\mathcal{Z}_{11,s}$ shows, that there is no uniqueness to the answers. As S is varied, 11, 7, 3 or only 1 cluster are obtained. It is theoretically possible to get a separate peak for each different data value, however when two or more data are equal, separate peaks cannot be distinguished even by making S extremely small.

Due to computation limitations, the exponent $2/S$ cannot increase over a certain limit; this is the reason, that the possible maximum of 11 peaks is not realized (the inner data group is too narrow in comparison to the sample range). But whether the “true” number of clusters is 7, 3 or 1 still remains to be determined. Instead of a single answer, the following issues

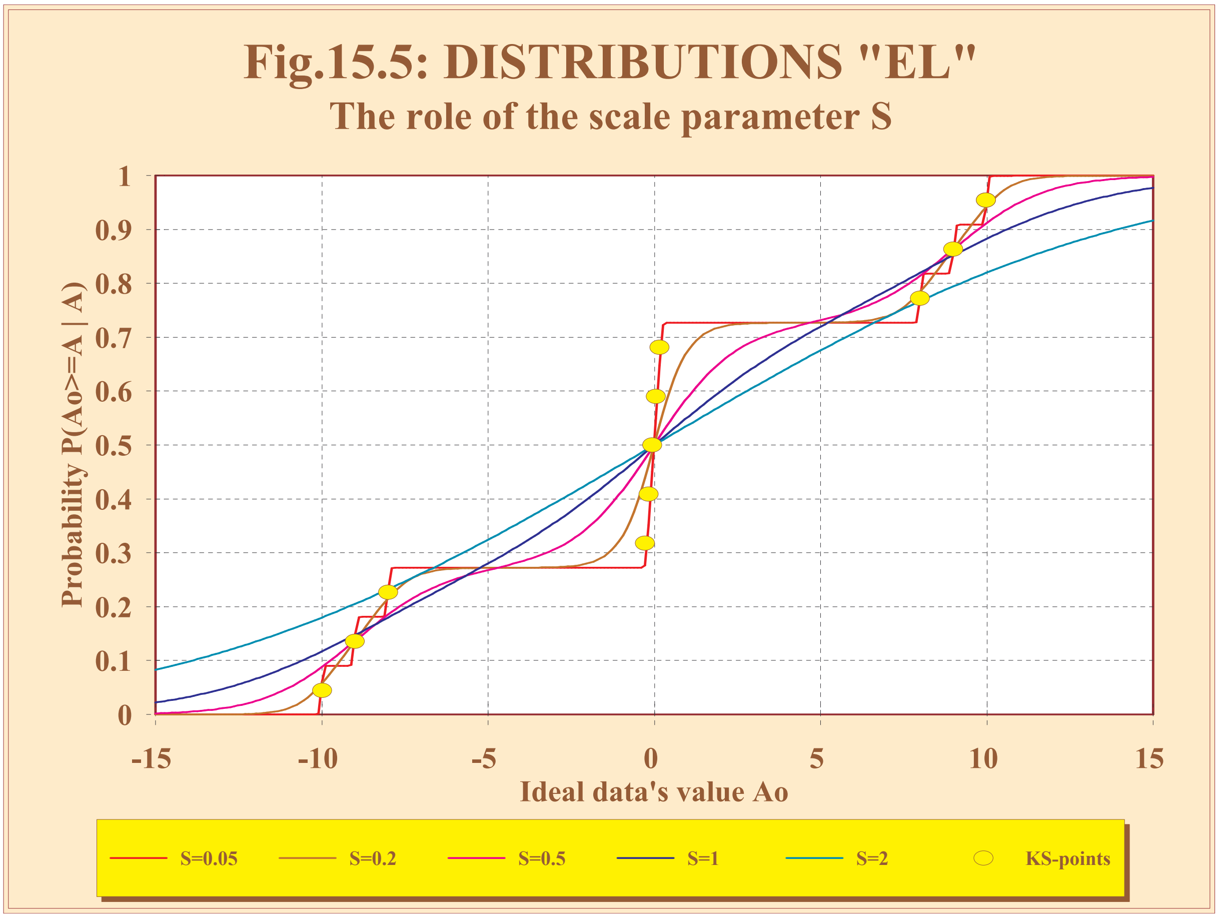


Fig.15.5: DISTRIBUTIONS "EL"

The role of the scale parameter S

must be considered:

- An objective and unique test for the homogeneity of a data sample cannot be accomplished by using the EL-function; the EG distribution function must be used. This will be considered in the following section.
- The choice of the S to use with the EL-distribution is a question of the *resolution power* of the analysis, which depends on the particular goal—in what detail do we want to see the inner sample structure. (This choice is analogous to that of using the zoom of a camera: "What is the right zoom level?").
- There are two important aspects involved:
  - It is possible to change the resolution power of the analysis by choosing a suitable value for S.
  - The number of separable clusters may be less then the number of data:
    - In some particular samples a specific number of clusters cannot be obtained at all. In Fig. 15.6, the density cannot have an

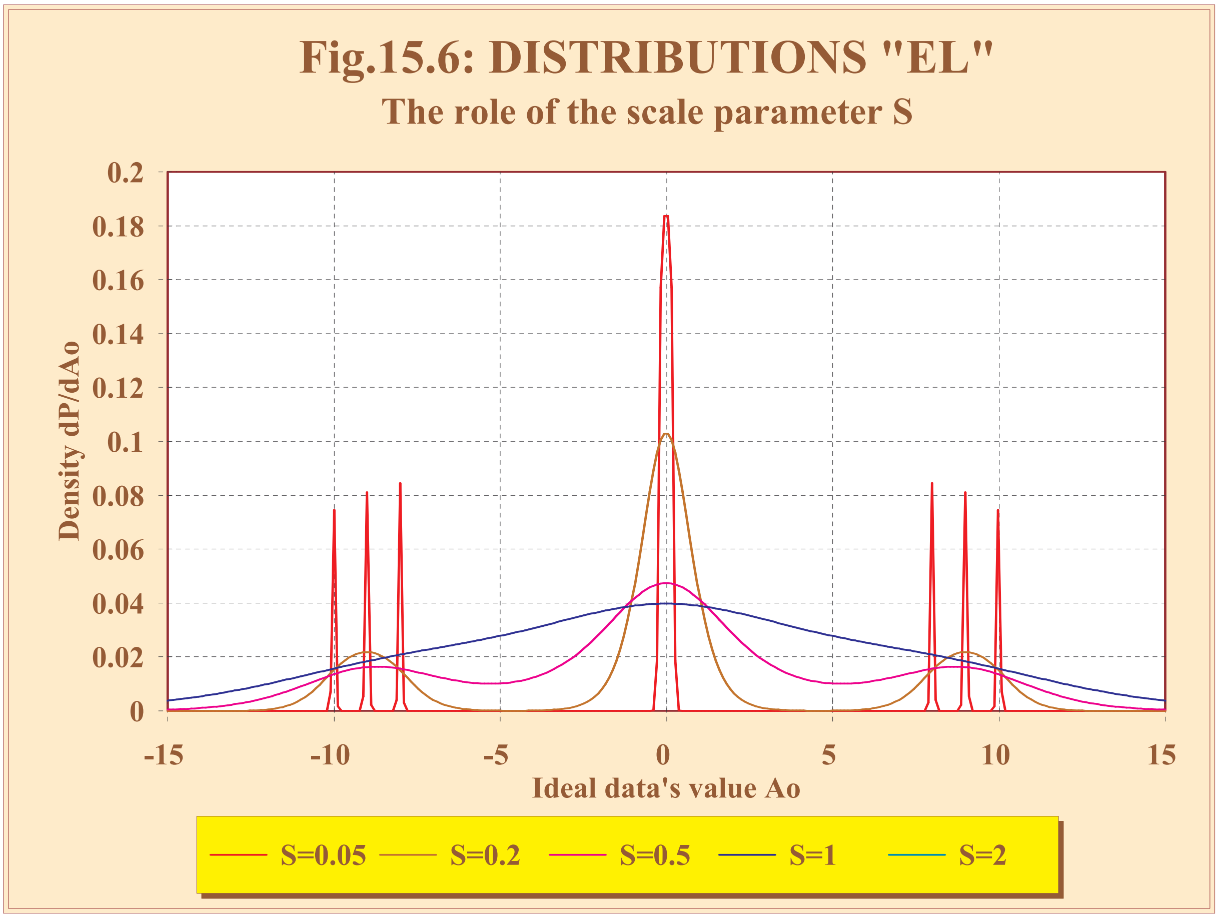


Fig.15.6: DISTRIBUTIONS "EL"

The role of the scale parameter S

even number of maxima because of its symmetry (in some cases this may be an important result of the analysis).

∗ Data, which have the same value, contribute to the same cluster.

In different applications, the size of the interval for S, over which the number of clusters remains constant, is not the same; therefore, the “best” number of clusters to retain may often be taken as that number of clusters, which corresponds to the widest interval usable without reducing the number, that can be distinguished.

Which data belong to each cluster is only a technical question. Once the number of clusters is established, it is easy to separate them by finding the local minima of the density curve. Data associated with each of the intervals between the minimum points are those, which form the cluster above the interval.

Fitting errors of the ELDF

As explained above, there is a benefit to be gained from reducing the value of the scale parameter, S, but this also results in a trade-off with respect to the residual entropy ratio, RE. This relationship is summarized in Tab. 15.1 for the data sample $\mathcal{Z}_{11,s}$:

Scale parameter S	**Residual entropy** RE
0.05	0.972
0.1	0.929
0.2	0.868
0.35	0.810
0.5	0.769
1	0.714
3	0.705

Tab. 15.1 The residual entropy of the EL-distributions for the symmetric data sample $\mathcal{Z}_{11,s}$.

Decreasing S improves the goodness-of-fit of the EL-distribution, but the paid price for this is a reduction in the smoothness of the distribution function. This can be seen in Fig. 15.4, where wavelets appear in the descending portion of the main density cluster at an already relatively high value of $S = 0.35$.

15.3.3 The Robustness of the ELDF

The high flexibility of the ELDF may lead to an impression, that this distribution function is not robust, because the notion of robustness connotes a feeling of rigidity contrary to the idea of flexibility. Such a conclusion is not valid, at least in the case of the ELDF, because the smaller the value of the scale parameter, the more flexible the ELDF is in adapting to the placement of the data, and the weaker the influence of data, which form an independent local cluster. There is a simple explanation: the width of each kernel becomes narrower as the scale parameter decreases. This independence of a local cluster with respect to a more distant cluster is *the local robustness.* A much more complex set of interactions of data treated with the ELDF will be analyzed in next chapter in connection with estimates of parameters of location. The above effects, related to local robustness, have an important application for the ELDF in *interval analysis*; it allows data to be split into several classes, ie to classify them with respect to their relationship to the best estimate of the 'central' value of the data sample. This problem is examined in section 16.4.

15.3.4 The Uniqueness of the EGDF

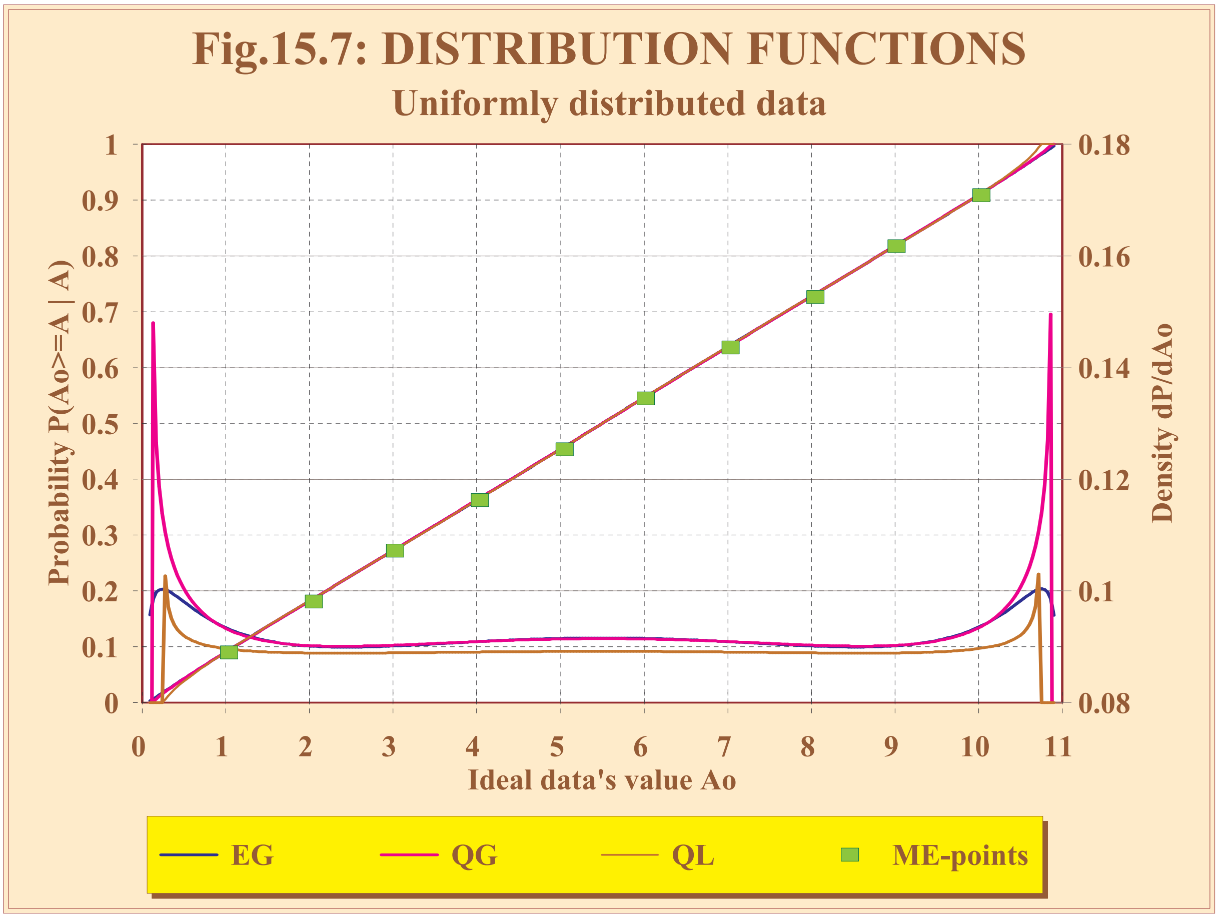
The estimating *global* distribution function (EGDF) behaves differently than the ELDF. Because of the normalization by the sample modulus, the scale parameter is unable to control the flexibility of the EGDF as it does in the case of the ELDF. The dependence of fitting errors on the scale parameter is not a monotonic function as it is with the 'local' function. When the parameter S is varied over a broad interval, the best value is always found, where the fitting criterion reaches its minimum. No advantage is gained beyond that point. This is the reason, why only the EGDF finds the best scale parameter and—if it is needed—the best estimate of the bounds of data support. This is important from several points of view:

- it simplifies applications—the scale parameter is chosen automatically together with the bounds for the data support,
- it enables hypotheses such as the homogeneity of data samples to be tested,
- the one and only (the best) EGDF for a homogeneous data sample can be accepted as the sole representation of the data's distribution and it provides unique values of probability for selected quantiles and unique quantiles for given probabilities.

In other words, the EGDF can be used as a model uniquely determined by a data set[9]. It may be a good model, but other methods may produce different models and the final choice of the model is always a matter of a careful analysis. The applicability of the EGDF to uniformly distributed data is demonstrated in Fig. 15.7.

The scale parameter and both bounds of the data support have been optimized to minimize the fitting criterion (in this case to minimize the entropy). The probability distribution is practically a straight line passing through all the ME-points. The EGDF's density is nearly constant over a broad interval between 2 and 9 and changes from this constant value only

[9] The uniqueness problem is examined in 15.3.9 in more detail.



when the bounds of the data support are approached.

15.3.5 Testing for Data Homogeneity

The fact, that there are two density maxima shown in Figure 15.7 may appear contradictory, because in theory a homogeneous data sample should only have a single maximum; and from the plot, the data sample surely appears to be homogeneous. The answer to this paradox is simple: the condition of uniqueness of a density's maximum is related to the distribution function defined over infinite data support. The process of calculating an EGDF passes through three important stages:

1. Transformation of the data (which are always defined over a finite data support) to infinite data support R_+. This transformation is parameterized by the lower and upper bounds (LB and UB) of the data support. These bounds may sometimes be given, but if they are unknown, their value is optimized together with that of the S, so as

to minimize the fitting criterion of the EGDF that is applied to the transformed data (over R_+).

2. Testing to determine the number of density maxima and for the density's polarity. The data sample is homogeneous only if the EGDF's density has a single maximum over infinite data support R_+. And the EGDF is a distribution function only if its density is non-negative over its full range.
3. Back transformation $R_+ \rightarrow (LB, UB)$ using the optimum value of S and the given or optimized bounds LB and UB.

The EGDF's distribution as shown in Fig. 15.7 results from the third stage (after back transformation to finite bounds). It can be shown, that over R_+ there is no doubt as to the data's homogeneity; the data sample's density has only one maximum.

There can be two causes for a data sample's non-homogeneity; the existence of:

1. outliers, or
2. clusters.

An *outlier* is a datum, the value of which is so different from the other data, that the EGDF's density has a local maximum near the datum's location. The ability of the EGDF to sensitively reveal this non-homogeneity is demonstrated in Fig. 15.8 and Fig. 15.9 for the data sample $\mathcal{Z}_{11}$.

The blue probability distribution function in Fig. 15.8 does not appear to change its form at the value of the outlier (-13.5), but this is only because the shape of the curve changes only minimally at that point, the vertical scale in either figure is too rough to be able to visualize the local "hill" of the density there, but numerical analysis shows, that it really exists, and that its value reaches a local maximum of 0.00083. This same result can be obtained with the EGDF over infinite data support.

The lesson is, that it is inappropriate to rely on what is seen on the screen or on a graph. The decision as to the homogeneity/non-homogeneity of the data must be made with the more sensitive numerical methodology.

The symmetrical data sample $\mathcal{Z}_{11,s}$ was used in Fig. 15.6 to show, that the ELDF can interpret the peripheral triples of data either as groups of three outliers (with a small S) or as two integrated clusters (using eg $S = 0.2$). There is no such ambiguity in the case of the EGDF (Fig. 15.10 and 15.11).

The disturbances caused by the peripheral data are shown directly by the EGDF's probability distribution, the small 'bumps' in the blue line

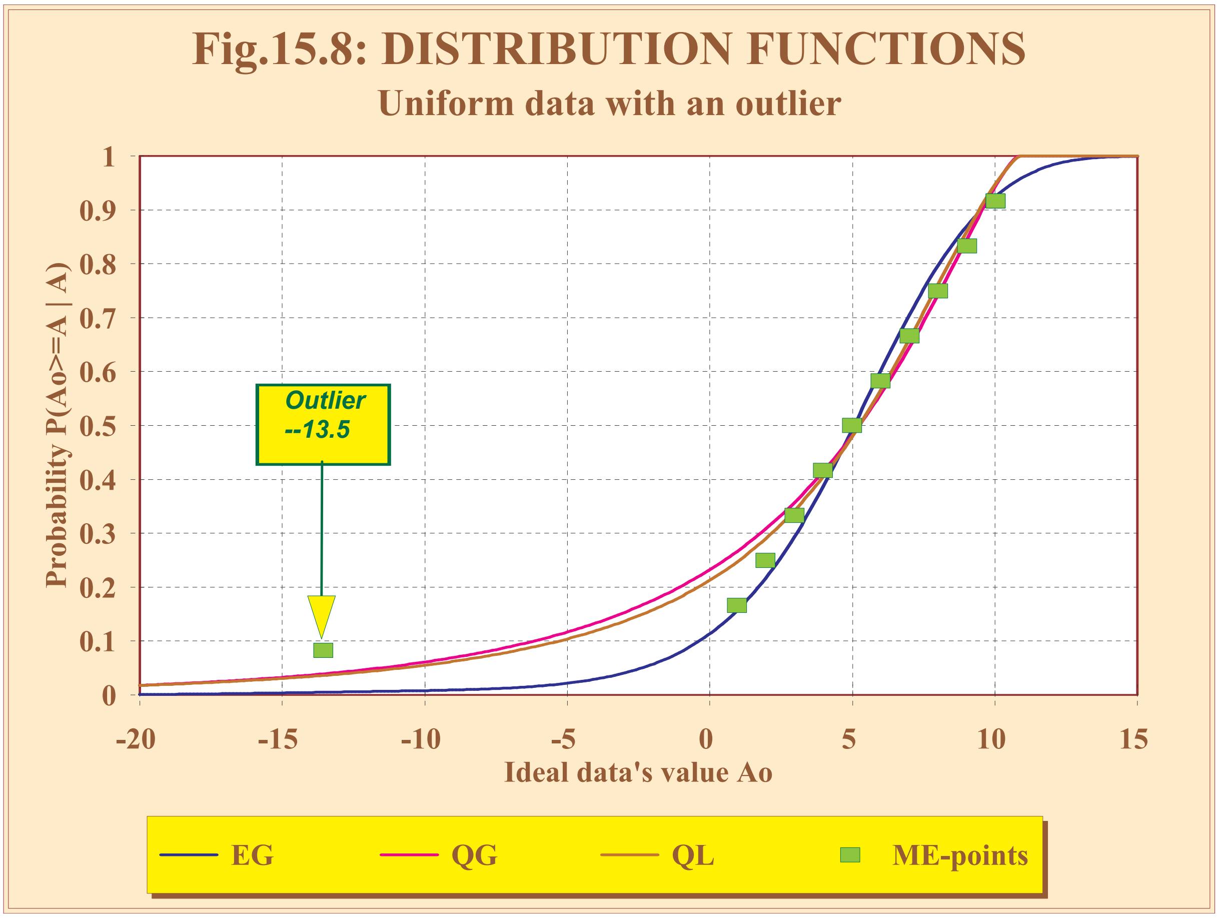


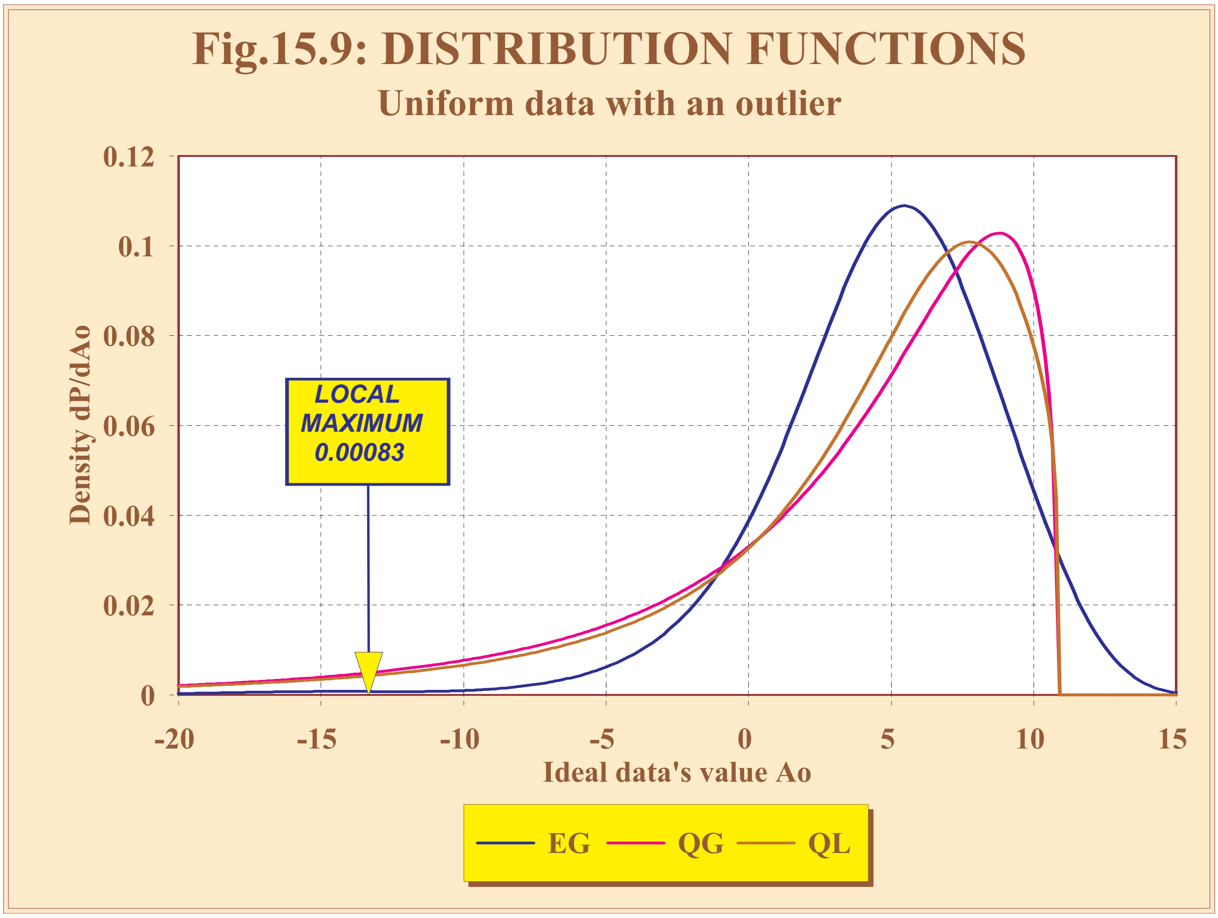
Fig.15.8: DISTRIBUTION FUNCTIONS

Uniform data with an outlier

of Fig. 15.10. The EGDF's density curve in Fig. 15.11 is even more impressive: the peripheral data are seen as triples of outliers, because they produce separated local maxima. Moreover, the density dips to negative values at these data points. The necessary condition for a distribution function—that it be non-decreasing—is violated. The blue function in Fig. 15.10 calculated with the EGDF's algorithm therefore cannot be used as a distribution function. However, it provides the useful and important information, that the peripheral data are separated outliers. Since the group of five central data appears in Fig. 15.11 as a single cluster, it can be expected, that if the six outliers were deleted, the resulting sample of five data would pass the homogeneity test.

15.3.6 The Robustness of the EGDF

Consider the examples shown above from a new point of view: the reaction of the blue distribution function (EGDF) to the existence of the outlier



−13.5 in Fig. 15.8 is so minute, that it cannot be seen in the graph. The distance between the blue line and the ME-point corresponding to the outlier is large, and the fitting error at this point is much greater than that of the data belonging to the main cluster. This feature illustrates *the robustness with respect to outliers* or *inner robustness.* The central data of the sample are accepted by the EGDF with a gnostic weight close to 1, while the peripheral data are given much less weight. This can clearly be observed in Fig. 15.10: the "inner" cluster of five data is completely assimilated by the blue distribution, while the outer triples of data are practically ignored.

15.3.7 Estimating a Sample's Boundaries

Consider the EGDF denoted as $P(\mathcal{Z}, \mathcal{Z}_{\prime}, \mathcal{S})$, where P is probability, the EGDF's value at the point Z_0 (quantile), $\mathcal{Z}$ is the data sample and S the scale parameter. Assume, that the data have already been transformed

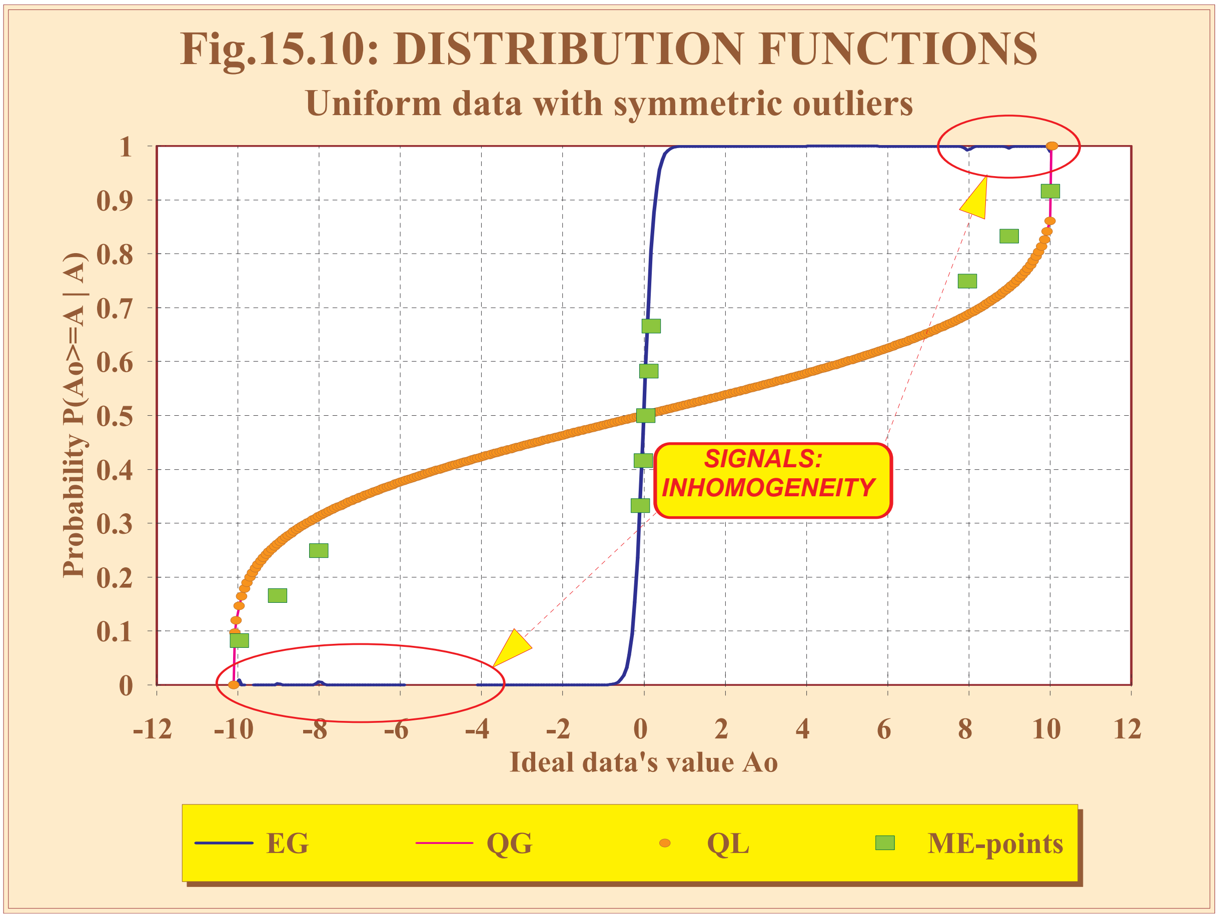


Fig.15.10: DISTRIBUTION FUNCTIONS

Uniform data with symmetric outliers

onto infinite data support (so that all data as well as all values of Z_0 are strictly positive). A homogeneous data sample is defined as one composed of only one cluster, therefore it has only one density maximum. If there are several clusters in a sample then each will have a separate maximum in the sample's density function. This motivates two important tests:

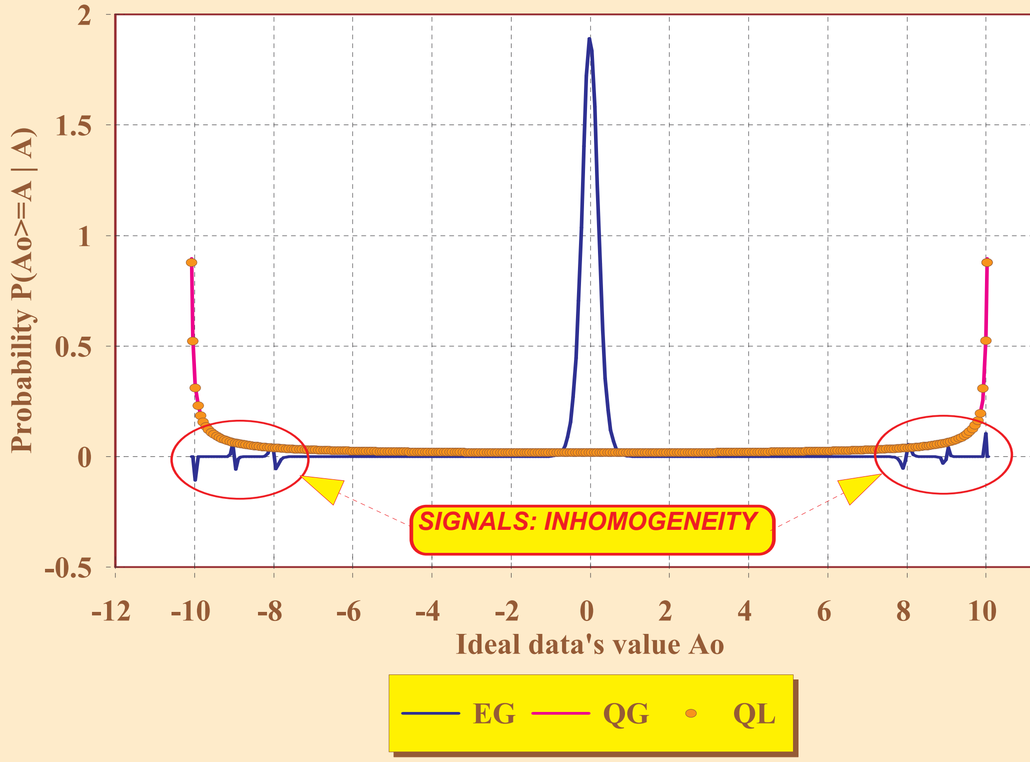
Definition 16: Let $P(\mathcal{Z},\mathcal{S}_{\mathcal{G}},\mathcal{Z}_{\prime})$ be the EGDF of a fixed sample $\mathcal{Z}$ of positive data and of an arbitrary positive quantile Z_0. Let S_G be the *global scale parameter*, which optimizes the EGDF's fit with the data. Let N_0 be the number of positive and finite solutions Z_x of the equation 15.39

$$\frac{dP}{dZ_0}|_{Z_0=Z_x} = 0. \tag{15.39}$$

The determination of the number N_0 will be called the *homogeneity test*. The positive statement ("the sample is homogeneous") will be considered as supported if $N_0 = 1$ and rejected with an $N_0 > 1$.

Fig.15.11: DISTRIBUTION FUNCTIONS

Uniform data with symmetric outliers



Let $\mathcal{Z}$ be the same data sample as above and S_G its global scale parameter. Let Z_ξ be a positive variable and $\mathcal{Z}' := \langle \mathcal{Z}, \mathcal{Z}_\xi \rangle$ the extended sample. Let $\mathcal{Z}$ and S_G be fixed.

Let $P'(\mathcal{Z}', \mathcal{S}_\mathcal{G}, \mathcal{Z}_\prime)$ be the EGDF of $\mathcal{Z}'$ abbreviated as $P'(Z_\xi, Z_0)$.

Let $Z1$ and ZL be respectively the smallest and largest datum in $\mathcal{Z}$. Let LSB be such a value of $Z_x i$, that all three relations

$$0 < LSB \leq Z1, \tag{15.40}$$

$$\frac{d^2 P'(Z_\xi, Z_0)}{dZ_0^2}|_{Z_\xi = LSB} = 0 \tag{15.41}$$

and

$$\frac{d^3 P'(Z_\xi, Z_0)}{dZ_0^3}|_{Z_\xi = LSB} = 0 \tag{15.42}$$

simultaneously hold. Let LSB be the only number satisfying these conditions. Then LSB is the *lower bound of the sample* $\mathcal{Z}$.

Let USB be such a value of the variable Z_xi, which satisfies both equations 15.41 and 15.42 and the relation $ZL \leq USB < \infty$. Let USB be the only number satisfying these conditions. Then USB is the *upper bound of the sample $\mathcal{Z}$.*
Numbers LSB and USB will be the *bounds of the membership interval* of the sample $\mathcal{Z}$, shortly the *sample's bounds.*

A number Z_{PM}, for which the relation

$$LSB < Z_{PM} < USB \tag{15.43}$$

holds, is the *potential member* of the sample $\mathcal{Z}$. Let Z be a given positive number. Then the process consisting of the following steps:

1. determination of the LSB,
2. determination of the USB,
3. verification, that 15.43 holds for Z,

is the *membership test.*

It is well-known, that a point at which the second derivative of a function reaches zero is an *inflection* point. The probability distribution function $P'(Z_\xi, Z_0)$ is dependent on the value of the extending variable Z_ξ, which plays the role of the additional datum to be tested. The first derivative $dP'(Z_0, Z_\xi)/dZ_0$ is the density. At the point Z_0, where the second derivative $d^2P'(Z_0, Z_\xi)/dZ_0^2$ passes through zero, the density reaches its local maximum or passes through its inflection point. The probability's third derivative $d^3P'(Z_0, Z_\xi)/dZ_0^3$ is negative in the former and zero in the latter case. Simultaneous zero values of both second and third derivatives signal, that the density function has passed through an inflection point. This takes place for a certain value of the parameter Z_ξ, which represents either LSB or USB. Homogeneity of the extended sample $\mathcal{Z}'$ is thus maintained if $LSB < Z_\xi < USB$.

The difference between the homogeneity test and the membership test can be explained by the question to be answered by the tests:

Homogeneity test: "Is the given sample $\mathcal{Z}$ homogeneous?"

Membership test: "Is a value Z_ξ a potential member of the given sample $\mathcal{Z}$?" In other words: "Will the homogeneous sample $\mathcal{Z}$ remain homogeneous after extension by Z_ξ"?

Note, that both homogeneity and membership problems are solved on the binary (yes/no) level. Since the estimating global distribution function is uniquely determined by data for each homogeneous data sample, the number N_0 and the bounds LSB and USB are also uniquely determined by the data. Both tests are objective, independent of the subject asking the questions. The outcomes of the tests are thus completely defined by the data.

Problems of this type also have a solution in statistics, however, the results of such statistical tests depend on the (subjective) choice of significance level and on the (subjective) choice of an a priori data model.

The membership problem has both theoretical and practical importance, eg in mathematics it plays a fundamental role related to notion of a set: "is x an element of a set $\mathcal{X}$?" Different answers may give rise to different mathematical concepts:

- In classical (Cantor's) set theory, membership is taken as a primitive notion ("everybody knows the right Y/N answer for any arbitrary x.")
- In contrast, the popular fuzzy set theory is based on the idea, that "everybody knows the value of the membership function—to what degree does the x belong to a fuzzy set $\mathcal{X}$."

In both of these examples the responsibility for the solution of the membership problem and for its consequences lies with the user of the method, on his subjective choice—just as in statistics.

Gnostic tests of homogeneity can be used to establish the membership of a particular value x. Practical applications for sample's bounds could include:

- Given a group of enterprises comparable in the sense, that the sample $\mathcal{R}$, consisting of a set of financial ratios R is homogeneous, within what bounds must the ratio R_x of another firm lie to allow its membership in the group?
- Given a collection of good products with quality parameters Q, which form a homogeneous sample $\mathcal{Q}$; what should be bounds of another product's quality parameter Q_x in order to be accepted as "good?".

The uniqueness and objectivity in the answer to these questions is undoubtedly desirable and the robustness of the EGDF with respect to outliers and peripheral data provides a reliable homogeneity test as well as estimates of a sample's bounds.

15.3.8 The Flexibility of the QLDF and QGDF

The additive composition of the unbounded quantifying irrelevances, together with transformation 15.32 results in natural behavior:

1. The QGDF exhibit limited flexibility like EGDF.
2. There is a best scale parameter and best estimates of the probability domain.
3. The QLDF has an unlimited flexibility like ELDF and can be used for the marginal analysis.
4. Both the QLDF and the QGDF are robust with respect to inliers.

Hence, the QGDF is also unique for a given data sample like EGDF. A comparison of the behavior of quantifying distributions with the EGDF, using the same examples that were considered above shows, that all three probability distributions taken with the same scale parameter coincide so closely in Fig. 15.7, that it is impossible to distinguish between their values. The densities are substantially different only in the vicinity of the bounds of the data support. The outlier (−13.5 in Fig. 15.8) raises the value of both the QLDF and the QGDF without affecting their monotonic character. The fitting errors of the data 1 and 2 are large, while the "peripheral" data 7, 8, 9 and 10 are modeled by the QLDF and the QGDF more precisely. Both probability distributions in Fig. 15.8 and densities in Fig. 15.9 document a similar behavior for the QL- and QG-distributions. The similarity of these distributions is also seen in Fig. 15.10 and 15.11: they are essentially identical in the case of symmetrical data samples.

The example in Fig. 15.8 shows, that both the QL- and the QG-distribution are more rigid than the EGDF—their slope (unlike that of the EGDF) does not change to reveal the presence of the outlier (−13.5) when the scale parameter and bounds of distribution's domain are the same.

15.3.9 The Robustness of the QLDF and QGDF

Figures 15.8 and 15.9 demonstrated the EGDF's *robustness with respect to outliers* or *inner robustness,* however the behavior of the quantifying functions is not the same. In Figures 15.10 and 15.11, both the QLDF and the QGDF ignore the inner cluster and try to model the peripheral triples of data, (especially the first and last ones). Such a behavior can be called *the robustness with respect to inliers* or *the outer robustness*. Instead of the S-form of the EGDF, the QL- and the QG- probability distributions take on a less usual form, which can be called the *reverse S-form*[10].

These two categories of gnostic distribution functions (E- and Q- types) provide an analyst the opportunity to interpret a given sample from two different perspectives, emphasizing either the inner or outer robustness of the analysis. Three questions can be raised in this connections:

1. Is there a need for robustness when analyzing real data?
2. If the response is positive, then are two mutually opposite types of robustness really necessary?
3. If this second answer is in the affirmative, then how should the type of robustness to apply to any arbitrary data set be chosen?

Question 1) has two aspects. The first one is related to data, which strongly deviate from a “main” distribution (either outliers or inliers). Any experienced analyst will answer the first question positively. Using a mass production process as an illustration: in spite of a high degree of automation, a high level of production control, and sophisticated quality assessment, defective products will persist. They must be identified and rejected. As discussed in Chapter 2, economic data include errors, and it is impossible to eliminate them without careful analysis. Even with the use of the most evolved technology, measurement errors cannot be avoided entirely.

The second aspect is connected with the problem of a priori assumptions with respect to statistical models of data. A method based on such assumptions risks, that the real character of data differ from the assumptions. To guard against the use of potentially false a priori assumptions, statisticians have labored for decades over the development of robust methods.

Both of these aspects support the idea, that robustness is necessary. When the problem of dependence of a data processing method on an a priori data model is considered, the task is clear—such dependencies should be minimized if the assumptions cannot be verified. What about robustness with respect to outliers/inliers? The need for both types of robustness can be demonstrated by example: When an inaccurate measurement technique is used to measure a certain constant quantity, the measurements are repeated in the intuitive belief, that results, which fall close to a central (eg average) value, are more reliable than the peripheral ones. This notion is based on the Law of large numbers. Such ideas gave rise to the first robust method of statistics, the so called *trimmed mean*, which has been used for physical measurements for centuries. The procedure is to

[10] However, the outer robustness of the quantifying DFs is related not only to the reverse S-form, it also manifests itself in the case of S-form quantifying distributions.

order the measurements and then to eliminate (trim) a percentage of the results on both ends of the ordered sample. The arithmetical average of the remainder of the sample (the trimmed mean) is then accepted as the best estimate of the true value of the measured quantity. In other words, the data belonging to the central part of the sample are given a full weight, while the peripheral data are awarded a zero weight. Such discrete weighting can be criticized by observing, that all data are composed partly of information and partly of "noise" or measurement error. The latter point of view would augur for a continuous weighting, giving weights to data with respect to their distance from the sample's center—the greater distance, the less weight. Such a weighting scheme is used in gnostics, but instead of being arbitrarily assigned, the weights are precisely determined by application of the theory. While the trimmed mean is (at least at first sight) a simple procedure, the gnostic methodology results in the preservation of more information. From the foregoing, it can be seen, that the long lived use of the trimmed mean shows, that there is a need for **inner** robustness in the structure of estimating methods.

Now consider another class of problems:

1. A very short term time series of the price of a security may be temporarily stationary with a modest volatility component; this could be termed its "normal" behavior. To be able to distinguish a departure from this state—a rise or a fall—would provide an opportunity to either buy or sell the issue on advantageous terms. The ability to quickly recognize a deviation from the normal price pattern would be a valuable trading tool.
2. Monitoring the quality of production is a similar problem. "Normal" results fluctuate around a required "central" value within some given bounds (tolerance). Exceeding these bounds requires an action—at least the removal of the bad product from the flow of normal products.
3. A third example is monitoring vital signs of a patient. A departure from the region of "normal" values requires an intervention.

The similarity of these examples lies in their underlying principle: the (frequently observed) "normal" data are much less important (or useful) for an analyst than those data, which represent a departure from the normal state. The "normal" data are close to the sample's center, while the "interesting" ones are peripheral data. If there were no volatility in the "normal" data, recognition of the change would be simple, but the more volatile the "normal" data are, the more difficult it is to establish a change in their mode of behavior. This means, that for a system, which should initiate any action, the “normal” data are “noise”, while the required information is contained in the outlying data. For such cases, it is natural to apply methods, which suppress the central part of a data sample and emphasize its periphery. In other words, under some conditions **outer** data robustness is also a desirable trait.

The choice then is closely related to the task of the analyst or that of an automated decision process. If the goal is to identify and describe the normal (quasi-stationary) behavior of a stock, characteristics of normal production, or the vital signs of a patient, then the inner robustness is required. If the objective is to design software for an automatic monitor to detect departures from the “normal” situation, then outer robustness is necessary.

There may also be situations, where the goal of analysis has not been established beforehand. An example is in exploratory analysis: the objective here is to explain the data in the best possible way. In such cases, it may be useful to apply both inner and outer robustness and then compare the results. The “right” robustness in such a case is that, which better explains the data. “To better explain data” may sometimes mean “to provide the fit of the data with the least fitting error”, while at other times it may be something else.

Referring once more to Fig. 15.10, consider first the blue EGDF distribution, which has the S-form and reveals the peripheral outliers thus manifesting its inner robustness. It is the best distribution function in the sense of minimizing the ME-criterion. But if a different criterion were to be applied, the situation could change.

Now if the MF-criterion[11] (which is general in the sense of its applicability to an arbitrary system of a priori data weights) is used instead, the three functions that are graphed, each take on a reverse-S shape. The QGDF and the QLDF are essentially on top of each other, while the EGDF (now using the MF criterion) generally follows the same pattern except for the central data; the global maximum of the sum of the fidelities is (0.8694) and it is obtained using the scale parameter $S = 6.31$. This EGDF has the reverse S form and it practically coincides with the QGDF and QLDF. The S-shaped distribution can also be obtained, but only with a much smaller scale parameter ($S = 0.182$). The resulting distribution function will correspond to a local maximum of the MF-criterion, which is only 0.67.

There is no contradiction between the two models. A careful analyst

[11] Maximization of the fidelities’ sum.

would make use of both of them: the more flexible to characterize the clustering of the data, and the more rigid to describe the overall view of the sample. To obtain either version is easy, it is a matter of the starting value of the scale parameter. The optimization algorithm will converge to the local optimum from a small initial S (say, less then 1.5) and to the global optimum from a larger initial S (greater then 1.8). The same multiple interpretation can also be used with the QGDF and QLDF.

15.4 Comparison of Distributions

To summarize the results set out above, it will be instructive to compare the numerical characteristics of the "rigid" distribution functions EGDF and QGDF, which maximize the mean fidelity of the fit (the MF-criterion). The optimal parameters of the distributions are S (the scale parameter) and LB and UB (the lower and upper bounds of the data support).

Para-	**Distribution Function**	
meter	**EGDF**	**QGDF**
S	3.76	4.54
LB	0.062	0.122
UB	10.95	10.88
MF	0.99999	0.99999

Tab. 15.2 Comparison of the maximum fidelity values MF obtained by the rigid distribution functions as applied to the data sample $\mathcal{Z}_{10}$. (Both distribution functions have the same nearly linear character.)

Although each of the functions was computed using a different value for the optimum scale parameter, all of them model this uniformly distributed data with the same (very high) precision (Table 15.2). The major difference stems from the different boundaries for the data support, LB and UB, particularly the lower bound; this different behavior can be seen in Fig. 15.7.

Para-	**Distribution Function**	
meter	**EGDF**	**QGDF**
S	1.15	2.14
LB	-13.5	-164.1
UB	23.55	10.80
MF	0.903	0.865

Tab. 15.3 Comparison of the maximum fidelity values MF reached by the rigid distribution functions applied to the data sample $\mathcal{Z}_{11}$.

This data sample is non-symmetrical (Table 15.3). The EGDF has the S-form (inner robustness). The QGDF manifests outer robustness. The large difference in the estimated bounds is caused by the different character of each of the distribution functions.

The symmetrical data sample $\mathcal{Z}_{11,s}$ may be used to show the multiple interpretation of the S-forms versus reverse S-forms.

Para-	**Distribution Function**	
meter	**EGDF**	**QGDF**
S	0.182	0.245
LB	-10.1	-10.3
UB	10.1	10.3
MF	0.670	0.664

Tab. 15.4A Comparison of the maximum fidelity values MF reached by the rigid distribution functions applied to the symmetrical data sample $\mathcal{Z}_{11,s}$. (Both distributions are of the S-form.)

The "small S" functions EGDF and QGDF violate the basic condition for being distribution functions, they are not everywhere non-decreasing. Their quality measured by the MF-criterion (although locally maximized) is low. This leads to the conclusion, that these S-form interpretations (Table 15.4A) cannot be accepted. It is useful to look for better results in a region of greater values of scale parameters.

Para-	**Distribution Function**	
meter	**EGDF**	**QGDF**
S	6.31	7.25
LB	-10.3	-10.3
UB	10.4	10.4
MF	0.869	0.870

Tab. 15.4B Comparison of the maximum fidelity values (MF) reached by the rigid distribution functions applied to the symmetrical data sample $\mathcal{Z}_{11,s}$. (All three distributions are of the reverse S-form.)

Both distributions cited in Table 15.4B seem to have the same S-form and to provide the same quality of fit. However, it would be incorrect to conclude, that they are identical. The EGDF is flexible enough to reveal the peripheral outliers (although not as completely malleable as the ELDF, which can approach the empirical function as closely as desired), while the QGDF "ignore" these data completely. These differences can be seen by exploring the behavior of the functions using a more refined numerical procedure.

However, the most important lesson is, that there can be several (local) solutions to the optimization task of parameters S, LB and UB. Some of the solutions can even be unusable and certain efforts may be necessary to find the global optimum.

The examples and comparisons of the distribution functions (EGDF and QGDF) optimized by the choice of the scale parameter show, that the estimating global distribution function (EGDF) has unique features, and that it has no substitute. When the iterative calculation of this distribution is initiated using a small scale parameter, a reliable test of the homogeneity of the data sample can be performed, and the robustness is of the inner type. If the iterative optimization process of the QGDF is started from a sufficiently large scale parameter, it manifests robustness of the outer type.

15.5 Cross-section Filtering

It is quite common in the field of data treatment to associate the notion of a filter with processing time series. The task is to make use of the whole signal to extract the desirable components and to suppress the residual "noise." The process consists of making use of a known regularity of the desired signal. When using gnostic distribution functions (which represent a model of the data's regularities), reliance is placed on both the theory (which shows how to create these functions) and experience (in manipulating the data, which make up the distributions). If data are related to the same time frame, *cross-section analysis* is used. An example of such a situation could be a data sample composed of certain financial ratios taken from financial statements of a set of firms from the same period. The distribution of such data reflects regularities relevant to the financial situation of the whole group at that point in time. Distribution functions—as a collective experience—can therefore be used to revise individual data since these, as a rule, do not exactly correspond to the sample's smooth distribution function. Four ways to discretely characterize data distribution were examined above (the empirical distribution function EDF, collections of KS-points and ME-points and the WEDF). Each of these systems of points represents a version of a discrete distribution function (DDF), which can be directly calculated from the data. Each of the four gnostic distribution functions ($**$DF) is a tool for smoothing the DDF. Discrepancies between the discrete and smooth representations of data may be used to suppress uncertainties in individual data, ie for cross-sectional filtering.

Definition 17: Let $\mathcal{Z}$ be a data sample consisting of data $Z_1, \ldots, Z_N$.
Let the DDF be the *discrete distribution function*, points (D_k) of which are evaluated from data $\mathcal{Z}$ using formulae 15.1, 15.3, 15.7 or 15.10.
Let $**$DF be any of the gnostic distribution functions ELDF, EGDF, QLDF or QGDF of the sample $\mathcal{Z}$ and let $P(Z_k)$ be the probability estimate at the point Z_k obtained as the value of this distribution function.
Let $\widetilde{Z_k}$ be the estimate of the true value of Z_k such that

$$P(\widetilde{Z_k}) = D_k \tag{15.44}$$

holds.

This procedure for obtaining estimates $\widetilde{Z_k}$ for $k = 1, \ldots, N$ is defined as *cross-sectional filtering.*

The ideas behind this notion of filtering are, that

1. the value of $\widetilde{Z_k}$ (obtained by inverting the distribution function for probability D_k) may be closer to the true value of Z_k, because it better corresponds to the direct representation of the data by the DDF than Z_k,
2. the $**$DF better describes the regularity reflected by the data than the DDF, because—due to its smooth character—it integrates the influence of all the cross-section data, compensates for uncertainties of individual data and awards to them optimal weights according to the ($**$DF's inner, outer or local) robustness.

It is thus obvious, that this cross-section filtering is **robust** and the specific kind of robustness may be chosen by the selection of the $**$DF.

15.6 Homo- and heteroscedascity

Homoscedasticity (heteroscedascity) refers to the circumstance in which the variability of a variable is equal (unequal) across the range of values of a second variable that predicts it. The variability of a kernel estimate is determined by the kernel's width. The kernel's integral is normalized to 1. Both width and amplitude of a kernel are determined by the scale parameter. The "equal" variability of a number of kernels is achieved when scale parameters of all kernels coincide. The heteroscedastic data case corresponds to unequal scale parameters of the kernels. There was an assumption of a constant scale parameter when different probability densities were considered above. However, this does not limit the application of the formulae to heteroscedastic cases, because a constant scale parameter is used for the entire kernel's form. The probability and density of heteroscedastic data will be modeled by kernels having different scale parameters.

15.7 Summary

Unlike parameterized families of statistical distribution functions, the gnostic distributions have no a priori prescribed form. However, this does not mean, that they are not parameterized; the primary parameters of gnostic distribution functions are data. The most frequently used secondary parameters are the scale parameter, and the lower and the upper bounds of the data support. These are 'secondary' in the sense, that they are estimated from data. The data thus play the full role of determining both the distribution functions and their densities. Again, this is the real thrust behind the gnostic 'credo' of "Let data speak for themselves."

To estimate the secondary parameters of distribution functions, it is necessary to solve the "goodness-of-fit" problem—to find parameters, which ensure the best correspondence of the gnostic distribution functions to the data sample as it is represented by its discrete distribution function.

Several types of discrete distribution functions for a primary representation of data were examined and several criterion functions were found suitable to obtain the best goodness-of-fit. The choice of criterion functions can be made from several gnostic criteria, which lead to robustness of the fit.

Four gnostic models of distribution functions were created from a generalization of the notion of the gnostic probability of an individual datum. The main element of these functions is the irrelevance of a data sample, which is obtained by the composition law (Axiom 2). There are two types of irrelevance in gnostics; these evolve respectively from the estimating and quantifying processes. Two composition rules result from Axiom 2 and these create the weighted and normalized weighted mean of the irrelevances. The local distribution functions use the weighted irrelevance, while the global functions are a result of using the normalized weights. The four versions of gnostic distribution functions are the EL (estimating and local, ELDF), EG (estimating & global, EGDF), QL (quantifying & local, QLDF) and the QG (quantifying & global, QGDF).

Formulae for these distribution functions reveal some interesting characteristics: the local ones (ELDF and QLDF) differ from the global gnostic distribution functions by its unlimited flexibility, which is controlled by the scale parameter. This feature can be used for marginal (univariate) cluster analysis to "zoom in" and get a detailed look at the data structure.

The global distribution functions (EGDF and QGDF) have limited flexibility and they are unique for each data sample in the sense, that the best fit can only be obtained by using an optimizing value for the scale parameter. These functions are based on the assumption, that the data sample is homogeneous (that its density has only one maximum). In the case of a non-homogeneous data sample there may be more than one density maximum or the density may even be negative. Given this possibility, the EGDF and QGDF are particularly suitable for conducting a reliable test for data homogeneity. Both global distributions are robust; the EGDF is robust with respect to outliers, the QGDF with respect to inliers. This innate robustness of gnostic distribution functions is a welcome feature especially in the treatment of small samples of widely spread data.

The following table summarizes the major characteristics of each of the distribution function:

	Distribution Function			
Attribute	**ELDF**	**EGDF**	**QLDF**	**QGDF**
Type of DF	Estimating	Estimating	Quantifying	Quantifying
Kind of DF	Local	Global	Local	Global
Composition of Kernels	WA	WAN	WA	WAN
Bounds of Data Support	Arbitrary	Optimized	Arbitrary	Optimized
Scale Parameter	Arbitrary	Optimized	Arbitrary	Optimized
Robustness	Local	Inner	Local	Outer
Flexibility	High	Low	High	Low
Formula for Probability	(15.25)	(15.29)	(15.33)	(15.37)
Formula for Density	(15.26)	(15.30)	(15.34)	(15.38)

Tab. 15.5 Comparison of the Features of the Four Distribution Functions.

The composition law for gnostic kernels in Tab. 15.5 is WA (arithmetic mean of gnostic kernels weighted by a priori weights) or WAN (normalized weighted arithmetic mean of gnostic kernels).

The suitability of gnostic distribution functions to different applications is summarized in Tab. 15.6.

Task	**Symbol**	**Suitable Distribution Function**			
		ELDF	**EGDF**	**QLDF**	**QGDF**
Estimate Probability	$P(Z_0)$	Y	Y	Y	Y
Estimate Density	$\frac{P}{dZ_0}$	Y	Y	Y	Y
Estimate a Quantile	$Z_0(P)$	Y	Y	Y	Y
Est. Location Parameter	LP	Y	Y	Y	Y
Est. Global Scale Parameter	S_G	N	Y	N	Y
Est. Data Support Bounds	LB, UB	N	Y	N	Y
Est. Sample's Boundaries	SB	N	Y	N	Y
Interval Analysis	IA	Y	Y	Y	Y
Cluster Analysis	CA	Y	N	Y	N
Cross-section Filtering	CSF	Y	Y	Y	Y
Test for Homogeneity	TH	N	Y	N	Y
Test for Membership	TM	N	Y	N	Y

Tab. 15.6 Applicability of the Four Distribution Functions. Y ... "Yes," N ... "No."

Chapter 16

Parameters of Distribution Functions

16.1 Parameters of Non-parametric Estimates?

Reality is richer than the words, which attempt to describe the experience. There are many more objects than words to characterize them, which is why one word can be used to define a myriad of things. The casual use of words in this way is imprudent in scientific discourse, where words purport to precisely define a process or an activity. The synonymity of a definition, which represents various activities can be confusing to the uninitiated, but this very often occurs over the period, over which a scientific discipline develops. What *statistics* means has been relatively clear for some decades; the sense of the word corresponds to a definition such as that given in [110]:

> *Statistics* is a collection of methods for planning experiments, obtaining data, and then analyzing, interpreting, and drawing conclusions based on the data.

There is no reason to doubt the validity of this definition as it applies to modern statistical applications, but the processes cited do not uniquely pertain to statistics. The definition is too broad to delimit the statistical framework, while—at the same time—distinguishing it from many other recently developed approaches to the same tasks. One can agree, that obtaining data is one of the tasks of statistics, but it is also applicable to many other methodologies: eg measurement theory describes in detail the process of “obtaining data” using mathematics (but not statistics), and derives conditions, under which this process is consistent. Neither statistics or measurement theory are subsets of each other; nor does statistics use the results of measurement theory to support statistical definitions or axioms. A rich survey of methods for planning experiments, analyzing and interpreting data and for drawing conclusions based on data include a plurality of theories, among others: Fuzzy Sets Theory, Rough Sets, Alternative Sets, Theory of Evidence, Belief Networks, Possibility Theory, Chaos Theory, Fractal Geometry, Non-standard Logics, Non-monotonic Logics, Default Reasoning, Temporal Reasoning, Approximate Reasoning, Multivalued Logics, Belief Updating, Tree Structures, GUHA Method[1], Knowledge Acquisition and Representation, Machine Learning, Inductive Methods, Neural Networks, Databases, Information Retrieval, Data Mining, Uncertainty in Cognition, Expert Systems and—last, but not least—Gnostics. All of these undertake the tasks set out in the above definition, but this does not mean, that they are all a part, a "chapter," of statistics, because these approaches are encountered as **non-statistical** methods. It has been said, that the present state of the art in this field is a "competition of paradigms," but, as yet no one has any idea, as to which of the methodologies will become the future "medalists." Therefore, it is difficult to say today, which of these approaches is more important[2]. In all of these procedures one encounters the notion of *chance*, when the task is to estimate an unknown element. There is not much doubt as to what this means in terms of ordinary human activity; however it typically demonstrates the labeling of different things with the same word. In expressions such as "a chance of winning," "a chance to explain," "a chance to take," "to meet by chance," "no chance," "by any chance," "to chance it," and in many other uses, the meaning of the word differs. The scientific (understand: statistical) notion of this word is closely related to the Law of Large Numbers, which behaves with a surprising consistency among events, that seem to occur "by real chance". The non-statistical "competitors" of statistics ordinarily avoid this—on the one hand fuzzy and on the other hand "reserved for statistics"—word and use another term, eg *uncertainty.* A similar nomenclature problem exists with parametric and non-parametric estimates. In the framework of statistics, a parameter is a numerical measurement, which describes a characteristic **of a population**, while a statistic is a numerical measurement, which describes some characteristic **of a sample** (a subset of the population). *Parametric methods* of statistics can usually be applied under conditions, where some fairly strict requirements (related to the population) are met. One of these is typically that the sample data come from a normally distributed population:

[1] GUHA stands for General Unary Hypotheses Automaton—a powerful method developed by Czech scientists J. Hájek and T. Havránek used to discover the logical interrelations in large masses of data.

[2] The major portion of the above list was taken from the call for papers of the 7-th conference on "Information Processing and Management of Uncertainty in Knowledge-Based Systems" held in Paris on July 6-10, 1998.

it is then reasonable to use as estimates of the population's parameters the arithmetic mean and the standard deviation. The parametric statistical methods can be called "distribution based", because a narrower definition for a statistical parameter would become something like "the parameter of a population's distribution function of a specific type." In contrast to the above, statistical non-parametric methods do not assume a particular distribution for a population and they are sometimes called *distribution-free methods*. The six most frequently used non-parametric statistical tests described in [110] are based on the ranks of ordered data samples or on patterns of sequences. The name "distribution-free" is not always justified, because to derive critical values for some tests of this type, it is necessary to apply the binomial distribution. "Non-parametric" cannot be strictly interpreted as "having no parameter", because all the statistics derived from these tests are parametrized at least by the size of the samples. Another example is the Parzen's kernel estimating method been discussed in previous chapters. It is suitable for the estimation of a broad class of distribution functions independent of their parameters. However, kernels do have parameters, which determine their form, width, height etc. This means, that the use of parameters by non-parametric methods is initially a confusing notion.

Gnostics is not based on the idea of a population. Instead, as has been discussed previously, it deals with data samples as given objects. The samples may be extended, because they are elements of a group of data, however, each extension is the subject of an investigation as to its impact on the characteristics of the sample, to test if—and to what degree—it is indeed a member of the sample. The notion of parameters is therefore related not to an assumed population, but to the data, that is being used. Gnostic distribution functions are estimates, parametrized primarily by data (much like estimates generated by some other methods), but they also employ several other parameters (the scale parameter S, the bounds of data support LB and UB, as well as the location parameter, which is dealt with in the next section). All these characteristics **are** estimated from the data. In this sense, gnostic distribution functions are parametrized only by data, they are not only "distribution-free" (in the sense of being based on a priori assumed distribution functions); they also are "parameter-free" in that they yield all the information necessary for estimation from the data itself. Despite this definition, the notion of parameters of distribution functions (such as S, LB, UB and others) will be used to distinguish different features of the estimated distributions. As such, these also will be "parameters of non-parametric methods."

16.2 Scale Parameters

The discussions in the previous chapters should have instilled in the reader a sense of the importance in the role played by the scale parameter. For the EGDF, QGDF and QLDF a unique scale parameter exists, which ensures the best possible goodness-of-fit. In the case of the ELDF the scale parameter determines the "resolution power" of the representation of the data sample's structure. Moreover, it also determines the width of the toleration interval of typical data intervals as used in interval analysis (to be discussed later in the chapter). It will be shown shortly, that the value of the scale parameter is closely connected to the degree of robustness of both uni- and multivariate gnostic models and to robust filters and predictors. To delve further in the idea, "Let data speak for themselves," these important parameters must be estimated from data "in the best way". However, different applications may require a different notion as to what the best scale parameter is.

16.2.1 Global Scale Parameters

The unique scale parameters, which optimize the distribution functions EGDF and QGDF (jointly denoted as **DF) will be called *global scale parameters*, because these distributions provide an overall view of the data of a sample. To do this, the global scale parameters satisfy the condition of the best fit. The notion of the best fit depends, on which type of discrete distribution function is to be fitted. The three approaches, the KS-points, the ME-points and the WEDF, were described in Chapter 15 along with several criterion functions. Three specific versions of the global scale parameter will be considered here denoted $S_{G,KS}$, $S_{G,ME}$ and $S_{G,MF}$.

Global Scale Parameter $S_{G,KS}$

In a simple case, when the bounds of data support are known, the procedure for estimating $S_{G,KS}$ as described in Chapter 15 can be written as a minimax task. Denoting

$$er_{m-} = **DF(S, Z_m) - EDF(Z_m)_- \tag{16.1}$$

and

$$er_{m+} = **DF(S, Z_m) - EDF(Z_m)_+, \tag{16.2}$$

where $Z_1, ..., Z_N$ are data, and where $**$ specifies the function (EG, QG or QL). Expression $EDF(Z_m)_-$ is the left hand limiting value of the empirical distribution function EDF at the point Z_m, while $EDF(Z_m)_+$ is its limit on the right side. Quantities er_{m-} and er_{m+} are thus the fitting errors. The task is

$$S_{G,KS} = \arg(\min_S(\max_m(\max(|er_{m-}|, |er_{m+}|)))). \tag{16.3}$$

This scale parameter has one advantage and several disadvantages. The advantage is, that it provides a clear statistical sense of the distribution's optimality: the distribution function minimizes the Kolmogorov-Smirnov statistic, which is familiar to statisticians. A statistician may need to be persuaded from time to time, that a particular gnostic distribution function is good enough to be used. A useful argument in this case is the value of the Kolmogorov-Smirnov statistic minimized by $S_{G,KS}$. Its value is frequently so small, that the KS-test rejects the hypothesis of a bad fit with a high level of significance. The main disadvantage of the $S_{G,KS}$ is, that the procedure for solving 16.3 requires the extremization of a non-smooth function. There are suitable algorithms for this, but they are neither very simple nor fast. In the more general and useful situation of unknown bounds, additional numerical problems may arise.

The $S_{G,KS}$ scale parameter was applied to the EGDF distribution function to compute all the estimates of gnostic location parameters for the comparison of robust methods cited in [62].

Global Scale Parameter $S_{G,ME}$

This type of global scale parameter is associated with the maximum entropy fit. Its value is determined by solving the maximization problem

$$S_{G,ME} = \arg(\max_S(F_1 * \ln(F_1) + \sum_{k=2}^{N}(F_k - F_{k-1}) * \ln(F_k - F_{k-1}))), \tag{16.4}$$

where

$$F_k = **DF(S, Z_k) \tag{16.5}$$

is the value of the distribution function $**DF$ at the data point Z_k. In this case the scale parameter, $S_{G,ME}$, is computed by finding the extreme values of a smooth and differentiable function, but its disadvantage is, that it is not suitable for data, which do not all have prior equal weights.

Global Scale Parameter $S_{G,MF}$

It is not difficult to extend the problem to the simultaneous estimation of $S_{G,MF}$ along with the bounds of the data support LB and UB. This global scale parameter maximizes the sum of the fidelities and is denoted $S_{G,MF}$; its value can be obtained by solving the equation

$$S_{G,MF} = \arg\left(\max_S \sum_{k=1}^{N} \frac{2}{(F_k/E_k)^{2/S} + (E_k/F_k))^{2/S}}\right), \tag{16.6}$$

where the E_k are values of the weighted empirical distribution function (15.9 and 15.10) and F_k is again 16.5. This scale parameter can also be simultaneously estimated with the bounds of the data support. As already noted, the values of the weighted empirical distribution function E are the same as the ME-points, when all prior data weights are equal, so that for this special case the scale parameter $S_{G,MF}$ coincides with $S_{G,ME}$. Because of its universality, the scale parameter 16.6 is a basic type of global scale parameter suitable for most applications.

16.2.2 Global Scale Parameter S_{L1}

A useful version of the global scale parameter can be obtained by solving the equation

$$S_{L1} = \arg(\min_S \sum_{k=1}^{N} \mid F_k - E_k \mid), \tag{16.7}$$

where the same symbols are used as in 16.6, and where the symbol $L1$ is used to recall the name of 'L1-approximation'.

16.2.3 Local Scale Parameter

When non-homogeneous data samples (those with a complex structure having several clusters of data) are to be analyzed, it cannot be automatically assumed, that the most suitable model is a single scale parameter, which is constant for all clusters. It often occurs, that such samples are a mixture of several subsamples, each of which represents a different object or process. Individual clusters may have a different width due to a different spread of data. The composite sample may thus need a “local” parameter dependent on a specific point on the data support rather than a single scale parameter, which is constant over the whole sample. The value of such a parameter characterizes the spread of data over the neighborhood of the point. Such a suitable local scale parameter results from the following theorem, which then characterize a region of the ELDF.

Theorem 16: Let $\mathcal{Z}_N$ be a sample of multiplicative data $(Z_1, ..., Z_N)$ defined over the infinite data support R_+ and let Z_{LP} be a location parameter of the sample not necessarily equal to the ideal value Z_0. Let $S_L(\mathcal{Z}_N, Z_{LP})$ be a local scale parameter of the sample. Now let $E_j(Z_k/Z_{LP}) = f_j(Z_k/Z_{LP}) - 1$ be the change in entropy of quantification (10.26) and $f_j(Z_k/Z_{LP})$ the corresponding Q-weight (9.10) for $c^2 = 1$) written as

$$f_j(Z_k/Z_{LP}) = \frac{(Z_k/Z_{LP})^2 + (Z_{LP}/Z_k)^2}{2} \quad (k = 1, ..., N). \tag{16.8}$$

Let

$$g(\mathcal{Z}_N, z, Z_{LP}, S) = \frac{4}{(q^{1/S} + q^{-1/S})zS}, \tag{16.9}$$

where $q = (\frac{z}{Z_{LP}})^2$, and where S is an abbreviation for the scale parameter $S_L(\mathcal{Z}_N, Z_{LP})$. Let

$$IE_j = \int_0^\infty E_j(z/Z_{LP})g(*)dz, \tag{16.10}$$

where $g(*)$ is the density 16.9).

Then:

A There exists exactly one scale parameter $S_L(\mathcal{Z}_N, Z_{LP}) < 2$ satisfying the condition

$$\frac{\Sigma_{k=1}^N E_j(Z_k/Z_{LP})}{N} = IE_j. \tag{16.11}$$

B This scale parameter can be obtained as the solution of the equation

$$\frac{\pi S/2}{\sin(\pi S/2)} = \frac{\Sigma_{k=1}^N f_j(Z_k/Z_{LP})}{N}. \tag{16.12}$$

Proof of Theorem 16: In order to prove 16.11 for the discrete quantifying weight it is sufficient to show, that the means of $f_j(Z_k/Z_{LP})$ (16.8) and of its continuous version $f_j(z/Z_{LP}) = ((z/Z_{LP})^2 + (Z_{LP}/z)^2)/2$ are equal (the means of the constant terms, which equal 1 on both the left and right sides of 16.11 cancel). Taking into account, that $dq = 2\frac{z}{Z_{LP}^2}dz$, one can write the integral

$$\int_0^\infty \left(\frac{z}{Z_{LP}}\right)^2 g(z)dz = \frac{2}{S}\int_0^\infty \frac{dq}{(q^{1/S} + q^{-1/S})^2}, \tag{16.13}$$

where $g(z)$ is the density (16.9). The latter integral is a special case of the integral

$$\int_0^\infty \frac{x^{t-1}}{(1+x^r)^2}dx = \frac{1}{r}\frac{(t-r)\pi/r}{\sin((t-r)\pi/r)} \tag{16.14}$$

(known from the literature [29]), which exists for $t < 2r$. Integral 16.14 therefore also exists for $S < 2$ and has the value shown:

$$\int_0^\infty \left(\frac{z}{Z_{LP}}\right)^2 g(z)dz = 2\frac{\pi S/2}{\sin(\pi S/2)}. \tag{16.15}$$

The differential of the reciprocal value $q' = 1/q$ is $-\frac{1}{q^2}dq$, so that

$$\int_0^\infty \left(\frac{Z_{LP}}{z}\right)^2 g(z)dz = -\frac{2}{S}\int_\infty^0 \frac{dq'}{(q'^{1/S} + q'^{-1/S})^2}. \tag{16.16}$$

Exchanging the integral's bounds in 16.16 to get the same direction for the integration path as in 16.13, summing the two equations and adding 1 results in IE_j (16.10.)

This expression should equal the arithmetical mean of the entropy changes of the data in accordance with condition 16.11. After substitution of 16.15 and its equivalent 16.16 into the continuous weight $f_j(z/Z_{LP})$, statement **B** 16.12 is obtained. The left hand side of 16.12 increases monotonically from 1 to infinity, when parameter S increases from 0 to 2. The right hand side of this equation does not depend on S and may take values between 1 and infinity depending on the data and the value of the location parameter Z_{LP}. Hence, given a data set, there exists exactly one $S(Z_{LP})$ for each value of the parameter Z_{LP}. Therefore statement **A**.

The idea expressed by condition 16.11 deserves further interpretation. The function $g(*)$, (16.9), is the probability density of the estimating local distribution function (ELDF). Integral IE_j (16.10) is therefore the integrated mean of the continuous quantifying entropy $E_j(z/Z_{LP})$, which is one estimate of the mean quantifying change in the entropy of the data sample. This estimate is dependent on the scale parameter S as is the density. A second estimate of the entropy's change is the arithmetical mean of the discrete entropy changes evaluated without prior knowledge of the scale parameter. Condition 16.11 requires, that both estimates be equal. The proper value for the scale parameter is then calculated by applying 16.12. The mean quantifying entropy change is robust with respect to inliers (but is sensitive to outliers). Estimates of the scale parameter obtained from 16.12 will therefore retain these same characteristics. However, obtaining an estimation procedure robust to outliers is not difficult; it is sufficient to recall, that equality $f_i = 1/f_j$ holds between the estimating and quantifying weights. The scale parameter estimate robust to outliers results from the solution of S of the equation

$$\frac{\sin(\pi S/2)}{\pi S/2} = \frac{\Sigma_{k=1}^{N} f_i(Z_k/Z_{LP})}{N}. \tag{16.17}$$

16.2.4 Scale Parameter for Required Fidelity

The global scale parameter[3] can only be used with distribution functions EGDF, QGDF and QLDF. Applying a local scale parameter instead would not make good sense, because there is no freedom in the choice of scale parameters for these distributions; they are to be used with the best—global—scale parameter. Because the local scale parameter is dependent on the spread of data in the neighborhood of a location parameter Z_{LP}, its local character is useful in some applications of the ELDF, but it may be undesirable in other cases. Under certain conditions, it may become necessary to find the value of a scale parameter, which will provide a given quality of fit for the whole data sample. A suitable measurement for such a condition is the mean fidelity of the fit

$$QF = \overline{f_i(EL(\mathcal{Z}_N, Z_k, S))/E_{MF,k}}_N, \tag{16.18}$$

where $E_{MF,k}$ is the k-th value of the weighted distribution function (15.9 through 15.11) of the MF-fit, and $EL(*)$ is the value of the ELDF of the data sample $\mathcal{Z}_N$ at the point Z_k. While formula 16.18 is used to evaluate

[3] Note, that using the adjectives 'global' or 'local' to describe distribution functions does not necessarily imply, that the same type scale parameter (global/local) is the parameter, that should be estimated.

the quality of the MF-fit, when the scale parameter S is given; in cases, where a specified quality for the MF-fit (QF) is desired, it can also be solved for S to find the scale parameter, which will provide the required quality. Such a scale parameter is denoted S_{RF}. There are interesting applications for this scale parameter. In marginal cluster analysis it may help in deciding, which level of resolution power should be chosen (or how many separate clusters should be revealed). The resolution power may be "normalized" by always requiring the same quality QF for the MF-fit of all the samples to be analyzed. When the similarity of samples is being tested by interval analysis (to be discussed in section 16.4), the samples' intervals can be made comparable in this manner.

16.2.5 Variable Scale Parameter

There is, of course, no reason to expect, that all the data in a sample will have a constant spread. In statistics, this condition is known as *heteroscedasticity*. Heteroscedastic data are frequently encountered in economic analysis, especially when the cross-section structure of groups of non-homogeneous objects is being examined. It is also true, that time series data cannot be automatically taken as having a time-independent spread. Gnostics provides tools to overcome this difficulty. The statistical notion of heteroscedasticity is based on the notion of variance. As previously explained, this measure of data spread is acceptable in gnostics only in the case of small data errors, as a limit to the mean data weights (14.54), the mean data entropy (14.56), mean information change (14.57) or mean squared irrelevance (14.58) under decreasing amounts of uncertainty. All of the cited formulae demonstrate the role of the scale parameter S. Moreover, the first derivatives of all four gnostic distribution functions (density functions) are proportional to the reciprocal value of the scale parameter (see 15.26, 15.30, 15.34 and 15.38), and the second derivative is proportional to $1/S^2$. Recall, that the second derivative of a function is a measure of its curvature. The local curvature of a gnostic distribution function is thus specified by the local scale parameter, and vice versa: the local scale parameter is determined by the local curvature of the distribution function. A low $S \Leftrightarrow$ a high curvature and low data spread, large $S \Leftrightarrow$ a flat form for the distribution function and a large spread. The local scale parameter can be estimated by solving 16.17 for the data sample being considered. Solutions can be obtained for different values of the parameter Z_{LP}, ie at different points of the data support. The greater the change in the local

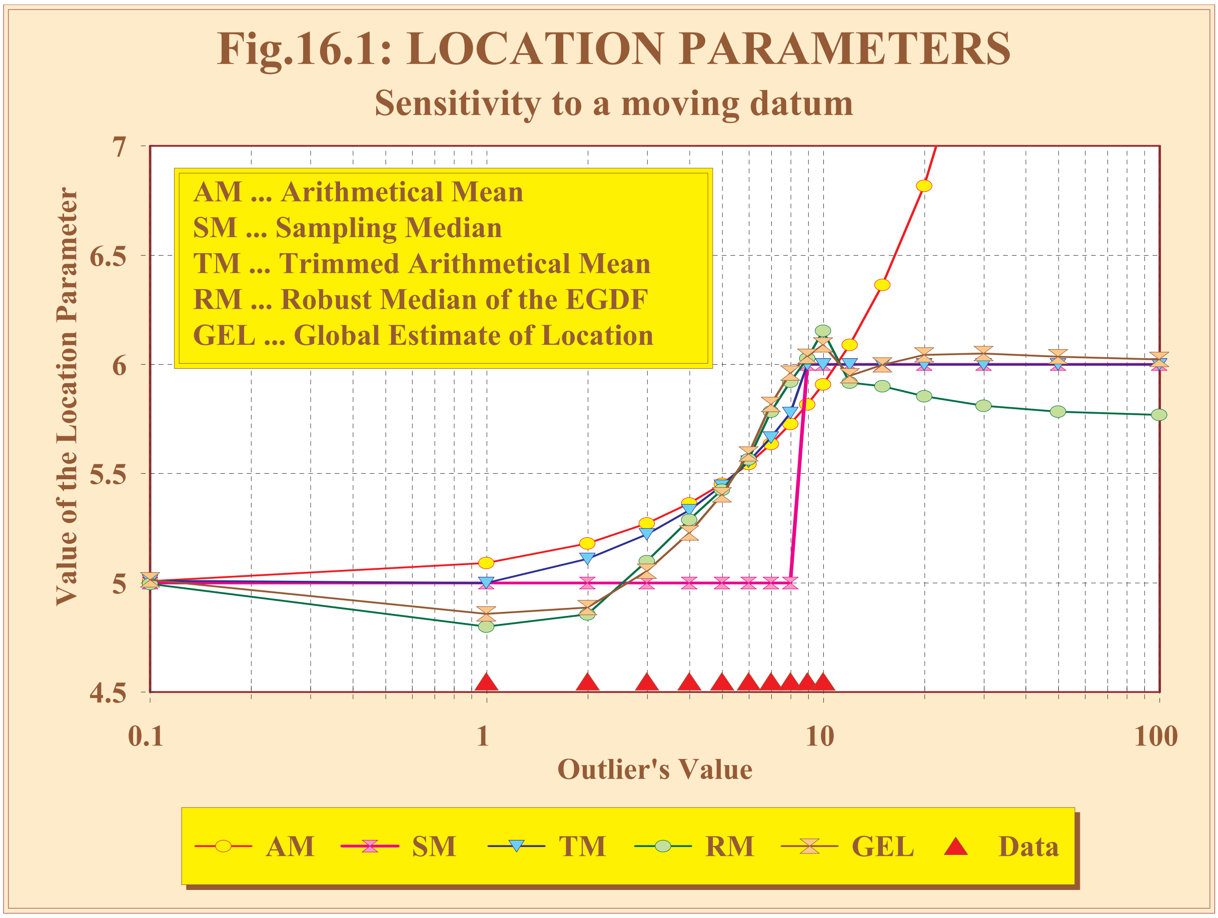
Ss, the larger the change in the data spread. This effect can be viewed more vividly by using the idea of the kernel estimation of densities and distributions dealt with in Chapter 11. Both the statistical kernel 11.6 and the gnostic kernel 11.7 are functions of the scale parameter S, which determines the width and height of the kernel: a large $S \Leftrightarrow$ a broad, flat and low kernel, a small $S \Leftrightarrow$ a narrow, sharp and high kernel. The former case is that of a strong data uncertainty (large spread), the latter corresponds to a small spread (better precision of data). The distribution functions and densities are obtained by superposition of kernels. Heteroscedasticity thus leads to changes in scale parameters and vice-versa. A practical observation related to the algorithmic use of the variable local scale parameter obtained by 16.17 is, that this equation does not warrant using the best fit to data, it gives only a relative profile $S_L(Z_{LP})$. To optimize the fit, it is necessary to introduce another parameter S_0, the value of which in $S_0 * S_L(Z_{LP})$ minimizes the fitting errors. Kernels can also be used in estimating global scale parameters in view of the fact, that the distribution function consists of kernels attached to individual data. In a heteroscedastic case, a scale parameter, which is a function of the ideal value Z_0 could be used rather than a constant (unknown) S for all kernels included in the optimization process. An example of such a function might be $S = S_0 \exp(\sigma Z_0)$. Both constants S_0 and σ are unknown, but they can be estimated using the extremization condition of the best fit. This simple function embraces several types of spreads: the homoscedastic case ($\sigma = 0$), decreasing and increasing linear cases (a small $|\sigma|$) and exponential changes (an arbitrary real number σ). The type of function, that applies in each case, can be chosen by inspection of the data fitting errors.

16.3 Location Parameters

The location parameter is the numerical characteristic of a data sample, which provides information, as to where the data are placed on the data support. There are many types of location parameters; the most frequently used in statistics are eg the arithmetic or geometric mean and the quantiles directly obtained from an ordered data sample (median, quartiles and others.) The minimum and maximum data value of a sample also indicates the breadth of the data. Analogously, the given or estimated bounds of the data support may also play this role. When a robust distribution function for a data sample is available, other location parameters such as robustly estimated quantiles (related to their probabilities) can be used. Useful location parameters such as quantiles, for which the density function reaches its maximum may also be derived from this function, ie the most frequently occurring value of the sample's data (the mode). Two such kinds of parameters will be distinguished, the *GEL (Global Estimate of Location)* and the *LEL (Local Estimate of Location)*. The former parameter is associated with the three versions of the gnostic distribution function (EGDF, QGDF and QLDF), for which a unique "best" scale parameter exists. The latter location parameter can be derived from the ELDF. It is important to emphasize, that the "location of the maximum density" in all these cases is informative only if it is specified, to which data support it relates.

16.3.1 The GEL vs. Traditional Location Parameters

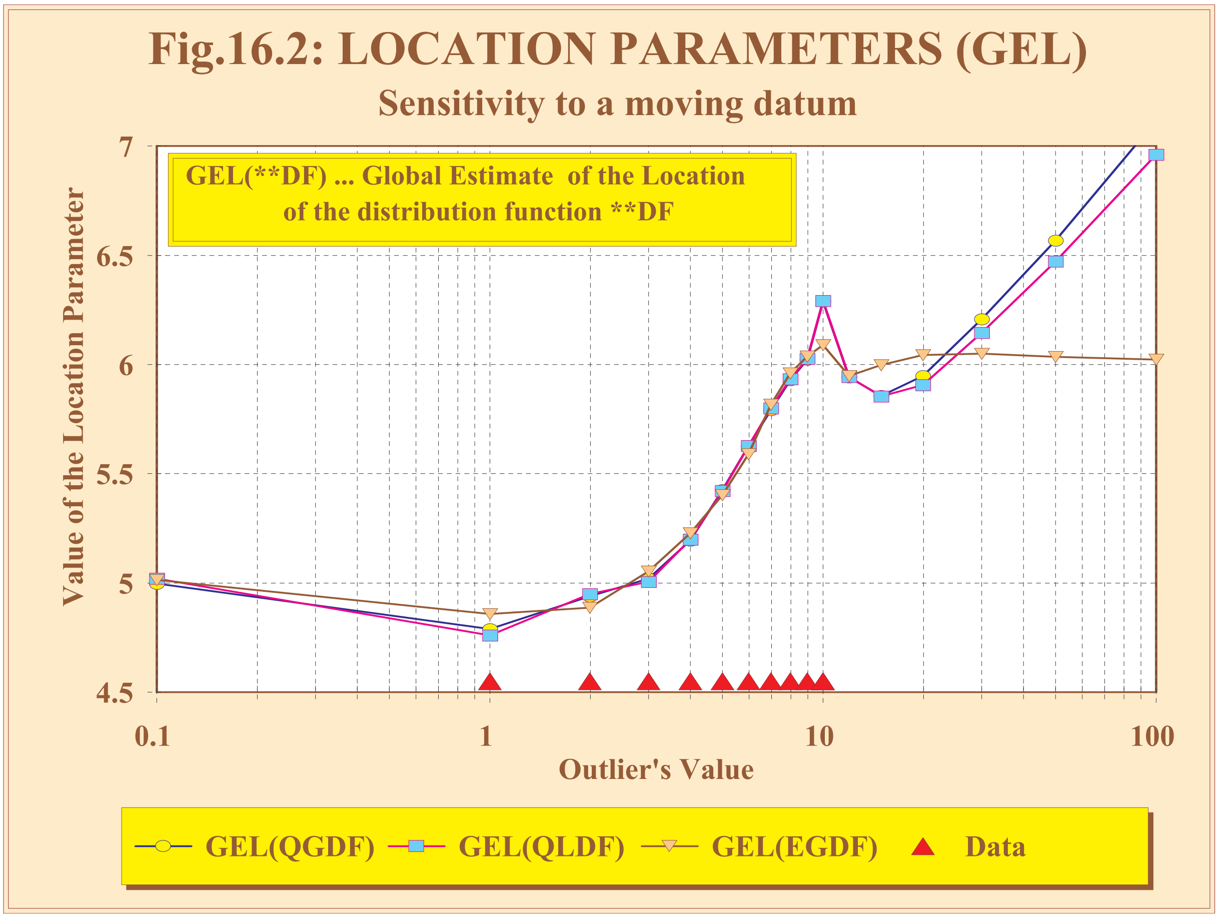
The behavior of the different location parameters can be illustrated by examples using the simple data sample $\mathcal{Z}_{10}$ of uniformly distributed data 1, 2, 3, ...,10 (Fig. 16.1).



These data are fixed, while an 11-th data element is free. Its value changes as it rises from a negligible value to infinity. When it is on either side of the range of the fixed data, the free value is an outlier. The red line shows the effect of these changes on the arithmetical mean (AM) of the 11 data taken together. When the outlier is very small, the AM approaches the value 55/11=5. An unlimited increase in the single outlier results in the unlimited increase of the location parameter. This effect is the well-known *unrobustness* of the arithmetical mean. The median, evaluated directly from the data, (the sampling median SM) shown by the magenta line is constant (5) until the point, at which the free datum reaches the highest value in the sample (10). At this point and for all larger values of the free datum the SM is again constant and equals 6. While the AM is oversensitive to large outlier values, the SM exhibits the opposite tendency and is completely insensitive to changes in the value of the free 11-th datum over broad intervals. The trimmed mean TM (blue line), obtained by omitting the smallest and the largest data elements, is constant outside the data range just as the SM. Within the range, it gradually increases from 5 to 6. The robust median (RM, green line, the quantile for which the EGDF reaches 0.5) and the global location parameter GEL (brown line) exhibit a surprising behavior: they both decrease with an increase in the outlier from a very small value to that of the smallest datum (1). This effect may be examined from at least two different points of view:

1. It may simply be accepted as the outcome of a mathematical theory, which results in maximizing the information contained in the data, or
2. the analyst may try to interpret, what it really means.

The former approach needs no further comment. An attempt to explain the latter can take a more intuitive form. With only the data $\mathcal{Z}_{10}$, the estimate of their “mean” value of 5.5 can be accepted. However, after it is learned, that there is an additional 11-th datum, with a value of less than 1, the estimate is revised and a lower “mean” is computed. There may also exist a contrary effect: if the additional value is a real outlier, which has a value much less than 1, the smaller it becomes, the larger the weight of the remaining (fixed) values in the sample and the weaker the outlier's effect on the mode. The dependence of the location's estimate on the position of the additional value is continuous. This is why the mode rises, when the free value approaches the value of the smallest datum (1). A similar explanation may be applied to an “overshoot” of both RM and GEL beyond the largest data value of 10. Note, that in the case of an outlier, which is well beyond the range of the fixed data, the application of the EGDF is limited by the potential appearance of a second density maximum. Both the quantifying distribution functions QGDF and the QLDF have a “global” character in that for them there is a unique (best) scale parameter, which is also the case with the EGDF. A comparison of the sensitivity of these three functions to an additional “free” datum is shown in Fig. 16.2.



There is no significant difference in the behavior of the three functions over the range of the fixed data (1 through 10) nor over the interval (0, 1). However, as the free data’s value increases beyond the last datum’s value, all three location parameters first decrease, then begin to rise. For both the QGDF and the QLDF the increase is accentuated, while for the EGDF it is much less pronounced and levels off at about 6. This effect is caused by the outer robustness of the quantifying distribution functions, which both give an increasing weight to the “outlier.”

16.3.2 A Detailed Comparison of Location Parameters

The fast pace in the growth of robust statistical theory over the postwar decades resulted not only in the development of many robust estimating methods, but also in the appearance of extensive numerical studies designed to compare their respective features and to evaluate the efficiency of the new techniques. In order to validate these tests, “artificial” data were used (where both the “true” and the “disturbance” components were known). The use of this procedure was defended using the argument, that it was necessary to know in advance, what the “true result” should be. However, researchers with a more realistic orientation objected, that in practice, the true character of data is never known beforehand, and that it would be difficult to judge, whether a “theoretically” good method would provide substantive results, when applied to real data. This was the objective behind S.M.Stigler’s [106] decision to test 11 types of robust statistical estimators of location using 16 independent samples of real data made famous for their historical measurement of well known physical parameters:

1. the parallax of the sun (Short 1763),
2. the mean density of the Earth (Cavendish 1798),
3. the speed of light (Newcomb 1882, Michelson 1879 and 1882).

The samples contained between 17 to 100 observations. The eleven robust estimating methods tested were: six “traditional” measures (sample mean, sample median, three trimmed means (10%, 15% and 25%) and Edgeworth’s L-estimator) and five “recently” developed parameters (outmean, three types of M-estimators (Huber P15, Andrews AMT and Tukey Biweight) and an adaptive estimator (Hogg T1))[4]. His motivation was, that since the true value of these measured quantities (which were unknown at the time the measurements were taken) have recently been estimated with a high precision:

> *The closer the realized value of an estimator to the current ‘true’ value of the estimated quantity, the better the estimator.*

Stigler concluded, that

> *Modern estimators are not worth the time necessary to compute them,*

and that

> *The smallest non-zero trimming percentage included in the study emerged as the recommended estimator* and *the mean itself*

[4] A description of these methods can be found in [106]

> *did rather well.*

A second study using contemporary analytical-chemistry data [94] also attempted to determine the most useful location parameter. In this endeavor, it was felt, that the current "true value" of the classical data was irrelevant, because of the possible existence of a bias, which could be larger than the variation across the data:

> *What is of importance is the variance of the location estimator used, since lower variance means, that the population location parameter is more precisely determined.*

A comparison of the variance of the estimators applied to both the classical and modern analytical-chemistry data resulted in the suggestion, that either severely trimmed means or modern robust estimators are required to obtain optimum performance. These conflicting conclusions suggested, that a comparison of the gnostic global estimate of location (GEL) with the 11 statistical ones using the same classical physical data [62] would prove instructive. Since the 12 estimating methods were to be compared using 16 samples of data, each sample was normalized, so that the task could be interpreted as the estimation of a single fixed quantity. The problem of the "true" data values was solved by associating the idea of an "expert board" of top statistical experts with the respective methodology of each of its members (Profs. Huber and Andrews, among others), whose outcomes are listed by their names in the table following. The authors of the classical methods are, of course, unknown, but it can be assumed, that they were the top experts of their time. It is therefore possible to accept in the same manner the ideas of "Prof. Mean, Prof. Median" and others. Thus there are 12 expert estimates of location parameters for 16 normalized independent data samples. Justification for this approach can be shown by the test, which demonstrates, that the distribution of the 12∗16 'expert estimates' around the mean value of 1 is undoubtedly Gaussian. The outcome of using these methodologies, ordered by the standard deviation of the estimating errors from the mean is summarized in Table 16.5.

Conclusions drawn from Table 16.5:

- The best results, which were close to the mean of the estimates of the whole "expert board," were provided by the gnostic estimator GEL.
- The higher percentage of trim, the better trimmed mean.
- Both the sample median and the arithmetic mean performed badly.

The last two conclusions are similar to those of [94]. In contrast, to [106], this study showed, that (some) robust estimators are really worth the time necessary to compute them. The standard errors of about 4% (GEL) or of about 6% (Hogg T1) have to be preferred over the more than 20% SE, which results from the use of the sample mean or the sample median.

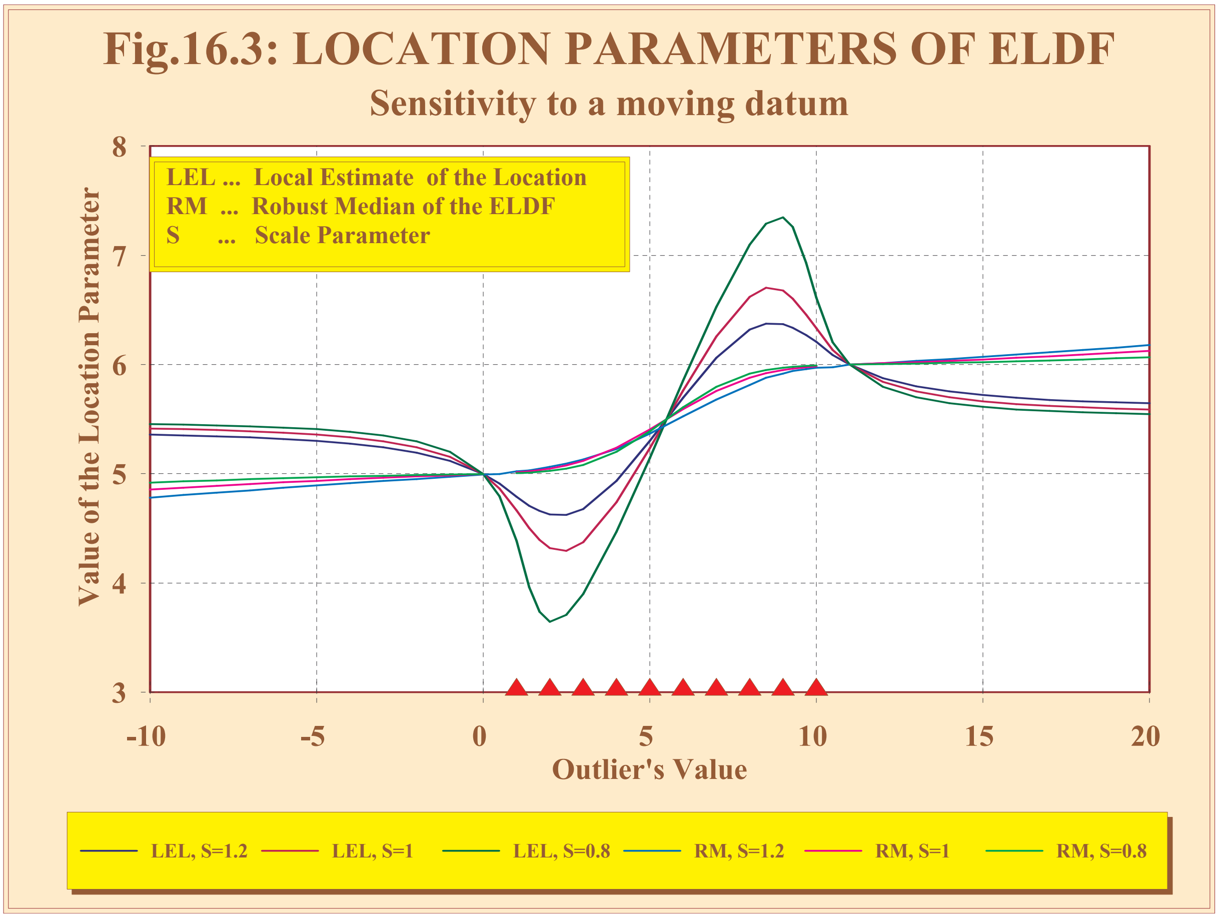
Estimator (method)	**Measure of the error**		
	Stand. deviation	**Mean error**	**Range of errors**
Gnostic (GEL)	0.038	−0.001	0.139
Hogg T1	0.061	−0.017	0.261
25% Trim	0.070	−0.029	0.261
Edgeworth	0.079	−0.011	0.273
15% Trim	0.104	0.032	0.447
Tukey Biweight	0.131	−0.043	0.631
Andrews AMT	0.147	0.025	0.660
Huber P15	0.210	0.083	0.856
10% Trim	0.211	0.097	0.821
Arithmet. mean	0.212	0.078	1.055
Sample median	0.278	−0.124	0.962
Outmean	0.610	−0.086	2.603

Tab. 16.5 Errors of 12 robust estimation methods of location parameters applied to 16 normalized samples of historical data.

16.3.3 Local Location Parameters (LEL)

The sensitivity of the estimating local distribution function ELDF to outliers has been shown to differ significantly from that of the EGDF (Chapter 15). It is no surprise, then, that the local estimate of the location parameter (LEL) behaves differently. Moreover, an examination of its behavior under a changing additional datum opens a path to a new way to classify data. Once again consider the same data sample, $\mathcal{Z}_{10}$, of 10 uniformly distributed data 1, 2, ..., 10 and once more extend it with an 11-th ‘free’ datum, which changes its value over a broad interval. The object of interest is the sensitivity of the estimating local distribution functions ELDF for three values of the scale parameter S (0.8, 1.0 and 1.2). The ELDFs have unimodal densities for these values of the scale parameter. The dependence of two location parameters (RM—the robust median and LEL—the local estimate of location) on the value of the additional free datum is shown in Fig. 16.3.

The behavior of the robust median is not surprising, it is reminiscent of the form of the trimmed mean from Fig. 16.1, only the sharp edges are



smoothed. A difference can also be observed outside of the data range, where the trimmed mean is constant (5 or 6), while the RM deviates from these values. Using 15.24 it can be shown, that the limit of the deviation in this case can reach $\pm Z_{11}$. The LEL's behavior is more interesting. When moving from left to right on Fig. 16.3, three portions of the LEL can be distinguished: one, which decreases to a minimum, an increase to a maximum and then once again a decreasing portion. These effects are considered in more detail in the following section and they are illustrated in Fig. 16.4.

16.4 Interval Analysis

16.4.1 Three Interesting Data Intervals

The following definitions will facilitate the analysis of the location parameters of a data sample derived from the estimation of the local distribution function (ELDF) of a data sample.

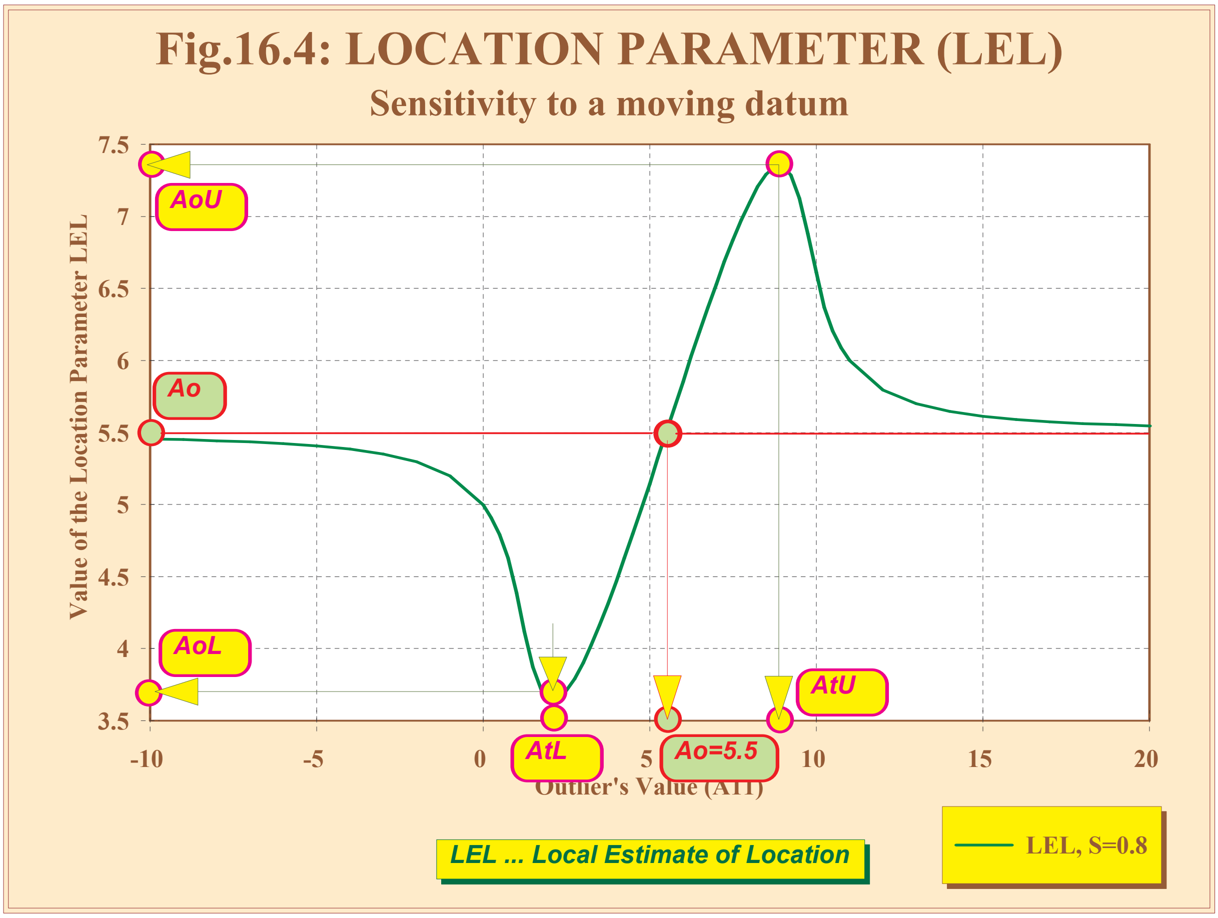


Fig.16.4: LOCATION PARAMETER (LEL)

Sensitivity to a moving datum

Definition 16: Let $\mathcal{Z}_N$ be a data sample of N fixed data elements defined over the infinite data support R_+.
Let w_k $(k = 1, \ldots, N)$ be a priori weights of fixed data.
Let Z_x be an $N+1$-th datum, which can take any arbitrary value from R_+, so that $\mathcal{Z}_{N+1}$ is the data sample $\mathcal{Z}_N$ extended by the variable datum Z_x, the a priori weight of which is 1.
Let the ELDF be the estimating local distribution function of the sample $\mathcal{Z}_N$ calculated for a scale parameter S, such that the ELDF's density is unimodal, and let Z_0 be the mode of the ELDF's density, ie location of the density's maximum. Let $Z_0(Z_x)$ be the main mode of the extended sample $\mathcal{Z}_{N+1}$. Denote Z_L the smallest and Z_U the largest finite value of Z_x, for which the equation

$$\frac{d(Z_0(Z_x))}{dZ_x} = 0 \tag{16.19}$$

holds. Introduce notation

$$Z_{0,L} = \frac{d(Z_0(Z_x))}{dZ_x}(Z_L) \tag{16.20}$$

and

$$Z_{0,U} = \frac{d(Z_0(Z_x))}{dZ_x}(Z_U) \tag{16.21}$$

Then the interval $[Z_{0,L}, Z_{0,U}]$ is *the tolerance interval of the mode* and interval $[Z_L, Z_U]$ is *the interval of typical data.* An analogous notation and terminology will be used for finite data support, onto which the points of infinite data support are transformed.

Figure 16.4 shows the function $\frac{d(Z_0(Z_x))}{dZ_x}$ transformed onto the finite support of additive data, ie as function $\frac{d(A_0(A_x))}{dA_x}$. All multiplicative data Z_* are represented as their additive transforms A_*. The graph represents the sample $\mathcal{A}_{10}$ of fixed additive data 1, 2, ... , 10, which is extended by a "free" eleventh datum ("outlier") A_{11}, whose value is indicated on the horizontal axis as $A11$. Values of the corresponding location parameter of the type "mode" are on the vertical axis. Several observations are in order:

O1 The mode of the extended data sample exactly equals the mode (A_0) of the original (non-extended) sample, when the value of the eleventh datum is $A_{11} = A_0$.

O2 Regardless of the value taken on by the eleventh "free" datum the location parameter (mode) will **always** remain within the tolerance interval $[A_{0,L}, A_{0,U}]$.

O3 The value A_0 as defined in **O1** above is the limit of the mode of the extended sample, when $A_{11} \to -\infty$ as well as when $A_{11} \to \infty$.

O4 The function $\frac{d(A_0(A_x))}{dA_x}$ increases only within the interval of typical data delimited by the points A_L and A_U.

O5 There may also be some peripheral data from the fixed sample, that are atypical if they lie outside the interval of typical data.

These features are preserved even when the outlier takes on an extreme value; under these circumstances, the distribution becomes bimodal, the second mode coinciding with the outlier's location. The form of the graph in Fig. 16.4 is considered "typical behavior" for the location parameter, when an increase in the value of a datum causes the location parameter to change also. It might have been expected, that the location parameter would increase as well; a view probably rooted in the habitual use of the arithmetic mean as the location parameter, where this expected behavior is natural. In the case of the LEL (Local Estimate of Location) defined as the mode of the ELDF such “natural” behavior is only observed inside the typical data interval, which is obviously determined by the sample of fixed data. If datum A_{11} is typical, the “board” of fixed data, “supports” increases in the mode as a result of increasing the value of A_{11}. However, a value of A_{11}, which is outside the interval of typical data is “rejected” by the “board”, because it is in conflict with the location of the fixed data. This “resistance” manifested by the decreasing mode is weak, when the outlier is very far from the fixed data—the weight of the outlier is small, but when the outlier approaches the bounds of the typical data, its weight increases and its influence on the mode reaches extreme values. It was shown in the foregoing sections, that the global distribution functions EGDF are robust with respect to outliers. It should now be clear, that the ELDF is also robust in this same sense. When the bounds of the intervals are revealed, they can all be used to undertake a new type of data analysis, *interval analysis*[5]. Making use of the fact, that a finite data support has a lower and an upper bound (LB and UB), and transforming the bounds of the intervals introduced in this section and the mode Z_0 onto the finite support, the additive data support can be split into subintervals defined by following bounds:

$$(-\infty, LB, A_L, A_{0,L}, A_0, A_{0,U}, A_U, UB, \infty).$$

Therefore, data may be classified by noting, within which of the eight intervals it falls. It is even possible to measure the probability of a datum being assigned to each of the subclasses (intervals), because the distribution function (ELDF) is available. This procedure permits different samples to be reliably tested as to identity, similarity and dissimilarity, which whets an interest in exploring the theory of data intervals.

16.4.2 Theory of the Data Intervals

The estimating local distribution function ELDF (shortly denoted as EL, 15.25) has the density $\frac{d(EL)}{dZ_0}$ (15.26). Under the theory of intervals, it is assumed, that $EL(Z_0)$ is defined over the infinite support R_+ of multiplicative data. (For applications to real data, which have a finite support, transformations of the resulting bounds onto the finite support is assumed.

[5] Kind readers are asked to again tolerate usage of a term, which already has its meaning elsewhere (to denote treatment of data given as intervals instead of 'mere' numbers). Taking of such data into account in estimation of gnostic distribution functions will be considered in Chapter 19.

The logarithmic scale naturally lends itself to graphing the probability and density over the infinite support. Then the density of the sample $\mathcal{Z}_N$ extended by Z_x according to Definition 16 is:

$$\frac{d(EL)}{d\log(Z_0)} = \sum_{k=1}^{N+1} \frac{1}{S} w_k f_k^2 \tag{16.22}$$

(resulting from 15.26), where

$$f_k = \frac{2}{(Z_k/Z_0)^{2/S} + (Z_0/Z_k)^{2/S}} \tag{16.23}$$

is the estimating weight (***the fidelity***), and where Z_{N+1} is Z_x. This density reaches its maximum, when equation

$$\frac{d^2(EL)}{(d\log(Z_0))^2} = 0 \tag{16.24}$$

holds. Both S and Z_0 are strictly positive finite numbers, therefore equation 16.24 may be rewritten after differentiation as

$$\sum_{k=1}^{N+1} w_k f_k^3((Z_k/\widetilde{Z_0})^{2/S} - (\widetilde{Z_0}/Z_k)^{2/S}) = 0, \tag{16.25}$$

where $\widetilde{Z_0}$ is the mode (the location of maximum density). According to the assumption, only N data (Z_k for $k = 1, ..., N$) are fixed, while the $N+1$-th datum denoted Z_x is a variable taking arbitrary values from R_+. The equation of the mode is therefore

$$\sum_{k=1}^{N} w_k f_k^3 * ((Z_k/\widetilde{Z_0})^{2/S} - (\widetilde{Z_0}/Z_k)^{2/S}) + f_x^3 * ((Z_x/\widetilde{Z_0})^{2/S} - (\widetilde{Z_0}/Z_x)^{2/S}) = 0. \tag{16.26}$$

Equation 16.26 may be used to show, that the observation **O1** (in 16.4.1) based on Fig. 16.4 is of a general nature. Indeed, in the case of $Z_x = \widetilde{Z_0}$ the second term in 16.26 vanishes. However, $\widetilde{Z_0}$ is the mode and the first term is zero also. The location parameters of the non-extended and extended sample always coincide, when the additional datum Z_x is equal to the mode of the initial (non-extended) sample. The stage is now set for a formal statement, which defines the intervals of the local estimates of location.

Theorem 17: Let $\mathcal{Z}_{N+1}$ be a sample of N multiplicative data defined over the infinite data support R_+. Let this sample be partitioned into a subsample $\mathcal{Z}_N$ containing data Z_k (their prior weights are w_k, $k = 1, ..., N$) and the $N+1$-th data element is Z_x. Let data $Z_1, \ldots, Z_N$ be fixed positive numbers having fixed weights and let a positive constant S be a scale parameter, such that the ELDF's density has only one mode $\widetilde{Z_0}$, which is a root of equation 16.26. Let Z_x be a positive real variable. Then the following formula holds:

$$\frac{d\widetilde{Z_0}}{dZ_x} = \frac{\widetilde{Z_0}}{Z_x} \frac{3f_x^4 - 2f_x^2}{\sum_{k=1}^{N}(3w_k f_k^4 - 2w_k f_k^2) + 3f_x^4 - 2f_x^2}, \tag{16.27}$$

where f_k and f_x are the same fidelities as in 16.23 and 16.26.

Proof of Theorem 17: Write equation 16.26 as an implicit function of two variables

$$F(\widetilde{Z_0}, Z_x) = 0. \tag{16.28}$$

From 16.26, function F is differentiable by both its arguments, therefore the total differential exists and equals zero:

$$\frac{\partial F}{\partial \widetilde{Z_0}} d\widetilde{Z_0} + \frac{\partial F}{\partial Z_x} dZ_x = 0. \tag{16.29}$$

Assume, that the derivative $\frac{\partial F}{\partial \widetilde{Z_0}}$ is zero. Then the second term in 16.29 must be zero, and the function $F(\widetilde{Z_0}, Z_x)$ does not depend on its arguments; however, it is also the second derivative of the distribution function ELDF (15.25). According to the assumption, the function must be a quadratic function of the ideal value $\widetilde{Z_0}$ and of a datum Z_x. Since this contradicts formula 15.25, the assumption is wrong, and the derivative cannot be zero. Derivative

$$\frac{d\widetilde{Z_0}}{dZ_x} = -\frac{\frac{\partial F}{\partial Z_x}}{\frac{\partial F}{\partial \widetilde{Z_0}}} \tag{16.30}$$

therefore exists. The numerator of this ratio is

$$\frac{\partial F}{\partial Z_x} = \frac{4}{SZ_x} * f_x^2 * (1 - 3h_x^2), \tag{16.31}$$

where $h_x = \pm\sqrt{1 - f_x^2}$ is the estimating irrelevance. Analogously, the denominator is

$$\frac{\partial F}{\partial \widetilde{Z_0}} = -\frac{4}{S\widetilde{Z_0}} \left(\sum_{k=1}^{N} w_k f_k^2 * (1 - 3h_k^2) + f_x^2 * (1 - 3h_x^2) \right). \tag{16.32}$$

Taking into account the relation between irrelevances and fidelities and substituting both partial derivatives into 16.30, the relation 16.27 is obtained.

It can now be seen, that Theorem 17 completely solves the problem of the bounds, that are needed for interval analysis:

Corollary 17.1: Let the conditions of Theorem 17 hold and let Z_L and Z_U be the lower and upper bounds of the interval of typical data. Then

$$Z_L = \widetilde{Z_0} * \left(\frac{\sqrt{3}-1}{\sqrt{2}}\right)^{S/2}, \quad Z_U = \widetilde{Z_0} * \left(\frac{\sqrt{3}+1}{\sqrt{2}}\right)^{S/2}, \tag{16.33}$$

$$\frac{Z_U}{Z_L} = \left(\frac{\sqrt{3}+1}{\sqrt{3}-1}\right)^{S/2} \tag{16.34}$$

and

$$Z_L * Z_U = (\widetilde{Z_0})^2 \tag{16.35}$$

Proof of Corollary 17.1: According to its definition, $\widetilde{Z_0}$ is a root of equation 16.26 for a given Z_x. As such, it is a function $\widetilde{Z_0}(Z_x)$. The derivative of this function results from Theorem 17 (16.27). The lower (upper) bound Z_L (Z_U) of the typical data's interval is defined as the point, where the function $\widetilde{Z_0}(Z_x)$ reaches its minimum (maximum), ie at the points, where the numerator of 16.27 equals zero:

$$f_x^2 = \frac{2}{3}. \tag{16.36}$$

There are two points satisfying this condition, namely the roots of the quadratic equation

$$\frac{2}{(\frac{Z_x}{\widetilde{Z_0}})^{2/S} + (\frac{\widetilde{Z_0}}{Z_x})^{2/S}} - \sqrt{\frac{2}{3}} = 0. \tag{16.37}$$

Identifying these roots with bounds Z_L and Z_U, one arrives at 16.33. Ratio 16.34 and product 16.35 of the bounds result directly from 16.33.

As a result of Corollary 17.1, the width of the typical data's interval is determined by only one parameter of the data sample, the scale parameter S. Moreover, the bounds Z_L and Z_U of this interval are placed on the data support so, that the mode Z_0 of the EL's density of (non-extended sample $\mathcal{Z}_{N-1}$) is the geometric mean of the interval's bounds (see 16.35).

16.4.3 Steps in Interval Analysis

Interval analysis of a data sample is comprised of the following steps:

1. Test of homogeneity of the data sample by means of the EGDF. If the sample is not homogeneous, it is split into homogeneous subsamples.
2. Estimate the EGDF's parameters LB and UB (bounds of the data support) and the scale parameter S.
3. Solve 16.26 (with $f_x = 0$ and with the EGDF's S) for the mode $\widetilde{Z_0}$ of the non-extended sample.
4. Calculate the bounds of the interval of typical data Z_L and Z_U using 16.33.
5. Calculate the bounds of the tolerance interval $Z_{0,L}$ and $Z_{0,U}$ by solving 16.26 first with $Z_x = Z_L$ and then with $Z_x = Z_U$.
6. Calculate probabilities $EL(Z_L)$, $EL(Z_{0,L})$, $EL(Z_{0,U})$ and $EL(Z_U)$.

There are two typical applications for interval analysis: determining, whether a datum belongs to a data sample and testing for the similarity between data samples. Given a datum Z_x, it can be determined if (and to what degree) it can be considered as a possible member of a given data sample $\mathcal{Z}_N$. Interval analysis of the sample (performed without the candidate datum) permits the following expectations as to Z_x's potential membership in $\mathcal{Z}_N$ to be estimated:

Highly probable: $Z_x = \widetilde{Z_0}$.
Within tolerance: $Z_x \in [Z_{0,L}, Z_{0,U}]$.
Typical: $Z_x \in [Z_L, Z_U]$.
Possible: $Z_x \in (LB, UB)$.
Improbable: $Z_x \leq LB$ or $Z_x \geq UB$.

Probabilities evaluated by means of the ELDF can be used to quantify these statements. Examples of simple applications include:

- For financial statement analysis, a set of ratios taken from a group of comparable companies form the sample $\mathcal{Z}_N$, while Z_x is a ratio of the same type for the candidate firm. How closely does the candidate's behavior follow that of the group taken as the "norm" for that specific measure of performance?
- In a quality control application, the parameters of a "normal" sample are $\mathcal{Z}_N$. The analysis would then determine, whether Z_x fits within the bounds of acceptable quality.

An additional situation, which Interval Analysis can resolve consists of determining the similarity between two samples, say $\mathcal{Z}_A$ and $\mathcal{Z}_B$. The *similarity* of the samples' states or processes can be classified by using the following criteria [6]:

High: $\tilde{Z}_{0,A} = \tilde{Z}_{0,B}$.
Within tolerance: $[Z_{0,L,A}, Z_{0,U,A}] \cap [Z_{0,L,B}, Z_{0,U,B}] \neq 0$.
Typical: $[Z_{L,A}, Z_{U,A}] \cap [Z_{L,B}, Z_{U,B}] \neq 0$.
Possible: $(LB_A, UB_A) \cap (LB_B, UB_B) \neq 0$.
Improbable: $(LB_A, UB_A) \cap (LB_B, UB_B) \equiv 0$.

Two data samples are very similar, when their modes coincide. Similarity "within tolerance" (a non-empty intersection of tolerance intervals) means, that there exist a single datum to each of the samples such that extension of both samples by their suitable datum would cause equality of their modes. The "typical" similarity (a non-empty intersection of the typical data's intervals) means, that some typical data of sample A may be typical data of sample B. The "possible" degree of similarity (non-zero intersection of the data supports) means, that there exist some "common" data in both samples.

16.4.4 A Link to Statistics

Within the framework of statistics, there is an idea, which plays a role complementary to probability, *the likelihood.* Probability is a theoretical construct, which is not directly related to any particular data. In contrast, with the likelihood function, a data set with unknown distribution or density parameters is tested to determine the *likelihood*, that it belongs to a specific distribution. Estimates computed using the maximum-likelihood method have favorable statistical features. A further development of this concept can be observed in robust statistical theory with *M-estimates* of location parameters (which maximize the likelihood function) ([33]). As shown in [77] and [78], formula 16.26 for the LEL estimator $\widetilde{Z_0}$—under certain statistical assumptions,—may be interpreted as a special case of the statistical robust M-estimator. This formal coincidence is further amplified below:

1. There are a large number of different M-estimators in robust statistics, each based on different statistical assumptions. It is therefore not easy to choose "the right one" to apply to a particular real data set (see Tab. 16.5 for an example). The gnostic LEL estimator does not need any prior statistical assumptions, because it is a product of a theory,

[6] The symbol $\cap$ corresponds to the interval's intersection and 0 denotes the empty interval.

which is independent of statistics. The LEL's form is unique and there is only one parameter (S) to be fixed.

2. From the point of view of science and contribution to general knowledge, it is a positive development, when a new autonomous theory (gnostics) coincides with special cases of a generally accepted old theory (statistics) even though the fundamental ideas of both theories are entirely different. Such "common boundaries" between gnostics and statistics have already been shown by the limit behavior of gnostic characteristics in the case of a weak data uncertainty. The formal similarity of the LEL and an M-estimator is more important, because it exists not only for weakly spread data, but also for a general case.
3. If the statistical assumptions (independent and identically distributed data) really apply to specific data, then the LEL also possesses the favorable statistical features proved in robust statistics for M-estimators. However, if the data are not "iid[7]", nothing can be said about the use of M-estimators, while the LEL is still valid.

16.5 Membership of a Data Sample

The idea of the sensitivity of a given sample's characteristics with respect to its extension by an additional "free" data item can also be applied to an estimating global distribution function (EGDF). Recall, that a data sample is considered homogeneous only if its EGDF, calculated for the unique scale parameter of the global type, is unimodal. Considering the notion of a sample's homogeneity in more detail, let it be noted at the start, that a data sample is created to satisfy the need for quantitative information about a subject or an event. Therefore, the subject or event of interest must first be qualitatively delimited. Data are thus "messengers" delivering a (more or less certain) message to the "client", the observer. It is natural to expect, that this transfer of information has some sense of "order": observations on an object, "A," should not be disturbed by data attributable to another object, "B." The membership problem is "easy" as posited as a primitive notion in classical set theory; it is assumed, that "everybody knows, whether any element is a member of a given set." The problem in Real Life is more complex. Example: is our kind reader sure to belong to the set of healthy people? In contrast, "fuzzy theorists" use a function to provide a vivid notion of the "degree of membership" of an

[7] Independent and identically distributed (a statistical condition).

element to a (fuzzy) set. Instead of the “sharp” yes/no determination used by the classicists, the required fuzzy statement takes the form of a number, which places the location of the statement somewhere between the two extremes of ‘yes’ or ‘no.’ However, it often occurs, that the observer only has the data and nothing further to help distinguish the data coming from “A” from that of “B.” Such necessary supplementary information may be unavailable. Data have their values and a data sample is specified by its distribution function. This explains, why the membership problem in statistics is solved by using estimated probability: highly probable data values are considered to be “members” of the data set, while data to which a low probability is attached are deemed to be “non-members,” or outliers[8]. The risk attached to making the statistical decision of member/non-member is measured by probability; and this risk corresponds to a (sometimes subjectively chosen) significance level: “Do you wish to ensure against membership by a non-member? Then increase the significance.” Subjectivity is also seen in the fuzzy approach, when the choice of the membership function is left to the user of the method. The major difference between the above methods and gnostics is, that gnostics solves the membership problem of a given (homogeneous) sample uniquely A data sample defines its global distribution function uniquely, the EGDF is completely determined by the data and by three other numerical parameters (the global scale parameter and the bounds of the data support), which are also uniquely estimated using the data. The homogeneity test (the EGDF’s unimodality) also has a “sharp” yes/no character. By leaving out data, which cause inhomogeneity, the rest of the data can be made into a homogeneous sample. Taking a set of homogeneous data, together with their scale parameter and the bounds of their data support, consider changes in the EGDF’s density caused by the addition of another datum, Z_x, which is transformed onto the infinite data support. Specify the EGDF as $F(Z_x) : (0, \infty) \to (0, 1)$. This function is continuous and unlimitedly differentiable, the density $D1 := \frac{dF}{d(ln(Z_x))}$ and derivatives $D2 := \frac{d^2F}{d(ln(Z_x))^2}$ and $D3 := \frac{d^3F}{d(ln(Z_x))^3}$ exist and are also continuous. Increases in Z_x to values sufficiently greater than the largest fixed data, or decreases in Z_x below the smallest member of the fixed data can lead to the formation of a second density’s maximum. (The density $D1$ over the infinite data support has always at least one maximum.) The boundary state between

[8] To be an outlier is not necessarily something bad. So, eg, an extraordinarily high profit for a company may be an outlier in a group of otherwise comparable companies.

the one-maximum and two-maximum situation corresponds to a Z_x value, at which exactly one additional inflection point exists in the density, $D1$. It is easy to see, that for an inflection point in $D1$, equation $D3(Z_x) = 0$ holds. This equation has **two** roots in the case of unimodal density $D1$ and **four** roots in the bimodal case. The boundary point is characterized by coincidence of the third and fourth inflection point. Such a "double inflection point" is easy distinguishable from the "ordinary" first and second inflection points, because in the double point the equation $D2(Z_x) = 0$ holds. There obviously exist two such double inflections, the lower and upper ones. These points determine the bounds of a homogeneous data sample and enable the idea of interval analysis to be further extended.

Definition 16A: Let $\mathcal{Z}_{\mathcal{N}}$ be a homogeneous data sample composed of N data, which after transformation onto the infinite data support form a vector $\underline{ZI}$. Let $\underline{W}$ be the vector of a priori weights of these data. Let S, LB and UB be optimum estimates of the scale parameter, lower and upper bound of the sample $\mathcal{Z}_{\mathcal{N}}$ based on $\underline{ZI}$ and $\underline{W}$. Let $F : (0, \infty) \rightarrow (0, 1)$ be the estimating global distribution function obtained as

$$F := F(\underline{ZI}, \underline{W}, S, LB, UB, Z_x), \qquad (16.38)$$

where $Z_x \in R_+$ is an additional variable, which extends the sample $\underline{ZI}$. Taking the above as $F(Z_x)$, while holding the function's other arguments fixed, then the values for Z_{LSB} $(0 < Z_{LSB} < \min(ZI))$ and Z_{USB} $(\max(ZI) < Z_{USB} < \infty)$, when equation

$$\left(S* \mid \frac{d^2F}{d(ln(Z_x))^2} \mid + \mid \frac{d^3F}{d(ln(Z_x))^3} \mid\right)_{Z_x = Z_B} = 0 \qquad (16.39)$$

holds, are correspondingly the *lower $(B = LSB)$* and *upper $(B = USB)$* bounds of the sample $\mathcal{Z}_{\mathcal{N}}$.

The sample bounds were defined over the infinite data support and are to be calculated in this format. Users, who would work with the more natural data form defined over a finite support should not forget to transform the bounds onto the natural scale. It is useful to consider an example. Ten normally distributed samples have been generated by the pseudo-random generator of S-PLUS for mean of 0 and a standard deviation of 1. Each sample consists of 10 data. Neither estimates (AVG) of the means nor estimates of the standard deviations (STD) are equal to their theoretical values because of the randomness. The sample estimates are summarized in Tab.16.6 along with the gnostic membership bounds LSB and USB.

It is seen, that the volatility of the estimates of both statistics is significant. Statistical control of normally distributed samples is ordinarily based on limits set as multiples of the standard deviations. Denote such a control limit by $M * STD$. Multiplication factor M is a function of the chosen significance of the statistical test, the limits being $AVG - M * STD$ and $AVG + M * STD$.

Ser.	**MEANS**	**St.Dev.**	**Gn. bounds**	
No.	**AVG**	**STD**	**LSB**	**USB**
1	-0.245	0.806	-3.79	3.68
2	-0.117	1.120	-3.88	3.87
3	0.096	0.784	-2.91	3.04
4	0.117	0.927	-3.97	3.82
5	0.484	0.586	-3.67	3.85
6	-0.120	0.889	-3.19	3.09
7	-0.026	1.042	-3.51	3.50
8	0.137	0.888	-3.64	3.68
9	0.588	0.756	-3.33	3.69
10	0.002	0.808	-3.30	3.41
MIN	-0.245	0.586	-3.97	3.04
MED	0.049	0.848	-3.58	3.68
MAX	0.588	1.120	-2.91	3.87

Tab. 16.6 Statistics and gnostic sample bounds for 10 normally distributed samples together with their minima (MIN), maxima (MAX) and medians (MED)

In the case of the ten series from Tab. 16.6 the estimated statistical control limits vary between $AVG \pm M * 0.586$ and $AVG \pm M * 1.120$ depending on the significance chosen. The relative range of the control limit defined as (max{STD}-min{STD})/median{STD} is thus (1.120-0.586)/0.848 = 0.630 for an arbitrary M. The gnostic sample bounds LSB and USB in Tab. 16.6 do not depend on the significance level and they are relatively stable in spite of strong random variations of the series. [9] Their relative

[9] To estimate their values, it was assumed, that the given data cannot have values outside of the interval $(-4, 4)$. For theoretical data distributed as $\mathcal{N}(\prime, \infty)$ the probability of appearing outside this interval is 0.000063. This probability is even more negligible in practice, where data are always bounded, "trimmed". The assumption, on which the sample bounds are based is therefore realistic.

range is 0.296 for the LSB and 0.226 for USB—less than a half of the statistical case. (Note, that this comparison neglects the volatility of the estimated mean, AVG).

It is seen from Tab. 16.6, that the volatility of the results is smaller if the gnostic methodology is applied. This is remarkable, because many people believe, that the application of pure Gaussian distributions raise the least doubts on the validity of classical statistical concepts.

16.6 Summary

All the information necessary for computing gnostic distribution functions is taken from the data. These functions are completely defined by the data and by parameters, which also are specified by the data. Three ("primary") parameters—the scale parameter and bounds of data support—determine the form of these distributions. Other ("secondary") parameters derived from the distributions provide numerical characteristics of the analyzed data sample.

To estimate the unique ("global") scale parameter, which provides the best fit of data for the distributions EGDF, QGDF and QLDF, it is necessary to solve the corresponding extremization problem. When the bounds of the data support are not given, they are found by optimization together with the optimum scale parameter. There are several methods, by which the quality of the data fit can be evaluated. Three types of best global scale parameters were examined based on the Kolmogorov-Smirnov minimax measure, the maximum entropy principle and on maximization of the fidelity of the fit.

The choice of a scale parameter for estimating a local distribution function ELDF depends on the task. The local scale parameter may be useful in evaluating the local behavior of the distribution function, eg within a cluster of data, to test the constancy of the scale parameter over a whole data sample or to treat a data series. An alternative to the local scale parameter is the scale parameter, which ensures a required quality for the fidelity fit. In particular, it may be applied to make local distribution functions of different samples comparable to each other, and in order to set the resolution power in cluster analysis.

The location parameters are important secondary parameters of distribution functions. While a quantile for an arbitrary probability can be useful in analyzes, the most frequently used numerical parameters of the sample's location are the distribution's median and mode. A sizable case study using well known historical data showed, that the mode of the estimating global distribution function EGDF is an excellent location parameter and its robustness exceeds that of many estimators used in robust statistics. The location parameter also plays an important role, when the estimating local distribution function (ELDF) is used in interval analysis. This technique allows the degree of association of individual data elements within the sample to be determined with respect to the whole sample. Another application is to evaluate the similarity between data samples.

The sensitivity of these operations depends on the choice of the scale parameter. Its unique estimation along with the bounds of a data sample with the EGDF provides an objective answer, as to whether a datum could be of "member" of a given homogeneous sample.

The suitability of different scale parameters to specific tasks is summarized in Tab. 16.7.

Task	**Scale Parameter**	**Description**
Kolmogorov-Smirnov's Test	$S_{G,KS}$	16.3
Estimate Probability	$S_{G,MF}$	16.6
Estimate Density	$S_{G,MF}$	16.6
Estimate a Quantile	$S_{G,MF}$	16.6
Est. Location Parameter	$S_{G,MF}$	16.6
Est. Data Support Bounds	$S_{G,MF}$	16.6
Est. Sample's Boundaries	$S_{G,MF}$	16.6
Test for Homogeneity	$S_{G,MF}$	16.6
Interval Analysis	$S_{G,MF}$	16.6
Comparison of Samples	$S_{G,RF}$	Subsection 16.2.3
Multivariable Modeling	$S_{G,RF}$, $S_{G,MF}$ or S_L	Chapter 18
Cluster Analysis	Chosen S	Subsection 15.3.2
Heteroscedastic Analysis	Variable S	Subsection 16.2.4
Filtering of Time-Series	S_L	16.17
Cross-section Filtering	$S_{G,MF}$	16.6

Tab. 16.7 Recommended Scale Parameters

Chapter 17

Gnostic Regression

17.1 Basic Concept

The term, *regression*, was introduced by Francis Galton in connection with his observation, that the average height of children tended to move or "regress" toward the average height of the population as a whole. This idea of "regression to mediocrity" (Galton's words) found a broad application in statistical modeling. As he defined the concept [26]:

> *Regression analysis is concerned with the study of the dependence of one variable, the* **dependent variable***, on one or more other variables, the* **explanatory variables***, with a view to estimating and or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.*

This description does not state explicitly, that the dependent variable is not deterministic, but a stochastic function, nor that in advanced regression models even the explanatory variables may be stochastic. Another important unstated characteristic is, that the model's structure (the mathematical description of the dependence based on some unknown parameters) is assumed to be given. The statistical estimation or prediction of the dependent variable is ordinarily completed by an analysis of variance of the results. As is the usual practice in statistics, the behavior of the stochastic components is assumed to be described by a priori given statistical models. In robust statistics, the a priori assumptions are weaker, but they still exist.

A regression function corresponding to the classical definition will be called an *explicit regression*, because it distinguishes between the explanatory variables and at least one explicitly expressed dependent variable.

The gnostic notion of explicit regression differs from the statistical formulation in following ways:

1. In statistics, data are viewed as random samples taken from a population, the statistical model of which is assumed to be known. A gnostic regression is based on the different gnostic definition of uncertain data and data samples (see Chapters 5 and 14).
2. Instead of using the usual statistical criteria (eg minimum variance, unbiasedness, etc.), gnostic regression procedures minimize one of the six gnostic measures of uncertainty for the system being examined (Chapter 10).
3. Classical regression models do not meet the requirements of robustness. The robustness of regression models in robust statistics is due to additional statistical assumptions, which have been made with respect to the data models. In contrast, gnostic regression models are naturally robust because of their inherent features, which result from the application of the selected gnostic criterion function.

The gnostic approach to the regression problem is comprised of the following steps:

A The selection of a model structure, which respects the available data and the goal of the modeling.

B The choice of a gnostic criterion function to evaluate the quality of the data explanation.

C Optimization of the model's parameters by extremization of the criterion function applied to an *ex ante* portion of the available data.

D An analysis of the residuals (modeling errors) by means of distribution functions and drawing conclusions as to the
   1. homogeneity of the data samples,
   2. suitability of the model structure to explain the data,
   3. need to separate the various data clusters,
   4. evaluation of the quality of modeling,
   5. need to carry out another iteration from step **A**.

E Validation of the quality of the model by testing the *ex post* portion of the data and comparing the distribution function of the results with the distribution function of the initial subsample.

The separation of the model identification phase (**C** and **D**) from the verification phase (**E**) by splitting the data into ex ante and ex post portions is a routine procedure in time-series analysis. It is highly desirable to apply this process to cross-section regression analysis, because it provides to the user an indication of the real efficiency of the modeling. The only obstacle may be a critical lack of data, but this constraint can be frequently overcome by using two similar data sets.

Gnostic methodology as it is applied to a broad and important class of regression models will be discussed in the subsections to follow.

17.2 Robust regression models

The formulation and solution of this problem is general enough to include both linear and nonlinear versions of cross-section models as well as dynamic models of interrelations between time series. Explanatory variables may be interpreted as *inputs* and the dependent variable plays the role of the *output* of the model.

17.2.1 Formulation of the Problem

The approach to the robust regression problem represents a generalization of the method originally published in [60].

Problem: Let F be a differentiable function of a known type,

$$F: \; R^M \times R^M \to R_+. \tag{17.1}$$

Let $\underline{C} \in R^M$ be a column vector of unknown but constant parameters. Let $\underline{x} \in R^M$ denote an explanatory (or input) vector of a (generally nonlinear) regression model

$$z = F(\underline{C}, \underline{x}), \tag{17.2}$$

the variable z being the dependent variable (or the output). Suppose, that neither the input nor the output is known precisely. Let the n-th true values $\underline{x}_{0,n}$ and $z_{0,n}$ be related by means of the equation

$$z_{0,n} = F(\underline{C}, \underline{x}_{0,n}). \tag{17.3}$$

Let $\underline{x}_n$ and Z_n denote the uncertain observations of $\underline{x}_{0,n}$ and $z_{0,n}$ (data) for $n = 1, \ldots, N$.

Given a twice differentiable function

$$D: \; R^1 \to R_+. \tag{17.4}$$

Let $c^2 \in \{1, -1\}$ and h_c be the irrelevance, either h_j (quantifying, $c^2 = 1$) or h_i (estimating, $c^2 = -1$), so that at the n-th point relations

$$q_n = \left(\frac{Z_n}{z_{0,n}}\right)^{1/S} \tag{17.5}$$

$$h_{j,n} = (q_n^2 - q_n^{-2})/2 \tag{17.6}$$

and

$$h_{i,n} = \frac{q_n^2 - q_n^{-2}}{q_n^2 + q_n^{-2}} \tag{17.7}$$

hold for an estimate $\tilde{z}_{0,n}$ of the true value $z_{0,n}$ and for a positive scale parameter S.

The **objective** is to find the estimate $\underline{\widetilde{C}}$ of the unknown vector $\underline{C}$ of parameters, so that the criterion function

$$\varphi := \sum_{n=1}^{N} D(h_{c,n}) \tag{17.8}$$

is minimized, ie

$$\underline{\widetilde{C}} = \arg\min_{\underline{C}}(\varphi(S)). \tag{17.9}$$

The scale parameter S is assumed to be given based on some other consideration. Alternatively, the minimization task may include optimization of the scale parameter's value (S_{opt}), so that

$$\underline{\widetilde{C}} = \arg\min_{\underline{C}}(\varphi(S))|_{S=S_{opt}}. \tag{17.10}$$

When choosing between these alternatives, it must be taken into account, that the scale parameter's value determines the degree of robustness of the results with respect to inliers ($c^2 = 1$) or outliers ($c^2 = -1$). Expression 17.10 is to be interpreted as the minimization of φ constrained by the condition, that S reaches its optimal value. This optimality should be based on a criterion function other than φ to exclude the trivial minimum $\varphi = 0$, which is obtained as $S \to \infty$.

17.2.2 Iterative Solution

The optimization problems 17.9 and 17.10 can be solved directly by numerical methods. However, to obtain a theoretical insight to the solution of at least the task represented by 17.9, an iterative solution is more useful.

Theorem 18: Let the assumptions and definitions of 17.1–17.8 hold. Given a scale parameter S and given the k-th iteration $\underline{C}(k)$ of the estimate of the parameter vector $\underline{C}$ together with estimates

$$\tilde{z}_{0,n} = F(\underline{C}(k), \underline{x}_n) \tag{17.11}$$

of the ideal values (17.3) for $n = 1, \ldots, N$.
Let $h_{c,n}$ be irrelevance (17.6) or (17.7) using q_n (17.5), where the estimates $\tilde{z}_{0,n}$ are substituted for the ideal values $z_{0,n}$.
Establish the column-vector (relative value of the gradient of the function F)

$$\underline{g_n} := \frac{1}{\tilde{z}_{0,n}} * \left(\frac{\partial F}{\partial C_1}, \ldots, \frac{\partial F}{\partial C_M}\right)^T_{\underline{C}(k),\underline{x}_n} \tag{17.12}$$

($n = 1, \ldots, N$), and scalars

$$D^{'}_{c,n} := \frac{\mathrm{d}D}{\mathrm{d}h_{c,n}}, \tag{17.13}$$

$$D^{"}_{c,n} := \frac{\mathrm{d}^2 D}{\mathrm{d}h^2_{c,n}}, \tag{17.14}$$

$$h^{(1)}_{c,n} := \tilde{z}_{0,n} \frac{\mathrm{d}h_{c,n}}{\mathrm{d}\tilde{z}_{0,n}}. \tag{17.15}$$

Then the k+1-th iteration, which decreases the criterion function φ (17.8) can be approximated by the formula

$$\underline{C}(k+1) = \underline{C}(k) + \left(\sum_n^N D^{"}_n (h^{(1)}_{c,n})^2 * (\underline{g}_n \underline{g}^T_n)\right)^+ \times \left(-\sum_n^N D^{'}_n h^{(1)}_{c,n} \underline{g}_n\right), \tag{17.16}$$

where all quantities on the right-hand side of 17.16 are estimated by substitution of $\underline{C}(k)$ and $\tilde{z}_{0,n}$ instead of $\underline{C}$ and $z_{0,n}$, respectively, and where the symbol $\underline{M}^+$ denotes the pseudo-inverse of a matrix $\underline{M}$.

Proof of Theorem 18: To take into account the uncertainty of the criterion, the criterion will be analyzed together with its variation $d\varphi$. Let the condition for minimization of the variated criterion be

$$d\varphi + d^2\varphi = 0. \tag{17.17}$$

Both differentials are functions of the irrelevance $h_{c,n}$, hence

$$\sum_n^N D^{'}_n (\mathrm{d}h_{c,n} + \mathrm{d}^2 h_{c,n}) + \sum_n^N D^{"}_n (\mathrm{d}h_{c,n})^2 = 0. \tag{17.18}$$

The second differential $\mathrm{d}^2h_{c,n}$ in the first sum can be neglected as a small quantity with respect to $\mathrm{d}h_{c,n}$. The irrelevances $h_{c,n}$ are functions 17.6 or 17.7 of $\tilde{z}_{0,n}$, therefore

$$\sum_n^N D^{'}_n h^{(1)}_{c,n} \frac{\mathrm{d}\tilde{z}_{0,n}}{\tilde{z}_{0,n}} + \sum_n^N D^{"}_n (h^{(1)}_{c,n})^2 \left(\frac{\mathrm{d}\tilde{z}_{0,n}}{\tilde{z}_{0,n}}\right)^2 = 0. \tag{17.19}$$

The following scalar product results from 17.11 and 17.12:

$$\frac{\mathrm{d}\tilde{z}_{0,n}}{\tilde{z}_{0,n}} = \underline{g}^T \mathrm{d}\underline{\mathrm{C}}. \tag{17.20}$$

The scalar product commutes ($\underline{g}^T \mathrm{d}\underline{\mathrm{C}} = \mathrm{d}\underline{\mathrm{C}}^T \underline{g}$). Therefore, the equation

$$\mathrm{d}\underline{\mathrm{C}}^T \left(\sum_n^N D^{'}_n h^{(1)}_{c,n} \underline{g} + \sum_n^N D^{"}_n * (h^{(1)}_{c,n})^2 \underline{g}_n \underline{g}_n^T \mathrm{d}\underline{C} \right) = 0 \tag{17.21}$$

should hold identically, ie for all $\mathrm{d}\underline{C}^T$. Taking approximately

$$\mathrm{d}\underline{C} \doteq C(k) - C(k+1) \tag{17.22}$$

and using the pseudo-inverse of the square matrix 17.16 is obtained.

There are good reasons to apply pseudo-inversion to obtain the solution, because it is possible, that the square matrix may be singular for several reasons:

1. A larger number of parameters may have been chosen (the dimension of the vector $\underline{C}$) than is necessary to explain the data.
2. A linear dependence of the rows/columns of the square matrix may result due to data uncertainty.
3. The limited precision of the numerical representation of vectors and operations available in the computer software may reduce the rank of the matrix.

The application of pseudo-inversion does not involve any additional problems and it offers some extra improvements:

1. When computing the pseudo-inverse by means of the SVD-technique (Singular Value Decomposition), there is full control of the numerical determination of the matrix by evaluation of the ratio of the smallest to the largest of the singular numbers.

2. A suitable dimension for the parameter vector (and for the model under consideration) can be established by using the information on the singular numbers.
3. The SVD-technique substantially improves the precision of the results, because it applies only linear (not quadratic) operations to the matrices.
4. This technique provides important (geometric) information (orthogonal bases of both row- and column-subspaces) about the vector subspaces occupied by the data matrices.
5. It can be shown, that an important advantage results from this orthogonality: the possibility of unifying the solution's scales, which simplifies numerical operations.
6. In the case of regularity, the pseudo-inverse matrix is identical to the inverse matrix.

The use of pseudo-inversion is thus a way of obtaining extended generality and better quality in the modeling.

17.2.3 Interpretation of the Iterative Solution

The results obtained above may be interpreted in terms of the ordinary least-squares method. This statement is neither obvious nor trivial. A more detailed explanation is therefore desirable. Expression 17.16 may be interpreted as

$$\underline{C}(k+1) - \underline{C}(k) = \left(\sum_{n=1}^{N} \underline{G}_n \underline{G}_n^T\right)^+ \left(\sum_{n}^{N} \underline{G}_n E_n\right), \tag{17.23}$$

where

$$\underline{G}_n := \sqrt{|D^{"}_n|}\, h^{(1)}_{c,n} \underline{g}_n \tag{17.24}$$

and

$$E_n := -\frac{D^{'}_n}{\sqrt{|D^{"}_n|}}. \tag{17.25}$$

There are obviously N column vectors $\underline{G}_n$. Let us compose an $N \times M$ matrix $\underline{X}$ using vectors $\underline{G}_n^T$ as the rows of $\underline{X}$, so that

$$\underline{X}_{n,m} = \underline{G}_{m,n} \quad n = 1, \ldots, N, \quad m = 1, \ldots, M. \tag{17.26}$$

It is well-known, that the singular value decomposition of the matrix X (which has the rank $R \leq \min(N, M)$) can be uniquely obtained as the product of three matrices

$$\underline{X} = \underline{UWV}^T, \tag{17.27}$$

where $\underline{U}$ is an $N \times R$-matrix, $\underline{W}$ is $R \times R$ and $\underline{V}$ is $M \times R$. It is also well-known, that matrix $\underline{W}$ is diagonal with ordered non-zero singular numbers d_r $(r = 1, \ldots, R)$, and that both matrices $\underline{U}$ and $\underline{V}$ are semiorthonormal, ie $\underline{U}^T\underline{U} = \underline{V}^T\underline{V} = \underline{I}(R)$, where $\underline{I}(R)$ is the $R \times R$ identity matrix. It is to be emphasized, that in order to obtain the decomposition 17.27, only the SVD-algorithm to matrix $\underline{X}$ needs to be applied. The semiorthonormality of $\underline{U}$ allows the n-th row $\underline{u}_n$ equation

$$\sum_{n=1}^{R} \underline{u}_n^T \underline{u}_n = I(R) \tag{17.28}$$

to hold. The n-th row of $\underline{X}$ is $\underline{G}_n^T = \underline{u}_n\underline{WV}^T$. Therefore

$$\sum_{n=1}^{N} \underline{G}_n\underline{G}_n^T = \underline{V}^T\underline{W}^2\underline{V} = \underline{X}^T\underline{X} \tag{17.29}$$

and

$$(\sum_{n=1}^{N} \underline{G}_n\underline{G}_n^T)^+ = \underline{V}^T\underline{W}^{-2}\underline{V} = (\underline{X}^T\underline{X})^+. \tag{17.30}$$

Using the matrix $\underline{X}$ and composing a column vector $\underline{E}$ of the components E_n, one may write

$$\sum_{n=1}^{N} \underline{G}_n E_n = \underline{X}^T\underline{E}. \tag{17.31}$$

Solution 17.23 may be thus written in the form

$$\underline{C}(k+1) - \underline{C}(k) = \left(\underline{X}^T\underline{X}\right)^+ \underline{X}^T\underline{E}, \tag{17.32}$$

which actually is the least-squares solution of the equation system

$$\underline{X}(\underline{C}(k+1) - \underline{C}(k)) = \underline{E}. \tag{17.33}$$

Returning to the original variables, the n-th equation of the system can be rewritten as:

$$FW(Z_n/z_{0,n}) * \underline{g}_n^T = E(Z_n/z_{0,n}), \tag{17.34}$$

where

$$FW(Z_n/z_{0,n}) = \sqrt{|D_n^{"}||h_{c,n}^{(1)}|} \tag{17.35}$$

results from 17.24 and $E(Z_n/z_{0,n})$ is E_n (17.25). System 17.33 of these equations is an iterative but linear substitute for the original, but more complex task of 17.9. This substitution was made to examine the behavior of the solution. Row-vectors of the matrix X (ie $\underline{G}_n^T$), which have the form of the left-hand side of 17.34 now play the role of the explanatory (input) vector variables of the now linear model 17.33, with the scalars E_n as its outputs.

Gradient g_n (17.12) does not depend on uncertainty (at least after iteration), but the multiplier FW (17.35) does, because the uncertainty causes the ratio $Z_n/z_{0,n}$ to decline from 1. The multiplier amplifies or attenuates the impact of the input vector g_n depending on the uncertainty, potentially suppressing its effect. The function FW (17.35) may therefore be called the *filtering weight.*

The difference between the observed value of the dependent variable and its value as provided by the model represents the error of modeling (the *residual error*) and it may be called an *additive residual.* To protect the quality of the solution of the regression from data errors, robust statistics creates residual functions known as *influence functions*, which are based on some a priori accepted assumptions about the nature of the data. The effect of these functions can be interpreted as nonlinear filters: instead of directly using the (linear) residuals, they are transformed (in a nonlinear way) by means of this influence function. The gnostic function E also plays the role of an influence function, but with some differences: instead of additive residuals, its argument is the *multiplicative residual* $Z_n/z_{0,n}$[1], which instead of being based on a priori statistical assumptions, is completely defined by the gnostic criterion function D (17.8), that is based only on theory.

The multiplicative residuals $Z_n/\tilde{z}_{0,n}$ determine the filtering effect applied both to the inputs (explanatory vectors, 17.24), and to the output (dependent variable, 17.25). This effect is specific; it is different for each equation represented by 17.34, because the error of each equation is different. It will now be shown how this filtering affects the robustness of the method.

17.2.4 Robustness of the Gnostic Regression

The results from 17.15 and 17.5–17.7 are such, that

$$h_{j,n}^{(1)} = -\frac{2}{S} f_{j,n} \tag{17.36}$$

[1] It is multiplicative, because it states **how many times** the observed value Z_n is larger/smaller than the true model output $z_{n,0}$.

and

$$h^{(1)}_{i,n} = -\frac{2}{S} f^2_{i,n}, , \qquad (17.37)$$

where $f_{j,n}$ and $f_{i,n}$ are Q- and E-weights (9.10), which are bound to the irrelevances by the relations

$$f_{j,n} = \sqrt{1 + h^2_{j,n}} \qquad f_{i,n} = \sqrt{1 - h^2_{i,n}}. \qquad (17.38)$$

There are three pairs of gnostic characteristics, which are interesting as criteria $D(h_c)$ (17.8) associated with Q- and E-entropy, Q- and E-information, and with the sources of the fields of these quantities. These are summarized in Tab. 17.1. A numerical multiplier in the definition of D does not change the results of the optimization. The sign of D is chosen to provide a positive first derivative $D^{'}_{n}$. Additive constants do not play a role in D, they can be omitted, because only derivatives D' and D” are needed for the solution. ,

Case	Gnostic characteristic	$D(h_{c,n})$	References
Q1	Sources of the Q-entropy's field	$h^2_{j,n}/2$	(10.47)
E1	Sources of the E-entropy's field	$h^2_{i,n}/2$	(10.48)
Q2	Q-information	$I_{i,n}$	(10.58), (10.56) (10.48)
E2	E-information	$I_{j,n}$	(10.52), (10.50) (10.42)
Q3	Q-entropy (Q-weight)	$f_{j,n}$	(10.26),(9.10)
E3	E-entropy (E-weight)	-$f_{i,n}$	(10.26),(9.10)

Tab. 17.1 Gnostic characteristics usable as criterion functions for the regression problem.

An examination of the particular formulae, which control the filtering effects derived from Tab. 17.1 may be of interest.

Case ^{2}	Filtering Weight	Error Function
	FW (17.34)	E (17.25)
Q1	f_j	h_j
E1	f_i^2	h_i
Q2	1	$f_j \arg\tan(h_j)$
E2	f_i	$f_i \arg\tanh(h_i)$
Q3	$1/\sqrt{f_j}$	$\sqrt{f_j} h_j$
E3	$\sqrt{f_i}$	$\sqrt{f_i} h_i$

Tab. 17.2 Filtering weights and error functions for different gnostic criteria D used in Tab. 17.1.

The above functions are depicted in Figs. 17.1 through 17.5. The variable on the horizontal axis in Figs. 17.1-17.4 (which show the three categories of Q- and E- weights and errors) is $\Phi_n = \ln Z_n/\tilde{z}_{0,n}$—the uncertainty measured as the additive residual error (the scale parameter equals 1 in these examples). Figure 17.5 compares the behavior of the six error functions. To provide a more comprehensive comparison, each of the error graphs also plots the Euclidean error.

So as to properly understand the behavior of all the filtering functions, it is useful to recall, that by equation 17.34, the filtering weight (FW) amplifies or attenuates the left-hand side of the equation, while the error function E is on the right-hand side. Both of these functions are dependent on the equation error (multiplicative residual) $Z_n/z_{0,n}$, but in a different way: Using a simple example of a linear regression, for which equation 17.3 reduces to

$$z_{0,n} = C_0 + \sum_{m=1}^{M} C_m x_{0,n,m}, \tag{17.39}$$

the gradient 17.12 is

$$\underline{g_n} := \frac{\underline{x_{0_n}}}{\tilde{z}_{0,n}} \tag{17.40}$$

with the row-vector $\underline{x_{0_n}}$ composed of elements $x_{0,n,m}$. The ordinary least-squares regression (OLS) is based on the assumption, that the explanatory vectors $\underline{x}_n$ are disturbed only by a “white noise,” ie that elements of the explanatory vectors[3] are not correlated. Such an assumption is unrealistic in many applications. So, eg, if the dependent variable were the profit of a firm and the $\underline{x}_n$s were other economic parameters such as financial leverage, a liquidity ratio, total asset turnover etc., then there will be uncertainties on both the input and the output sides of the regression equation and they will surely be correlated. In the case of a “proper” choice for the criterion function D, the filter weights (FW) will suppress the input uncertainties and the error function (E), the uncertainties of the output variable. The resulting model will thus be robust with respect to both input and output uncertainties. Different criteria D will result in different types of robustness (inner/outer) with respect to the inputs and/or outputs.

Consider the case of Q-regressions (Q1, Q2 and Q3, in Tab. 17.1), where the filtering weights have the form of Fig. 17.1. Three qualitatively different behaviors of the FW functions are documented:

[3] In statistics, there are generalizations of the OLS methodology for a more general case of disturbances, but these assume knowledge of the correlation matrix of the disturbances, which is not always available. If an estimate of this matrix is used instead, it is likely, that there will be problems with the traditional method, which will result in unrobust estimates.

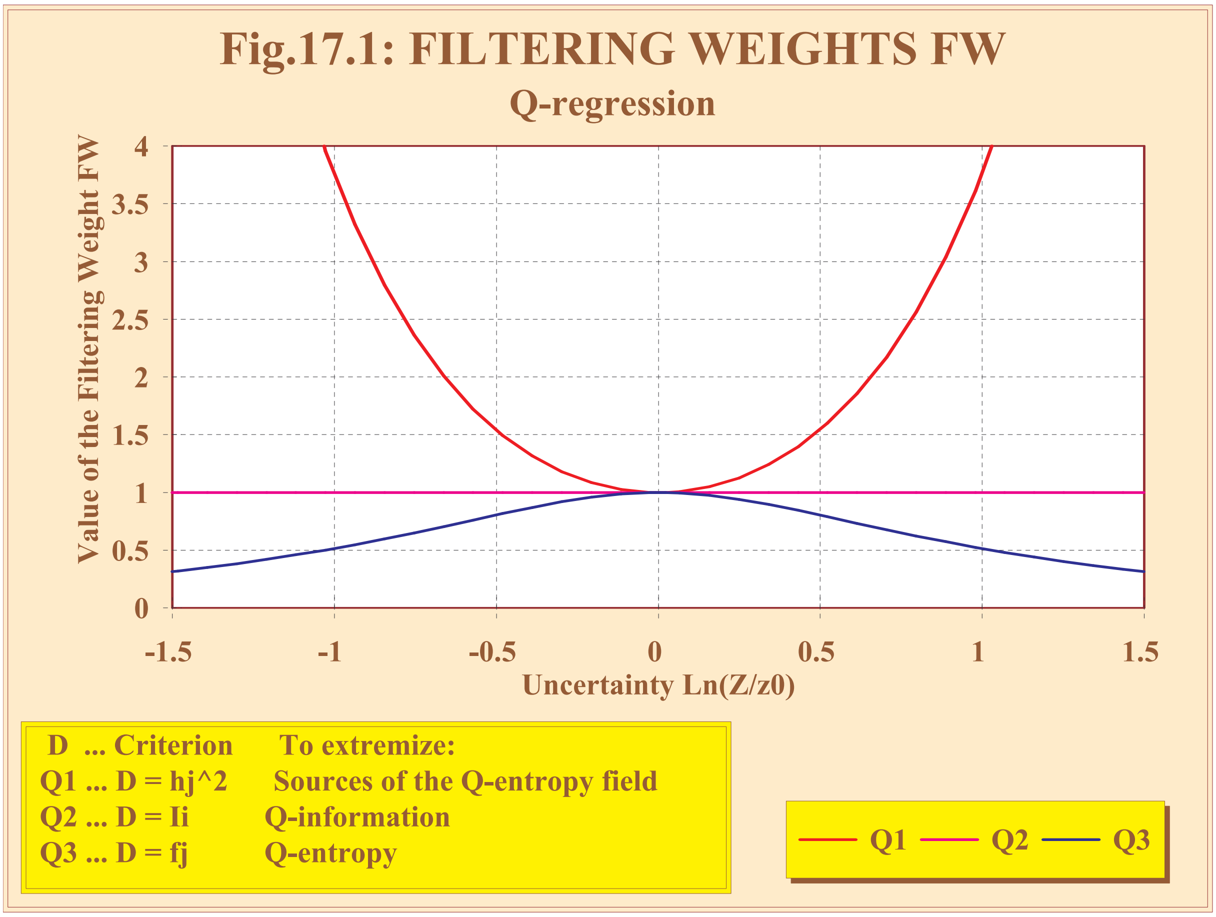


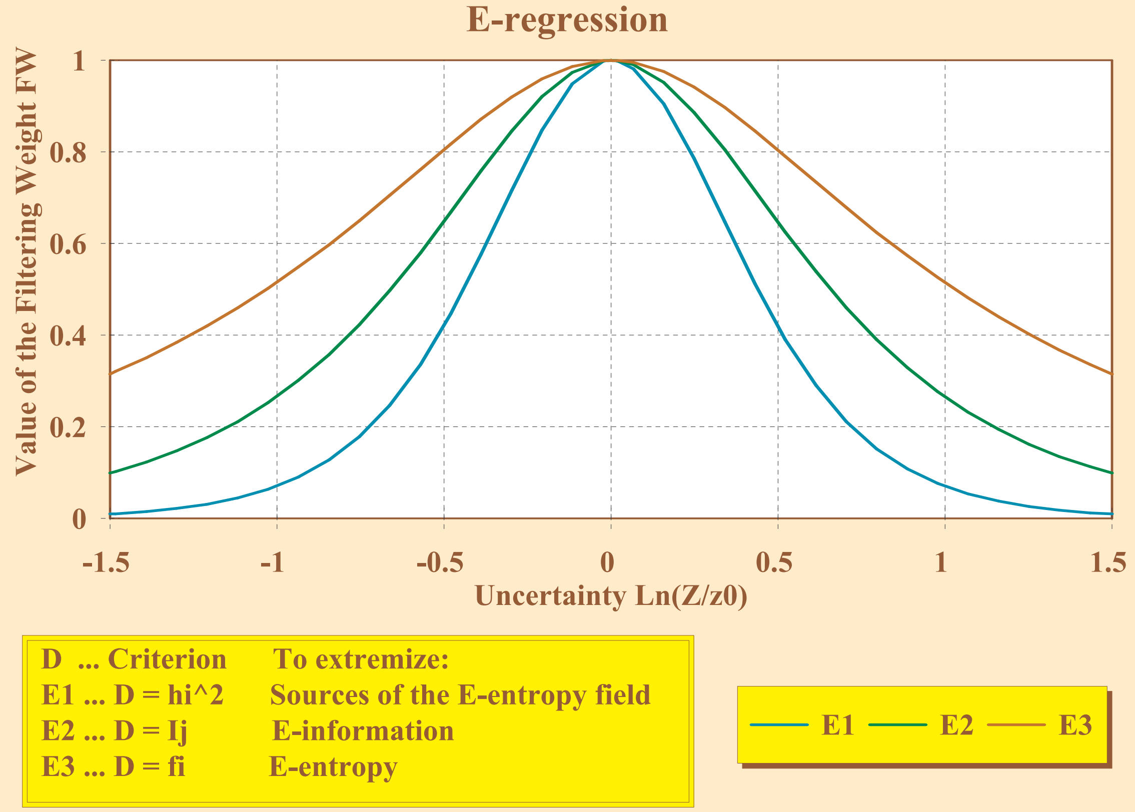
Fig.17.1: FILTERING WEIGHTS FW

1. The FW of criterion Q2 is neutral to input uncertainty: the filtering weight is constant as in the case of the classical OLS (Ordinary Least Squares) method,
2. Unlike OLS, a strong outer robustness is manifested in the case of Q1.
3. An inner robustness results in case Q3.

For E-type regressions, all three cases E1, E2 and E3 (Fig. 17.2) lead to the same kind of (inner) robustness but with a different intensity. In all (Q- and E-) cases, the FWs converge to 1, when the multiplicative residual approaches 1 (zero value of the additive residual Φ) which means, that for very weak uncertainty, the differences between it and the OLS-method vanish.

It can be shown, that—for very weak uncertainties—all (Q- and E-) error functions, E, converge to the linear function $2\frac{Z_n-\tilde{z}_{0,n}}{\tilde{z}_{0,n}}$. This can be called Euclidean relative error and it proves, that all six considered cases of gnostic regression are consistent with the classical regression methodology,

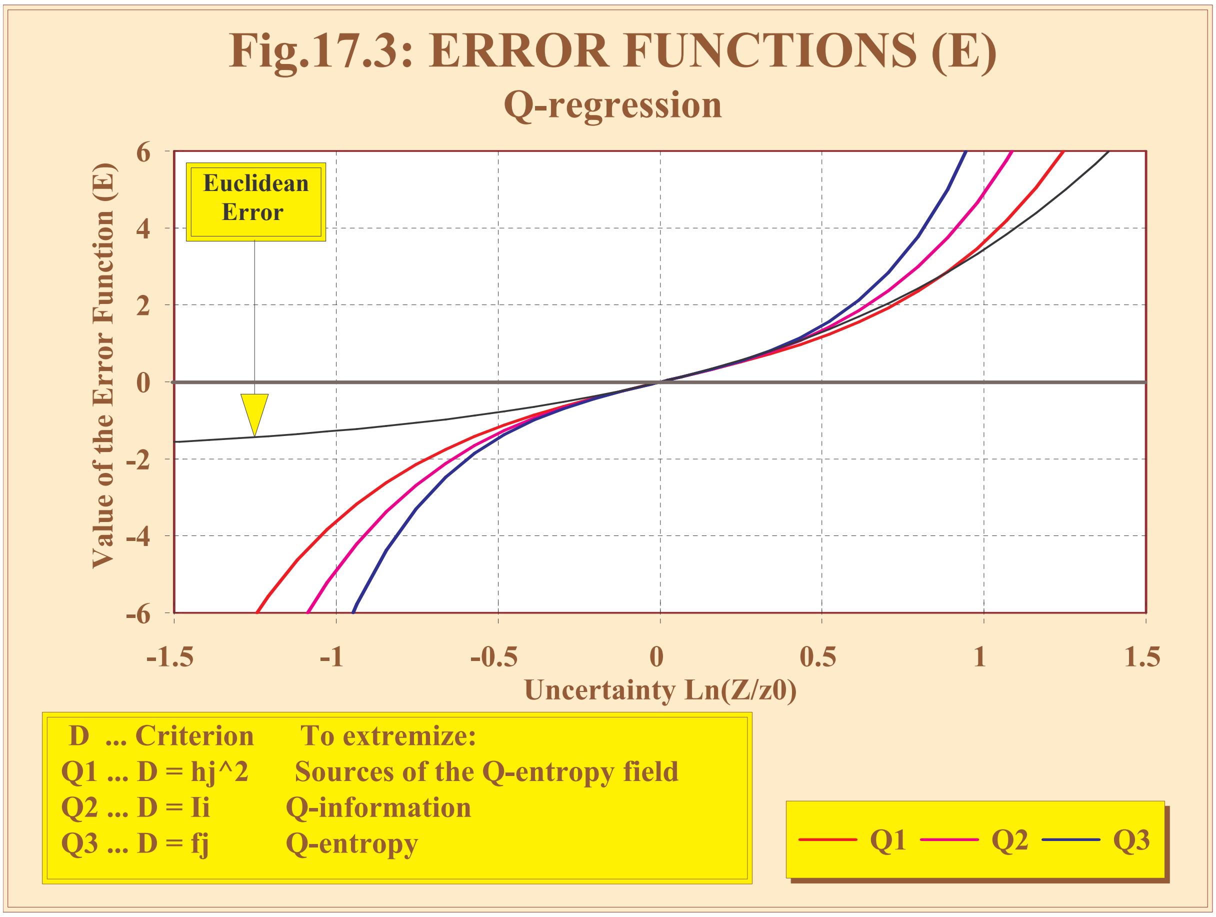
Fig.17.2: FILTERING WEIGHTS FW



if uncertainties are sufficiently weak. Moreover, the gnostic regressions were derived without the usual statistical assumptions, they have their own axiomatic justification, which makes them more generally applicable and they yield robust results.

The error functions E may be interpreted not only as “influence functions” (as already mentioned) but also as definitions of a Riemannian metric: they determine how an (output) error should be measured. To compare their behavior with that of the Euclidean case, Figs. 17.3–17.5 also show the Euclidean metric. For weak uncertainties one can again see the coincidence of the curves with the Euclidean one, however there are large differences, when strong uncertainties are present.

In all three Q-cases, the E-function strongly amplifies the effect of uncertainties thus ensuring outer robustness. It is worth noting, that there are three different combinations of input (IR) and output (OR) robustness for Q-regressions:



1. Q1 - IR: outer, OR: mixed,
2. Q2 - IR: neutral, OR: outer,
3. Q3 - IR: inner, OR: strong outer.

The robustness OR for Q1 has been called ‘mixed’, because for some middle intensity of uncertainty (Φ between roughly 0.25 and 0.85) the E function rises more slowly than the Euclidean line (more like inner than outer robustness), but for large uncertainties it clearly manifests outer robustness (see Figure 17.3).

These differences between the types of input and output robustness in the Q-versions of D can be useful in applications, where the makeup and intensities of the input and output disturbances are different.

In contrast, for E-regressions, all input (FW, Fig. 17.2) and output (E, Fig. 17.4) filtering effects demonstrate an inner robustness, but with different intensities. There are other effects to be noted with respect to the error functions in Fig. 17.4:

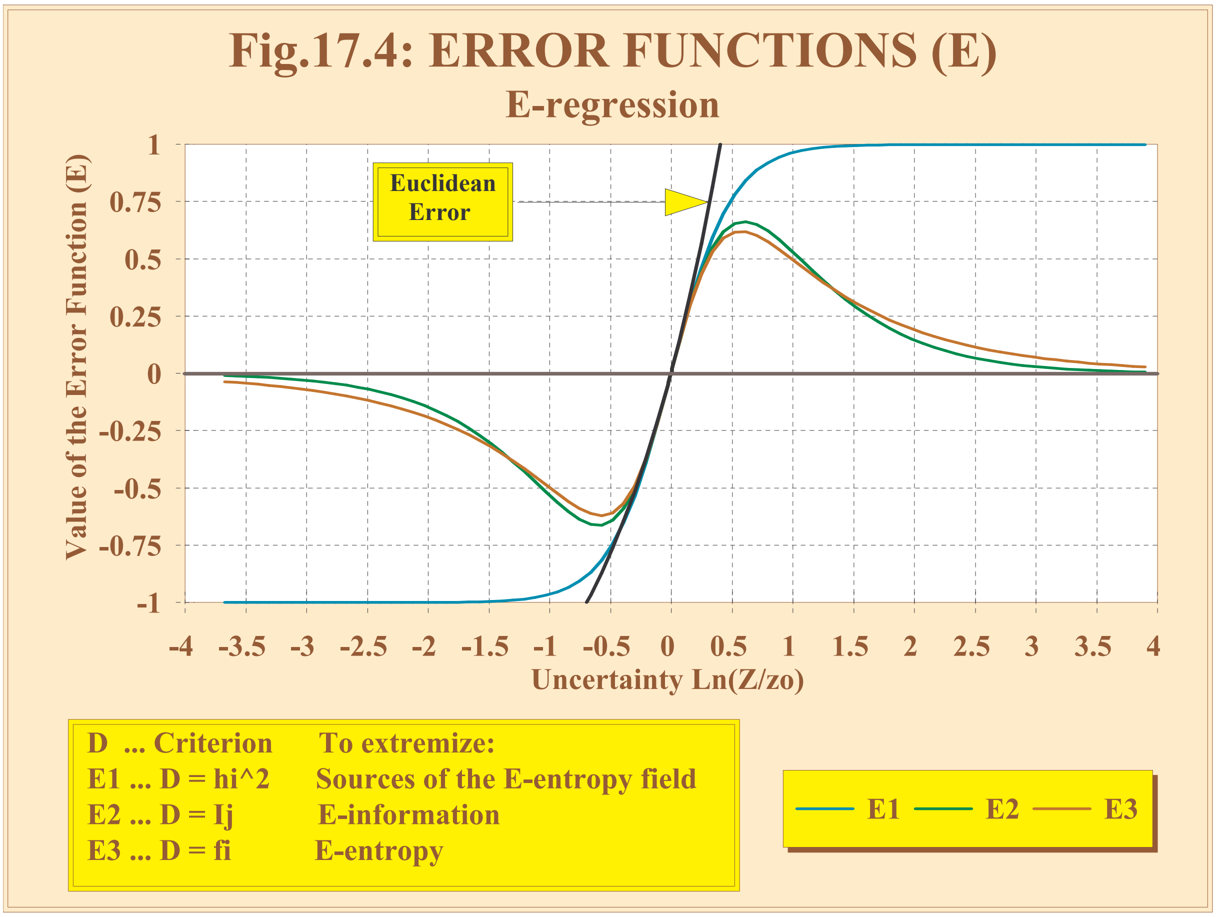


Fig.17.4: ERROR FUNCTIONS (E)

E-regression

D ... Criterion	To extremize:
E1 ... D = hi^2	Sources of the E-entropy field
E2 ... D = Ij	E-information
E3 ... D = fi	E-entropy

1. Differences between the E2 and E3 cases are very small, when only the error functions are considered: the type and the effect of output robustness is nearly the same although the criterion functions differ. This, of course, does not mean, that there is complete equivalence, because as shown in Fig. 17.2, the intensity of inner robustness is different.
2. There is a qualitative difference between the error function of E1 and the other two functions: the former behaves in a “saturating” way, while the form of the E2 and E3 error functions is “redescent”[4]. The saturating filter takes all uncertainties beyond a certain limit as “the same”, while the redescent filter completely attenuates very large uncertainties.

All six E-functions are shown in Fig. 17.5 plotted against a horizontal axis of multiplicative residuals $\frac{Z_n}{\tilde{z}_{0,n}}$ so as to compare their deviations from those

[4] To apply the term used in robust statistics to describe such influence functions.

of the Euclidean function (which shows its true linear nature).

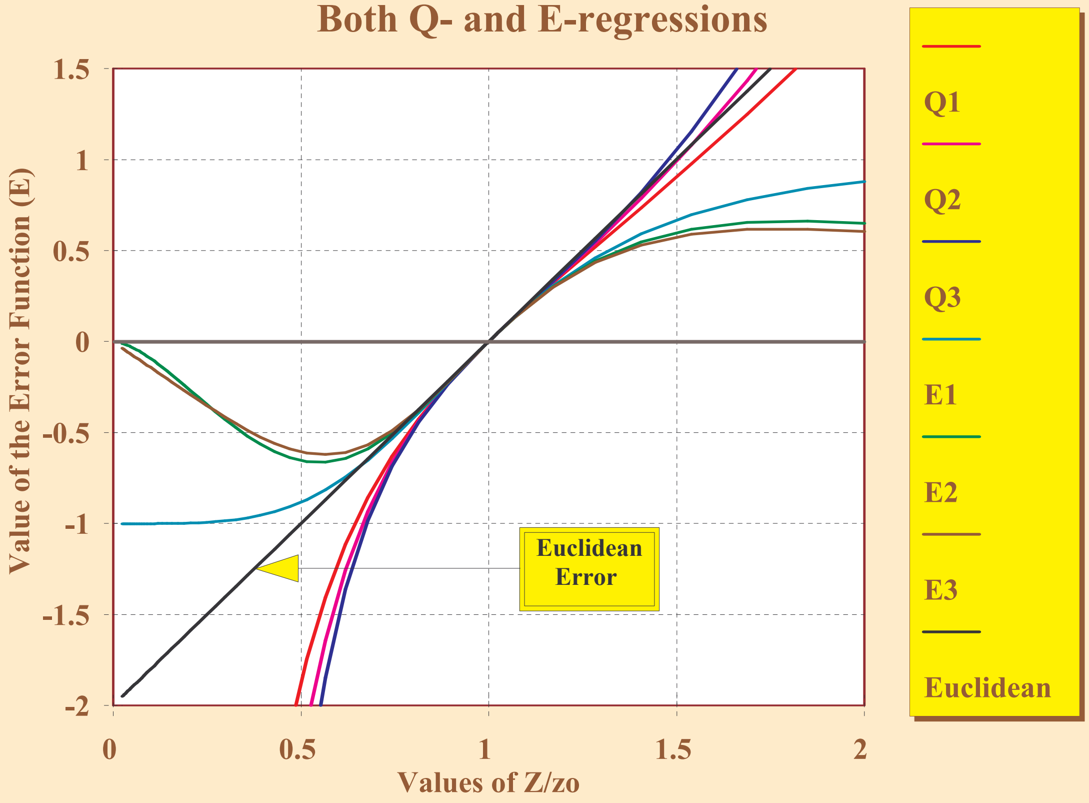


Fig.17.5: ERROR FUNCTIONS (E)

Both Q- and E-regressions

A natural question at this point relates to the choice to be made between the six types and intensities of robustness: "Which one is the best?" Each serves its own purpose and a choice may be alternatively based on

- the prior experience of an analyst with particular types of input and output data,
- the data 'speaking for themselves:' that is, to run procedures using all six versions of the gnostic criterion functions sequentially or in parallel and to measure the quality of the results of such a 'pilot' analysis so as to determine the best D, which can be subsequently applied to analogous situations.

Recall, that all six approaches are optimal, each in its own strictly defined and theoretically justified sense. An important observation should be made as to the geometrical aspects of the measuring errors connected to the regression problems. Euclidean geometry is not curved, it measures errors linearly. However, each of the gnostic error functions shown in Fig. 17.5 represents a nonlinear measuring method, which corresponds to a certain Riemannian (curved) geometry. This curvature is what provides robustness to the measurement process: the smaller the local radius of curvature, the stronger the effect of robustness. As Fig. 17.5 shows, the curvature is different at different points along the line. The slope at each of these points is determined separately, by the value of the individual datum. Another factor also influences the curvature at all the data points, the scale parameter (S), the value of which is estimated through optimization of the quality of the model's results (see Chapter 16). All this again manifests the gnostic credo: "Let the data speak for themselves."

The local curvature is determined by the gnostic functions D, which result from the fundamental features of real data as elements of the commutative group. The metric of uncertain data space is thus determined by the nature of data as mapped real quantities, by "some objectively existing regularities" as specified by Riemann a long time ago.

17.2.5 Example of a Gnostic Filter

To demonstrate the power of the approach, consider a gnostic regression, which applies the technique to a univariate case: time series robust filtering. The criterion for case E1 (see Tab. 17.1 and Tab. 17.2) has been chosen. The model 17.11 in this case has the form

$$\widetilde{Z}_{0,n} = C * Z_n, \tag{17.41}$$

where $\widetilde{Z_0}$ is the estimate of an unknown ideal data value, which should represent all uncertain data Z_n. This quantity thus plays the role of the filtered value. Parameter C is unknown. The $k+1$-th iteration to the solution 17.16 reduces to

$$C_{k+1} = C_k + \frac{S\sum_{n=1}^{N} f_{i,n}^2 h_{i,n} x_n}{2 f_{i,n}^4 x_n^2}, \tag{17.42}$$

where $h_{i,n}$ is the n-th estimating irrelevance and $f_{i,n}$ the estimating weight. Multiplying by x_n and using 17.41 one arrives at

$$\widetilde{Z}'_{0,n} = \widetilde{Z}_{0,n} + \frac{S\sum_{n=1}^{N} f_{i,n}^2 h_{i,n}}{2 f_{i,n}^4}. \tag{17.43}$$

If the data sample is a regular time series, a recursive filtering formula may be desirable to track possible changes of the current "mean" value.

A suitable method is “exponential forgetting”, which can be achieved by using a factor $\beta < 1$, so that instead of 17.43 the recursive formula

$$\widetilde{Z}_{0,N} = \widetilde{Z}_{0,N-1} + \frac{SN_{N-1} * \beta + Sf_{i,N}^2 h_{i,N}/2}{SD_{N-1} * \beta + f_{i,N}^4} \tag{17.44}$$

is applied for $N = 1, 2, \ldots$, while

$$SN_0 = SD_0 = 0. \tag{17.45}$$

To demonstrate the function of such a filter, a simulated time series has been prepared, the elements of which were generated as additive data $a = D0 + \mathcal{N}(1, 0.05)$, ie a constant $D0$ and a normally distributed pseudo-random variable with mean of 1 and standard deviation of 0.05. At random times, 20% of these data were additively contaminated by Cauchyan disturbances $d = \mathcal{C}(0, 10)$. Data to be filtered were transformed onto R_+ by exponentiation ($Z = \exp(a + d)$). The length of the series was 200, but it was composed of 4 portions of 50 elements. The quantity $D0$ was given values 0, 3, 0, 3 within the partial intervals to demonstrate the filter's transients.

The result of this simulated experiment is shown in Fig. 17.6, where the output of the linear recursive filter

$$\widetilde{Z}_{0,N} = \frac{\sum_{n=1}^{N-1} Z_n * \beta + Z_N}{(N-1) * \beta + 1} \tag{17.46}$$

with $\beta = 0.6$ is shown. It is obvious, that a linear filter is unsuitable with these gross disturbances.

Next a statistical robust filter of the L-type (known from the literature [104]), which is based on moving medians having for an odd degree V the form

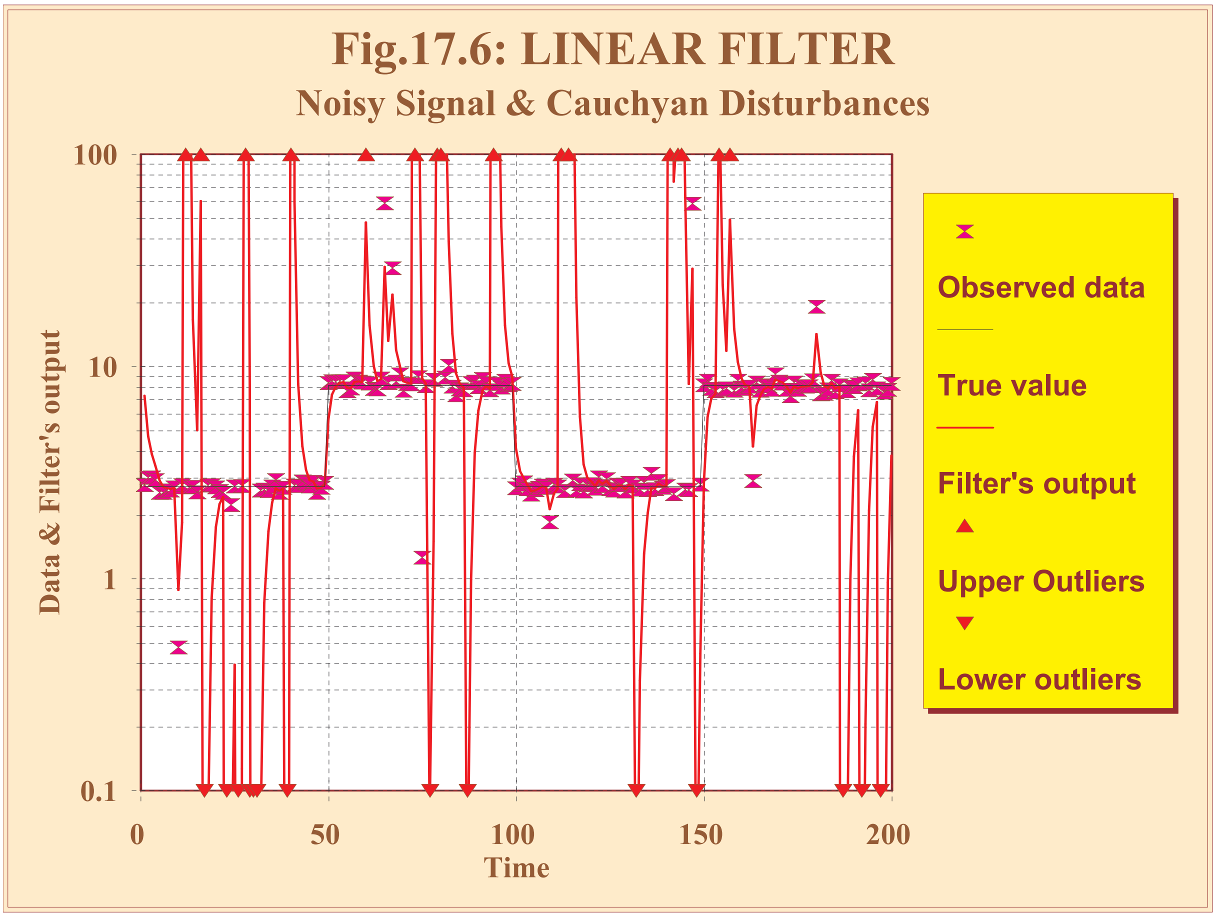
$$M(V, K) = \text{median}(Z_{K-U}, \ldots, Z_K, \ldots, Z_{K+U}), \tag{17.47}$$

where $U = (V - 1)/2$. The recursive formula of filter 53H is

$$\widetilde{Z}_{0,N} = \frac{M(5, N-2)}{4} + \frac{M(5, N-1)}{2} + \frac{M(5, N)}{4}. \tag{17.48}$$

When it is applied to the data used in Fig. 17.6, this filter suppresses the outlying data in most cases as shown in Fig. 17.7. However, it fails when there are strong disturbances following each other over short time intervals.

On the other hand, the gnostic filter (shown in Fig.,17.8) using formula 17.44 performs well even under these difficult conditions.



17.3 Non-traditional Regression Models

17.3.1 Implicit versus Explicit Regression

Returning to the definition of regression analysis cited at the beginning of the chapter, it was noted, that the separation of one of the variables as the 'dependent' one from the other ('explanatory') variables is based on knowing the nature of the dependence. There is also another important assumption: that the dependent variable to be modeled is only one-way. In other words, the 'dependent' variable has no influence on the 'explanatory' variables, there is no 'feed-back.' Regression models of this type can be called *explicit*, because they are based on explicit equations. There are at least three problems with respect to explicit regression models:

1. In the real world strictly one-way dependencies are the exception rather than the norm.

Fig.17.7: ROBUST FILTER 53H

Noisy Signal & Cauchyan Disturbances

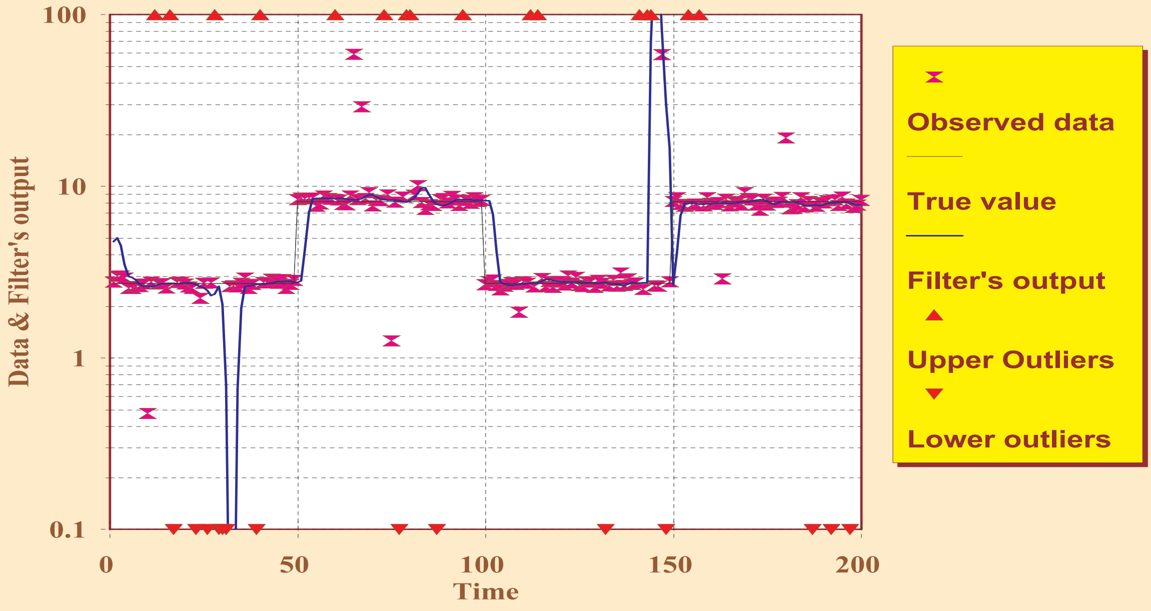
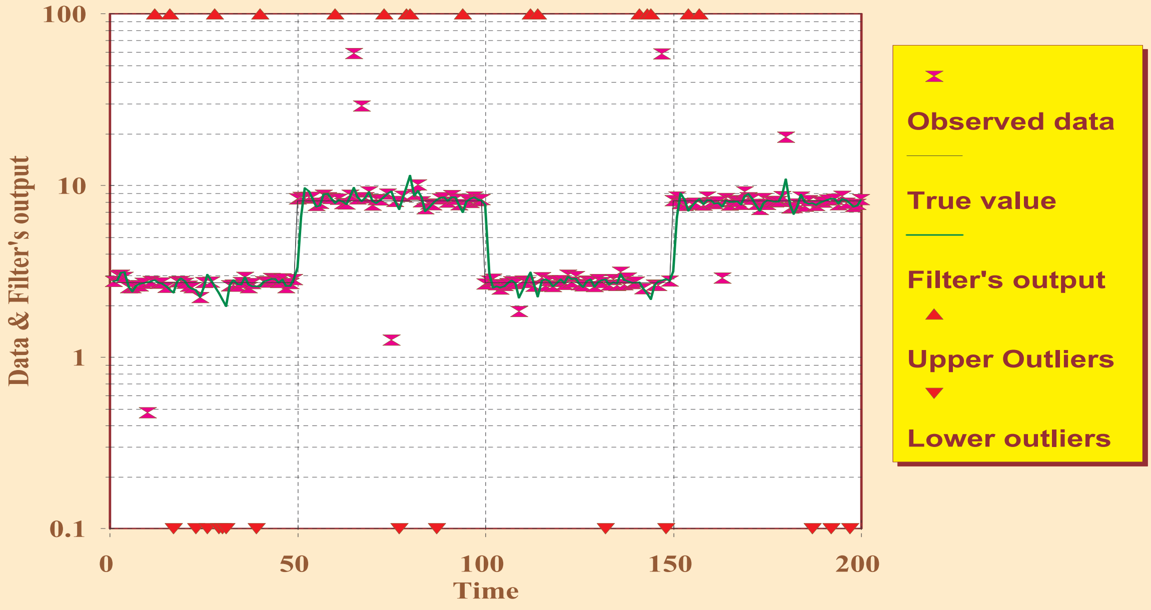


Fig.17.8: GNOSTIC FILTER

Noisy Signal & Cauchyan Disturbances



2. It is not always possible to solve the model's equations with respect to the 'dependent' variable, even if such a variable exists.
3. Solutions of overdetermined systems of equations—such as explicit regression equations—can be only approximate. This results in inconsistencies if one attempts to exchange the roles of 'dependent' and 'explanatory' variables.

Indeed, the mathematical definition of a dependence as a function easily introduces a strictly one-way action: 'arguments → dependent variable.' Modeling real processes is frequently far from easy, because each of the variables being considered is dependent on others. Vivid examples can be found in financial statement analysis:

Profitability is a frequent object of interest to analysts and financial managers, who are interested in discovering, how it depends on other factors such as financial leverage, various turnover relationships, working capital, etc. A regression with profitability as the dependent variable is expected to estimate these effects:

$$R_{PR,k} = C_0 + C_1 * R_{RWC,k} + C_2 * R_{TATO,k} + C_3 * R_{FL,k}. \quad (k = 1, \ldots, K) \tag{17.49}$$

Each of the variables $R_{*,k}$ are financial ratios of individual firms: PR is profitability, RWC the relative value of working capital, $TATO$ total asset turnover and FL financial leverage. All of these financial parameters are mutually dependent. When there is a good profit margin, a manager may decide to decrease financial leverage, to improve liquidity, to accelerate the total asset turnover by additional investment (if the demand exists) as well as to take other positive measures, because there are sufficient financial resources. This illustrates the multidimensional "feed-backs", which cause the explanatory ratios to be dependent on the profitability as well as on each other. There are methods in control theory to solve such problems, but their application could be even more complex than the initial problem.

The solution of an explicit regression task (17.49) suffers from another serious drawback: Since the firm's growth is also dependent on the level of the working capital, why not evaluate the dependence of working capital on the other financial parameters using a regression such as 17.50 with working capital as the dependent variable?

$$R_{RWC,k} = c_0 + c_1 * R_{PR,k} + c_2 * R_{TATO,k} + c_3 * R_{FL,k}. \tag{17.50}$$

Once the parameters have been estimated, the profitability can be expressed as

$$R_{PR,k} = (R_{RWC,k} - c_0 - c_2 * R_{TATO,k} - c_3 * R_{FL,k})/c_1. \qquad (17.51)$$

The problem here is, that the coefficients $1/c_1$, $-c_0/c_1$, $-c_2/c_1$, and $-c_3/c_1$ are not the same as those of the regression 17.49.

This unpleasant inconsistency is easily explained. A system of regression equations ordinarily contains more equations than unknown coefficients ($K > 3$ in this case) so as to minimize the uncertainty of the solution. The solution will then depend on using an optimization method and the result is, that the uncertainty of the observed values of the variables in equations 17.49 and 17.50 is suppressed differently in each regression even if the same optimization method is applied. Obviously, when different results are obtained by the same method from the same data, deciding, which of the two solutions is the true one, is difficult.

The conclusion to be drawn here is, that dividing the variables of an explicit regression into categories of 'dependent' and 'explanatory' is seldom theoretically consistent, because any of them can be either 'dependent' or 'explanatory' depending on how the problem is set out. Such an analysis introduces an asymmetry into the solution, which leads to inconsistencies in many practical cases.

The desired symmetry in the roles of all the variables can be achieved by using the *implicit form* of a regression. Note, that if $C_0 \neq 0$, then equation 17.49 can be rewritten in the form

$$R_{k,N}/C_0 - \sum_{n=1}^{N-1} C_k/C_0 * R_{k,n} = 1 \quad (k = 1, \ldots, K) \qquad (17.52)$$

or—in a new notation—

$$\sum_{n=1}^{N} C'_n * R_{k,n} = 1 \quad (k = 1, \ldots, K). \qquad (17.53)$$

Here, all the variables play the same role. Once a solution for coefficients C'_* is obtained, 17.53 can be used to express any desired variable as an explicit function of the others without any danger of inconsistency. Further, a single 'universal' solution is preferable, rather than a different one for each "dependent" variable.

The problems associated with explicit regressions in economic analyzes are due to the fact, that the inner interactions of the variables of an economic system are nearly as complex as those of a living organism. Indeed, it is impossible to state, that any one single parameter (such as eg temperature, blood pressure, pulse rate, electric potential, number of blood cells, or the composition of fluids) of a living creature is “only dependent” on other variables and does not also influence them. These same problems also exist in other application fields. The above example using financial statement analysis was linear; however, the same difficulties are also present in the more usual non-linear cases. Problems of the same nature also exist in other application fields. It would seem, that some important Laws of Nature are formulated in an implicit form, and that there are similar non-linear interdependencies in economics and in the other social sciences. Returning to the gnostic approach to the regression problem, it is evident, that the discussed formulation can include both explicit and implicit forms. The explicit case was considered in detail, but to undertake an implicit regression, it is only necessary to substitute a vector of “1’s” for the observed ‘output’ values, $z_{0,n}$.

Calculations of an implicit regression can require modifications of algorithms ordinarily used in variance analysis, because they assume a non-zero variance of the ‘dependent’ variable.

17.3.2 Regression in Probabilities

It was shown in subsection 14.3.6, that a univariate regression is closely connected with the idea of the (linear) similarity of samples measured by the cross-covariance of the ‘dependent’ with the ’explanatory’ variables normalized by the variance of the latter. This interpretation can also be applied to the multidimensional linear regression:

Explicit regression: A linear combination of explanatory variables should be similar to the dependent variable.

Implicit regression: A linear combination of explanatory variables should be similar to a constant (eg to 1).

The OLS (Ordinary Least Squares) solution of the regression problem is a matrix composed of the cross-covariances and variances of all the variables. Using the reasoning of subsection 14.3.6, it can be concluded, that the OLS method resulting from the ordinarily defined covariances and variances is based on the Euclidean measure of 1D (one-dimensional) errors and MD (multidimensional) lengths. The application of Riemannian geometry of the gnostic type leads to a linear relation 14.46 between irrelevances. The

MD extension of this relation is

$$h_c(y_n) = \sum_{m=1}^{M} C_m * h_c(x_{n,m}), \tag{17.54}$$

where the index m identifies each of M explanatory variables and $n = 1, ..., N$ denotes the number of the equation. For $N > M$, the equation system is overdetermined and its OLS solution applies gnostic covariances and variances such as those in 14.46, but this time in the matrix form. As shown in 14.3.6, the linear dependence of irrelevance on probability 10.42 in the estimating case enables the equations 17.54 to be rewritten in the form

$$P(y_n) = C_0 + \sum_{m=1}^{M} C_m * P(x_{n,m}), \tag{17.55}$$

where $P(\alpha)$ is the probability of the value α, and where C_0 is the constant

$$C_0 = 1 - \sum_{m=1}^{M} C_m. \tag{17.56}$$

Regression 17.55 will be called a *regression in probabilities* in both linear and nonlinear applications.

The relation 10.42 also establishes the linear dependence of improbability on the quantifying irrelevance. An analog to equation 17.55 can therefore be used for improbability in the quantifying case. The difference between the estimating and quantifying version will lie in the type of robustness:

estimating irrelevances: probabilities, variances and covariances are robust with respect to outliers.

quantifying irrelevances: improbabilities and other quantifying characteristics are robust with respect to inliers.

Variables in regression models may frequently have a different range of values. Financial leverage $\frac{Total\ Debt}{Total\ Assets}$ can take on values in the interval $[0, 1]$, while the price-earnings ratio $PE = Price/EPS$ could theoretically reach any value in the interval $(-\infty, \infty)$. Not only do these different ranges make the interpretation of model coefficients difficult, but the magnitude of their values is not comparable, the scales are different. Other problems arise from the physical dimensions and the measurement units of the variables: some (eg financial leverage) are dimensionless, some may be expressed in percentage form (eg profitability), while others are generally given in physical units (a turnover dependent on a time unit, book value of equity on a monetary unit). All of these effects contribute to difficulty in making valid comparisons between the coefficients of the regression model, in which variables are quantified using different physical units. Further, the use of “natural” measurement units for variables prevents an easy survey of dependencies. It is therefore desirable to unify the measuring scales. All of these desirable effects can be obtained by using the probabilities instead of ‘natural’ variables:

1. The range of values of all the transformed variables is the same: $[0, 1]$.
2. All transformed variables are dimensionless.
3. All the coefficients of a (linear, implicit) regression are directly expressed as the weights of the transformed variables (probabilities), they are comparable.
4. The probabilistic transformation made by means of the EGDF (Estimating Global Distribution Function)— due to its inner robustness— efficiently filters the data, so that the effect of outliers is suppressed.
5. Important by-products of the application of EGDFs are:
   (a) Estimates of the bounds of the values of all variables are obtained.
   (b) Possible data inhomogeneity is revealed and eliminated.
   (c) Data censoring can be accounted for (see Chapter 19).

The effect of robustness due to the EGDF differs from the effects of the filtering weights FW and the error function E (described in 17.2.3) in an important aspect. Both latter functions are dependent on the multiplicative residual $Z_n/z_{0,n}$, ie on “inner” relations, ‘input/output’ of the n-th equation. One can therefore speak of *row-wise filtering.* Note, that the FW gives the same weight to all of the components of the gradient $\underline{g}$ 17.34, but the robustness of the EGDF affects each element of this gradient in a different way, because it redistributes the weights within a column, thus realizing *column-wise filtering.* In the case of an explicit regression in probabilities, the values of the error function E are also substituted by their probabilities; they are therefore also filtered column-wise. This means, that a gnostic regression in probabilities ensures the **double filtering** of the uncertainty of data. The effect of double filtering can be expected to contribute to the robustness of the method.

The notion of gnostic covariances was introduced with the modulus of a data sample (Definition 14, 14.19 and Corollary 15.3, 14.21). The usefulness of these notions is emphasized by their connections to the measurement of dissimilarity analyzed in 14.3.6 and by their application to regression in probabilities.

17.4 Summary

The gnostic approach to modeling interdependencies between variables differs substantially from statistical methods:

1. no a priori statistical assumptions on data are applied,
2. all variables are represented by data samples, which are the only information given on the variables,
3. both dependent and explanatory variables may contain unknown uncertainties,
4. instead of a mean or average value for the dependent variable, its distribution function is estimated or predicted,
5. the use of a very practical and advantageous regression in probabilities is theoretically justified,
6. the implicit regression model extends the application field of the ordinary (explicit) versions,
7. the gnostic characteristics used as criterion functions ensure, that the model will possess the unusual qualities of inner/outer robustness, and the minimization of losses of information.

The problem formulation is general enough to cover not only linear, but also nonlinear multivariate regression functions applied both to cross-section and time-series models. The influence functions of errors were examined for six types of gnostic criteria by using an iterative solution for the equations resulting from a variated model function. It was shown, that this approach provides several useful kinds of influence functions, which ensure a saturating, redescent or expanding reaction to data uncertainties. The nonlinearity of gnostic criteria leads thus to effects, that are comparable to those of nonlinear filters resulting in the protection of the estimated model parameters in both explanatory and dependent variables against the influence of uncertain components. The characteristics of these filters are optimized, because they extremize the chosen gnostic optimality criterion. This choice determines the type of robustness of the model.

Classical covariances can be interpreted as “by-products” of the solution of the regression task. Their drawback is the sensitivity to outliers, especially when they are estimated from small data samples. Gnostic regression models use robust covariances originally introduced in connection with the modulus of a data sample (see chapter 14).

Chapter 18

Optimal Numerical Operators

18.1 A Side-step to Statistics?

The formulation of the regression problem in the previous chapter was general enough to cover not only static (cross-section) problems, but also dynamic problems such as uni- and multivariable time-series processing. In practice, there are a broad spectrum of problems to be solved, therefore it is useful to accept as broad a definition of "processing" as possible. From the previous chapter, it might appear, that there were only three classes of tasks to be considered, the explicit as well as the implicit regression, and the regression in probabilities. The "dependent" or "output" variable was z in the former and a constant (1) in the two latter cases. To show, that the concept set out in 17.2 really covers a more extensive range of problems, the more detailed analysis which follows will demonstrate, that special numerical operators are needed, and that an appeal to statistics will be useful.

18.1.1 Diversity of Modeling Problems

An explanatory vector $\underline{x}$ can take on a different character depending on the nature of the problem. Consider three examples of an "input" vector of length M:

1.
$$\underline{x}_{CS,n,t} = \langle x_{n,1}(t), x_{n,2}(t), \ldots, x_{n,M}(t)\rangle \quad (n = 1, \ldots N). \qquad (18.1)$$

The set of M different input values of the n-th variable describes the vector's instantaneous **state** at time t. The $N \times M$ matrix composed of these rows characterizes the state of the N objects, the *cross-section* of the group.

2.
$$\underline{x}_{TS,t} = \langle x(t), x(t-1), \dots, x(t-M+1)\rangle. \tag{18.2}$$

This vector describes the "history" of a certain input variable, its value at time t along with the $M-1$ lagged values. If there are T such vectors ($t = 1, \dots, T$), then the $T \times M$ matrix composed of these rows characterizes the *dynamics* of the variable.

3.
$$\begin{aligned}\underline{x}_{CSTS,M,N,t} = \langle & x_1(t), x_1(t-1), \dots, x_1(t-M+1),\\ & x_2(t), x_2(t-1), \dots, x_2(t-M+1),\\ & \dots, \quad \dots, \quad \dots, \quad \dots,\\ & x_N(t), x_N(t-1), \dots, x_N(t-M+1)\rangle.\end{aligned} \tag{18.3}$$

This vector describes the "history" of a group of N objects, their values at time t and their $M-1$ lagged values. If there are T such vectors ($t = 1, \dots, T$) then the $T \times (N \times M)$ matrix composed of these rows describes the *dynamics of the cross-section.*

The "historical" vectors 18.2 and 18.3 apply a *moving window* strategy, which assigns to all $M-1$ lagged components the same a priori weight as the index value (equal to 1), while the higher lags receive a zero weight; they are "completely forgotten." While this concept has some merit, it is not completely satisfying since there easily arise questions such as, "What is the difference in significance between the values of $x(t-M+1)$ and $x(t-M)$, which justify the inclusion of the former, while excluding the latter?" A useful method for overcoming this problem is *exponential forgetting*, which was introduced in the example of the gnostic filter (Section 17.3). The concept is to apply a *forgetting factor* β ($0 < \beta < 1$), such that eg the sum of the vectors' $\underline{x}_{TS,t}$ of the exponentially "forgotten" components has the form

$$\Sigma_{x,t} = \sum_{m=0}^{M-1} \beta^m * x(t-m). \tag{18.4}$$

This technique and its modifications result in a smoothing of the forgetting effects. Another advantage is the recursive character:

$$\Sigma_{x,t+1} = x(t+1) + \beta * \Sigma_{x,t}. \tag{18.5}$$

The model (17.2) can be thus interpreted as

$$z_* = F(\underline{C}, \underline{x_*}), \tag{18.6}$$

where $\underline{x}_*$ is 18.1, 18.2 or 18.3, and where z_* denotes one of the many possible forms. The simplest representation of the dependent variable z_* is related to a cross-section analysis using an explicit regression with explanatory vectors having the form of 18.1, while the dependent variable is another parameter of the same (n-th) object at the same time t. (Example: The dependence of the return on assets (ROA) on other financial ratios for the same year.) However, there are a variety of alternatives:

1. The explanatory vector is 18.2 and the dependent variable is $z_* = \tilde{x}(t)$, ie the best estimate of the last value of the time series. This is the case of *filtering.* Example: given a series of volatile observed share prices, estimate a recent "true" value.
2. The same explanatory vector, but the dependent variable is now represented by $z_* = \tilde{x}(t+\tau)$, where $\tau > 0$. This is the case of *time-series prediction.* Example: given a series of sales of a product, predict a future value of sales.
3. The explanatory vector is now 18.1 and the dependent variable is $z_* = \tilde{x}_{n,M+1}(t+\tau)$, where $x_{n,M+1}$ is another parameter of the n-th object and the task is the *prediction of the object's parameter based on a given cross-section.* Example: given sets of financial ratios of an industry for a year Y, predict industry sales for the year $Y + \tau$.
4. The explanatory vector is 18.3 and the dependent variable is $z_* = \tilde{x}_{n,M+1}(t+\tau)$, where $x_{n,M+1}$ is another parameter of the n-th object and the task is to *predict an object's parameter based on the given time series of cross-sections.* Example: given sets of financial ratios of an industry for years Y, $Y-1, \ldots, Y-M+1$, predict the share price of the n-th firm for the year $Y+\tau$.

In all the above examples, the data are directly observable. There may be an objection, that the datum $x(t+\tau)$ has not yet occurred at time t, however, there are ordinarily three time intervals associated with the treatment of time series([27]):

1. the model estimation period,
2. the ex post forecast period,
3. the ex ante forecast period.

Over the first period, the initial portion of the "historical" time series of explanatory and dependent variables is used to estimate the model's parameters. The rest of the historical time series is used for verification of the model's applicability and for the estimation of its errors. The values of the dependent variable are thus known for both of these phases. The model is not really "used" until the third period, when a forecast of *future* values is desired: given the input, predict the (unknown) output. Information on the model's errors can be purged from the third period, but only after the true value of the output becomes known, however these techniques are usable only when the data are directly observable.

There are tasks, when either the variables' input or output values are not directly observable; these must be mathematically derived from data. Let us confine ourselves to **linear operations** on continuous variables and to their discrete (numerical) representations, to *linear numerical operators.* Although constrained, the playground for this game is still large. An example is in order: it is well-known, that investments in science, research and development, in technology, know-how, skill of employees, marketing expenses and other factors accelerate sales after a time interval. However, this acceleration is not directly observable, it must be mathematically derived from the time series of sales. In the continuous case, acceleration is proportional to the second time derivative of the state function. In the case of discrete function values (which are given at regularly distributed time intervals), the acceleration is proportional to the second difference of this function. However—as it is well known—differentiation amplifies noise (or data errors). In order to minimize this effect, in a statistical analysis *redundant formulae* are developed, ie formulae, which make use of more data than is necessary for a purely analytical minimum.

There are cases, when neither the dependent variable nor the explanatory variables are directly observable. Example: the short-term financial situation of a firm is dependent on the rate of cash inflow and outflow. These events are discrete (and perhaps highly variable), so that to obtain reliable estimates of these flows, they must be smoothed. Such a model can eg evaluate the current need for short-term financial requirements based on smoothed cash flows and their rates of change. The smoothing instruments (filters) are numerical operators, which treat the time series of cash flow. In other fields of application a need for directly unobservable quantities can also exist. An example might be that of a feedback control system, which uses velocity, acceleration and/or integral signals for its stabilization and optimization.

This need for numerical operators can also be met, when implicit models are used. Regularities with respect to relations between variables can be sometimes expressed by an equation, which has zero on its right-hand side. (This "dependent" variable becomes $z_* \equiv 1$ after exponentiation of the equation transforming the task onto the infinite data support R_+.) These situations occur in the case of certain balancing statements. Some of the elements to be balanced may not be directly observable, in which case they must be drawn out from the time series by numerical operators.

Because gnostics has been presented as a method, which surpasses statistics in yielding information from small samples of strongly dispersed data, one may wonder, whether it is consistent to apply statistical instruments such as (statistically) best numerical operators in combination with gnostic methods. The answer is positive:

1. data series can be subjected to robust filtering by gnostic methods before numerical operators are applied so as to minimize the effects of possible gross data errors,
2. this robust filtering of data reduces data uncertainty to such a low level, that statistical methods—as shown by gnostics—yield results, which approach the best possible quality,
3. gnostic methods allow the reliable estimation of covariances needed for some numerical operators.

It will be seen from the material that follows, that the most effective analysis of this type will result from the joint use of both statistical and gnostic procedures, because the advantages of each methodology can be exploited. There are, then, good reasons to consider in more detail the problem of optimal numerical operators, by which the directly unobservable quantities can be derived from data. This will not only provide instruments useful to the extension of the application field of the gnostic methods, but also offer a better insight into the eternal battle of man versus uncertainty.

18.1.2 A Short History of the Problem

An important contribution to the solution of the problem of filtering random sequences was made by A.N. Kolmogorov in 1941 ([46]). This class of problems was also secretly studied during World War II in connection with the development of radar. Some of the results of this work were published in the famous post-war book of Norbert Wiener ([116]). Wiener's filtering problem assumed a random signal, which had the form of a stationary continuous function. This concept was extended by L.A. Zadeh and J. Ragazzini, who added a non-random component to the continuous signal ([119]). The optimality of the linear forms of least-squares estimates was proved by the Neumann-David theorem [20]. The rapid development of digital computers then turned attention back to discrete time series and R.E. Kalman's recursive filter ([41]) opened a new line of development to the state-space approach, which was later shown by P. Swerling ([107])

to be a version of the known recursive solution of the least-squares problem. A discrete version of the Zadeh-Ragazzini problem was considered by M. Blum ([10]) and generalized by P. Kovanic ([47]). Different aspects of treating discrete signals composed of a stationary random component and a non-random non-stationary component were analyzed in [68], [38], [10] and [50]. A generalization of the Gauss-Markov theorem [69] subsequently extended the application field to the estimation of correlated signals.

In 1954, V.M. Semyonov ([98]) made a substantial contribution by considering a mixture of a "useful" random signal (represented by a polynomial, which had random coefficients) and a stationary random noise. Unlike the traditional concept of BLUE (the Best Linear Unbiased Estimate), Semyonov's approach dealt with unconstrained minimum variance estimates. Such estimates are ordinarily biased, but their variance can be lower than the well-known Cramer-Rao low bound of BLUE. The minimum-penalty concept [52] and [53] allowed the integration of all nonrecursive least-squares methods into one generalized estimate. This method permitted both unconstrained and unbiased estimates to be obtained as extreme cases of a generalized estimate by choosing the value of the penalty, which determined the weights of both types of estimates. The universality of this approach makes it useful as a base for considering the theory of optimal linear operators.

18.2 The Minimum Penalty Estimate

18.2.1 Definitions and Notation

The matrix form will be used to represent the data samples for cases of both time series and data sets, which are not regularly distributed in time or space. As usual, $\underline{M}^T$ is the transposed matrix $\underline{M}$ and $\underline{M}^+$ is the Moore-Penrose pseudo-inverse ([89]) of $\underline{M}$.

Definition 17: Let $\mathcal{S}_t$ be a set of N real numbers t_n $(n = 1, \ldots, N)$, such that $t_B \leq t_n \leq t_E$. Let $\mathcal{I}_t$ be the interval $[t_B, t_E]$ of real numbers. Consider a function $\xi(t) : \mathcal{I}_t \to R^1$. The vector $\underline{x} = \langle \xi(t_1), \ldots, \xi(t_N) \rangle$ will then be called the *vector representation* of the function ξ over the support $\mathcal{S}_t$.

Consider a set $\mathcal{S}_x$ of M functions $\xi_m(t)$ $m = 1, \ldots, M$ of the aforementioned type. Let $\mathcal{L}$ be a linear functional $\mathcal{L} : \mathcal{S}_x \to R^1$. Then the row-vector $\underline{L} := \langle \mathcal{L}(\xi_1(t)), \ldots, \mathcal{L}(\xi_M(t)) \rangle$ will be called the *numerical operator.* Let the data model be a real column vector

$$\underline{Y} := \underline{Y}_x + \underline{Y}_e, \tag{18.7}$$

where $\underline{Y}_x$ is the vector representation of the true and $\underline{Y}_e$ of the error component of a function represented by data, which may be made up of both random and non-random elements. Let the means of the random components be zero.

Denote the mathematical expectation of a random matrix $\underline{Q}$ by $\widehat{\underline{Q}}$. It is assumed, that all the first and second statistical moments of the variables under consideration are known and constant. Let $\widehat{\underline{Y_x Y_e^T}} = 0$. Let $\underline{A}$ be a zero-mean random column-vector, for which relation

$$\widehat{\underline{AA^T}} = \underline{I} \tag{18.8}$$

(with the identity matrix $\underline{I}$) holds, so that the covariance matrix of the information component $\underline{Y_x} = \underline{XA}$ (for a non-random $N \times M$ matrix $\underline{X}$) is

$$\widehat{\underline{Y_x Y_x^T}} = \underline{XX^T}. \tag{18.9}$$

Let the covariance matrix of the noise component be

$$\underline{B} := \widehat{\underline{\hat{Y}_e \hat{Y}_e}^T}. \tag{18.10}$$

and let the required result of the estimation be

$$Z_x = \underline{LA}. \tag{18.11}$$

Let a scalar

$$\|\underline{Q}\| := \text{trace}\{\sqrt{\widehat{\underline{Q}}}\} \tag{18.12}$$

be a measure of a random matrix $\underline{Q}$. Let a constant (row-vector) operator $\underline{W}$ be applied to perform the estimation by means of the linear form

$$\underline{Z} = \underline{WY} + \underline{C} \tag{18.13}$$

with a constant vector $\underline{C}$. Let the *estimation error of the first kind and of the second kind* be respectively

$$e_x := \|\underline{W}\underline{Y_x} + \underline{C} - \underline{Z}_x\| \tag{18.14}$$

and

$$e_y := \|\underline{WY} + \underline{C}\|. \tag{18.15}$$

Let the *penalty* p be a scalar

$$p := p_x e_x^2 + p_y e_y^2, \tag{18.16}$$

where p_x and p_y are non-negative weights. Let

$$r := p_y/(p_x + p_y). \tag{18.17}$$

A *centralized* random vector is then denoted by $\hat{\underline{V}} := \underline{V} - \widehat{\underline{V}}$.

Using centralized variables simplifies (18.13) to

$$\hat{\underline{Z}} = \underline{W\hat{Y}}. \tag{18.18}$$

The decomposition in 18.9 is not unique, but applies to all $\underline{X}$, which satisfy 18.9. Columns of the matrix $\underline{X}$ can be considered as vector representations of the already mentioned functions $\xi_m(t)$. Matrix $\underline{X}$ defines a subspace of the vector space $R^N \times R^N$ by its columns: it may be called the *basis* of this subspace.

The rank of the covariance matrix $\underline{B}$ can be full (equaling N) or less, depending on the features of the noisy data component $\underline{Y}_y$, which contaminates the data. The errors of treatment of both data components are given a weight by the penalty to express the preference of one kind of error over the other.

18.2.2 The Main Result

Theorem 19: Let the definitions and notation introduced above hold. Then the minimum penalty estimator has the form

$$\underline{W} = \underline{L}(r\underline{I} + \underline{X}^T\underline{B}^+\underline{X})^+\underline{X}^T\underline{B}^+. \tag{18.19}$$

More detail on this theorem and on its more general versions as well as its proof can be found in [52] and [53]. The most important types of the vector $\underline{L}$ are given in [50]. To more fully understand its role, interpret the elements of the matrix $\underline{X}$ (numbers $x_{k,m}$) as values of the set $\mathcal{S}_x$ of M differentiable functions ξ_m defined above, so that $x_{k,m} = \xi_m(t_k)$ for $k = 1, .., N$ and $m = 1, .., M$. Cases of special interest are determined by the following components of the numerical operators $\underline{L}$:

Analysis: $L_{A,l} = 1$ for $m = l$, $L_{A,l} = 0$ for $m \neq l$, $(m, l = 1, .., M)$. The results of estimation are weights A_m of functions $\xi_m(t)$.

Smoothing: $L_{S,m}(t_*) = \xi_m(t_*)$ $(m = 1, \ldots, M,\ M < N)$. This operation includes **filtering** $(t_* = t_N)$, **smooth interpolation** $(t_1 \leq t_* < t_N)$ and **smooth extrapolation** $(t_* < t_1$ or $t_N < t_*)$. (There would be no smoothing effect for $N = M$. Such a case would correspond to ordinary Lagrange's interpolation or extrapolation.) The quantity t_* will be called the *target point.*

Smoothing differentiation:

$$L_{D^n,m}(t_*) = \frac{d^n \xi_m(t)}{dt^n}|_{t=t_*}. \tag{18.20}$$

The resulting derivatives can be interpolated or extrapolated depending on the chosen value of t_*.

Smoothing integration:

$$L_{I,m}(t_*) = \int_{t_0}^{t_*} \xi_m(t) dt. \tag{18.21}$$

Smoothing convolution: Given a kernel function $f : R_1 \to R_1$. Then

$$L_{C,m}(t_*) = \int_{t_0}^{t_*} f(t) \xi_m(t_* - t) dt. \tag{18.22}$$

It is well-known, that such convoluted integrals simulate the reaction of a dynamic object (the impulse response function of which is $f(t)$) to the input ξ_m.

18.2.3 On Signal-to-Noise Ratios

An important role is played in the formula for optimal linear operators 18.19 by the term $\underline{X}^T \underline{B}^+ \underline{X}$, which can be interpreted as a *matrix signal-to-noise ratio* of the data being considered. Indeed, the matrix $\underline{B}$ is the covariance matrix of noise 18.18, while $\underline{X}$ is the square-root of the covariance matrix of the "signal" (informative, true) component. For some square matrices $\underline{M}_1$ and $\underline{M}_2$, both expressions $\underline{M}_1 \underline{M}_2^+$ and $\underline{M}_2^+ \underline{M}_1$ have some of the features of a matrix ratio. However, these expressions may be asymmetric even in cases, when both matrices are symmetric.

In contrast, by definition, expression $\underline{X}^T \underline{B}^+ \underline{X}$ is always symmetric. It also has the important feature of a matrix ratio, that multiplication of both the "numerator" ($\underline{X}^T\underline{X}$) and "denominator" $\underline{B}$ by a full-rank matrix does not change the expression's value. The impact of the matrix signal-to-noise on the optimal linear operator increases with rising covariances of the signal components and decreases with increasing noise. The weaker the noise, the better the quality of the results obtained by the operators. This statement can also be supported by consideration of the estimate's errors as discussed in [52] and [53].

What has been discussed thus far in this chapter as well as what follows in the next several sections represents the application of statistical principles. Even so, this approach provides a clue, which leads to a better understanding of the gnostic method. The matrix operator $\underline{W}$ attaches weights, which are dependent on the matrix signal-to-noise ratio of the data: bad ratio—small weights—low confidence in the data. Covariance matrices represent estimates of the volatility of the whole population, from which the data originate. The matrix signal-to-noise ratio evaluated from covariances is therefore a **collective** characteristic of the relationship between two populations. The weights given to individual data by $\underline{W}$ do not distinguish between good and bad data. But since some good individual data invariably must be present, if the collective signal-to-noise (matrix) ratio is bad, then the confidence ascribed to **all** the data is low regardless of their individual quality. Therefore some of the information borne by good data cannot be used.

A bad collective behavior is not the "fault" of the good individual components, but all members of the sample are "punished" to the same degree by the low weights because of the "collective fault." It seems logical, that if individual weights could be attached, more information could be extracted from the data. This idea was one of the important starting points of gnostics many years ago. Indeed, gnostic weights are attached to data depending on the value of the ratio Z/Z_0, ie on the ratio, which includes both the "useful signal" Z_0 as well as the uncertainty—noise component of the observed value Z. Z/Z_0 is therefore also a version of the signal-to-noise ratio, but a strongly individual one, which reflects the particular datum's specific quality with respect to the "collective" quantity Z_0.

18.3 Special Cases of the Minimum Penalty Estimate

18.3.1 Constrained (Unbiased) Estimates

Limiting the preference of the error e_x to e_y corresponding to $p_y = 0$ (ie $r = 0$) reduces 18.19 to

$$\underline{W} = \underline{L}(\underline{X}^T \underline{B}^+ \underline{X})^{-1} \underline{X}^T \underline{B}^+. \tag{18.23}$$

When this operator is applied to data in accordance to 18.18, it is known to yield asymptotically unbiased minimum variance estimates. The most important special cases are:

Ordinary Least Squares: Let

$$\underline{B} = \varsigma^2 \underline{I} \tag{18.24}$$

with $\varsigma^2 > 0$. Let the result of the estimation be the vector $\underline{A}$ of weights of $\underline{X}$'s columns in decomposition $\underline{Y_x} = \underline{XA}$. This can be achieved by using the numerical operators of analysis $\underline{L}_{A,l}$ for $l = 1, \ldots, M$. The required estimate will then result from 18.19 in the form

$$\tilde{\underline{A}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}, \tag{18.25}$$

ie it will be the *ordinary least squares estimate* (OLS).

Neumann-David Estimate of Linear Forms: Let $\underline{L}$ be an arbitrary real vector of dimension M. Let $\tilde{\underline{A}}$ be the OLS estimate 18.25 and let

$$\tilde{z} = \underline{L\tilde{A}}. \tag{18.26}$$

Then—according to the Neumann-David theorem ([20])—the estimate $\tilde{z}$ will be BLUE (the best linear unbiased estimate) of the linear form $\underline{LA}$.

Gauss-Markov estimate: The substitution of $L_{A,k}$ into 18.26 for $k = 1, \ldots, L$ enables the set of L operators $\underline{W}$ to be obtained in the form of the $L \times N$ matrix

$$\underline{W}_A = (\underline{X}^T \underline{B}^+ \underline{X})^{-1} \underline{X}^T \underline{B}^+, \tag{18.27}$$

which can be called the *analyzer.* Applying the analyzer to data in accordance with 18.18 provides the generalized Gauss-Markov estimate ([69]).

Generalized Discrete Zadeh-Ragazzini Estimate: Let the covariance matrix $\underline{B}$ be regular and the data be of the form

$$y_k = \sum_{m=1}^{M} a_m \xi_{k,m} + s_k + n_k \quad (k = 1, .., N), \tag{18.28}$$

where $\xi_{k,m}$ are values of some given independent functions $\xi_{t,m}$, and a_m are non-random but unknown coefficients. Numbers s_k are values of a useful, informative random function $s(t)$ and n_k is noise. The required result of the estimation is the numerical representation of the function $\mathcal{L}(\sum_{m=1}^{M} a_m \xi_{t,m} + s(t))$ at a given point $t = t_*$. The *a priori* information is given by

- the covariance matrix $\underline{B}$ of the noise,
- the $N \times M$ full-rank basis matrix $\underline{X}$ composed of vectors $\xi_{*,m}$,
- the operator $\mathcal{L}$.

This task implicitly assumes that the vector composed of s_ks belongs to the subspace, which has the basis $\underline{X}$. The estimator derived in [47] therefore reduces to 18.19, into which $r = 0$ is substituted.

18.3.2 Unconstrained Estimates

It is well-known, that the BLUEs (Best Linear Unbiased Estimates) are efficient, ie they reach the lower bound of Cramer-Rao inequality for variance. However, this is true only among **unbiased** estimates. Biased estimates may be better than BLUEs in the sense of having smaller variances. Estimators, which reach an unconstrained minimum of variance were reviewed in [107]. Such an estimator was first presented and published in 1954 by V.M. Semyonov in [98]. A generalized version of the unconstrained estimates can be obtained from 18.19 by the substitution $r = 1$. This value of the penalty means, that the error component e_x is completely ignored: an absolute priority is given to the error e_y.

An interesting modification of the unconstrained estimate represents the *ridge regression* ([4]), which provides an estimate of the weighting vector A in the following way:

$$\underline{\tilde{A}} = (K\underline{I} + \underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{\hat{Y}}. \tag{18.29}$$

This method did not result from a theoretical notion, but from the experience, that introducing the term $K\underline{I}$ with a positive K decreases the estimate's variance. However, by using formula 18.19 instead a theoretical justification for this type of estimator can be obtained: it will be optimal (in the sense of minimizing the penalty) in the case of uncorrelated noise ($\underline{B} = \underline{I}$) and for K interpreted as the relative penalty r. It is useful to note, that the variance of the minimum penalty estimate depends on the matrix signal-to-noise ratio $\underline{X}^T\underline{B}^+\underline{X}$.

By choosing a penalty ratio of $r \neq 0$ and $r < \infty$ one can reach a compromise between the conflicting requirements of minimum variance and zero bias. The best choice depends on the characteristics of the data being considered and it can be made by the evaluation of errors e_x and e_y, when a particular data set is being used.

18.4 Applications of Numerical Operators

18.4.1 The Method of Static Programing

The formula for optimal numerical operators 18.19 depends on the data values in a complex way, because the data determine covariances, from which the operator is calculated. However, the situation changes significantly, when it is possible to assume stationarity of the object or process under consideration. In such a case, all variances and covariances converge to constants with increasing data volume. The operators then become fixed matrices/vectors, which can be estimated "once forever" (valid until the occurrence of large scale changes in the data structures). If the "new" data entering the formula 18.18 can be assumed to originate from the same population as the "old" ones (from which the operator was determined), operation 18.18 is linear with respect to the "new" data. It reduces to the ordinary scalar product of vectors or to the matrix product of matrices. The operator is constant and repeatedly multiplies the new operands—data vectors.

In this manner, the repeated application of numerical operators can be separated from the "design" of operators. A typical flow of data treatment then comprises two stages:

Calculation of optimal numerical operators:

- estimation of the mean values needed for the centralization of vectors,
- estimation of the covariance matrix $\underline{B}$ (18.9) of the noise,
- estimation of the covariance matrix $\widehat{\underline{X}\underline{X}^T}$ (18.9) of the useful data components and calculation of its "square-root" matrix $\underline{X}$,

- calculation of the vector $\underline{L}$ according to Definition 17,
- calculation of all elements of the formula 18.19 with the relative penalty r as a variable,
- optimization of the operator by finding the most suitable value of the relative penalty r with respect to the particular data and both of the errors e_x and e_y,
- calculation and storing the operator $\underline{W}$.

Application of operators: Repeated operation of 18.18 on new data vectors/matrices.

This approach (called *static programing* [48]) was developed some decades ago and was motivated by the necessity to minimize the time and memory requirements of computers, when these were rising much faster than the computers' capacity. In spite of the enormous progress in computing technology, these constraints still persist as an economic problem. The execution of an operation on a computer or the storage of a byte is incomparably cheaper now than in the early days, but it still costs money. On the other hand, the demand for an increasing volume of computations to be performed has risen so fast, that the product of their cost and the required number of operations surely has not fallen. Therefore, ideas, which lead to economical programs do not lose their value. "Old ideas" do not always mean "bad ideas."

The scheme of static programing is easily extended with the use of robust gnostic filters (see Chapter 17):

First, static phase: calculation of optimal numerical operators.

Repeated phase:

1. Robust cross-section or time-series filtering.
2. Application of operators to filtered data.

Robust filtering suppresses gross data errors thus making the data suitable for optimal linear operations.

18.4.2 Main Classes of Applications

There are several classes of applications, which differ by the structure of the data samples to be treated. Samples can consist of numerical representations of smooth functions of a non-periodical or periodical character as well as non-smooth functions or functions defined only by covariance matrices. The choice of the basis of the informative data subspace (of the matrix $\underline{X}$) must reflect the nature of the data. The following important cases are set out in more detail.

A Polynomial Basis

Smooth and differentiable continuous functions can be approximated by polynomials of a sufficient order. A typical example is the Taylor's expansion. If the informative component of the signal has no special features (such as periodicity) and if the signal can be assumed to be sufficiently smooth, then the polynomial approximation is suitable. The tasks considered above relate to discrete representations of continuous functions. In the case of a polynomial approximation of the signal, the components of the data vectors and the matrices of the discrete model are also polynomials. To get columns of the basis $\underline{X}$ to be independent (theoretically), it is sufficient to use elements $X_{n,m} = t_n^m$ $(m = 0, 1, \ldots, M)$, where t is the independent space- or time-variable of the model.

Applications using a polynomial basis suffer from the following problems:

1. The degree of the polynomial must be optimized. A low degree increases the approximation error of the smooth data components. A high degree causes a high volatility of the approximation errors, which results in the strong amplification of the noise components.
2. Functions t^m and t^k are theoretically independent for all $k \neq m$. But this does not mean, that they are independent, when they are represented numerically in a real computer. Depending on the length of the code (bits used to represent numbers), each computer sets a limit on the polynomial's degree. Those exceeding this limit are not represented properly by the computer; this can cause errors in the calculations of the numerical operators or even a failure of the calculations.
3. A polynomial basis is not suitable for signals with large slopes and/or with sudden changes of slope. These would require a polynomial of a degree higher than the specified limit.
4. For the same reason, the polynomial basis is not suitable for signals, that have several maxima and minima. Periodic or other multimodal basis functions are preferable in such cases.

In spite of these limitations, the application of polynomial bases for numerical operators in combination with a robust gnostic filter can be very efficient.

Periodical Basis Functions

Common practice has been to treat periodic functions with Fourier transforms, particularly after the FFT (Fast Fourier Transforms) became available. Periodic signals exist in many application fields and their processing solves important tasks. An advantage of the Fourier transform is, that for a sufficiently long data series, it provides estimates of the signal's harmonic components without a priori knowledge of the basic frequency. There are also drawbacks to using this technique:

1. It assumes stationarity of the series and a constant sampling interval.
2. It cannot be applied to a series of periodic components superimposed over an unstable drift.
3. Its precision depends on the length of the series, which is to be approximated or examined.
4. It is sensitive to outliers, its results are unrobust.

A drift of the signal's "mean" value violates the basic assumption of the Fourier series, that all elements of the series are mutually orthogonal. A drifting non-periodic component is not orthogonal to the harmonic functions $\cos(mt/T)$ and $\sin(mt/T)$ (where T is the basic period and $m = 1, \ldots, J$ with a sufficiently large J). Moreover, the longer the series, the better the Fourier approximation (theoretically), but the lower the (practical) hope, that the temporary mean remains constant.

There are methods available to (at least partially) escape these problems in the form of programs cited in [103]:

De-meaning: Subtracting the series' mean from the series before application of the Fourier transform.

Detrending: Subtracting a linear trend (estimated by means of the least squares method) from the series before applying the Fourier transform.

Tapering: An operation applied to a de-meaned or detrended series to reduce the *leakage phenomenon* in spectral estimates. Leakage occurs, when there is a large amplitude peak at a particular frequency f. Then the spectral estimates at frequencies near f can be higher than expected. Tapering consists of multiplication of the series' values by numbers w taken from the closed interval $(0, 1)$. The weights w are close to zero at the ends and close to one in the central part of the series.

Smoothing: Smoothing can be applied either to a periodogram (square of the discrete Fourier transform) or to autocovariances obtained from the periodogram by the inverse Fourier transform. The usual tools are moving (running) windows and bandwidth filters.

These methods suffer from unrobustness. All these transformations have a linear character. They do not reflect the individual uncertainty of a datum and as such, they cannot cope with gross data errors. Moreover, the parameters of tapering sequences as well as of smoothers are chosen subjectively. This means, that these methods can produce results, which are unrobust and far from information optimality.

Some analysts use an alternative method of de-meaning or detrending: they form new series consisting of differences between neighboring values of the original series. The first difference rids the series of a non-zero mean, the second difference removes the linear drift, the third difference takes care of a quadratic drift, and so on. The problem is, that differentiation strongly amplifies noise and errors in the data.

Consider now the extended scheme of static programing with a basis matrix X composed of three kinds of columns:

- cosines $(\cos(mt_n/T),\ m = 1, \ldots, J,\ n = 1, \ldots, N)$,
- sines $(\sin(mt_n/T),\ m = 1, \ldots, J,\ n = 1, \ldots, N)$,
- powers $(t_n^d,\ d = 0, \ldots, D,\ n = 1, \ldots, N)$,

where T is the period of the basic harmonic and t_n are values of an independent space- or time-variable, the values of which do not have to be uniformly distributed (the sampling interval is not necessarily constant). There is a freedom in selecting the harmonic functions, not all m from 1 to J need to be included; some of them can be left out as uninteresting. It is assumed, that the basic period T is known, either by previous experience or by enforcing it. The latter case frequently occurs in experiments, when an object is subject to periodical disturbances and its reaction is to be analyzed. An example can show the real efficiency of this approach based on experiments performed on an extensive power-distribution system [54]–[55]:

At the end of the seventies, the electrical power generating and distribution systems of many Central- and East-European countries including Soviet Russia were interconnected to form a grid. This large system helped to provide an efficient redistribution of power between countries, when the national systems experienced demand peaks for power at different points in time. Because the quality of electrical power as measured by its frequency decreases, when the network is overloaded, this international cooperation enabled surplus power generating capacity to be used wherever it existed in other countries. This activity had both commercial and technical aspects, because it led to frequency stabilization over the complete system.

The problem to be solved was to determine the response of the network's frequency to a local increase of production or consumption of electricity. The difficulty of this task was compounded by the enormous size of the system: for instance, the emergency switch-off of a power generating station in East Germany, that represented a negative disturbance of 1000 MW was not picked up by the measuring and registering apparatus of the network's Central Dispatching in Prague. The effect was buried in the permanent volatility of the network's generated power and demand. It became obvious, that a "step-response" experiment (which had been sufficient for such measurement in the past) was inapplicable under these new conditions. Therefore a decision was made to test the idea of numerical operators prepared to analyze a periodic signal superimposed over a polynomial. A periodic power disturbance was introduced using hydroelectric stations, which could increase or decrease their power from the technological minimum to full capacity and back again within a fraction of a minute. The test consisted of a set of power cycles made up of a two minute sequence at full power followed by a reduction to minimum power for two minutes. One experiment was comprised of three or four such cycles immediately following each other. The amplitude of these power disturbances represented only about 0.1% of the total power of the complete grid. The "signal" matrix ($\underline{X}$) was composed of 5 harmonics (sines and cosines) of the base period (4 minutes) and a quadratic polynomial with unknown coefficients.

In spite of the relatively weak disturbance, which was deeply imbedded in the volatility of the power output, the periodic response of the network's frequency as well as the polynomial drift were reliably identified and estimated with a surprisingly high accuracy. Moreover, it was possible to estimate the transfer function of the "network's frequency/power" and to estimate the distribution of the disturbance between the different international transmission lines.

At least three favorable factors led to this success:

1. the simplicity and statistical optimality of the approach,
2. the known value of the base period enforced by known power impulses,
3. a practically zero probability of observing a similar periodic component in random "noise" within the "natural" volatility of the power and frequency.

Another successful application of the harmonic & polynomial analyzer was in the acoustic analysis of the state of different means of transport. Diagnostics such as these can serve to discover and identify faults through the analysis of changes in acoustic spectra.

This methodology can also find interesting applications in economics. Imagine eg a large super- or hyper-market, which has a need to identify the reaction of sales to price changes. Sales fluctuate over an ever changing mean value. Three or four very weak price "impulses" can be introduced similar to the power fluctuations seen above. An analysis of the response of sales using the harmonic & polynomial analyzer could reliably reveal the sensitivity of the market (price elasticity of demand) to price changes.

Operators for Automatic Monitors of Processes

The supervision of processes in many fields of application (economics, medicine, technology, production quality assessment etc.) represents one of the most typical uses of computers for the treatment of time series. Such tasks include not only filtering (providing a type of moving average), but also a broad selection of other functions based on the recognition of several possible states of the object to be monitored. The automatic device must reliably distinguish “abnormal” situations from the desirable “normal” ones. Moreover, it is frequently necessary to determine, which one of a foreseen set of “abnormal” situations took place, and to initiate the proper response from the supervising system. Numerical operators used in a scheme of static programming are a suitable tool for the purpose, but only when the basis matrix $\underline{X}$ is adequately defined. The task of filtering, of checking limits, of estimating rates of change (trends) or acceleration can ordinarily be solved by operators based on polynomial or harmonic versions of $\underline{X}$. However, there are transient processes, which cannot be decomposed into simple smooth and differentiable functions and their numeric representations. Examples of such processes are considered below.

Basis Matrix Determined Experimentally

Using the basis matrices described above can be justified with a priori knowledge of the nature of the data samples or the data series. If such knowledge is missing or if it is not reliable, a direct application of the formula (18.9) can help to define the basis matrix. A necessary condition before taking such a step is, of course, an experimental determination of the covariance matrix of the informative data components.

18.4.3 Intelligent Sensors

Starting with computers of the second generation, a typical approach to the control and supervision of complex technological processes has been through the use of large central main-frame computers, where all information gathered from a mass of simple peripheral sensors via centralized multiplexers and analog-to-digital convertors is concentrated in the central CPU. Such a centralization—although advantageous in many respects—also has serious problems:

1. Danger of technical faults: A breakdown of the central computer could cause a serious outage affecting the whole system. The installation of stand-by computers involves additional complex electronic equipment, which further increases operating and software problems.
2. Danger of software faults: The complexity of the system and of its functions requires support by extremely complicated programs. It is

not possible to verify the operation of the software under all conceivable operating conditions, so that an unexpected situation could occur, under which the software would fail.

3. Vulnerability of the system: A single center for all monitoring, emergency, and control functions could easily be damaged or suffer an attack with vital consequences for the system.

For these reasons, initial centralized control lead to a tendency toward decentralization. The recent development of integrated electronics has allowed, not only simple monitoring tasks, but also many more sophisticated, ‘intelligent’ functions of the main-frame computer to be decentralized. It is currently possible to create *intelligent sensors*, which integrate a sensor with a microprocessor, taking over many important activities of the central ‘brain’ of the system as well as allowing new functions to be locally controlled:

1. reliable self-diagnostics of each sensor,
2. robust filtering and/or prediction of the process level,
3. reliable classification of the states of the observed object (‘everything is O.K.,’ ‘significant trend,’ ‘fast transient,’ ‘emergency condition of i-th class’ etc.),
4. efficient automatic control of local subsystems,
5. activation of local warning or emergency systems,
6. ‘post-mortem’ records (archiving of the process as it develops during emergency conditions),
7. on-line estimation of probabilities of some events or states.

The decentralization of automated decisions allows a significant simplification of the central supervising computer and its programs. The application of gnostic algorithms ensures informational optimality and enhanced reliability for the management of intelligent sensors, especially when a robust gnostic filter is combined with the application of properly selected numerical linear operators. The basis matrix X provided for an intelligent sensor/monitor can include columns, which represent the following suitable functions ([51]):

- polynomials of a low order to monitor a smooth drifting level and to derive rates of change and acceleration,
- single outlier $(0, \ldots, 0, 1, 0, \ldots, 0)$,
- step function $(0, \ldots, 0, 1, \ldots, 1)$,
- ramp function $(0, \ldots, 0, 1, 2, 3, \ldots)$.

Experimental runs of such a monitoring system can be used to obtain gnostic distribution functions of the “amplitudes” (weights) of these components to establish bounds for “normal” situations based on accessible probabilities and to set limits for automatic signaling of different “abnormal” conditions. These states can then be classified using information, as to which of the components exceeded their limit. A broad application field for intelligent sensors and monitors is currently being developed in connection with environmental control.

For example, the discharge outlet for industrial waste into rivers as well as factory chimneys can be fitted with automatically monitored autonomous devices independent of the plant supervisors. The principal objective of these policies would be to deliver intelligent and reliable signals on dangerous situations to competent authorities. Many environmental norms are extremely low and the quantities to be controlled are barely measurable. Informational efficiency of intelligent sensors is therefore an important factor in these kinds of applications.

Algorithms for intelligent sensors can be used not only for decentralized control, but they can also be useful as program modules of a main-frame, which surveys a large number of variables. An example would be monitoring the prices of stocks or commodities in financial markets to facilitate decision-making. The high effectiveness of gnostic intelligent sensors has been proved by several applications to real data:

- Real-time buy/sell decision making on an on-line currency exchange system.
- Robust monitoring and prediction of dust concentration in a cleanroom of a producer of high-tech chips.
- A monitor for low-level radiation.
- A fast and precise weighing machine for mobile operation.
- A robust sensor for an adaptive control system working under random disturbances.
- Acoustic monitors for diagnostics of complex machines.

18.4.4 Identification of Models

Numerical operators can be used to derive variables, which are not directly measurable, but which are needed as input or output variables for a model. A simple example may be a system of Lotka-Volterra equations, which has the form

$$\frac{1}{x_k}\frac{\mathrm{d}x_k}{\mathrm{d}t} = e_k + \sum_{m \neq k} G_{k,m} x_m + \sum_m H_{k,m} y_m \quad (m, k = 1, \ldots, M) \tag{18.30}$$

with parameters e, G and H and variables x and y, which are functions of the independent variable t. Variables x characterize the state of the system and y represent external (eg control) actions. As is well known from the literature ([85]), many real processes can be modeled by these equation systems.

To simplify estimation of the parameters, all variables x and y are to be observed at uniformly distributed points $t_1, \ldots, t_N$ and the unobservable derivatives $\frac{\mathrm{d}x_k}{\mathrm{d}t}$ are estimated with 18.18 using operators 18.20. Numerical representations of all the variables are thus prepared for estimation of the system's parameters by means of the gnostic explicit or implicit methods described in Chapter 17.

The solution of some problems using numerical operators is incomparably more efficient than when classical tools are employed. This was demonstrated in [49] by an example, which identified the *buckling* of the neutron flux in a nuclear reactor. The simplest (so called one-group) model of the space distribution of the neutron flux is

$$\nabla^2 \Phi(\underline{x}) + B^2 \Phi(\underline{x}) = 0, \tag{18.31}$$

where $\Phi(\underline{x})$ is the neutron flux at a point $\underline{x}$, ∇^2 is the Laplace's operator and B^2 is the important physical parameter (buckling), which is to be estimated by measuring the flux at a sufficient number of points. The standard approach is to make use of eigen-functions of the operator ∇^2 and to look for B^2, which ensures the best fit of the theoretical function with the data. This theoretically correct method imposes serious difficulties in practice. The problem is, that eigen-functions are well-known for a number of "perfect" geometrical objects such as spheres, cylinders, cubes and parallelepipeds. However, nuclear reactors rarely have such perfect forms. Another difficulty is connected with boundary conditions, which are never as ideal as those used in mathematical illustrations. As it is shown in [49], all of these problems can be resolved by using numerical operators, which convert a system of partial differential equations 18.31 written for measuring points into a system of linear algebraic equations. Both a "global" value (the mean value valid for the reactor as a whole) as well as local values within zones, which differ in physical parameters (fission rate, moderation, absorption), and that therefore contribute differently to the chain reaction can be determined in this way.

18.5 Examples of Numerical Operators

18.5.1 Differentiating Operators

In order to examine more closely the behavior of numerical operators, let:

1. the basis matrix $\underline{X}$ be of a polynomial type with columns t^m, $(t = 1, 2, \ldots, N,\ m = 0, \ldots, M)$,
2. the noise be uncorrelated,
3. the covariance matrix be $\underline{B}$ with constant variance V on the main diagonal and zeros elsewhere,
4. the relative penalty value be denoted r, and
5. the components of the numerical operator $\underline{W}$ be W_n $(n = 1, 2, \ldots, N)$.
6. the vector $\underline{L}$ define two linear operations:
   (a) $D1$ (value of the first derivative taken at the center of the observation interval $\langle 1, 2, \ldots, N \rangle$, ie at the target point $(N + 1)/2$) and
   (b) $D2$ (value of the second derivative at the same point).

Under these special conditions (uncorrelated noise), the variance of the result of the operation will be $V' = V * \sum_{n=1}^{N} W_n^2$. Values of this variance together with the operators $\underline{W}$ for the operation $D1$, zero relative penalty r, and for the target point (t_*) placed at the center (median point) of the observation interval are shown in Tab. 18.1. These operators will provide precise values, when they are applied to arbitrary linear functions (ie for $M = 2$.)

M	t_*	W_1	W_2	W_3	W_4	W_5	W_6	W_7	V'
2	1.5	-1	1						2.0
2	2.0	-0.5	0	0.5					0.5
2	2.5	-0.3	-0.1	0.1	0.3				0.2
2	3.0	-0.2	-0.1	0	0.1	0.2	1		0.1
2	3.5	-0.1429	-0.0857	-0.0286	0.0286	0.0857	0.1429		0.0571
2	4.0	-0.1057	-0.0714	-0.0357	0	0.0357	0.0714	0.1071	0.0357

Tab. 18.1 Numerical operators for the first derivative of a linear function in the center of the observation interval.

The effect of increasing redundancy is demonstrated by the last column of Tab. 18.1. Since a linear function is defined by two constants, an N of at least two is necessary for the numeric operator to properly evaluate the first derivative of a linear function. The redundancy is zero for $N = 2$. Absolute values of the components of the operator $\underline{W}$ as well as the estimate's variance V significantly decrease with increasing N. There also is a drawback to increasing N: the time lag of the estimate increases with N. If this is an important consideration, the problem can be remedied by increasing the data density.

The shortest length of the operator $\underline{W}$ needed to estimate the second derivative (operation $D2$) of a polynomial of the second order ($M = 3$) is three. However, due to the effect of symmetry, operators with $N \geq 4$ will also be suitable for polynomials of the third order. Examples of these operators are shown in Tab. 18.2 for the operation $D2$ at the center of the observation interval and zero penalty r.

M	t_*	W_1	W_2	W_3	W_4	W_5	W_6	W_7	V'
2	2	1	-2	1					6
3	2.5	0.5	-0.5	-0.5	0.5				1
3	3	0.2857	-0.1429	-0.2857	-0.1429	0.2857			0.2860
3	3.5	0.1786	-0.0357	-0.1429	-0.1429	-0.0357	0.1786		0.1070
3	4	0.1190	0	-0.0714	-0.0952	-0.0714	0	0.1190	0.0476

Tab. 18.2 Numerical operators for the second derivative of quadratic and cubic functions in the center of the observation interval.

This task is more difficult than the previous one, because precise results are required with polynomials of up to the third order. These higher analytical requirements result in an increased variance. For polynomials up to the fourth and fifth order, this increase in the estimate's variance is even more pronounced as can be seen in Tab. 18.3.

M	t_*	W_1	W_2	W_3	W_4	W_5	W_6	W_7	V'
4	3	-0.0833	1.3333	-2.5	1.3333	-0.0833			9.82
5	3.5	-0.1042	0.8125	-0.7083	-0.70833	0.8125	-0.1042		2.35
5	4	-0.0985	0.5076	-0.1439	-0.5303	-0.1439	0.5076	-0.0985	0.857

Tab. 18.3 Numerical operators for the second derivative of polynomials up to the fifth order in the center of the observation interval.

18.5.2 Impact of Correlation

It can be concluded from Tabs. 18.1–3, that redundancy in numerical operators can significantly reduce the variance of the estimate: a larger operator size = increased operator's length = more points than necessary are considered = increased redundancy = lower analytical precision, but decreased variance.

The results in Tabs. 18.1–3 were obtained under special conditions: zero value of the penalty r, and for uncorrelated noise. Under more general conditions, the noise correlated matrix has the form of the *Toeplitz matrix*. This matrix reflects a stationary noise with an exponential autocovariance function. If V is variance, then the crosscovariance of the m-th and n-th terms of the series is

$$\operatorname{cov}(m,n) = V * \exp\left(-\vartheta * |m-n|\right) := V * DecrB^{|m-n|}, \tag{18.32}$$

where

$$DecrB := \exp\left(-\vartheta\right) \quad (\vartheta \geq 0) \tag{18.33}$$

denotes the constant decrement of the series.

In the case of a correlated noise, a more general formula

$$V' = \underline{W}^T \underline{BW} \tag{18.34}$$

of the estimate's variance is used.

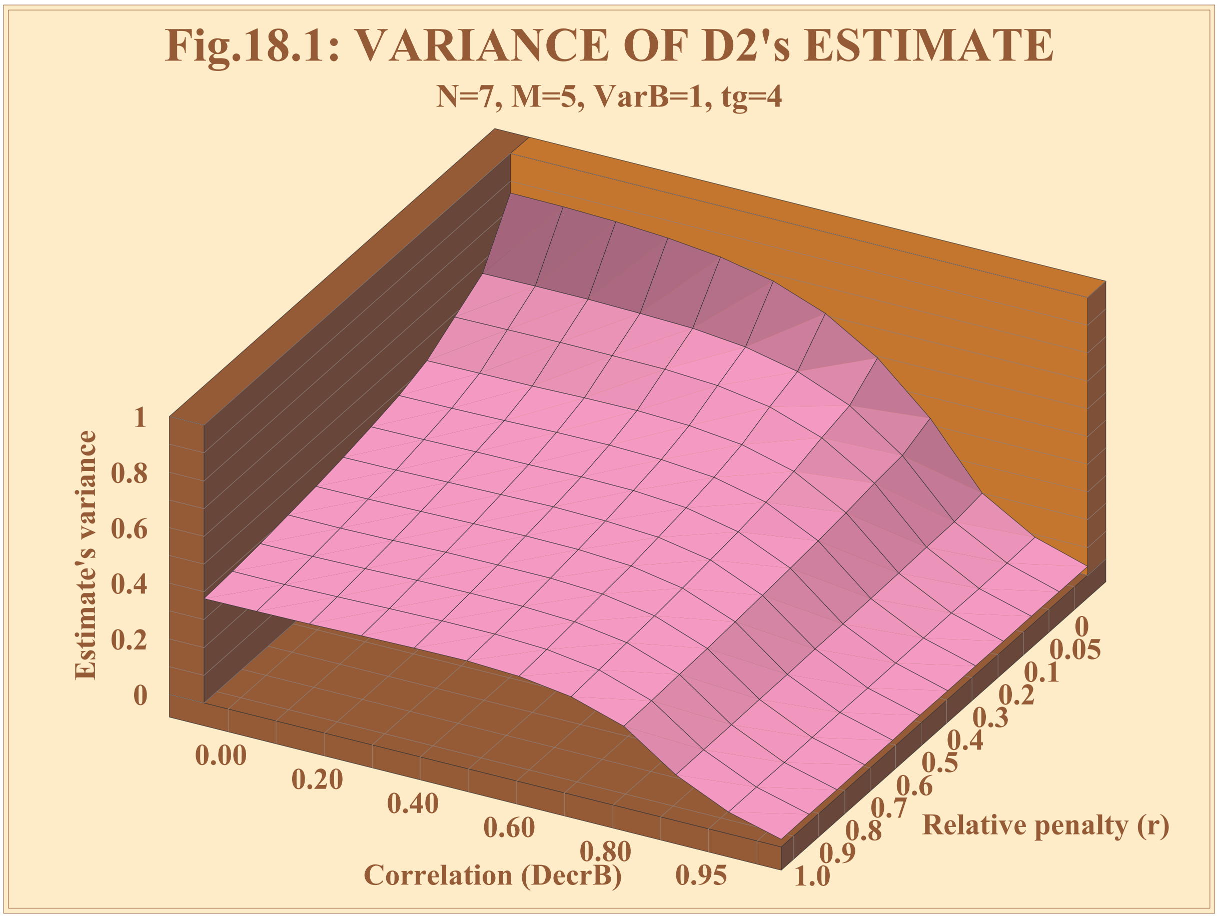
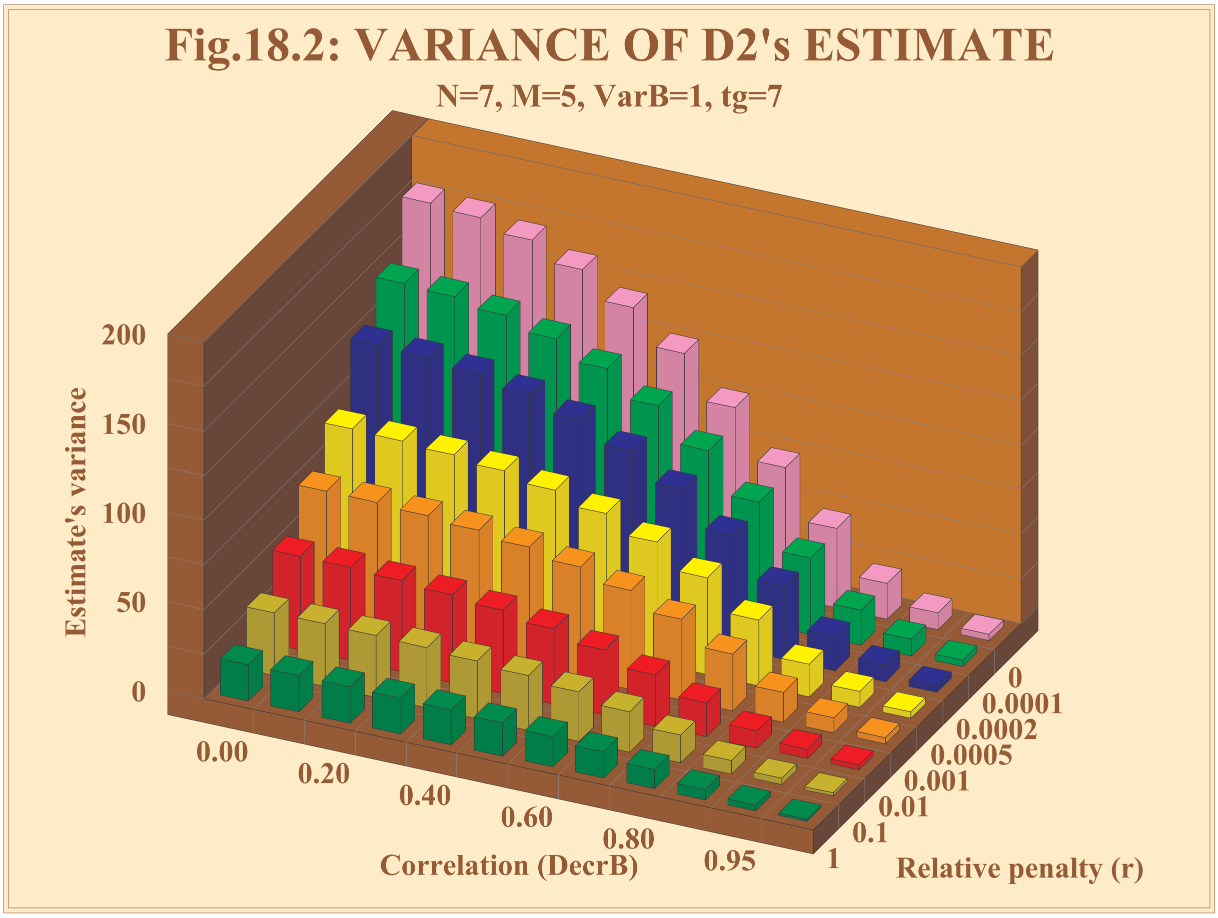


Fig.18.1: VARIANCE OF D2's ESTIMATE

N=7, M=5, VarB=1, tg=4

The dependence of V' on the correlation due to the decrement $DecrB$ and on the penalty r is shown in Fig. 18.1 and Fig. 18.2 for $N = 7$ and for polynomials up to the fifth order. The noise variance V denoted in the graphs by the symbol $VarB$ equals one in both cases. When the terms of the series, to which the operator is to be applied, are $\langle t_1, \ldots, t_N \rangle$, the (central) target point, for which the second derivative is to be estimated, is $t_* = 4$ in Fig. 18.1, however $t_* = 7$ (the end point of the interval) has been used for Fig. 18.2. A comparison of both graphs demonstrates, that a change in the target point has a strong impact on the estimate's variance. (Choosing the central point as a target leads to minimum variance).



This general tendency of decreasing an estimate's variance with increasing data correlation (as demonstrated by both graphs) can be easily explained: a correlated noise is subjected to an inner regularity, which causes the noise values to be less unexpected. Unfortunately, this tendency cannot be easily used in practice, because noise correlation is rarely under the control of a data user. But since the estimation technique must be chosen by the analyst, increasing the penalty from a zero value results in a significant decrease in the estimate's variance, especially in cases of weak data correlation. As shown in Fig. 18.2, this effect is much stronger for the unfavorable case of $t_* = 7$.

18.5.3 Effect of the Penalty

In both Fig. 18.1 and Fig. 18.2 the autocovariance V was equal to 1. It is obvious from 18.19, that the effect of a non-zero penalty r will be greater the larger the autocovariance V. This is due to the above noted effect of the “signal-to-noise matrix ratio” $\underline{X}^T \underline{B}^+ \underline{X}$, which decreases with increasing variance V. (This variance equals the sum of the elements on the main diagonal of $\underline{B}$). The dependence of D2’s variance on both V and r is shown in Fig. 18.3.

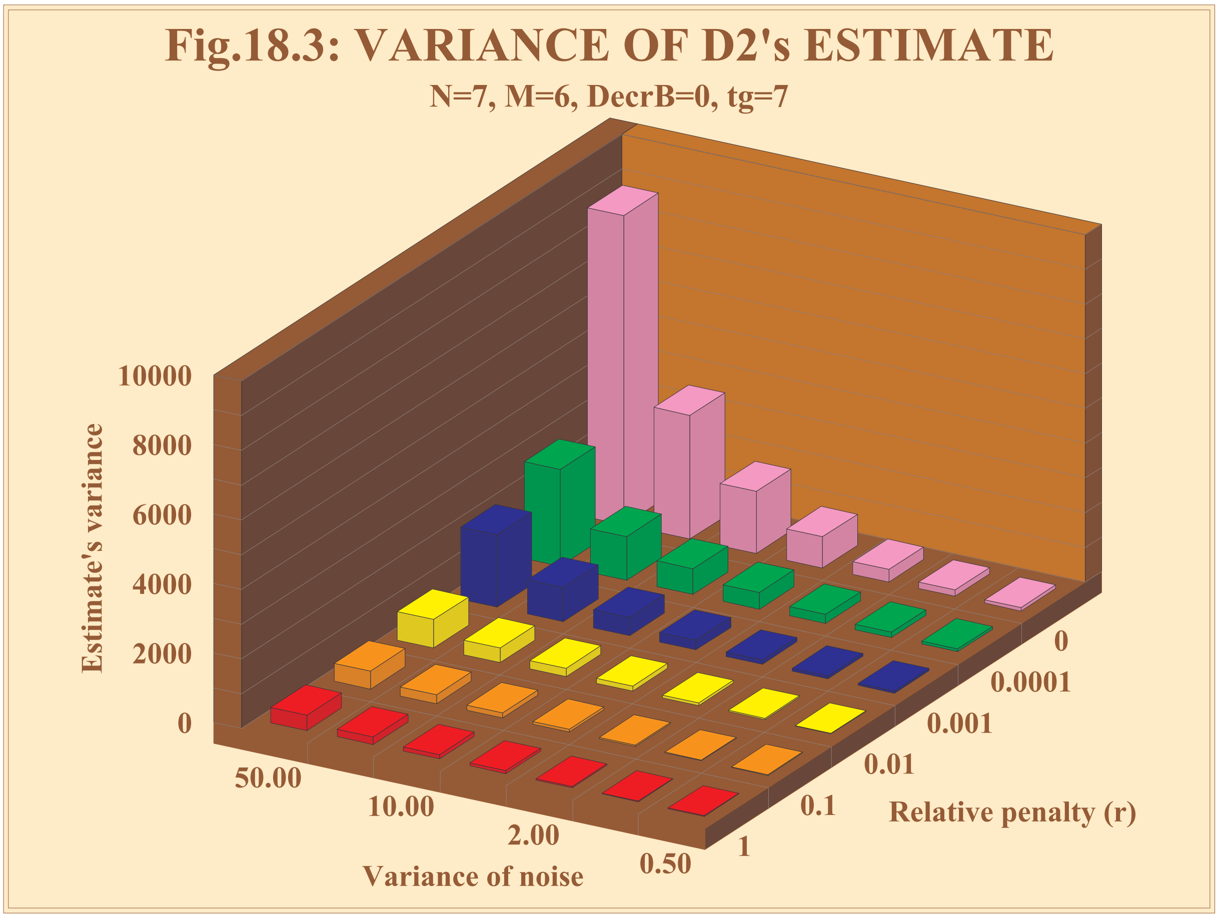


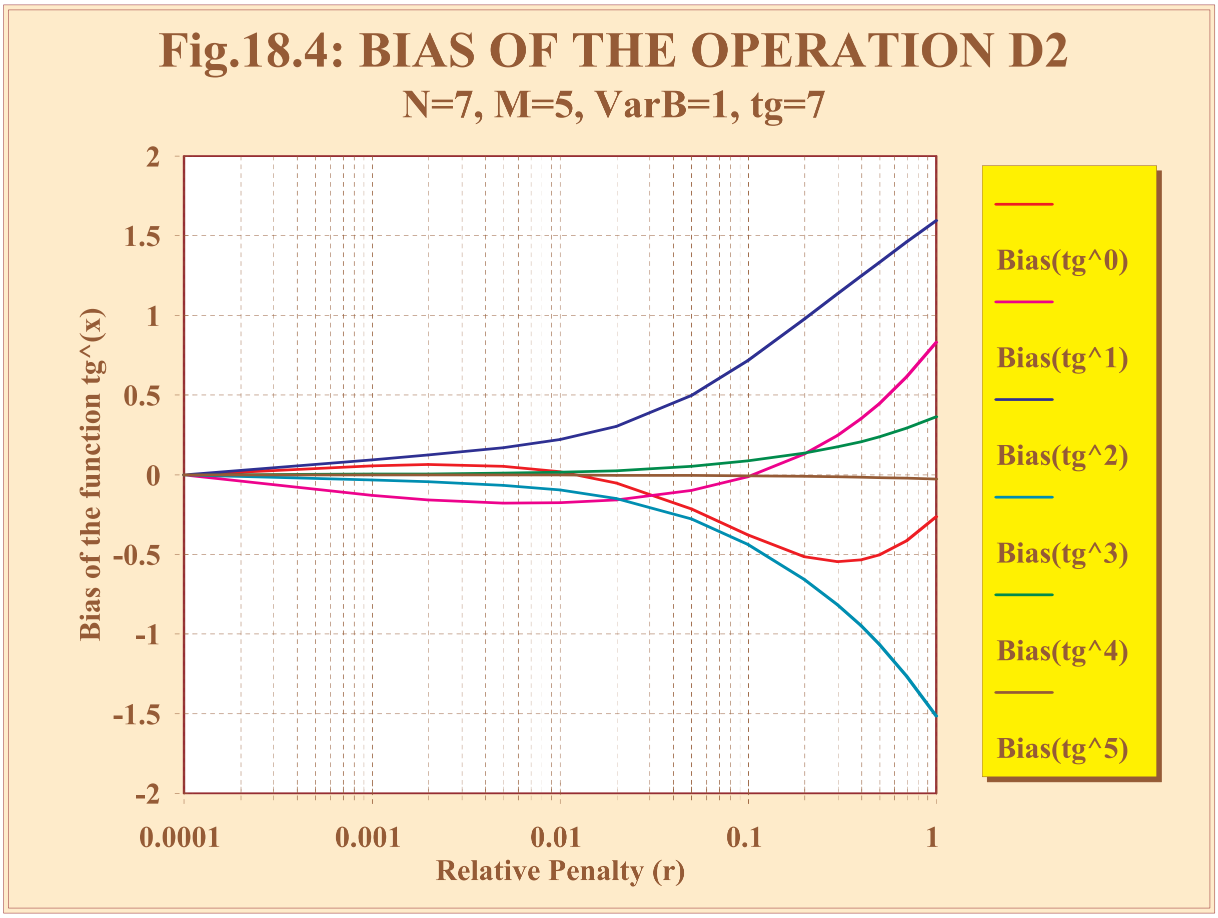
Fig.18.3: VARIANCE OF D2's ESTIMATE

N=7, M=6, DecrB=0, tg=7

The lesson given by all three graphs seems to be clear: a retreat from the strict requirement, that an estimate be unbiased, can significantly decrease its variance. However, when a non-zero penalty r is applied, how much bias can be expected? There are two issues, which should be considered here:

1. The dependence of the bias on the penalty.
2. Expectation of the bias.

Once again, using the estimate of the second derivative's value (of polynomials of up to the fifth order) with the target point at the end of the observation interval ($t_* = 7$), the length of the operator N equal to 7, and the variance of the uncorrelated noise $V = 1$, the first part of the question is answered by Fig. 18.4.



Each graph labeled $Bias(t_*^k)$ shows the dependence of the error $\underline{WY} - t_*^k$ of the estimate 18.18, when the data vector's components Y_k are exactly t^k $(k = 1, \ldots, N)$. Two further observations can be made:

1. As it was shown in Fig. 18.3, even a very small positive value for the penalty r of the order of magnitude 0.0001 causes a large decrease in the estimate's variance.
2. Fig. 18.4 shows, that such a small penalty leads to a very small bias in all functions t^k.

The answer therefore is, that it is worthwhile to use a positive, but small value for the penalty r.

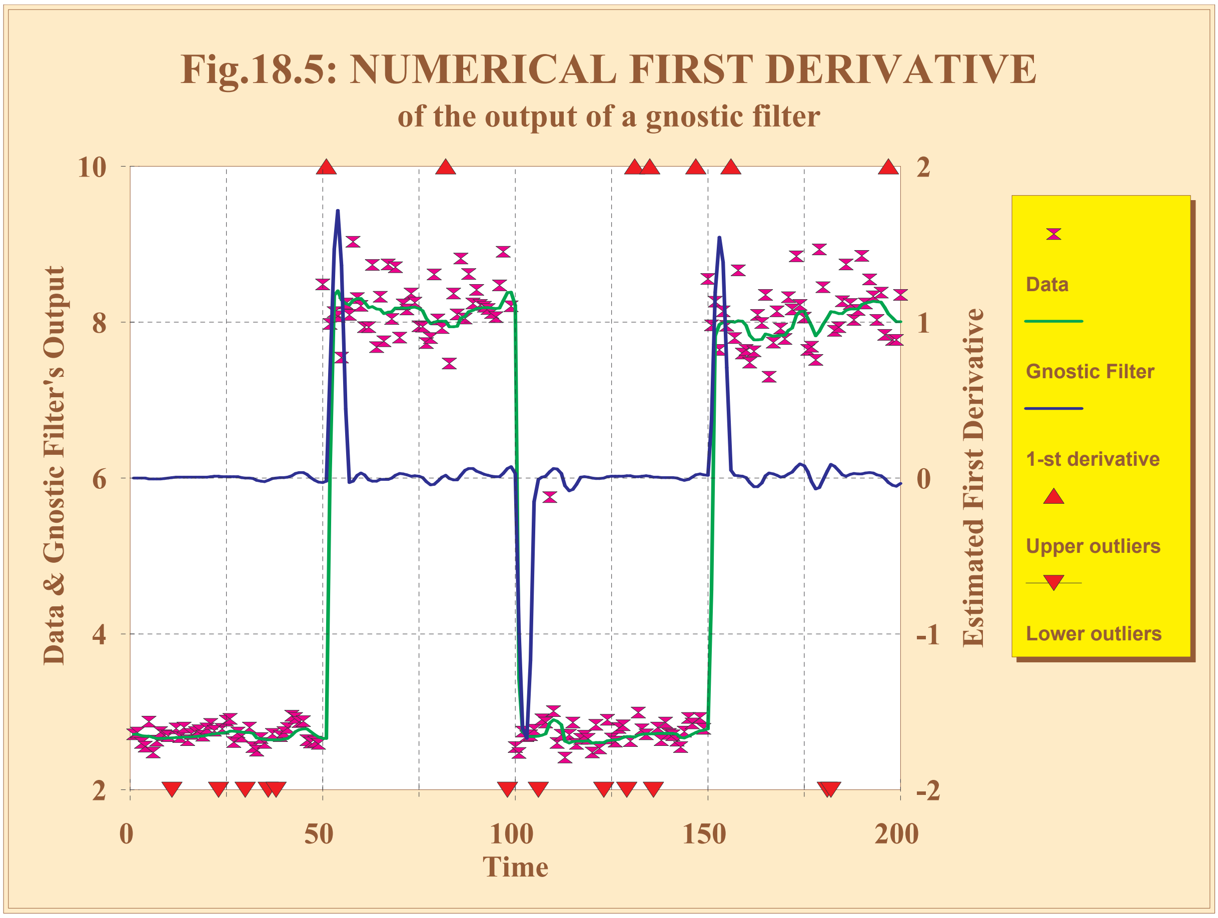
The second part of the answer is, that since the errors shown in Fig. 18.4 correspond to cases of pure polynomial data, the expectation aspect must be considered: how frequently should such data be expected? The basis matrix $\underline{X}$ was defined in 18.9 as the matrix square-root of the covariance matrix of the true (information) component of the data vectors. All the examples above assumed, that $\underline{X}$ was composed of functions t^k, and that the assumption made in 18.8 was valid: the covariance matrix of the random weights of functions t^k is the identity matrix. This assumption is not quite realistic for the covariance matrix of real data may differ. The probability of obtaining data components of the type t^k would then be lower, which would also lower the corresponding analytical error of the estimate. The answer to the second question therefore supports the decision to apply a non-zero penalty value.

18.5.4 Hybrid: Gnostic Filter Plus Numeric Operator

The main gnostic characteristics of uncertainty (irrelevance and gnostic weight) have been shown to be inherently connected with the application of a non-Euclidean geometry. It is known from Riemannian geometry, that under certain conditions, Riemannian metrics of curved spaces approach the Euclidean metric valid in tangential linear spaces, but only at points, which are sufficiently close to the tangent point. This is why irrelevance approaches Euclidean error, when errors are sufficiently small. Under the same conditions, the gnostic weight of a datum approaches the deviation of square error from 1. According to the gnostic composition axiom, irrelevances are to be composed additively and the same holds for weights. This is why, when uncertainty is weak, the gnostic characteristics of a sample's uncertainty approach linear functions of the most popular statistical characteristics, the arithmetical mean and the sample variance. The uncertainty of the statistics of a sample decreases with an increasing sample size; this fact can lead analysts to the false conclusion, that when a vast volume of data is to be treated (as eg in the case of data series, where the number of observations increases permanently), the advantages of gnostic methods cease because of the possibility of obtaining arbitrarily precise estimates by increasing sample sizes. However, the opposite is true because of the need to consider not only estimating errors, but also the timeliness of the estimate, that will result.

Indeed, it would be unreasonable to expect, that a quantity to be monitored could remain relatively constant, allowing the formation of a large enough data pool to permit very precise statistics to be computed. If such a steadfast data series did exist, there would be little need to measure and control it! The requirement to deliver warning or emergency signals within an acceptably short time intervals requires, that time limits be established for the collection of data with a corresponding reduction of sample sizes. The first economic principle "Time is money" applies here because of the possibility of losses resulting from a delay in the reception of the signal. A second economic aspect, the cost of data, should also be considered. An increase in data density can increase the size of a data sample available within a given time interval and therefore the precision of the measurement, but unfortunately, costs of more frequent measuring intervals ordinarily rise faster than the value of the obtained benefit.

The conclusion is, that in practice, a lack of data (and the need for really efficient methods to treat them) can be escaped only rarely. This justifies the idea of the joint usage of a gnostic filter with numerical operators. An example of the effects of such a combination is shown in Fig. 18.5.

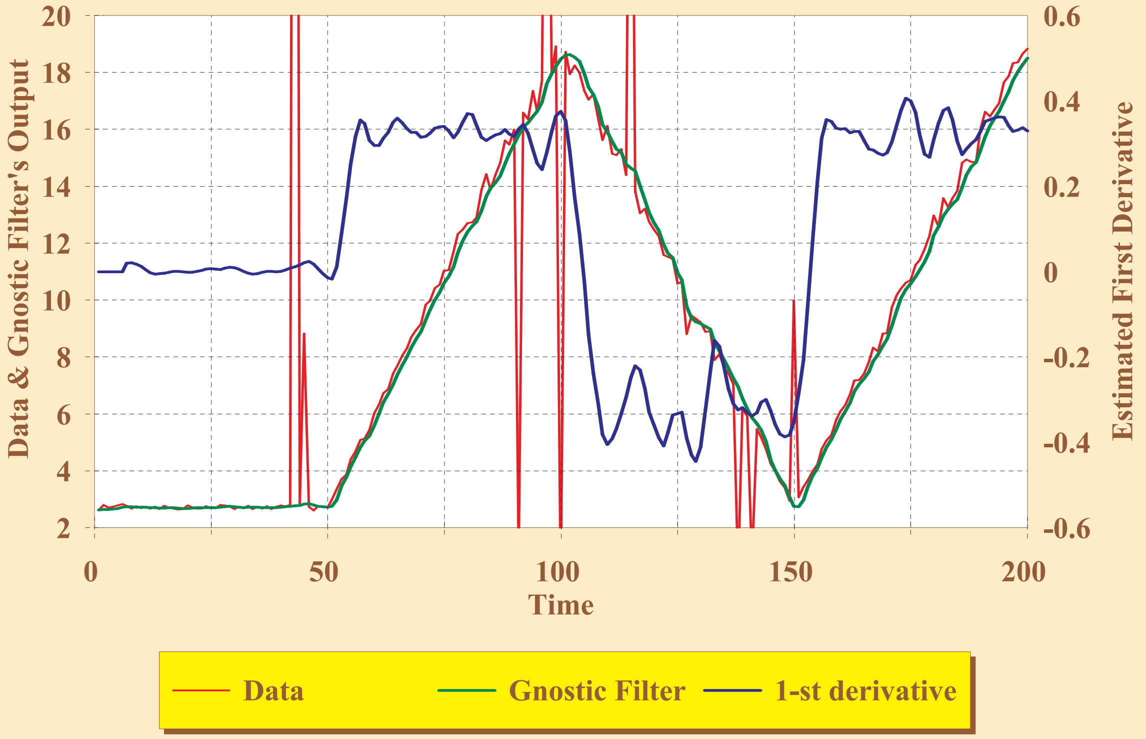


The simulated time series and the gnostic filter are similar to that in Fig. 17.8, but the output of the robust filter is treated by a numerical operator in accordance with 18.18. The operator $\underline{W}$ was obtained by using 18.19 for $N = 7$, a polynomial base of $M = 1$, uncorrelated noise with a variance of 50, and a penalty of $r = 0.0001$. The value of the first derivative (operation D1) was estimated at the central point of the moving window of the data series ($\underline{Y}$). As can be seen in the figure, the hybrid system can provide information on sudden changes in the process level with only a small delay, and with a quality sufficient for the reliable initiation of action. The poor quality of the data is compensated by the favorable dynamics of the hybrid system.

The more complex example of a data series with a rapidly changing trend is illustrated in Fig. 18.6. While the gnostic filter captures only the broad trend reversals, the first derivative provides a timely notification, that conditions are becoming more volatile.

Fig.18.6: NUMERICAL FIRST DERIVATIVE

of the output of a gnostic filter



18.6 Summary

A broad family of statistical estimates, which minimize several versions of quadratic criteria are calculated as scalar products of two vectors. The first vector can be interpreted as a numerical operator, while the second one is an operand. It was shown, that the vector-operator is determined by taking into consideration the type of (linear) operation to be performed, the mean values and covariance matrices of both the informative and uncertain data components, and by the target point (the point, at which the value of the estimated variable is to be evaluated). When the data model does not change (the means and covariances are constant), the numerical operator is a constant vector. It is therefore not necessary to recalculate it before it is applied to a new data operand. Instead, it suffices to repeatedly apply this "ready-made" numerical operator to the new data vectors. This mode of operation, called "static programing," thus reduces the estimation to the scalar product of a constant vector (or matrix) with variable data vectors. To obtain the best possible effect of such operations, it is necessary to choose the operator properly.

Statistical theory offers various methods suitable for the purpose, which can be divided into two classes based on the requirements of the user:

1. unbiased minimum-variance estimates (or BLUES—the best linear unbiased estimates),
2. unconstrained minimum-variance estimates.

Both of these classes have their pros and cons. The former estimates are unbiased, when they are applied to a specific class of data, but they do not reach the minimum variance obtainable by the latter estimates, which also can suffer from substantial mean errors. The conflicting features of these "extreme" classes lead to the creation of a more universal technique, the minimum-penalty estimate. By the choice of extreme values for the penalty, both of the extreme cases can be obtained. Moreover, properly chosen non-extreme values for the penalty can increase the advantages of using both extreme classes. Examples show, that the resulting effect can be very significant.

There are practical tasks, where using numerical operators is especially useful, eg when a variable is needed, that is not directly observable, such as the trend or acceleration of processes, where special measuring instruments are expensive or do not even exist.

All linear estimating methods suffer from unrobustness. This is why it is preferable to use numerical operators jointly with a robust gnostic filter. Such a combination is suitable even for heavy-duty data processing tasks.

Chapter 19

Consideration of Censored Data

19.1 Censored Data

Although not stated explicitly, there has been a hidden assumption with respect to all the data sets considered in the previous chapters: they were made up of the exact values given (observed, measured) or values within a relatively narrow spread of the given value. More exactly, the probability density of a single measured datum was assumed to have one of two forms:

1. that of an extremely narrow impulse (19.1) placed at the point of the observed datum A_k,
2. that of the gnostic kernel estimate (11.8), the location of which was determined by the ideal (additive) datum's value A_0 with a width defined by the scale parameter S.

The former case deals with the design of one of the empirical distribution functions (a step function, eg WEDF 15.8–15.11, while the latter is applied to get smooth distribution functions ELDF (15.24), EGDF (15.29), QLDF (15.33) and QGDF (15.37). These were shortly denoted $**$DF. All of these distribution functions were constructed by using either the estimating or the quantifying gnostic kernels. The appearance of the smooth functions depends on the form of the empirical distribution function due to the requirement of the goodness-of-fit. If there are restrictions with respect to the empirical distribution function, then these also exist and they have an impact on how the smooth distribution functions look. It is therefore important, that these (still hidden) assumptions be revealed:

Denote $\Theta(A) = dP/dA$ the probability density of $P(A, A_0, S, AL, AU)$ over the finite support $[AL, AU]$ of an additive observed/measured data A, given ideal data value A_0 and scale parameter S. Further denote $\delta(A) :=$

$R^2 \to R_+$ the Dirac's[1] (impulse) function, such that

$$\lim_{(A_-,\ A_+ \to\ A)} \int_{A_-}^{A_+} \delta(A)dA = 1. \tag{19.1}$$

Then the contribution of a datum A_k (the a priori weight of which is W_k) to the density of the empirical distribution function is $\Theta_k = W_k\delta(A_k)$. In other words, the assumptions

$$(\forall A)(AL < A < A_k)(\Theta(A) = 0) \tag{19.2}$$

and

$$(\forall A)(A_k < A < AU)(\Theta(A) = 0) \tag{19.3}$$

hold to satisfy the previously accepted condition

$$\int_{AL}^{AU} \Theta(A)dA = W_k. \tag{19.4}$$

Data, which satisfy these requirements, are *uncensored data.* The construction of a local bounded "Parzen's" type of kernel to such data is straightforward. However, it is not difficult to find data, for which either one or even both of these assumptions (19.2 and 19.3) are violated. Such data will be called *censored.* Three types of censored data are possible: *right-censored*, *left-censored* and data censored from both sides, *interval data.* Just as for uncensored data, the condition of 19.4 is satisfied for all three types of censored data.

19.1.1 Right-censored Data

In practice it is possible, that not only the number A_k itself is given as a member of the data set to be treated, but that additional information about it is available from a knowledge of the measuring process or from the nature of the particular datum. Such a situation can be defined as:

1. Condition 19.2 holds.
2. Instead of 19.3, statement

$$(\forall A)(A_k \leq A < AU)\left(\Theta(A) = \frac{W_k}{AU - A_k}\right) \tag{19.5}$$

characterizes the nature of the event quantified by A_k. Such data A_k will be called *right-censored.* Examples include:

[1] Paul A. M. Dirac, English physicist (1902–1984), and Nobel Prize winner for his work on quantum mechanics.

Survival data: Positive data T_k are the life-times of a group of objects. A "life-time" is the span from the object's "birth" T_0 until its "death." The uncensored lifetime is defined completely by the number T_k, which is fixed by the already known end of the object's life. However, in all the rest of the cases T_k is used to describe only the end of the observation period, not the end of life. The object 'lives on:' it is only known, that it "survived" time T_k, but there is no evidence of its "death." These data are to be read as "the life-time of the k-th object is **at least** T_k".

Prudence in accounting: The principle of conservatism in accounting means being cautious or prudent and making sure, that net assets and net income are not overstated. The entries in the corresponding financial statements are therefore rather right-censored than uncensored.

Measurements off the scale: A measured variable exceeded the maximum value of the scale of the measuring system.

Tax expense: In the real world people tend to estimate their tax obligations on the low side. This means, that tax paid (T_k) can be interpreted as the 'actual obligation was at least T_k.

To analyze the impact of censored data on the gnostic distribution functions, it is useful to recall, that these functions are obtained so as to ensure their best fit with the system of ME-points (15.7) related to the sample's ordered data. In the case of repeated data values this system has the form of the weighted empirical distribution function (WEDF), which has been already been examined for uncensored data. The next step is a generalization, which will also take into account censored data.

Consider the same additive form of data A_k as in 19.1–19.5 together with their a priori weights W_k.

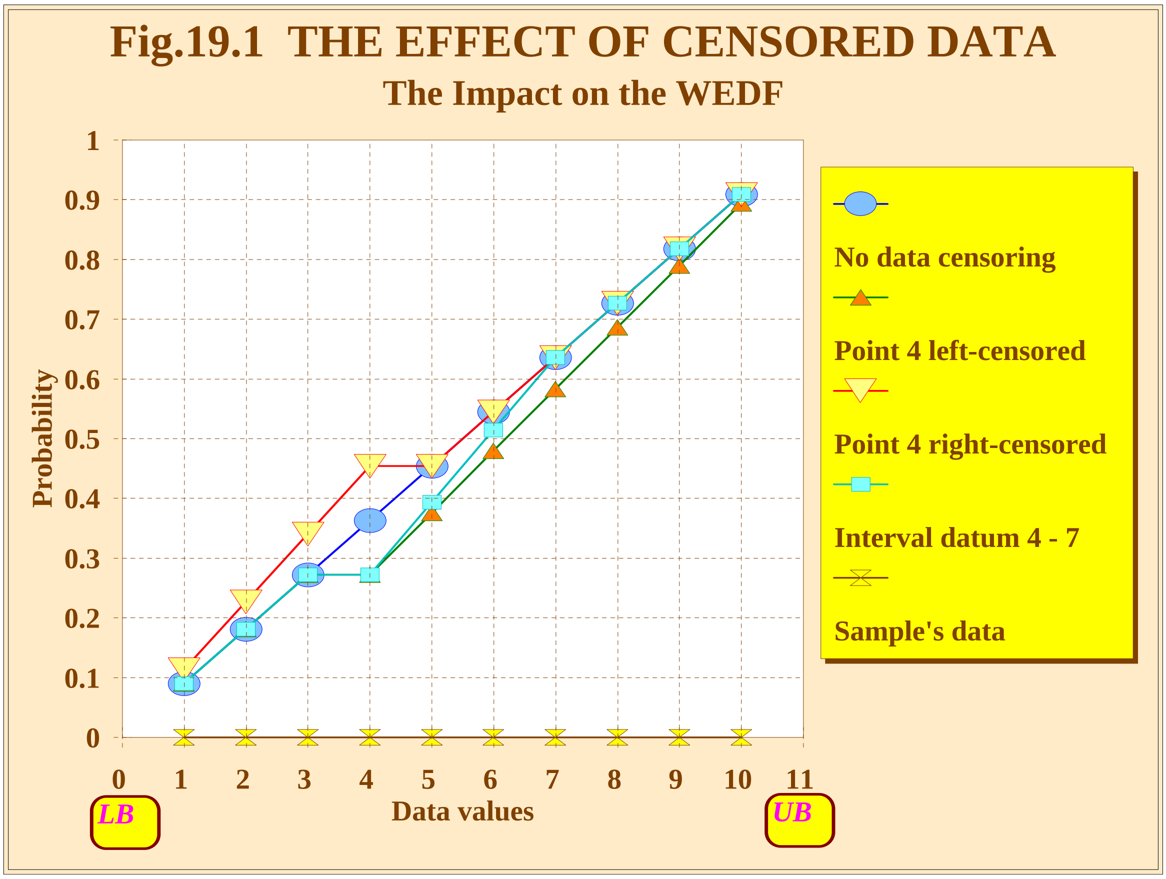
The WEDF is a cumulative function determined by the integral of the probability density. The density of all three types of censored data is constant over a limited interval and zero outside of it. The contribution of each individual data element to the WEDF is therefore linear in the former and constant in the latter case. Namely, in the case of an right-censored datum A_k (the a priori weight of which is W_k) the contribution $\Delta WEDF_k$ to the WEDF is:

$$(\forall A < A_k)(\Delta WEDF_k = 0) \tag{19.6}$$

$$(\forall A \geq A_k)\left(\Delta WEDF_k = W_k \frac{A - A_k}{AU - A_k}\right), \tag{19.7}$$

where AU is the upper bound of the data support.

Consider a simple example of the impact of an right-censored datum on the form of the WEDF shown in Fig. 19.1.



The sample is composed of ten uniformly distributed data $D_{10} := \langle 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \rangle$. If all the data were uncensored, the WEDF would be the diagonal straight line shown in the figure (the bounds of the data support are $AL = 0$ and $AU = 11$). If one data item ($A_4 = 4$) is right-censored, there is no effect for $A \leq A_3$, but further on, the WEDF does not rise between A_3 and A_4 and its values for all $A > A_4$ rise linearly, but stay below the graph that represents uncensored data.

19.1.2 Left-censored Data

Left-censored data can also be encountered under various conditions. Examples:

LIFO inventory valuation in periods of rising prices: Under the LIFO (last-in, first-out) method, the costs of the last goods purchased are charged against revenues as the cost of the goods sold, while the inventory account is based on the costs of the oldest goods acquired. When prices rise, the inventory is thus undervalued.

The cost of acquisition of an item of inventory can sometimes represent the upper bound of possible prices of an item in inventory such as goods, which are out of fashion or technologically below the current state of the art.

Expert's underestimation eg "not worth more than A_k".

Insufficient sensitivity: Imagine an automatic measuring system installed to check the concentration of dangerous gases. Some results are below the detection threshold of the sensor, but they cannot be ignored, because they represent the most desirable conditions of the monitored system. Such measurements can be therefore treated as left-censored data.

The selling price of a good is in most cases higher than the upper estimate of the good's true value. If its price is T_k than one can interpret it as 'the actual value is anywhere between zero and T_k'.

The contribution of left-censored data A_k to the WEDF can be evaluated in the following way:

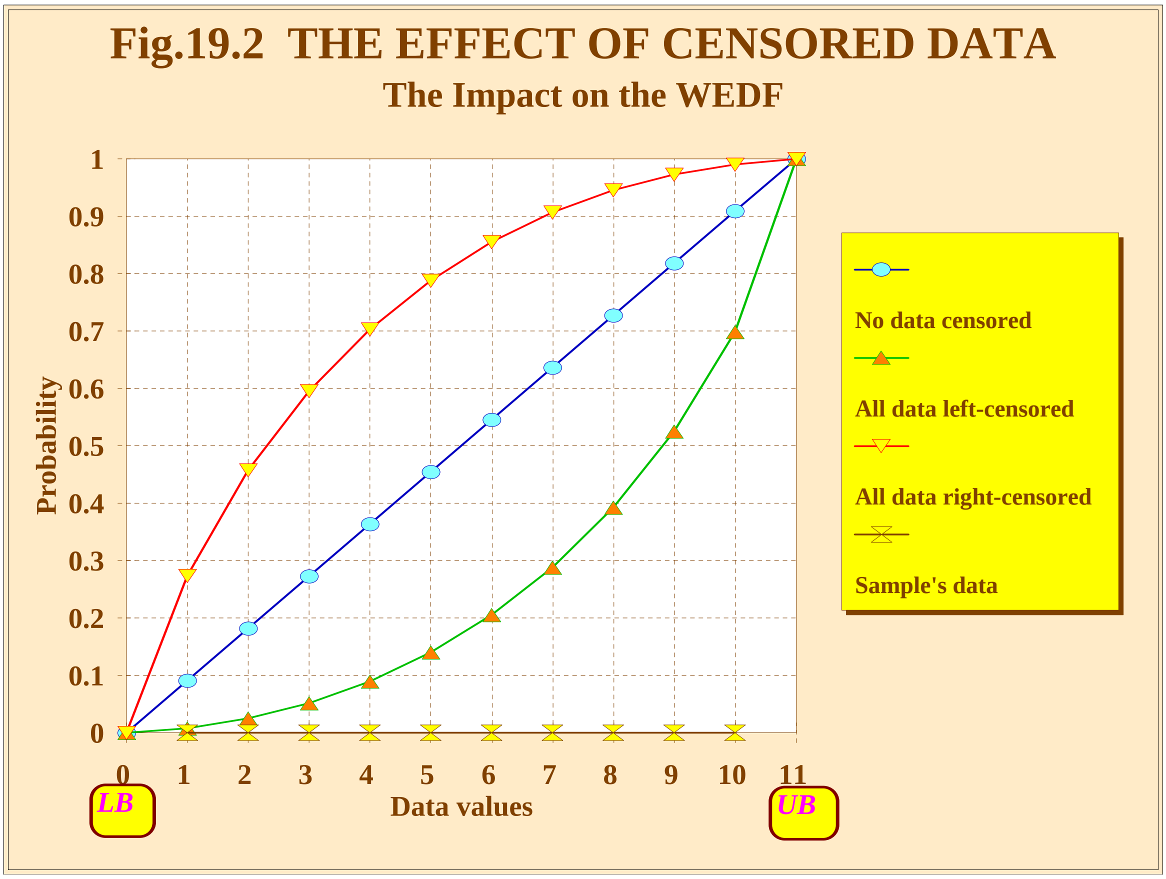
$$(\forall A \leq A_k)\left(\Delta WEDF_k = W_k \frac{A - AL}{A_k - AL}\right) \tag{19.8}$$

$$(\forall A > A_k)(\Delta WEDF_k = W_k). \tag{19.9}$$

The effect of left-censoring of the datum $A_4 = 4$ can be observed in Fig. 19.1, where the other data of D_{10} are unchanged. The distribution function rises faster due to the left-censoring effect than in the case of uncensored data over the data interval $[1, 4]$. The effect of censoring then vanishes after the point A_5.

19.1.3 Interval Data

The judgment of members of an expert board can just as easily either over or underestimate the value of an object. Each point would normally represent an uncensored datum, however, estimates could be in the form of "not less than A_k, but not more than A_m." In other words, the datum could be of interval nature, censored from both sides. Such data are not as rare as it would seem. Consider two nontrivial examples:



Product's quality: The distribution function of a parameter Q characterizing the quality of comparable products from different producers is to be calculated. There are two types of data possible:

- Measured values of Q (the uncensored data).
- Specific results of measurements of Q, values of which are not provided, but that are only counted—due to a reliable automatic quality control system—to surely fall within the interval $[Q_k, Q_m]$. Such data would then be accounted for as interval data.

Market prices are based on estimates of the true current value of goods. Assume, that both parties to a deal are well informed about the market situation. It is natural to expect, that the asking price (set by the seller) will be more than his estimate of the true value, while the bidder (the buyer) will try to establish the price under his estimated true value. The bid price can thus be viewed as the lower and the asked price as the upper bound of the interval of acceptable prices.

Three relations account for an interval datum spread from A_k to A_m:

$$(\forall A < A_k)(\Delta WEDF_k = 0) \tag{19.10}$$

$$(\forall A)(A_k \leq A \leq A_m)\left(\Delta WEDF_k = W_k\frac{A - A_k}{A_m - A_k}\right) \tag{19.11}$$

$$(\forall A > A_m)(\Delta WEDF_k = W_k). \tag{19.12}$$

Assume, that $A_k = 4$ and $A_m = 7$. Then the impact of an interval datum A_4 on the WEDF of the data sample $\langle 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \rangle$ is demonstrated in Fig. 19,1: there is a displacement to the right between A_3 and A_4, which is followed by a linear return to the line of uncensored data. Note, that it is not an estimate of the fourth datum, that is lower, but that the probability of being at least 4 has declined since the measurement could be in error on either side.

It is instructive to examine the WEDF, which results, when the whole sample consists of censored data. As shown in Figure 19.2, when all data from the uniformly distributed sample D_{10} is taken as interval data, spread from A_k to A_{k+1} $(k = 1, \ldots, 9)$ the WEDF is the same as for uncensored data.

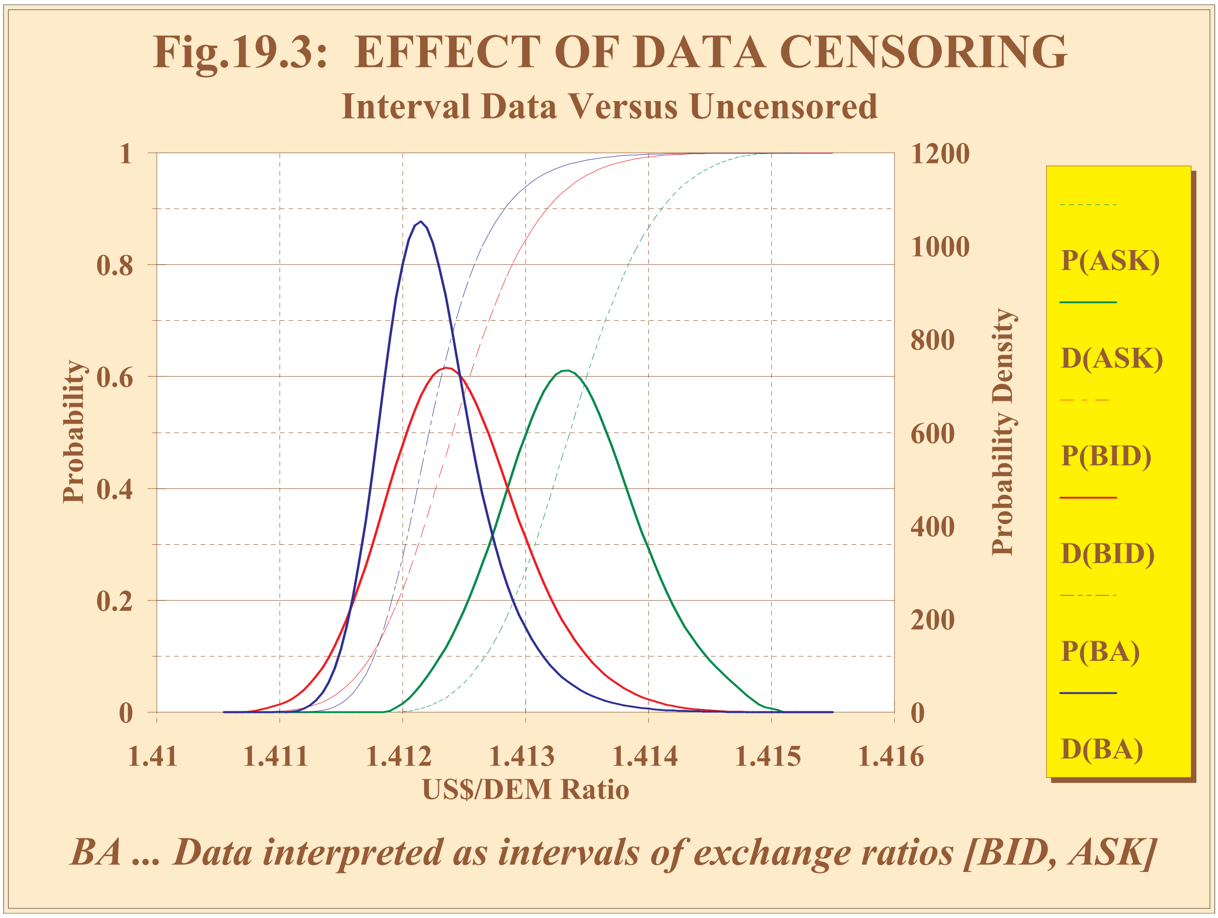
In contrast, when the data are all right-censored, the WEDF would have the form of a curve running below the straight line corresponding to uncensored data. For left-censored data, the WEDF has the opposite form, above the straight line. The effects caused by data censoring thus can be significant.

19.2 Censored Data from the Currency Market

A real example, taken from currency exchange trading, presents the concept of censored interval data using a foreign currency electronic market data series[2]. Fifty consecutive observations of the US$/DM exchange ratios for both ASK and BID prices for Oct. 10, 1992 were taken for the analysis. The probability distributions and densities of the EGDF type of both series are shown in Fig. 19.3.

It is interesting to note, that the spread between the two ratios, measured in Fig. 19.3 as the horizontal distance between both distributions, is close to 0.001. Since the ask and bid ratios in the series appear only in pairs, because they correspond to actually realized transactions, assume that the respective ratios represent the bounds of the estimated interval of

[2] Data HFDF93 were made available to the world scientific community by Olsen & Associates, Research Institute for Applied Economics, Seefeldstrasse 233, CH-8008 Zurich, Switzerland, E-mail: hfdf@olsen.ch



acceptable values of the ratio. It is then possible to analyze both uncensored series as only one series of interval data. The probability distribution and density obtained in this way are in Fig. 19.3. Two interesting conclusions result from the comparison of these distributions:

1. The volatility of the data interpreted using the "interval concept" is significantly less than that of the individual ask and bid series taken separately.
2. The location parameters (the distribution's median and density mode) of the interval data are slightly lower than those of the two single distributions.

The conclusion is, that the concept of interval data can be useful in establishing more reliable real time values for exchange ratios.

19.3 Summary

The estimating technique of gnostic distribution functions permits additional information related to the data's character or origin to be put to use. This procedure distinguishes four classes of data: uncensored, right-censored, left-censored and interval data. Censored data are distributed over a not negligible subinterval of the data support. In spite of the fact, that censored data are in a way more uncertain than uncensored data, classifying data sets in this manner can increase the value of the analysis by extracting a greater amount of information from the results of the treatment.

Part III

APPLICATIONS

Chapter 20

Letting Data Speak for Themselves

20.1 Introduction to Part III

As promised in the Preface, Part III of this book should help the reader become familiar with applications using gnostics. The authors have in mind two categories of readers; both are interested in knowing, that the methodology works. The first will want to understand why and how, and the second, who in the most part will be practitioners, will be satisfied, that it does work. Nevertheless, it will do no harm for practitioners to have a feel for what is behind the algorithms and formulae which they are going to apply.

Those of the first category who have studied chapters 1-19, already know how the theory was developed, and can omit reading this section which is primarily addressed to the second group. Our aim is to use simple terminology and to constrain ourselves to using commonly understood notions to express the theory's ideas.

Gnostics is a mathematical theory of uncertain data which should be applicable to both individual data and to small samples of such data in order to support practical needs by robustly mining information from data in an optimum manner. Several notions are to be emphasized here:

- Mathematical theory and its applicability,
- Uncertainty of data,
- Individual data and small samples,
- Information,
- Optimality,
- Robustness.

Mathematical theory: Historically, the initial objective of mathematics was to serve the needs of ordinary life by establishing values and quantities for the barter of goods and to develop rules for measurement, comparison, and exchange. During the solution of these problems, it appeared, that in properly achieving this objective, mathematics created abstract notions and found ways to use them. This process of evolution led to the complete independence of mathematics from its immediate practical application to everyday problems. Historical experience proved the fruitfulness of such development, because many of the initially completely abstract ideas and methods of mathematics found practical applications centuries after they had been developed: complex variables, the axiomatic approach to geometry resulting in non-Euclidean geometries, etc.

Mathematics can be described as a building properly erected over its foundations: definitions and axioms. “Properly” means to start building with nonconflicting axioms and to apply to them consistent reasoning. Generally speaking, these are the only requirements and the value of a mathematical theory is considered to be greater, the more mathematically interesting its outcomes are. But, when mathematics is used to model real processes, it must base its reasoning on realistic assumptions, which would lead to the applicability of the obtained results.

This is the case with gnostics. Its first axiom is based on notions and principles of measurement theory (which developed over time in parallel with markets). To exchange goods, it was necessary to develop certain rules for the measurement and manipulation of quantities. The main result of measurement theory is, that the measured or counted objects are elements of sets having a special structure. Summarizing, it is possible to state, that this main axiom evolved from something very practical, the development of the exchange of goods.

Uncertainty of data: Data are numerical images of quantities. The process of forming these images (quantification) realized under practical conditions is subjected to different disturbances which make the images imprecise, corrupted by uncertainties. In gnostics, data uncertainty is not interpreted as a random effect, but as the impact of unknown real factors. The uncertainties are thus also images of real quantities and as such they are subjected to the same regularities, as true (undisturbed) images of quantities. Each observed datum is a composition of these two elements and the role of the data treatment is in the separation of these two components. Shortly: uncertainty =

lack of knowledge.

Individual Uncertain Data: When describing the random nature of uncertain components of data, mathematical statistics relies on the main regularity of collective uncertainty, the Law of Large Numbers. Unfortunately, this approach cannot be consistently applied to individual data and to small samples. However, many practical problems deal with a need to treat very limited amounts of data. It is therefore necessary to discover regularities, to which the individual uncertain data are subjected. Gnostics shows, that such regularity exists and results from the special nature of data's structure. Everyone knows, that a thrown stone describes a precisely defined path. Analogously, gnostics derives from the first axiom the path, along which the observed datum moves under the influence of uncertainty. The movement of the tossed stone depends on its momentum and kinetic energy, which is under the control of the laws of Newtonian mechanics, and which can be described by using familiar Euclidean geometric concepts.

The movement of the uncertain data along its path, forced by changing uncertainty, is in a way similar: the data's error (called irrelevance, 9.6) and its weight (9.5) are linked to the changes in uncertainty. The data's error quantifies the size of the uncertainty and the weight is a measure of the data's truthfulness. The point is, that to treat individual uncertain data, it is necessary to have a means for the evaluation of the data's error and weight. As shown in Part I, this evaluation/measurement is realized by using non-Euclidean geometrical concepts. The choice of geometry to be applied to some particular data is not a decision to be made by the users of the method, the proper geometry is determined objectively, by the data themselves.

Information: Classical information (Shannon's) is based on a complete probabilistic description of a message; this means, that it is inherently connected with the statistical notion of uncertainty of mass events and as such it cannot be used to evaluate information carried by an individual datum. In contrast, gnostics derives a formula (10.64, which quantifies the loss of information caused by the uncertainty of the individual datum. The information in a small data sample can then be obtained by the composition of information carried by the sample's data. No a priori statistical model of data is necessary for this to occur. Everything needed is taken from the data themselves.

Optimality: The problem of optimality is linked to a requirement to do things in the best possible manner. To find a mathematical solution to this task, a criterion, which can be reasonably optimized is needed. In many instances, the role of such a quantity is given to the sum of squared errors. But, in this case, the problem is, that such a solution has some undesirable features and it is not quite clear, why squared errors and not some other function of errors should be used. Data treatment should result in information and there is no simple connection between squared errors and information. This is why gnostics interprets the optimality problem as maximization of information gained from the outcome of the data treatment. Moreover, as shown by gnostics, this optimality is really the best possible way to treat data, because any variation from the path (along which gnostic data treatment moves the data) would result in less information being delivered. This type of optimality is similar to the minimum use of fuel by a rocket or satellite, which moves by inertia along its elliptical orbit.

Robustness: The modern expression, 'robustness of a data treatment method', means a reduced sensitivity to undesirable data. There can be two different types of undesirable data. When the required information is carried by data, which are close to the central value of the sample, the undesirable disturbances are caused by outliers, ie by data, the values of which are far from the sample's center. However, the opposite may also be true: the central data may be noisy, while the information is carried by data, which are beyond the lower and upper boundaries of the undesirable data. Fortunately, all gnostic formulae have two alternatives; one provides robustness with respect to outliers, while the other is robust with respect to inliers. The first is called estimating, while the latter is quantifying. The choice depends on the goal of the task to be performed.

It can be shown, that the notions and rules of mathematical statistics are closely connected to Euclidean geometry and to Newtonian mechanics (see Chapters 7 and 13). As discussed in Part I (Chapter 7), there are important linkages between gnostics and Einsteinian relativistic mechanics. The importance of these ties lies in the proven possibility to support composition rules for uncertain data by something as universally accepted as the relativistic law of conservation of energy and momentum.

Although gnostic methodology fundamentally differs from that of statistics, it does not mean, that there is an unsurmountable gap between both approaches. The opposite is true: under the condition of gradually fading uncertainty, gnostic formulae approach the classical statistical characteristics of uncertain data. There are two analogies to this effect:

1. Imagine a plane tangential to a curved surface at some arbitrary point. A pattern of points on the surface can be projected on the plane to approximately depict its features if the pattern is sufficiently close to the point of contact. Nonlinear relations between points on the surface are then approximated by linear relations on the plane.
2. When velocities are low, the formulae of relativistic mechanics approach those of Newtonian mechanics.

Relativistic physics is a theory, which deals with processes that have very high velocities. In a similar sense, gnostics is a theory designed to treat uncertain data with strong uncertainties.

It is a tedious task to extract gold from an ore, however the value obtained from the gold recovers the effort expended in obtaining the refined product. Uncertain data are another type of ore and information has the price of gold. To obtain refined information, the data must be treated by a tedious process as well. There are many ways to treat data, but when a decision is made to use gnostic methods, it is necessary to be prepared to use complex nonlinear formulae, difficult optimization procedures, and to develop adequate efforts to thoughtfully interpret the results. Fortunately, modern computer technology easily permits the rapid and efficient manipulation of complex mathematical structures, but to serve a useful purpose, the results obtained by these mechanical procedures must still be interpreted by the efficient and diligent application of the human brain's ability to reason and draw rational conclusions from these outcomes.

20.2 Objectivity versus Subjectivity

Quantitative data are the numerical images of real quantities. There are three stages in the quantitative recognition of these real objects:

1. measurement,
2. analysis,
3. interpretation.

The real, but unknown, measured quantity is, of course, the most objective value, that could theoretically be obtained. However as the quantification is undertaken, the degree of objectivity attached to each image obtained by the observer progressively decreases due to imperfections in the process.

Methods for measurement are developed in parallel with the progress of technology so as to best satisfy the requirements of objectivity. Even in fields such as economics, where technical measurement methods are generally not applicable, there are regulations, which attempt to establish the quantification of objects on a reasonably objective level. It can be thought, that data produced by an objective (or an objectively oriented) measurement process are objective, and that the deviation of such data from their true (precise) value is due to uncertainty.

The true and uncertain components of data are—in a way—inseparable and both are objective in nature. The true value is objective, because it depicts an objective quantity. In turn, the uncertainty is objective, because it reflects objectively the existing imperfections in the measurement processes and the impact of unidentified objective factors on this process. However, this does not mean, that a separation of the true and uncertain value is automatically objective. This is the task of analysis, the conclusions of which can be both subjective and objective.

The last statement may give rise to objections: analysis comes about through mathematics and mathematics has strict rules, which harbor no subjectivity. While this is true, when the context is related to the consistency of mathematical reasoning, derivations, proofs and operations, all the "truths" of mathematics stem from its assumptions and these can be either realistic or unrealistic. The problem lies in the choice of a definition for the notion of *truthfulness*. Everything that results from consistent reasoning based on a system of nonconflicting axioms/assumptions is "truthful" for mathematicians. A similar status is also given to the idea of existence. An object always exists unless its existence has been proved to be contradictory to a truthful statement. No requirements for "realism" are posed in mathematical assumptions. Even the notion of "real" has a different sense in mathematics than in ordinary parlance: real numbers "really exist" in the same sense as that of imaginary or complex numbers. Such an abstraction and its isolation from the "real world" imparts to mathematics its exclusive creative power and its inner purity. However, there also is a substantial drawback to "being allowed" not to be realistic: the door is open to false judgements, which are derived from unrealistic assumptions.[1] In data analysis, this danger threatens the clarity and objectivity of mathematical models of uncertainty and the analytical methods to be applied to data. Some data treatment methods make use of notions, which have

[1] It is interesting to note in this connection, that mathematics is frequently classified as a member of the family of *natural* sciences.

a legal and precise meaning in mathematics, but when they are applied to real objects or processes they are not always well justified. Examples: the ideas of infinity, normality, randomness, dependence, independence, existence, homogeneity, membership, similarity, model.

The purpose of data analysis is to support decision making with respect to real objects and real processes. Results of data analysis are only useful if they are realistic and objective. Since analysts are human and can have their preferences as to methodologies as well as subjective expectations as to the results of their analysis, their points of view can be prejudiced and biased. In order to draw objective conclusions from an analysis, all subjective views must be suppressed so as to accept the dominating role of data. Data represent reality and they decide if the model resulting from the analysis is objective. The primary rule for letting data speak for themselves is thus not to violate data by imposing an a priori model.

It is during the third stage of data analysis, interpretation of the results, that the importance of objectivity becomes paramount. It necessarily relies on the subject's (the analyst's) interpretation. The analytical stage uses computers, which are objective to the same degree as the assumptions, on which the analytical methodology is based, therefore "realistic assumptions ⇒ objective results of analysis." In contrast, interpretation must be entrusted to a specialist in the discipline, who must correlate the analytical results with collateral information, as well as his/her experience, and general know-how. The degree of objectivity in interpretation depends not only on the skill and unbiasedness of the interpreter, but also on the capability of the analytical technology used to provide him/her with persuasive arguments, that support these conclusions.

20.3 Realism of Notions

Because data are real, when they "speak for themselves", they cannot reveal something unrealistic. "Responses" to "Questions" asked of data should therefore be free of 'not quite realistic' notions.

20.3.1 Infinity and zero

The primary (historical) role of mathematics was to provide quantitative and logical models of the real world and of its processes, ie of things that can be seen or imagined. However, the free (formal) development of mathematics soon led to the crossing of this boundary and notions of infinity and of its "reciprocal," zero were introduced through the concept of induction, one of the most potent mathematical methods. These ideas appeared to be necessary for the creation of logically closed mathematical structures, which are necessary in mathematics, but are not always suitable for good models of reality. A real quantity can be vast, but it is always bounded. Some amount of money can be large, but the idea of infinity is not necessary in order to visualize some larger amount; it is only sufficient to add one cent to the previous sum. Real things are bounded by their inherent nature. The extraordinary power of mathematical infinity lies in its capability to exceed an arbitrary quantity. The application of mathematics to practical problems can lead to the necessity for a symbiosis of realistic bounded functions defined over finite domains with their unrealistic mathematical images defined over the infinite unbounded horizon. This is the case in gnostics, which models distributions of real (bounded) data by means of abstract, theoretical distribution functions defined over the infinite domain, then transforms them back into the finite data support prior to their use in applications.

The notion of zero is another mathematical invention. As a member of a structure of real numbers it also displays great power: it can "destroy" any finite number by multiplication, it reproduces itself. This remarkable feature is generalized in algebra by designating it as an *ideal*[2]. If infinity can be accepted as a model of "everything," then zero models "nothing." It is difficult to relate these notions to anything real[3]. Children easily understand the idea of a set containing a finite number of objects, but the notion of an abstract empty set is much more difficult for them to comprehend; try to explain how an empty set of dinosaurs differs from an empty set of children!

This discussion becomes much more practical, when it is applied to data analysis. There is no device, which can measure infinity. Therefore, it can be concluded, that infinite data do not exist. Despite this, mathematical statistics frequently use distribution functions (eg Gaussian distributions) defined over the infinite data support. Non-zero probabilities are thus attached to data values, which are not real. This means, that an unrealistic model is used with the expectation of fitting real data.

[2] Abbreviation for 'ideal element,' a term used in number theory.

[3] An interesting contradiction: Ernst Zermelo, one of the pioneers of set theory interpreted zero this way: "There exists a (fictitious) set, the null set, 0, that contains nothing at all." Fictitious ... how do fictitious things exist?

The finite bounds of real data have more than quantitative consequences for modeling. They can change the nature of the model. The most popular (normal) distribution function has the well-known S-form and a bell shaped density. However, some processes have very different distribution functions, where the density can be sharply cut off from one or even from both sides. A non-zero probability density is in such cases only given to a bounded interval over the data support. The estimation of these bounds can provide the analyst with information of fundamental importance, which would otherwise have been lost.

Another frequent application of the notion of infinity is in connection with the Law of Large Numbers and the Central Limit Theorem. An unlimited increase in the sample size is acceptable in theory, but outside of some physical applications, can rarely be achieved in practice, particularly in economics and finance. In the analysis of mass events, the substitution of infinity for a very large finite value leads to working mathematical models, that result in only negligible errors. But it is a quite different situation in the case of costly data, or when it is impossible to increase the number of data. The mathematical trick of letting the number of data rise without limit cannot be reasonably justified in these situations and a non-statistical model of uncertainty has to be used.

Zero data can actually be generated by a real measuring device; even so, they deserve special attention. The range of a measurement instrument is bounded and the measuring range must be properly selected. The measured value should not exceed the upper bound of the range. On the other hand, it must not be so small as to be close to the lower bound. The problem is, that the relative precision of the instrument falls with decreasing measured value.

Another problem with zero data is its potential double meaning. Zero data can be interpreted not only as “value below the measurement threshold”, but also as “measuring instrument or communication line failure.” Zero data have a doubtful informative value even in quantification processes not realized by technical means. An example can be a statistical survey based on random sampling. Observing a zero frequency of occurrence for an event can be interpreted not only as “non-existence of the event”, but also as “too rare to be detected.”

Zero data value can result from additive operations on real (additive) data. A “natural” support for additive data is therefore a bounded interval of real (positive, negative and zero) numbers. Strictly positive (multiplicative) data exclude both infinite and zero data values and their data support is a bounded interval of positive numbers. Infinite bounds of data supports for distribution functions can also be used, when they correspond to data, but their realistic interpretation should be "a very distant value" rather than an "infinite" one.

When an assumption of bounds for the data support is accepted, a question arises as to how these are established. The bounds can sometimes be determined by the nature of the event under consideration. In other cases, they must be estimated, which leads to a corresponding requirement for the proper analytical tools.

20.3.2 Normality

The notion of *normal* in statistics is reserved for the well-known Gaussian type of probability distribution. The assumption of normality is a source of many popular statistics and tests and it is supported by the Central Limit Theorem. The problem is, that this assumption is frequently applied to non-limit cases, which have neither an infinite number of data nor infinite data support. Many users of statistics rely on the statistical idea, that the sample means can be reasonably approximated by a normal distribution for samples of size larger then 30. It is easy to show, that this statement is more of a superstition than a provable fact, because it does not take account of possible outliers. The inclusion of a data item with an extremal value in a sample can distort the distribution function of much larger samples. The influence of large, but finite outliers can be completely eliminated in the case of an infinite sample size. A potential objection to the possibility of large outliers could be, that data are always bounded in practice. This is true, but then what reason can there be for using the normal distribution defined only over an infinite data support?

The statistical notion of normality is far from the common understanding of the meaning of the word: conforming with the accepted standard or norm; natural; usual; regular; average; ordinary or what is expected. The Gaussian distribution is favored due to the sufficiency of its two easily estimated parameters, the mean and the standard deviation. However, there is no reason to see it as a standard or norm. It can be natural, regular and usual in one application field and unacceptable in another. Consider two examples of the frequent misuse of the normal distribution:

1. In financial statement analysis: the number of comparable enterprises, which can be used for statistical comparison is always low, frequently not exceeding ten and rarely exceeding 30 or 40. Strong outliers can be

expected even though the data are surely bounded. To be realistic, the estimation of risks and chances associated with data values should be based on actual distribution functions and not on an a priori assumed model.

2. In production quality assessment: there is no reason to assume, that an assessment of good quality is subject to a normal distribution. The actual distribution of qualitative parameters is dependent on the production and control technology. Infinite deviations from the mean are impossible. Decisions as to good/bad products should be therefore based on the actual distribution functions of the good products.

It must be concluded, that decisions on the type and parameters of distribution functions should be entrusted to the data themselves.

20.3.3 Randomness, Dependence and Independence

A mathematical model of randomness is a suitable tool for the simplification of models of mass events, where there are simple interactions, that lead to confusing chains of causes and effects. Typical examples of such application fields are the dynamics of gases, plasmatic states of matter and the theory of nuclear reactors. It is inappropriate to think of this type of randomness as chaotic, and/or maximally disordered. The opposite is true: these dynamics result from compliance with the strict laws of Nature, particularly those of mechanics, electrodynamics and nuclear physics. The independence of many regular microscopic events thus leads to a strict interdependence of the measurable macroscopic parameters of the environment such as temperature, pressure, concentration, distribution and flow of mass. In the case of nuclear reactors, the macroscopic effects of microscopic movements include the rate of the fission reaction and thusly the reactor power. The “random” origin of microscopic uncertainty results in a high degree of certainty on the macroscopic level. There are also application fields, where the data uncertainty has nothing in common with this (random) character of events, and where the assumption of the random cause of a real effect would be absurd. Again using data from financial statements: they most certainly are uncertain as it was set out in chapter 2, but an investor would not be satisfied by an explanation, that profit fell because of random events. Each individual change in assets, liabilities, cash flows or expenses is documented and is accounted for. As such, all of these effects have identifiable causes and could be traced if complete information were available. This, of course, is not the case for an observer, who is given only a limited access to insider information. For him the data are uncertain, and this kind of uncertainty is due to a lack of knowledge about the object and/or process characterized by the data.

A similar comment can be made about product quality control. It would be hard to believe, that a defective product was shipped from the factory randomly, or that its defects originated by a chance. Current procedures in product quality include the systematic registration of all factors, that influence the quality of output, so that the causes of potential defects can be discovered immediately and even traced after a complaint surfaces.

Because of substantial differences in such situations from true random ones, the following cautions are appropriate, when:

1. there is only one (or a small number) of possible causes for the effect being considered,
2. the size/intensity of such disturbances or causes is substantial.

In such cases,

1. a small number of strong disturbances in a limited data sample thwarts any hope of correctly applying Gaussian distribution functions,
2. the treatment of such data requires, that they be checked for homogeneity and for the presence of outliers.

The interpretation of normality as something usual, frequently occurring, natural or ordinary suggests a 'jeu de mots': "it is not normal to find a normal (Gaussian) distribution in many application fields." Or "normal distributions in many applications are not normal."

In non-mass processes even the notions of independence and dependence take on a different character. The stochastic independence of molecular collisions at different space points again relates to a large number of particles and only the average of these minute effects is of importance. However, a strongly disturbed datum in a small data sample has a non negligible impact, for which there is no compensation from a large number of opposite effects. The dependence of a sample's characteristics on strong disturbances (outliers) must be taken into account and eliminated by provisions to increase the robustness of the data processing method in use.

The use of mathematical functions as models of dependence between variables is limited in practice. The ordinary notion of a mathematical function has the "one-way" nature of $Argument \Rightarrow Value$ or $If \Rightarrow Then$. Using the cybernetic notion of a black-box or system, such a function can be characterized by an interaction $Input \Rightarrow System \Rightarrow Output$. In many real systems the chain $Cause \Rightarrow Effect$ (especially in multivariate cases) is much more complex, because it is impossible to specify the dependency of one variable on others, while stating, that it does not in turn influence them. (This is eg the case of the standard formulation of the statistical regression model with one dependent and several explanatory variables.) Finding such an extraordinary situation in financial statements (such as changes in assets or liabilities), which have no effect on the value of other items would be a rarity indeed[4]. The feed-back cybernetic model $Input \Rightarrow System \Rightarrow Output \Rightarrow Feed-back\ System \Rightarrow System's\ Input$ is more suitable with the feed-back leading from all output variables back to all the variables in the input. The mutual independence of variables in a real system can only be an illusion.

20.3.4 Membership and Homogeneity of a Data Sample

There are several typical reasons for creating and analyzing data samples:

1. The same quantity is repetitively measured and analyzed so as to suppress its uncertainty or that of the measurement channel.
2. The changes in a single quantity is monitored over time by analyzing a time series of measurements.
3. Several quantities are measured at the same time and analyzed to obtain the cross-section (static) model of the relationship between the variables.
4. The time series development of several quantities is measured to analyze the dynamics of these objects.

In all of these cases, only a single object or a single class of objects is to be delimited as the object of interest. Data samples, which satisfy this requirement are classified as *homogeneous*. They are thought of as a single true quantity, a class of similar true quantities or as a time series of true quantities. The space points of variables, which represent these quantities, form a narrow cluster or a time series of narrow clusters. Data uncertainty increases the spread of these points (width of the clusters), but does not change the character of the distribution functions and/or the number of clusters. The proper distribution function of a homogeneous data sample has thus a single mode for the probability density function (only one maximum for density functions having a “bell form” or not more than two for other cases[5]). This definition can be accepted as a criterion for a

[4] This is not related to a single statement, but to a time series of statements of a company or to cross-section analysis of statements of a group of comparable companies.

[5] There is no contradiction in this statement, because the “non-bell forms” are caused by the bounds of the finite data support and the homogeneity is always tested with densities transformed onto the infinite

sample's homogeneity, however it is also theoretically based on the fundamental feature of the global gnostic distribution function, which reliably signals the sample's inhomogeneity by the occurrence of another density maximum/maxima.

The potential inhomogeneity of a data sample gives rise to the following issues:

- Is a given uni- or multidimensional sample homogeneous?
- If the above is affirmative, then what are the bounds of membership of this sample, ie those distances from the sample's minimum and maximum values, which define the interval of values, within which an additional data item could be included without causing the data sample to become inhomogeneous?
- If the sample is not homogeneous at the start, then how should it be decomposed into homogeneous sub-samples (clusters)?

A special case of inhomogeneity is that caused by the problem of an outlier: "To be or not to be an outlier?" Or more particularly: "Is this extremal data item a 'legal' member of that given data sample?"

The usual solution to the membership problem can be trivial or sometimes extraordinarily complex, but it always plays a fundamental role. In classical set theory the problem of whether an element belongs to a set is posed as a primitive notion, ie "everybody knows if this element belongs to that set." On the other hand, in the theory of fuzzy sets the notion of the "membership function" is introduced: "nobody can decide with certainty if this element belongs to that set, but everyone can determine the value of the membership function, and the degree of membership of the element." Both these approaches are obviously subjective. A statistical solution of this problem depends on two subjective choices: the assumed distribution function and the selected significance of the test. In the gnostic case, the decision on membership is objective: non-outliers are data, which are located within the bounds of the membership interval. These bounds are estimated by using the data.

20.3.5 Similarity, Comparability, Models

The notion of similarity as presented in elementary geometry is based on simple rules of proportion. A more advanced concept takes two figures as similar if they coincide after some "allowed" transformations, eg shifts and

data support, where all densities have a "bell form."

rotations. Similar objects can be compared and ordered: this triangle is larger than that (similar) triangle. Comparisons and ordering need to be made with respect to an aspect common to the objects. Similar triangles can be ordered by their linear size as well as by their area, while eg triangles and circles by area or perimeters. Similarity is a base for modeling. A model has to be similar to the modeled object, but its use is reasonable only if the model is simpler than the object. Only the most important aspects of similarity are modeled. Modeling therefore involves a necessary element of abstraction and introduces additional uncertainty.

The goal of data analysis is to draw as much information from the data as possible, so that the largest number of characteristics of the object can be recognized. An analysis is successful, when the resulting model provides a sufficient degree of similarity to the object. This is not an easy task:

1. Real objects have many characteristics. Which of these are necessarily decisive for the model?
2. What degree of abstraction and simplification is acceptable?
3. Since data are uncertain, how can this uncertainty be minimized in the model?
4. How can the quality (truthfulness) of the model be evaluated?
5. Which aspects of the model can and should be used for comparison and ordering of all of the possible models?

The required depth of the analysis will determine the significance and the degree of simplification to be applied to the selection of the model's characteristics. These requirements depend in a large part on the available resources as well as on the type of analysis to be undertaken. For example, in financial statement analysis, judgements about a company are based on a set of financial ratios, which are evaluated on the basis of the mean or median value of data taken from "similar" firms in the same industry. A formal classification of industrial activity is necessarily different from a 'real' classification. Multiple univariate impressions are not comparable to the real multivariate images of an object. Simple rules of proportion cannot lead to the real characterization of the complex state and behavior of a firm. Primitive methods such as these yield primitive and sometimes confusing results, which can only be considered as a general characterization of a firm, not as tools for financial planning and control.

The need to protect the model against "bad" data calls for robust modeling methods, which are equivalent to using a geometry of a curved space. The best choice of geometry is the one, which results in the best fit with "good" data. This again means, that the data should lead to the geometry chosen for their own treatment. Methods used in data analysis must be data oriented, so that they:

1. can adapt the geometry of robust measurements to data,
2. are able to determine the significance of each aspect of the different objects, and
3. can evaluate the quality of the model by ordering and comparing the objects provided by the data.

20.3.6 Censored Data and Trimming

The common feature of censored and trimmed data is, that both suffer from a lower informative value. The informative incompleteness of a censored data item is objective and not a result of the action of a subject (censor). Censored data do not have a firm value (which answers the question of how much or how many), but rather have a limit of the type “at least”, “not greater/more than” or an interval of values. An analysis should not neglect such data, because they also contain information. Simple examples support this statement: “this diagnosis gives the patient a chance of survival of not longer than...”. Or “this painting is worth at least...” The requirement for taking into account censored data enhances the quality of the analysis, but the construction of the necessary algorithms can lead to difficulties.

Unlike censored data, trimming results from the subjective decision of a person, who reserves for himself the right to declare some data as ‘bad.’ Such action can only be justified if reliable information is available excluding the chance, that the data item carries any useful information. Otherwise trimming represents a risk of loss of information. Data are costly not only due to the cost of measurement, but also from the potential information, which they can carry.

There are different rules for trimming, some better than others. Consider the example of a trimmed arithmetical mean. If eg 5% of the smallest and 5% of the largest data values of a symmetrical sample are trimmed, then if there are outliers, they will be cut off, otherwise the precision of the estimated mean will be slightly diminished by not taking into account 10% of the good data. But what if symmetrical trimming is applied to a non-symmetrical data sample? Another more dangerous example of trimming is an a priori fixed “window”—an interval containing all the acceptable values, while all other data are excluded. Such a drastic limitation of “data’s rights” can lead to serious distortions in the results of the data treatment. It is not the data, but rather the analysts, that speak for themselves in such cases. To prove, that such practice exists, it is sufficient to note the forms of so called influence functions, nonlinear data weights of the M-estimates, which are used in statistics for robust parameter estimation of a linear regression. There were eg 10 such functions available in the S-PLUS[6] statistical package (version 4.5). Five of these functions apply influence functions with sharp edges, ie points of nonexistent derivatives. It is hard to believe, that some data values are not entitled to continuous derivatives for their influence function, and that it is reasonable to assume an infinite response to an infinitesimal change. This criticism is not directed at Insightful, whose programmers are just including the more commonly used routines in robust statistics.

Respect for data as the messengers of reality should be manifested by including censored data and by refraining from using trimming and—when trimming is really reasonable—by applying the Czech proverb: "Measure twice, cut once."

20.3.7 Optimality

A statement of the type "This object is the best" is an empty statement until the criterion of optimality is specified. Optimality criteria can be either purely formal or realistic. A formal criterion function or functional is one, which is defined and accepted without taking into account, whether it represents anything real.

For centuries, a popular criterion for data fitting has been the quadratic function. If asked for a reason for this popularity, a mathematician would probably answer, that its advantage is a comfortable numerical treatment, because differentiation leads to a linear function. Looking for the extremum is therefore reduced to solving a system of linear equations. Simple calculation was very important, when no computers were available. A different point of view, offered by a physicist, would note, that in many instances, energy is quantified by a quadratic function of such real quantities as velocity, current or voltage. If these quantities were to model data errors, then minimization of the quadratic function would minimize the errors in the quantification of these energy sources. Such an interpretation is realistic. A quadratic criterion function thus has a double nature: it is both a good formal and a good realistic criterion. However, there is a serious problem: a quadratic error criterion implies the use of Euclidean geometry,

[6] S-PLUS is a registered trademark of Insightful Corp.

which is not always applicable in data treatment because of the resulting unrobustness of estimates.

Ordinarily, the idea of optimality is closely connected with the nature of the application being pursued. A screwdriver is the best instrument for setting screws, but for cleaning teeth a toothbrush is preferred. The purpose of data analysis is to draw information from data, therefore a natural measure of the performance of a data treatment is the amount of information presented by the results. Since data are uncertain, an alternative criterion is the minimization of uncertainty. Therefore, preferred means of data treatment are those, which maximize information drawn from data, or that minimize the uncertainty of the results, eg the entropy. As shown in gnostic theory, the use of such criteria requires a non-Euclidean geometry, but it leads to robust estimates.

There is also an optimization problem, which is different from simple data fitting. In physics there are two formulations of the Laws of Nature. One takes the form of second order partial differential equations, while the other is obtainable by double integration of the equations along a suitable (**optimized**) path. (Two popular examples of such formulations are: 1) Equations of Newtonian mechanics and 2) Maxwell's equation of the theory of electromagnetism). The extraordinary importance of the integral formulation is, that it enables the problem of the optimality of natural processes to be posed and solved by means of calculus of variations. A simple example of a variation principle deals with the inertial movement of a satellite along its path. Theoretically, no energy is spent in the continuation of this movement, while each variation in the path "costs" energy from the firing of the rocket motor. The path is thus optimal in the sense, that it minimizes the satellite's mechanical impulse and energy. Methods of data treatment can also be thought of as "leading" the data along a path. As shown in Chapter 12, optimality in the case of gnostic transformations of data is based on a proof, that along the path of these changes, information drawn from data is maximized and entropy is minimized.

20.4 Advanced Data Analysis

The foregoing has prepared the way to introduce the concept of *Advanced Data Analysis*, which will be interpreted as an analysis respecting the objectivity of data and aiming to draw out the maximum of information, while letting the data decide the fundamental problems of their treatment.

The basic rules of advanced data analysis include following:

1. Do not violate the data by
   (a) subjecting them to unjustified a priori models or distribution functions,
   (b) trimming the data sample,
   (c) imposing on them behavior in accordance to non-smooth functions,
   (d) not respecting their finiteness.
2. Make use of all available data by
   (a) including censored data,
   (b) including suspected outliers,
   (c) including suspected inliers,
   (d) excluding data only after proving their negligible impact on results.
3. Let data decide the
   (a) bounds of data supports,
   (b) outlier/inlier ('membership') problem,
   (c) sample's homogeneity,
   (d) structure of inhomogeneous data samples,
   (e) the metric of their space,
   (f) determination of their own weights,
   (g) proper use of models.
4. Do not shun the use of good non-statistical methods, when statistics fails, or when its application is not appropriate.
5. Use distribution functions instead of point estimates for data characteristics.
6. Take as similar/comparable only objects, which behave in accordance with the same model.
7. Do not blame randomness for effects. Try to explain their causes by using the data.
8. Prefer robust estimation and identification methods over unrobust ones.
9. Select the desired kind of robustness (inner/outer) with respect to any given task.
10. Apply realistic criteria (information/entropy) to optimization.
11. Respect theoretically proved optimal paths for data transformation.

Gnostics has been developed (as described in Part I and Part II of this book) as a theoretical basis for algorithms, which satisfy the requirements of advanced data analysis. Extensive examples of the application of gnostic methods are given in Part III.

20.5 Summary

Good data treatment methods should be based on a mathematical theory, which makes precise, what is assumed and explains how a statement results from the assumptions, which have been made. However, mathematics is an abstract world of its own and has no obligation to use only notions, which correspond to the non-mathematical structure of reality. Results derived from mathematics are not necessarily applicable in practice. There exist thus two different worlds: the real world, where Nature rules and men live and the ideal world of mathematics. In the former, things exist such as they are, while in the latter things are created by propositions such as: "Let . . . be . . . ." From this point of view, the nature of data is strange, perhaps even schizophrenic. They originate in the real world (by measurement) as real objects and "penetrate" through the wall, which isolates the two worlds, into mathematics and there become mathematical objects to be analyzed. The results thus obtained must once again pass through the barrier to be interpreted and applied in real life. This special character of data requires, that the limitations of mathematical notions and methods applied to data be recognized: they must be realistic. These limitations can be summarized as rules to undergird the main idea, that data are objective elements of data analysis. They must be allowed to speak for themselves. These rules are presented as a requirement for any method of advanced data analysis.

Chapter 21

Gnostic Advanced Data Analysis

The requirements for Advanced Data Analysis were set out in Chapter 20 as a set of rules, which raise the level of data analysis over that of oversimplified manipulations of data. Primitive methods, eg univariate ratio analysis based on point estimates of statistics accepted as the 'recommended ratio's values', are popular and widely accepted in economic analysis, because:

- "everybody" understands them and is able to apply them in practice,
- becoming familiar with them is not difficult,
- instruction in these methodologies is provided in a wide range of institutions from high schools through universities,
- they are supported by many economic text books, and
- "everybody" uses them.

On the other hand, overly simplified methods, which are inadequate for the complexity of the problems to be solved, provide more difficulties than good solutions, because they:

- cannot really (efficiently) solve the problem,
- waste scarce information and increase the cost of the analysis by not making use of all the information incorporated in the data,
- give rise to the unjustified illusion, that "things are under control",
- cannot set the stage for a deeper analysis nor provide means to conduct it,
- give their users a false sense of confidence in managing economic propositions.

Data analysis is a substantial element of control in economics. Faulty analysis leads to inefficient control. There are a number of criteria, which can be used for the optimization of the control function and economic criteria are among the most important. However, one quasi-economic principle should be avoided at all costs: economy of thought, which will seldom lead to the desired result. Those, who can subscribe to the idea, that the economical use of information is the paramount criterion for a successful analytical effort, should review the use of the tools, that gnostics provides for this purpose, and which are set out in this chapter.

21.1 Weight and Irrelevance of an Individual Datum

All of the ideas and tools developed for gnostics are tied to data weight (9.5) and irrelevance (9.6) and have a simple geometric interpretation based on the rotation operator of the Minkowskian (9.3) or the Euclidean (9.4) planes. As explained in Chapter 9 (9.2.2), the data weight signifies the relative importance of each individual element of information (the datum), while the irrelevance is a nonlinear measurement of its uncertainty/error. A second interpretation of these parameters is developed in Chapter 10 (10.3.2):

- The deviation of the gnostic data weight from the full weight of 1.0 represents the change of entropy caused by uncertainty (10.26), which in turn permits
- the irrelevance to represent the gradient of the entropy field (10.29 or 10.30). The significance of the entropy field is that by double integration of its sources (written as **??** and **??**), the gnostic probability and improbability of each datum becomes linearly dependent on the irrelevance (10.42) as is seen in section 10.5.

There are two versions of data weights and irrelevances: the quantifying and the estimating types (Chapter 9). The estimating version frequently used in gnostics takes on a working form of 15.12, in which 15.13 is substituted.

The following rather remarkable features of data weights and irrelevances follow from the theory:

1. They are both derived from the realistically established Axiom 1 (5.2, 5.6 and 5.7).
2. They are both inherently connected to a certain non-Euclidean (Riemannian) geometry (9.14 and 9.15).
3. Both are arguments of realistic characteristics of uncertainty, entropy (10.26) and information (10.42).
4. Data weights and irrelevances parametrize the observed datum's path as it develops under the influence of uncertainty (see 12.2, where Theorem 12 proves the extremality of this path).

5. The variational theorems (Theorems 13 and 14) prove, that the quantification and estimation, which follow this path, are extreme in the sense, that
   (a) for quantification, they maximize the entropy increase (12.26) and information loss (12.30), and
   (b) in the case of estimation they maximize the entropy decrease, (12.27) which creates an increase in information (12.31).

   The foregoing as it relates to data weights and irrelevances proves the optimality of measuring the uncertainty of each individual datum in this manner.
6. Theorem 5 (7.2) proves the isomorphism of a linear mapping (7.10) of the quantifying data weights and irrelevances onto the momentum and energy of a relativistic particle. This mapping is invariant with respect to the Lorentz transformations, which shows, that they are valid for all magnitudes of data uncertainty.
7. The mapping, supported by the Momentum and Energy Conservation law of relativistic physics, leads to Axiom 2, which states, that uncertain data should be composed by means of the additive composition of data weights and irrelevances (13.13).

21.2 Gnostic Distribution Functions

21.2.1 Their Origin

Unlike other approaches to uncertainty, gnostics does not introduce the notion of probability as an a priori given building block of the theory. In gnostics, probability is not connected with the idea of collective or even mass events; instead, the function 10.42 (or in an explicit form 11.7), which manifests the features of probability, is derived in gnostics by double integration of the sources of the entropy field (10.62). This equation has a fundamental importance, because it describes the process of the mutual conversion of entropy into information and vice versa. These ideas are derived from Axiom 1 (5.6 and 5.7). This explains why the probability distribution of an individual datum also is a result of the theory and not an assumption. The single parameter of this probability is irrelevance (see 10.42). By Axiom 2 (13.13) it is required, that the irrelevances of individual data be composed additively to obtain the irrelevance of a data sample. However data from homogeneous functions are composed differently than those from inhomogeneous samples: the irrelevances of an inhomogeneous sample are simply averaged as in 13.13 but for the case of a homogeneous sample, the average irrelevance is normalized by the sample's modulus (14.8) as in 14.13. Moreover, there are two kinds of irrelevances, the estimating and quantifying ones. Therefore, when composing the probability distribution functions of individual data, four probability distribution functions are possible: the

ELDF estimating local distribution function: 15.25 with its density 15.26,

EGDF estimating global distribution function: 15.29 with its density 15.30,

QLDF quantifying local distribution function: 15.33 with its density 15.34,

QGDF estimating local distribution function: 15.37 with its density 15.38.

The form of these functions is dependent on three parameters (the scale parameter S 15.13 and the lower (LB) and upper (UB) bounds of the data support (15.23)). The optimum values of these are to be optimized to ensure the best godness-of-fit (section 15.1).

The differences in behavior of these functions and the proper manner of their use is explained in Chapter 15 and summarized in Tab. 15.5 and 15.6. All of these functions are of the type $R_+ \leftrightarrow (0,1)$, ie they are theoretically defined over the infinite data support $R_+ := (0,\infty)$ and they are continuous and differentiable. Since real data are bounded, they are to be transformed onto the infinite interval using formulae 15.20 through 15.23.

As shown in Tab. 15.5, the four types of **DFs[1] are robust in different ways. This feature makes them suitable for the different kinds of applications summarized in Tab. 15.6.

21.2.2 Uses of Distribution Functions

The four kinds of distribution functions can be used for

- estimation of the probability p of each quantile z,

$$p = **DF(z), \tag{21.1}$$

- estimation of the quantile Z for each probability p, so that 21.1 holds, ie

$$z = ((**DF)^{-1})(p), \tag{21.2}$$

[1] The symbol ** designates either EL, EG, QL or QG.

- estimation of the probability density d of each quantile[2],

$$D^{(1)} = \frac{d(**DF)}{d\ln z}, \tag{21.3}$$

- cross-section filtering as will be seen in section 21.6,
- estimation of the location parameter LP (or of location parameters) of a data sample, such that

$$\left(\frac{d^2(**DF)}{(d\ln z)^2}\right)(LP) = 0, \tag{21.4}$$

- estimation of higher derivatives of distributions,

$$D^{(K)} = \frac{d^K(**DF)}{(d\ln z)^K}, \tag{21.5}$$

  where up to 3 K are required for the EGDF in some tasks while up to 4 K are needed for some applications of the ELDF.

These distribution functions can be used to treat not only data, which have numerical form but also data given as intervals (*censored data*) of three types:

- right-censored data (19.6, 19.7),
- left-censored data (19.8, 19.9),
- interval data (19.10–19.12).

Censoring results from an imperfection in the data or incomplete measurement. Even so, in some applications the information provided by these data can be significant.

21.2.3 Special Applications of the EGDF

The significant inner robustness of the EGDF makes it suitable for several important tasks:

1. robust estimation of the location parameter 21.4 (see also 16.3.1),
2. robust estimation of extreme risks and chances, ie
   (a) robust estimation of probability for extreme quantiles (those closest to the data support bounds LB or UB),
   (b) robust estimation of quantiles for extreme probabilities (those approaching 0 or to 1),

[2] Densities and their derivatives are derived by $d\ln z$ instead of the simple dz to ensure correspondence of the derivatives' form with the independent variable depicted on the logarithmic axes

3. simultaneous robust estimation of such values of the scale parameter S_g of the global type and of the data support bounds LB and UB, which minimize the data fitting error,
4. robust estimation of bounds LSB and USB of the membership interval of a data sample,
5. robust testing of data samples for homogeneity.

The optimal values of parameters S_g, LB and UB are found by solving equation

$$f_{max} = \max_{Sg,LB,UB} \sum_{m=1}^{N} f\left(\frac{EGDF(Z_m)}{E_{MF,m}}\right) \tag{21.6}$$

with respect to these parameters. Function $f(*)$ is the fidelity, ie the estimating data weight $f(*) := 2/((*)^2 + (*)^{-2}))$ (15.12), points Z_m are the sample's data ($m = 1, \ldots, N$) and $E_{MF,m}$ is the weighted empirical distribution function calculated by 15.8 through 15.10. This method, described in more detail in 15.1.5, is called the *maximum fidelity fit.* In some applications, due to the nature of the task, one or more of these parameters may be known. In such cases, these fixed values are used in the application of 21.6.

The test for homogeneity of a data sample (also described in 15.3.7) consists of the determination of the number of roots Z of equation

$$\frac{d(EGDF)}{d\ln z}(Z) = 0. \tag{21.7}$$

The sample is homogeneous if this equation has only one root Z on the infinite data support R_+.

The bounds LSB and USB of the membership interval have been justified by Definition 16 (15.3.7) and they can be found as values which satisfy the inequalities

$$0 < LSB \leq Z_{min} \qquad Z_{max} \leq USB < \infty, \tag{21.8}$$

while they are roots of both equations 15.41 and 15.42.

The sample characteristics LB, LSB, LP, USB and UB, obtained by the EGDF, allow any arbitrary positive number Z to be classified with respect to its possible relation to a data sample $\mathcal{S}$:

- $0 < Z \leq LB$ … improbable value in $\mathcal{S}$,
- $LB < Z < LSB$ … lower outlier in $\mathcal{S}$,
- $LSB \leq Z \leq LP$ … lower or central member of $\mathcal{S}$,
- $LP < Z \leq USB$ … central or upper member of $\mathcal{S}$,

- $USB < Z < UB$ ... upper outlier of $\mathcal{S}$,
- $UB < Z < \infty$ ... improbable value in $\mathcal{S}$.

This method of classification for data (*global interval analysis*) can be transferred onto the finite data support by back transformation of the five boundary points using the inversions of functions 15.20 through 15.23. It is worth noting, that global interval analysis relates only to homogeneous data samples, and that it is unique when the global scale parameter SG is used.

21.2.4 Special Applications of the ELDF

As shown in 15.3.1, the flexibility of the ELDF (15.25) is unlimited and is completely under the control of the user by choosing the scale parameter S. This feature can be used in *marginal cluster analysis,* described in 15.3.2, which is applicable to both homogeneous and inhomogeneous data samples and consists of repeated calculations of the ELDF with different values of S to reveal the inner structure of the data sample. Clusters of data manifest themselves by separate peaks in the probability density. Parameter S decides the resolution power of the analysis and the number of separate clusters to be shown: the smaller S, the more density peaks. The user determines how detailed the structural information must be. The most suitable technique is an interactive interpretation of the ELDF and data density visible on the screen. Once the necessary S is determined, an inhomogeneous sample can be decomposed into individual (hopefully homogeneous) clusters by cutting off the clusters at the points of their density's minima.

The technique of marginal cluster analysis has an important extension. In combination with robust modeling it can be used for *robust multivariate cluster analysis* as described in section 21.8.

Just because the ELDF has unlimited flexibility does not mean, that it is unrobust. As shown in 15.3.3, the peaks of the individual clusters manifest a significant local robustness with respect to a datum which is potentially added to a given data sample. The location parameter Z_0 of a cluster is the root of the equation:

$$\frac{d^2(ELDF)}{(d\ln z)^2}(Z_0) = 0. \tag{21.9}$$

With a data sample of $N-1$ fixed data, the parameter Z_0 varies depending on the value of a free N-th data item ZX which extends the sample.

A special value of the location parameter Z_{00} is defined by relations

$$Z_{00} := \lim_{ZX \to 0} (Z_0) \equiv \lim_{ZX \to \infty} (Z_0). \tag{21.10}$$

There are four subintervals and one significant point in the interval $(0, \infty)$, where the value of ZX may vary. If

1. $0 < ZX < ZL$, parameter Z_0 **falls** from Z_{00} to $Z0L$ as ZX increases.
2. $ZL \leq ZX \leq Z_{00}$, parameter Z_0 **rises** from $Z0L$ to Z_{00} as ZX increases.
3. $ZX = Z_{00}$, then $Z_0 = Z_{00}$.
4. $Z_{00} < ZX \leq ZU$, parameter Z_0 **rises** from Z_{00} to $Z0U$ as ZX increases,
5. $ZU < ZX < \infty$ parameter Z_0 **falls** from $Z0U$ to Z_{00} increasing as ZX increases.

The points ZL, $Z0L$, Z_{00}, $Z0U$, ZU together with the bounds of data support are determined by the ELDF and the process is called *local interval analysis.* These points are estimated using 16.33 together with the location parameter as defined by 21.9 applied both to the original and the extended sample. A complete description and solution of the problem is described in 16.4.3. Classification of data by local interval analysis establishes the following intervals:

- (LB, ZL) ... less than typical data,
- $[ZL, ZU]$ ... typical data,
- $[Z0L, Z0U]$ ... tolerance interval of the location parameter $Z_0(ZX)$,
- (ZU, LU) ... greater than typical data.

The adjective "typical" results from the reaction of $Z_0(ZX)$ to changes in ZX: both changing in the same direction—both decreasing or both increasing for a ZX in the typical interval as if these values for an additional data item were "accepted" by the data of the sample. If ZX moves out of the interval of typical data, then the changes in Z_0 go in the opposite direction, such values of the additional datum appear to be "rejected" by the sample's data. The tolerance interval is also worth a mention: it shows the bounds of possible changes in $Z_0(ZX)$ when ZX changes over the interval from zero to infinity. The tolerance interval is always a narrow subinterval of the (finite) interval of typical data: infinite changes in ZX result in only small bounded changes in $Z_0(ZX)$. This is a manifestation of the high robustness of the location parameter Z_0 (21.9).

Local interval analysis provides an efficient method to compare two data samples and to determine the degree of their similarity as shown in 16.4.3.

21.3 Robust Estimation of Scale Parameters

Scale parameters play an important role in gnostic applications. They can be thought of as parameters that determine the curvature of the data space. This in turn establishes the metric of the space which then controls the degree of robustness of the algorithms. Several types of scale parameters are used in gnostics:

- global scale parameters which ensure the best goodness-of-fit for the EGDF in the sense of the
  - minimum Kolmogorov-Smirnov statistic, $S_{G,KS}$ (16.3),
  - minimum entropy, $S_{G,ME}$ (16.5),
  - maximum fidelity, $S_{G,MF}$ (16.6), and the
  - minimum mean absolute values of fitting errors, S_{L1} (16.7,
- global scale parameters which ensure the best goodness-of-fit for the QGDF based on the same criteria as in the case of the EGDF,
- the local scale parameter $S_{loc}(Z_0)$ which fits the ELDF at the point Z_0 by satisfying equation 16.12,
- the recursive local scale parameter $S_{rec}(Z_0)$, the recurrent version of $S_{loc}(Z_0)$,
- the scale parameter S_{RF} which ensures a required fidelity for the ELDF's fit to data (16.18),
- the variable scale parameter usable for heteroscedastic data (data with varying spread) and for cluster analysis (16.2.5),
- a free scale parameter which is chosen by the user to conduct marginal analysis.

Using a global scale parameter results in a unique form for the EGDF; it is optimal with respect to the corresponding criterion function. The most universal scale parameter (maximum fidelity) is $S_{G,MF}$, the value of which can be taken as a reference. To achieve the L1-optimality, the scale parameter S_{L1} can be used. Local scale parameters are applied to the ELDF. The parameter S_{rf} is particularly applicable to local interval analysis, the results of which depend on the scale parameter. To standardize this analysis, the following steps are recommended:

1. Fit the data sample using the EGDF,
2. determine the mean fidelity of this fit, and
3. find S_{rf} which ensures the same mean fidelity by fitting the ELDF, finally
4. apply this S_{rf} to the ELDF to conduct local interval analysis. Alternatively: set the required fidelity subjectively.

Estimation of global types of scale parameters is robust with respect to outliers because of the robustness and uniqueness of the EGDF. In the case of the QGDF, global scale parameters are robust with respect to central data (inliers). Local scale parameters S_{loc} and S_{rec} make use of mean data weights. They can be robust either with respect to outliers or inliers—depending on whether estimating or quantifying weights are used.

As previously noted, the ELDF's degree of robustness can be controlled by changes in the scale parameter. For the EGDF and the QGDF the scale parameter is determined uniquely. Hence, robustness is also uniquely established in these cases.

21.4 Robust Variance, Covariance and Correlation

The notion of covariance is predominant in gnostics, and the variance and correlation are obtained as special applications of the gnostic covariance. Just as the other fundamental notions of gnostics, covariance is derived from the two gnostic axioms, and it is obtained from consideration of the different forms of the data sample's modulus, in particular from the form shown in 14.21. Here, covariance appears as the arithmetic mean of the product of the individual pairs of irrelevances of the elements making up the data sample. This—together with the role it plays in the calculations of the modulus—justifies the name *auto-covariance* for this expression (14.19). Its immediate generalization is *cross-covariance* (14.20). As in statistics, the products in both types of covariances are formed by irrelevances of data "distance" (lag) which is constant. When this distance is zero, the products become the squares of irrelevances, and the result is the gnostic variance (14.28). Gnostic correlations (14.29), are just as in statistics covariances normalized by the product of the square roots of the two variances. All gnostic covariances, variances and correlations are based on irrelevances, of which there are two versions: quantifying and estimating. Both of them are robust: the quantifying version with respect to central data (inliers), and the estimating version with respect to outliers.

The relationship of these gnostic characteristic to the analogous statistics results from the analysis of the limits associated with weak uncertainties: very small errors. As it is shown in 14.60, in this special case, gnostic covariances converge to statistical covariances. The advantage of the gnostic characteristics lies in their more universal applicability to strongly dispersed data (large uncertainty/errors), in their robustness, and more specifically, in the ability to choose between the two kinds of robustness.

All gnostic covariances are dependent on the scale parameter which determines their robustness. They also are dependent on the location parameter of the data sample; it is possible to say, that they are “centralized”.

The availability of robust covariances opens the way to improving the results obtained by well-known statistical methods (such as eg Principal Component Analysis, Factor Analysis and Discriminant Analysis) all of which make use of statistical covariance matrices. The application of gnostic covariances improves these methods by making them more robust.

21.5 Robust Modeling

The robustness of a model is determined by the metric used to evaluate modeling errors (residuals). The minimization of the sum of squared residuals (the Ordinary Least Squares method) applies Euclidean geometry to a rectilinear space. As is well-known, this results in unrobust models. The data space has to be curved, so that error measurement is dependent on the specific point in the space. The problem becomes how to determine the geometry and the resulting curvature.

In gnostic applications to modeling, the geometry and curvature are determined by the data which are treated. In order for this to be accomplished, gnostic criterion functions are applied to the modeling errors and extremized by the model's parameters. As discussed in Chapter 17, six realistic criterion functions are available for this purpose:

1. the sources of the Q-entropy's field (gnostic Q-variance)[3],
2. the sources of the E-entropy's field (gnostic! E-variance),
3. Q-information,
4. E-information,
5. Q-entropy (quantifying data weight),
6. E-entropy (estimating data weight).

The properties of these functions together with references to theoretical formulae are summarized in Tab. 17.1. Models which result from the application of quantifying criterion functions are robust with respect to the inliers of the data samples, while applications using an estimating criterion lead to robustness with respect to outliers.

[3] As has been past practice, the symbol Q- defines ‘quantifying’ and E- ‘estimating.’

Gnostic criterion functions are applicable to both linear and non-linear models. The special case of a linear regression model is analyzed in detail in Chapter 17 and it shows, that the effect of using gnostic criteria is equivalent to the application of certain non-linear filters to the dependent variable and certain non-linear weights to the explanatory variables. The characteristics of these filtering and weighting functions are summarized in Tab. 17.2.

An interesting innovation to robust modeling is the application of probabilities of the data instead of directly using the data themselves (17.55). This is possible, because of the availability of robust gnostic distribution functions for each sample. Advantages of using probabilities include the:

1. unification of the ranges of all variables in the model to $(0, 1)$,
2. unification of the physical dimensions of the variables (all are dimensionless),
3. possibility of including the effects of the finite bounds of data supports,
4. possibility of using censored data.

It is well-known, that the application of the Ordinary Least Squares method to the linear regression problem is closely connected with the covariance matrix of all the explanatory variables and with the cross-correlations of the explanatory/dependent variables. When modeling in gnostic probabilities, robust gnostic covariances are used instead of the unrobust statistical ones. This is possible due to the linear relation which exists between probability and irrelevance.

21.6 Robust Filtering and Prediction

The application of gnostic criterion functions to filters can lead to desirable robustness. The following classes of gnostic filters can be distinguished:

1. filtering of one-dimensional time series,
2. cross-section filtering,
3. filtering of multidimensional time series.

Gnostic criterion functions can be used to solve the filtering task. Examples are given in next chapters.

Cross-section filtering is a direct application of a gnostic distribution function. Given a data sample composed of measurements made on a selected variable in a group of comparable objects at a certain moment in time, the data are uncertain and it is required to obtain robust estimates of their true values. Cross-section filtering (eg using the EGDF) consists of the following steps:

1. the calculation of the EGDF and verification, that the objects are really comparable (that the data sample is homogeneous as shown by this distribution function),
2. preparation of the weighted empirical distribution function (WEDF) using formulae 15.8–15.10 and application of the WEDF to convert data $A_1 \dots A_N$ to empirical probabilities $P_1 = WEDF(A_1) \dots P_N = WEDF(A_N)$,
3. use of the EGDF to find quantiles $Q_1 = (EGDF)^{-1}(P_1) \dots Q_N = (EGDF)^{-1}(P_N)$[4]. These quantiles are robustly filtered values of data $A_1 \dots A_N$.

The distribution function characterizes relations between the objects as reflected by the values of the analyzed variable. Example: a data sample is formed from the sales of companies belonging to the same industry. These companies are (collectively) subjected to the same external disturbances such as changes in market conditions, taxes, interest rates, prices of energy and raw materials etc. This results in a certain interdependence of the sales of the enterprises which is reflected in the distribution function. But there are also individual factors which cause deviations from the distribution function. Cross-section filtering suppresses these individual deviations and shows what the data would be like if they were subjected only to the common regularity manifested by the distribution.

21.7 Robust Multidimensional Cluster Analysis

There are important practical tasks which cannot be solved unless objects can be compared. Example: ratio analysis used to estimate the financial position of selected firms. While general guidelines exist, there are no theoretical methods which establish values considered as “standard,” “obligatory,” or “healthy.” The problem is, that these ratios are dependent on the specifics of the industry, on the state of the whole economy, on relations which a firm has with its sources of funding, its size and on other factors. Moreover, one ratio cannot provide a complete view of a firm, individual ratios are mutually dependent and deviation of one ratio from a “standard” can be compensated by the deviations of other ratios from their “norms”. However, for financial control as well as for investment decisions,

[4] The symbol $(EGDF)^{-1}$ denotes the inversion of the function $EGDF$

judgments are necessary and these are ordinarily based on inter-firm similarity. Success can be expected only when the comparable are compared. What is the meaning of the phrase, “these firms are comparable?” There are too many aspects of similarity and comparability, which can be used.

A need for comparisons also exists in other application fields. An example based on the famous (and surely realistic) novel, Arthur Hailey’s “Wheels[5]:” “Are cars produced on Mondays comparable with cars produced on Wednesdays or Thursdays?” Or similarly: “Is production quality of the day work shift comparable with the night shift?” And one more: “Is the quality of this product comparable with the product of a competing firm?”

It is not difficult to generalize all these examples: the description of a group of objects is given by multidimensional data samples which can be depicted as points in a multidimensional space. These points form spatial “clouds.” The problem occurs, when there is only one compact cloud or several, more or less separated, clouds. In other words: are the multidimensional data samples homogeneous or inhomogeneous, formed as a “mix” or superposition of several homogeneous samples/clusters? And in the latter case: how should the inhomogeneous multidimensional data sample be decomposed into homogeneous subsamples?

To solve this problem, gnostics applies the following steps:

1. conduct a robust estimation of a multidimensional model (Chapter 17),
2. calculate the modeling errors (residuals),
3. perform a marginal analysis of the sample of residuals using the ELDF to highlight the main peak of data density,
4. identify the data which creates the main peak (main cluster),
5. extract the main data cluster and test it for homogeneity. If it is inhomogeneous, then extract its main cluster and then analyze all the remaining clusters by restarting the iteration from point 1 until all the clusters have been isolated.

This procedure is based on the idea, that there exists comparability/similarity of **multidimensional** objects if their behavior can be approximated by the **same multidimensional model.**

[5] Garden City, N.Y., Doubleday, 1971

21.8 Robust Ordering in a Multidimensional Space

Another view of multidimensional homogeneity can be applied in a qualitative sense: this object compares or does not compare to another object. This interpretation is also important, because it is the basis for the ordering of objects: “this object compares favorably to that object” or even “this object is better than that object.” Ordering multidimensional objects is a complex problem:

1. Even the existence of a well-ordered set of objects is a non-trivial problem: tennis player A beats B who beats C who beats A. Who is the better player?
2. A measure, a multidimensional criterion function or another evaluation system is necessary to order objects in multidimensional space.
3. To minimize the impact of data uncertainty on the order, a robust system must be used.

Gnostics can fulfill these requirements in the following way:

1. Determine the positive, desirable direction of change for each of the variables in the model.
2. Modify the variables, so that an increase in each of them causes the positive effect. (The modification could sometimes consist of simply changing the sign of the variable or in using its reciprocal value.) If there is a desirable optimum for the variable, a new variable, equal to the negative distance to the optimum can be introduced.
3. Find the robust multidimensional model of the implicit type (17.53, in section 17.4.1).
4. Calculate the modeling errors.
5. Order the modeling errors, then order the objects according to the order of their errors.

In some instances, the objects to be ordered will be an inhomogeneous data sample. In such cases, an alternative procedure can be applied:

1. Inhomogeneous ordering accepted: the summary scores of the individual clusters are calculated to get the inter-cluster order.
2. Inhomogeneous ordering rejected: multidimensional cluster analysis performed followed by inner-cluster ordering with respect to the model of each cluster. Inter-cluster ordering can be then realized by additional analysis of the models of clusters.

When modeling errors coincide, the final order can be established by taking into account additional parameters of the objects (parameters, that have not yet been employed in the analysis).

21.9 Summary

Analytical tools which have been developed for gnostics satisfy the requirements resulting from the categorical principle, “Let the data speak for themselves.” As such, they can serve as instruments for Advanced Data Analysis in application fields where it is impossible or not economical to make available large samples of highly precise data.

Chapter 22

Traditional Financial Statement Analysis

It is not the authors' intent to dwell on the mechanics of Financial Statement Analysis, but instead to examine areas, where often the conclusions are not justified, and to consider pitfalls, which may not altogether be clearly understood by the practitioner.

22.1 Current State

22.1.1 Current Usage and Potential Application

Financial statement analysis is frequently used as a method for diagnosing the financial health of a firm. But a diagnosis should not be the goal per se, rather it should serve as support for decision making, and in particular for the establishment of policies, which specify the amount/intensity of actions necessary for the improvement of a firm's financial position, or at least for the maintenance of its status quo. In other words, financial statement analysis should be viewed as a necessary condition for efficient financial control. Unfortunately, the general acceptance of such a role is overly optimistic due to several misconceptions as to the outcome and utility of commonly used treatments, because of:

1. A paucity of hard results in the application of this kind of financial analysis to real problems.
2. Reliance on theories of efficient markets, which limit the motivation of analysts to use the methodologies intended to predict market performance. These issues will be taken up in chapter 25.
3. Technical difficulties in the application of advanced mathematical

methods.

None of these comments or others of similar nature completely describe the problem. Several further thoughts:

(a) Unsatisfactory outcomes often result from the application of inadequate methods.

(b) Unsatisfactory outcomes, when **some** methods are used, does not condemn **all** methods. The progress of science is continually providing new approaches and new opportunities.

(c) Mathematical difficulty is always relative to the level of education of the user. It is evident that an electrical engineer needs to understand the Maxwell equations of the theory of electromagnetism in spite of their "difficult" mathematical form (partial differential equations, notions of field theory etc.). Why then do so many, who work in the economic field, feel that it is sufficient to limit themselves to knowledge of summation, subtraction and proportions in order to describe economic processes, which are much more complex than physics?

(d) Even complex mathematical methods can result in working algorithms that run on computers. Executing such an algorithm by pressing a key is not more difficult than starting the calculation of a proportion. The interpretation of results requires more thought but—as already mentioned—economy of thought is not a very sound policy.

However, a critical view of the role of financial statement analysis as it is currently practiced suggests that the potential for its use is not generally recognized. Most university textbook in both accounting and financial management do not cover the problems discussed below until a student reaches a very advanced level. Therefore the average "technician" or the decision maker in a typical size firm has had only a limited exposure to the rich literature in the specialized field of financial statement analysis, which is found in professional journals and bibliographies such as those cited in [93] or [24].

The objective here is to bring to the attention of the reader those methodological problems that occur and that are frequently glossed over, and to introduce a new non-statistical robust methodology for data treatment, which could improve numerical analysis in **all** applications. These include those found in financial statement analysis, which also provides a source of interesting real data. Given the shortcomings, which are discussed in the following sections, it is no wonder then that the potential power of exact methods is used much less for decision making than are rules of thumb.

22.1.2 The Most Widely Used Accountancy Data

The elemental building blocks of financial analysis are the firm's balance sheets, income statements, cash flow statements and notices; these are generally prepared by the firm's accounting department using a standard set of rules. A balance sheet can be thought of as a compilation of **stocks** of various assets and liabilities at a point in time. In contrast, an income statement describes a **flow** of activities over the period of time, which links two consecutive balance sheets.

These accounting data reflect the company's historical performance; while there is some interest in knowing the path followed to bring the firm up to the present, analysis from the financial manager's (or an investor's) viewpoint is primarily concerned with future performance and therefore the prediction of balance sheets and income statements (proforma) for subsequent periods is also necessary.

The choice of data taken from financial statements to use as inputs to an analysis depends on the goal and depth of the desired results. The most frequently used data are summarized in Tab. 22.1 together with their customary symbols:

Symbol	**Parameter**	**Source**
ADA	Accumulated Depreciation and Amortization	Balance Sheet
AR	Accounts Receivable	Balance Sheet
CA	Current Assets	Balance Sheet
CF	Cash Flow	Statement on Cash Flows
CL	Current Liability	Balance Sheet
$COGS$	Cost of Goods Sold	Income Statement
DA	Depreciation and Amortization	Balance Sheet
DIV	Dividends	Income Statement
EAT	Earnings After Taxes	Income Statement
$EBIT$	Earnings Before Interest and Taxes	Income Statement
EBT	Earnings Before Taxes	Income Statement
IE	Interest Expense	Income Statement
INV	Inventory	Balance Sheet
NS	Net Sales	Income Statement
NSO	Number of Shares Outstanding	Notice
PS	Share Price	Financial Market
RE	Retained Earnings	Balance Sheet
TA	Total Assets	Balance Sheet
TEQ	Total Common Equity	Balance Sheet
TL	Total Liabilities	Balance Sheet

Tab. 22.1 Data most frequently used for financial statement analysis

This is of course only a very narrow extract of potentially available data; eg the COMPUSTAT[1] tape includes 350 annual data items (as well as 178 quarterly ones).

22.1.3 The Main Techniques

The traditional treatment of accounting data includes the following techniques:

1. **Common Size Statements** (percentage analysis): assume, that the data are presented as matrices. One line of such a matrix contains entries representing a selected data item corresponding to different periods, ordered with respect to time (years or quarters).
   (a) *Horizontal analysis:* data in each succeeding period are compared with those of the preceding period. The changes are displayed in per cent and/or as differences.
   (b) *Vertical analysis:* a line is designated as the base (100%) and entries on other lines are expressed as percentages of this reference.

   Common Size Statements thus provide a view of changes in both the statements' individual entries and their inner relative structure.
2. **Ratio analysis:** probably the most widely used financial analysis tool; it attempts to characterize the relation between variables as proportions.
3. **Sequential decomposition of ratios:** the DuPont Chart's thrust of presenting a ratio as a product of other ratios or of their reciprocal values or using the sum or difference of selected data entries in the ratio's numerator results in a multilevel pyramidal structure, that depicts the contributions of the many detailed ratios to the ratio at the summit, which is generally the return on equity[2].
4. **Data scoring:** the computation of scores based on **diagnostic formulae:** a system designed to calculate and interpret a diagnostic 'score' using a weighted sum of several ratios. There are two different approaches to the determination of the weights:
   (a) a *mathematical* approach (eg Altman's method (1968), revised in [1] (1984)), or
   (b) the *heuristic* approach.

[1] Compustat is a registered name of the Standard and Poors Corporation

[2] As shown in [14] and simplified in [64], pyramidal decomposition can also be applied to show the sensitivity of the top ratio to variations of the partial ratios located on lower levels.

5. **Time series analysis** of either raw data from the statements or ratios to provide information on changes, trends or reversal points in the development of the data elements.
6. **Cross-section analysis:** the comparison of the operating results of one company with those of a group of comparable companies.

Percentage analysis provides only very limited information by comparing recent results with those of past periods. Whatever judgments are made on the value of parameters and on the state of the firm's financial structure are thus relative as well as subjective.

The quality of results in both time series and cross-section analysis strongly depends on the quality of the used analytical methods. So eg conservative unrobust methods applied under gross disturbances result in gross distortions rather than in usable and accurate outcomes.

There is, however, a path to the mutual comparison of firms offered by the ratio method, which—together with other procedures—deserves more detailed consideration.

22.1.4 Basics of Ratio Analysis

A popular belief exists, that entries taken from accounting documents bear some relationship to each other, which can be examined simply by computing ratios relating the two variables. Further, if this expected relationship is thought to vary over some range for *all* or *most* firms of a group, then comparing a target firm's ratios with those of similar companies can provide some information as to whether the firm of interest has "better" or "worse" ratios than its peers. The usual caution expressed at this point is, that a ratio by itself provides very little information, and that in order to draw meaningful conclusions, trends over time must be examined, both for the firm, as well as for the set of comparables[3].

A very large number of possible pairs of data can be chosen from the 350 variables found in Compustat, but only a small number of ratios have useful economic interpretation. There are five principal categories of ratios, that are ordinarily used to examine the financial position of a firm:

1. profitability (return),
2. liquidity,
3. activity (turnover),

[3] Among others, industry ratios are published by Robert Morris Associates, Dun & Bradstreet, and the Federal Trade Commission.

Aspect of Analysis	Examples of Ratios	
	Ratio	**Formula**
Profitability:	Return on Assets	$ROA := EBIT/TA$[a]
	Return on Equity[b]	$ROE := EBIT/TEQ$
	Net Margin	$MGN := EAT/NS$
Liquidity:	Current Ratio	CA/CL
	Quick or Acid	$(CA - INV)/CL$
	Working Capital to Total Assets	$RWC := (CA - CL)/TA$
Activity:	Total Assets Turnover	$TATO := NS/TA$
	Accounts Receivable Turnover	$ARTO := NS/AR$
	Inventory Turnover[c]	$COGS/INV$
	Inventory Turnover	$SALES/INV$
Financial Structure:	Liability Ratio	TL/TA
	Financial Leverage	TA/TEQ
	Times Interest Earned	$EBIT/IE$

Tab. 22.2 Examples of Popular Ratios

[a] Taken before interest and taxes for both ROA & ROE to emphasize the business nature of the measure.

[b] Both ROA and ROE can also be evaluated by using EBT or EAT. There is a need in applications to consider the definition of the earnings.

[c] preferred formulation

4. financial structure/leverage,
5. market position.

While the first four groups represent the view “from inside” the firm, the last one is a result of viewing the firm from “outside,” by the market. The most popular financial ratios reflecting the view “from inside” are summarized in Tab. 22.2.

The analyst will choose the ratios to use with respect to the goal of the analysis. However, his/her choice is limited by the availability of data, which is determined by the analyst’s relations with the firm. Many useful ratios require more detailed information than can be found in published financial statements. This is not only because a firm will not disclose all the necessary data, but also due to the time factor: many figures change daily, while the financial statements of publicly held companies are generally published four time a year and with a sensible delay.

The “inner” ratios are thus computed from an “outside view,” and are complemented by several typical ratios, which reflect the market valuation of the firm:

- *Earnings per share (EPS):* Net income minus senior claims[4] on earn-

[4] Eg preferred stocks.

ings divided by the weighted average number of common stock shares.

- *Price-earnings ratio (P/E):* Market price of a share of the company's stock divided by earnings per share.
- *Dividend yield:* Dividend (annualized) per common share divided by market price per common share.
- *Book value per share:* Common stockholder's equity divided by the number of common shares outstanding at the end of the period.
- *Dividend payout ratio:* Dividend per common share divided by net income minus preferred dividends.
- *Total return (TR):* Current share price minus lagged share price + dividend per share all divided by the lagged share price.

A more detailed explanation together with a discussion of the limitations of ratio analysis can be found in the literature (eg [15], [93] and [64]).

22.1.5 'Trivial' Ratio Analysis

The popular and seemingly transparent idea of simplifying the modeling of a firm's financial affairs to a small set of "standard" ratios has led to an evisceration of financial statement analysis by reducing it to the application of several simplistic rules/myths:

1. Accountancy provides a true and fair view of the financial position of a firm. It is therefore sufficient to take data from the financial statements as they are to use them as input for the analysis.
2. The relationship between two financial parameters of a firm can be accurately modeled by their ratio.
3. Two firms of different size have approximately equal ratios for analogous parameters.
4. Two companies in the same industry are comparable in economic terms.
5. There are 'standard' financial ratios (or 'benchmarks') for these 'comparable' firms.
6. These 'standards' can be obtained as the means or the medians over the industry.
7. The 'analytical' portion of the analysis reduces to a discussion of the deviations of individual ratios from the 'standards.'

These ideas can be contested on a number of grounds:

1. As discussed in Chapter 2, the provision of a true and fair view is more of a lofty ideal of accountancy rather than a reality. There are

serious obstacles to the use of raw accounting data without having made a number of adjustments to represent a more accurate cash flow picture of the target's operations. In order to undertake this task, it is necessary to understand both the assumptions underlying the preparation of the relevant statements as well as to be comfortable with the major rules and principles governing the preparation of these documents. For a review of these issues, see Rees [93]. It also must be understood, that there is no way to completely remove the uncertainty (the unexplained components) from the data. Therefore the employed analytical methodology must be capable of dealing with uncertain data.

2. A ratio is a good model only, when the relation really represents a direct proportionality in the variables; this happens rarely. As mathematical models, ratios suffer from several serious problems (small denominator, problem of the sign, etc.)
3. The equivalence of ratios requires, that financial variables be linearly dependent on a firm's size with a zero intercept.
4. The notion, that there could be 'standard' ratios, is illusory since the value of the ratios is subjected to the impact of a large number of internal as well as external factors. This can be easily confirmed by the wide spread of ratios in seemingly similar firms.
5. Information gained from a set of ratios taken from single point estimate (eg mean, median or trimmed mean) is insufficient to judge the actual behavior of the data based on their symmetry, possible bounds, homogeneity and/or inner structure, expected values, outliers, character and size of the spread, etc.
6. The results of such an oversimplified, fuzzy and incomplete data analysis cannot be used for serious decision making.

This trivial approach is frequently used as a means for "window dressing" in annual reports rather than as a basis for financial control.

An important caution is to avoid blindly using 'rules of thumb' to evaluate ratios. There may be (and in many cases there probably are) valid reasons for serious deviations from what could be thought of as the norm, eg liquidity measures for a large public company with easy access to money markets and established sources of standby credit would be significantly lower than for a small rapidly growing company, which has recently gone through an IPO (Initial Public Offering) and has therefore not yet established a track record. Another point to keep in mind is, that many companies do not operate within a single industry; the 'comparable group' is then a weighted average of the various business segments in which they operate.

Moreover, an entire industry group's performance may be substandard, so that average ratios really reflect poor measures. Of course, it is also useful to keep in mind, that a mean value is nothing to brag about, there are as many firms with better numbers as there are with worse ones!

22.1.6 'Comfortable' Diagnostic Formulae

There is a method from mathematical statistics, *discriminant analysis*, which, under precisely defined statistical assumptions, establishes a multivariate space for a surface of a given type. In the special case of a hyperplane the procedure is called *linear discriminant analysis.* The discriminant surface becomes linear iff two independent multidimensional Gaussian populations are to be separated. To calculate the discriminating hyperplane, the covariance matrices of both populations are required. When this approach is used to predict financial distress, many users know it as *Altman's method* due to the pioneering work of Edward I. Altman over the period 1968 through 1984 ([1]). Altman's application separated two data sets composed of:

A Five financial ratios ($R_{n,1} := (CA - CL)/TA$, $R_{n,2} := RE/TA$, $R_{n,3} := EBIT/TA$, $R_{n,4} := NSO * PS/TL$ and $R_{n,5} := NS/TA$ (see Tab.,22.1), taken from the financial statements of the n-th company, $n = 1, \dots N$) of N comparable companies, which had exhibited good financial characteristics over the past several years.

B The same ratios for M companies, which developed symptoms of financial distress over the same time period.

As shown in [1], the application of linear discriminant analysis to these data resulted in the following model:

$$Z_k = 0.717 * R_{k,1} + 0.847 * R_{k,2} + 3.107 * R_{k,3} + 0.420 * R_{k,4} + 0.998 * R_{k,5}, \quad (22.1)$$

that by an experimentally established rule a k-th tested firm with $Z_k > 2.9$ could be considered as being in a good financial condition, while a $Z_k < 1.2$ would signal a danger of financial distress. Values of Z_k between these thresholds would be considered as a 'grey' zone with an undetermined result for the test.

This method is well justified theoretically and can work in practice, but only under the following conditions:

1. The data are really mutually independent and belong to a multivariate normal distribution.
2. A sufficient amount of data of both types is available.
3. All the companies represented by the data are really comparable in economic terms with the tested firm.
4. The coefficients of the discriminating formula 22.1 are recalculated for data relevant to the time period of interest.

A critical review of these conditions leads to the conclusion, that from a practical standpoint, the utility of this method is very limited. However, the simplicity of calculation and the seeming transparency of equation 22.1 is tempting enough to motivate its unwitting application with consequent undesirable developments:

- It is unfortunate, that many economists use the model 22.1 without reference to some or all of these constraints under the impression, that a 'scientifically justified' method is being used.
- There are ample examples in the literature, which describe modifications of the Altman formula proposed by authors, who refer to 'experience,' 'praxis,' 'heuristics' and 'it can be shown that', which result in different weighting coefficients and a different choice of ratios.

This kind of 'willy-nilly' black box pseudo-analysis leads to a lack of appreciation for and degrades the respect, which properly applied financial statement analysis can engender.

22.2 More Problems in Ratio Analysis

Several techniques mentioned above are based on the use of ratios. Ratios of good quality are thus a necessary condition for acceptable results from these methods. Therefore, those factors, which have an impact on the quality of ratios, require more detailed consideration.

22.2.1 The Size Problem with Ratios

The definition of a company's size may be based on total assets, sales, shareholders' equity, etc. These parameters are therefore size-dependent . . . parameterized by the size criterion used. When a group of companies (an industry, for instance) is being examined, the range of these characteristics is generally fairly wide, the parameters of "large" companies exceeding those of “small” ones by several orders of magnitude. The **intent** of ratio analysis is to permit firms to be compared independently of their size, therefore if the size of a company is measured by its total assets, sales, shareholders’ equity or other quantitative parameters, then the expectation is, that the relationship between these values is linear with respect to size. Under these circumstances, the basic underlying assumptions are:

1. *all quantitative parameters of a company are proportional to a size parameter*,
2. *the value of a selected ratio for any pair of comparable companies is equal.*

Denoting the size parameter for the $k-$th company by ξ_k, a pair of line items from the financial statements of the $k-$th company can be written in a general case as functions $N_k(\xi_k)$ and $D_k(\xi_k)$. The former assumption can be thus written as

$$N_k(\xi_k) \equiv \xi_k * n_k \qquad D_k(\xi_k) \equiv \xi_k * d_k, \tag{22.2}$$

where n_k and d_k are arbitrarily chosen size-independent parameters. If relation 22.2 holds, then the ratio

$$\frac{N_k(\xi_k)}{D_k(\xi_k)} \equiv \frac{n_k\text{‘}}{d_k} \tag{22.3}$$

really does not depend on the size because the size factor cancels out.

Under the second assumption a comparison of the $p-$th and $r-$th company then has the form

$$\frac{n_p}{d_p} \equiv \frac{n_r}{d_r}. \tag{22.4}$$

It should be noted, that these conditions, in Euclidean geometry, define similar right triangles with sides n and d . In another geometry the conditions for similarity could be different. Put another way, using ratio analysis, a comparison of the economic processes of two enterprises appears to be mathematically equivalent to the comparison of two similar triangles. Why should this be true, and why should Euclidean geometry be chosen as the medium of comparison from the many other geometries which exist?

Assumption 22.2 is not sustainable because it imposes two conditions:

1. Both numerator $N(\xi)$ and denominator $D(\xi_k)$ of the ratio are **linear** functions of the size ξ.
2. The *intercept* (constant term) of each of the linear functions is zero.

There is neither theoretical support nor experimental justification for these statements. To illustrate, data taken from 55 companies included in the U.S. chemical industry (excerpted from financial statements for the fiscal year ending Jan. 1, 1996) have been taken to evaluate the *Net Working Capital to Total Assets* ratio

$$RWC = \frac{CA \ - \ CL}{TA}, \tag{22.5}$$

which is frequently used as a rough measure of a firm's liquidity and also plays an important role as a control tool in financial management. If the assumptions 22.2 and 22.4 were true then the ratios (22.5) for each of these companies should be equal. Upon verification, it is seen, that these values are spread over the broad interval from 0.0031 to 0.536.

In the sections that follow, the implications of the mathematics of ratio analysis will be clarified, and the necessity for always examining the behavior of a ratio's trend over time should become starkly evident.

22.2.2 Non-zero Denominator

Even grade school children are aware, that division by zero is undefined. It is possible, that this trivial rule may be inadvertently violated by naive users of ratio analysis. Taking return on equity as an example (ROE), the value may not always be positive (if there have been cumulative losses). On occasion, this comes about, because equity is no longer positive. The owners of a firm are not generally pleased under these circumstances, but this state does occur from time to time. Mathematically, the case of equity close to zero or even exactly zero leads to unintended consequences for the meaning of the ratio's value. An example can be found in the set of chemical companies, that is being examined. One company (Comp.X) reported total equity of $18.6M on Jan. 1, 1996. Several earnings measures are also shown in Tab. 22.3 along with values of the corresponding ROEs expressed as a percent ($ROE = 100 * Earnings/Equity$).

The notation is as generally accepted:

$EBIT$ ... earnings before interests and taxes,

EBT ... earnings before taxes,

EAT ... earnings after taxes,

TEQ ... total common equity.

Numerator		Denominator		100*Ratio (%)
Name	**Value**	**Name**	**Value**	*ROE (%)*
EBIT	136.3	*TEQ*	18.6	732.8
EBT	115.5	*TEQ*	18.6	621.0
EAT	7.16	*TEQ*	18.6	38.5

Tab. 22.3 Example of misinterpretation of *ROE* ratios, Comp.X for 1995. (Values are in $M).

Looking only at the *ROE* values, one could easily mistakenly evaluate Comp.X's activity in 1995 as having been extraordinarily successful. An *ROE* based on *EAT* of 38.5%! Reality is more pessimistic: the favorable *ROE* values came not from a large profit, but were due to a very small value of equity. Indeed, the firm reported total assets (TA) of $588 M. Taken together with the equity, financial leverage (TA/TEQ) was an extraordinarily high 31.6. Such a high value would be surprising even if the company were a bank. With only slightly less equity, 0, instead of the actual $18.6 M, all three *ROE*'s would take on infinite values. Even though division by zero is prohibited—as we see—division by "small" numbers should also be avoided due to the unrealistic values that result. However, what is a "small" number? The methodology of ratio analysis does not give a clue; statistics has instruments to test hypotheses on outlying data values, but such tests are based on an even more questionable premises about the data's probability distribution, which in practice is rarely known a priori. Moreover, statistical tests are subjective—the value of the statistical significance of the test must be chosen by the analyst; it is therefore possible, that the same hypothesis, tested on the same data might not be rejected at some significance level, while it would be at others.

Division by "small" numbers can also cause another difficulty with ratios: the problem of the sign. The denominator may change its sign, when passing from positive values through zero to negative ones. The denominator's sign change causes the ratio's sign to change and can lead to other problems as seen in Tab. 22.4 for another chemical company (Comp.Y), which reported total equity of $-96.4 M as of 1 January 1996.

Numerator		Denominator		100*Ratio (%)
Name	**Value ($M)**	**Name**	**Value (MM$)**	*ROE (%)*
EBIT	218.6	*TEQ*	-96.4	-226.6
EBT	-9.34	*TEQ*	-96.4	9.69
EAT	-22.5	*TEQ*	-96.4	23.3

Tab. 22.4 Example of misinterpretation of *ROE* ratios,

Comp.Y in 1995.

A before tax return on equity of -226.6% can under no circumstance be looked upon favorably, but the (in this case) corresponding *ROE* of 23.3% ordinarily would reflect a very successful year, until it was recognized, that this return was produced by the division of two negative values, both equity and earnings!

These examples reveal another source of data uncertainty: not only from a lack of information, but also due to a methodology, which is not suitable to the data to be analyzed. Under certain circumstances, a better representation of the relationship under study may be obtained from the ratio's reciprocal as shown in the following subsection. However, once again the reader is cautioned, that the results must be theoretically acceptable to have any economic interpretation.

22.2.3 The Ratio or its Reciprocal?

The foregoing examples are so trivial, that the spurious results would have been recognized as such and thus ignored during a "manual" analysis. But what about a computer treating the data automatically? Is it easy to establish the 'thresholds' for 'bad' data?

Non-trivial problems arise, when the P/E (Price Earnings) ratio is used. The P/E ratio is a popular tool, which purports to state how many times current earnings an investor would be willing to pay for a share of a company's stock. While the concept has merit, the measure does not precisely provide the desired information. Earnings are at best a current measure although it is more likely, that they represent the profit of a past period; they do indeed 'belong' to the shareholders, in the sense, that their proportional ownership entitles them to a claim on their share. On the other hand, stock price by definition is the present value of all expected *future* cash flows. The P/E is sometimes expressed with respect to the expected value of the next period's earnings; while this is conceptually more accurate, to be completely faithful to the idea, the denominator should be the present value of the *next period's earnings.* But these are not known, they must be predicted introducing yet another source of error.

Therefore, not only do the general comments in the previous section apply, but new problems are also introduced. Moreover, the stock price is a strictly positive quantity, while earnings may be positive, zero or even negative. A meticulous analyst might remedy the situation by ignoring the negative and “too large” values of a P/E ratio[5]. But even so, how are the bounds of “too large” values of the ratio established? If an industry ratio is based on the frequently used arithmetic average of the ratios of a group of companies, how badly is it distorted by the inclusion of such inappropriate values?

Suppose the Earnings/Price ratio were used instead. Define the following ratio:

$$E/P_{\%} = 100 * \frac{EPS}{PS}, \tag{22.6}$$

where EPS is once again earnings per share and PS is the stock price. Using data on the same 55 chemical companies, we find, that the range of $E/P_{\%}$ is much smaller $[-21.0, 14.7]$, while the P/E ratios ranged between $[-62.1, 64.7]$. When choosing a ratio, it should be taken into account, that the volatility of a ratio may be more or less favorable than that of its reciprocal.

The $E/P_{\%}$ ratio is sometimes used by financial analysts as an estimate of the return, which should be received from an investment in a stock. The simple idea is, that one invests PS; then the rate of return should provide $E/P_{\%}$ to the investor in payment for the use of his or her funds. Using the $E/P_{\%}$ as an *opportunity rate,* then the return on investments of various risks can be compared. While some of the strictly *mathematical* problems of ratios are alleviated by using the $E/P_{\%}$, there remain some serious interpretation difficulties from the finance point of view. While the same timing problems, that were discussed with the P/E exist, even if these are properly corrected, the E/P formula represents a perpetuity, and therefore only provides the correct **return** if the firm does not grow and therefore pays out as current return (dividend) 100% of its earnings each year forever. Most companies retain some of their earnings in order to grow: to buy new plant and equipment, fund research for new products, promote its existing product line, etc. Since, as noted above, stock price depends on expectations of future earnings, these retentions play a major role in the growth component of stock price and return. Therefore only in the no growth scenario does the ratio represent a real return and can it be legitimately used as a ‘cost of equity capital,’ ‘required return’ or ‘discount rate’ for the firm’s cash flows.

Instead of a tangible return, then, $E/P_{\%}$ is more of the form of a Return on Investment (ROI) and should be understood to be such except in the

[5] In hindsight, should some of the P/E’s of ‘new economy’ stocks of the late 90’s have been rejected as too high? The lessons from the ‘go-go’ market of the early 70’s seem to have been ignored.

narrowly defined case noted above. Nevertheless, since the formulation is often characterized as a real return, it should be kept in mind, that at least three different 'returns' can be envisaged: define E_0 and P_0 as the current values of earnings and stock price, then:

- $(E_0/P_0)_{\%}$ is the 'return', that is expected, if earnings remain unchanged over the next period and if the investment continues to be held. This is the most commonly used variant and its advantage is, that both values are known with certainty even though its importance from a theoretical perspective is dubious.
- $(E_{(t+1)}/P_0)_{\%}$: a better idea from the theoretical point of view (and it frequently is seen in the financial press), but to be completely faithful, the present value of $E_{(t+1)}$ should be used. Having to estimate the next period's earnings introduces uncertainty (as well as picking a proper discount rate if the present value is used).
- $(E_0/P_{(t-1)})_{\%}$: can be taken as the investor's ROI over the past period. The values are known, and it is a solid historical return on the investor's stake. But it says nothing about future expectations.

To express a true return, the cash flows accruing to an investor must be measured over a set period of time. Therefore, the (total) return is then the profit (excess of proceeds over the beginning period price), which could be realized on selling a share of stock plus any dividend received over the intervening period. Expressed in mathematical terms:

$$TR := \frac{PS_{(t+1)} - PS_t + Div}{PS_t} \tag{22.7}$$

Put another way, the return given the risk of the investment is the familiar dividend growth model first proposed by Myron Gordon [28], for which the return **required** to persuade a rational individual to invest is: $Ke = \frac{Div_1}{PS} + g$ or in its more usual form: $PS = \frac{Div_1}{Ke-g}$, where:

- Ke represents the cost of equity capital, the return that is *required* given the risk of the investment.
- Div_1 is the dividend expected over the next period,
- PS is the current stock price, and
- g is the expected capital gains yield $(\frac{PS(t+1)-PS(t)}{PS(t)})$ attributable to the investment of the retained portion of the owners' funds.

If one can assume, that the risk has remained constant over a period of time, then the changes in PS can be explained as reactions to the ex-post changes in dividend yield and the changes in growth, which have occurred over that time[6].

A more useful interpretation for the $E/P_{\%}$ ratio might be the return, which the *company* earns given the average *market* price of its stock. But using it under these circumstances, its 'hybrid' nature must be emphasized: the firm generates the earnings, a portion of which, the dividend, is paid out, while the rest is retained for growth ... to generate future dividends and other cash flows, which may or may not materialize. The stock price, on the other hand, is set by the stockholders, given their expectations of the firm's future growth and earnings potential. So, the measure is neither a measure of current return to the stockholder, nor an indication of the firm's capability to efficiently employ its assets or its invested capital, and neither party has 'control' over the value, which $E/P_{\%}$'s may assume[7].

It will be shown in Chapters 24 and 25, that there exist serious problems in using ratios such as P/E or $E/P_{\%}$ rooted in a different origin of quantities E and P. Earnings are **internal** (from the point of view of the firm's) evaluation of its performance, while share price is an **external** appreciation established by the market. The former is calculated following strict regulations, documented and mostly audited, while the latter results from the "free game" of market forces. The volatility of market valuations can therefore substantially exceed that of the earnings.

22.2.4 The 'Standard' Values of Ratios

Assume, that $\tilde{r}$ is an estimate of the ratio r consisting of a set of K couplets of parameters N and D. Let us consider the model

$$\widetilde{N_k} = \tilde{r} * D_k \tag{22.8}$$

for all $k = 1, ..., K$ instead of trivial individual "model" (ratios)

$$N_k = \frac{N_k}{D_k} * D_k. \tag{22.9}$$

Both of these relations can be interpreted geometrically, each individually as a point (N_k, D_k) in a plane with coordinates (D, N). These points have a radius vector, the slope of which is α_k, determined by $\tan(\alpha) = N_k/D_k$. In the case of the model 22.8, the tangent is equal to $\tilde{r}$.

[6] Another caution here, if the model is used to estimate the share price, then g is the perpetual growth rate, a very difficult value to estimate.

[7] Of course, the same comment applies to the P/E ratio as well.

The quality of the model $\tilde{r}$ depends on the manner by which it was estimated: the better the method, the better the representation of the data by model 22.8. Using the same 55 chemical companies, then if the initial assumptions of the ratio method were true, all 55 points representing the value of the relative working capital ratio should lie on a single straight line from the origin (the slope of the radius vector of each point would coincide). As is obvious from Fig. 22.1, reality is quite the opposite! The two red arrows point to the slopes, which correspond to the smallest (0.0031) and largest (0.536) ratio. The vectors of each of the remaining points all lie within the sector bounded by the red arrows.

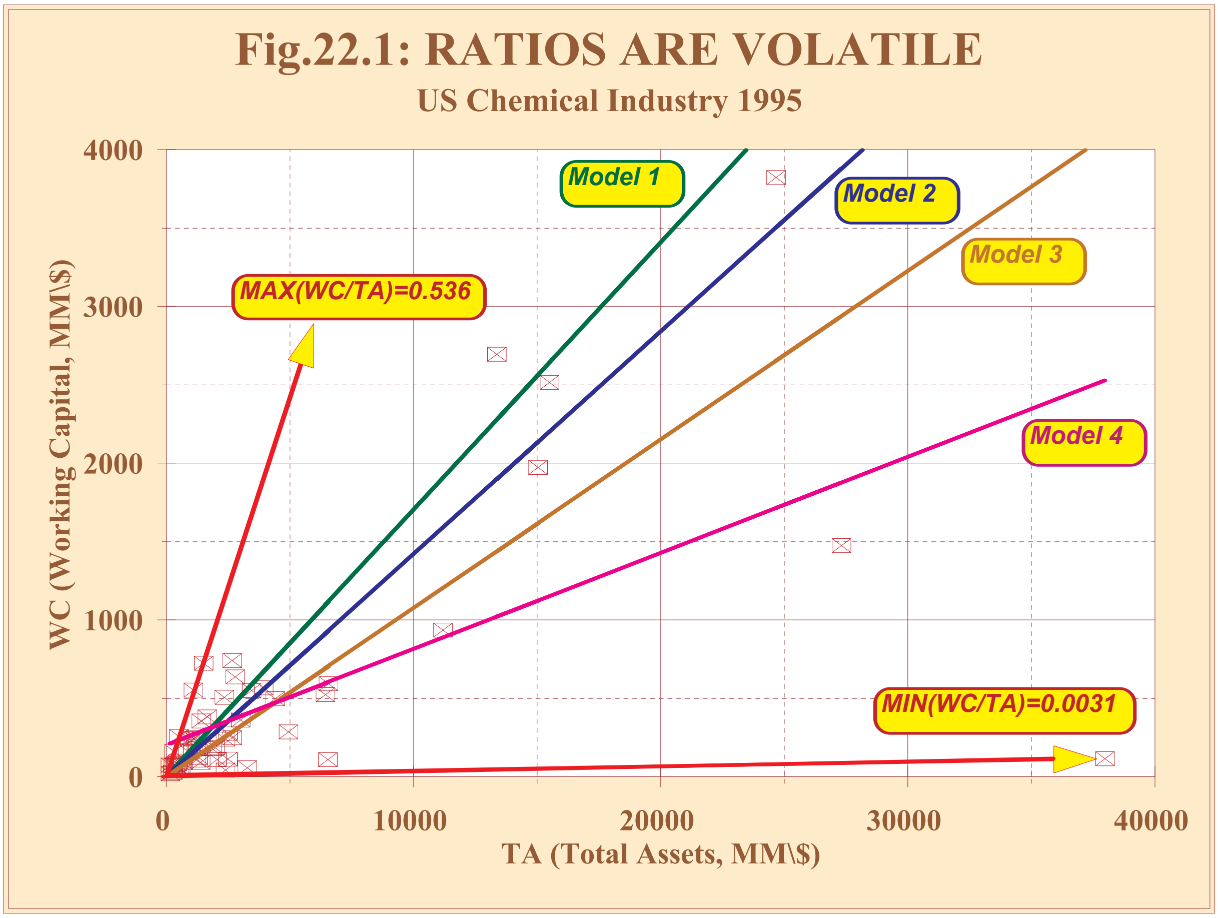


Fig.22.1: RATIOS ARE VOLATILE

US Chemical Industry 1995

The most popular and perhaps the most primitive method, which can be used to estimate the model $\tilde{r}$, is that of the arithmetical mean,

$$\tilde{r}_1 \equiv \bar{r} = \sum_{k=1}^{K} r_k, \tag{22.10}$$

where r_k is the $k-$th firm's ratio. This model, denoted Model 1, and represented by the green line in Fig. 22.1 provides a value for the mean: $\bar{r} = 0.170$. Other alternative models, which could be used include:

$$\tilde{r}_2 = Median(r_k) \quad (k = 1, .., K) \tag{22.11}$$

and

$$\tilde{r}_3 = \frac{\sum_{k=1}^{K} N_k}{\sum_{k=1}^{K} D_k}. \tag{22.12}$$

The second model is the median of the ratios, e.g. the middle of the ordered values. The third model is obtained by consolidating (summing up) all of the individual data to form one very large enterprise, which represents the entire industry. The fourth model estimates the ratio by using the popular Ordinary Least Squares (OLS) method, which is available on most spreadsheet and commonly used statistical packages. Here, the model is: $WC_k = K_0 + K_1 * TA_k$, where K_0 is the constant and K_1 is the slope of the line.

These values are shown on Fig. 22.1 for the given data as:

1. $\tilde{r}_1 = 0.170$ (the green line),
2. $\tilde{r}_2 = 0.142$ (the blue line),
3. $\tilde{r}_3 = 0.108$ (the brown line).
4. With the OLS model, the equation is $WC = 202.7 + 0.0613 * TA$ (the magenta line):
   (a) The model's statistical quality measured by the R^2 is only 0.379, which indicates, that it explains only a small portion of variance.
   (b) Moreover, the *intercept* (the constant K_0) is non-zero. This is counter to the previously noted assumption of the necessity for a zero intercept.

Model 1 was obtained by computing the arithmetical average of the individual ratios, and it has the steepest slope of all the models. The outcome demonstrates a typical result for this method, because the arithmetical mean is very sensitive to large addends, which leads to *non-robustness with respect to outliers.* Model 4, obtained by the OLS method in this instance, has the smallest slope. It also manifests a strong non-robustness, but this time with respect to so called *influential points.* This is caused by the point representing company DD, which has the largest value of total assets, but also the least working capital. It is located in the graph's lower right corner. The median (model 2, blue line) is statistically the most robust of the models tested, but it does not explain the widely spread data very well. The "consolidated" model 3 (brown line) is strongly influenced especially by the respective values of the large companies. Again, company DD has a major impact on the value of $\tilde{r}_3$.

Conclusions drawn from the preceding example as illustrated by Fig. 22.1 are as follows:

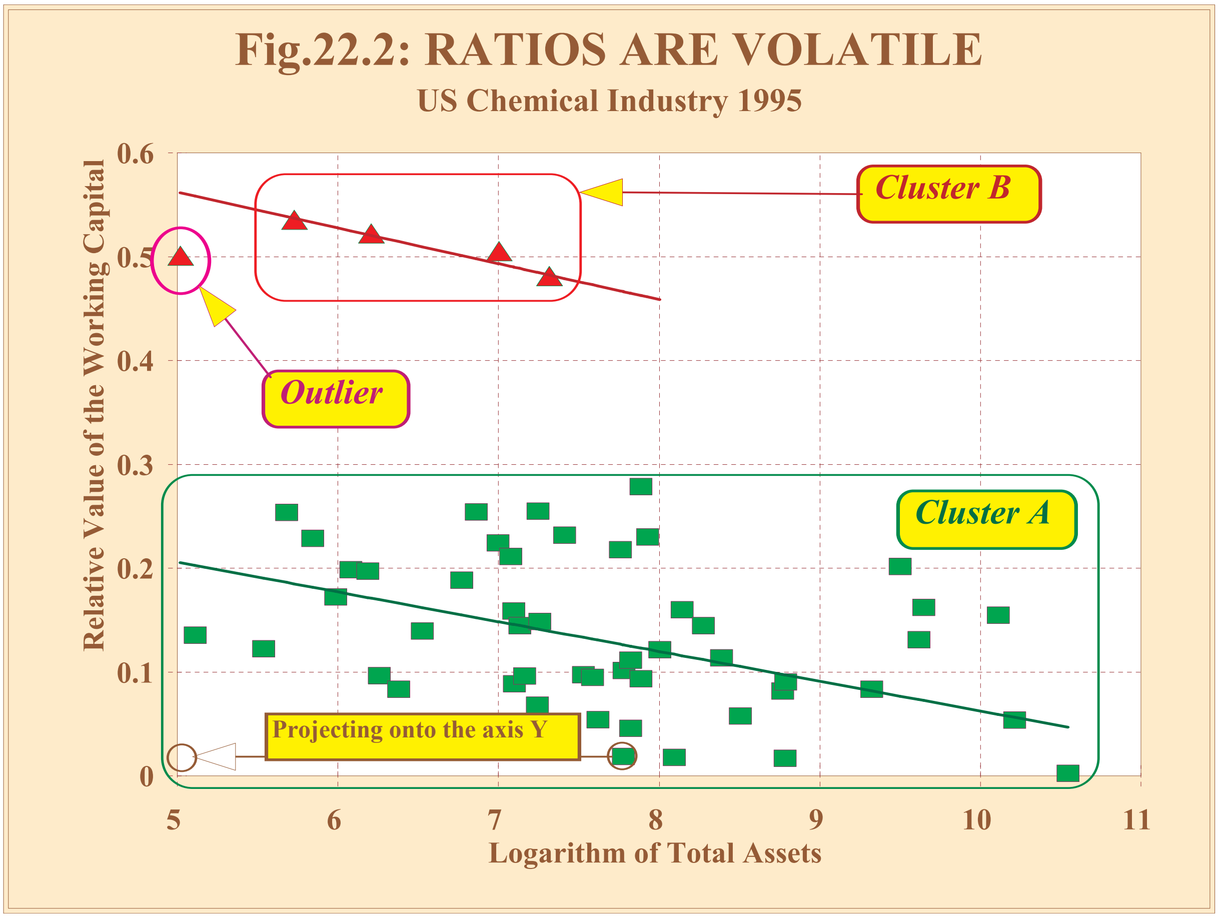
1. Real data do not support the assumptions inferred by the ratio analysis method. A financial ratio calculated for a group of companies in the same industry is not necessarily the same for each firm. On the contrary, the ratios may differ significantly.
2. None of four popular methods for estimating a "typical" ratio value provides a suitable explanation of the data's behavior.

By using the logarithms of the variable values in an effort to improve the above results, the large collection of points representing the smaller companies, which are crowded around the origin in Fig. 22.1 should be "diluted." It is once again noted, that if assumptions 22.2 and 22.4 are true, then the ratios 22.5 calculated over all 55 companies should have the same value and the points representing the ratios should all lie on a single strictly horizontal straight line in Fig. 22.2.

(The $TA-$independence of the ratio would imply the $\log(TA)-$independence.) Just as in the case of Fig. 22.1, the real picture provided by Fig. 22.2 is much more complicated than what is assumed by the ratio method:

- The values of ratios RWC do not lie on a horizontal straight line, but they are spread over a broad interval, which reflects the presence of a strong data uncertainty.
- The fact, that the values of the ratio are split into two autonomous clusters (the main—lower, of 49 companies, green—and the peripheral—upper, red, four companies) infers a strong inhomogeneity in the sample.
- The form of the clusters does not support the idea of $\log(TA)-$independence and the significant dependencies can be modeled by **two** decreasing straight lines.
- There can be individual outliers, which belong in neither cluster (such as LRI with 0.5 in the magenta circle of Fig. 22.2).

The strong data uncertainty observed in Fig. 22.2 motivates a question as to the sources of instability of the ratios. Gnostics rejects the "statistical" interpretation of the random nature of data "disturbances." Instead, gnostics states, that the data uncertainty is due to a lack of information. With additional information, data uncertainty would be less, and the re-



sult would be a better explanation of the data's behavior. The economic characteristics of a company cannot be described by considering a number of individual parameters independently, because the mathematical representation of the firm is a multivariate complex. Each of its parameters is influenced by others and changing the value of a parameter is likely to induce changes in the value of others. This said, then changes in a parameter can be **explained** by the changing values of other parameters, which is possible only when multivariate modeling is employed.

Clearly, it is illusory to think, that there may be a single "recommended value" for a ratio $\tilde{r}$, which can be used as a "standard" for making a judgement about the financial health of a company. This position is further supported by examining the conditions, under which point estimates can be used.

22.2.5 Sufficiency of Point Estimates

Another important, but hidden assumption of naive ratio analysis is, that only two statistics, the arithmetical mean and the standard deviation are needed to describe the characteristics of a collection of data. Such an assumption is only justified for a very narrow choice of special probability distributions, the most popular and better known of which is the *normal (Gaussian)* distribution. For this distribution, these two statistics are sufficient estimates (eg they provide a "good" approximation to the true distribution). However, as it has been repeatedly stressed, there is no reason to expect, that data taken from financial statements will fit this model. A quick look at Fig. 22.2 reveals, that these data plot in two separate clusters as if they belong to two data sets rather than only one; the distribution is far from that of a normal form. A look at the components of the ratio under consideration further demonstrates the general insufficiency of using point estimates of the location parameter to explain simple financial data. Several types of location parameters for the three variables (RWC, CA and CL), which are related through 22.5, are shown in Tab. 22.5 together with the robust means and lower and upper bounds of data support, which have been estimated by gnostic methods see Chapter 16).

Variable	**Location parameter**				**Bounds**	
name	*A-mean*	*Mode*	*Median*	*R-mean*	*LB*	*UB*
CA/TA	0.404	0.369	0.388	0.402	0.099	1.37
CL/TA	0.233	0.233	0.234	0.234	0.073	0.43
$(CA-CL)/TA$	0.171	0.113	0.142	0.163	-0.02	∞

Table 22.5: Point Estimates of Location Parameters

Notation:
CA/TA . . . the relative value of the current assets,
CL/TA . . . the relative value of the current liabilities,
$(CA-CL)/TA$ . . . the relative value of the working capital,
A-mean . . . the arithmetic mean,
Mode ... the location of the maximum of the probability density,
Median ... the quantile of 50% probability,
R-mean ... the robustly estimated mean of the data,
LB ... the lower bound of the data support,
UB ... the upper bound of the data support.

The data in Tab. 22.5 suggest several observations. The variables are neither normally nor lognormally distributed because of the finite bounds of their domains. In the case of a normal distribution, the data support would be $(-\infty, \infty)$ and $(0, \infty)$ in the lognormal case.

Symmetry is another typical feature of both normal and lognormal distributions; all three location parameters must converge to the same value. The coincidence of all four location parameters for CL/TA shows, that the distribution of this variable is symmetrical; however, for the others, the differences between the several types of location parameters reveal a substantial asymmetry in their respective distributions. Hence, the idea of the universal normality of the data distribution must be rejected. It appears obvious, that the characterization of a data distribution by means of a single or some few point estimates is difficult or even impossible.

The main problem of point estimates is, that they only characterize the sample with respect to one aspect. This can be good information only when the data distribution is known a priori (before obtaining the data) and if the distribution is of a "suitable" type. For other distributions, point estimates rarely say anything useful. This statement is supported by the point estimates of ratios CA/TA and $(CA-CL)/TA$ in Tab. 22.5 and even more impressively by the results summarized in Tab. 22.6.

Characteristics	**Year**	**Mode**		**Median**	**R-mean**	**A-mean**	**MSQE**
EPS	1988	0.71		1.19	1.59	1.79	1.93
Stock Price	1995	16.3		15.7	18.6	19.6	15.2
Total Return	1991	0.26		0.37	0.44	0.44	0.37
Dividends paid	1988	0.16	0.69	0.34	0.62	0.56	0.69

Tab. 22.6 Examples of point estimates

Trying to use the differences between individual point estimates is not an acceptable method for obtaining an idea of the form of the sample's distribution. The case of dividends paid is especially unsuited to this purpose. The density distribution of this characteristics—as will be shown in the next chapter—contains two dominating clusters of data. The maxima of their respective densities are 0.16 and 0.69 as shown in Tab. 22.6. The arithmetic mean 0.557—a value in between the clusters, in the density's valley provides no useful information. Another serious problem with the use of the arithmetic mean is related to high values of mean square errors ($MSQE$): the ratios of $AM/MSQE$ are so small, that a potential hypothesis of a non-zero value of the mean AM must be rejected from the point of view of statistics. But could this "hard zero" value for EPS be accepted by economists?

22.2.6 Economic Comparability

To make a judgement about the financial condition of a firm, its economic parameters are compared with those of other similar firms. The difficulty lies in choosing a group of firms, which are truly alike. Just because a number of companies are assigned, the same Standard Industry Classification (SIC) code provides no assurance, that each of these will have the same internal structure or employ the same operating policies. The following are some of the more important features among numerous characteristics, which could be used to differentiate between industry and sub-industry groupings:

- The breadth of products or services offered, which influence the total market exposure of the firm, since few of the sectors of the market behave in the same way or in a synchronous manner.
- The type of inputs used (raw material, energy, prefabricated goods, union or non-union labor, etc.).
- The technology employed: assembly line, large automated machinery, labor intensive processes, retail operation, etc.
- The company's image and reputation, and relative position in its markets.
- Size: not only in terms of relative dominance in market share, but a large established company has access to resources not available to others. Further a firm with a substantial international presence can advantageously employ the resources available to any of its components to dampen other local economic problems asynchronously.
- The existence of strong industry lobbies to assist in obtaining better operating conditions as well as other comparative advantages over both local and international competitors.
- Geographic location determines the distances from sources of raw materials, availability and cost of transport facilities, etc.
- Corporate laws and Market Regulation can provide comparative advantages to operating in certain countries or states.
- Links to the Financial Sector: in those countries, where banks are not prohibited from owning other types of firms, advantageous sources of non-competitive financing may exist.

These points are obviously just a few among those that could be mentioned, but they demonstrate, that there is really only a very small chance of finding a group of firms anywhere that are rigorously comparable to each other. It is, however, possible to use another approach: to begin the analysis with a group of companies primarily chosen by applying some formal criteria: industry membership, technology, etc., but leaving the final choice of firms really comparable to a robust modeling technique, which provides this information as an *outcome* of the analysis, rather than having it imposed as an a priori assumption.

22.2.7 Opportunities and Risks

One of the main goals of a free market environment is to provide **opportunities** to all of the potential participants, who are ready to bear the corresponding **risks**. Opportunities and risks form an inseparable couplet—there are no opportunities without accepting a risk. Moreover, the greater the opportunity, the greater the risk. Both components can be measured by assessing the probability of success associated with the project, eg by the formal quantification of opportunities versus risks. One can clearly recognize the difference between the two following statements:

1. "Take this opportunity, it has a good chance of success with only a modest risk."
2. "This opportunity has a 90% chance of yielding five million dollars, but there is also a 10% probability of losing fifty million."

The first statement is fuzzy and the corresponding decision must be based on a subjective understanding of what constitutes a good opportunity and a modest risk. The quantification of both the amount of the gain or possible loss, and an assessment the likelihood of either of these outcomes is missing. Whatever progress made in the development of the craft of financial analysis, it does not yet include any methodology, which can evaluate probability by any means other than subjective inference. Indeed, to infer probability from data, one must work with data distribution functions, but the necessary tools to estimate these structures are not available in the "simple" financial statement analysis framework, which has been discussed thus far. We have been left trying to **solve complex problems with inadequately simple methods**. The following chapters will demonstrate, that gnostic methods are well suited to deal with the complexity of the real problems of economic analyses.

22.2.8 Feed-back Effects

A firm is a complex multidimensional system influenced by many uncontrolled factors. There are a large number of inputs only partially controllable by the management. All the variables of this system are mutually dependent and there are different feed-backs, which cause changes in output variables to be reflected back to the system's input without any intervention by management. This complicates the control function, because actions taken by management result not only in the required reactions of the systems, but also in undesirable changes. Officers in all armies know and respect the old principle "No command can be given without communication." With respect to a firm this same principle has to be interpreted as "There can be no financial management without financial information." But the financial information must be extracted from the firm's financial data and the databases of other firms by means of financial statement analysis. Ideally, the value of financial statement analysis should be recognizable from the impacts of analytical findings on the efficiency of the firm. The chain of causes and effects should be: past and recent data → financial statement analysis → recommendations for control → financial control based on the recommendations → increased efficiency of the firm → better new data. A desirable side effect should be: a rising importance of the role of the financial statement analysis in decision making.

22.3 Summary

Financial statement analysis has the potential to be a source of vital information for the efficient financial control of a firm. However, there are a number of obstacles to the fulfillment of this positive role: the conservatism of analysts, the attempt to solve complex problems by using inadequate and simple methods, an insufficient level of knowledge of the theoretical background of the employed methods, ignorance of the limitations of methods, that can provide satisfactory results, underestimation of the results of thorough analyzes, and most importantly, not using the results of financial statement analysis for management's real decision making and financial control, ie not closing the feed-back loop: firm's current data → financial statement analysis → financial control.

The solution to this problem is to set aside the current inadequate methodology, to begin to learn more progressive and effective analytical methods, and to follow through and apply these to financial control.

Chapter 23

Advanced Fin. Statement Analysis I.

As described in Chapter 21, there are three levels of gnostic data analysis:

1. Marginal (one-dimensional) analysis based mainly on the application of distribution functions.
2. Pair analysis: examination of relations between two variables by means of correlation coefficients.
3. Multidimensional analysis of models characterizing the mutual dependence of several variables.

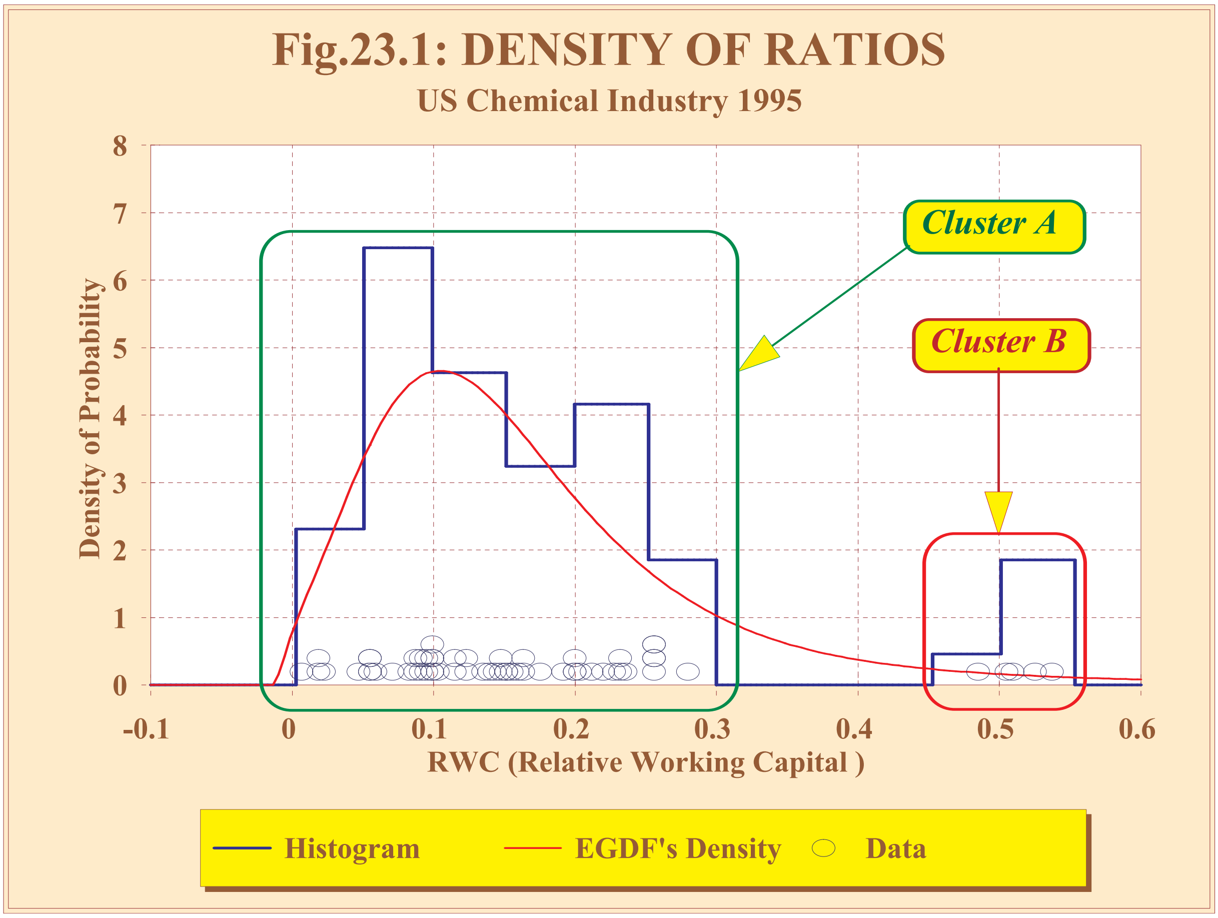
This chapter provides examples of marginal and pair analysis from financial statement data, while the multidimensional problems will be taken up in Chapter 24.

23.1 More on Distribution Functions

Four types of gnostic distribution functions (ELDF, EGDF, QGDF and QLDF) were introduced rather formally in Chapter 15 as outcomes of the mathematical theory. Their features were also studied using mathematical language. However, before beginning to run applications, it is important to master the notion of distribution functions on a level, such that their use for data treatment becomes just as natural as using a knife to cut bread. To assist a user in gaining this necessary familiarity, a series of examples has been prepared to illustrate the use of this not yet sufficiently understood tool.

Figure 23.1 recasts the same relative working capital data shown in Fig. 22.2, but after some simple manipulations.

Imagine, that all the data points in Fig. 22.2 are horizontally projected



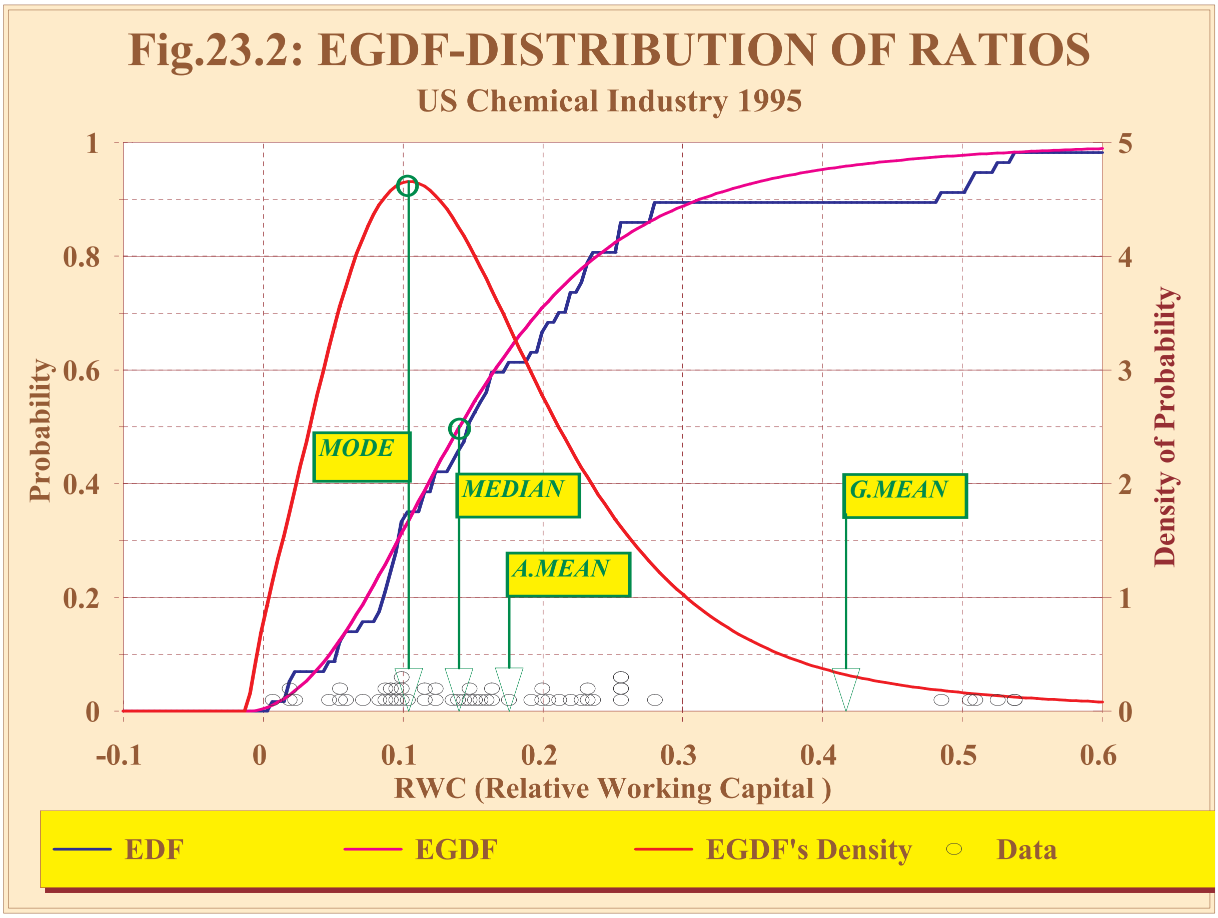
from their position within the chart onto the left vertical (Y) axis labeled 'Relative Value of the Working Capital' and marked by small circles. Now rotate the chart clockwise by through 270^o to construct a histogram over the markers. This results in Fig. 23.1, which is a simple way[1] of getting a rough view of the density of the data (the blue line). The previously noted clusters A and B are once again highlighted in green and red. While the use of "discrete windows" in the histogram provides a general idea of the density, it is only approximate. The shape of the data is better represented using a smooth EGDF density function[2] (the red curve). The density curve illustrates the EGDF's robustness to peripheral outliers and clusters—it preserves its unimodality in spite of cluster B. This can be interpreted as

[1] The histogram consists of rectangles with a width of 0.05 and heights proportional to the number of data falling within this interval. The total surface is normalized to equal 1.

[2] As set out in Chapter 15, gnostics makes available several kinds of smooth data distribution functions along with their densities. Only the **estimation** functions will be discussed here, therefore it is sufficient to distinguish between the global (EGDF) and the local (ELDF) functions.

the rejection of a possible hypothesis, that cluster B causes the sample to be inhomogeneous. The EGDF's density function confirms, that all the data belong to the data set, and that the sample is homogeneous.

The red line in Fig. 23.2 shows the global probability distribution of the Chemical Industry's Relative Working Capital Ratio, while the magenta line plots its density[3].



The blue step function is the empirical distribution function (EDF), which describes the underlying data. The greater separation on the right side between the smooth and step functions is due to the high (B) cluster: although it does not contradict the assumption of homogeneity, its data differ from that in the main cluster. The form of the EGDF in Fig. 23.2 leads to several observations:

1. The distribution function cannot be approximated by a normal (Gaussian) distribution function: Unlike a normal distribution,

[3] Note, that the scale used for the probability density is different from that used in Figure 23.1.

(a) the density is non-zero only over the bounded data support (for data values exceeding the lower bound LB=-0.0127).

(b) the form of the density function is far from symmetrical with respect to the mode ('central' value).

2. Although an occurrence of strongly negative working capital (less than the LB) is not expected, the possibility of zero or slightly negative working capital is not excluded.
3. Different kinds of point estimates of the mean value of the working capital are widely different: the
   - mode of the density function (the location of its maximum, the most frequent value) is 0.105,
   - robust mean of the data (data values weighted by the probability density) is 0.139,
   - median of the distribution function (the quantile for a probability 0.5) is 0.140,
   - sample median is 0.142,
   - arithmetical mean is 0.170,
   - geometrical mean is 0.406.

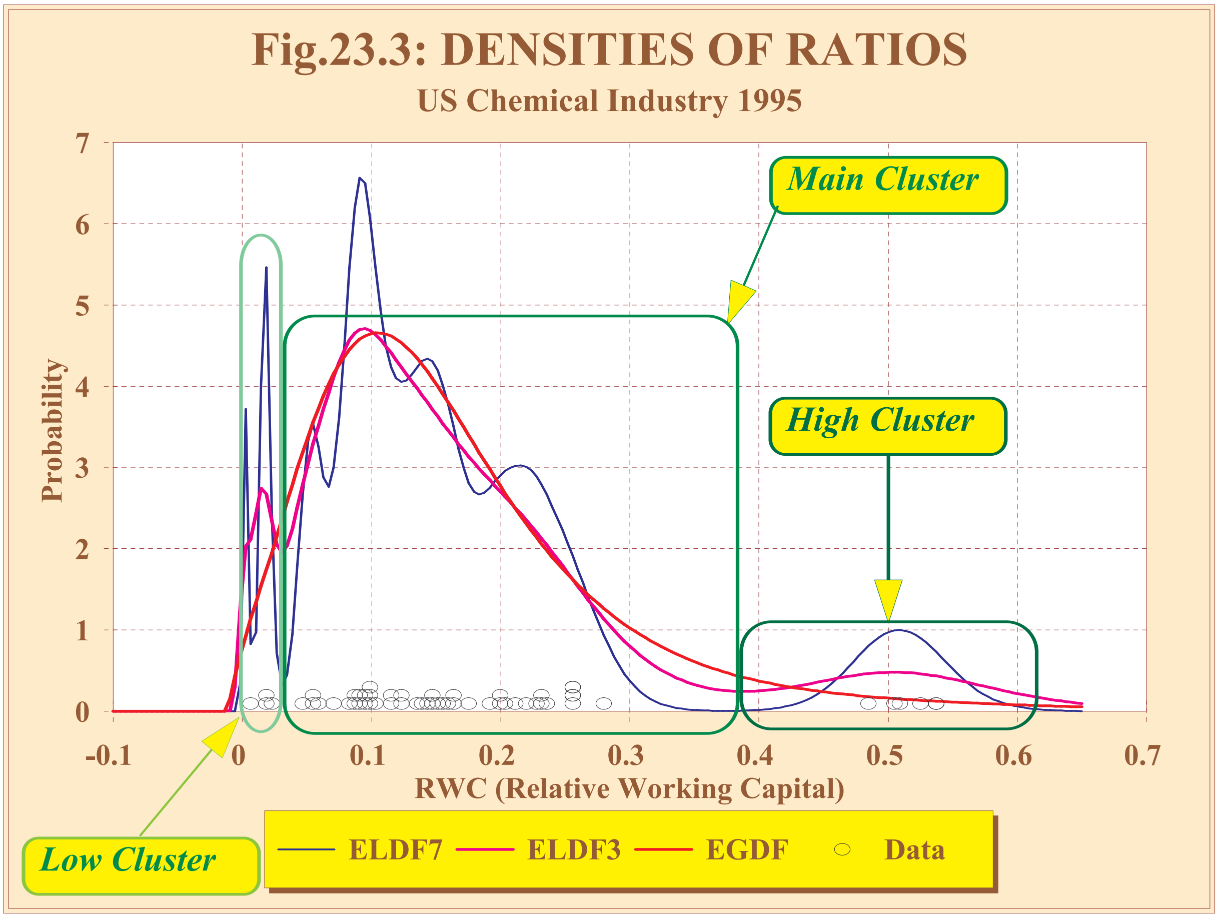
   The large values of the arithmetic mean and especially of the geometric mean are due to the high data cluster.
4. The slow decay of the density curve with rising working capital shows, that the number of companies, which maintain large working capital (high liquidity) exceeds the number of companies, which can operate with a low liquidity.

Each of the point estimates provides only a limited information on the data sample and it would be difficult to obtain an idea of data behavior even by considering all these point characteristics. Unlike this, the distribution function visualizes the data samples and enables not only these, but all others data characteristics (including probabilities) to be obtained.

Applying the local distribution function (ELDF) will provide a closer look at the inner structure of the data sample. This is shown in Fig. 23.3.

Recall, that the amount of detail revealed by the local function is controlled by the value chosen for the scale parameter (S). Two values are used here:

1. The magenta curve (ELDF3), calculated for $S = 0.82$, brings out the three principal clusters, which were revealed in a preliminary way by the EGDF. The bounds of these clusters are now seen to be determined



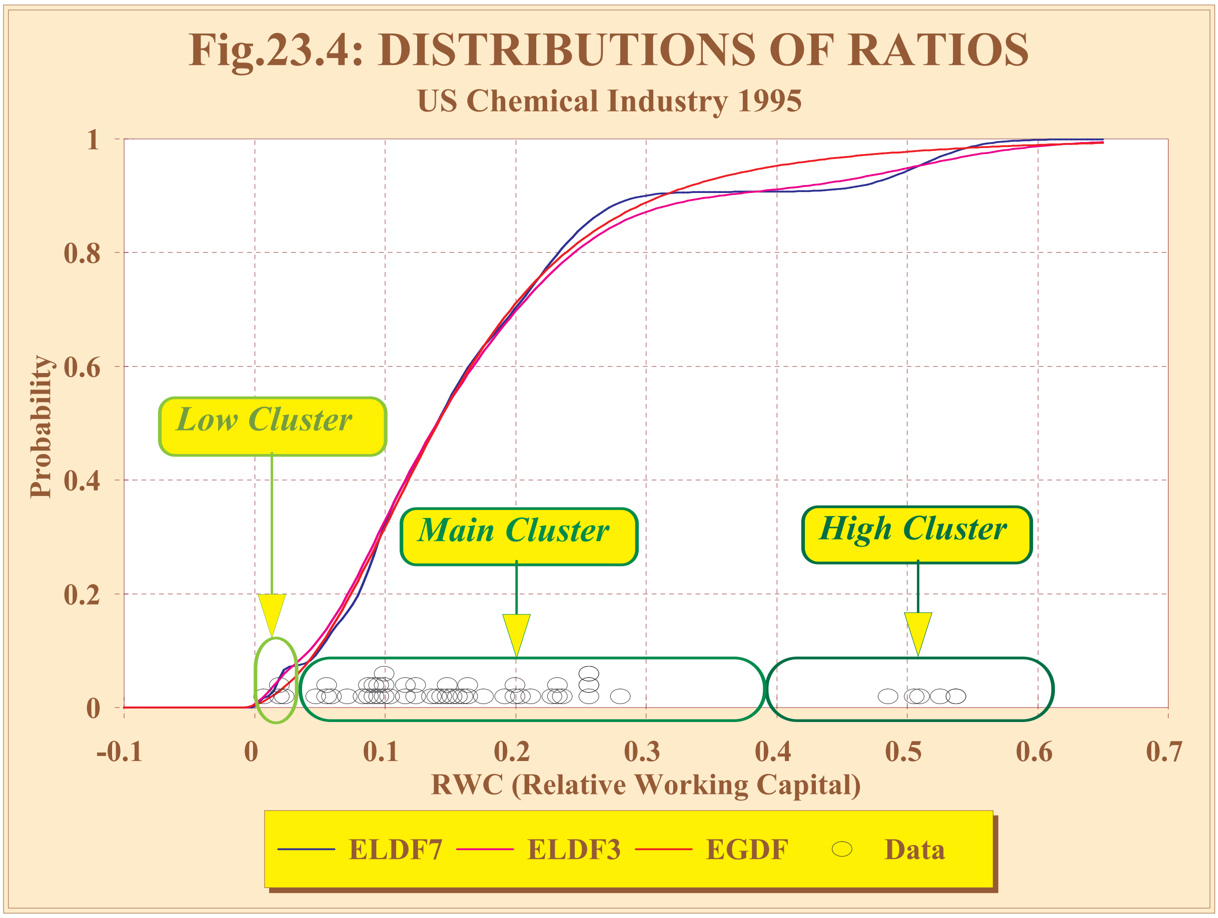
by the local minima of the magenta density function: the interval of values $RWC < 0.031$ belongs to the low cluster, while the interval $RWC > 0.39$ highlights the high cluster leaving the broad main cluster to cover the interval between the two peripheral ones. (The previous cluster A has now been split into two subclusters, low and main.)

2. The blue curve (LDF7) with $S = 0.33$ shows, that the sample has seven clusters:
   - The previous low cluster is now a single outlier and a cluster, which includes three companies.
   - The upper cluster contains five firms.
   - The ratios representing the remaining firms are all incorporated in the main cluster, which now displays four small "bumps" that could be further expanded if greater detail were desired.

Identification of firms belonging to each of the clusters results from the analysis, too.

The previously shown density of the EGDF is superimposed in red for comparison. All three densities are normalized in the sense, that the integral of the area under the curve is 1.

The probability distributions in Fig. 23.4 reveal much smaller differences between the three cases than the densities: the probability of finding ratios of RWC from roughly 0.05 through 0.23 will differ only slightly under either of the three scenarios.



For the remainder of the interval, the value of the probability or quantile read from the EGDF will represent a **filtered** quantity or the value, which would be obtained, were all the data to behave with the same regularity.

A general methodological conclusion can be drawn from the comparison of Figs. 23.3 and 23.4: Distribution functions are suitable for the estimation of probabilities for given quantiles or quantiles for given probabilities (as well as for the estimation of several types of data bounds), but the density functions are more useful for the analysis of a sample's structure and its homogeneity/inhomogeneity.

23.1.1 Distribution Functions as Semi-invariants

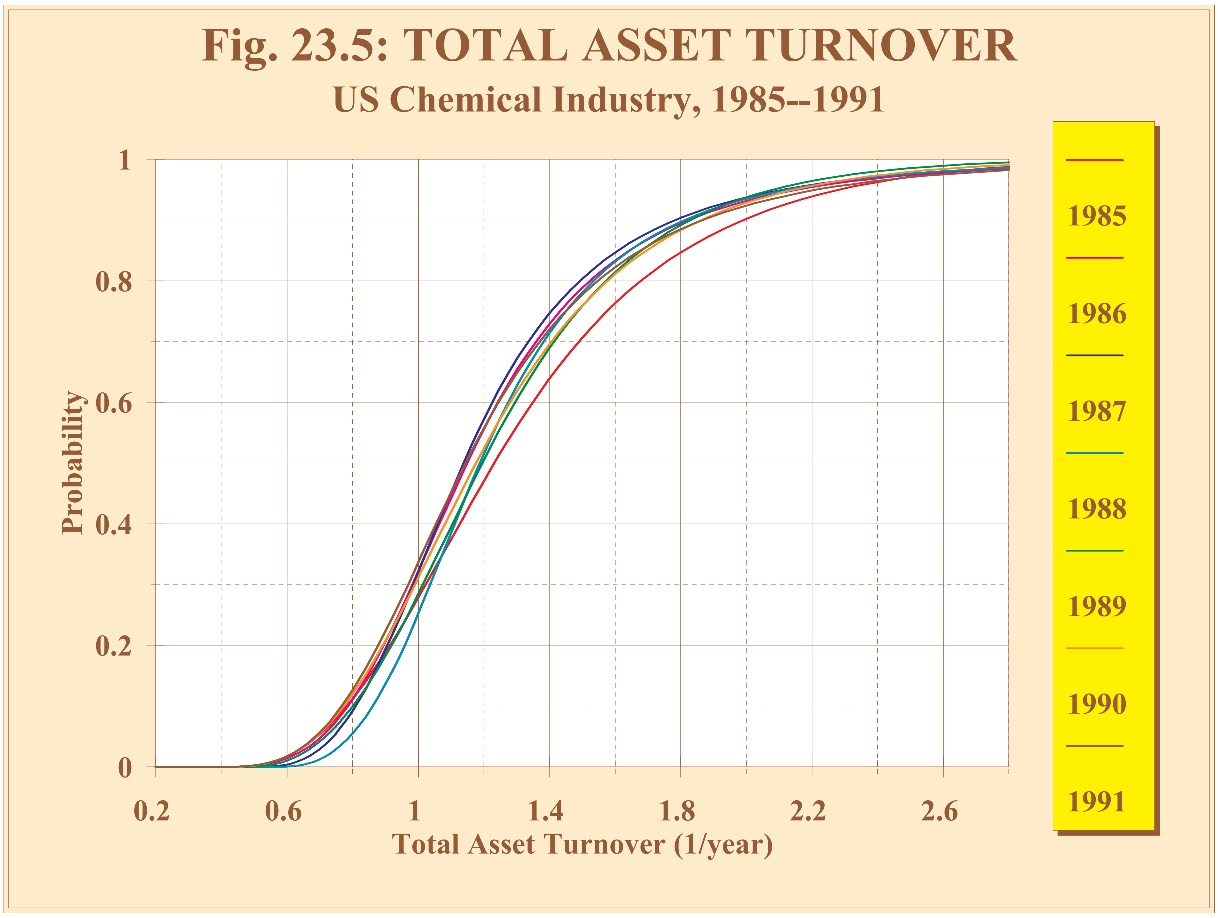
One radio station in Great Britain frequently uses a weather forecast of this type: “The weather tomorrow will be much the same as today’s.” Experience has shown, that this prediction is not less efficient than that obtained by means of satellites, computers and meteorological theories. The explanation is simple: the weather in the UK is fairly consistent (some say bad). The reason for this quasi-regularity is the stabilizing effect of the Gulf Stream. To predict, this regularity—an **invariant**—is used.

The best invariants are mathematical formulations of Laws of Nature. So, eg, equations based on the Energy and Momentum Conservation Law enable the paths of satellites to be predicted with a high precision. In statistics, the natural Law of Large Numbers leads to the Gaussian distribution. Unlike what is found in mechanics, the regularity expressed by this distribution function is not a basis for the prediction of individual events, but only the reliable prediction of mass events is possible. The Law of Nature formulated by the two gnostic axioms leads to optimized distribution functions of individual uncertainty and—due to the composition law—to the distribution functions of data samples, which are not necessarily large. Therefore, there are two sources of regularity for gnostic distribution functions:

1. Laws of Nature, on which the gnostic axioms and the gnostic technology for the construction of distribution functions are based.
2. Laws of Nature, which control the processes, from which data originate, and which “feed” the data by providing information (about them).

The former regularity is completely invariant, independent of any particular data. The latter regularity should be invariant, when the process that produces the data is unchanged. In such a case, the distribution functions of data samples obtained under ‘unchanged’ conditions should also be unchanged. But this requirement must be relaxed, because both the processes and the observed data are subjected to uncertainties. Therefore the need for a precise replication of distribution functions is tempered to require only an acceptable degree in their similarity. On the other hand, if the source of the data changes its state, then the data treatment process should no longer produce invariant results. It is for this reason, that the heading above is labeled *semi-invariant* (in the sense, that the features of the data processing tools are 'conditionally invariant').

Figure 23.5 illustrates such a situation by demonstrating the regularity and repeatability of the probability distribution functions EGDF of the Total Asset Turnover ($TATO$) ratio taken for 50 companies in the US Chemical Industry.



The seven lines each represent annual data over the period 1985–1991. The distributions form a narrow band and so support the thesis of a strong underlying regularity that is only slightly disturbed by events specific to any one year. The activity of each company, measured by its $TATO$, depends on internal factors such as changing market position, assortment of goods produced, development of technology, etc. As seen in the data, the ratio ($TATO$) of each company changes every year. But, in spite of these **individual** changes, the distribution functions for the industry **as a**

whole are very similar. The regularity and repeatability of the process has thus been demonstrated.

The similarity of distribution functions can be viewed in more detail by using the corresponding density curves, which are more sensitive to differences between the forms of the probability distributions. This is shown in Fig. 23.6.

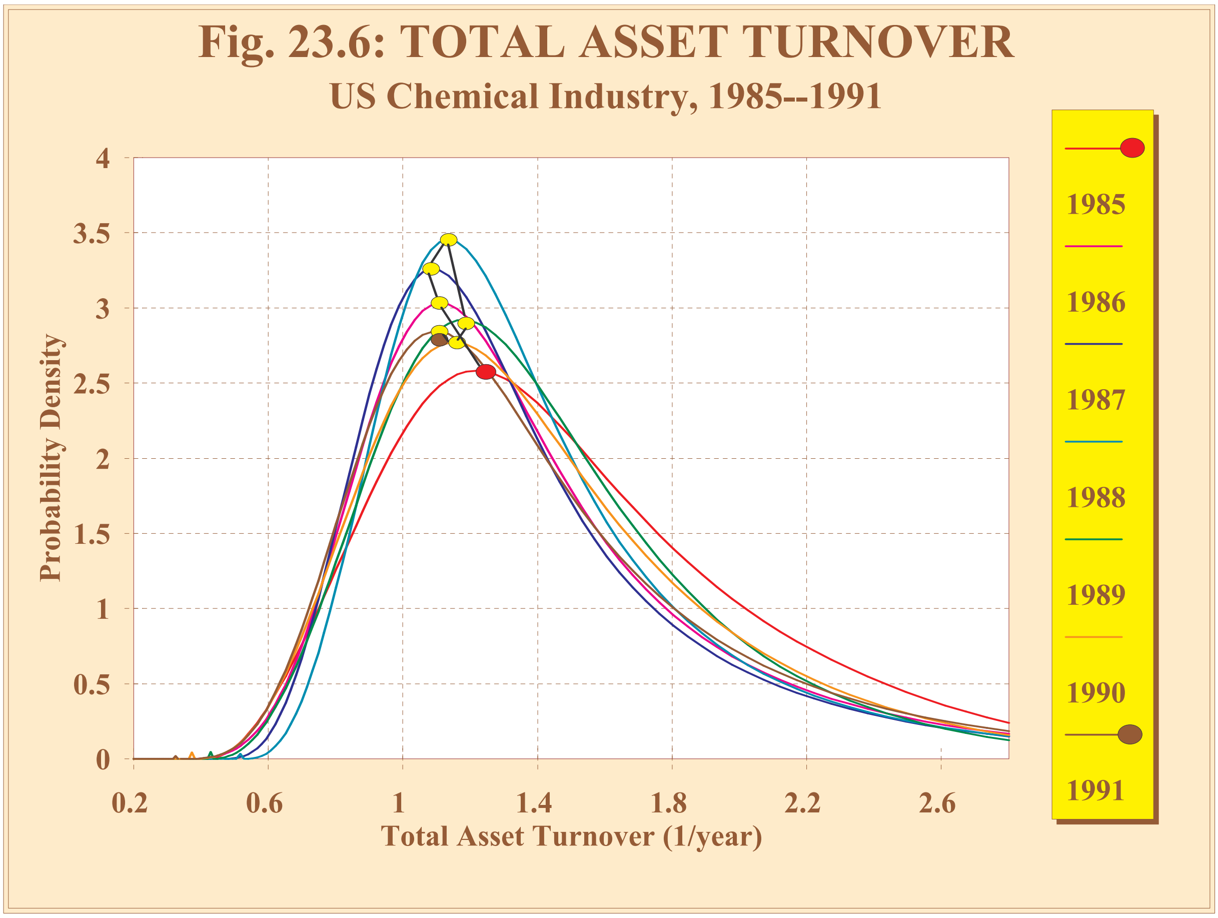


Fig. 23.6: TOTAL ASSET TURNOVER

US Chemical Industry, 1985--1991

The densities are proportional to the reciprocal of the scale parameter $1/S$ (see 15.30). The scale parameter increases with the spread, which causes the maxima of the densities to be dependent on the spread of the data: the narrower the spread, the higher the maximal point in the figure and the steeper the probability distribution in Fig. 23.5.

The positions of the maximal densities (one of the location parameters characterizing a ‘central’ value) in Fig. 23.6 are close each to other: the ‘central’ (most frequently occurring) values of the $TATO$ were only slightly changed over the time interval 1985–1991. Moreover, the heights of all the density curves (the spreads of $TATO$ values) are also similar, which illustrates the small spread in the distributions' slopes. To highlight the changes in the curves and trace the process, the densities' maxima are marked by yellow circles connected by a black line. The path of the maxima follows an elliptic movement with a dominating vertical component: the main changes occurred in the spread of the data rather than in the locations of the densities' maxima ('means' of the $TATO$). Moreover, the path traced by the maxima is smooth, which indicates, that the changes in the curves were caused by real changes in the process.

A careful look at the density curves reveals small 'bumps' (additional small maxima) close to the left extremities of several curves. As has already been seen, such effects are caused by inhomogeneity in the data and are due to lower outliers: $TATO = 0.317$ in 1986, again in 1987 with 0.467, 0.528 in 1988, and again in 1989 (0.437) and in 1991 (0.332). In each of these cases there was only one outlier, the other data were significantly larger: the second smallest $TATO$ ratio in 1987 through 1990 was associated with only one company (0.739, 0.754, 0.709 and 0.724) and belonged to the homogeneous main cluster. The repeated presence of a company as a lower outlier should be surely explained by the specific nature of the company's industrial activity. The reader should note, that neither of these inhomogeneities affected the overall similarity of the distribution functions.

This shows, that the distribution functions manifest a high robustness with respect to uncertainties in the data. However, a question arises at this point as to whether this robustness in the distribution functions might be an obstacle in detecting real changes in the informative content of the data. This problem is given more detailed consideration below.

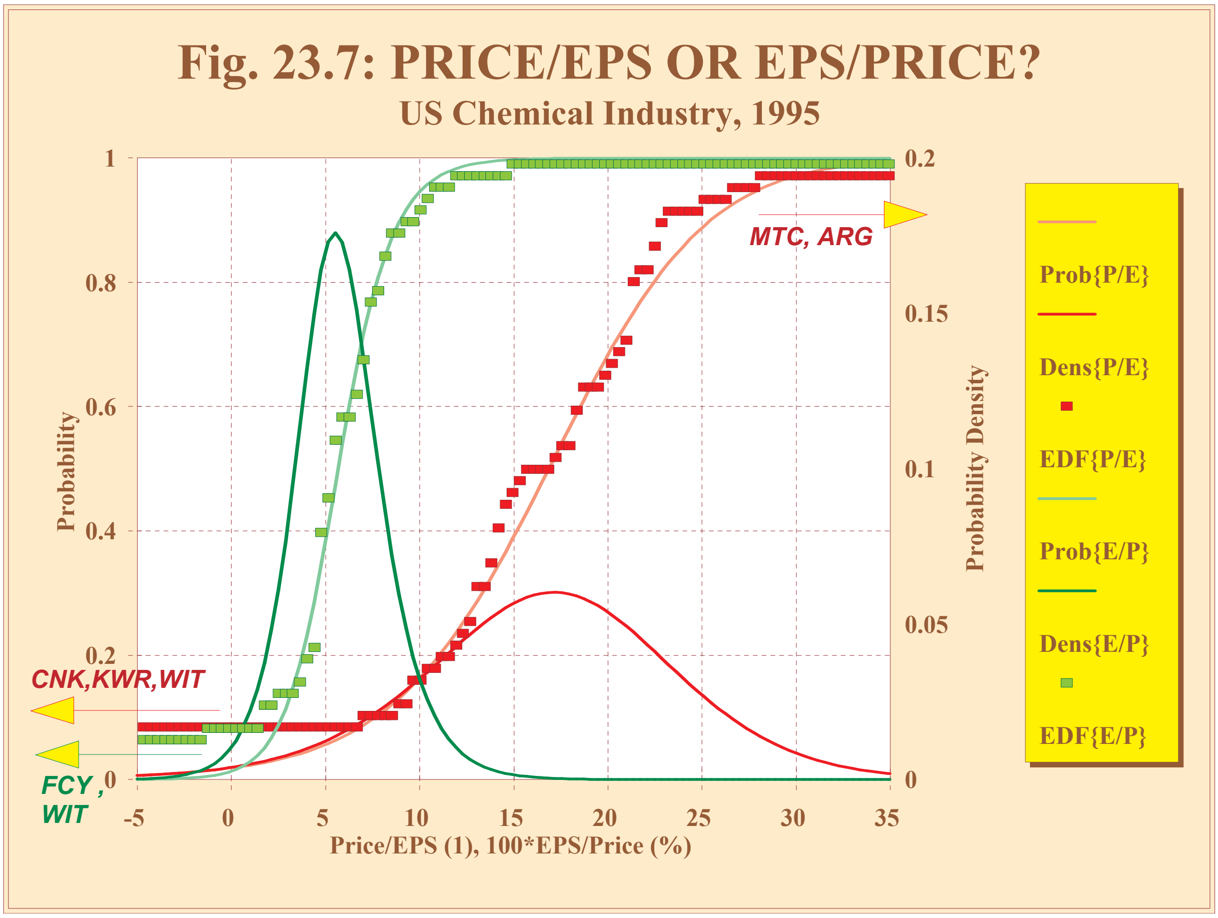
23.2 Direct Application of Distributions

23.2.1 Sensitivity of Distribution Functions to Information

There are two mutually contradictory requirements to an efficient data treatment process: it should be robust with respect to uncertainty, while being sensitive to information carried by the data. The former requirement translates to: "a reduced sensitivity to data is desirable", while the latter means, that "a greater sensitivity to data is desirable." The objective of this section is to demonstrate, that gnostic distribution functions are a suitable tool to resolve this seemingly 'schizophrenic' problem.

As demonstrated in the previous chapter, the amount of information inferred from the popular P/E measure or its reciprocal, E/P is limited. But, these measures can still be useful to illustrate the gain in information, that can be obtained from the data by using gnostic procedures. In Chapter 22, both the ratios were illustrated by point estimates and it was shown, that the spread of the former measure was substantially larger than that of the latter version. More details can be obtained by expanding the analysis to use distribution functions.

Fig. 23.7 plots the gnostic global probability, density, and empirical distribution functions for these ratios (PS/EPS in red; EPS/PS in green)[4].

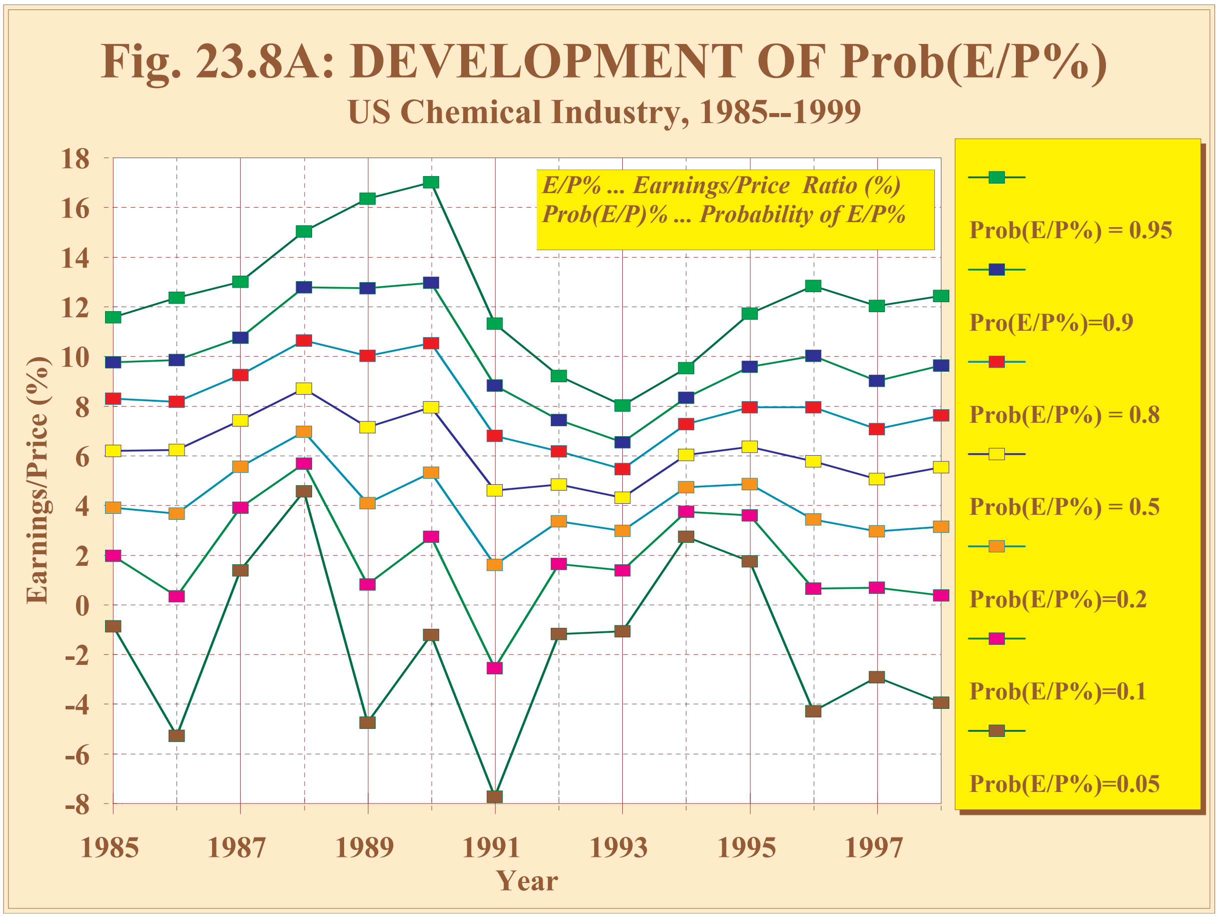


Expressing the latter in percent, so that the same horizontal scale fits both ratios, the values of $Price/EPS$ are seen to be spread over a re-

[4] The data are for the US Chemical Industry for 1995.

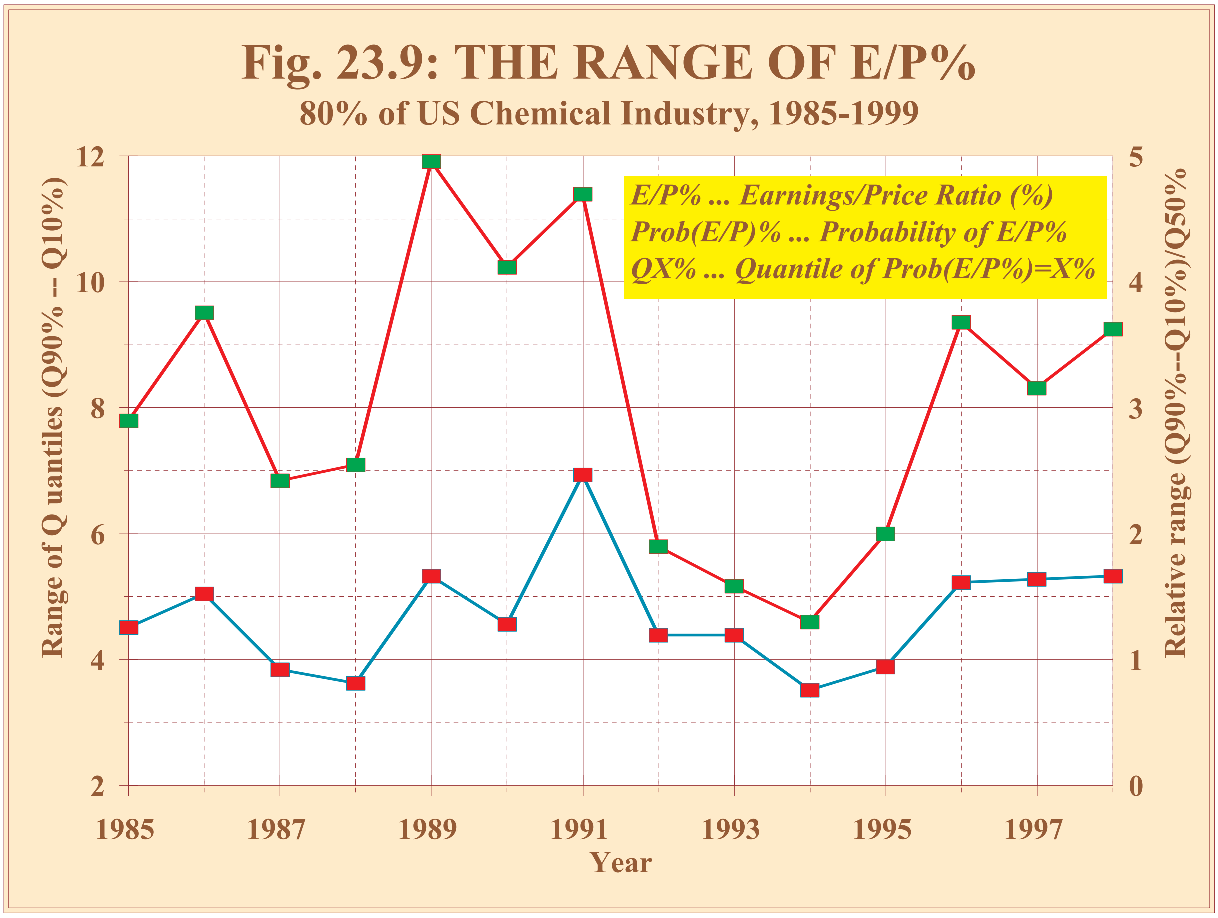
markably broader interval than those of $EPS/Price\%$: the P/E is much more volatile than its reciprocal. The most frequently expected value (the probability density's maximum) for the $EPS/Price\%$ reflects a "return" of 5.68%. Before attempting to venture an economic interpretation of this result it will be useful to examine the time development of the data.

The global probability density functions of the $E/P_{\%}$ were calculated for each of the fourteen years, 1985 through 1999, and their medians (quantiles of probability 0.5, robust location parameters) were estimated together with the quantiles corresponding to probabilities 0.05, 0.1, 0.2, 0.8, 0.9 and .95. The values for these quantiles are shown in Fig. 23.8A, each connected by a straight line.



As an example of the interpretation of these graphs, the pink mark on the light blue line denoted Prob$(E/P_{\%}) = 0.2$ for 1996 lies on the horizontal scale line of 3.44%. This means, that in 1996, 20% of the companies had an

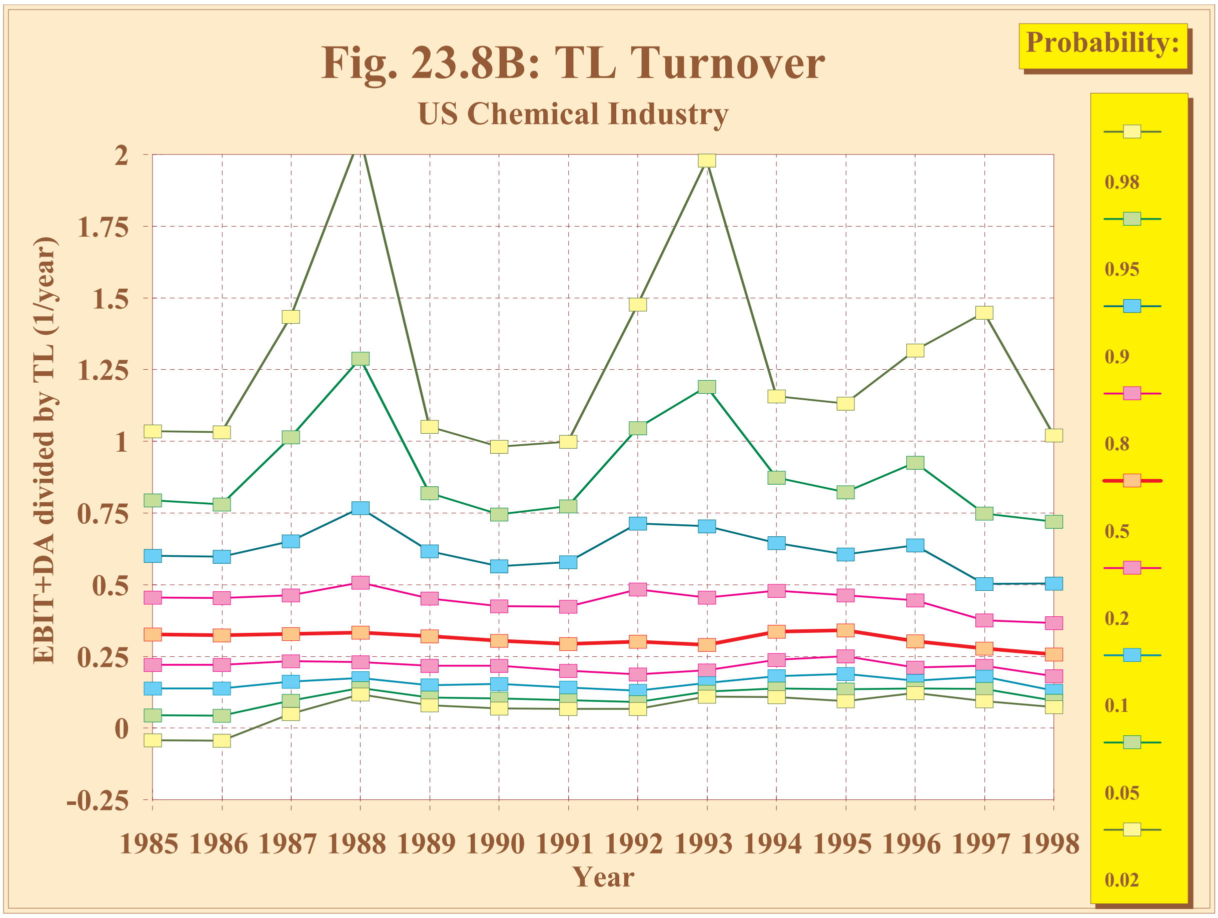
$E/P_{\%}$ ratio equal to or less than 3.4%[5]. The dark blue line thus connects the robust medians of the quantiles, ie points for which Prob($E/P_{\%}$) = 0.5 holds. The vertical distance between the quantile of Prob($E/P_{\%}$)=0.9, which equals 10.0% in 1996, and the quantile of Prob($E/P_{\%}$)=0.1 of 0.7%) can be taken as a measure of the volatility of the variable $E/P_{\%}$. Therefore the volatility in 1996 was $10.0 - 0.7 = 9.3\%$. This measure has a simple interpretation: it corresponds to the range of values of the volatility of $E/P_{\%}$ of 80% of the companies in the US Chemical Industry in 1996. The time development of the range of volatility is depicted in Fig. 23.9.



It rose eratically until 1991; then sharply decreased to rise again after 1994. The figure shows the development of volatility for the 90% (red) and the 10% (blue) quantiles using two scales (the absolute range and its relative value with respect to the median.

[5] Only the 50 companies forming the main homogeneous cluster of the data samples were taken into account for the estimation of these graphs. The values of $E/P_{\%}$ are thus robust mean values not influenced by outliers.

This technique of using "quantile lines" provides an analyst with condensed easily readable information with respect to the development of complex processes. Further examples are given in Figs.23.8B and 23.8C:



There has been recent interest in using $EBITDA$ (Earnings before Interest and Taxes plus Depreciation and Amortization) to estimate cash flow as a predictor of a company's expected performance and its ability to service debt[6]. Such a variable as $EBITDA$ could be used to introduce a useful ratio, $TLTO$ (Total Liability Turnover):

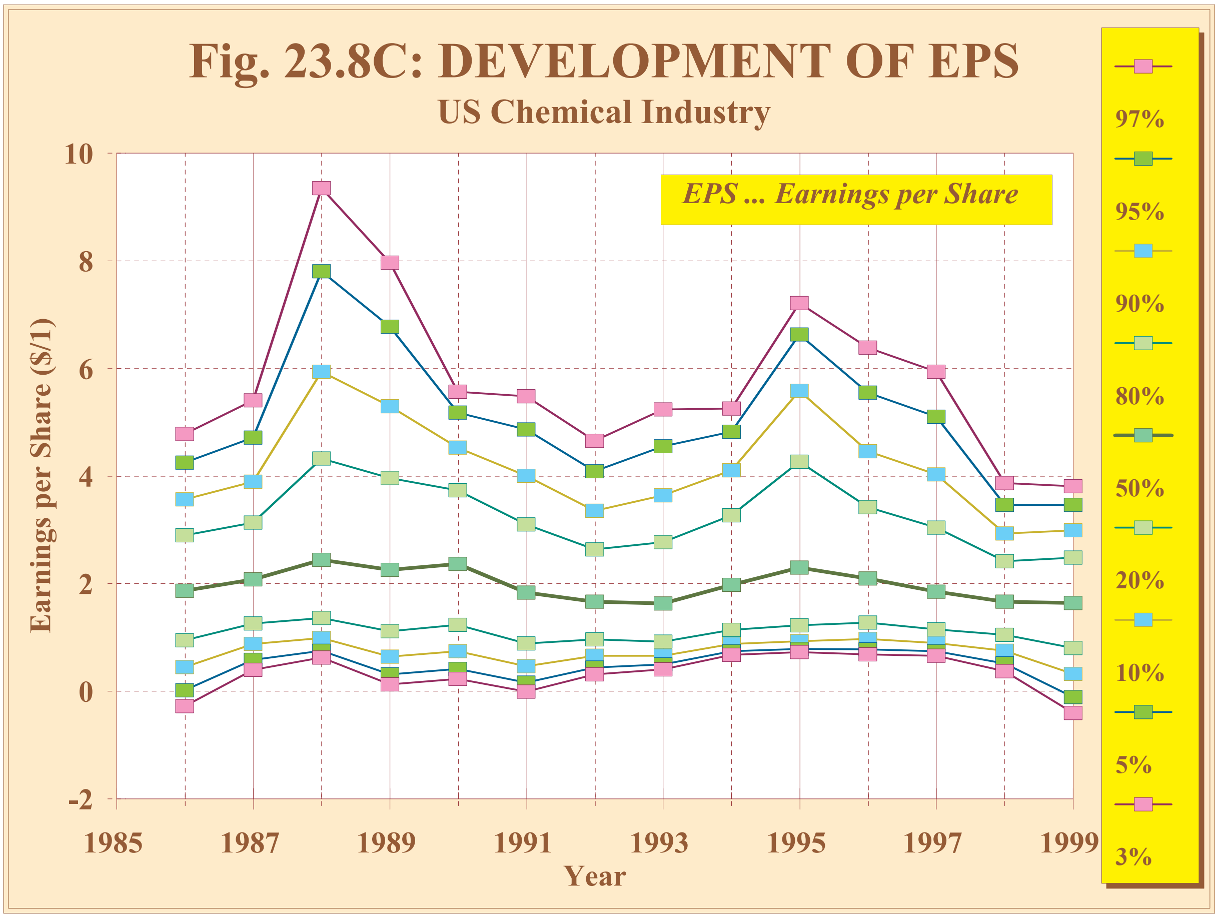
$$TLTO := \frac{EBIT + DA}{TL}, \tag{23.1}$$

where TL stands for Total Liability. This measure provides a useful insight into the "fitness" of the industry as illustrated in Fig. 23.8B: there were three periods of substantial acceleration in the total liability turnover

[6] See for instance: "Putting EBITDA In Perspective," Moody's Investors Service, June 2000.

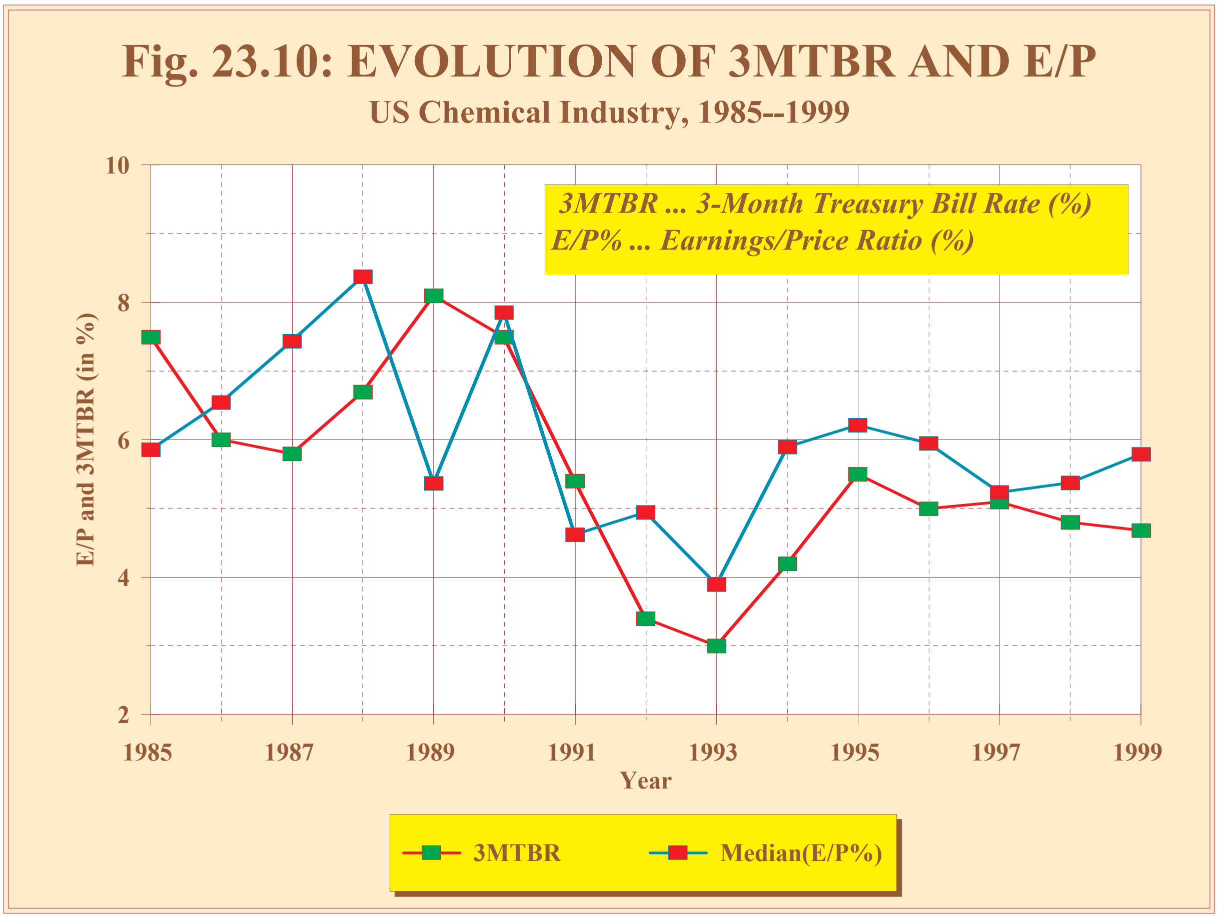
for the “better” half of the US Chemical Industry (1986-1988, 1991-1993 and 1995-1997). A noteworthy asymmetry of the distribution functions of $TLTO$ can also be observed in this graph: the probability density (the density of quantile lines) above the red median line is much smaller than below the median, which signifies, that large values of $TLTO$ are expected more often than small ones.

Determining the causes of the acceleration in $TLTO$ is not trivial. Fig. 23.8C shows, that the behavior of the $TLTO$ curves in Fig. 23.8B over the period 1986-1991 were in a close correspondence to the development of the Earnings per Share (Fig.23.8C) over the same period.



This could be explained as “better earnings ↔ faster $TLTO$.” However, the period of improving earnings (1993-1995 in Fig. 23.8C) lags the period of accelerated $TLTO$ by two years. A simple hypothesis does not work; it is necessary to look for further factors, which influence both ratios.

It is obvious from Figs. 23.8 and 23.9, that—unlike the nearly constant distributions of the $TATO$'s in Figs. 23.5 and 23.6—the distributions of the $E/P_{\%}$ ratio manifest different behavior in both their mean values and their spreads, showing a large fall of the median in 1989 below the nearly smooth drift level. To show, that this picture corresponds to economic reality, the $E/P_{\%}$'s median has been plotted in Fig. 23.10 (blue line) with the risk free rate (3 month T-Bill rate denoted $3MTBR$, red line).



The 3MTBR reflects the activity of the whole US economy, while the $E/P_{\%}$ relates only to the results of the chemical industry. These graphs thus compare the "returns" of the Chemical Industry with the risk free rate of interest. It would be unusual for both curves to coincide exactly, because they each express different things, however, what is surprising is the similarity of the nearly synchronous behavior of both processes, with the difference between the two 'rates of interest' exceeding two percent only once, in 1989.

The results presented in Figs. 23.7 through 23.10 allow a return to the problem of the economic interpretation of the ratio $E/P_{\%}$. This ratio—as well as its more popular reciprocal P/E—suffers from its already noted “hybrid” nature: the earnings E referring to the past and price P reflecting the expectations of future performance. However, a number of conclusions can be drawn from the results that have been summarized above:

1. The volatility of the P/E is substantially greater than that of the $E/P_{\%}$ (Fig. 23.7), therefore the mathematical structure of the relationship between these two variables can be more accurately depicted through the use of the latter ratio instead of the more popular P/E.
2. The $E/P_{\%}$ ratio, determined as the median of the distribution function of the main homogeneous cluster can be interpreted as the expected mean return on an investment of P in shares of the companies making up the cluster if no significant change in performance occurs. The role of this ratio with respect to the group of comparable companies tracks along with the role played by risk free rate 3MTBR as an economy wide indicator (Fig. 23.10).
3. The distribution functions of the ratio $E/P_{\%}$ allow the quantiles of $E/P_{\%}$ to be robustly estimated and attached to the probabilities shown in (Fig. 23.8) so as to give an economic interpretation even during volatile periods.
4. The application of these quantiles leads to the robust characterization of the ranges of the ratio (Fig. 23.9), so that they can be used to make judgments as to the volatility of the returns.
5. The different time aspects of the $E/P_{\%}$ ratio’s numerator and denominator do not appear to impose any difficulty in arriving at an economic interpretation of the trend of the time series. See Fig. 23.10 and compare with the path traced by the three month T-Bill rate.

It can be thus concluded, that gnostic distribution functions—while robustly suppressing the data’s uncertainties—sensitively and reliably reflect real changes in the objects, from which the data are produced.

23.2.2 Bounded Data Supports

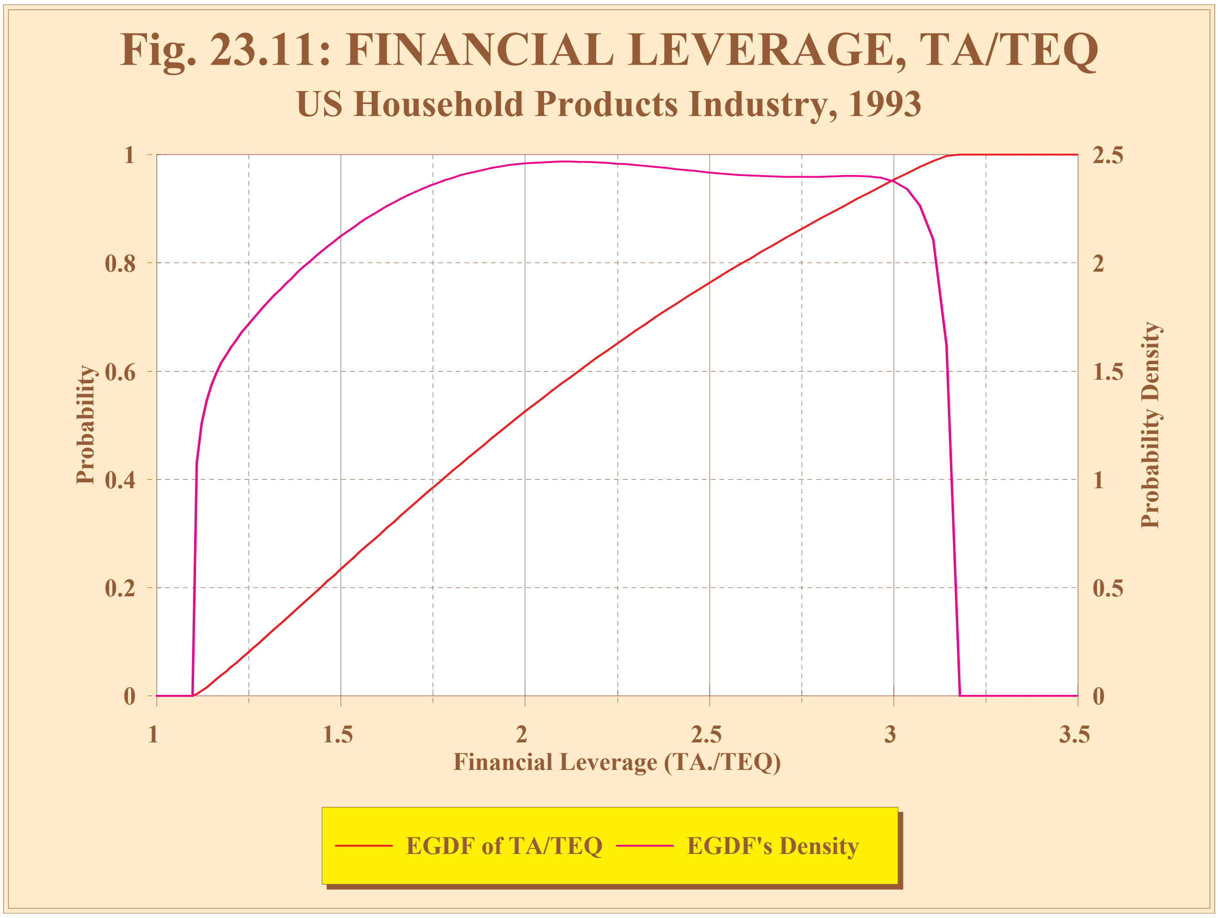
The concept of bounded data supports was illustrated in Figs. 23.5 and 23.6 using the estimation of the distribution functions EGDF of the *Total Asset Turnover* ratio for the time period 1986–1991 ($TATO$, see Tab. 22.2). It leads to following conclusions:

1. The smallest observed value in the sample of $TATO's$ was 0.308, the largest, 2.485.
2. The estimated lower bound LB of the data support was not breached until 1986, when the value of 0.299 was reached.
3. The estimated upper bound UB of the data support was never over 2.485 until it reached 11.64 in 1985.

The data supports were thus always bounded and the probability, that a chemical company would operate with a value of $TATO$ outside the interval [0.299, 11.64], was zero during this time period. An attempt to apply a normal or lognormal distribution could in no way describe this behavior.

This illustration shows, that the existence of finite bounds for data support together with a particular form for the data sample can substantially change the nature of the distribution function. The probability density can be very far from the popular bell curve of the normal distribution function. Consider two further examples:

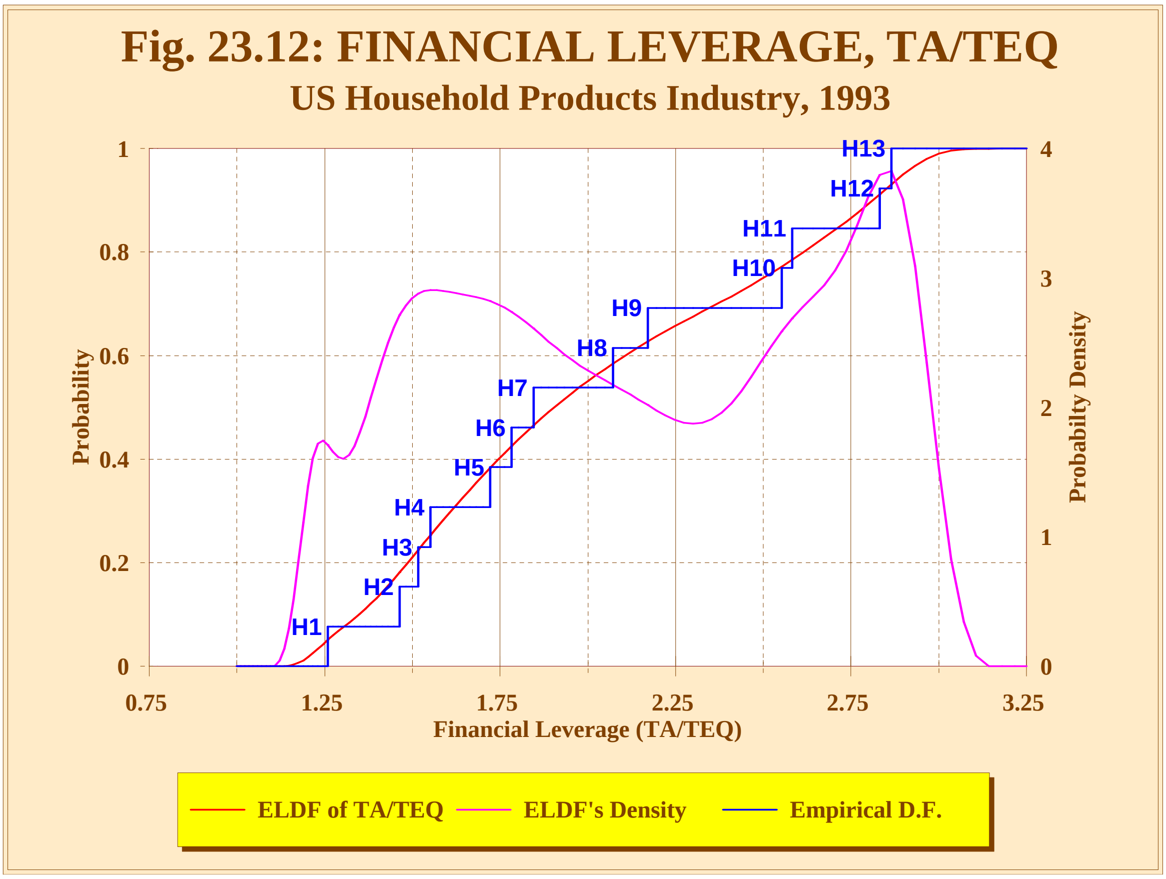
1. *Financial Leverage* defined as ratio TA/TEQ (total assets to total equity) is a measure of the ability of a company to increase its capital through the use of borrowed funds, and therefore to increase the return from its own equity by using long-term liabilities. The probability distribution EGDF and density of this ratio are shown in Fig. 23.11 for 13 companies of the US Household Product Industry in 1993.
   The data support is bounded at $LB = 1.10$ and $UB = 3.15$ and the distribution is close to uniform with probability increasing nearly linearly and with the density sharply rising from the LB and sharply falling to the UB. A more detailed insight into the data structure is offered by the ELDF (Fig. 23.12), which explains the unusual form of the global view of Fig. 23.11 as an interaction between two tendencies:
   (a) The sharp bounds of the density function infer, that there are leverage 'norms' in this industry, and that financial leverage is kept between two strict bounds, neither too low nor too high.
   (b) Financial management policies are adapted to take advantage of the companies' strengths allowing large firms to employ liabilities to an extent, which would be risky for the firms in the lower cluster in Fig. 23.12.
2. The relative value of the working capital (RWC) introduced by 22.5 is a measure of the company's liquidity, its ability to cover its short-term liabilities. The data sample of these ratios evaluated for the



same Houshold Product Industry sector for the same period is so inhomogeneous, that a distribution function of the EGDF-type does not exist. Its ELDF distribution is shown in Fig. 23.13 (red line) together with the EDF (Empirical Distribution Function, 15.1, the blue step function of the type used in statistics).

There again are finite bounds for the data support ($LB = -0.0465$ and $UB = 0.475$). The negative value of LB says, that a negative liquidity ratio can be encountered even in such a well-established industry, but its probability of occurrence is only about 0.115. Once more, the form of both continuous functions in this figure has no resemblance to the normal curves. The probability has two sharply rising sections (close to the LB and the UB) and the density function therefore has a U-shape.

The bounds of data support can have a strong impact on the shapes of the distribution functions and their densities, and they should not be neglected



in the analysis. In practice, densities can range over a broad scale of forms from concave through uniform to convex. It is obvious, that neither information on the data support's bounds nor on the form of probability distributions and densities can be obtained from the point estimates of the usually employed statistical methodology.

23.2.3 Example of a Hard Data Bound

All data bounds illustrated in Figs. 23.11–23.13 were estimated from data. In the terms defined in Chapter 15 they can be characterized as **soft**. The existence of such bounds should be recognized by experienced economists, but their values are a priori unknown. Soft bounds are not fixed, so that eg the distributions of the same kind of financial ratios for different groups of firms can differ. Hard bounds of data support are unlike the latter and are determined 'once forever' by the nature of the quantities to be investigated. They are fixed and known even without having to look at the real data. As such, they represent important a priori information, which has to be

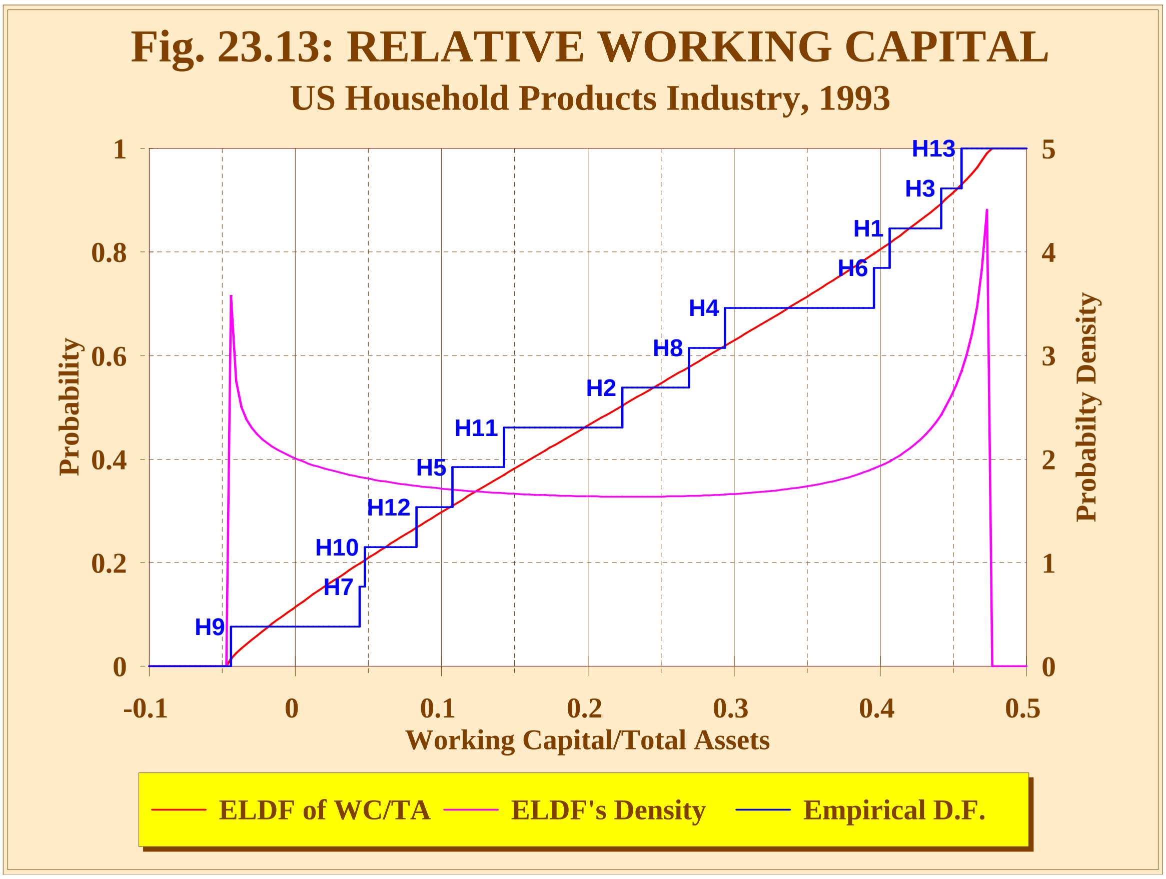


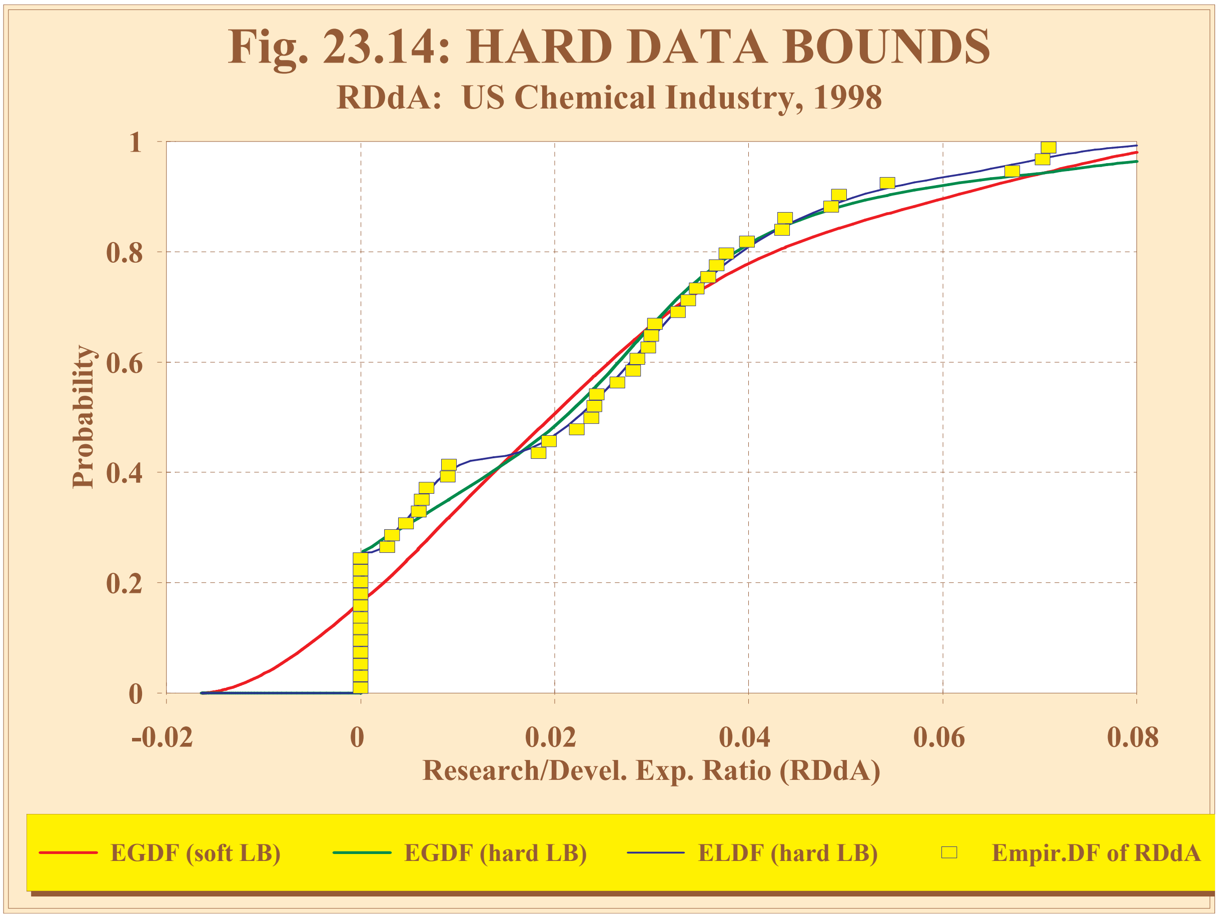
Fig. 23.13: RELATIVE WORKING CAPITAL

US Household Products Industry, 1993

used in constructing algorithms for these distribution functions.

In Fig. 23.14, the distribution function of the ratio RD/TA (Research and Development Expenses divided by the Total Assets) is shown as an example of a hard data bound.

Twelve out of forty seven companies in the US Chemical Industry in 1998 reported zero expenses of this type. Thus a zero value for this ratio is probable. On the other hand, it is obvious, that negative expenses do not exist, which means, that the data support for RD/TA is a semi-closed interval. The red line in Fig. 23.14 shows the EGDF calculated for the data available, while the existence of the hard low bound is ignored. Non-zero probabilities are unrealistically attached to negative RD/TA's by this distribution function. Moreover, this EGDF does not satisfactorily explain the data, the empirical distribution of which is drawn in Fig. 23.14 with square boxes. The remedy is simple: the subsample of zero RD/TA is qualitatively different from that of non-zero values, the former has a simple discrete distribution function, while the latter can be characterized by a



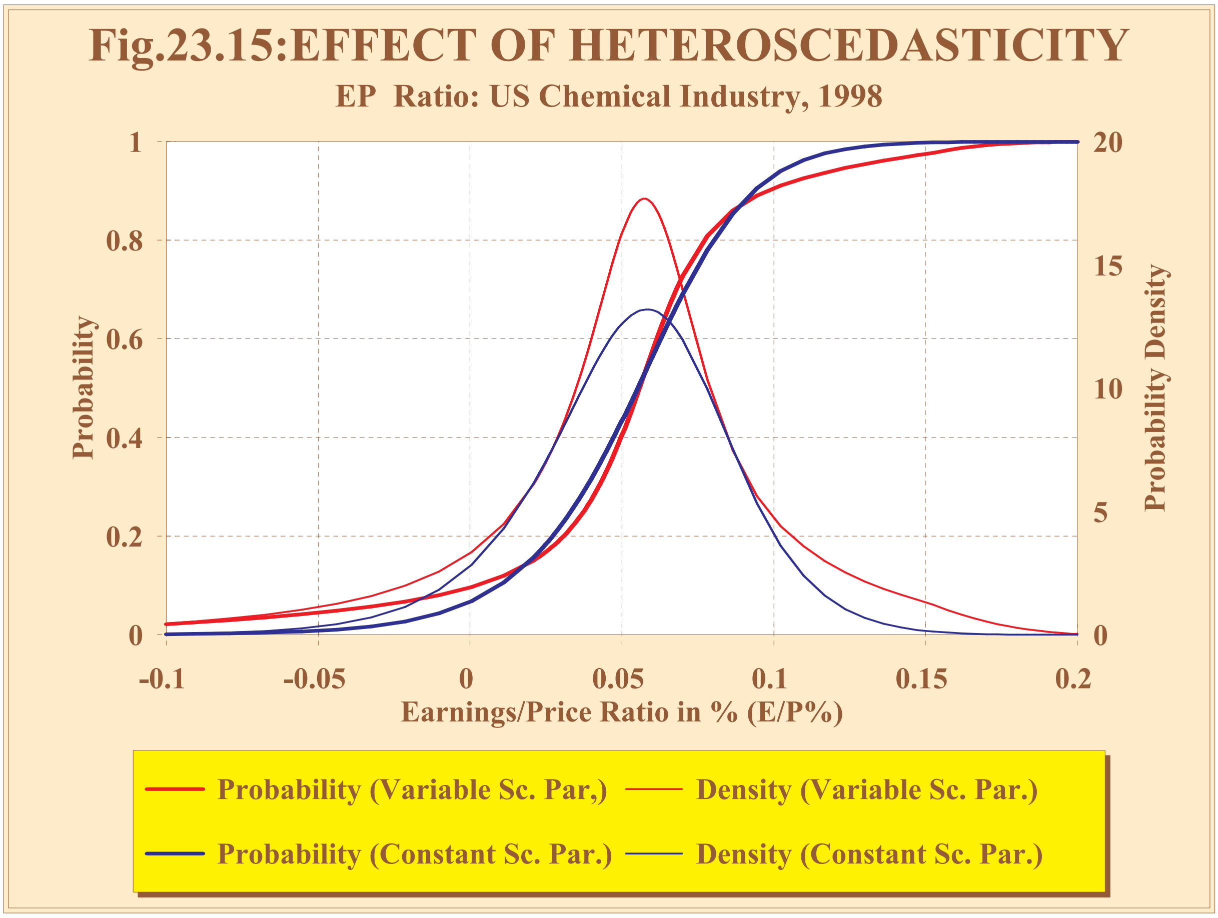
continuous EGDF designed over an open data support. Both distributions are combined and shown in Fig. 23.14 by the green line. If a closer explanation of data is required, the ELDF can be obtained by the analogous combination shown by the blue line.

A lesson is learned here: if one has a priori information (such as a particular type of probability distribution function) it is not necessary to estimate a gnostic distribution function. However, if such sure information as the existence of hard data bounds is available, it is not reasonable to ignore it.

23.2.4 Heteroscedastic samples

The complexity of economic processes makes it unrealistic to accept simplifying assumptions such as stationarity, ergodicity, normality, etc., which are so welcome in theoretical studies. Specifically, there is no certainty, that the spread of economic data is constant (or—using statistical language—that the data are homoscedastic). Experience shows rather the opposite. As already discussed in Chapter 16, data spread is characterized in gnostics by the scale parameter S. Its value can be estimated by, among others, the technique described in 16.2.2 (the local scale parameter). It is useful to get a feel for the real effect of having variability in the spread of local data.

The $E/P_{\%}$ (earnings price) ratio can serve as an example as illustrated by Fig. 23.15.



The bold blue line is the EGDF optimized under the assumption of a constant scale parameter and the other blue line is its density. The red lines are distributions using variable scale parameters obtained by two steps:

1. The dependence of the local scale parameter $S_L(Z_0)$ on the location (Z) in the sample was estimated by solutions of 16.17.

2. To ensure optimality of the fit, the best weight S_w was found by using variable scale parameters obtained as $S_w * S_L(Z_0)$.

As Fig. 23.14 shows, the effect of variability in the scale parameter is not negligible. The density curve is sharper than that of the constant S, documenting the steeper rise of probability in the central part of the data support, while the probability of the peripheral data values is larger.

Important questions arise: are the red curves in Fig. 23.14 in a way 'better' than the blue ones? And if so, then more generally: is it always better to consider the data 'heteroscedastic' and to use variable scale parameters? To keep the principle 'Let data speak for themselves' and to extend the experience, a comparison of 16 financial ratios has been made and the quality of data fit was measured by several indicators, two of which are the most important, the relative information quality ($InfQ$) and the mean absolute probability error ($MAPE$), each determined in the following way:

1. Denote points on the Weighted Empirical Distribution Function given by the data E_n and probabilities from the EGDF for the same data D_n by $P(D_n)$. The multiplicative evaluation of the quality of the approximation of the WEDF by the EGDF at the data point D_n is determined by the ratio $r_n = E_n/P(D_n)$. The additive characterization of the fitting error is $d_n = E_n - P(D_n)$.

2. Calculate the estimating irrelevances $h_n = (q_n - 1/q_n)/(q_n + 1/q_n)$ for $q_n = r_n^{2/S_n}$ (see 9.11) using the (constant or local) scale parameter S_n. The probability of r_n is then $(1 - h_n)/2$ (10.60) and its informational evaluation is $I_n = -p_n * \ln(n) - (1 - p_n) * \ln(1 - p_n)$ (10.56), the maximum of which is $\sqrt(1/2)$. The relative information in r_n is therefore $I_n/\sqrt(1/2)$. The relative information quality ($InfQ$) of the fit is the arithmetic mean of these quantities determined for all the sample's data.

3. The mean absolute probability error $MAPE$ is simply $\frac{\sum_{n=1}^{N} d_n}{N}$, where N is the number of data.

Using these measures one can get a positive answer to the first question: the application of a variable scale parameter to the ratio RD/TA is better than using a constant parameter in the sense, that a higher information quality $InfQ$ and a smaller probability fitting error $MAPE$ are obtained. The answer to the question of the general applicability of this concept is more sophisticated: it depends on the nature of the data to be analyzed. So eg from the 16 types of financial ratios of the US Chemical Industry in 1998, which were considered (see table 23.4), only in 8 cases did the usage of a variable S lead to an improvement in the information quality of the EGDF, while in 5 cases the quality was worse and in 3 cases practically the same. The additive error $MAPE$ decreased in 5 cases, increased in 5 cases and in 6 cases remained practically unchanged. However, as shown in table 23.1, some of the improvement noted were not negligible.

The conclusion is obvious: some data are homoscedastic, while others are heteroscedastic and the difference can be substantial. Which model should be chosen in any particular case? Let the data decide: try both methods and retain the better result.

Scale parameter	**Ratios**					
	RD/TA	TL/TA	DA/TA	EPS	PS	$E/P_{\%}$
	Relative Information Quality ($InfQ$)					
Fixed	0.946	0.895	0.929	0.940	0.968	0.910
Variable	0.965	0.982	0.964	0.979	0.983	0.963
	Mean Absolute Probability Error ($MAPE$)					
Fixed	0.043	0.037	0.037	0.031	0.025	0.037
Variable	0.031	0.023	0.034	0.021	0.024	0.028

Tab. 23.1: Effects of variability in the scale parameter (‘heteroscedasticity’)

The four examples in Fig. 23.16 show the extent of the changes in the variable S.

There is the typical U-shape of the curve $S(Z_0)$ with the minimum over the points corresponding to the center of the data support. The curve can be nearly symmetrical (such as for CA/TA) or asymmetrical (like EAT/TA, the right end of which is much higher than the left one). The relative ‘depth’ of the curves is different for different ratios: changes in the EGDF’s curvature can be more or less pronounced depending on the nature of the data.

23.2.5 Marginal Analysis

The term *Marginal Analysis* is interpreted in gnostics as the application of local distribution functions ELDF to a uni-variate data sample for the pur-

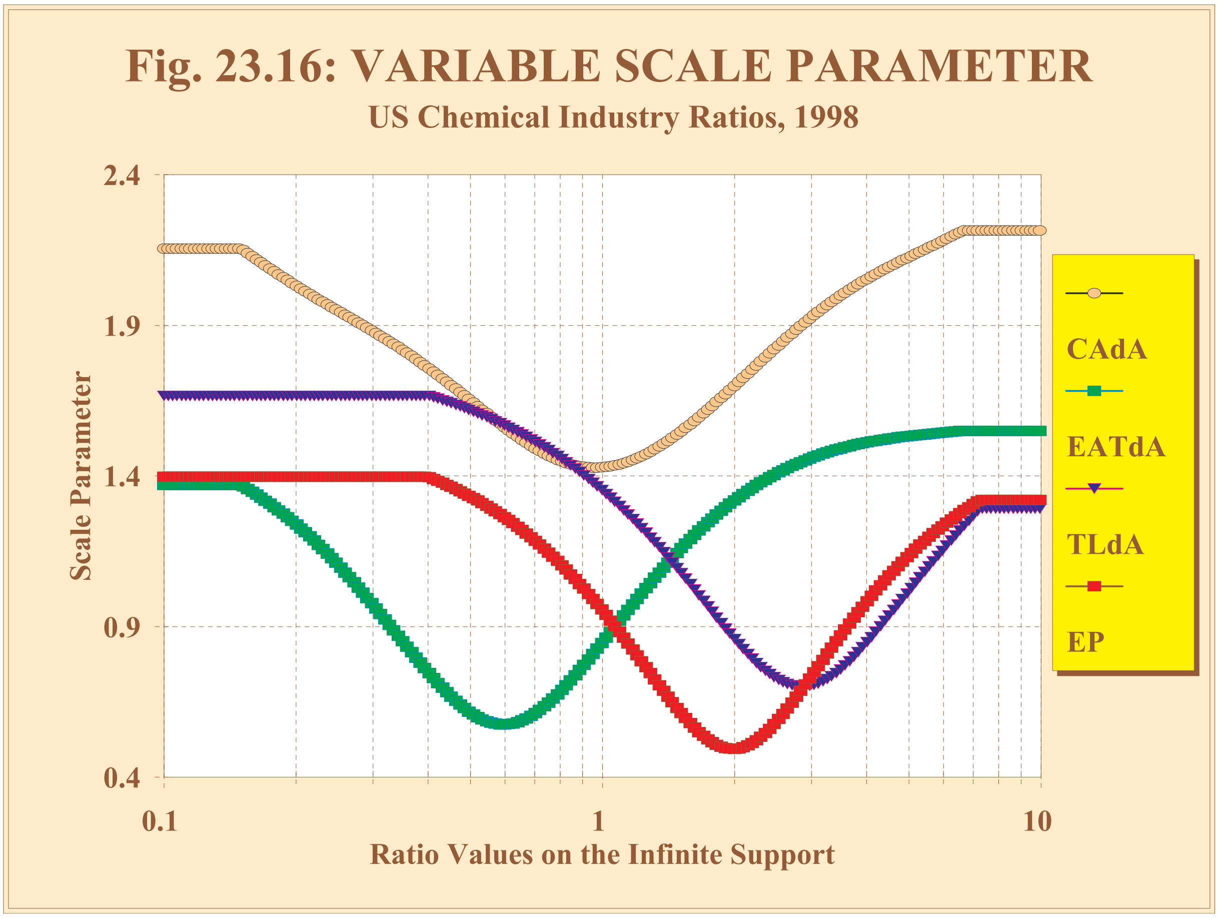


Fig. 23.16: VARIABLE SCALE PARAMETER

pose of decomposing it into individual subsamples/clusters. Such a decomposition helps to reveal the inner structure of the data sample. Marginal analyzes can be used to examine both cross-sections as well as time series of cross-sections to yield information on the development of the samples' inner structure.

A practical example of marginal analysis was given in Fig. 23.13, where a sample of RWC ratios appears to consist of one cluster of companies with high RWC concentrated close to the bound UB and another one of companies, which operated with low RWC found near the lower bound LB. The point separating these clusters is near the distribution's median ($RWC = 0.23$).

The labels in Figs. 23.12 and 23.13 locate the companies within the individual clusters. A comparison between both decompositions offers interesting conclusions:

1. Both companies H12 and H13 are members of the **high** cluster of

financial leverage TA/TEQs and of the **low** cluster of RWCs. Both these facts infer, that these companies are stable enough to operate with a high level of liabilities and low liquidity. This can be explained by taking into account other data such as the size factor: Sales of H12 and h11 were respectively 56.3% and 13.2% of the sales of the whole industry; they thus represented a market concentration of 69.5%. On the other hand, their total equity was only 57.7% of the total of the industry's book value, while total debt was 77.3% of the collective figures. This situation as reflected by the graphs deserves a comment in the style of George Orwell[7] "All companies are equal, but some are more equal than others!"

2. A company's size is not always the decisive factor: the relatively small H10 with only 1.89% of the industry's sales also had a low RWC and high TA/TEQ like the giant H12. A more detailed analysis would be necessary to decide if these 'abnormalities' signal something positive or negative. A first guess might be, that it is 'operating on the edge' and is about to sink. A historical analysis of the data would be required to provide more clues, but the condition could also result from sound management and astute financial policies.
3. There were also other ratio combinations in both of the extremal clusters, for example:
   (a) High RWC and high TA/TEQ (H13).
   (b) High RWC and low TA/TEQ (H3,H1, H4)[8].

A financial manager seeking to learn, what level of financial leverage or liquidity ratio is 'normal', must recognize, that the 'norms' are different for different classes of firms. The first question to be answered must be that of comparability: does my company compare with H12 or is it more like the rest of the industry?

Another example of marginal analysis is seen in the development of dividends per share in the US Chemical Industry over the time period 1986–1998. Probability densities of the ELDF-type are shown in Fig. 23.17.

An examination of the time sequence of the curves gives the impression of an animated cartoon: the curves change their forms in a continuous movement, while preserving their basic bi-modal features. This permits the

[7] "All animals are equal, but pigs are more equal."(Orwell: The Animal Farm, New York, Harcourt, Brace & Co., 1946)

[8] Both of these make sense, and are probably due to managerial preference: in the first case, keeping some cash or securities sourced from long term liabilities, in the second, being rather risk averse and ready to respond quickly to changing market conditions. Again, more information is needed, but the ratio analysis suggests, where to look.

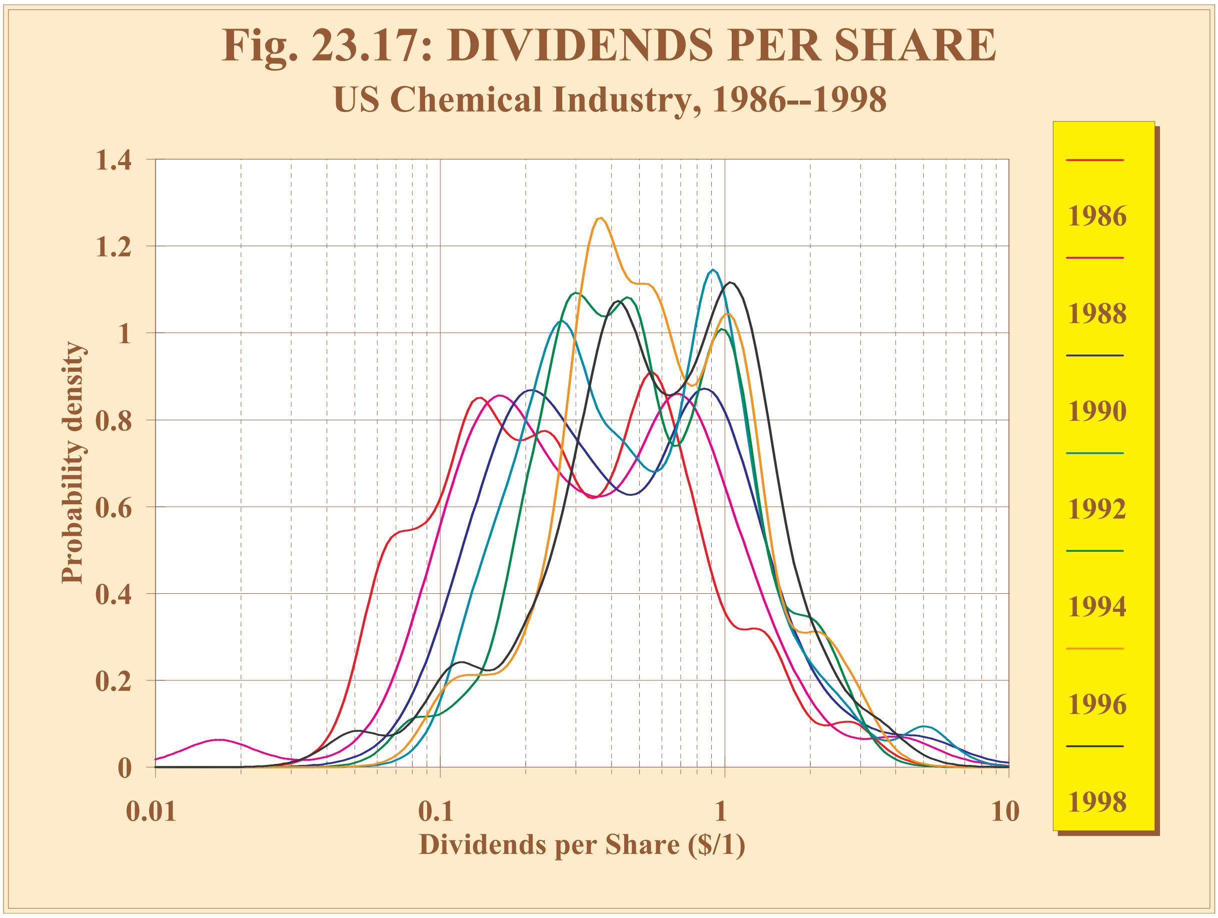


Fig. 23.17: DIVIDENDS PER SHARE

US Chemical Industry, 1986--1998

companies to be split into two main clusters (L and H) representing those with dividends, that are lower/higher than the intra-peak minimum[9]. The trend in these maxima and of the valley separating them can be followed in Table 23.2:

The most frequent dividend levels in each group rose monotonically over the time period 1986–1998. Both prominent peaks in Fig. 23.17 move from left to right with the passage of time. To show this movement in more detail, the points of maxima are marked by yellow ellipses in Fig. 23.18 along with arrows to emphasize the development.

The height of the maxima again follows the changing spread (maximum up ⇔ spread down).

These results suggest the following conclusions:

[9] Generally speaking, young growing companies pay little or no dividends. Firms, which have established themselves, and which wish to signal their ‘arrival’, but that also need to use internal funding for growth, balance the payment of dividends with their need for retentions, while mature enterprises with solid access to external funding sources are able to payout a higher proportion of their earnings.

Year	Typical Dividends		
	Max. of L	Sep. point	Max. H
1986	0.14	0.34	0.56
1988	0.16	0.36	0.69
1990	0.20	0.47	0.85
1992	0.27	0.56	0.91
1994	0.30	0.67	0.98
1996	0.37	0.77	1.05
1998	0.42	0.64	1.05

Tab. 23.2 Typical Density Points for Dividends Paid

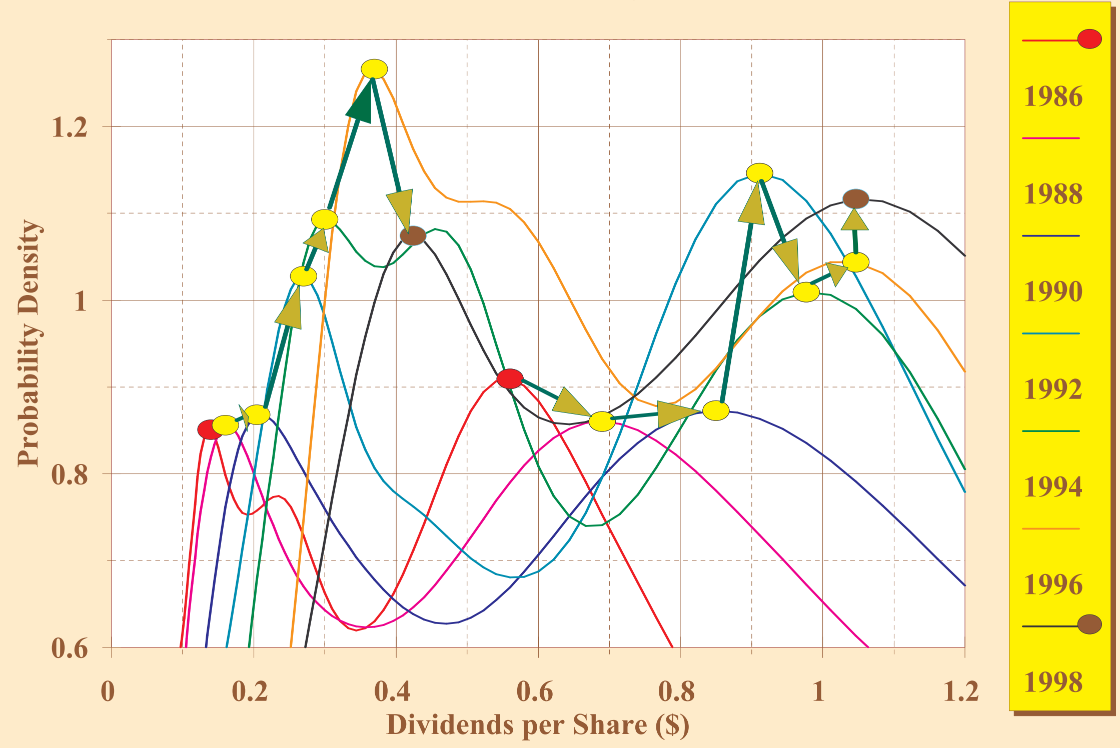


Fig. 23.18: DIVIDENDS PER SHARE

US Chemical Industry, 1986--1998

1. The dividend policy of the industry is similar to that of other industrial groupings and falls into three categories:
   (a) No dividends.
   (b) A compromise between dividends and retentions (the lower cluster in Fig. 23.17).

(c) High payout (the upper cluster in Fig. 23.17).

2. The membership of firms in a particular group tended to persist over a long time period.
3. Over time, the changes in the observed structure were continuous and consisted only of a systematic increase in payout.

As a comparison, the development of the dividend yield (dividend paid divided by the share price) was analyzed in a similar way. All dividend paying firms fell into only one cluster with the maxima of probability density drifting in 1986–1998 over the interval from 0.018 to 0.032[10].

23.2.6 Homogeneity of Data

The idea of homogeneity of a data sample is closely connected with Axiom 2, the data composition axiom, which states, that a single (equivalent) datum for a data sample is obtained by averaging the gnostic weights and the irrelevances of the sample's data. However, gnostic data weights are cosines and irrelevances are sines and their averages must be normalized by the sample modulus (14.4) to obtain the equivalent (14.3). This normalization is the necessary condition for obtaining a true equivalent, but it is not sufficient. Although the equivalent (normalized) weight can always be calculated (for all data samples), it will not always be a true equivalent. This can be demonstrated through an analysis of the equivalent's relationship to the probability density of the distribution function EGDF (15.29), which is obtained by a normalized average of irrelevances. Each distribution function must be—by definition—non-decreasing, ie its density must be non-negative everywhere. As mentioned in Chapter 14, this condition holds only when the data of the sample are not spread too widely. In other words, an EGDF does not exist for all data samples. With an increasing spread of data, the EGDF's density inescapably falls to local negative values. It is a continuous and unlimitedly differentiable function, therefore its local decline can take place only on an interval between two density maxima. This implies, that *(negative density)* $\Longleftrightarrow$ *(non-existent EGDF)* and *(negative density)* $\Leftarrow$ *(at least two density's maxima)* hold. The converse of this statement is not true, because there can be two density maxima with a much smaller data spread, when a positive segment of the EGDF passes through a maximum and starts decreasing. At such a point, the EGDF's density passes through a local maximum, which does not necessarily co-

[10] During this same time period, the (monthly) average dividend yield on the S&P Index ranged between 0.015 and 0.038.

incide with the global maximum. At the "critical" point, which discriminates between one-maximum and two-maxima density forms, the density function has an inflection point. Such points are mathematically defined and—if they exist—can be calculated. Data samples with a one-maximum (unimodal) density function are taken in gnostics as *homogeneous*. The requirement for a unimodal density is of course much stronger than the condition for the density's nonnegativity. This means, that EGDF can be applied to tests of homogeneity based on a strict condition of unimodality.

This idea of homogeneity has a vivid interpretation. If it is assumed, that the distribution of uncertainties about a true value is unimodal, then the appearance of a bimodal distribution of observed uncertain data shows the existence of two true values. Multimodal distributions thus belong to mixtures from two or more data sources of possibly different nature. The analyst's task in such cases is similar to that of a dog, who has to choose between pursuing two or even several rabbits.

The important characteristics of a data sample only make good sense and can be properly estimated, when the sample is homogeneous. Homogeneous samples are therefore preferred for most analyzes. Unimodality can be tested easily to establish homogeneity or the lack thereof in a sample by calculating and checking the EGDF and its density. An inhomogeneous data sample consists of at least one homogeneous cluster (subsample) plus one or more other clusters of data. The process of isolating the homogeneous "kernel" (cluster) of a data sample of a general type includes following steps:

1. Calculation of the unique EGDF of the given sample by estimating its scale parameter of the global type ($S_{G,MF}$) simultaneously with the bounds (LB and UB) of the data support by means of optimizing the fit of WEDF (Weighted Empirical Distribution Function) by EGDF.
2. Calculation of the EGDF's density defined over the infinite data support.
3. Establishing the number $Nmax$ of this density's maxima.
4. If $Nmax = 1$ then STOP (the sample is homogeneous), else continue.
5. Find the largest of the maxima (determine the main cluster denoted M).
6. Find the density's minima on both sides of the main cluster.
7. Take the data from the interval between the minima to contain the members of the M-cluster, the data located under the lower minimum as the L-cluster and that found over the higher minimum for the U-

cluster.

8. Iterate by applying operation 1 to the M-cluster.

This process is repeated until the homogeneity test is affirmative. The sample's data are thusly classified as belonging to one of three categories: Main (M), Lower (L) and Upper (U). It is useful to have a look at a practical example of such a classification applied to financial statement analysis.

23.2.7 An Example of Homogeneity

As a demonstration of a test for homogeneity, 16 financial ratios and other financial characteristics of 47 companies of the US Chemical Industry have been calculated and analyzed for 1988. The companies are listed in Tab. 23.3 by their ticker symbols and the ratios taken for analysis are tagged by a number, which is explained in Tab. 23.4. The symbols used for the ratios are in the form of N/D (read 'N divided by D'), where N is the numerator and D the denominator of the ratio. The N and D components are identified in the same manner as in Tab. 22.1. Symbols L, M and U represent membership respectively in the Lower, Main and Upper clusters.

The companies shown in Tab. 23.2 all have at least one ratio out of the main cluster.

Before interpreting the results of the clustering, it should be made clear, that being assigned to the cluster L does not always mean, that the firm is considered 'to be worse than the others.' The economic meaning of the placement depends on the interpretation of the parameter which is considered. The same situation applies to the cluster U. To be 'low' in current liabilities is not always bad and to work with a 'high' financial leverage is not always good. In many cases ratios do have an optimum interval of values, but it does not always coincide with 'what the majority does.' (An everyday example: is it a good idea to spend as large a portion of one's "free" time watching TV as the majority of society is reported to do? This introduces the problem of how to interpret membership in the main cluster (M), which is closely connected with a notion of **normality**. To examine this problem, it is necessary to recall the notions of the data sample's bounds (LSB and USB) introduced in 15.3.7.

In gnostics, the word "normal" bears no relation to its statistical mean-

Comp.	**Number of the Ratio in Tab. 23.4**															
Tick	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T1	M	U	M	M	M	M	M	M	M	M	U	M	M	M	M	M
T2	M	M	M	M	M	M	M	M	M	U	U	M	M	M	M	M
T5	M	M	M	M	M	M	M	M	M	U	M	M	M	M	M	M
T7	M	M	M	M	M	M	M	M	U	M	U	M	M	M	M	U
T8	M	M	M	M	M	M	M	M	M	U	M	M	M	M	M	M
T10	M	M	M	M	U	M	M	M	M	U	U	U	M	M	M	M
T11	M	U	M	M	M	M	M	M	M	M	M	M	M	M	M	M
T13	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	U
T17	M	M	M	M	M	M	M	M	U	M	M	M	M	M	M	U
T19	M	M	M	M	M	M	M	M	M	M	U	M	M	M	M	M
T20	M	M	M	M	M	M	M	M	U	M	L	M	M	M	M	L
T22	M	M	M	M	M	M	M	M	M	M	M	M	U	M	M	M
T25	M	M	M	M	M	M	M	M	M	U	M	M	M	M	M	M
T28	M	M	M	M	M	M	M	M	M	U	U	M	M	M	M	M
T31	U	M	M	M	M	U	M	M	M	M	U	M	U	M	M	M
T33	M	M	M	M	M	M	M	M	M	U	M	M	M	M	M	M
T34	M	M	M	M	M	M	M	M	M	U	M	M	M	M	M	M
T35	M	M	M	M	M	M	M	M	M	M	M	M	U	M	M	M
T37	M	M	M	M	M	M	M	M	M	U	U	M	M	M	M	M
T39	M	M	M	M	M	M	M	M	M	U	U	M	M	M	M	M
T40	U	M	M	M	M	M	M	M	M	M	M	M	U	M	M	M
T42	M	M	M	M	M	M	M	M	M	M	M	M	U	M	M	M
T43	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	L
T44	M	M	M	M	M	M	M	M	M	U	U	M	M	M	M	M
T46	U	M	U	M	M	M	M	U	M	M	M	M	U	M	M	M
T57	M	M	M	M	M	M	M	M	M	M	M	M	M	U	M	M

Tab. 23.3 Clustering by homogeneity of data (US Chemical Industry, 1998). Ticks: see Tab. 23.7.

Identification of Financial Measures			**Cluster size**		
No.	**Name**	**Symbol**	**L**	**M**	**U**
1	Rel.Current Assets	CA/TA	0	44	3
2	Rel.Current Liabilities	CL/TA	0	45	2
3	Total Assets Turnover	$TATO$	0	46	1
4	Rel.Deprec.& Amort.	DA/TA	0	47	0
5	Rel.Capital Expenses	CX/TA	0	47	0
6	Rel.Retained Earnings	RE/TA	0	46	1
7	Rel.Research & Devel.	RD/TA	0	47	0
8	Return on Assets	ROA	0	46	1
9	1/(Financial Leverage)	TL/TA	0	44	3
10	Accum.Depr.& Amort.	ADA/TA	0	36	11
11	Earnings per Share	EPS	1	36	10
12	Stock Price	PS	0	46	1
13	Rel.Working Capital	RWC	0	41	6
14	Dividend Yield	DIV/PS	0	46	1
15	Total Return	$TOTR$	0	47	0
16	Earnings/Price Ratio	E/P	2	42	3

Tab. 23.4: Numbers, names and occurrence of financial ratios in Tab. 23.2

ing, which addresses Gaussian distribution functions. The gnostic definition of normality states that:

Definition: Normal data D of a homogeneous sample are those, which satisfy the relation $LSB \leq D \leq USB$, where LSB and USB are the bounds of the sample (membership bounds).

In other words, 'to be a member of the (homogeneous) main cluster M' means 'to have a normal value.' This is close to the sense, which the word takes in common language (usual, ordinary, what is expected, conforming to an accepted standard). And again, 'being normal' does not automatically mean 'to be good' or 'to be bad.' Such an evaluation must be based on other information.

The gnostic interpretation thus takes as normal all the members of the main, homogeneous cluster of the data sample. A preference for such a concept lies in the following:

Universality: A main cluster can be found in all data samples. In a homogeneous sample the main cluster contains the whole sample, all data are normal. For inhomogeneous samples the main cluster can always be found by the application of gnostic algorithms. Normality is a characteristic of all types of distribution functions, not just another name for the Gaussian type.

Objectivity: As shown in Chapter 15, parameters LSB and USB are determined uniquely from the data without any intervention or a subjective choice (such as the significance level of statistical tests) made by the analyst.

Robustness: Parameters LSB and USB are estimated by the EGDF, which is robust, and are therefore also robust.

Predictive power: Parameters LSB and USB are solutions to the membership problem; they delimit an interval of values, in which a datum must lie to be a member of the homogeneous cluster (the main cluster), ie to be normal. This interval relates not only to data already existing in the sample, but also to possible future data, which originate from the same source.

There also is a danger of misinterpreting the outcomes of this procedure (which also applies to any other methods) of marginal analysis: the limitations of univariate information. Any parameter of a multidimensional object can be classified as being ‘abnormal’ from the univariate point of view, but this abnormality can be compensated by the abnormality of another parameter and it can thus disappear, when a multidimensional approach is used. Another flaw in this classification scheme is, that it is very rough: to know, that an object is normal, does not imply optimality in its position, it can be located near the center of the main cluster as well as close to its low or high boundary value. Finer measures must be applied to evaluate the object’s position among its normal, comparable neighbors.

While this step of marginal analysis of ratios has these and other limitations, it can still yield some useful information.

As can be seen in Tab. 23.4, there were only four ratios, which appeared to be homogeneous (DA/TA, CX/TA, RD/TA and TR). The comportment of all the tested enterprises was comparable from these points of view, with no apparent ‘excesses.’ The ‘silver medal’ was shared by $TATO$, RE/TA, ROA, PS and DIV/PS with only one excessive case. The ‘bronze medal’ belonged to CL/TA with two excessive cases. The Earnings per Share (EPS) ratio was at the other end of the ‘ladder’ with only 36 members in the M-cluster, 1 substandard (L) and 10 in the high cluster. The strong inhomogeneity of EPS can be explained by the fact, that this ratio has no directly interpretable economic meaning for a cross-section analysis and comparison. Because the number of shares in each company is different, the value of EPS is not useful in making relative comparisons of real economic performance between firms. This is also true for an analysis of a single firm, when prior earnings are not restated after there has been a significant change in shares outstanding.

A rough indication of the relative performance of individual companies can be drawn from Tab. 23.3. Only 21 companies (44.7%) had all of their 16 ratios in the M-cluster. These ‘completely normal’ companies are omitted from Tab. 23.3. The membership pattern of twelve others put all but one of their ratios in M (11 in U and 1 in L). Thirty three companies out of 47 can be thus considered as completely or nearly ‘standard’ for their industry and thus comparable to each other.

23.2.8 “Normal” Values for Financial Ratios

The important problem of establishing the ‘right’ (‘sound’, ’standard’, ‘recommended’, ‘desirable’, ‘proper’) values for financial ratios and other financial parameters was discussed in Chapter 22. The use of the arithmetical mean of industry ratios was rejected because of the bad robustness of this statistic. Some analysts use the sample median as a more robust statistic, but it ignores the form of the data distribution. A British company ([34]) periodically publishes a national compilation of three statistics for 16 principal financial ratios for 206 industries: the median and the lower and upper quartiles. A quartile is the median of half of an ordered sample, the three statistics are therefore denoted as quantiles Q1, Q2 and Q3. A knowledge of all three quantiles provides certain information about the distribution's symmetry and allows four classes of values of a ratio R to be distinguished: $R \leq Q1$, $Q1 < R \leq Q2$, $Q2 < R \leq Q3$ and $Q3 < R$. These reviews could provide more useful information if they were completed by the minimum (Min) and the maximum (Max) values for the ratios observed over the time period considered. In such a case two other categories could be introduced ($R < Min$ and $Max < R$) to make the bounds of the categories 'the lower quarter' and 'the upper quarter' more precise. A financial manager seeing, that 'his' ratio appears to be of an 'not yet observed' value, would have thus a motivation for contemplation. Of course, if he would use the bounds LSB and USB, a sharper signal were be given: 'your ratio is abnormal' in the sense of membership[11]

To compare the two kinds of bounds, Tab. 23.5 was prepared for the same set from the US Chemical Industry for 1998 used above along with Tab. 23.6. The sample quantiles presented in this table are order statistics obtained directly from samples of **all** the ratios. Without the minimal and maximal values, they allow a firm's ratio to be classified only with respect to one of the four broad intervals. This does not lead to any desired 'recommended' or 'sound' values of ratios, but only states, that 'your ratio is less than that of 50%, but it is exceeded by 25% of your competitors', which says nothing about the 'normality' of the ratio.

A more complete and reliable set of information can be obtained through the following steps:

1. Determine the main clusters ('M' as in Tab. 23.3) for all the ratios by isolating the comparable firms in the data sample.
2. Determine the minimal and the maximal values of the ratios belonging to the main cluster.
3. Calculate the EGDFs of the main cluster.
4. Estimate the membership bounds (the lower and upper bounds of the homogeneous samples, LSB and USB) of the main clusters using the

[11] The sample's bounds are mostly wider than the interval of observed data [min, max].

EGDFs.

To complete the comparison with the traditional analysis presented above, the robust gnostic quantiles of the EGDF were computed for probabilities of 0.25, 0.5 and 0.75 along with the LSB and USB, which show the value of the ratios separating the clusters L,M and U in Tab. 23.3. These results are shown below in Table 23.6.

Some of the ratios are positive by definition. However, numerical estimation of their LSB can result in a slightly negative value, because the procedure 'knows' only the data and extrapolates their distribution. In such cases zero values of LSB were assigned in Tab. 23.6[12].

Several observations can be directly drawn from the last two tables:

- The sample estimates of quartiles are theoretically robust with respect to outliers. But a comparison of the tables does not confirm this expectation: there are large differences between the quartiles determined by the two methods, eg $Q1(EPS) = 0.960$ vice 0.754, $Q3(ADA/TA) = 0.475$ vice 0.355, $Q3(EPS) = 2.040$ vice 1.613 etc.

[12] A better alternative can be the application of the 'hard bound' approach already discussed above.

Ratio	Min	$Q1$	$Q2$	$Q3$	Max
CA/TA	0.213	0.319	0.374	0.472	0.712
CL/TA	0.099	0.191	0.236	0.287	0.448
$TATO$	0.368	0.890	1.017	1.306	2.035
DA/TA	0.015	0.036	0.049	0.059	0.084
CX/TA	0.000	0.033	0.049	0.069	0.114
RE/TA	-0.147	0.129	0.273	0.471	1.052
RD/TA	0.000	0.003	0.024	0.036	0.071
EAT/TA	-0.071	0.026	0.056	0.084	0.309
TL/TA	0.151	0.532	0.638	0.702	0.966
ADA/TA	0.053	0.274	0.320	0.475	0.703
EPS	-1.950	0.960	1.380	2.040	5.830
PS	5.625	19.844	28.625	41.250	90.94
RWC	-0.062	0.064	0.113	0.221	0.513
DIV/PS	0.000	0.0113	0.0203	0.0323	0.0703
$TOTR$	-0.805	-0.254	-0.095	0.026	0.320
EP	-0.124	0.042	0.056	0.069	0.151

Tab. 23.5: Ranges and sample quantiles of financial ratios, US Chemical Industry, 1998

Ratio	LSB	$Q1$	$Q2$	$Q3$	USB
CA/TA	0.115	0.299	0.372	0.449	0.749
CL/TA	0.044	0.185	0.231	0.277	0.456
$TATO$	0.000	0.841	1.047	1.254	2.123
DA/TA	0.000	0.037	0.048	0.058	0.084
CX/TA	0.000	0.030	0.049	0.068	0.138
RE/TA	-0.398	0.129	0.271	0.422	1.082
RD/TA	0.000	0.004	0.020	0.039	0.087
EAT/TA	-0.701	0.030	0.058	0.085	0.192
TL/TA	0.142	0.501	0.615	0.715	1.059
ADA/TA	0.053	0.230	0.296	0.355	0.569
EPS	-7.939	0.754	1.216	1.613	2.744
PS	0.000	19.52	28.56	38.34	85.15
RWC	-0.208	0.056	0.111	0.166	0.406
DIV/PS	0.000	0.012	0.021	0.031	0.045
$TOTR$	-4.066	-0.287	-0.114	0.036	0.550
EP	-0.033	0.037	0.055	0.071	0.107

Tab. 23.6: Quantiles and membership bounds of measures estimated by distribution functions EGDF for the homogeneous main cluster M

The figures in Tab. 23.6 were obtained after the outliers had been eliminated by creating a homogeneous sample. The sample estimates of quartiles in Tab. 23.5 are therefore unreliable due to distortion by outliers.

- The quartiles do not provide useful bounds for the classification of acceptable values for the ratios, because 25% of them are lower than $Q1$ and 25% exceed $Q3$. There does not appear to be a valid reason for such large portions of the data to be designated as being ‘excessive’ or ’unusual.’ Tab. 23.6 shows, how far the sample bounds LSB and USB can extend from Q1 and Q3. The sharp selectivity of the sample bounds used to delimit normal values was demonstrated in Tab. 23.3 and in Tab. 23.4.

The role of sample bounds to establish normal values for ratios can thus be played by LSB and USB with theoretically justified and realistic outcomes. It can be seen from Tab. 23.4, that splitting the data into three clusters (L, M and U) by using the sample bounds LSB and USB depends on the form of the sample's distribution function. The result can be a very restrictive choice of ‘normal’ values as in the case of AAD/TA or EPS (with only 76.6% normal data), or an extremely tolerant classification as for DA/TA, CX/TA, RD/TA and $TOTR$ (with no abnormal data). This is a sharp contradiction to the current practice applied eg in production quality assessment, where a product is accepted as ‘normal’ if its qualitative parameter ϵ lies within a tolerance interval bounded eg by probabilities $Pr\{\epsilon\} = 0.03$ and $Pr\{\epsilon\} = 0.97$. Assume, that there is a sample of 100 products to be tested. Also assume, that the probability distribution of the parameter for this sample has been estimated; the distribution will—as a rule—differ from the 'normal' distribution in the statistical sense. Such a difference can be interpreted in at least two ways: It is caused by the

1. random sampling effect,
2. presence of the 'abnormal' products.

There is no reason to throw out 6 percent of the product in the former case, because all of them may be normal. Nor does the latter case provide a reliable base to justify the loss of 6 per cent, because the probability distribution of the 'abnormal' products is unknown and it can have many different forms depending on the causes of the abnormality. The use of such an approach to establish normality bounds for financial ratios is not less problematic. There is no 'fundamental' set of events, nor is there an (infinite) population, the distribution of which could be used as a reference. Further, there is no reason to expect, that a given (fixed) portion (eg 94%) of all firms performs well ('normally'), while the rest do not. A decision as to **whether** all firms behave in a comparable ('normal') way should be based on an analysis of a particular data sample. This is entrusted to the homogeneity test in gnostics. The percentage of normal data is then an objective outcome and not a subjective a priori assumption of the analysis. The same can be said with respect to the classification (M/L/U) for individual data.

23.2.9 Marginal Rating

Financial ratios are numbers and—as such—they can be ordered by their values. Order number $N_{m,n}$ of a ratio $R_{m,n}$ (of the m-th kind) of an n-th company can be used as rough information on the financial position viewed with respect to that ratio. Some ratios—eg as return on assets—have a monotonic nature 'the larger the better.' In such a simple case the 'worse' firm is the first firm and the 'best' is the last one. The 'average' position is near the median. However, many other ratios have a more complicated character: too small a value as well as one too large is unfavorable. The 'best' position is then somewhere near median and the performance can be measured by the absolute difference of the firm's order and the median's order. However, knowledge of the orders does not quantify the distance, and does not answer the question 'how far is ratio $R_{m,x}$ from the ratio $R_{m,y}$.'

Distribution functions of ratios are more suitable for this purpose: they establish the same order (because distribution functions are monotonic), but enable distances between positions to be quantified. Examples are shown in Tab. 23.7.

Let $R_{m,n}$ be a financial ratio's value of the n-th kind presented by the m-th firm. Entry $P_{m,n}$ in Tab. 23.7 reads 'the probability of not exceeding $R_{m,n}$ is $P_{m,n}$.' This relates to the first four columns, which characterize liquidity (RWC), activity ($TATO$), return on assets (ROA) and the reciprocal value of the financial leverage (TL/TA). Value $P_{1,1} = 0.18$ thus says, that only 18% of companies of the US Chemical Industry were expected to have RWC less or equal to AKZOY's, while 82% exceeded AKZOY's value. It is to be noted, that the probabilities were estimated by means of the local distribution function ELDF, because a part of the sample was inhomogeneous.

Once the ratios have been computed and ordered, the probabilities from Table 23.7 allow the meaning of the notion 'the smallest' or 'the largest' to be quantified. It is seen, for example, that the company GRA has the smallest $ROA(-0.0712)$ and SIAL the smallest $TL/TA(0.1511)$, WDFC the largest $ROA(0.3085)$ and SHLM an RWC of 0.5131. Note, that after using probabilities, these extremal values are not equivalent: exceeding the largest ROA can be expected only in 1.06% of cases, while exceeding the largest RWC can occur much more frequently, 3.08% of the time. Similarly, not exceeding the smallest ROA (-0.0712) is to be expected in 1.46%, while results less or equal to the TL/TA can appear in 3.66% of cases. The probability of an excessive value for a ratio is important in estimating certain losses and chances. A spread of estimates in the range 1:3 is not negligible.

Probabilities also enable the spread of individual ratios to be quantified. For example consider the probabilities of RWC: the order number for LI's value of 0.0969 is 17 and for ARG (0.0974) is 19. The appearance of a ratio between these values can be expected in only 0.14% of all cases. However, the distance between order number 40 (CEM with 0.264) and order number 42 (IFF with 0.414) is given a probability 0.113, which corresponds to an occurrence in 11.3% of cases. The same two-step movement along a sequence of ordered values and the exactly measured spreads differ by a large order of magnitude. These results show, that the marginal distribution functions of ratios are a more suitable instrument for making a judgment about ratios than is the marginal ordering the same ratios.

Table 23.7 further provides a perspective as to the trends in a firm's operational policies. Retained Earnings and Accumulated Depreciation and Amortization represent the past, TATO, ROA, Depreciation & Amortization Expense and Total Return mirror the present, and Capital Expenditures and R&D give an idea of the firm's expectations of the future. All these ratios can be evaluated using the 'the larger the better' principle. The same is valid for their probabilities, because these are monotonously increasing functions. Therefore the following composite indicators have been constructed and are seen in the last three columns of the table:

- $(Pr\{RE/TA\} + Pr\{AAD/TA\})/2$ (past position),
- $(Pr\{TATO\} + Pr\{DA/TA\} + Pr\{ROA\} + Pr\{TOTR\})/4$ (recent position),
- $(Pr\{CX/TA\} + Pr\{RD/TA\})/2$ (future position),

where $Pr\{Ratio\}$ denotes the probability of the *Ratio*.

The figures shown in the columns of the table can be used not only to make a judgment about the order of firms, but also to get an idea of their development through time. The following three-point patterns can be discerned in the table:

- A continuing high level: T15 with 0.70/0.72/0.69.
- A continuing low level: T26 with 0.33/0.36/0.33.
- Sustained long-term advance: T1 with 0.41/0.59/0.76 and T30 with 0.24/0.32/0.66.
- Sustained long-term decline: T31 with 0.67/0.54/0.27.
- Current peak: T42 with 0.28/0.47/0.19 or T11 with 0.47/0.62/0.48.
- Current trough: T21 with 0.47/0.09/0.45.

	PROBABILITIES OF BASIC RATIOS				TIME ASPECT		
TICKER	RWC	TATO	ROA	TL/TA	Past	Present	Future
T1	0.18	0.50	0.44	0.85	0.41	0.59	0.76
T2	0.25	0.15	0.61	0.53	0.60	0.56	0.44
T3	0.40	0.37	0.29	0.67	0.25	0.37	0.43
T4	0.30	0.88	0.80	0.46	0.72	0.78	0.68
T5	0.24	0.37	0.57	0.46	0.70	0.59	0.59
T6	0.78	0.18	0.24	0.13	0.62	0.33	0.73
T7	0.54	0.69	0.85	0.94	0.31	0.58	0.18
T8	0.03	0.14	0.86	0.52	0.60	0.46	0.70
T9	0.72	0.42	0.27	0.59	0.52	0.31	0.68
T10	0.25	0.24	0.47	0.61	0.86	0.47	0.72
T11	0.20	0.81	0.55	0.61	0.47	0.62	0.48
T12	0.32	0.70	0.91	0.33	0.55	0.85	0.73
T13	0.68	0.37	0.56	0.83	0.72	0.50	0.78
T14	0.74	0.74	0.52	0.66	0.31	0.55	0.31
T15	0.68	0.88	0.67	0.57	0.70	0.72	0.69
T16	0.59	0.70	0.20	0.58	0.36	0.48	0.26
T17	0.40	0.71	0.69	0.95	0.42	0.60	0.18
T18	0.84	0.17	0.39	0.25	0.30	0.37	0.23
T19	0.55	0.39	0.46	0.48	0.33	0.39	0.55
T20	0.08	0.10	0.01	0.95	0.28	0.10	0.44
T21	0.10	0.02	0.14	0.91	0.47	0.09	0.45
T22	0.89	0.46	0.95	0.11	0.68	0.54	0.80
T23	0.75	0.75	0.48	0.37	0.48	0.51	0.47
T24	0.40	0.63	0.52	0.59	0.24	0.46	0.51
T25	0.77	0.44	0.38	0.33	0.65	0.43	0.55
T26	0.64	0.80	0.22	0.56	0.33	0.36	0.33
T27	0.14	0.37	0.15	0.82	0.56	0.42	0.86
T28	0.52	0.51	0.68	0.41	0.67	0.61	0.61
T29	0.42	0.62	0.23	0.49	0.34	0.51	0.64
T30	0.49	0.07	0.08	0.64	0.24	0.32	0.66
T31	0.93	0.95	0.48	0.30	0.67	0.54	0.27
T32	0.33	0.41	0.25	0.53	0.55	0.45	0.43
T33	0.53	0.36	0.43	0.28	0.75	0.38	0.42
T34	0.37	0.34	0.12	0.64	0.66	0.35	0.58
T35	0.94	0.55	0.38	0.33	0.53	0.50	0.48
T36	0.57	0.27	0.58	0.20	0.60	0.60	0.69
T37	0.41	0.47	0.82	0.46	0.85	0.63	0.36
T38	0.23	0.82	0.67	0.32	0.48	0.71	0.34
T39	0.45	0.47	0.88	0.39	0.70	0.72	0.54
T40	0.97	0.93	0.72	0.13	0.34	0.60	0.19
T41	0.43	0.64	0.57	0.40	0.20	0.57	0.22
T42	0.91	0.29	0.86	0.04	0.28	0.47	0.19
T43	0.45	0.53	0.09	0.59	0.09	0.35	0.34
T44	0.28	0.24	0.47	0.57	0.79	0.50	0.77
T45	0.67	0.80	0.73	0.40	0.49	0.68	0.67
T46	0.96	0.98	0.99	0.06	0.44	0.70	0.21
T47	0.17	0.29	0.26	0.67	0.41	0.29	0.34

Tab. 23.7: Basic financial ratios and time aspects of US Chemical Industry, 1998 evaluated by means of probabilities

Investors are especially interested in the evaluation of prospects. Therefore the relationship of the columns Present and Future can be useful in this respect. An optimistic outlook is given for CEM (0.33/0.73) or MIL (0.42/0.86) and pessimistic ones can be seen for WDFC (0.70/0.21), SHLM (0.60/0.19), GGC (0.60/0.18) or CNK (0.58/0.18). Because the effect of both capital expenditures and research and development have long term trends, it is difficult to evaluate the accuracy of such estimates using data for a limited period of time.

Although such results can be helpful, their one-sided character as well as the subjectivity of weighting the ratios in terms of the time aspect can be improved only by using multivariate modeling. This topic will be examined in the next chapter.

23.2.10 Interval Analysis

The technique of interval analysis was described in Chapter 16 in connection with the examination of local distribution functions (ELDF). The procedure consists of the calculation of five characteristic values for a homogeneous data sample:

- AL ... The lower bound of the interval of typical data.
- $A0L$ ... The lower bound of the tolerance interval of the location parameter.
- $A0$ ... The location parameter (the location of the probability density's maximum).
- $A0U$ ... The upper bound of the tolerance interval of the location parameter.
- AU ... The upper bound of the interval of typical data.

This notation is for additive data. In the case of multiplicative data these bounds will be designated as ZL, $Z0L$, $Z0$, $Z0U$ and ZU.

Recall, how these parameters were defined: Assume, that a data sample of additive data $\mathcal{A}$ is homogeneous (or has been made homogeneous). Let N be its size. Let scale parameter S be chosen so, that the ELDF's density be uni-modal. Let A_0 be the location parameter defined as the quantile of ELDF's density maximum for this sample. Let A_x be another datum added to the sample. Let $A0(\mathcal{A},A_x)$ be the location parameter of the extended sample of $N+1$ data. Fig. 16.4 shows, how $A0$ reacts to changes in A_x: There are three cases of equivalence $A0 = A_0$ ($\mathcal{A}$ fixed):

1. for $A_x = A_0$,
2. for $A_x \to -\infty$,
3. for $A_x \to \infty$.

The location parameter of the extended sample thus coincides with that of the original sample, when the additional datum is equal to it or if it is extremely far from all the data of the original sample. As seen in Fig. 16.4, there are three ways, that the function $A0(\mathcal{A}, \mathcal{A}_\S)$ can behave depending on the value of A_x. It can:

fall when A_x increases from $-\infty$ to AL,

rise when A_x increases from AL to AU,

fall when A_x increases from AU to ∞.

Only the central portion of this range of values can be accepted as typical behavior—with an increasing location parameter given an increasing value for the additional data item. The reaction of the location parameter (in the first and third sections), when the extra datum lies beyond the upper and lower bounds is ‘strange’, ‘unnatural.’ It gives the impression, that the original data of the sample ‘defend’ the original sample against an encroaching ‘stranger,’ by ‘pushing’ it away. This is why only the interval $[AL, AU]$ can be called the *typical interval* of data.

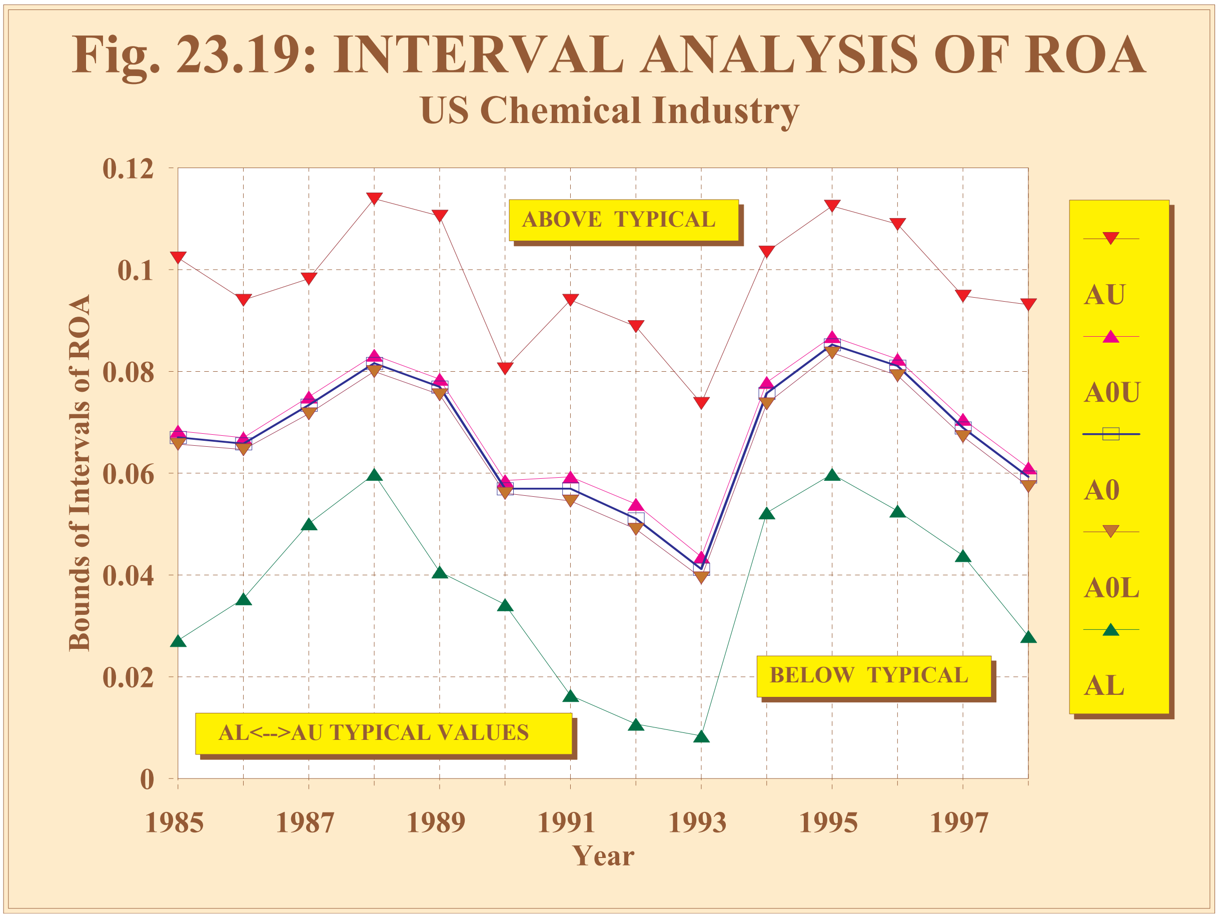
There are two other important points in the curve $A0(\mathcal{A},A_x)$:

1. $A0L = A0(\mathcal{A},AL)$,
2. $A0U = A0(\mathcal{A},AU)$.

To get a feel for the significance of these points, it is sufficient to consider the arithmetical mean of the extended sample, when A_x approaches $-\infty$ or ∞. In the former case, as a location parameter it reaches $-\infty$ and in the latter case ∞. In contrast, changes in the location parameter $A0(\mathcal{A},A_x)$ are bounded by AL (minimum) and AU (maximum) for any arbitrary value of A_x. This interval is called the *tolerance interval* of the location parameter

$A0$ and it shows, that this location parameter is **robust** with respect to an outlier.

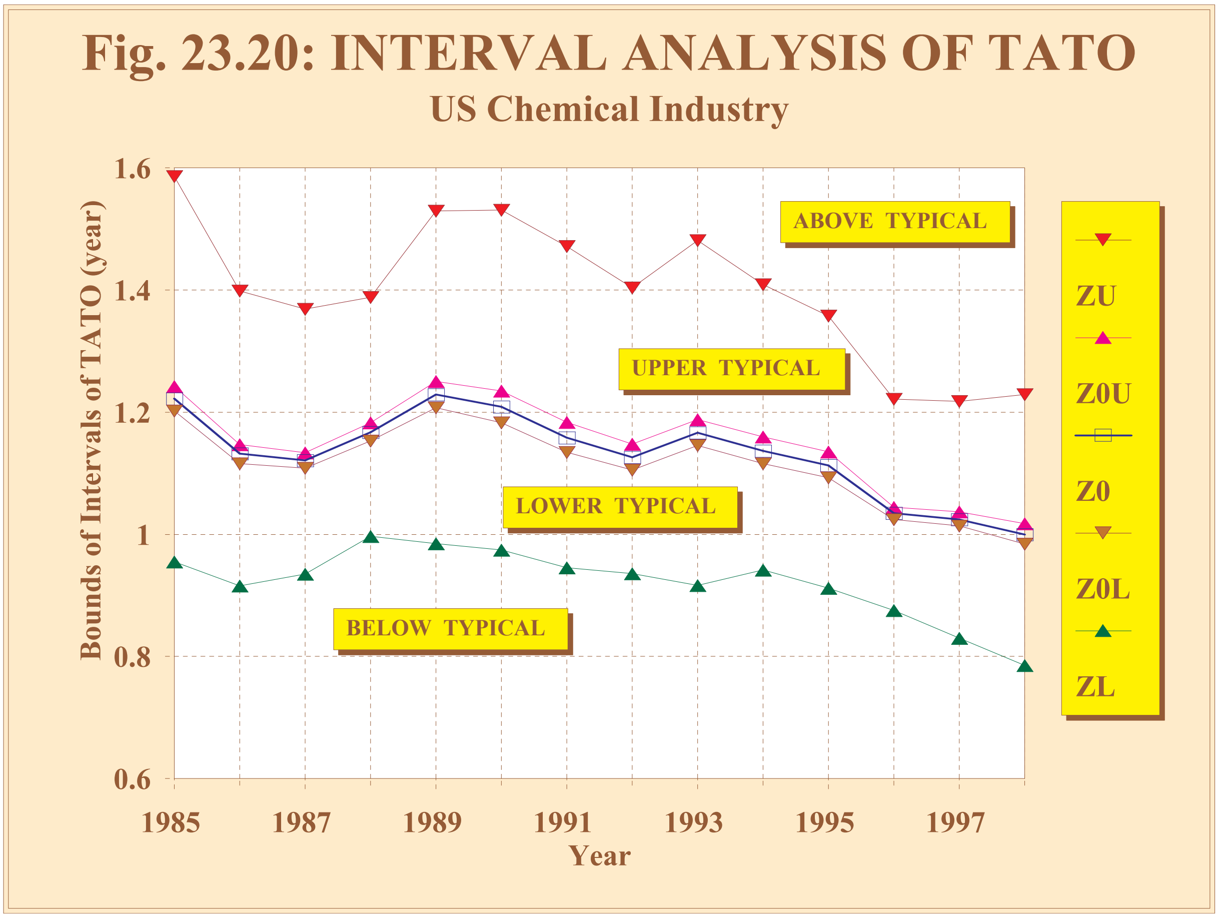
Fig. 23.19 shows how these characteristics evolve in the case of ROA (Return on Assets) defined as EAT/TA (Earnings after Tax divided by Total Assets).



The location parameter $A0$ (blue line) traces two waves with maxima in 1988 and 1995 and a global minimum in 1993. (Similar waves were observed in the probability distributions of $E/P_{\%}$ in Figure 23.8. The bounds of the typical interval (the lower—green—and the upper—red) follow the value of the median with some local deviations caused by the varying spread of data.

An analogous example related to multiplicative data ($TATO$, Total Asset Turnover) is in Fig. 23.20.

It shows, that this ratio is more conservative than the ROA. The curves in both pictures demonstrate the meaning of the statement 'this ratio has



a typical value' in a precisely defined gnostic interpretation. So eg, in 1998 the typical EAT/TA was between $AL = 0.028$ and $AU = 0.093$, while 'the most typical' ($A0$) was 0.059. The typical values of $TATO$ in 1998 were between 0.786 and 1.229 with the most frequently expected value $Z0 = 1.000$.

The narrow band of $A0L$ and $A0U$ ($Z0L$ and $Z0U$) shows how robust the $A0$ ($Z0$) is with respect to the extension of the sample by another datum with an **arbitrary** value: the location parameter changes only slightly.

An important practical application of interval analysis is to compare samples and to determine their degree of similarity. This is also described in Chapter 16. Using these criteria, most of the annual changes in the ratios can be classified as due to 'typical' similarity of the samples (ie typical intervals in subsequent years overlap) and in some cases ($TATO$ 1986/1987, 1989/1990, 1996/1997 and EAT/TA 1985/1986, 1990/1991) are all 'within tolerance' (with overlapping tolerance intervals). Such a compilation of the similarity of characteristics could be especially useful, when ratios of different industries are to be compared.

It is instructive to examine in more detail examples of the role played by interval bounds and to compare them with the sample bounds (LSB, USB), which establish the interval of 'normal' values as well as with the bounds of data support LB and UB (Tab. 23.8 and Tab. 23.9):

LB	Min	LSB	AL	$A0L$	$A0$	$A0U$	AU	USB	Max	UB
-1.468	-0.071	-0.071	0.028	0.057	0.059	0.061	0.093	0.192	0.309	0.195

Tab. 23.8 Summary of characteristic points of the additive data sample EAT/TA (ROA) for 1998

As shown in Tab. 23.4, the sample of ROA (EAT/TA) ratios was slightly inhomogeneous with 46 items classified as M (normal) and with 1 in U, an outlier (WDFC). This is why the interval $[LSB, USB]$ is within the span of the interval $[Min, Max]$. In contrast, the interval LSB, USB is wider than $[Min, Max]$ in the case of the homogeneous sample of multiplicative data TATO (Tab. 23.9).

Both tables demonstrate, that typical intervals are substantially narrower than the interval of normality ('membership' interval). This result shows, that for a ratio to be classified as 'typical' is a stricter requirement than 'to be normal.'

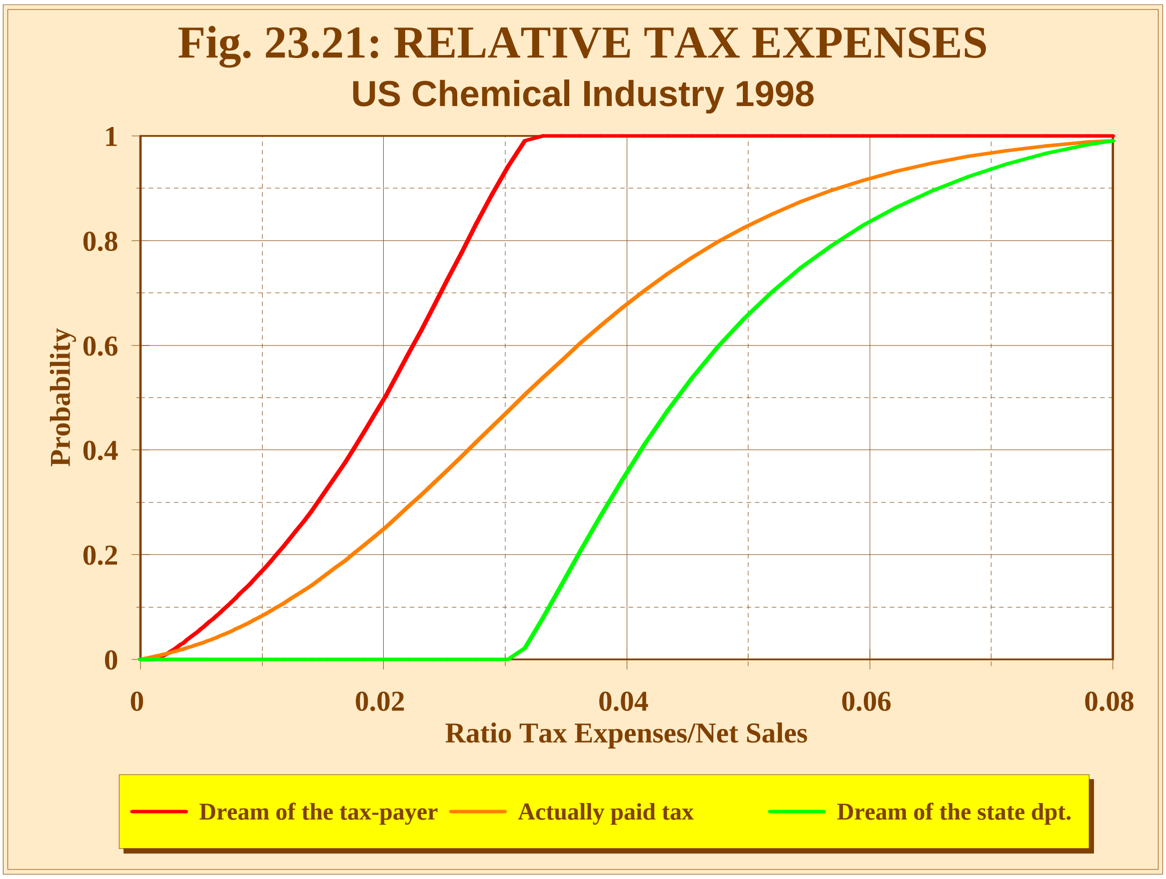
LB	LSB	Min	ZL	$Z0L$	$Z0$	$Z0U$	ZU	Max	USB	UB
0	0	0.368	0.786	0.984	1.000	1.018	1.229	2.035	2.123	2.136

Tab. 23.9 Summary of characteristic points of the multiplicative data sample TATO for 1998

Interval analysis together with sample bounds and bounds of data support enable several classes of data within a sample to be distinguished: subnormal, normal but subtypical, lower typical, upper typical, normal but above typical and above normal. However, data values smaller than LB and larger than UB are improbable, ie unexpected.

23.2.11 Effect of Censored Data

As discussed in Chapter 19, some data can suffer from incomplete definition by being given either one-sided bounds, which limit the inclusion of possible data values, or some other indefinite intervals rather than specific numbers. Such statements are of the type 'at least X' (right-censored data) or 'not larger than X' (left-censored data) or even 'exceeding X_1, but less than X_2 (interval data). To show, that even such 'second-rate' data can provide useful information, an example using income tax is instructive. To eliminate the size comparability problem, the ratio of tax expenses to net sales is introduced. The brown line in Fig. 23.21 is the probability distribution EGDF of this ratio ('relative tax expenses,' ie tax expenses divided by sales), which is calculated for the homogeneous 'main cluster' of companies in the US Chemical Industry, which consists of 36 firms.



This distribution reflects actual values of the relative tax burden as presented in their financial statements for 1999. Since no one wishes to pay more tax than is required, the firm's accountants do their best to ensure, that all legal loopholes are utilized, so that only the minimum tax consistent with the 'fair play' rules of accounting is paid. In this context, tax paid can be thought of as the upper bound of real tax obligation. Assuming, that the all of the considered ratios of relative tax expenses are left-censored, the distribution function of the thusly interpreted ratios is shown in Fig. 23.21 (red line). The "philosophy" of this approach could be called "dreams of the taxpayer", who pays only the minimum possible and never more than the official duty. However, their might be the opposite point of view, that of the state department believing ("dreaming"), that the taxpayers take their civil duties as serious, as to pay "as much as possible and never less than the legal duty. Several observations can be based on this graph:

1. The effect of this type of 'censoring' is not negligible. The median of 'true' values of relative tax expenses can be as low as 0.0201 (when the taxpayer's dream would prevail) or as high as 0.0444, when the official optimism became reality. (Robust median 0.0316 corresponds to the officially declared and paid values.
2. The probability, that tax expense in the Chemical Industry exceed 9.7% of net sales is practically zero. This limiting value is not influenced by the censoring. However, the mean value of the right-censored ratios (0.0471) exceeds that of those uncensored (0.0331) because of a higher concentration of large values close to the maximum.
3. Some companies reported negative tax expenses (due to losses or carry-forwards from previous years), but these were excluded as causing inhomogeneity. As seen in Fig. 23.21, the probability of negative taxes were therefore zero even for the left-censored data.

This illustration shows, that using the technique of censored data provides some interesting interpretations, and that it can be used to estimate effects, which cannot be quantified by any other method.

23.3 Robust Correlation Coefficients

23.3.1 The Need for Robust Correlations

Correlation coefficients are parameters, which characterize the degree of 'similarity' of two data samples, say, $\mathcal{A}$ and $\mathcal{B}$[13]. As such they are closely related to bilinear approximations of relations between a pair of variables.

It is assumed, that data form a sequence or series of pairs $\langle a_k, b_k \rangle$, where $a_k \in \mathcal{A}$, $b_k \in \mathcal{B}$, $k = 1, 2, \ldots$. Typically, the application of correlation coefficients is a preliminary stage of multidimensional modeling to select those variables, which have the strongest relation to the dependent variable. The strength of the relationship is quantified by the absolute value of the correlation coefficient of the explanatory/dependent variable. The stronger the correlations, the higher the quality of the model. When using data

[13] In geometric terms, the classical correlation coefficient is the cosine of the angle between two multidimensional vectors $\mathcal{A}$ and $\mathcal{B}$

contaminated by uncertainties, which include outliers, additional requirements exist: the methodology for estimating correlation coefficients should be robust. An example based on financial statement data illustrates the importance of robustness in the estimation of correlation coefficients.

Four ratios were created from data taken from the financial statements of 43 companies in the US Chemical Industry for 1998:

1. $ROA = EAT/TA$... Return on Assets (Earnings After Tax divided by Total Assets),
2. $EPS = EAT/SO$... Earnings per Share (Earnings After Tax divided by the number of common shares outstanding),
3. DIV... common dividends paid per share,
4. PS... market price of a share.

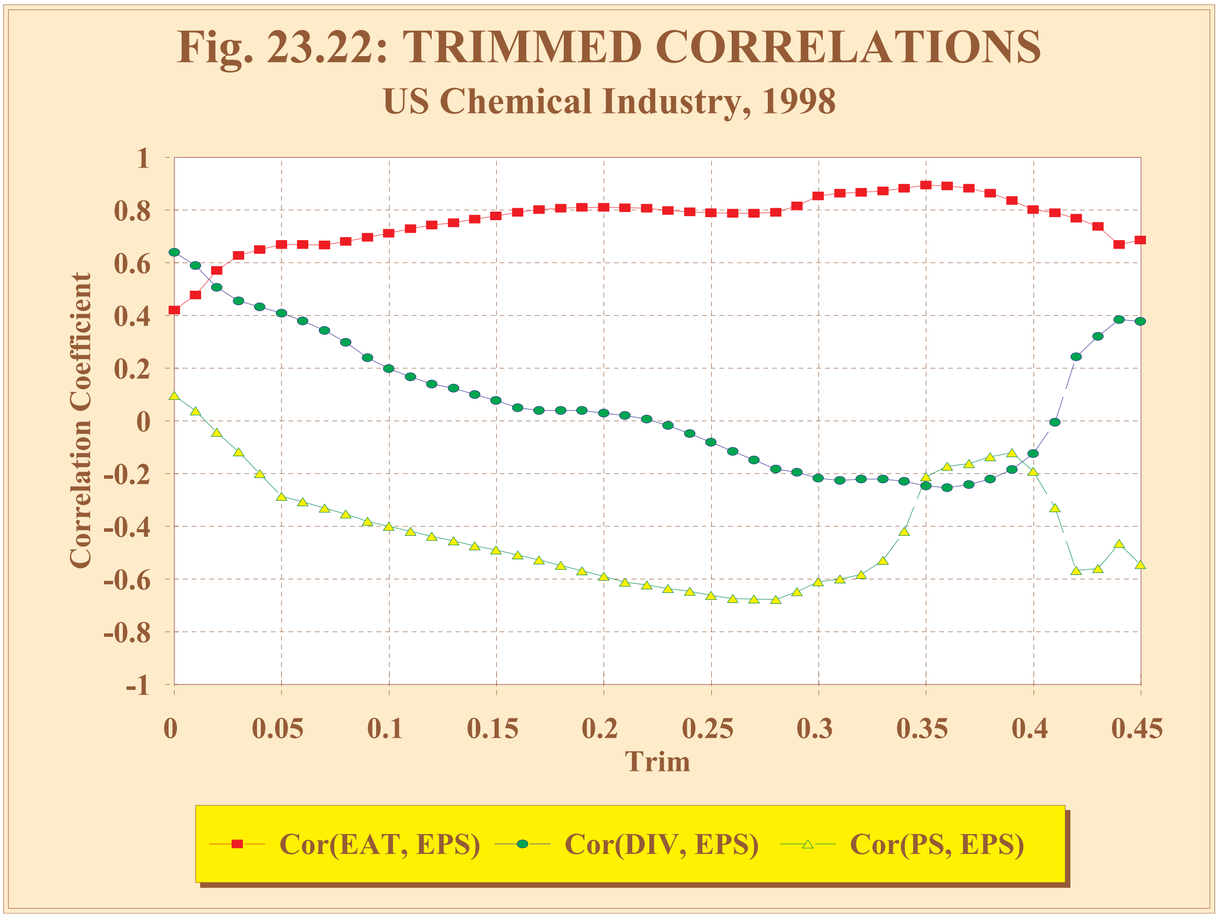
There is no doubt, that these ratios are interdependent, but the requirement is to evaluate the degree of such mutual dependence by using correlation coefficients. The data surely contain uncertainties, therefore the usual Pearson's (unrobust) method will not be useful and robust estimation methods are desirable.

The initial attempt uses the oldest robust statistical methodology of trimming: a specific percentage of the ordered data is cut off from both sides of the ordered series. The results obtained by using the program for trimmed estimation of correlation coefficients available in the S-PLUS package ([103]) are shown in Fig. 23.22.

It is obvious, that by choosing different level of trimming, very different estimates can be obtained. There are two other robust estimates of correlation coefficients available in S-PLUS, Kendall's and Spearman's rank correlation methods. These were also applied to the correlations between the four ratios; the results of all three statistical approaches are summarized in Tab. 23.10 and they are compared with both the unrobust and the gnostic estimates. The values obtained by trimming are presented as an interval of the outcomes given for different levels of trimming from zero through 0.45. The classical (Pearson's) correlation coefficients are a special case of trimmed values, for zero trimming.

The table suggests the following observations:

1. The critical value for the classical (Pearson) estimates at the significance level of 0.01 is 0.371. Therefore, Pearson estimates 0.418 and 0.645 are highly significant, but as can be seen from the robust estimates, they are unacceptable because of their corruption by outliers.
2. The strong influence of outliers is demonstrated by the broad range



of trimmed estimates. It is obvious, that practically any desired value can be obtained by choosing a particular level of trim.

3. The critical value for the Spearman's robust estimate at the significance level of 0.01 is 0.380. Both pairs of variables (ROA, EPS) and (DIV, EPS) can therefore be considered as correlated, while (PS, EP) is not. However, neither the unrobust Pearson nor the robust Kendall estimates correspond to the robust Spearman estimates.
4. There is no contradiction between Spearman and gnostic estimates.

The last observation confirms the validity of the theoretically justified gnostic estimates of correlation coefficients. Recall, that in Chapter 14 it was shown, that gnostic correlation coefficients are inherently connected to the gnostic probability distribution functions. Indeed, irrelevance h is the simple linear function $2 * p - 1$ of the probability p (see 10.60). Irrelevance is therefore the deviation of the probability from its "neutral" value of $1/2$. The gnostic correlation coefficient is the cosine of the angle between the vectors of irrelevances (14.20). It is thus a bi-linear function of probabil-

Correlated	Method				
Variables	Trimming	Pearson	Kendall	Spearman	Gnostic
(ROA, EPS)	$[0.421, 0.896]$	0.418	0.482	0.668	0.640
(DIV, EPS)	$[-0.252, 0.645]$	0.645	0.295	0.401	0.401
(PS, EPS)	$[-0.667, 0.098]$	-0.049	-0.114	-0.146	-0.140

Tab. 23.10: Comparison of unrobust (Pearson's) and robust estimates of correlations between four financial ratios (1998)

ities of 'synchronous' occurrence of observed data. Gnostic distribution functions are robust, hence gnostic correlation are robust as well.

23.3.2 Inertia in Financial Ratios

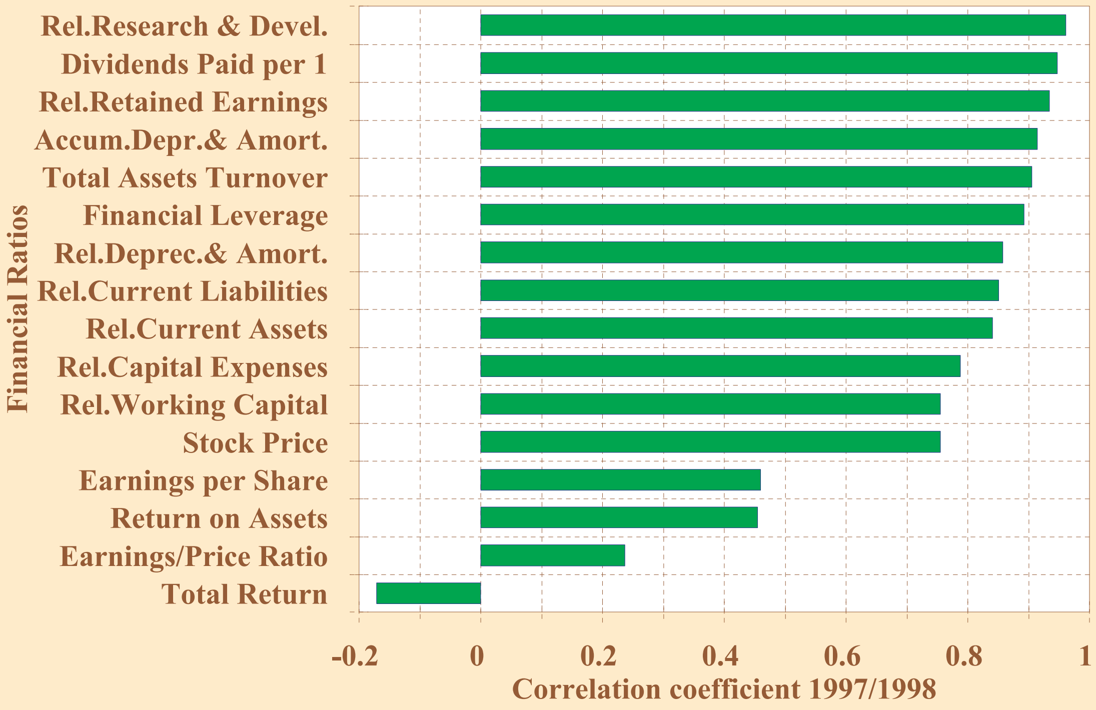
Auto-correlation functions of time series can be obtained as sequences of correlation coefficients estimated for different time lags between the elements of the series. The form of an auto-correlation function reveals the basic features of the mechanism, which generates the data. One such feature is inertia, which limits the probability of rapid changes in the observed variable. In cases involving financial data, inertia can be influenced by policy decisions such as those that establish long term strategies for financial control, such as eg the maintenance of research & development investment or dividend payout at a constant level. Fig. 23.23 demonstrates, that this effect really exists and evaluates it for the principal financial ratios.

The ratios are plotted on the vertical axis and the length of the columns corresponds to the gnostic estimates of correlation coefficients between the ratio's value in 1997 and 1998. The strongest correlations (Research & Development, Dividends Paid, Retained Earnings, Accumulated Depreciation & Amortization) can be explained by the continual influence of management policies. In other cases such as Total Asset Turnover, Financial Leverage, etc. the "inertia" limiting rapid change is not only based on management's decisions, but also caused by external factors, which constrain the ability of managers to intervene. In any case, substantial inertia and 'conservatism' exists in 12 of the 16 financial parameters. The more random (and unexpected) behavior of EPS, ROA, $E/P_{\%}$ and TR can be explained especially by their strong uncertainty, which is—in the case of TR—even amplified by using the stock price change.

As might be expected (and will be seen), strong autocorrelation favor

Fig.23.23: INERTIA IN FINANCIAL RATIOS

US Chemical Industry, 1997/1998



useful predictions.

23.3.3 Correlations of Financial Ratios

A positive/negative correlation coefficient in some variables u and v means, that an increase/decrease in u was 'frequently' accompanied by an increase/decrease in v. The opposite also holds, because the correlation coefficient is a symmetric function of both variables. Such tendencies to similar/opposite behavior can be caused by different situations:

1. random coincidence,
2. one-sided causal relation: variable u is an effect caused by v (or the converse),
3. multilateral causal relation: both u and v are effects of some other related causes.

The probability of random correlation decreases with the number of pairs considered, and it is therefore rare. However cases of both No.2 and 3 can be seen in practice and it is not easy to distinguish, which of the two occurs in any particular case. Variables, which have a deterministic or stochastic functional relationship are correlated, but the opposite is not necessarily true. This is the reason, that a significant correlation does not necessarily imply a causal relation[14].

In complex systems such as the economics of a firm, the interactions between financial ratios are surely subjected to multidimensional models and cannot be thought of as pairs of isolated one-sided cause-effect relations. The identification of multidimensional models is therefore the best way to reveal the structure of interactions between variables of a complex system. Even so, **some** information can be obtained by an analysis of the correlation structure and as an example we return to the ratios for the US Chemical Industry for 1998.

The most significant correlations between the sixteen ratios for 1998 are worth being discussed further[15] The largest non-diagonal element of the 16x16 correlation matrix, (0.754), is for CA/TA and RWC. This can be easily explained, because RWC equals the difference between CA/TA and CL/TA.

A less trivial correlation is that of $TATO$ with CA/TA (0.512): the shorter the production cycle—the greater the need for current assets. The strong negative correlation (-0.584) between financial leverage TL/TA and the liquidity measure RWC is not surprising: firms, which have a sound financial base can afford to operate with high financial leverage and low liquidity. However, a large amount of total debt can be accompanied by decreasing return (increasing interest expense): there is a strong negative correlation (-0.426) between TL/TA and ROA.

[14] There is an extensive literature on this subject, for instance, see:

1. "Commercial Crises and Sunspots," Jevons (1888) revisited in Sheehan & Grieves, Southern Economic Journal, V48, Jan. 1982, pp. 775-777: Using data from 1889 to 1979, the U.S. economy has a significant impact on sunspots, but the reverse is not true.
2. "Econometrics – Alchemy of Science?" David Hendry, Economica, V47,Nov. 1980, pp. 387-406: Rainfall in Great Britain is a good predictor of inflation.
3. "Spurious Regressions in Econometrics," C.W.J. Grainger & P. Newbold, Journal of Econometrics, V2, July 1974, pp. 111-120. Nonsense regressions due to autocorrelation problems.
4. In Scandinavia, statisticians once found a close correlation between the number of births and the presence of storks in Scandinavia. While this condition was not accepted as a proof, that babies are brought by storks, the idea still lingers in today's childrens' literature.

The lesson is, that while the numbers relate a mathematically sound result, the analyst must ensure that the relationship being tested has a sound practical connection.

[15] The correlation coefficients cited here have been determined by gnostic formulae.

All three ratios, which characterize future planning (CX/TA, RE/TA and RD/TA) are positively correlated: 0.377 (CX/TA with RE/TA), 0.385 (CX/TA with RD/TA) and 0.438 (RE/TA and RD/TA). On the other hand, high debt is not favorable for a forward-looking strategy: the correlation of TL/TA with RE/TA is negative (-0.313), but dividends paid reflects this strategy positively (correlation coefficient of DIV with RE/TA is 0.411 and with RD/TA 0.405) (dividends are a demonstration of management's positive expectations for the future). High correlations between EPS and EAT/TA (0.621) and EP (0.526) are a consequence of the definition of these ratios, they all are functions of the same variable (earnings).

It is also interesting to see, how the individual financial ratios are correlated with stock prices and to compare these correlations: PS with DIV (0.636), PS with EPS (0.579), PS with RD/TA (0.340), PS with EAT/TA (0.333), PS with ADA/TA (0.326), PS with RE/TA (0.293) and PS with TL/TA (-0.272). These figures help to get a grip on how the market "weighs" the individual characteristics of the firm's financial position. A high correlation (0.533) between PS and TR can give some analysts a preference for TR (total return) over other indicators: their point of view is close to the market's. Somewhat surprising is the practically negligible correlation (-0.044) between the liquidity measure RWC and PS, and only 0.055 with TR. This is also due to the definition of RWC as $CA/TA - CL/TA$ and the high **positive** correlations of PS and TR with **both** CA/TA and CL/TA. The positive evaluation of sufficient current assets seems to be straightforward, but the case of CL/TA is different. However, these 'strange' correlations—together with the positive correlation (0.344) of CL/TA with RD/TA, $TATO$ (0.229), EPS (0.225) and DIV (0.218)—signal, that the minimization of current liabilities is not a really sound idea.

The cross-correlations between the most fundamental ratios are shown below in Tab. 23.11:

23.3.4 Causal Interactions of Financial Ratios

In all these illustrations, the strong correlations referred only to 'parallel' or 'synchronous' changes in variables without consideration of the causal aspects. The cause must always precede the effect. There is no symmetry in correlation coefficients with respect to time. The non-zero correlation coefficient of a sequence $\langle u(t,1),\dots,u(t,N)\rangle$ and a previous sequence $\langle v(t+d,1),\dots,v(t+d,N)\rangle$ with a positive d can be interpreted as a causal effect of u onto v, but not the converse. If a ratio's values $R1(1997)$ taken from the 1997 data are strongly correlated with a ratio $R2(1998)$ then it is possible, that a causal effect could contribute to the correlation coefficient $\mathrm{cor}(R1(1997), R2(1998))$ but not to $\mathrm{cor}(R1(1998), R2(1997))$. When these coefficients are equal, then the correlation must be assigned to the impact of other variables (cross-correlations). An example is the strong correlation (0.631) of $DIV(1997)$ with $PS(1998)$ and nearly the same (0.619) correlation of $PS(1997)$ with $DIV(1998)$. Neither conclusion: 'high share prices were caused by good dividends' nor 'high dividends were caused by high stock prices' would be appropriate. These values would be rather characterized by 'good firms pay high dividends **and** are well priced by stock market', while the classification 'good firm' is based on an appreciation of other operating characteristics. But the correlation matrices of the 16 ratios considered for 1997 with their values in 1998 (and the matrix of 'opposite' correlations) offers examples of noticeable causal effects:

- $\mathrm{cor}(EPS(1997), DIV(1998)) = 0.658$ vice
  $\mathrm{cor}(DIV(1997), EPS(1998)) = 0.320$:
  'high EPS this year will increase dividends next year',
- $\mathrm{cor}(EPS(1997), PS(1998)) = 0.520$ vice
  $\mathrm{cor}(PS(1997), EPS(1998)) = 0.347$:
  'high EPS this year will increase PS next year',
- $\mathrm{cor}(TL/TA(1997), PS(1998)) = -0.330$ vice
  $\mathrm{cor}(PS(1997), TL/TA(1998)) = -0.021$:
  'high debt this year will decrease PS next year',
- $\mathrm{cor}(PS(1997), E/P_{\%}(1998)) = -0.356$ vice
  $\mathrm{cor}(E/P_{\%}(1997), PS(1998)) = 0.021$:
  'high stock prices this year will cause low $E/P_{\%}$ next year',
- $\mathrm{cor}(EAT/TA(1997), DIV(1998)) = 0.357$ vice
  $\mathrm{cor}(DIV(1997), EAT/TA(1998)) = 0.099$:

Ratio	*RWC*	*TATO*	*EAT/TA*	*TL/TA*	*TOTR*	*PS*
RWC	1.000	0.325	0.192	-0.584	0.055	-0.044
TATO	0.325	1.000	0.393	-0.185	0.266	-0.110
EAT/TA	0.192	0.393	1.000	-0.426	0.473	0.333
TL/TA	-0.584	-0.185	-0.426	1.000	-0.308	-0.272
TOTR	0.055	0.266	0.473	-0.308	1.000	0.533
PS	-0.044	-0.110	0.333	-0.272	0.533	1.000

Tab. 23.11: Robust correlation matrix of fundamental ratios

'high ROA this year will raise dividends next year'.

The asymmetric behavior of these figures shows, that some causal effects can be identified by an analysis of covariances. It should be noted, that using this method, the already discussed 'strange' positive correlations of CL/TA with 'favorable' ratios cannot be explained by the effect of a single ratio. The strongest (1997,1998) correlation of $CL/TA(0.850)$ occurred with the last year's value saying 'it is difficult to change current liabilities in a short time interval, there is a substantial inertia in this variable.' The complexity of the multidimensional cause/effect structure manifests itself in this case. The conclusion, which results from this and other similar experiences is, that the relations between elements of financial statements and financial ratios must be examined in a multidimensional mode (which is the subject of the next chapter).

23.4 Summary

Marginal analysis of data based on the application of gnostic distribution functions to financial data is a way to obtain a broad selection of useful information:

- A rich choice of universally applicable distribution functions, which are derived without a need for subjective a priori assumptions possess predictive power due to the objective reflection of the regularities in the data. These provide a robust quantification of the risks of various strategies provided directly by the data.
- robust estimates of bounds of data supports as the expected range of potential data, which could originate from the source, which generated the given sample,
- robust measures for the location and spread of the data sample,
- objective testing of data for homogeneity with a unique outcome determined only by the data,
- procedures to extract the main homogeneous cluster from an inhomogeneous data sample and to accomplish a robust marginal cluster analysis,
- robust estimation of sample bounds (ie of the 'membership' interval of data values potentially acceptable as belonging to the given homogeneous data sample). This can be used as a range for 'normal' data defined by a given data sample,
- robust estimation of the bounds of 'typical' data, in a subinterval of

the membership interval and to compare the degree of similarity of data samples,

- evaluation of effects caused by incomplete definition of data (effects of censoring),
- taking into account the possible heteroscedasticity of data,
- robust estimation of correlation coefficients with applications to:
  - reveal and quantify the interdependence of observed variables,
  - evaluate the inertia in characterizing processes,
  - analyze possible causal effects between variables.

Chapter 24

Advanced Fin. Statement Analysis II

Three modifications of the regression problem together with the expected effects resulting from the choice of the approach were described in Chapter 17:

1. robust explicit modeling,
2. robust implicit modeling,
3. robust probabilistic models.

This chapter will examine the outcomes of these methods using real data.

24.1 Objective (Mathematical) Rating

There is no doubt, that economists understand the notion, and the practical significance, of *rating* various quantities. But as frequently happens, the popularity of an idea is no guarantee of the clarity of its sense. Economic literature is replete with statements, that fundamental analysis, making use of both quantitative and qualitative information, rather than mathematical procedures should be used as a basis for rating, relegating numerical analyzes to the margin. This point of view defines rating more as an art than as a scientifically based technology. Judging art is, of course, highly subjective; on the other hand, the generally accepted notion of rating is "a measurement of how popular or good something is" ([80]). Measurement is a typical technical procedure, which tends toward maximum objectivity. The problem of measuring economic objects or processes is, that—unlike in physics—there is no instrument or "rule" available. The much touted hope, that statistics would somehow find a solution has not been corroborated in practice. Mathematical gnostics offers new ways to process uncertain data; it then makes sense to try to use gnostic instruments for this purpose. The objective is not to reject fundamental analysis, but to combine it with mathematical methods to attain a much broader and more objective solution.

The complexity of economic measurement includes (among other things) the necessity to take into account many factors, therefore multivariate measurement techniques must be applied. The first problem to be examined is the manner, in which objects defined over a multidimensional space could be ordered.

24.1.1 Multivariate Ordering of Objects

Mathematics has shown, that a universally applicable method of ordering in a multivariate space does not exist. Take eg two points X and Y in the M-dimensional linear space R^M. The location of these points within the space is uniquely defined by their coordinates. But to reveal other important features related to these positions requires, that additional instruments be defined: elemental relations between the points are needed. Three of these can be formulated as questions:

1. *Is X equivalent to Y?* Formally: does the relation $X \simeq Y$ hold?
2. *Does X precede Y?* Formally: does the relation $X \prec Y$ hold?
3. *How far is X from Y?* Formally: which geometry is to be applied to measure the distance between the two points?

A mapping is needed to answer these questions: a *criterion function*, $CrF : R^M \to R^1$, which establishes the rules for ordering:

$$(\forall X, Y \in R^M)(x = CrF(X), y = CrF(Y), (x, y \in R^1)) \qquad (24.1)$$
$$((X \simeq Y) \iff (x = y))$$

and

$$(\forall X, Y \in R^M)(x = CrF(X), y = CrF(Y), (x, y \in R^1)) \qquad (24.2)$$
$$((X \prec Y) \iff (x < y)).$$

Denoting $\mathcal{D}_M(X, Y)$ the distance between M-dimensional objects X, Y and $\mathcal{D}_{G,1}(x, y)$ the distance between real numbers x and y measured using a univariate geometry G, the definition of the former distance may have the form

$$(\forall X, Y \in R^M)(x = CrF(X), y = CrF(Y), (x, y \in R^1)) \qquad (24.3)$$
$$(\mathcal{D}_M(X, Y) \ := \ \mathcal{D}_{G,1}(x, y)).$$

The problem has thus been reduced to a (seemingly) elemental mathematical task: to chose a proper function CrF and a suitable geometry G. The real complexity of the situation arises, when 'a proper' is understood as 'the best' and 'suitable' as 'the natural'. Something can be 'the best' only in a certain sense, but how can the preference of 'senses' be established? And as to geometry: according to Riemann's point of view, the choice of geometry (of the metric) should be determined by something objective, eg by a Law of Nature. Are there any reasonable answers to these questions?

The problem is far from a purely academic discussion. Taking an example from financial statement analysis: X and Y are sets of several financial ratios from two companies. Are the financial/operational characteristics of the companies equivalent? If not, then which company's position is better? How far are they apart?

Other important problems of this type exist, eg the overall quality of a car. In this case the X and Y are sets of decision parameters for cars such as power, acceleration, mileage, weight, reliability, maximum speed, safety, and price among others. Are two cars equivalent? Or which of the cars is better? How much better is one car than another?

It has been said, that "the invisible hand" of the free market automatically takes over all the 'duties' of functions CrF and $\mathcal{D}_{G,1}(x, y)$ by establishing market prices eg for shares. "Two shares are equivalent if their price/earnings (P/E) ratios are equal." "A share is better if its P/E ratio is lower." But there is a 'Catch 22' here: A share is better than that of another **equivalent** company if its P/E is lower. If P/E's are equal does growth or economic potential play a role too? Or are all these things already imbedded in the numbers by the 'invisible hand?' The use of the Price/Earnings ratio as if it were all encompassing in this respect is paradoxical:

1. Low P/E stocks are thought to be 'cheap', while those with a very high ratio are believed to be overpriced for the earnings being delivered. Therefore the investor is hoping, that the P/E of his shares will increase substantially. This, however, requires that earnings remain relatively modest, while the stock price increases. Is this realistic? If earnings double, the price must rise by a greater proportion for the ratio to increase. Does this happen? Of course! Not only for those, that are bid up because of excellent potential, but also those related to them by some sort of 'kinship.' Look at all the speculative bubbles over time; but when they burst, prices come tumbling down fast and far, the 'good' along with the 'dogs.'

2. If a stock is purchased, because its P/E is attractive and the price is up by a small amount, the P/E will likely increase. Is its quality now a worse than before, does it now have greater risk?
3. Low P/E shares can have low earnings and consequently low prices, because they are not expected to perform well and are therefore fully priced. On the other hand, some low P/E stocks could be 'stars' if only they were noticed and properly valued. How can one differentiate between them?
4. For stocks actively followed by analysts, the P/E reflects the market's recognition (and the current psychology), but it doesn't measure, what the investor receives.
5. Alternatively, the Total Return (TR) is shareholder-oriented, but infers the firm's performance characteristics only indirectly through the change in stock price.

This latter alternative, which takes into account all the gain brought by a stock, the TR, reflects both approaches and might serve as a better barometer. "Hence, the order of preferential financial positions is determined by the order of Total Returns." The "distance between the financial position of two firms is measured by the difference in their Total Returns."

Although they are reasonable and thus usable, both of these approaches suffer from drawbacks:

1. Both PE and TR are very volatile, especially when quarterly data are used.
2. They combine two different views, that "from the inside" (by employing accounting data such as earnings or dividends) with the view "from the outside" (by using stock prices). Stock prices, in an indirect way also show how the market reflects the accountancy data. However, the current financial position of firms over the corresponding time periods is more stable than the evaluation by the market at the same time. A combination of the "inner" valuation with the "outer" one can lead to unnecessarily high volatility as shown by the examples below.
3. They do not reveal the "mechanism", by which the market's evaluation is established, nor how the weights to be given to the individual elements of the financial statements are obtained.
4. The uncertainties in their values (and risks connected with their application) are not explicit, they are 'hidden' in the price component.

These problems can be solved by using multidimensional models, along with cross-section and time-series processing.

24.1.2 Cross-section Ordering and MD-rating

The frequently used term *industry composite measure* is not generally suitable for detailed economic analyzes of the type considered here:

1. It ordinarily denotes a point estimate (a single number), but the advanced analysis uses more complex characteristics to describe economic behavior (distribution functions, multidimensional models etc.).
2. The term ‘industry’ implies the idea, that all firms in such a set are comparable. The advanced analysis shows, that an industry can be inhomogeneous in the sense, that economically incomparable firms are included, while comparable firms exist in different industries and are not taken into account.
3. The term ’composite’ can be interpreted as the use of some elemental statistical techniques (point estimates) such as weighted arithmetic means or sample quantiles to summarize the characteristics of the firms. The advanced analysis makes use of more sophisticated methods.

Instead, a more flexible notion of *cross-section analysis* will be applied and illustrated through the use of financial statement analysis: Take a multidimensional set of elements from financial statements of different firms defining their state of accounting at a given point in time along with a set of market indicators at the same moment. The ‘cross-section’ is the data set and the ‘analysis’ is an application of the advanced methods to the set. There are no a priori assumptions on the homogeneity of the set nor on comparability of the firms; decisions with respect to these features should be internal to the analysis. The choice of members for the set to be analyzed is therefore arbitrary, based on intuitive expectations, on hypotheses, additional information etc. Three variants of an analysis of this type will be distinguished by applying different points of view:

1. **Inner:** Only financial statement data for the firms, which form the cross-section, are analyzed to evaluate the financial position of firms.
2. **External:** Only market indicators are used for valuation.
3. **Both internal and external:** Both financial statement and market data are employed.

The subject of this section is only one of the tasks of the cross-section analysis, namely the ordering of multidimensional objects based on a real criterion function. Specifically, the method will be illustrated by ordering the financial statements of firms based on several ratios made up from their respective statements. The results of this ordering will be called the *MD-rating*.

Consider once more the basic financial ratios, the relative current assets CA/TA, relative current liabilities CL/TA, total asset turnover $TATO := NS/TA$[1], financial leverage TL/TA and return on assets (before tax and interest) $ROA := EBIT/TA$. The initial aim is to obtain an inner ordering, without taking into account the market's evaluation.

To examine the reliability of such an approach, this method will be applied not to a single cross-section, but to a time series of 44 cross-sections formed by quarterly data from two US industries covering the period from the second quarter of 1990 (90Q2) through the first quarter of 2001 (1Q1). The order number of a firm at a point in time quantifies its position in the cross-section. It is thus a relative rating with respect to the other firms. The time series of a single firm's order numbers represents a record of its relative financial history over the period analyzed. To obtain a continuous record of this type, only those firms, which had complete data in all the 44 cross-sections were included, leading to the retention of 88 firms (of which 30 belonged to the US Chemical Industry and 58 to several segments of the High-Technology Industry). A 'mixture' of firms from obviously different industries was chosen to see if there are any regularities, which are common to industries of a different nature.

The key element of the method is the application of the system of explicit regression equations in probabilities

$$\begin{aligned} C_{q,0} + C_{q,1} * Pr\{CA/TA_{q,k}\} + C_{q,2} * Pr\{CL/TA_{q,k}\} \\ +C_{q,3} * Pr\{TATO_{q,k}\} + C_{q,4} * Pr\{TL/TA_{q,k}\} &= \\ Pr\{EBIT/TA_{q,k}\}, \end{aligned} \tag{24.4}$$

where q denotes the sequential number of the quarter (q=1,...,44), $C_{q,0}$ through $C_{q,4}$ are the unknown model coefficients, $Pr\{R_{m,q,k}\}$ is the probability of the m-th ratio's value of the k-th firm ($k = 1, \ldots, 88$) at the q-th time. Such an equation system was set up and solved for each from 44 quarterly cross-sections.

Because the ratios entering this equation system are volatile, the volatility

[1] Note, that these ratios depend on the time unit used to define net sales NS. The following numerical examples will use quarterly data resulting in values of the $TATO$ equal to about a quarter of the annual $TATO$. Quarterly $TATO$, along with the other ratios will be 25% of the annual value only if sales and costs are distributed equally over the year.

of the system's solutions was minimized by using regression equations in probabilities and a robust method of solution was applied to make use of the double filtering effect[2] so as to maximize robustness.

The following steps were taken to obtain the MD-rating:

A Ordering within an n-th cross-section:

1. Form the data matrix M_n of ratios from financial statements for the n-th quarter, so that the k-th row is composed of ratios $CA/TA_{n,k}$, $CL/TA_{n,k}$, $TATO_{n,k}$, $TL/TA_{n,k}$ and $EBIT/TA_{n,k}$ denoted generally by Rn, m, k $(m = 1, \ldots, 5)$.
2. Estimate the parameters of the distribution functions of the EGDF type (subsection 15.2.5) for each of the 5 columns of the matrix M_n (ie for $m = 1, \ldots, 5$):
   (a) The global scale parameter $S_{n,m}$ (see subsection 16.2.1).
   (b) The lower and upper bound ($LB_{n,m}$ and $UB_{n,m}$) of the data support (subsection 15.2.2).
3. Perform column-wise filtering of all the ratios' probabilities $Pr\{R_{n,m,k}\}$ by calculating estimates of probabilities $\widetilde{Pr}\{R_{n,m,k}\} = EGDF(R_{n,m,k})$ for all $m = 1, \ldots, 5$ and $k = 1, \ldots, 88$ using the estimated parameters.
4. Using the gnostic robust method, estimate the coefficients $C_{n,0}, \ldots, C_{n,4}$ of the n-th cross-section by solving equation system 24.4.
5. Substitute the coefficients into 24.4 to compute the probable values of $EBIT/TA$ denoted $\tilde{M}_{n,k}$ estimated by the model[3].
6. Order the estimates $\tilde{M}_{n,k}$ in an ascending manner to determine the score $Sc_{n,k}$ of the k-th firm as its order number. Express the relative score in per cent form, $Sc_{n,k}\% := 100 * Sc_{n,k}/K$, where K is the total number of firms and equals 88.

B Repeat step **A** for all cross-sections (all quarters $n = 1, \ldots, 44$ from 90Q2 through 01Q1)[4].

To interpret the results of the ordering (scores, MD-rating) properly, it is necessary to take into account, that a higher score does not yet necessarily

[2] Probabilities were estimated by means of the EGDF (Estimating Global Distribution Function, see section 15.2 of Chapter 15), which filters the data of the regression system matrix column-wise, then a robust method was used to solve the system 24.4, which introduced a row-wise filtering effect.

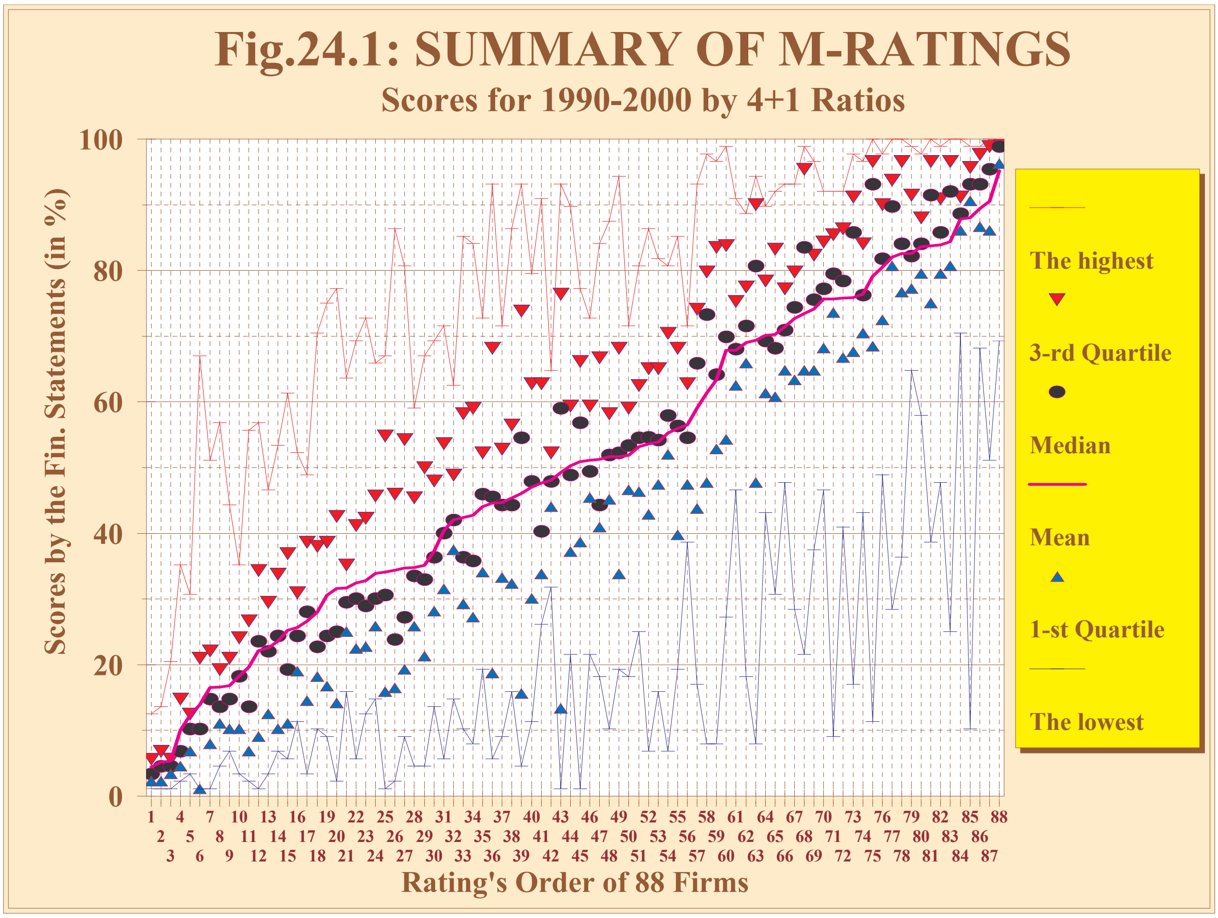
[3] Note the difference between $\widetilde{Pr}\{EBIT/TA_{n,k}\}$ and $\tilde{M}_{n,k}$. The former quantity results from the application of operation A.3, it is the value of the distribution function of the **actual** returns. In contrast, the latter variable is obtained from the model, it estimates "what should the return be from the point of view of the whole cross-section."

[4] The symbol $XYQZ$ defines "the Z-th quarter of the year 19XY". 01Q1 is the first quarter of 2001.

mean ‘a better financial position.’ Such an implication would hold only for economically comparable firms, ie for firms, for which their important financial variables fit the same multidimensional models. However, the homogeneity of the cross-section (ie the comparability of all its members) is not generally assumed, when proceeding with the ordering. The final evaluation of the financial positions requires, that **both** multidimensional ordering **and** a selection of comparable firms be performed. This second task for the cross-section analysis (the multidimensional cluster analysis) will be examined in Section 24.3.

24.1.3 Examples of the MD-rating

This ordering, performed on 44 cross-sections of 88 firms, resulted in 44 sets of 88 scores. Before considering the history of MD-ratings for individual firms, it is instructive to examine a summary of the average histories shown in Fig. 24.1.



Each of the 88 firms is attached to a point on the horizontal axis determined by its mean score over the entire time period. On the vertical axis, six symbols show the scores' statistics characterizing the distribution (EGDF) of the ratings obtained for each firm during the whole period being considered:

1. the lowest rating reached (the smallest score, the blue line),
2. the lower quartile of all the 'historical' scores (the blue triangles),
3. the arithmetical mean of scores (the magenta line, which was used to position the firms along the horizontal axis),
4. median of all scores (black circles),
5. the higher quartile (the red triangles),
6. the highest rating reached (the maximal score, the red line).

The scores and their statistics are in per cent form, so that the smallest is $1 * 100/88\%$ and the highest is 100%. The following conclusions can be drawn from Fig. 24.1:

- The approach discriminates sensitively between the lower and the higher MD-ratings: if the results were purely random, then all 88 averages of the 44 'scores' would closely approach 50%. However, the smallest average score obtained was that of Comp.B, equal to 4.4% (spread from 1.1% through a maximum of 12.5%). The highest long-term mean of scores (Comp.A) was 95.2%.
- The half of all scores falling between the first and third quartiles is mostly concentrated around the median and their spread is not very wide.
- The other half of the scores reaches both very low and very high values: the financial situation of any individual firm can deviate substantially from its average position.
- Even firms with very good average financial positions sometimes decrease to very low values: eg firm with an average performance of 88% fell once to 10.2% before recovering.

24.1.4 Monitoring the MD-rating

Multidimensional cross-section ordering enables the relative financial situation of a firm to be monitored and timely warnings of possible danger to be generated. To demonstrate this, the 'financial history' lines of three firms out of the 88 are depicted in Fig. 24.2R:

Comp.C: Performed well until 91Q1 then oscillated around 70% for about

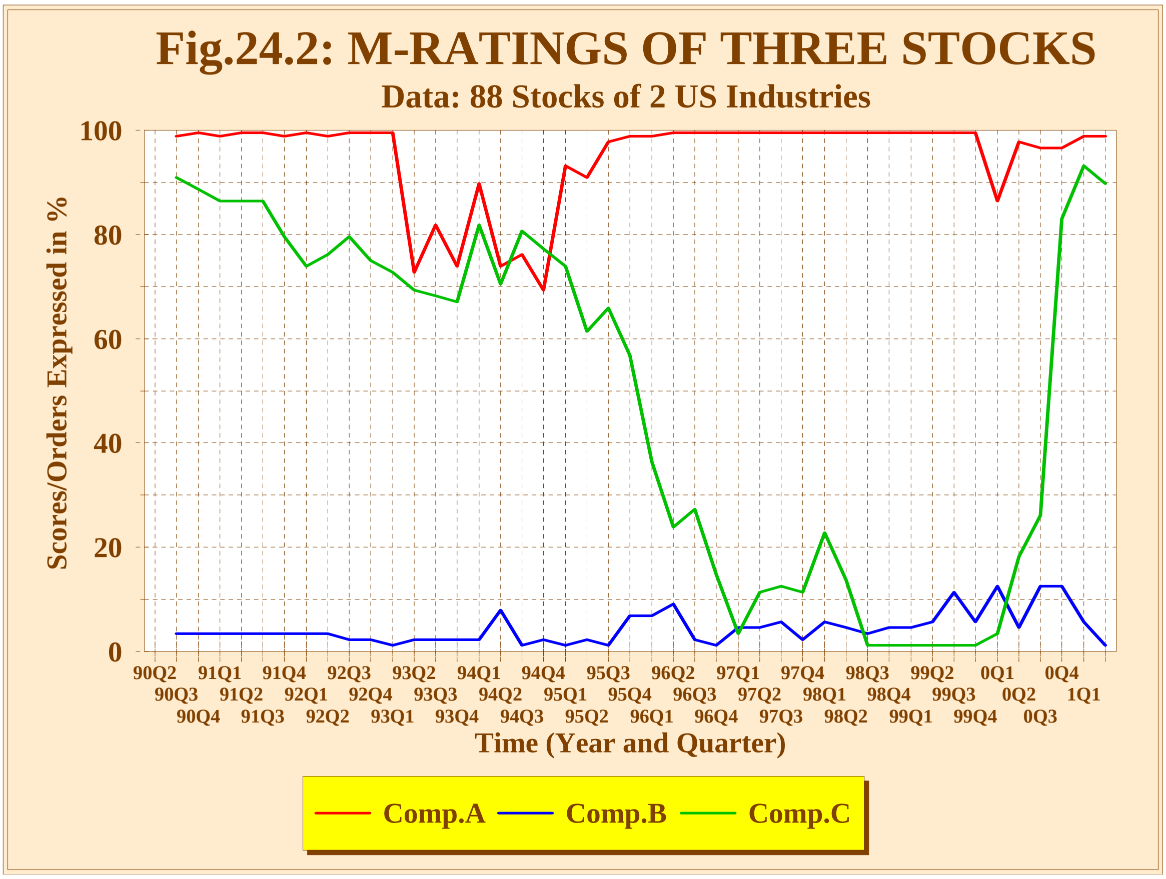


Fig.24.2: M-RATINGS OF THREE STOCKS

Data: 88 Stocks of 2 US Industries

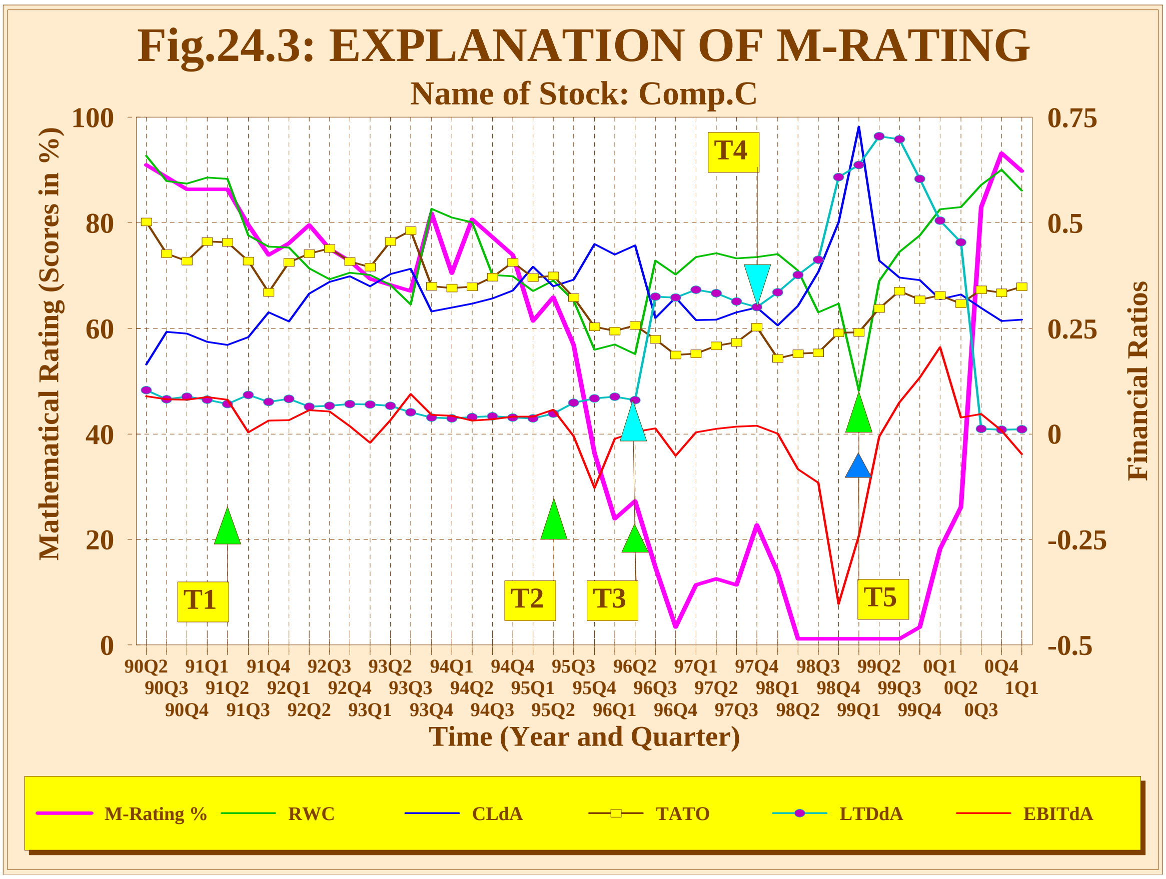
four years before falling steeply to completely recover from the lowest level over 3 quarters starting after 99Q4.

Comp.A: Dominating for 11 quarters at near 100% until 92Q4, when it fell to about 80% for two years to return to its dominant position, which lasted essentially for the rest of the period.

Comp.B: Maintained a long-term low position with only weak gradual improvement.

The goal of the analysis is not only to reveal, **that** something occurred, but also determine, **why** it happened: what was the cause of the effect. As an example of such an explanatory role for the analysis, the case of Comp.B can be used (Fig. 24.3).

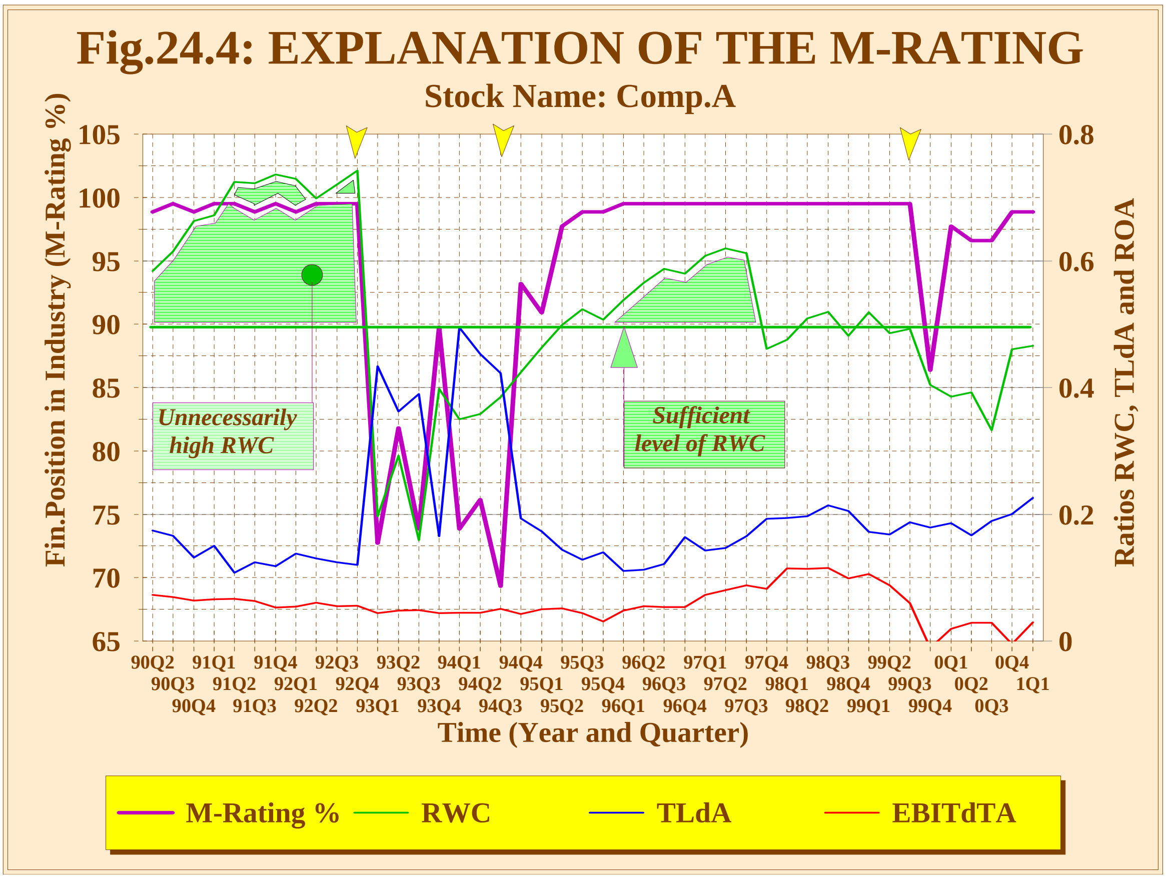
The critical points in time are marked by flags with arrows: At T1 the decline in performance started with a decrease in relative working capital RWC due to rising current liabilities CL/TA. The curve of the MD-rating continued to closely follow that of the working capital until the break point T2, where a sharp decline in RWC pulled the MD-rating under 40%. At 95Q4 the level of working capital is quite low and the overall valuation



continues to fall. At T3 (96Q2) another negative shock (a sudden increase in long-term debt (LTD/TA) pushes the valuation down to the bottom. However, increase in working capital helps to gradually return it to over 20% (T4). At 97Q4, a real financial stress can be seen to be developing by the simultaneous increases in both current and long-term debt accompanied by a deep slump in the return $EBIT/TA$. Another break-point (T5) opens a phase of a fast recovery: decreasing CL/TA accompanied by substantial increases in working capital helps the $EBIT/TA$ to return to black numbers. Increasing return enables working capital to further increase and the long-term debt to fall to the level it had held during the 'better' years before the end of 1994. All the other ratios also return to their 'good old' values and the MD-rating reflects this comeback.

This example shows how strongly the working capital and financial leverage affect the performance of a firm. This finding can be confirmed using the example of Comp.A (Fig. 24.4): its dominating position within the cross-section over the period 90Q2 through 92Q4 and 95Q4 through 99Q3 coincides with periods of a low financial leverage TL/TA (under 0.2) and a high level of the working capital (RWC over 0.5).

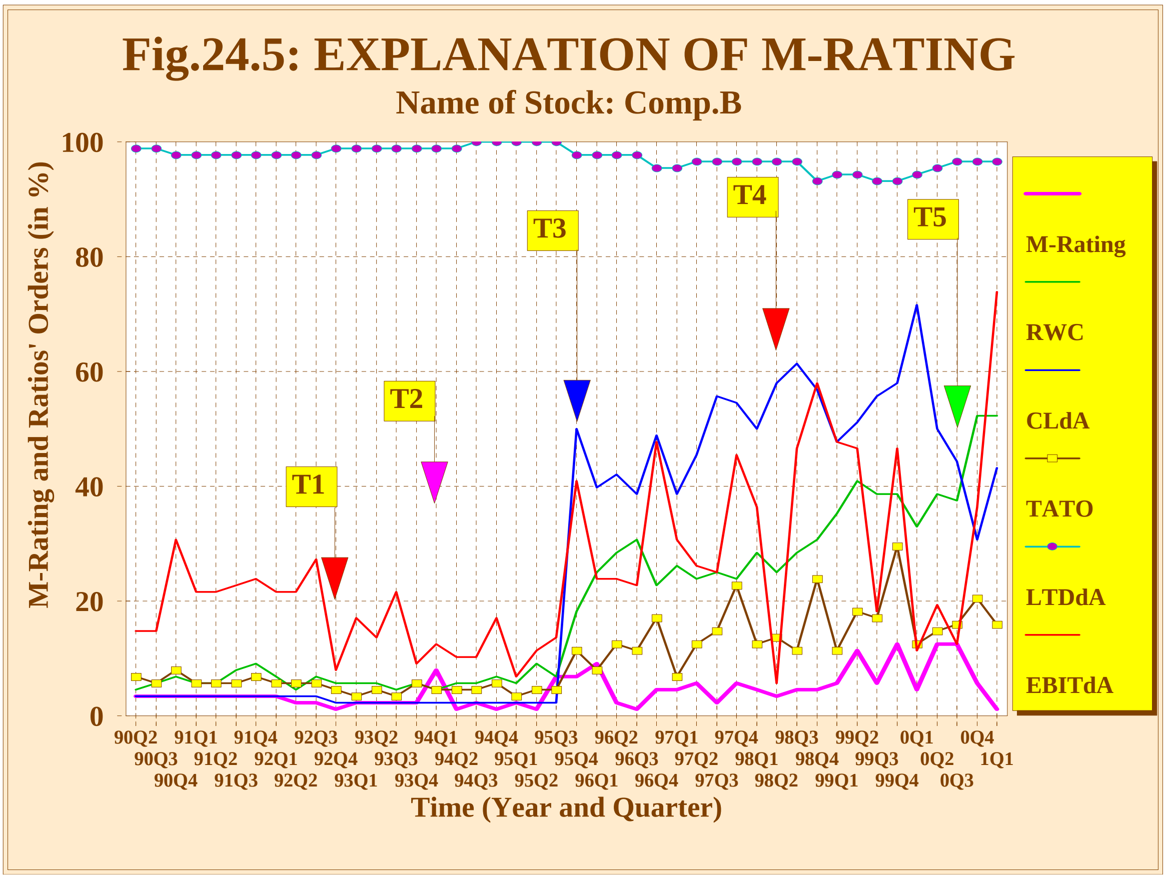
Fig.24.4: EXPLANATION OF THE M-RATING



The temporal loss of Comp.A's leading position (92Q4 through 95Q4) and after 99Q3 corresponds to a sharp fall in RWC and a striking increase in TL/TA. Moreover, as emphasized in Fig. 24.4 by green fields, the graphs lead to a determination (for this firm) as to the 'necessary and sufficient' level of working capital: there were periods (1. before 92Q4 and 2. 96Q1 through 97Q3) of unnecessarily high RWC, which exceeded the level about 0.5 sufficient to maintain a rating of 100% over the quarters 97Q4–99Q3.

The low MD-rating of Comp.B also requires an explanation, which is presented in Fig. 24.5, where order numbers (in %) of the main ratios are plotted together with the graph of the MD-rating (magenta line).

The main cause of the low valuation is seen at first sight: the long-term debt's order number was close to or even equal to 100% over the whole time interval considered. (Remember the permanently negative (and strong) impact of the LTD/TA, that has previously been noted). However, in spite of this prevailing effect, the graphs in Fig. 24.5 reveal two entirely different phases of Comp.B's financial history: until 95Q3, there were no signs of a tendency toward improvement, $EBIT/TA$ gradually fell. The



temporal changes of the order numbers' patterns observed at T1 and T2 can be explained by actions of the majority of the cross-section rather than by changes in policy instituted by Comp.B's management. However, at T3 (95Q4) the situation suddenly and significantly changed: a strong increase in $EBIT/TA$ enabled to start increase in the working capital RWC in spite of suddenly grown current liability CL/TA. This initiated an acceleration of the total assets turnover $TATO$ and a modest gradual improvement in the MD-rating.

The graphs reflect a strong year periodicity probably caused by the operating cycle, but the overall progress, reflected by the rising MD-rating, is obvious. (The sharp slump of $EBIT/TA$ in 98Q2 (T4) was corrected in 98Q3 and 98Q4). At T5 (0Q3) a sharp increase in $EBIT/TA$ was probably initiated by a rapid decrease in current liabilities CL/TA and by a positive step up in working capital RWC, but the MD-rating did not respond and began to fall: the extreme financial leverage and high current liabilities together with the still low activity ratio ($TATO$) did not allow Comp.B to keep up with the overall performance of the other members of the cross-section.

This example demonstrates the usefulness of the MD-rating: it is sensitive enough to recognize changes in the relative financial situation of a firm in a timely manner and to attract a manager's attention to apparent causes of the changes.

24.1.5 Impacts and Contributions of Explanatory Variables

To quantify the roles of the individual ratios in creating the return $EBIT/TA$, relations of the type

$$\frac{\partial(AVG(Pr\{EBIT/TA_q\}))}{\partial(AVG((Pr\{R_{m,q}\}))} = C_{m,q} \tag{24.5}$$

can be employed using the results from 24.4 after the equations have been averaged over all k. Indeed, the equation system is linear and therefore it can relate not only the probabilities of the ratios of the m-type (eg $R_{m,q,k}$ for the k-th individual firm in the q-th quarter), but also the (arithmetic) averages of these probabilities over all 88 firms. The roles of the model coefficient $C_{m,q}$ is therefore obvious: it evaluates the *partial impact* of the mean of the ratio's $R_{m,q,k}$ probabilities on the probability of the return $EBIT/TA_q$ in the q-th quarter. These impacts/sensitivities change depending on q, because there is a separate cross-section model for each of the quarters.

These partial impacts can be used to introduce some other characteristics, which demonstrate the influence of individual ratios in balancing equations 24.4, these are the *partial contribution* and the *mean contribution* of the m-th explanatory variable. Let the general form of the linear regression model 24.4 be

$$C_{q,0} + \sum_{m=1}^{M} C_{q,m} Ex_{q,m,k} = De_{q,k}, \tag{24.6}$$

where $C_{q,0}$ through $C_{q,M}$ form the q-th model, $Ex_{q,1,k}$ through Exq, M, k are values of the explanatory variables of the k-th object (eg probabilities of the ratios of the k-th firm) and $De_{q,k}$ are values of the 'dependent' variable of this object (which equals 1 in the case of an implicit regression). The partial contribution of the m-th explanatory variable of the k-th object on the $De_{q,k}$ is then

$$co_{q,m,k} = C_{q,m} Ex_{q,m,k} \tag{24.7}$$

and the mean contribution of all K objects (eg firms) is

$$\overline{co}_{q,m} = C_{q,m} * \frac{\sum_k^K Ex_{q,m,k}}{K}. \tag{24.8}$$

These characteristics are applicable to both implicit and explicit linear regression models. An application using these notions will illustrate their utility.

Consider once more the time series of 44 quarterly cross-sections (30 companies in the US Chemical Industry and 58 firms in different sectors of the US Information Technology Industry). Apply the robust method to the 44 systems of 88 explicit linear regression models in probabilities

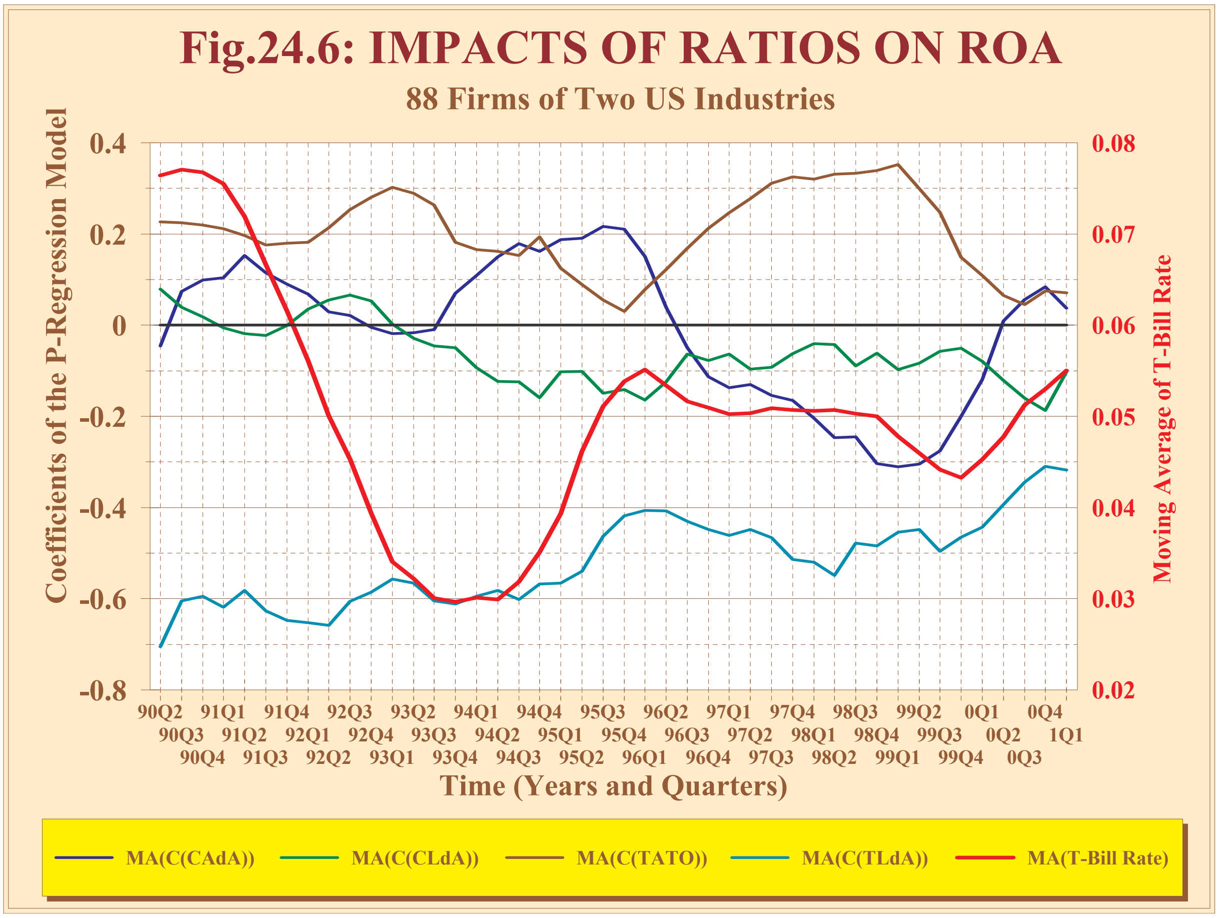
$$\begin{aligned} C_{q,0} + C_{q,1} * Pr\{CA/TA_{q,k}\} + C_{q,2} * Pr\{CL/TA_{q,k}\} \\ +C_{q,3} * Pr\{TATO_{q,k}\} + C_{q,4} * Pr\{TL/TA_{q,k}\} \quad &= \\ Pr\{EBIT/TA_{q,k}\} \end{aligned} \tag{24.9}$$

for $q = 1, \ldots, 44$ and $k = 1, \ldots, 88$. The time series of the model coefficients (ie of the partial impacts of the ratio's probabilities on $EBIT/TA$'s probability, shortly 'impacts') are depicted in Fig. 24.6 together with the T-Bill Rate.

To suppress the quarterly volatility, all five curves are smoothed by a four quarter moving average. Coefficients are labeled, so that eg $MA(C(TATO))$ is 'the moving average of the coefficient $C_{3,q}$ of equation 24.9 at the q-th quarter'. (Indexes 1,. . . ,4 identify ratios CA/TA, CL/TA, $TATO$ and TL/TA and the averaging period is: $q-3$ through q.)

The graphs in Fig. 24.6 infer the following:

- The strong sensitivity of a firm's financial position to working capital ($RWC = CA/TA - CL/TA$) observed in Figs. 24.3 and 24.4 was mainly due to changes in current assets CA/TA (the dark blue line), because the current liabilities had only a weak and relatively small impact as shown by the green line in Fig. 24.6.
- The strongest (and always negative) impact on the return $EBIT/TA$ was that of the financial leverage TL/TA (the light blue line). The long-term trend of this impact was remarkable: the coefficient $C_{4,q}$ began under -0.6 and oscillated around this value until 94Q4 to eventually rise to -0.3 after 0Q4.
- There are strong similarities between the graphs of the T-Bill Rates and the impact of the current assets $C_{1,q}$. This can be interpreted as



"there is an interdependence between the price of short-term money and its impact on returns." An examination of the turning points in both curves provides a good indication of the direction of the cause/effect: changes in $C_{1,q}$ occur earlier (as if the statement, "the FED carefully keeps its finger on industry's pulse to adjust—with a delay of several quarters—the cost" was confirmed). The curves also show, that a T-Bill Rate of about 0.05 is too high from the point of view of industry (causing the negative partial impact of the current assets on the return), but a decrease to around 0.045 is sufficient to spur a recovery.

The graphs in Fig. 24.6 reflect the time series of quarterly cross-sections (88 firms). It can also be useful to have a look at the long-term means and standard deviations of these impacts. These are shown in Tab. 24.1.

Mean Contribution	**Ratio**	AVG	STD
$\overline{C_0}$	Intercept	-0.006	0.007
$\overline{C_1 * Pr\{CA/TA\}}$	CA/TA	-0.012	0.176
$\overline{C_2 * Pr\{CL/TA\}}$	CL/TA	-0.065	0.103
$\overline{C_3 * Pr\{TATO\}}$	$TATO$	0.197	0.112
$\overline{C_4 * Pr\{TL/TA\}}$	TL/TA	-0.508	0.119

Tab. 24.1: Long-term Mean Values of Contributions of Ratios to the Return's Probability and their Standard Deviations (STD).

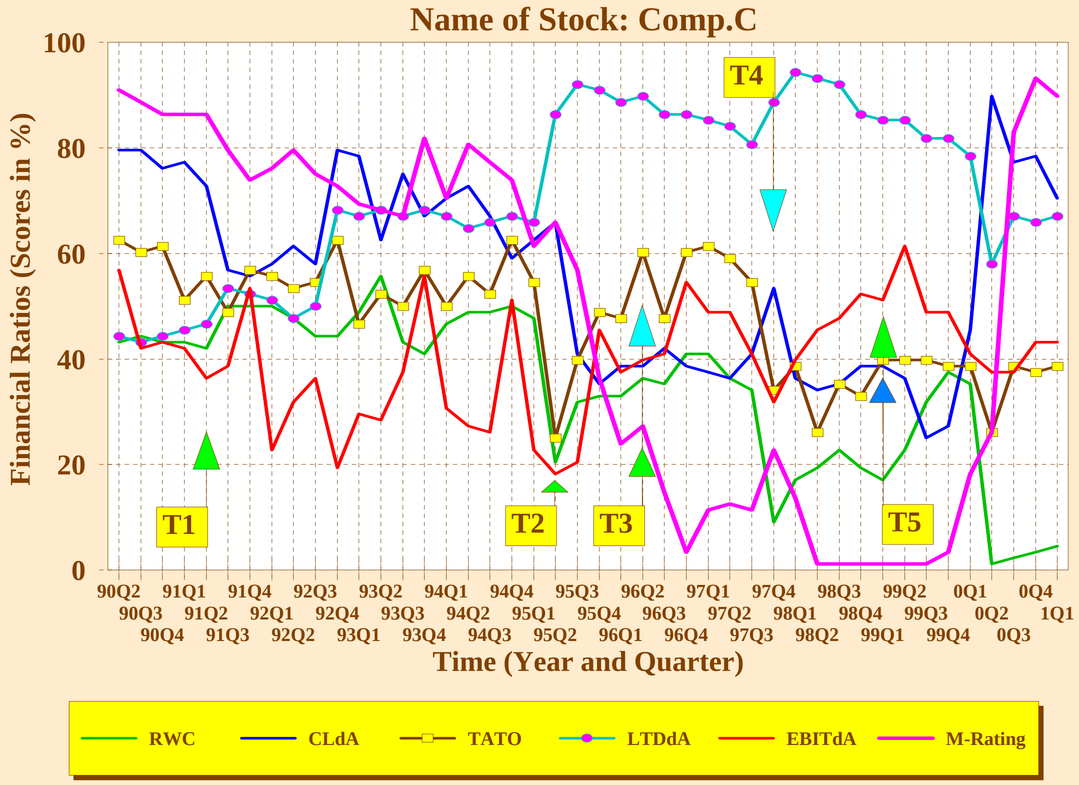
The symbol $\overline{C_0}$ denotes the mean contribution of the intercept (coefficient $C_{q,0}$) to the probability of the return in 24.4. The long-term evaluation summarized in Tab. 24.1 is instructive:

1. The intercept's contribution to the probability of the return's values is practically negligible. This is a good finding, because a substantial intercept ordinarily results from having neglected the non-linearity of the relation and/or from a poor choice of explanatory variables. It could be shown, that linear regression of the ratios would yield a much less acceptable model. This confirms the superiority of the regression in probabilities, at least in this application.
2. As shown in Fig. 24.6, current assets can affect the return both positively or negatively especially depending on the price of short-term money. However, the long-term mean effect (Tab. 24.1) of CA/TA is slightly negative. Its volatility (measured by the STD) is high, but this is caused rather by the slow, but large, oscillations (probably due to changes in the cost of short-term money) rather than by a random structure for this ratio.
3. The far strongest long-term negative impact on return is from financial leverage, TL/TA.
4. Although useful in completing the over-all picture, the long-term mean values of the models' coefficients do not provide an analyst with dynamic information comparable to that shown by Fig. 24.1 through Fig. 24.6.

24.1.6 Multi-marginal Ordering?

The close correspondence of the MD-rating (multidimensional ordering) to the behavior of the individual ratios (observed in Fig. 24.3 and 24.4) might suggest, that such a complex procedure is superfluous, and that the financial position can be simply evaluated by ordering several individual ratios, ie by “multi-marginal analysis.” Figure 24.7 illustrates such an attempt, again using the Comp.C: instead of the values of the ratios as in Fig. 24.3, here, the ratios’ order numbers are used to reveal their relative position within the cross-section.

Fig.24.7: MULTI-MARGINAL ORDERING



The critical points, T1–T5, are also plotted as before. A comparison of the two graphs leads to the following conclusions:

Similarities: A certain similarity of forms is observed between the graphs of the MD-rating (Fig. 24.3) and of the CL/TA’s score (Fig. 24.7) although there is no similarity between the values of CL/TA and the MD-rating in Fig. 24.3. Only the critical points T2 and T4 in Fig. 24.7 attract attention to the fact, that striking changes are in progress in the two processes.

Dissimilarities: There is no clear resemblance between the forms of the graphs of the ratios’ values in Fig. 24.3 and their order numbers in Fig. 24.7. Changes are sometimes in the opposite direction.

Valuation: It would be very difficult to arrive at an overall valuation of the financial situation using only the multi-marginal view of

Fig. 24.7, while the MD-rating allows the three distinctly different phases (90Q2–94Q2, 94Q3–99Q3, 99Q4–1Q1) to be easily recognized in Fig. 24.3.

Such a negative evaluation of the ability to achieve an overall rating by multi-marginal analysis does not mean, that this procedure is not useful. Table 24.2 lists the arithmetic means of scores over the period 90Q2–01Q1 (in per cent form) for the three firms, Comp.B, Comp.C and Comp.A.

	Firms					
	Comp.B		**Comp.C**		**Comp.A**	
Ratio	AVG	STD	AVG	STD	AVG	STD
CA/TA	14.8	12.8	33.0	15.0	31.2	9.2
CL/TA	26.2	24.3	56.1	17.7	61.1	16.7
$TATO$	10.2	6.2	48.0	10.7	56.0	13.0
RWC	19.1	14.4	34.5	14.7	30.3	15.2
$EBIT/TA$	25.4	15.1	39.9	10.7	32.4	17.2
LTD/TA	97.3	1.9	71.4	16.0	73.4	10.6
$TOTR$	48.8	25.9	47.9	18.0	46.2	28.3
EP	59.0	27.0	61.3	18.7	58.9	10.6

Tab. 24.2: Marginal Ordering of Financial Ratios of Three Firms

Notation:AVG …Arithmetic Means of Scores, STD …Standard Deviation of Scores (while 'scores' are order numbers expressed in per cent form.)

The first six lines of Tab. 24.2 represent the 'inner' view based only on the financial statement data, while the last two are 'outer' valuations obtained by ordering the market's data (Total Returns and Earning/Price ratios). The most striking result from the first group is, that $LTD/TA = 97.3 \pm 1.9$[5] says, that Comp.B's financial leverage was (with a small STD) nearly always very close to the top of all 88 firms forming the series of cross-sections. All the other ratios: CA/TA, CL/TA, $TATO$, RWC and $EBIT/TA$ were significantly lower for this firm than the cross-section means (50%). This explains, why Comp.B's financial position, evaluated by the MD-rating, was so low. On the other hand: this result demonstrates, that results of the 'automatic' mathematical rating cannot be taken as absolutely valid. The lowest MD-rating does not necessarily mean, "Comp.B's financial position is consistently the worst of all 88 firms," because alternative interpretation of the results can exist:

[5] The notation is $AVG \pm STD$, ie the arithmetic mean plus or minus the standard deviation.

"Comp.B is not comparable with the majority of the 88 firms, because it is substantially different from them in the sense of availability of credits." Similar results were demonstrated in Chapter 23, where some firms, eg Procter & Gamble and Colgate-Palmolive, were shown to be performing very well with much larger financial leverage and much lower working capital than the others. The problem of the comparability of firms thus arises and as previously noted, will be examined in Section 24.3.

On the other hand, the joint ordering of all firms is reasonable even though some firms are not quite comparable to each other: changes in financial position within the cross-section signal, that "something happened with respect to the firm's financing." Such a message can serve as a useful warning to the financial manager. If this type of cross-section analysis were included into the firm's information system as a routine function, such a warning together with knowledge of the actual model of the cross-section could serve as an efficient tool for financial management.

The results summarized in Tab. 24.2 lead to other observations:

- None of Comp.A's mean ratios approaches extreme values (0 or 100%). In other words, a candidate for the highest MD-rating does not need to have ratios with extreme values. Rather, "golden means" and an "optimal mix" of ratios are preferable.
- Multi-marginal analysis based on the long-term means of ratios cannot distinguish Comp.C's entirely different policies from those of Comp.A (which has the highest evaluation), although it can recognize the extreme behavior of Comp.B. Monitoring the time-series of MD-ratings (such as Fig. 24.1) is thus desirable.
- The market evaluation using $TOTR$ and EP cannot separate the highest, Comp.A, from the lowest, Comp.B. The average scores of all three firms under comparison are close to the mean and their variance is too large to make any valid decision as to differences between firms.

The conclusion is, that multi-marginal analysis is insufficient by itself to give a reliable rating to firms although it can provide analysts with additional useful insight into the data.

24.1.7 Financial Market Generates Chaos

Figures 24.1–24.6 presented a robust multidimensional ordering within cross-sections formed by firms in two industries, which lead to useful insights into the development of the financial positions of firms. Note, that these results are based entirely on the accounting data, ie they represent an "inner" point of view. Only the T-Bill Rate, as an "external" factor, was used to characterize its impact on the financial position of firms. Market impact was reflected only indirectly, by the total asset turnover $TATO$. Market valuation factors such as $TOTR$ and E/P did not enter into the model. Efficient market theory postulates, that the market absorbs all information both directly and indirectly related to the economy, using this to establish valuation for stocks and firms. The accounting data are only a part of this information and an important question to be asked is: "How far–if at all–can the use of market valuation improve the 'inner' valuation?" The high volatility of the market indicators in Tab. 24.2 does not foretell great success in this respect but the problem deserves a more detailed analysis.

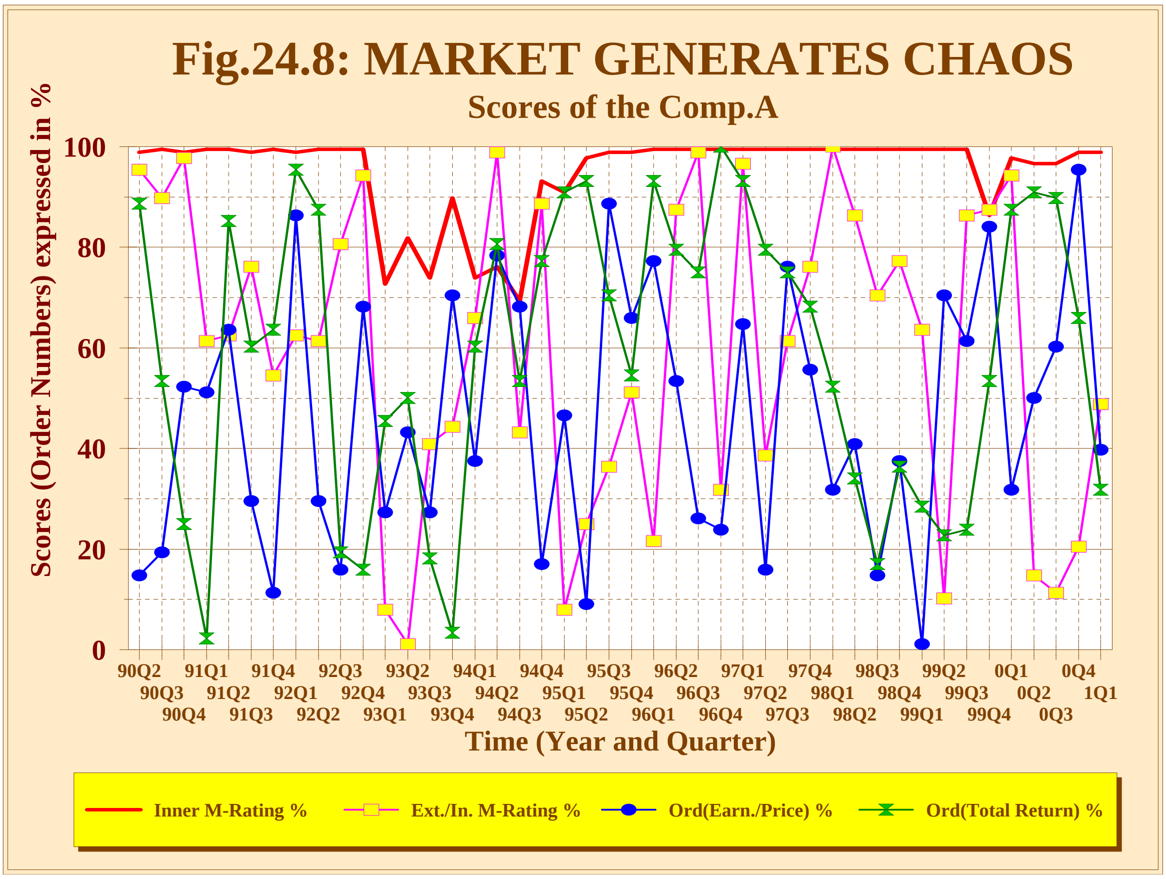
The "external" (market) valuation of firms represented by the total return ($TOTR$) can be taken in account by extending the model 24.4 by this variable:

$$\begin{aligned} C_{q,0} + C_{q,1} * Pr\{CA/TA_{q,k}\} + C_{q,2} * Pr\{CL/TA_{q,k}\} & \\ +C_{q,3} * Pr\{TATO_{q,k}\} + C_{q,4} * Pr\{TL/TA_{q.k}\} & \\ +C_{q,5} * Pr\{EBIT/TA_{q,k}\} &= Pr\{TOTR_{q,k}\}, \end{aligned} \tag{24.10}$$

where again q identifies the year quarter and $k = 1, \ldots, 88$ the firm. Equation 24.10 characterizes the impact of probabilities of accounting ratios CA/TA, CL/TA, $TATO$ and $EBIT/TA$ on the probability of the market indicator $TOTR$. The relationship will be analyzed in the same way as before, ie using multivariate ordering. The results for one firm are presented in Fig. 24.8: The heavy red line is the same as in Figs. 24.2 and 24.4: the M-Rating of Comp.A obtained by using only accounting data (model 24.4).

The yellow squares connected by the thin magenta line show the results obtained by using model 24.10. The method is the same as in the case of 24.4 (robust multidimensional ordering), but inner (accounting) data were complemented by the market valuation ($TOTR$). Again using the notation $AVG \pm STD$ a long-term statistical comparison of the two time series of valuations can be made: while 95.2 ± 8.8 characterizes the quality of the purely inner valuation, the "mixed" outer/inner valuation was 59.8 ± 30.1. In other words, the former can, but the latter cannot be taken as significantly different from the general mean 50%.

The quarterly values of cross-section orders of $TOTR$ and E/P are also shown in Fig. 24.8, their statistics were presented in Tab. 24.2. Because of



the high volatility, it is impossible to judge, which of the three approaches using the market data is the best. As shown in Fig. 24.8, all three can sometimes concur as to the firm's quality (94Q2), while at other times they simultaneously disagree (1Q1), but most frequently, the three measures diverge.

Instead of improving the results of the inner valuation, the additional information that was hidden in the market data distorts the previous results. This rather disappointing outcome is not because the quality of the external information 'absorbed' by the market is low, but rather by the 'market psychology', which encourages traders to react to news, whatever the validity of the new developments may later turn out to be. The chaotic character of this 'shoot first and sort it out later' approach is the most likely source of the volatility of the market measures cited above. The development of techniques tailored to the efficient and rapid extraction of only **useful** information from the vast amount of data instantaneously available, is badly needed, so that the information treatment used by managers and traders for decision making can be improved. The robust MD-rating described and illustrated in this section could become the nucleus of a better approach to these problems than what is currently in use.

24.2 Monitoring a Single Firm's Economics

The multidimensional cross-section monitoring set out in the previous section allows the relative financial position of a firm to be compared with that of a number of other firms. The MD-rating "measures" the relative overall performance through time by the order number assigned to the firm by the process, but only in terms of, "over the past n quarters the performance of firm X was evaluated by a higher score than those given a lower order number." Therefore, the order number is not directly interpretable as an evaluation of the quality of performance such as: "the higher the order number the better financial position **now**;" further analysis is needed to establish, **why** this and not another MD-rating was obtained. The most useful outcome of this approach, when it is applied to a time-series of cross-section data, is a reliable indication, that the firm's financial position within the cross-section has **changed**. Information of this nature should attract the attention of the financial managers, lead to a search for the causes of the change, and to the taking of corrective measures.

Could robust multidimensional analysis be used to track the behavior of a single firm? Such a potentially useful approach has several drawbacks: it is insufficiently timely since cross-section accountancy data are available only on a quarterly basis. Effective financial management requires a continuous flow of information so as to be able to take immediate action. This is achievable within a firm, where up to date information systems register all pertinent information needed for financial control; the 'only' problem is to learn how to extract the relevant information from the rich data flows. There is also an alternative approach based on robust multidimensional models of the interdependence between financial ratios.

To show, that such a method could yield useful results, the quarterly data of Comp.C, the company that was previously used in the cross-section analysis (Fig. 24.3) will be used.

Let $R_{m,t}$ $(m = 1, \dots, M,\ t = 1, \dots, T)$ be a matrix, the columns of which are composed of M ratios measured at time t and let an integer be $L > M$. The following set of L equations is assumed to hold at each time $t \geq L$:

$$\sum_{m=1}^{M} k_{m,t} * R_{m,t-\tau} = 1, \qquad (24.11)$$

where $\tau = 0, \ldots, L-1$. The matrix R thus represents an M-dimensional time series of financial ratios bound to an implicit linear regression model $k_{m,t}$ with a zero intercept. This model is time-dependent, it is valid at the instant t for the 'moving average' (of length L) of R's columns. By applying a robust modeling method to each of the $T-L+1$ moving averages it is easy to obtain

1. a sequence of $T-L+1$ models $k_{t,m}$ ($t = L, \ldots, T$, $m = 0, \ldots, M$), ie of partial impacts,
2. a sequence of $M*(T-L+1)$ of the fractional contributions of the explanatory ratios to 1 (see 24.7),
3. a sequence of $L*(T-L+1)$ residuals (equation errors)

$$e_{t,\tau} = 1 - \sum_{m=1}^{M-1} k_{m,t} * R_{m,t-\tau}. \qquad (24.12)$$

From all the residuals, the value obtained when $\tau = 0$, is the most relevant, because it is the difference between 'what would be expected as a smooth continuation of the process' and 'what really resulted due to the current values of the explanatory variables.' Each sudden change in an explanatory variable is thus reflected by this residual. Figure 24.9 is an example of such residual monitoring for two firms, Comp.C and Comp.A.

According to relation 24.12, a positive model error signals 'the model yields less than what would be expected by the moving average'. As seen in Fig. 24.9R, negative residuals dominated in the case of Comp.C, while Comp.A has the opposite tendency. Comparing these results with Figures 24.3 and 24.4, it is seen, that large residuals occur especially when sudden changes in ratios take place after several 'quiet' periods (Comp.C at 97Q3 or 99Q4 or Comp.A at 94Q1 and 0Q1). To properly use Fig. 24.9, one must understand, that 'quiet' does not mean, that the ratios' values remained constant, but that the relations between the ratios did not change. This is because the effects of different ratios can compensate for each other's changes.

More sensitive signals of change can be derived from the contributions of individual explanatory variables. Indeed, the 'dependent' variable (the constant 1 in the case of implicit regression 24.11) is input from the explanatory variables ($k_{m,t} * R_{m,t-\tau}$). Its composition changes with changes in these variables (ratios $R_{m,t-\tau}$) and corresponding changes in the model coefficients $k_{m,t}$. An example of such changes in the structure of the equation's right-hand side can be seen in Fig. 24.10.

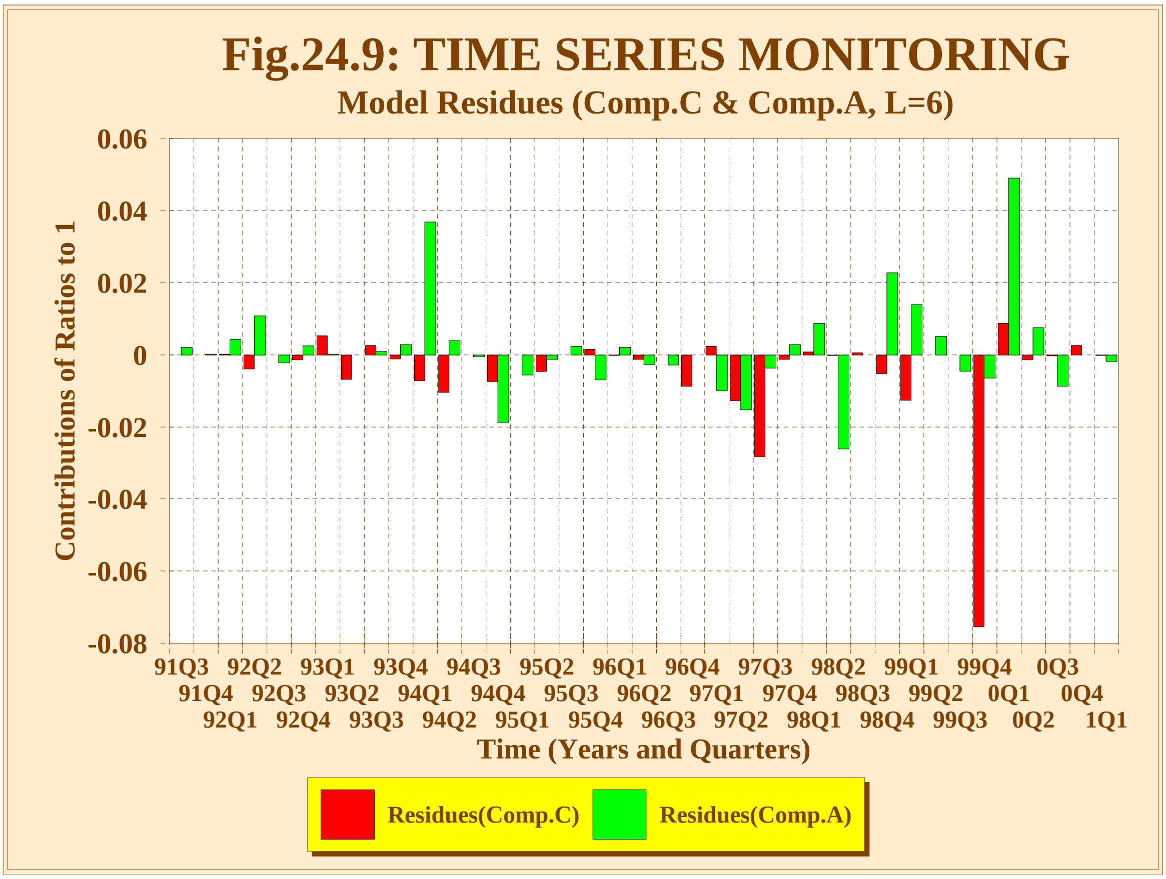


Fig.24.9: TIME SERIES MONITORING

Model Residues (Comp.C & Comp.A, L=6)

The sum of the contributions of all the explanatory variables/ratios always equals 1 (24.11), but the make up of these contributions will change over time. Because the sum must be 1, large positive impacts from a ratio must be compensated by a strong negative impact in another ratio or ratios. The total length of the columns in Fig. 24.10 is determined by the sum of the absolute values of all the impacts. This length is close to 1 for relatively undisturbed cases (smooth changes in ratios), while sudden changes are manifested by a large increase in total length such as 92Q3, 93Q3, 94Q1–96Q3 and 99Q1–0Q2. A comparison with the graphs of the ratios in Fig. 24.3 reveals, that tall columns properly signal real changes in the ratios. Moreover, the changing structure of the mean contributions is reflected by the color patterns in Fig. 24.3, which in some cases show the exchange of roles between individual ratios: a positive contribution becomes negative and vice versa.

There also is an alternative to monitoring the structure of the impacts shown in Fig. 24.10 based on a geometric idea: the M values of the contributions at a moment t define a point $(p(t))$ in an M-dimensional space, which moves at the next instant to $p(t+1)$. An evaluation of the changes

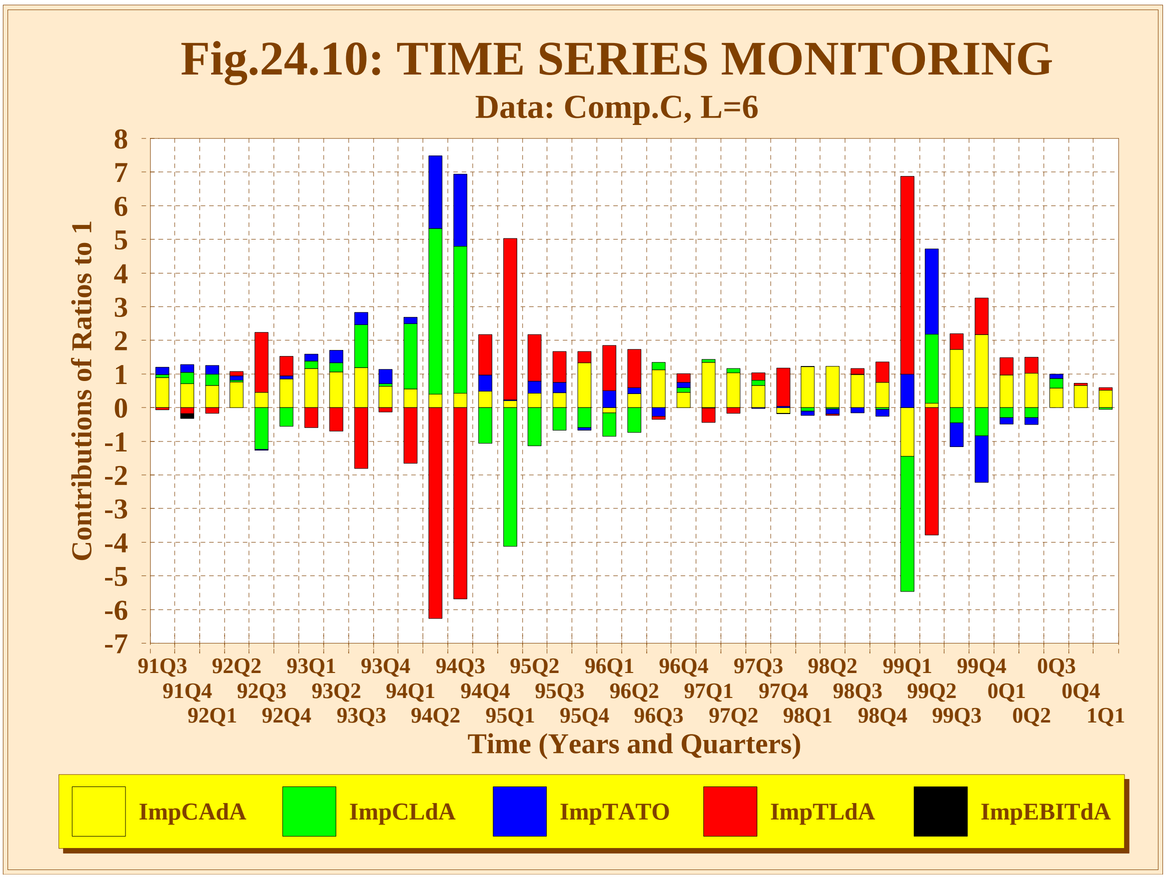
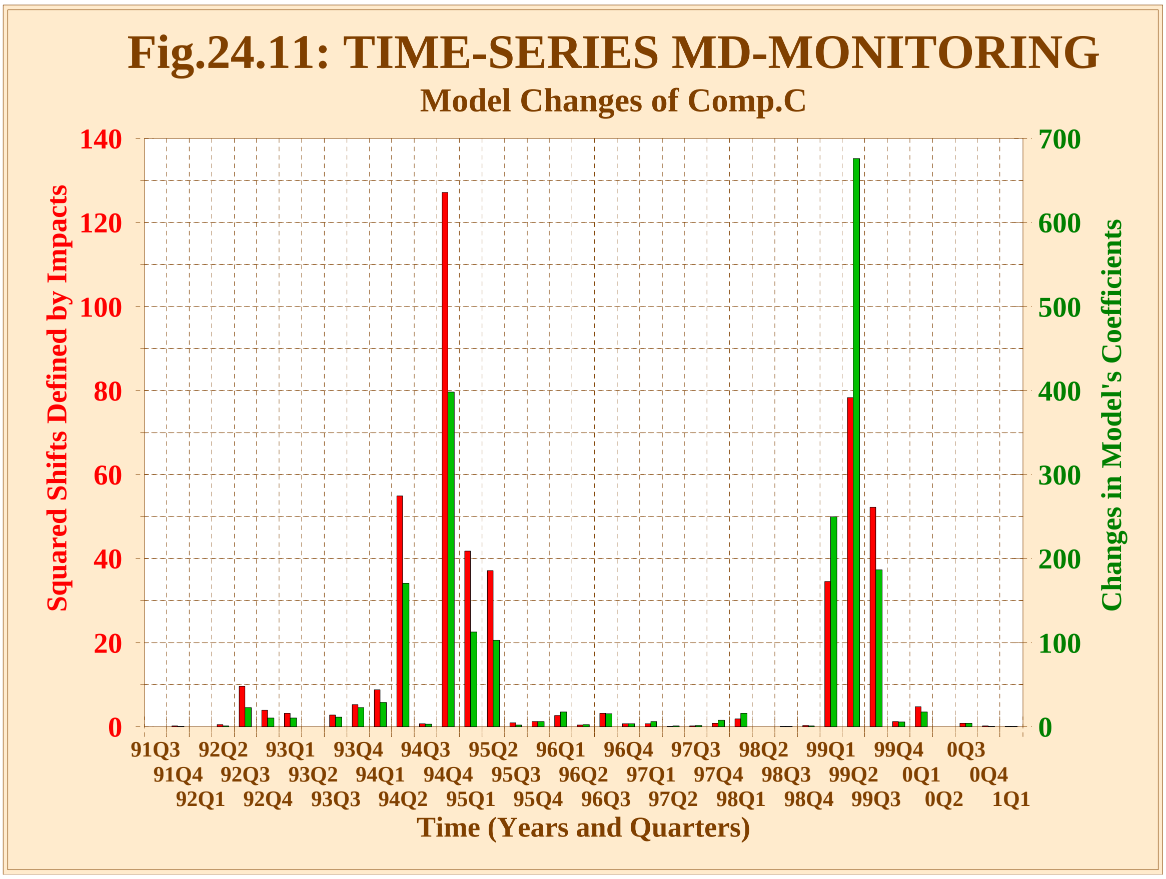


Fig.24.10: TIME SERIES MONITORING

in the contributions' structure can be obtained by calculating the distance between points $p(t)$ and $p(t+1)$. Of course, it is necessary to choose a geometry, but in this case, it is a simple decision: Euclidean geometry is suitable, because the bad robustness of its scalar product is advantageous here; it helps to detect deviations from the smooth path of the representative point caused by the 'outlying' behavior of a ratio. The red columns in Fig. 24.11 illustrate the use of the squared distances (sums of changes in contributions) in monitoring the 5-dimensional time series of ratios of Comp.C.

These can be compared with the sums of squared changes in the model coefficients (partial impacts), which are also shown in Fig. 24.11 (green columns). The information provided here differs from the previous views due to its different sensitivity to disturbances (compare periods 94Q2–95Q2 with 99Q1–99Q3). The changes in the model's coefficients play a role in both cases, reflecting changes in ratios indirectly, however changes in the contributions are also directly reflected in the ratio's values. A question of the type 'which is better' can be answered from the point of view of practice with 'both of them.' However, there is one theoretical aspect favorable to



contributions: they are expressed in a dimension-less manner and all are directly comparable as parts of the same whole (of 1 in the implicit case).

The general conclusion from the examples in Figures 24.9–24.11 is, that robust multidimensional models applied to a time series of ratios of a single firm can provide valuable information on the changes in the financial situation of the firm. This information is not in conflict with that obtained by cross-section analysis (the MD-rating), which requires data on all firms in the cross-section to be available. When these latter results are evaluated, it must be remembered, that the sporadic quarterly data series were used in both cross-section and time series analysis. For a firm's managerial team, much more frequent observations would be available, greatly enhancing the practical utility of the monitoring.

24.3 Explicit Multidimensional Clustering

24.3.1 Intra-industrial Clusters

Although marginal analysis yields useful information on the behavior of complex objects, there still is a substantial difference between multiple usage of uni-variate models and true multivariate models, because only the latter can lead to the characterization of the interdependence of all the variables. Multivariate analysis naturally involves the steps of marginal and pair analysis as well as specific multidimensional techniques. It is not surprising then, that the problems of data inhomogeneity and the robustness of the methods described on the level of partial analysis also have an impact on multivariate modeling. For example, recall Figures 23.12 and 23.13 showing through marginal analysis, that firms with different working capital and leverage policies fell into different clusters. It was then put forth, that this was probably because of different means of access to financial markets. It then follows, that a model of the dependence of relative working capital on financial leverage in these groups would also differ.

Assume, that the idea of "economically comparable firms (firm A $\sim$ firm B) behaving in accordance to the same multidimensional model" has been accepted. Recall, that inhomogeneity of working capital and financial leverage for firms of the chemical industry was shown in Chapter 23. Therefore the natural implication "marginal inhomogeneity $\Leftrightarrow$ multivariate inhomogeneity" infers, that not all firms belonging to the same industry are economically comparable, and that it cannot be expected, that a single model will be a universally good representation of all firms in an industry made up of several clusters, each represented by a different model. As shown in previous sections, a single model of all firms in the two US industries could be useful for a special purpose (particularly, for M-ordering of financial positions) although the possible inhomogeneity (and incomparability) of all the ensemble of firms was not analyzed. This shortcoming of the MD-rating provides only relative and approximative results, suitable especially for monitoring significant changes in the relative financial positions of firms. To make really reliable comparisons, a more detailed analysis is needed based on multidimensional clustering.

The problem that is faced is to decompose an inhomogeneous multidimensional data sample (eg of financial statement data of firms of given industries) into clusters of comparable firms (ie firms, which behave according to the same model).

Let a simple explicit (linear) model representing the mutual dependencies of several ratios such as the following be chosen for a group of K companies:

$$EBIT/TA_k = K_0 + K_1 * RWC_k + K_2 * TATO_k + K_3 * TL/TA_k, \quad (24.13)$$

where $k = 1, ..., K$ is the order number of the company listed alphabetically by ticker symbol, and $EBIT/TA$ represents earnings before interest and taxes (Compustat's data 178) divided by total assets (data 6). The variable RWC is the working capital divided by total assets 22.5, $TATO$ is the Total Asset Turnover (relative value of net sales, data12/data6) and TL/TA denotes the ratio of total liabilities to total assets (data18/data6). The choice of variables is one of several that could be made and it is motivated principally by the ability to interpret the meaning of the ratios directly:
$EBIT/TA$ ... the gross business return on assets,
RWC ... liquidity,
$TATO$ ... activity, and
TL/TA ... financial leverage.
Model 24.13 then describes the principal aspects of a firm's financial health.

It is possible, that there will be no useful statistical solution to the system of equations 24.13, when real data are used to identify the model's coefficients. This is illustrated with the previously used data set of the 54 companies in the US Chemical Industry for 1995. The coefficients of the model 24.13 estimated by the basic statistical Ordinary Least Squares method (OLS) are shown in Tab. 24.3 together with the main qualitative statistical characteristic of a linear model, the multiple *R-Square* (R^2). This statistic (sometimes called *the multivariate correlation coefficient*) evaluates the relative portion of the dependent variable's variance explained by the model. The model's precision is given by the fitting error, MAE (Mean Absolute Error), which is shown in the last column. It is preferred here to the standard deviation (STD), because the latter statistic is not suitable for models, which treat strongly uncertain data.

Model	**Coefficients**				**Quality**	
Method, Size	K_0	K_1	K_2	K_3	R^2	*Fitting MAE*
OLS, 54	0.142	0.00781	0.0320	-0.0804	0.141	0.0335
Gnost., 54	0.291	-0.142	-0.00860	-0.224	—	0.0336
Gnost., 32	0.2921	-0.143	-0.00857	-0.225	—	0.0109
Gnost., 10	0.271	-0.137	-0.00495	-0.232	—	0.0157
OLS, 10	0.271	-0.137	0.00505	-0.231	0.995	0.00151

Tab. 24.3 The search for a model of the main cluster 1

The first line of the table with the very low value of *R-Square* and the unacceptably large fitting error shows, that there would be no statistically acceptable model in the form of 24.13, which could explain the behavior of all 54 sets of ratios. However, this result does not exclude the existence of a subset (a cluster containing a smaller number of companies), for which a good model can be found. Companies, which behave in accordance to the same model are **comparable**—they react in the same way to the same changes in their "input" variables. A set of such companies is **homogeneous**. Accepting this point of view, the first line in Tab. 24.3 can be interpreted as a demonstration of the **inhomogeneity** of the Chemical Industry taken as a whole.

The second line is obtained, when the coefficients of the above model are estimated using robust gnostic methods. The MAE of the model is increased slightly due to the robustness of the procedure, which assigns smaller weights than OLS to outlying points, while emphasizing the "central" data in the most dense region of the multidimensional "cloud." This is shown in Fig. 24.12 by the blue line, which depicts the global view of the whole cloud's density (the density of the global distribution function (EGDF) of the 54 modeling errors). The range of the modeling errors is broad with a minimum of -0.098 and a maximum of 0.186. A more detailed structure for these errors is obtained by means of the local distribution function (ELDF), which is shown by the red line. The effect of the method's robustness is clearly seen: the model emphasizes the most dense area of the cloud, which contains 32 companies, cluster "A". An *R-Square* is not computed, because the modeling method is not a statistical regression. The notion of statistical variance loses its meaning, when data with gross errors are treated.

The second step is to identify the 32 companies, which make up cluster "A", to extract them from the initial data set, and to find their robust model (line 3 in Tab. 24.3). The density of errors for this model is shown by the magenta line. Note, that the MAE decreases significantly from 0.036 to 0.0109. It is also seen in Fig. 24.12, that the ELDF of the 32 "kernel" companies has a narrow central cluster ("B") of 10 companies.

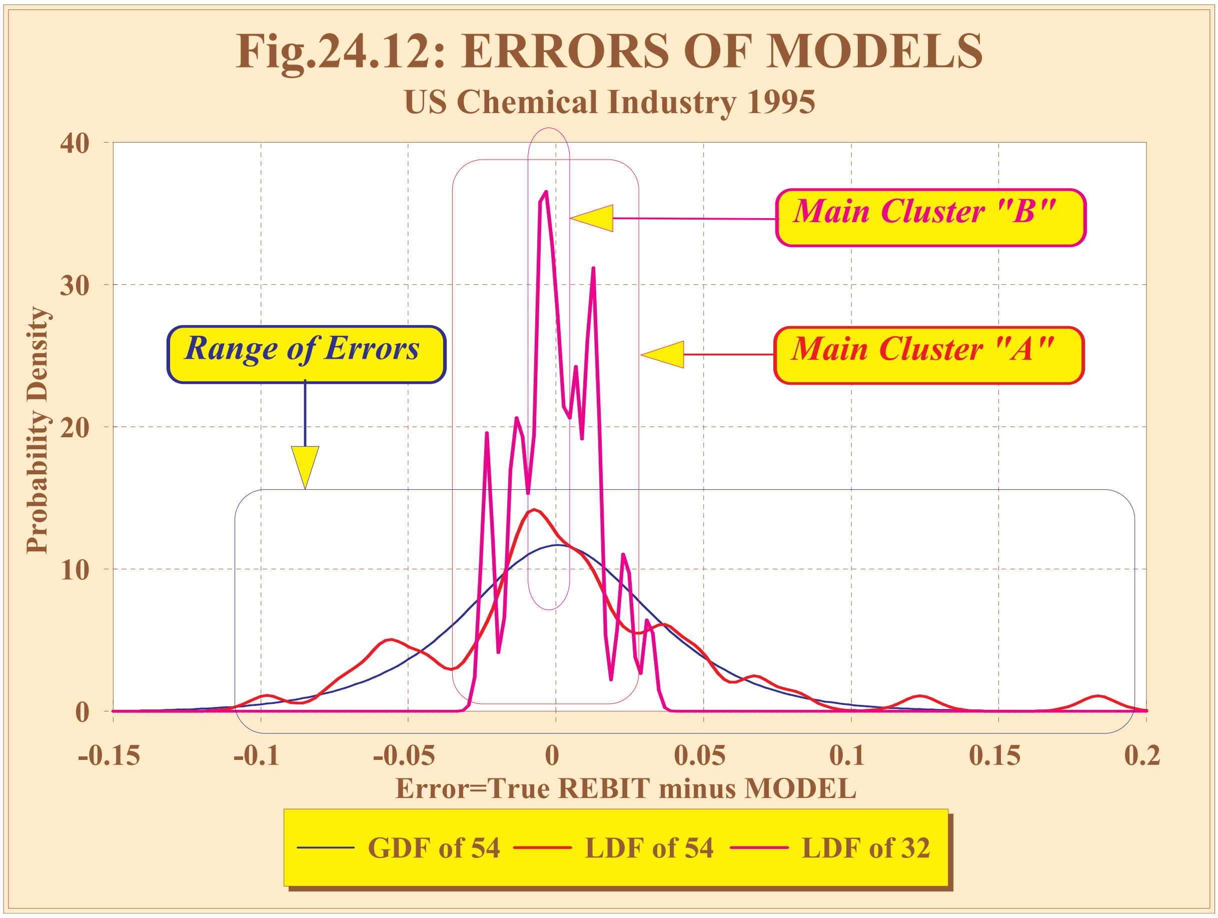
After separating the 10 companies of cluster "B," the desired sub-sample of comparable companies is obtained. This cluster's model coefficients (line 5 of Tab. 24.3), obtained using OLS, closely approaches those values estimated by the robust method (line 4 of Tab. 24.3). The near coincidence of these values demonstrates, that the model now explains the data's behavior very well and this is further supported by the OLS model's *R-Square* of 0.995 and by the very low MAE (0.00151). This whole process, which results in the isolation of cluster "C" from clusters "A" and "B", is illustrated by the three density curves shown in Fig. 24.13 using a more detailed scale of errors.

The concept of multivariate gnostic clustering consists of repeating the robust modeling procedure to gradually extract individual homogeneous clusters, thus decomposing the original inhomogeneous sample into its components.

The "first round" of the multidimensional clustering described above consists of the following steps:

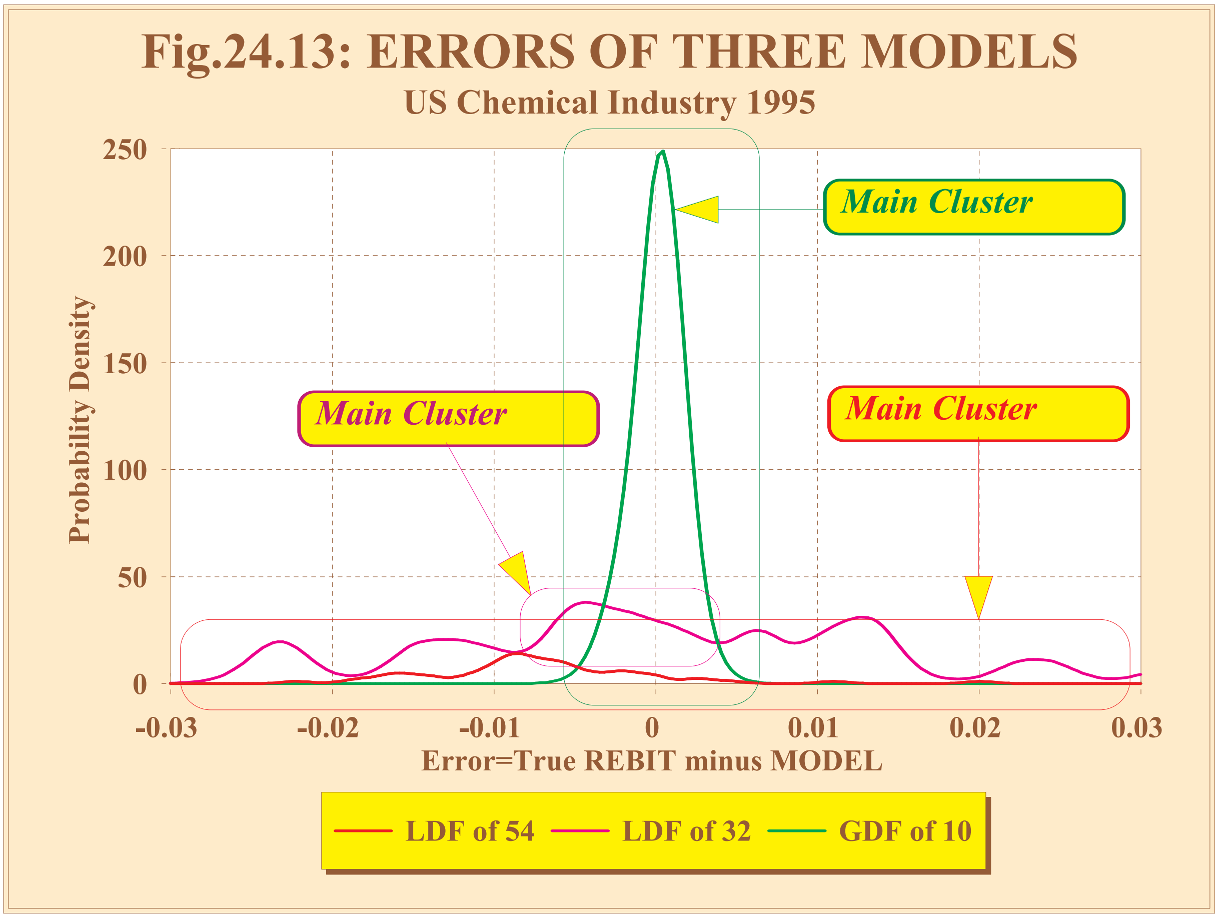
1. The coefficients of the model 24.13 are estimated for all 54 companies using a robust methodology.
2. The 54 sets of explanatory variables are substituted into 24.13 to obtain estimates $\widetilde{EBIT/TA_k}$ for the observed values of $EBIT/TA_k$ $(k = 1, \ldots, 54)$.
3. The vector $\widetilde{EBIT/TA_k} - EBIT/TA_k$ of residuals (modeling errors) is calculated for all 54 values of k.
4. The density of the local distribution function (ELDF) of the residuals is calculated using a sufficiently small value for the scale parameter to reveal the individual clusters as autonomous "hills" (the red density curve in Figs. 24.12 and 24.13.
5. The highest of the density's "peaks" and the density's minima closest to the main peak are found and the data between the two minima are accepted as members of the main cluster A (the dim red frame in Figs. 24.12 and 24.13 delimits the interval of cluster's A data).
6. Cluster A's data are extracted from the original data sample for further analysis (there are 32 data in A).

The same process is applied to cluster B's data, its main cluster C is



formed by 10 data. The data bounds of cluster B are delimited by the magenta frame in both Fig. 24.12 and Fig. 24.13.

After having extracted the homogeneous cluster C (number 1) containing 10 companies, the remaining 44 data form another inhomogeneous data sample. The same procedures are repeated and gradually clusters 2 (14 companies), 3 (9 companies), 4 (8 cos.), 5 (5 cos.) and 6 (6 cos.) are isolated. Two companies remain, which exhibit behavior so far removed from the six models, that they must be considered as not belonging to the rest of the industry. An examination of their data shows extremely low $EBIT$ (0.065 and 0.060), which is not "compensated for" by the expected value of the other parameters RWC, RNS and/or $RTLB$. The characteristics of the six cluster models are summarized in Tab. 24.4.



Cluster		Coefficients				Quality	
No.	Size	K_0	K_1	K_2	K_3	R^2	STD
1	10	0.270	-0.137	0.00505	-0.231	0.995	0.0022
2	14	0.0861	0.221	0.0556	-0.0650	0.997	0.0024
3	9	0.333	-0.0686	-0.0583	-0.222	0.988	0.0057
4	8	-0.0959	0.128	0.159	0.0717	0.984	0.097
5	5	0.303	-0.0680	-0.129	-0.0310	0.986	0.0082
6	6	-0.784	0.583	0.418	0.760	0.998	0.0043

Tab. 24.4 Review of the models of six homogeneous clusters of comparable companies in the US Chemical Industry, 1995.

The models of the individual clusters are all different and it is difficult to find any regularity in their coefficients at first sight. However, their statistics, R^2 and STD, show that they all explain the data very well indeed.

The arithmetic averages of the parameters of the individual clusters are shown in Tab. 24.5.

Cluster		Mean Ratios			
No.	Size	$EBIT/TA$	RWC	$TATO$	TL/TA
1	10	0.111	0.215	1.137	0.587
2	14	0.135	0.162	0.964	0.625
3	9	0.127	0.224	1.182	0.551
4	8	0.136	0.130	1.075	0.606
5	5	0.149	0.106	1.001	0.589
6	6	0.157	0.156	1.069	0.530
All	54	0.130	0.170	1.067	0.591

Tab. 24.5 Mean of the ratios for the six homogeneous clusters of the US Chemical Industry, 1995

This table shows, that there are large differences between the means of the individual clusters. This is not particularly surprising since each cluster represents a different set of "ratio behavior" patterns. To visualize the different features of each cluster, it is useful to evaluate the **mean impact** of the different ratios on the mean before tax business return on assets ($AVG(EBIT/TA)$) as $K_x * AVG(RX)$, where K_x are the model's coefficients from Tab. 24.4 ($x = 1,\ 2,\ 3$) and $AVG(RX)$ represents the mean value of the ratios taken from Tab. 24.5 ($X = RWC$, $TATO$ and TL/TA). These results are shown in Fig. 24.14 together with the mean values of $EBIT/TA$ ("Profitability"). "Liability" represents the product $K_3 * AVG(TL/TA)$ (financial leverage), "Liquidity" stands for $K_1 * AVG(RWC)$ and "Activity" is modeled by $K_2 * AVG(TATO)$. The averages are taken over all members of each cluster. The order of the rows in the diagram is chosen so as to prevent one column from being obscured by another and the clusters are ordered by the values of the debt impact (Liability). This then provides a "snapshot" of the manner, in which each cluster's policies contribute to the firms' profit.

As the last graph shows, the mean relative returns ($AVG(EBITdTA)$) for each individual cluster differ only slightly. The most striking differences are seen in the impact of debt ($AVG(TL/TA)$), which is strongly positive in cluster 6 and weakly positive for cluster 4, while the others manifest a negative effect from debt with an increasing negative intensity from cluster 5 through cluster 1. A possible economic interpretation is, that since the firms in cluster 6 have such low relative financial leverage, they could substantially increase their debt burden and still enhance their profitability. This is also supported by the fact, that the members of this cluster already have the highest value of profitability as well as a large positive contribution to profit from liquidity and activity. (The sum of the magenta, blue and green columns of a cluster, would equal the red column only after the inclusion of the intercept, which is not shown in the diagram).

The influence of financial leverage in clusters 4, 5 and 2 is close to neutral; given the combined policies employed by these firms, the amount of leverage taken on doesn't appear to provide any significant contribution to profitability.

The situation in clusters 3 and 1 is less clear: there are potential signs of financial distress (the effect of debt is not offset by the other components in the composite model). The negative impact of liquidity supports such a conclusion. Yet profitability is at a respectable level, therefore one must search for other causes to explain the apparent success of these firms.

The negative impact of activity in clusters 5 and 3 once again leads to a search for a possible cause. Tables 24.4 and 24.5 imply, that the reasons could be different in these two cases: $TATO$ has the highest mean value of all the variables in each of the clusters, but the negative coefficients of the equation of clusters 3 and 5 indicate, that even a high level of sales can be excessive and not increase efficiency. The low liquidity levels and the negative impact of total liabilities on these firms would tend to infer, that the problem probably lies with excessive plant and equipment (fixed assets), perhaps recent acquisitions, which have not yet contributed to sales.

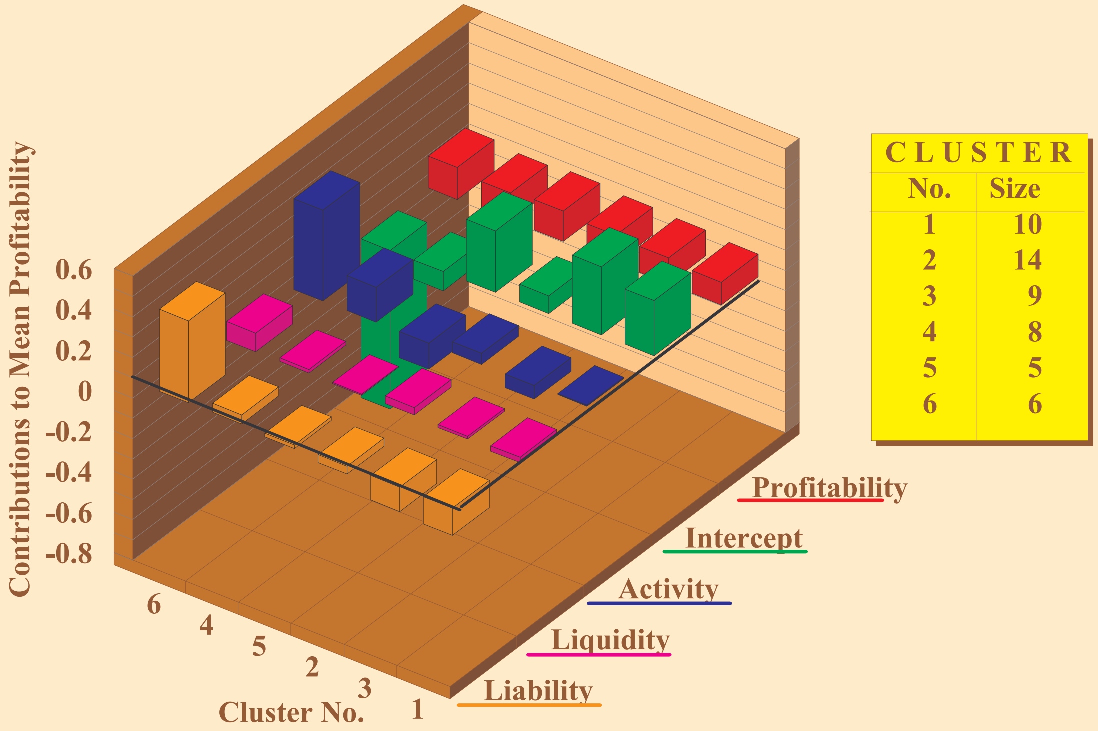
An analysis based on criteria such as those displayed in Fig. 24.14 identify the features (operational policies), which can be used to classify firms as members of one cluster or another, and it highlights the differences between clusters.

General conclusions such as the following might be drawn:

- Firms in cluster 6 offer maximum return on assets due to the most efficient use of low financial leverage, a medium total asset turnover and a relatively low liquidity. This cluster contains the best performing companies.
- Cluster 4 is formed by good performing firms; their policies indicate no problems.
- Companies in cluster 3 and 1 are the poorest performing firms in the Chemical Industry, although they manifest the highest activity and liquidity and do not appear to use excessive financial leverage. The impact of these (not bad) financial parameters on the (still not bad) profit is, however, negative thus warning of potential problems, which could lead to financial stress.

Fig.24.14: MULTIVARIATE CLUSTERING

US Chemical Industry (1995)



CLUSTER	
No.	Size
1	10
2	14
3	9
4	8
5	5
6	6

- Firms in clusters 5 and 2 can be classified as performing “on average” between the “good” in 6 and 4 and the “bad” in 3 and 1.

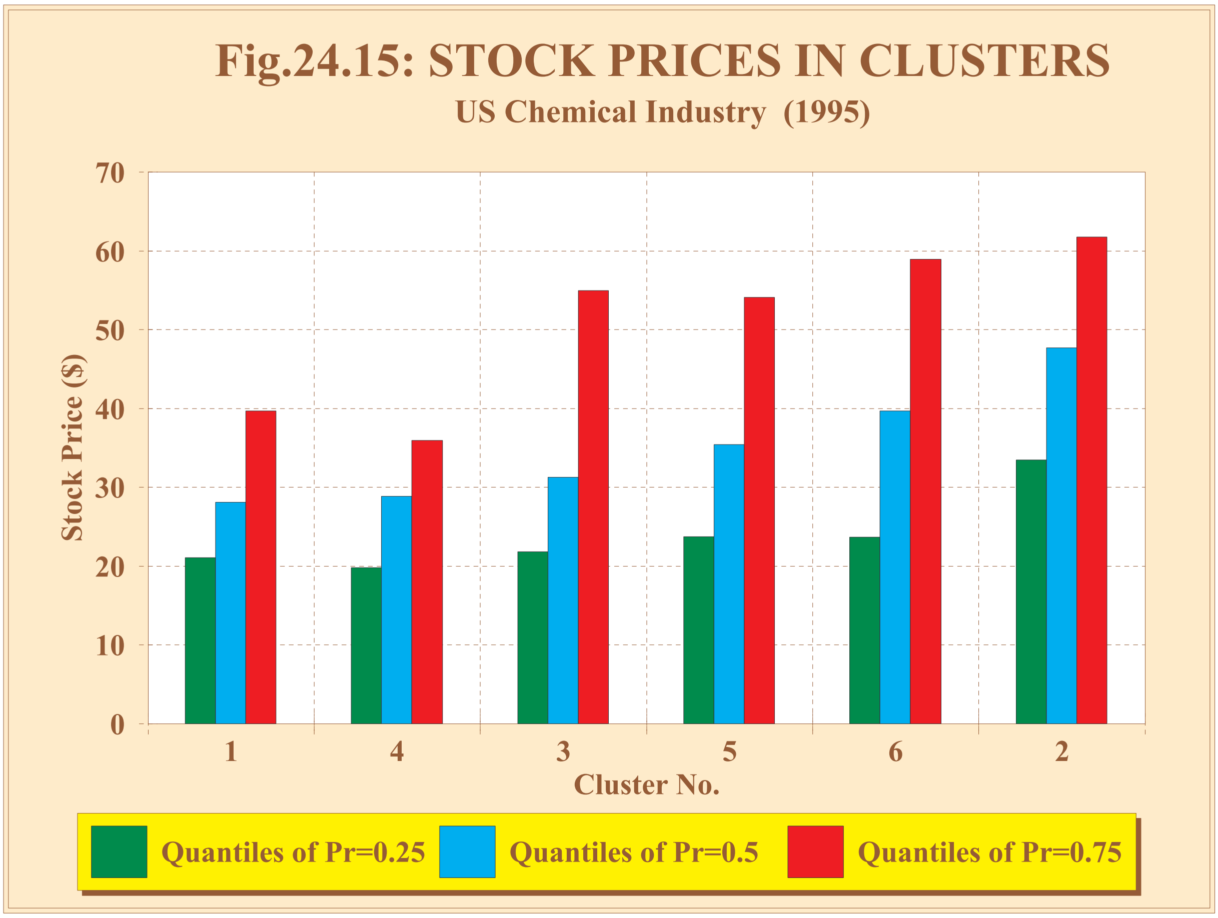
This technique is useful for the selection of firms, which behave similarly, and are therefore comparable to each other; it then presents a solution for one of the problems set out earlier: how to identify groups of comparable firms within an industry. The startling conclusion, that can be drawn here, is that the ‘membership” of a firm in an industry is no guarantee, that it will compare in economic terms with all the other members. Only a portion of the firms appear to operate in a similar way (in accordance to the same model) and in extreme cases there may be no comparable firm. This inhomogeneity is probably caused by the application of the purely qualitative classification criteria used for the definition of an industry.

The real danger, which results from this intra-industrial incomparability and clustering of firms is relying on the usage of “typical”, “normal”, “healthy” or even “recommended” values for financial ratios obtained by averaging over a whole industry. Using the means of parameters of economically incomparable firms cannot provide useful or accurate results.

24.3.2 ‘Outer’ versus ’Inner’ Classification of Clusters

Classification by Stock Prices

Taking the analysis from the investor’s standpoint, a subjective evaluation of the riskiness of the investment is necessary; the market prices risk, and the resulting stock price is a reflection of the investment community’s assessment of the firm’s future performance. Stock price then becomes a natural candidate for one of the external measures of a firm’s performance. The application of this idea to the six clusters identified above will now be considered along with the results, which are summarized in Fig. 24.15.



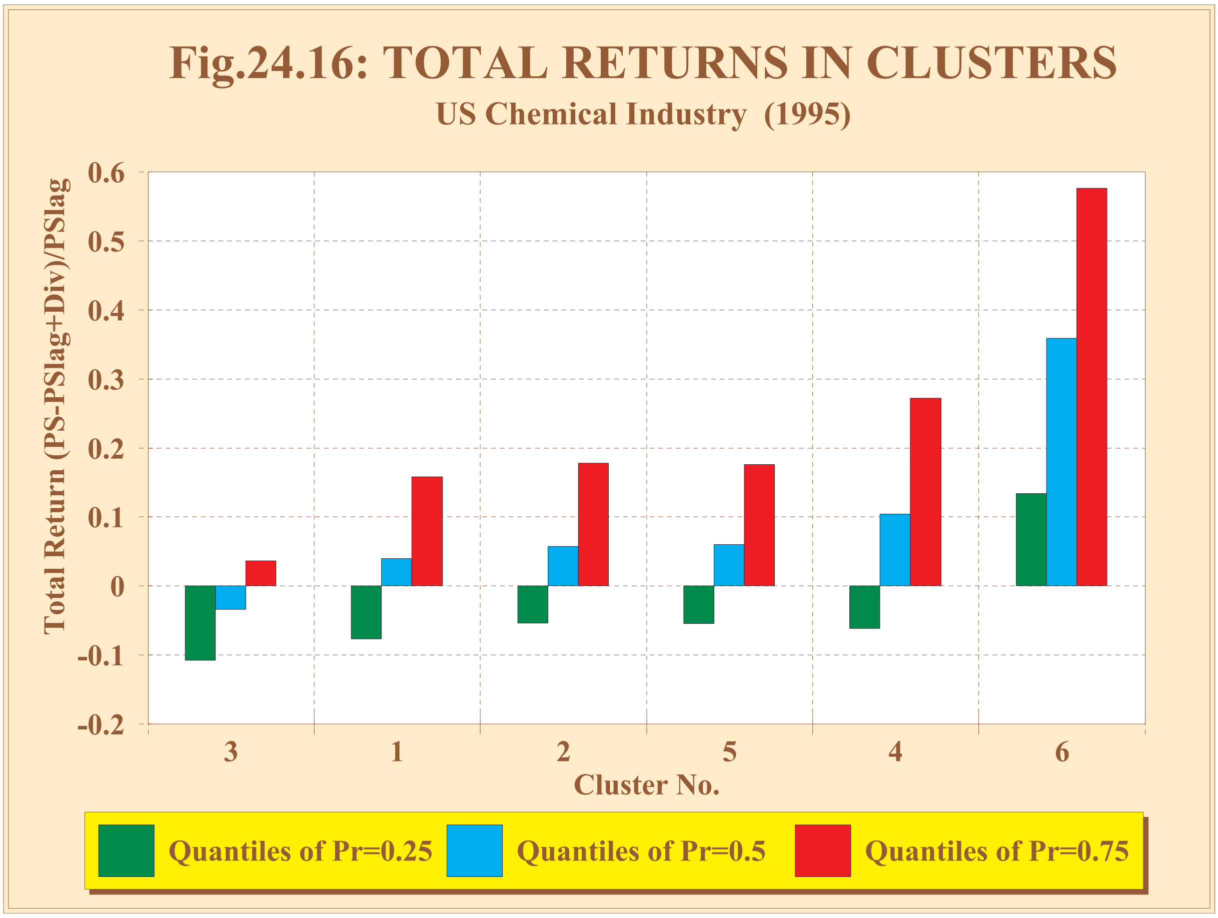
Three quartiles (quantiles corresponding to probabilities 0.25, 0.50 and 0.75) are shown in the columns of Fig. 24.15. These were robustly estimated using the EGDF of stock prices of the firms in the six previously identified clusters. The performance ranking of the clusters based on Fig. 24.14 and on their mean profitability $EBIT/TA$ (Tab. 24.5) are not in a full agreement: the best cluster (6) comes in second according to the latter valuation, cluster 3, which previously was one of the worst now surpasses the second best (4) (which becomes the second worst in Fig. 24.15). This approach, incorporating market pricing, does not strictly correspond to the valuation based on the 'inner' or accounting based view.

It must be recognized, that just as EPS, stock prices are functions of the number of shares outstanding and saying, that the stock price of A is better than B's or improved more, etc. is not a very good performance indicator. A more useful alternative is total return, defined as the relative annual change of the stock price plus dividend per share (or other cash flow to the investor). This benchmark takes into account both investing strategies, the growth-oriented and the dividend based. A strong correlation of the total return with stock prices was observed in Chapter 21 although a full interdependence was not shown. It therefore seems reasonable to take the market's view into account through this measure. Regardless of its size, or the number of shares outstanding, a firm returning 10% on its stock price at the beginning of the year performed as well as another, that earned the same amount, and better than another, that provided only 7%.

Classification by Total Returns

Total returns using closing stock prices and the annual dividend paid can be computed for each of the previously identified clusters and the robust distribution functions (EGDF) of returns for each cluster can be obtained to provide a second measure for a market based performance evaluation. Just as in Fig. 24.15 for stock prices, the three quartiles for total returns are shown in Fig. 24.16.

Since the composition of the six clusters was made from financial statement data without reference to either market price or total return, any support from a market based evaluation for *these same rankings* can be taken as strong support for the idea, that the multivariate analysis does reveal substantial differences in the economic behavior of firms in the same industry. This coincidence of 'inner' with the 'outer' valuation can seem to be conflicting with Fig. 24.8, which manifested the rather chaotic behavior of the market indicators. This seeming contradiction can be resolved by emphasizing following:



1. The indicators' values shown in Fig. 24.8 were related to individual firms, while Fig. 24.16 characterizes mean values of a group of firms, which form the cluster. The averaging suppresses the volatility.
2. The valuations in Fig. 24.8 took into account **all** the firms and did not respect differences between clusters, while Fig. 24.16 takes the cluster structure into account.

A comparison of Fig. 24.14 through Fig. 24.16 offers the following comments:

1. the ordering of clusters by stock prices differs from the ordering by total returns, and is not particularly useful, as explained above,
2. the valuation of clusters by total returns discriminates the cluster's performances significantly more sensitively than a valuation by stock prices,
3. the ordering by total returns is in full correspondence with the valuation based on Fig. 24.14: cluster 6 is absolutely the best, cluster 4 is the second best and clusters 3 and 1 are the worst with cluster 2 and

5 performing ‘on average.’ The reversal of the positions of clusters 3 and 1 can also be explained:

- According to the total return, cluster 3 is worse than cluster 1, because more than 50% of its members earned negative total returns. In figure 24.14, cluster 1 was in the lowest place, because the impact of financial leverage was slightly stronger ($K_3 * AVG(TL/TA) = -0.587 * 0.988 = -0.580$) than in cluster 1 ($-0.551 * 0.995 = -0.548$). However, the firms in cluster 3 suffer from the strong negative impact of the asset turnover. In this respect, both valuations put cluster 3 in the last place.

Multivariate clustering based on financial statement data has thus shown, that several subgroups in the same industry may not be comparable using simple ratio analysis, and it has revealed the possible causes of such incompatibility showing both the existence of and the possible causes for the inability to directly compare them on the basis of simple ratio analysis. When clusters of firms are identified as above, the firms within each cluster are comparable to each other, and they behave in accordance to the same model. These individual models then provide a base, which can be used to:

- determine the operating policies of the member companies,
- provide a performance evaluation,
- rate each company as a potential investment candidate.

As a by-product of the analysis, it can be stated, that (in this case) the rating view based on the individual firms’ financial statement information was consistent with the outside view, the market’s evaluation based on total returns, while an appraisal based on stock prices gave somewhat conflicting results.

However, to forestall overestimation and misinterpretation of these results, the following cautions are offered: To belong to a certain—eg the ‘best’ cluster No.6 or the ‘worst’ one No.3—does not automatically mean ‘to be better/worse than all members of the other clusters.’ What it really means is, that they ‘react in a similar way as other members of the same cluster to some impulses and differently from members of other clusters’. Equations 24.13 would hold even after multiplication of both sides by a non-zero constant (lower/higher inputs $\Leftrightarrow$ outputs). This means, eg, that firms in a certain cluster can have different values for the dependent variable. The coincidence of the inner and outer valuation (demonstrated by Fig. 24.16) is shown for three quartiles.

The ranges of total returns within a cluster can be very broad. This prevents a reasonable ordering: total returns in the second best cluster No.4 range from -0.11 through 0.93, while total returns in the 'best' No.6 are from -0.01 through 0.74. It is thus difficult to confirm the "global" or "universal" superiority of No.6 over No.4 as could seem in Fig. 24.16. Hence, demonstrated clustering does not provide a complete ordering of all the firms' performances, it 'only' delimits groups of firms similar in economic behavior. A transparent example can be given from ordinary life: cars differ by their power/weight ratio. A car from the most powerful (and expensive) class/cluster accelerates faster than a car from the cluster of weak (and cheaper) cars. Cars from the same cluster react to increasing accelerator pressure in the same proportion. But, this does not mean, that they always run at the same speed. It depends on the driver, some are slow, while others take the vehicle up to its maximum speed.

The MD-rating and the multidimensional cluster analysis obviously solve different problems and answer different questions. They should be used together, and completed with an economic interpretation of the outcome of the computations.

24.4 Implicit Multidimensional Cluster Analysis

24.4.1 Clustering in Implicit Models

As discussed in Chapter 17.4, there are two problems with using explicit regression models:

1. the feed-back problem, and
2. conflicts between seemingly equivalent equations.

Both of the problems related to explicit equations can be eliminated by using implicit models. The existence of an implicit model can be easily shown for the case, which is being illustrated. It was seen, that a system of six homogeneous clusters satisfying explicit equations of the type 24.13 can be obtained by multidimensional clustering. The homogeneity of the clusters' models (their statistical acceptability) has been verified. Moreover, an entirely independent check with mean values of total returns showed, that such a partitioning of the firms is reasonable. Equation 24.13 can therefore be taken as 'granted' and used as a point of departure for further reasoning.

In order to measure more precisely the influence of the explanatory variables, 24.13 will be modified slightly before going over to an implicit model:

- profitability will now be evaluated by net profit defined by the ratio EAT/TA,
- instead of using the liquidity ratio RWC $(CA/TA - CL/TA)$, the addends will appear as separate variables to be allowed to generate their own weighting coefficients, because there is no real reason to expect, that they affect 'working capital' by the usually accepted a priori weights +1 and -1. Now, these variables will show their true influence themselves,
- another change concerns the total liabilities TL (the sum of current liabilities CL and long term debt LTD). A new ratio LTD/TA will be used instead.

The explicit equation 24.13 is thus rewritten as follows:

$$\begin{aligned} EAT/TA_k \;=\; & K_0 + K_a * CA/TA_k + K_b * CL/TA_k + \\ & K_2 * TATO_k + K_c * LTD/TA_k \end{aligned} \tag{24.14}$$

for some $k = 1, \ldots, K$. Two notes, which pertain to the intercept (constant) K_0 are in order:

1. The actual interdependence of each of the ratios is not necessarily linear. In these cases, the linear equations model depicts the relations between projections of the variables' deviations from a fixed point onto a tangential plane. The location of the point is determined by the intercept's value.
2. The variables on the right-hand side of the equation cannot completely explain the behavior of the dependent variable. Therefore, the value of the intercept incorporates the impact of factors not included in the model.

In any event if K_0 is not zero, equations 24.14 can be divided by K_0 to obtain the desired implicit equation

$$\begin{aligned} K_1^{'} * EAT/TA_k + K_2^{'} * CA/TA_k + K_3^{'} * CL/TA_k + & \\ K_4^{'} * TATO_k + K_5^{'} * LTD/TA_k \;=\; & 1 \end{aligned} \tag{24.15}$$

The new coefficients, $K_1^{'}, \ldots, K_5^{'}$, will give all the variables in the equation the same 'right' to contribute to the joint result. None of the variables play a ‘special’ (dependent) role. After the coefficients are estimated, the dependence of each ratio on the others can be obtained from the model with no inconsistencies.

The computation of the optimal implicit model is analogous to the explicit case with one exception: Statistical programs, which evaluate the multidimensional correlation coefficient (R^2) need a non-zero variance for the dependent variable to be substituted into the denominator of the proportional ratio of variances. This is impossible if a constant plays the role of the ‘dependent variable’, because its variance would be zero. Therefore, if a statistical verification of clusters, which result from the implicit analysis is desirable, it is to be performed on the explicit equivalent of the implicit regression.

The main clusters of the modeling errors of the local distribution’s (ELDF) density are determined and isolated just as in the explicit case and the inhomogeneous data sample is decomposed into homogeneous clusters. The coefficients $K_*^{'}$ for the implicit equation system 24.15 applied to the US Chemical Industry for 1997 are shown in Tab. 24.6.

Cluster	**Coefficient for the Ratio**				
No.	CA/TA	CL/TA	$TATO$	LTD/TA	EAT/TA
I	1.231	0.639	-0.120	1.197	0.904
II	0.310	1.526	0.203	0.680	-0.046
III	0.264	0.823	0.174	0.849	2.636
IV	1.275	0.692	-0.164	1.250	0.427
V	1.437	1.050	-0.117	0.359	2.065
VI	2.060	-2.659	0.728	1.590	-4.941
Cluster	**Mean Contribution of the Ratio to 1**				
No.	$K_1^{'} * \overline{CA/TA}$	$K_2^{'} * \overline{CL/TA}$	$K_3^{'} * \overline{TATO}$	$K_4^{'} * \overline{LTD/TA}$	$K_5^{'} * \overline{EAT/TA}$
I	0.588	0.168	-0.164	0.339	0.068
II	0.134	0.418	0.225	0.227	-0.004
III	0.121	0.217	0.195	0.259	0.209
IV	0.437	0.145	-0.193	0.586	0.031
V	0.616	0.263	-0.136	0.138	0.118
VI	0.823	-0.552	0.629	0.510	-0.410

Tab. 24.6 US Chemical Industry, 1997: Coefficients of the implicit equation 24.15 and mean contributions of the ratios to the right-hand side value (1) for each cluster

The results presented above suggest the following:

1. Imagine, that the mean value of the ratio $\overline{CL/TA}$ of cluster I has increased by 1%. This increment multiplied by the coefficient $K_2^{'}$ (0.639) will increase the ”share” of the ratio by 0.00639. However, the same relative increase in LTD/TA would increase its contribution

by 0.01297, ie nearly twice as much. This implies, that changes in non-current liabilities are a more powerful tool to control the financial condition of a cluster I firm than changes in current liabilities. It is easy to see in the upper part of Tab. 24.6, that an analogous relation also holds for cluster IV, while in clusters II and V the relation is in the opposite direction: current liabilities provide a much stronger effect than the long term ones. The situation in cluster VI is entirely different: the contribution of the two types of liabilities have opposite signs, they "pull on the opposite ends of the rope."

Taken from the perspective of a financial manager, assume, that he considers the effect of acquiring an additional asset ΔA in the most effective way. The denominators of all the ratios will change to $A+\Delta A$, while the changes in the numerators will be proportional to coefficients $K_k^{'}$. If the manager knows, to which cluster his firm recently belongs, then the upper part of Tab. 24.6 solves his problem.

2. In providing the weights for CA/TA and CL/TA, coefficients $K_1^{'}$ and $K_2^{'}$ are far from the ordinary equality ($K_1 = -K_2$), which is used to establish the relative value of the working capital (RWC). This does not mean, that the difference $CA/TA - CL/TA$ cannot be used to evaluate the firm's liquidity, but it gives warning, that the effect of the ratios is not simple nor is it constant through time, assumed as a given, and known once for ever. Instead, as is seen from the diversity of weights in Tab. 24.6, a careful data-based analysis is necessary in each particular case.
3. Complementary information is offered in the lower part of the Tab. 24.6, where the contributory "share" of each ratio to the sum (1, see 24.15) is presented as the product of the equation coefficients $K_*^{'}$ and the ratios' average value taken over each of the clusters[6]. It is easy to see, that clusters I, V and VI are "CA/TA-dominated" in the sense, that this ratio manifests its strongest effect. Using the same term, it is possible to conclude, that clusters III and IV are "$TATO$-dominated" and cluster II is "CL/TA-dominated". A complete description of the differences between clusters is of course more complicated and it is given by the clusters' full models.

[6] Observant readers will find, that the sum of each column in the rows of the lower part of Tab. 24.6 slightly deviates from 1. This discrepancy is caused by rounding.

24.4.2 Comparison of Clusters by Odds

It was shown in Fig. 24.16, that clustering based exclusively on "inner" information taken from the financial statements of chemical firms for 1995 was in agreement with the "external" evaluation produced from the total returns of the firms in these individual clusters. This comparison made use of three quartiles estimated from the global distribution functions EGDF; but this is not the best use of the powerful instrument represented by these functions. A knowledge of the quartiles does not permit an answer to the most important question on the mind of an investor: what are the chances of making a successful investment, or the flip side: what are the risks? A formal statement of the question is: "given an industrial group (a cluster of comparable firms) and choosing an acceptable level of total return TR_a for a firm from this cluster, what is the chance of exceeding TR_a?"

Chances are measurable by probabilities. To go over from data to probability is easy in gnostics because of the availability of distribution functions, which can robustly estimate each desired probability directly from the data without the necessity of an a priori assumption on the model. However, many people are more accustomed to evaluate chances and risks in terms of odds and transforming the relation between the probability p of an event to odds in favor OIF of this event is very simple:

$$OIF = \frac{p}{1-p}. \tag{24.16}$$

The following steps lead to the answer:

1. Perform a multidimensional cluster analysis of the target industry to find the cluster (X), to which the firm to be tested belongs.
2. Estimate the global distribution function EGDF of TR for cluster X.
3. Define the event as exceeding the value TR_a by a firm from cluster X and determine its probability $p := Pr\{TR > TR_a\}$.
4. Calculate the odds in favor of this event using 24.16.

Such an analysis is demonstrated below with data from the US Chemical Industry for 1997 and 1998.

To make a useful comparison of the results over the two years, it is necessary to take into account the substantial changes in market conditions over this period (Tab. 24.7).

The arithmetic means of the characteristics in Tab. 24.7 illustrate how

Year	$\overline{CA/TA}$	$\overline{CL/TA}$	$\overline{TATO}$	$\overline{NCL/TA}$	$\overline{ROA}$
1997	0.423	0.245	1.137	0.351	0.076
1998	0.399	0.241	1.078	0.378	0.062
Year	$\overline{EPS}$	$\overline{PS}$	$\overline{DIV/PS}$	$\overline{TOTR}$	$\overline{E/P}$
1997	1.932	37.409	0.018	0.181	0.053
1998	1.558	31.510	0.023	-0.138	0.050

Tab. 24.7: Means of ratios and indicators for the US Chemical Industry for 1997 and 1998.

different the market performance was in the two consecutive years. The mean financial statement ratios were only a little worse, but the market indicators declined significantly. The fall of stock prices is useful in establishing, that ’98 was a down market year, but it is not clear, what changes there were in total number of shares outstanding, which puts a cloud on the EPS figures. The most sensitive decrease was in the total returns (which fell from 0.181 to -0.138) especially because of a tumble in stock prices. The fact, that most firms try to maintain stable dividend policies, even with declining earnings, can be seen by the increase in mean dividends paid, which increased from 0.697 to 0.726 (to signal, that ‘things aren’t all that bad’). This, together with decreased stock prices, resulted in increased dividend yield (DIV/PS, dividends paid divided by the stock price) from 0.018 to 0.023.

The real lesson here is in the E/P: Those, who held the stock over the two year period took a bath, but the measure says, that investment return didn’t change appreciably; what it really is saying is, that if you buy the stock ‘now’ (31 Dec. ’98) and the earnings don’t go down in ’99, then the return will be around 5% unless the stock price increases, but isn’t the latter, what the investor is hoping will happen? Not a very useful barometer!

The reaction of OIF of the clusters, shown for both years in Fig. 24.17 is interesting.

In 1997, the odds in favor of exceeding an acceptable value of total return (set to 0.10) in all but one cluster were substantially greater than for 1998. Since the odds in favor of success in throwing a coin is 1 (corresponding to a probability 0.5), the data in Fig. 24.17 show, that investing in the five higher clusters at the end of 1997 had a much better chance of success than buying the two worst ones,‘f’ and ’g’. The odds for the best, cluster ‘a,’ were nearly 7. In contrast, the odds were less then 3 for all clusters in 1998, even for the best performing one. The chances of the worst clusters

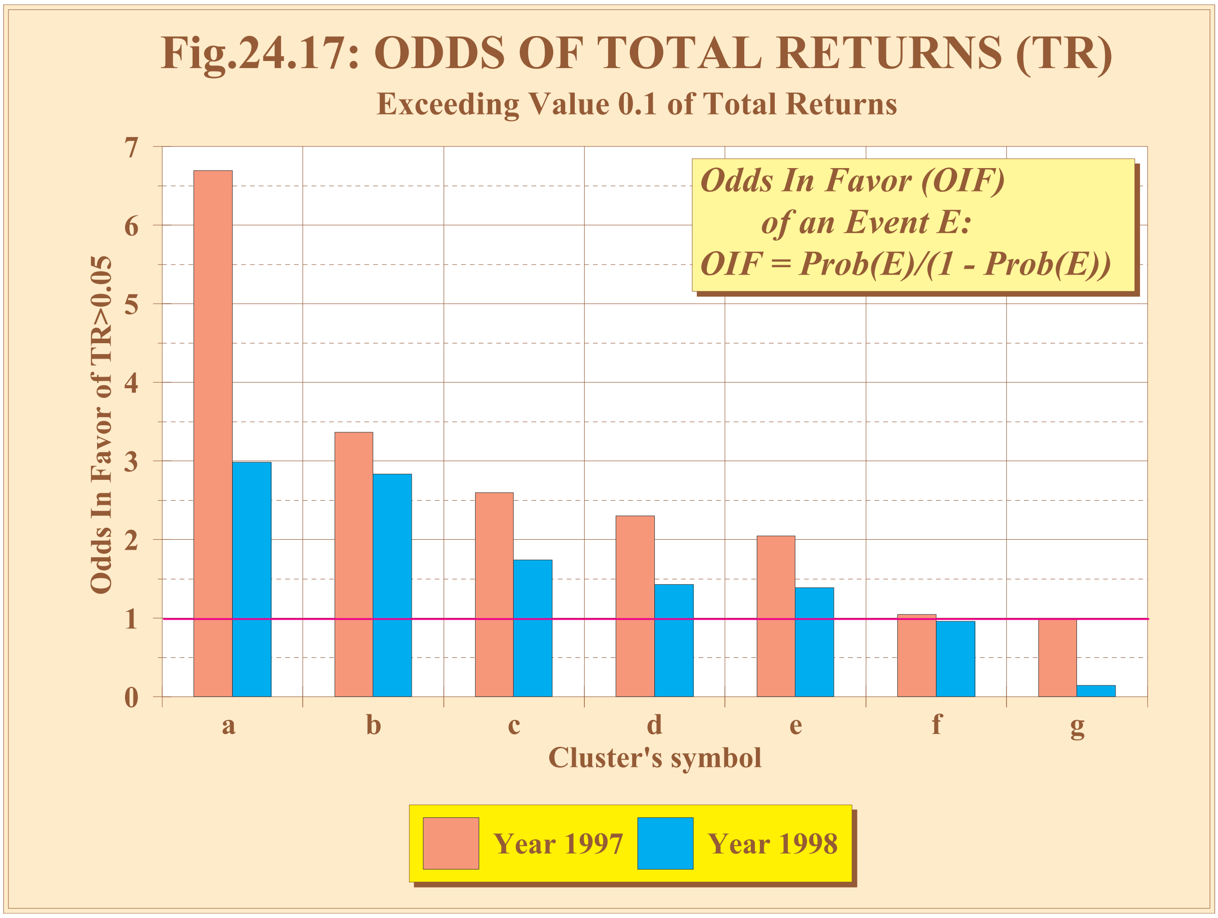


Fig.24.17: ODDS OF TOTAL RETURNS (TR)

Exceeding Value 0.1 of Total Returns

'f' and 'g' were no better than throwing a coin. These findings emphasize the importance of careful analysis.

It can again be concluded, that

1. clustering based only on financial statement information identified clusters of firms, which—evaluated through independent external appraisal by the market—lead to reasonable conclusions with respect to possible investor's decisions,
2. the odds of success for different clusters in the same industry and in the same year were significantly different by an order of magnitude,
3. the methodology was successful in both stable and transitory states of the market.

The clusters in Fig. 24.17 were ordered only by odds, but the composition of the clusters in both years was not necessarily made up of the same firms. The order of odds does not coincide with the order of the mean values of total returns, because the probability of an event depends not only on the mean value, but also on the form and breadth of the distribution function (on the volatility of the variable).

24.4.3 A Technical Note

The procedure of multidimensional cluster analysis described above and illustrated by Figs. 24.12 and 24.13 is based on the robustness of the modeling, which emphasizes the main cluster over the interval of the most 'dense' occurrence of data. It enables purposeful manipulations with groups of data, which lead to the isolation of the main cluster. The most efficient way to apply this approach is interactive: to depict the residuals' density on the screen, to decide on the break-points between the main cluster and the rest of data and to take out the main cluster manually step by step. The disadvantage is the necessity, that the analyst personally participates in the selection. There is an alternative suitable for full automation of the process using the following cycle:

1. Robust estimation of the multidimensional explicit or implicit model,
2. calculation of fitting errors (residuals),
3. finding the largest absolute residual,
4. removing the 'worst' data vector,
5. repeating this sequence until an error or the least acceptable number of data vectors is reached.

Although this simple approach frequently leads to usable results (as demonstrated by Fig. 24.14 and other examples) it suffers from the disadvantage, which can be demonstrated by the simple implicit equation: let C_1 and C_2 be coefficients of the implicit regression model, and R_1 and R_2 the corresponding ratios. In an ideal case equation $C_1 * R_1 + C_2 * R_2 = 1$ holds. However, it would also hold, when the 'contributions' to 1 is $C_1 * R_1 + C$ and $C_2 * R_2 - C$. In such a case, the firms in the same cluster could have a different structure for their contributions to 1. To prevent this non-uniqueness from occurring, another step can be included into the above cycle: "determine and remove the data vector, for which the absolute deviation of a contribution (say, $C_k * R_k$) from the mean is maximum." This approach will be called the *double elimination.*

24.4.4 Inter-industrial Comparability

The existence of mutually incomparable firms within an industry was established through multidimensional cluster analysis. This renders unusable the popular idea, that “typical” values for financial ratios really exist, and that they can be obtained as means or medians of the ratios of firms belonging to the industry. The variety of factors, which determine the ‘membership’ of a firm within a cluster is broad and colorful and cannot be reduced to a simple official classification bestowing membership. An important question arises in this connection: does membership in an industry automatically imply, that the firm is not economically comparable to firms in another industry?

In other word: do firms, which can be described by the same multidimensional financial model, exist in different industries? The intra-industry incomparability, which has been confirmed, adds complexity to financial statement analysis. Were there a positive answer to the question of inter-industrial comparability, the task would become much simpler. To gain further insight into this matter, consider a mixture of 136 firms, 35 of which belong to the US Chemical producers and the remainder to the Information Technology and other branches of the High-Technology Industry. The data source is the Compustat tape financial statements for the 1-st quarter of 2001. There is no conflict to the fact, that the two industries are dissimilar, they differ by raw, technology, products, market orientation, tradition, dynamics, interdependence with other sections of the economy, scientific background and many other aspects.

The first step is to verify the inhomogeneity of the whole set of 136 data vectors, which are composed of the ratios CA/TA, CL/TA, $TATO$, TL/TA and ROA defined as $EBIT/TA$. An attempt to estimate the explicit linear regression model of $EBIT/TA$ explained by the other ratios yields an R-Squared[7] of 0.088, which means, that a hypothesis of possible linear dependence is to be rejected. The standard error of the fit, 0.039, would also be unacceptable.

An implicit multidimensional cluster analysis results in the separation of 22 clusters, the parameters of which are summarized in Tab. 24.8. This clustering was performed using the automatic double elimination process repeated twice: only 12 of the clusters obtained in the first run had a satisfactory quality (these are identified by numbers in the first column of Tab. 24.8). The data of the ‘bad’ clusters were subjected to a second run, which yielded the results denoted by letters. The series 1, 2, ... and A, B, ... correspond to the order, in which the clusters were obtained.

[7] This statistics is called *Multiple coefficient of determination* or *Multidimensional correlation coefficient*. It estimates the part of variance of the dependent variable explained by the model.

Cluster		**Industry**		**Quality**	
No.	Size	Hi-Tech.	Chem.	R^2	STD of Y
CL	Size	HT	CH	R^2	STD
1	6	1	5	0.9985	0.0012
2	6	5	1	1.0000	0.0009
A	6	5	1	0.9664	0.0155
4	6	3	3	0.9640	0.0042
5	6	6	0	0.9849	0.0228
B	6	2	4	0.9337	0.0109
C	6	4	2	0.9990	0.0016
D	6	3	3	1.0000	0.0001
E	6	6	0	0.9910	0.0124
F	6	2	4	0.9561	0.0094
11	6	5	1	0.9655	0.0103
12	6	3	3	1.0000	0.0001
G	6	4	2	0.9940	0.0025
H	6	4	2	1.0000	0.0002
15	6	1	5	0.9989	0.0011
I	6	4	2	0.9945	0.0032
17	6	6	0	0.9999	0.0004
18	6	6	0	0.9862	0.0097
19	6	4	2	0.9357	0.0047
20	6	5	1	0.9968	0.0119
21	6	6	0	0.9971	0.0040
J	6	4	2	0.9816	0.0067

Tab. 24.8: Cluster structure of the mix of firms from two industries, 2000Q1

Recall, that the R-Squared (R^2) and the standard fitting error (STD) were calculated for the explicit version of the regression models of individual clusters. Models of all clusters are obviously acceptable from the statistical point of view. Even in the case of the least precise cluster, B, the R-Squared is not under 0.93. The worst standard error of the fit is 0.0228, (cluster 5). The total number of firms, that were clustered, is 132; four firms did not fall into clusters. The reason is obvious, at least with the outlying one, which had 'record-breaking' low $TATO$ (0.043) and very low CA/TA and CL/TA (0.047 and 0.060). In the other case , zero was given for both CA/TA and CL/TA and CL/TA was very high (0.93, the third highest and uncompensated by other ratios).

The principal finding from Tab. 24.8 is, that there really are firms, which operate according to the same multidimensional model (being thus comparable), although they belong to different industries. Out of the 22 clusters only five were 'clean', composed only of companies belonging to a single industry. The question of the economic comparability of firms is thus far from trivial: an industry does not represent a homogeneous collection of firms. The intra-industrial comparability of firms is not automatically warranted, while inter-industrial comparability does exist. Therefore, a multidimensional analysis is necessary before a judgement as to comparability can be made; this complicates matters, but on the other hand, the existence of inter-industrial comparability might simplify an industry wide analysis: analyzing a sufficiently representative number of firms from different industries would perhaps yield a finite variety of multidimensional models of ratios. A company could be matched to a cluster by identifying the model with the smallest 'distance' from the vector of ratios of the firm; this would then allow it to obtain a classification as a 'similar firm,' as a member of the subject cluster. It is to be emphasized, that this comparability is based on the similarity of vectors of impacts, the components of which are summed to approximate 1 (the 'dependent' variable on the right-hand side of the modified implicit equation 24.15). In other word, comparable firms 'generate' the addends on the left-hand side of the implicit equation in a similar way. This structure is thus the firm's 'economic image'. All addends contributing to 1 are dimensionless and the vector can be thought of as a point in an abstract vector space endowed with a Euclidean metric. The distance between points attached to individual firms (or clusters of firms) can be easily calculated eg as the mean square difference between the vectors' components to answer the question "how far is firm A from firm B". To illustrate the idea of inter-industrial comparability, consider a large computer firm CF1, which was assigned to cluster No.4 together with three companies from the US Chemical Industry (CI1, CI2 and CI3). Three other well-known computer firms CF2, CF3 and CF4 each fell in different clusters. It is useful to have a look at the distances between CF1 and all these companies (Tab. 24.9).

Company	Industry	Cluster	Distance from CF1
CI1	US Chemistry	4	0.074
CI2	US Chemistry	4	0.062
CI3	US Chemistry	4	0.065
CF2	US Hi-Tech Ind.	E	0.126
CF3	US Hi-Tech Ind.	17	0.203
CF4	US Hi-Tech Ind.	21	0.273

Tab. 24.9: Inter- and Intra-industrial Distances of CF1, 2000Q1

The mutual distances of the three other computer firms are not much less than that from CF1: CF2-CF3 0.114, CF2-CF4 0.263 and CF3-CF4 0.324. These examples confirm the thought, that firms, which are technologically very different, can be similar in terms of their economic behavior, while firms that are similar from the technological standpoint can operate with very different economic characteristics. It can be concluded, that the multidimensional economic model is thus not uniquely determined by the industry, to which firms officially belong.

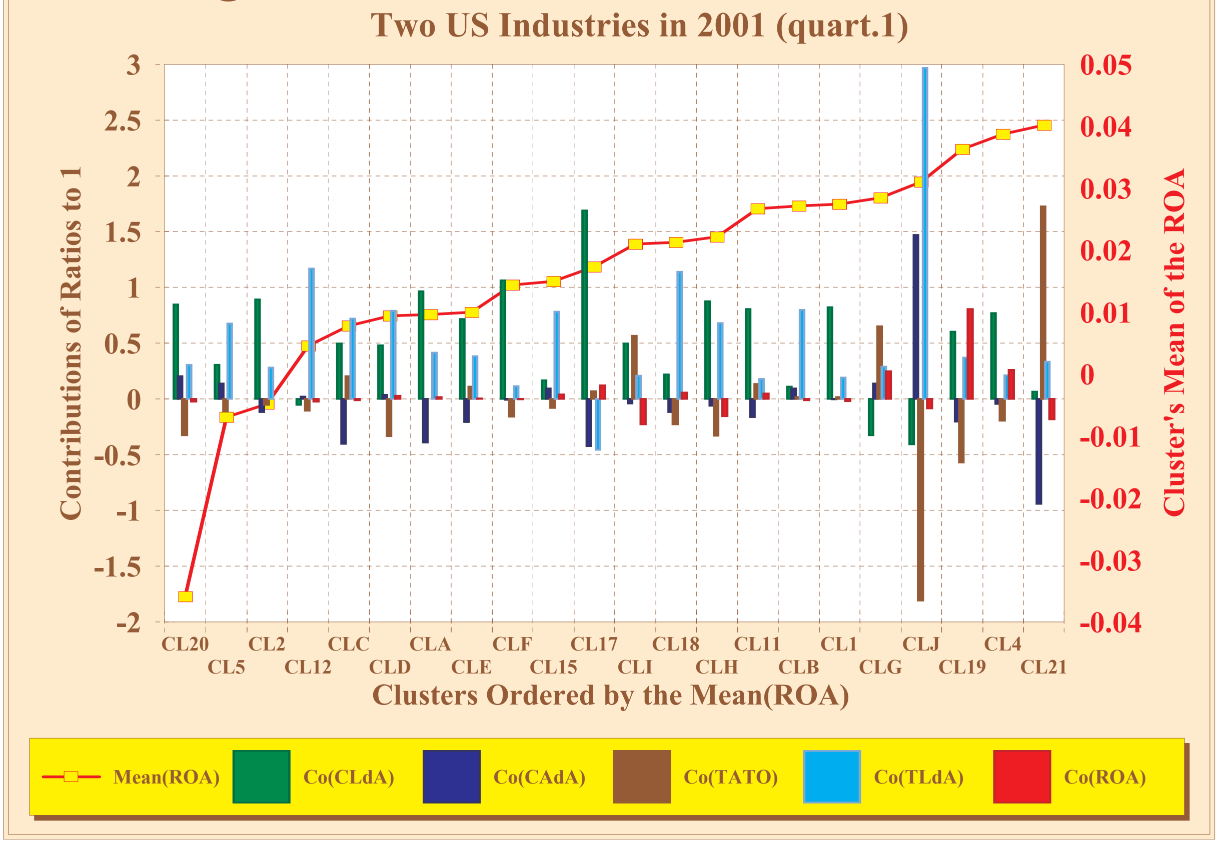
24.4.5 Structure of Clusters

A rough description of the clusters identified in Tab. 24.8 can be obtained from Fig. 24.18.

The clusters are ordered by the mean returns $EBIT/TA$ of the firms, from which they are composed. (These values are shown by the red line). The color of the columns reflects the clusters' structure by the mean contributions to 1 of the individual ratios of its members. It is easily seen, that the color patterns of the clusters are unique, with no duplication in their makeup. In each case, the sum of the lengths of the columns in each cluster (when respecting the sign of contributions) is 1 (plus or minus the relatively small modeling error). The size of columns gives a hint as to the volatility of cluster's parameters. So eg in the case of the cluster J (denoted CLJ) the sum which equals 1 is obtained as the difference of several contributions, which have opposite signs. Small relative variations in such 'fighting' components can result in a strong overall variation in the sum. Clusters with long columns of opposite signs can be thus expected to be more volatile than clusters, the columns of which are positive (eg CLB and CL1).

The size of a column leads to a judgement as to the dominating role of some ratios. So eg clusters CL12, CL15, CL18 and CLB are TL/TA-

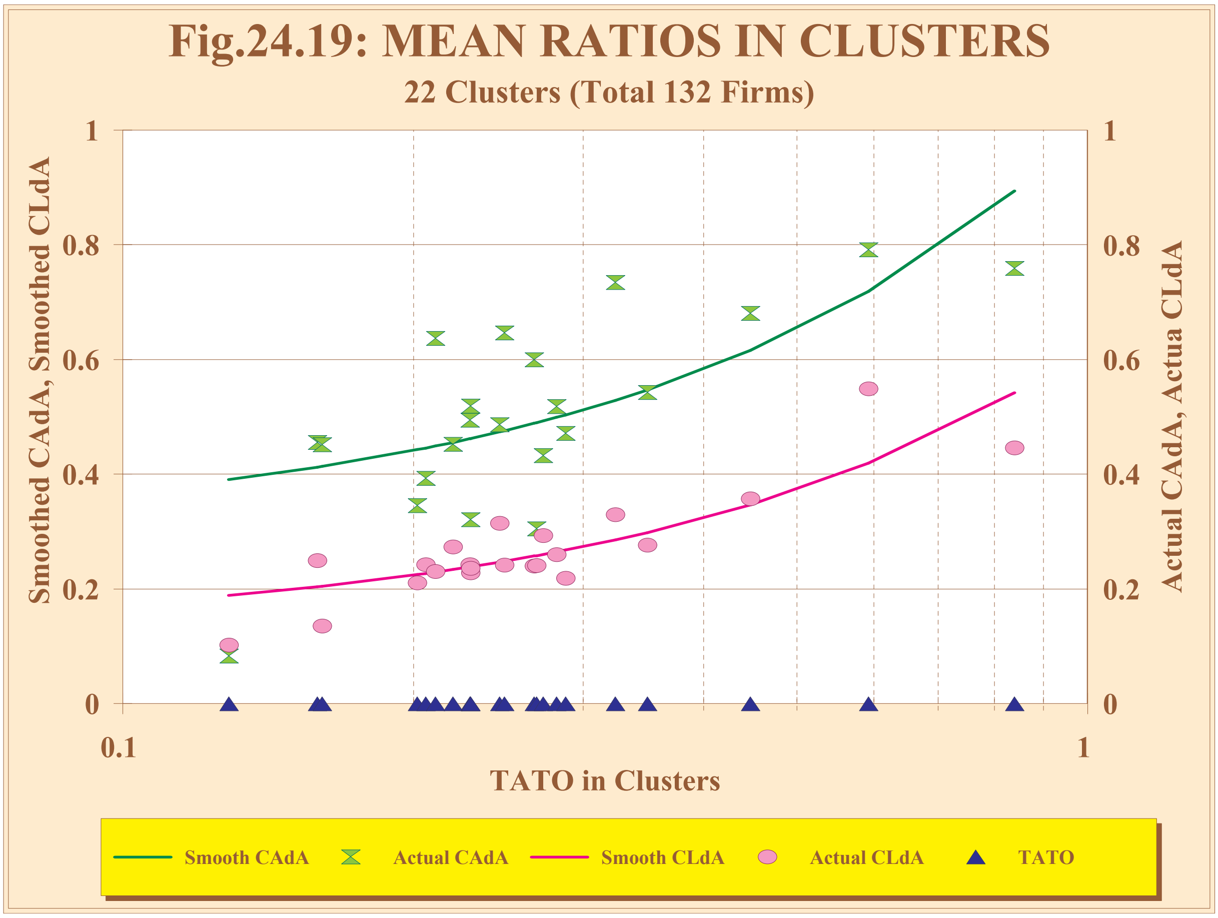
Fig.24.18: COMPARISON OF CLUSTERS



dominated, while clusters CL2, CLF, CL17, CL11, CL1 and CL4 are CL/TA-dominated.

Useful information can be conveyed by signs of the contributions. The role of total liabilities (TL/TA) is positive in nearly all the clusters with the exception of CL17, for which the partial impact (coefficient of the regression model) is negative, while the ratio's mean is always positive. The $TATO$ ratio is also always positive, but its partial impact is negative in 12 clusters; this ratio must compensate for the strong positive effects of the others. The prevailing role of CA/TA and CL/TA is positive. This type of information can be valuable to a financial manager as he tunes the intensity and direction of his control activity.

The high variability, which was noted, might give an initial impression, that there is no regularity in the relationship between the financial ratios. However, it is shown in Fig. 24.19, that a regularity does exist, but that it is a general tendency rather than a strict interdependence.



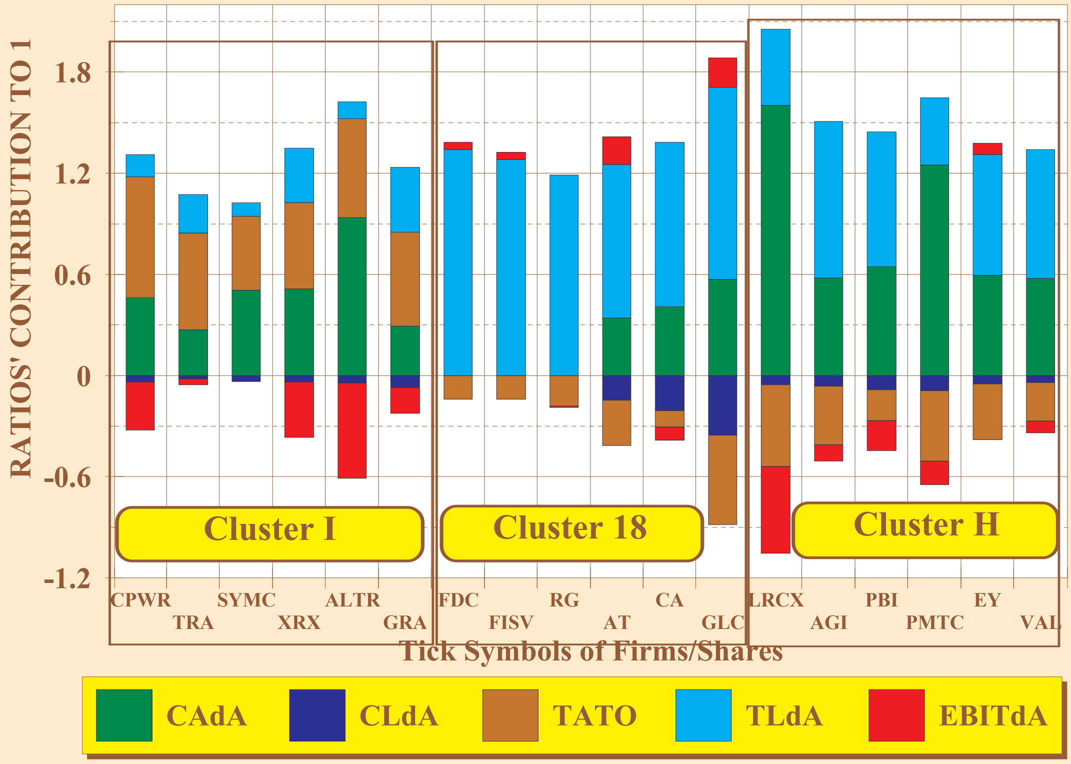
The smooth lines were obtained by a linear fit, which approximates the dependence of CA/TA and CL/TA on $TATO$ (their curvature is due to the logarithmic scale applied to $TATO$). The economic interpretation of these results is straightforward:

1. The faster total assets turnover, the higher the current assets and current liabilities that must be kept available.
2. The general tendency is to maintain a reasonable level of relative working capital equal to the difference $CA/TA - CL/TA$ (the space between the green and red lines).
3. The general tendency is, that working capital increases with increasing $TATO$ from about 0.2 to more than 0.3.

A close look at individual pairs of CA/TA and CL/TA in Fig. 24.19 justifies the use of the expression ‘general tendency’, because the relations between these quantities in individual cases vary significantly. Examples of the variability of ratios within a cluster are shown in Fig. 24.20.

Fig.24.20: STRUCTURE OF 3 CLUSTERS

Clusters of Firms in Two US Industries



Firms in clusters I, 18 and H earn approximately the same return $EBIT/TA$, but the makeup of their ratios is different: the contributions of $TATO$ are large and positive in CLI, but negative in CL18 and CLH. The role of TL/TA is dominating in CL18, but much weaker in CLI. The most intense effect of CA/TA is in CLH. It is also obvious, that membership in a certain cluster does not mean, that the right-hand value, 1, is obtained in the same manner, the similarity of 'patterns' of the contributions is volatile, although the partial impacts (regression coefficients) are identical for all firms in each cluster. These are all shown in Tab. 24.10 together with the impacts in clusters CL4, CLE, CL17 and CL21, to which CF1, CF2, CF3 and CF4 belong.

To complete the comparisons, the mean values of the ratios in these same clusters are given in Tab. 24.11. Note, that ratios in the last four lines of Tab. 24.11 are **not** the individual values of CF1, CF2, CF3 and CF4 for the first quarter of 2001, but the arithmetical means over the clusters, to which these companies belong.

Cluster	CA/TA	CL/TA	$TATO$	TL/TA	$EBIT/TA$	**Hint**
CLI	1.09	-0.17	3.55	0.37	-10.89	In Fig. 24.20
CL18	2.64	-1.15	-1.76	1.67	2.67	In Fig. 24.20
CLH	1.93	-0.23	-1.51	0.95	-7.00	In Fig. 24.20
CL4	1.48	-0.17	-0.69	0.41	6.69	CF1's cluster
CLE	0.97	-0.64	0.33	0.87	0.28	CF2's cluster
CL17	2.23	-0.95	0.08	-0.82	6.99	CF3's cluster
CL21	0.14	-6.92	10.71	1.38	-4.59	CF4's cluster

Tab. 24.10: Partial impacts of ratios in several clusters

Cluster	CA/TA	CL/TA	$TATO$	TL/TA	$EBIT/TA$	**Hint**
CLI	0.456	0.250	0.159	0.559	0.021	In Fig. 24.20
CL18	0.083	0.103	0.129	0.681	0.021	In Fig. 24.20
CLH	0.452	0.274	0.220	0.710	0.022	In Fig. 24.20
CL4	0.518	0.260	0.282	0.517	0.039	CF1's cluster
CLE	0.735	0.330	0.324	0.438	0.010	CF2's cluster
CL17	0.759	0.446	0.842	0.553	0.017	CF3's cluster
CL21	0.452	0.136	0.161	0.243	0.040	CF4's cluster

Tab. 24.11: Arithmetical means of ratios in several clusters

The data presented in both tables demonstrates, why and how these four companies are not comparable from the point of view of their financial statements.

24.4.6 Interval Analysis of Ratios

The notion of interval analysis was introduced and explained in Chapter 16, section 16.4. It is a method, which examines the reactions of the local distribution function ELDF to the extension of the data sample by an additional varying datum. Due to the special kind of robustness associated with the ELDF, its location parameter (quantile of the local maximum of probability density denoted $Z0$) remains in the *toleration interval* $[Z0L, Z0U]$ even if the additional datum were to vary from $-\infty$ through $+\infty$. Moreover, the location parameter's reaction to an increasing added datum typically is to move in the same direction as the added datum only if the datum's value does not exceed the bounds of the *typical interval* of the data sample (ZL and ZU). All the typical points can be defined mathematically and an algorithm can be used for their estimation. These bounds for the data of quarter 1Q1 are summarized in Tab. 24.12 together with the lower and upper bounds (LB and UB) of the data support. The latter values were estimated by using the distribution function EGDF along with the scale parameter used to calculate the bounds of the other intervals. Given the methodology, that was used to establish the clusters, it is not surprising to find, that all five ratio samples were found to be homogeneous, although the data stem from two entirely different industries.

Point	CA/TA	CL/TA	$TATO$	TL/TA	$EBIT/TA$
LB	0.005	0.047	0.001	0.000	-0.115
ZL	0.385	0.203	0.159	0.384	0.0030
Z0L	0.505	0.263	0.225	0.532	0.0187
Z0	0.509	0.266	0.227	0.535	0.0189
Z0U	0.512	0.268	0.229	0.538	0.0191
ZU	0.659	0.353	0.315	0.702	0.0357
UB	2.922	5.006	1.812	1.363	0.558

Tab. 24.12: Bounds of data intervals

The table demonstrates how narrow the toleration interval is: the location parameter is very robust to large changes in the added datum. The intervals of typical data are relatively broad, their width depends on the variability of data. The financial parameters of the firms in each cluster can be classified and visualized by introducing the following symbols to represent the bounds (R represents the ratio's value):

LT:= $ZL \leq R < Z0L$. . . lower typical value,
CT:= $Z0L \leq R \leq Z0U$. . central value,
HT:= $Z0U < R \leq ZU$. . . higher typical value,
$> T$:= $ZU < R < UB$. . . over typical value.

An interval analysis applied to the ratios from the financial statements of the two US industries for quarter 1Q1 resulted in a distribution of ratios with the intervals shown below in Tab. 24.13.

Interval	CA/TA	CL/TA	$TATO$	TL/TA	$EBIT/TA$
$< T$	38	51	32	32	31
LT	36	21	33	39	40
CT	0	2	0	0	0
HT	29	30	31	38	31
$> T$	33	32	40	27	34

Tab. 24.13: Incidence of ratios within specified intervals

With a total number of $136times5$=680 occasions for 332 cases (roughly 1/2) the ratios fell into the interval of typical data. In 184 cases the ratios were under typical, and in 166 cases they were over typical. The distribution is thus not far from being symmetric.

Using the firms shown in Tabs. 24.9 through 24.11 as examples:

Firm	CA/TA	CL/TA	$TATO$	TL/TA	$EBIT/TA$
CF1	LT	$>T$	HT	$>T$	HT
CF2	$>T$	$>T$	$>T$	HT	LT
CF3	$>T$	$>T$	$>T$	HT	$>T$
CF4	$<T$	$<T$	LT	$<T$	$>T$

Tab. 24.14: Classification of ratios of well-known companies in the US High-Tech Industry

From the point of view of interval analysis, the main (and important) difference between CF2 and CF3 is in their return on assets: with similar (relatively high) levels for four ratios, CF3 obtained a higher return.

Not less interesting is the case of CF4, which had a return greater than the typical bound although the four 'explanatory' ratios were rather low.

The results of interval analysis, although based on exact numerical analysis, provides a qualitative evaluation of a firm's financial position. However, the advantage of this approach lies in its synoptic form.

24.5 A Case Study: Dirty Financing

The power of advanced financial statement analysis can be demonstrated by an application to an actual problem:

Countries, that pass through the transient phase from socialism to a market economy must overcome problems not only of an economic nature, but also of a moral and judicial character. A suitable example is the "invention" of an alternative way of financing a commercial activity by using "cheap credit" hidden in the slow payment of liabilities. Countries, which have a long history of an established market economy generally have regulatory provisions to protect against such misuse of commercial credit relationships. These include contract discipline, legal restrictions and the development of a business philosophy, which imbeds not only the Old Testament rule "an eye for eye, and a tooth for a tooth," but also the New Testament Golden Rule "do unto others as you wish them to do unto you." However, in countries with transitory economies, both legal and moral environments are (at least temporarily) in flux and "dirty" financial manipulations are more common. Even where capitalism has long thrived, changes in the structure of the economy open similar loopholes, which are rapidly exploited[8]. This generally goes far beyond 'zero balance' accounts

[8] In Europe, after WW-II, when the individual remained pretty much on a cash basis, the expansion of the food distribution industry and the early emergence of supermarkets, for instance, allowed these

or the relatively rare practice of using a distant subsidiary to handle local payables. But the idea is still the same: to expedite collections and to slow down outflows thereby maximizing the float. Such abuses would be extremely expensive in an established economy, both from the legal standpoint as well as from retaliation by commercial associates. Moreover, it goes against the "going concern" generally accepted principle of accountancy, which assumes the long term viability of the firm and the good will necessary for this to come about. But the penalties are nowhere near as severe in countries, where a large, fast, one time profit is possible, because firms were established with that idea in mind, and where a court takes years to decide commercial conflicts.

The most efficient protection against such undesirable trading partners is to recognize them in time and to avoid them. Such identification is possible by using financial statement analysis. The amount of current receivables CR divided by operating costs OC incurred for the accounting period (say, a year) represents the time of a current receivables cycle, the reciprocal value of current receivables' turnover. Current liabilities CL divided by operating income OI is an estimate of the duration of a revolution of liabilities, the reciprocal value of current liabilities turnover. The lower CR/OC, the faster the payment for a firm's products/services, the more desirable the clients. The lower CL/OI, the better the behavior of a firm with respect to its "creditors."[9] The ratio

$$RI = \frac{CL/OI}{CR/OC} \tag{24.17}$$

can be called *relative insolvency.* The ordinary notion of insolvency is defined ([25]) as:

> *Inability of a person to pay debts as they fall due. This is sometimes called <u>practical</u> <u>insolvency</u>, when the term <u>absolute</u> <u>insolvency</u> is used to mean, that liabilities of a person are greater than the assets of the person.*

The case characterized by the ratio RI differs from the definition cited

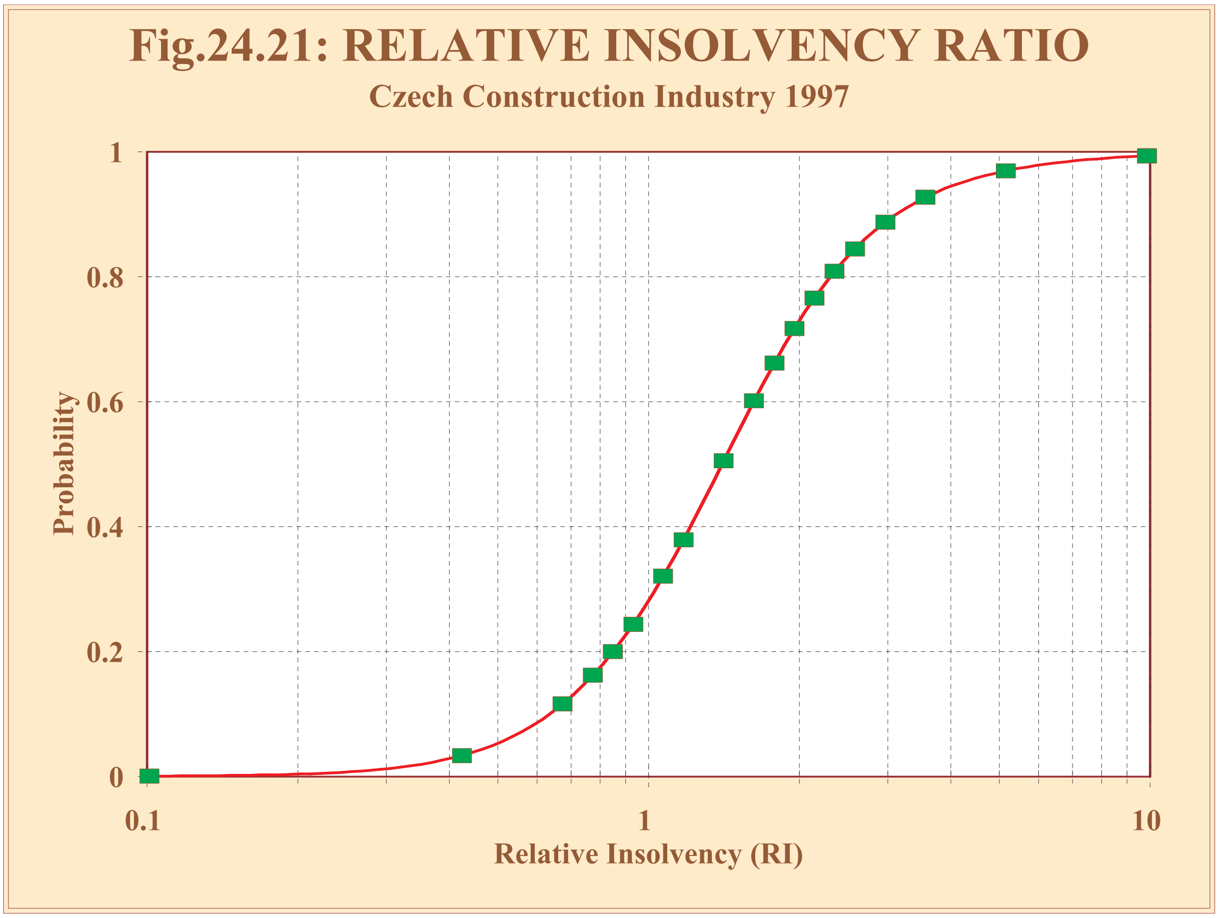
institutions to use 90 day commercial credit to fund their rapid growth at very little cost until the use of individual credit became more prevalent. The enactment of Commercial Credit Codes lead to a reduction in the opportunity for abuses of this kind.

[9] Many firms use 'standard measures' to track their performance in these areas: the Collection Period $(AR/(S/360))$, where AR is the value of receivables, and S is sales in the period in question. The analogue for payments is $(AP/(COGS/360))$, where AP are accounts payable and $COGS$ is cost of goods sold. Because these ratios are not very useful in cases, where sales vary substantially over time (seasonality or other reasons), there are variants to this, which relate receivables/payables to sales/$COGS$ for the month, in which they were created.

above for insolvency, because the debtor **is able to pay, but is unwilling** to do so. The applicability of this type of insolvency is thus more subjective than it is objective. Therefore, the lower the value of the ratio, the more solvent the firm in terms of its ability and *willingness* to pay its debts.

A numerical example based on data from a sector of the Construction Industry in the Czech Republic for 1997 is used to illustrate the problem as well as to suggest a potential means of solution by using advanced financial statement analysis[10].

The global distribution function EGDF of the relative insolvency (RI) in Fig. 24.21 shows, that the problem really exists, as only about 30% of the values of RI fell under 1.



This means, that in more than 70% of the firms, creditor-debtor relations could be characterized as using their partners as a source of 'dirty

[10] The data do not cover all firms in the industry, but only 29, which used the financial services of a bank, which was kind enough to make the statements for 1997 available for the analysis.

credits.' The green squares symbolize the positions of individual firms identified by the bank's client numbers (the data for three firms out of a possible 32 were not complete.) Recall, that $RI = 1$ corresponds to equal turnover of current receivables and liabilities; the larger RI, the more abusive the policies of the debtors. As the figure shows, the imbalance between turnovers can be really significant with RI far exceeding 1. A multidimensional cluster analysis can be enlightening as the following implicit regression model demonstrates:

$$K_1{*}RI_k{+}K_2{*}TL/NS_k{+}K_3{*}TATO_k{+}K_4{*}CE/CL_k{+}K_5{*}EBIT/TA_k = 1, \tag{24.18}$$

where $k = 1, \ldots, 29$, and where the listed financial relations were used:
$RI \ldots$ relative *insolvency* index 24.17,
$TL/NS \ldots$ total liabilities divided by net sales, ie the estimated *lifetime of debt*,
$TATO \ldots$ total asset turnover (net sales divided by total assets), a measure of *activity*,
$CE/CL \ldots$ cash and cash equivalents divided by current liabilities, a measure of *liquidity*,
$EBIT/TA \ldots$ earnings before interest and taxes divided by total assets, a measure of *profitability.*

The mean value of each ratio in every homogeneous cluster divided by the median of all the clusters is shown in Fig. 24.22.

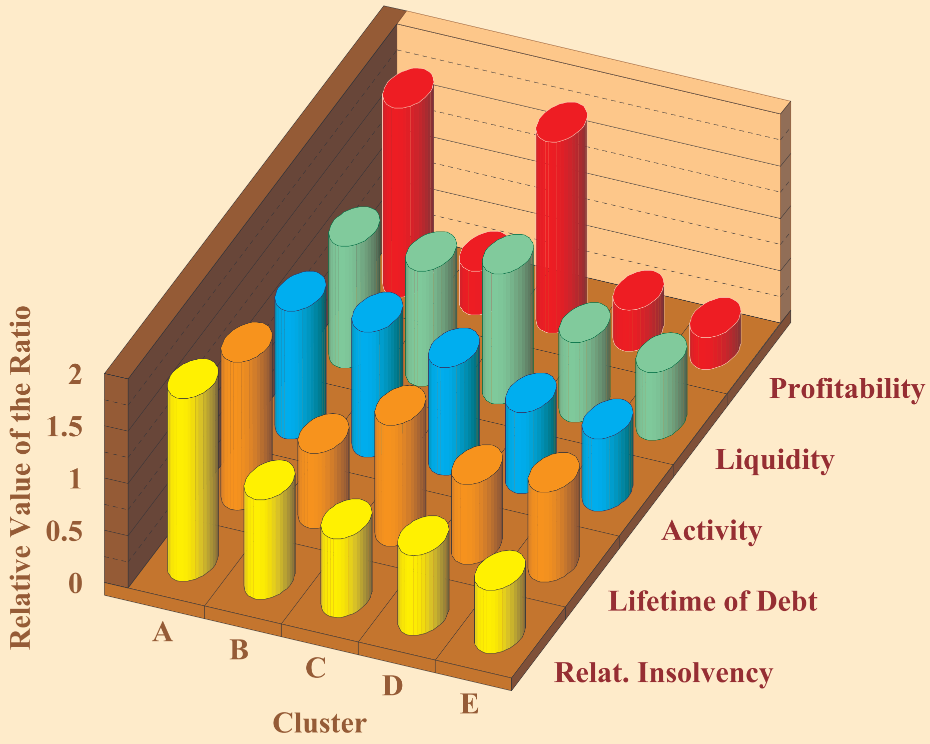
Clustering revealed earthshaking differences in the behavior of firms belonging to different clusters:

1. Firms in cluster A manifest the best financial position in the sense of having maximum return on assets and the highest liquidity and activity. In spite of this, these firms are the worst partners in terms of highest relative insolvency: they could pay, but don't want to. The longest life of debt signifies high financial leverage and supports the impression of the risky behavior of the A cluster.
2. In contrast, firms in cluster C, which have practically the same financial position including financial leverage display much better consideration toward their creditors with a lower RI.
3. The lowest RI is found in cluster E, the firms of which are in the worse financial position with respect to their core ratios.

Since the decomposition of firms into clusters is known, further information can be obtained by calculating the distribution functions EGDF of

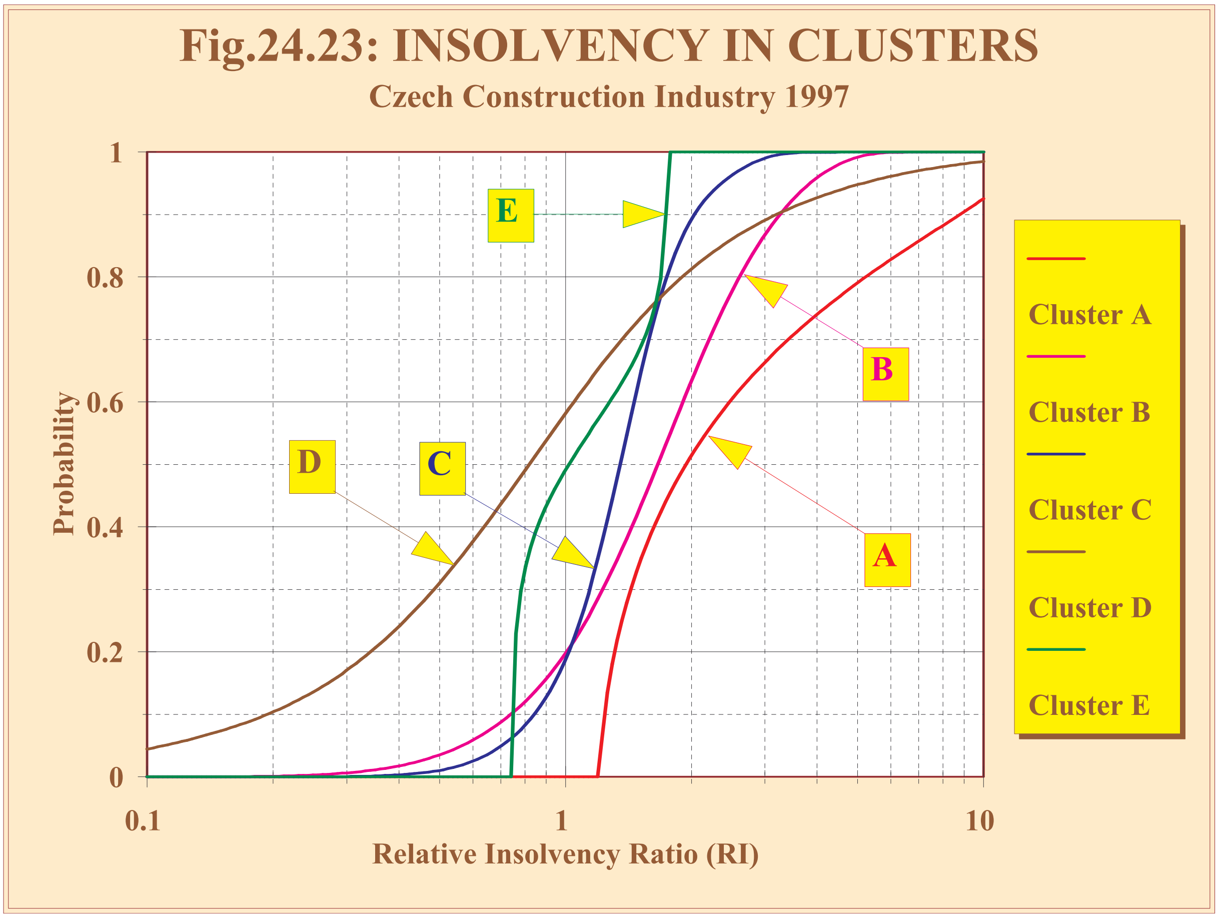
Fig.24.22: RATIOS IN CLUSTERS

Czech Construction Industry 1997



RI within each cluster. These results are in Fig. 24.23:

1. The distribution of RI for cluster A affirms the worse expectations: the probability of RI reaching a value of 1 or less is **zero** and it is very high, that extreme values of RI will be attained; eg the probability of exceeding a value $RI = 9$ is about 0.1!
2. The distribution functions reveal more detail of the firms' character than the mere mean values of RI: Members of cluster C with the second best financial performance and a modest mean value of RI leave 80% of their creditors waiting for much longer than they themselves wait for the payment of their receivables. The percentage in cluster B is the same, but their creditors' wait is even longer. The most favorable probability (about 0.6) of having an RI of less then 1 is for firms of cluster D, but their distribution function has the largest spread on both sides of 1: such high volatility can be interpreted as doubtful reliability for prediction.
3. A nearly ideal and balanced behavior is shown by the distribution



function of cluster E, which has an unusual form, sharply bounded from both sides. Large deviations from 1 thus have a probability of zero.[11]

A potential associate might interpret the above as:

1. Stay away from firms of type A: they are dominated by self interest and will make poor trade partners.
2. Be cautious in developing relations with types B and C, particularly the former, due to their apparent systematic recklessness.
3. While firms of type D have the best overall distribution of RI, the volatility of the group is very high, and it would be wise to examine an individual firm's distribution function over time before entering into a long term commitment.
4. Firms in cluster E are the most stable and appear to display the best long term reliability even though their performance vis a vis their

[11] Note in Fig. 24.22, that the cluster's E profitability was the worse of all clusters. An ugly and sad thought intrudes: doesn't this amount to a penalty for being honest among dishonest fellow-travellers?

financial statements is marginal; ie they may not be doing very well for themselves, but perhaps because they are in need creditor support, they are willing to behave properly.

These conclusions pose a nearly philosophical question: does the best financial position of cluster A and the worst for cluster E come about **in spite** of the bad/good ethical policies of the firms or **because of** this behavior? There is only one certainty: such large scale of behavioral patterns cannot have a long life.

The importance (or weight) of each of the five ratios with respect to each other can be seen in Fig. 24.24.

One last thought can be tendered with respect to the operating policies cited before leaving this example: The figure shows the proportional importance of the elements of the implicit equation 24.18. The strongest impact is that of relative insolvency, which points to the fact, that testing for the presence of such a “phoney insolvency” should not be neglected in

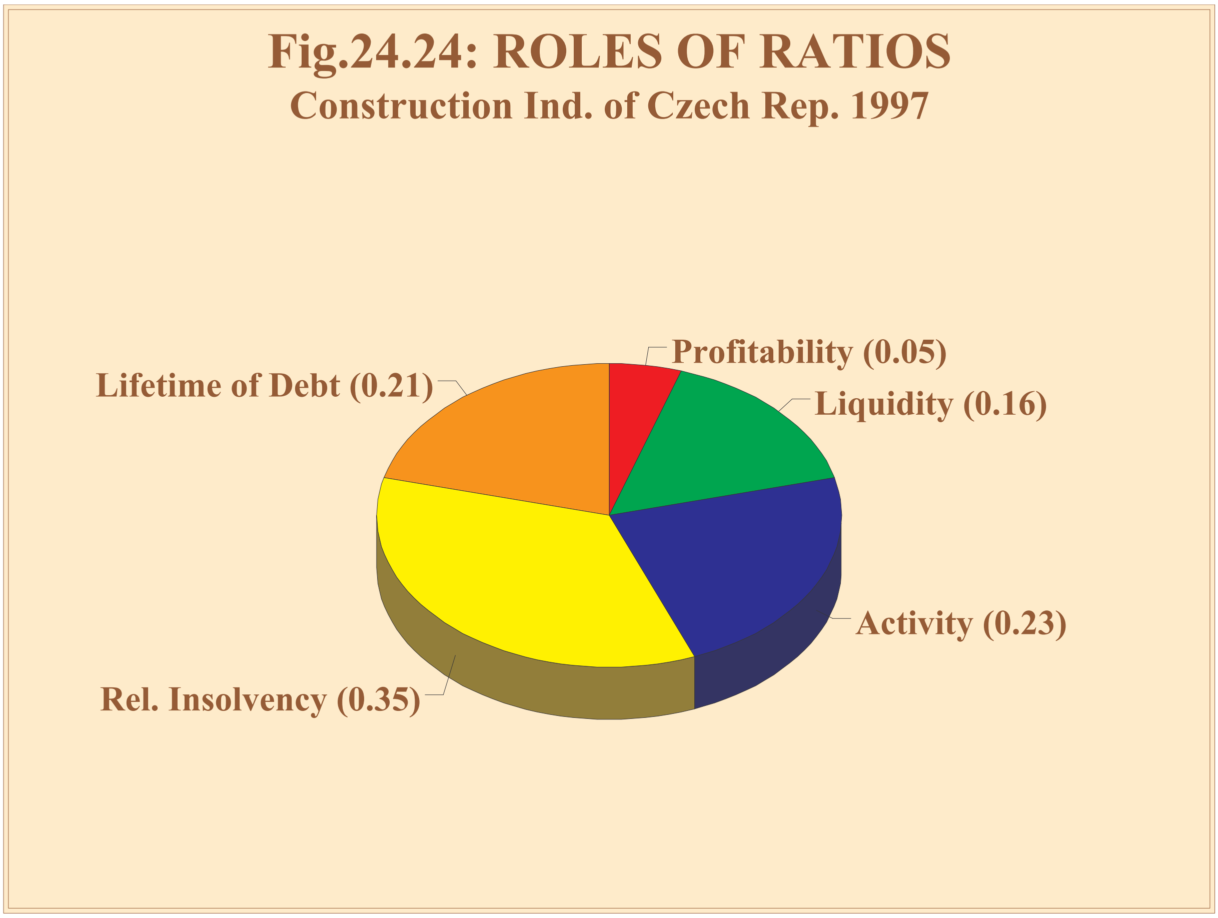


Fig.24.24: ROLES OF RATIOS
Construction Ind. of Czech Rep. 1997

the historical evaluation of the Czech economy, and indeed it could prove valuable in the analysis of other economies in transition.

24.6 On the Internal Information System of a Firm

The recent state of information technology is advanced enough to make the idea of a firm without an internal information system unthinkable. Highly developed information systems are offered on the market and high costs of both purchasing and operating these systems are generally considered as being fully compensated by the convenience they provide to firms' management. However, a critical analysis of the functions of these systems results in a certain degree of scepticism: they suffer from imperfection in the final treatment of the rich databases they create and maintain. They ensure collection of data on all levels of a firm's activity, classify the data and accumulate them in databases, make data available both for analysts and some relatively simple automated functions such as report or statement composition, but the high level of automatic data processing currently available for application to technological control and information systems is far from a reality in the present state of business information systems. The functions of a module of financial statement analysis (if any exists) are limited to what was critically analyzed in Chapter 22.

One should expect, that the information system of a firm based on recent know-how in computers and networks automatically includes a module to perform the tasks of

1. evaluating the financial position of the firm,
2. preparing recommendations to assist financial managers in optimum decision making of the financial management,
3. monitoring cash flows,
4. automatic warning of unusual or even dangerous states of the firm's economic status,
5. both operational and long-term predictions,
6. etc.

In other word, such a module should perform the functions of advanced financial statement analysis and other functions illustrated by examples provided in this book. It can be expected, that performing these tasks will require (among others) the 'continuous' (as frequent as possible) monitoring the firm's multidimensional data series as well as checking the firm's relative financial position within the cross-section of its economic environment.

24.7 Summary

A broad spectrum of the tasks of advanced financial analysis can be solved using the means provided by robust gnostic methodology, especially the use of gnostic distribution functions and robust regression models (both explicit and implicit types) together with a regression in probabilities:

- Robust multidimensional modeling of interactions between the financial parameters of firms,
- multidimensional ordering of the financial position of firms leading to an objective (mathematical) rating,
- reliable and sensitive cross-section analysis of groups of firms,
- robust monitoring of multidimensional time series,
- multidimensional cluster analysis of groups of firms resulting in subgroups of economically comparable firms,
- classification of the financial positions of firms based on robust interval analysis,
- creation of software capable performing the tasks of advanced financial statement analysis in the environments of the internal information systems of firms.

The real effects of gnostic multidimensional analysis were demonstrated by examples of applications to real data.

Chapter 25

Contributions to Market Analysis

25.1 Industry Comparisons

25.1.1 Cross-section Analysis

Recall, that the first stage of the cross-section analysis (the marginal analysis in section 23.2.5) dealt with the probability distribution of a given variable (eg a financial ratio of a given type) estimated using data taken from a group of objects (eg an industry) at a fixed point in time (eg a year or a quarter). If they are needed, homogeneity tests are run on the data samples at this stage and possible inhomogeneities are eliminated by means of marginal (univariate) cluster analysis. The second stage of the cross-section analysis focuses on the time series of these marginal (one-dimensional) distribution functions. The most detailed insight into the structure of the cross-section data can be obtained by using a multidimensional structure analysis with robust modeling techniques.

The application of marginal analysis to illustrate the development of the distributions of the total returns (TR) of the US Chemical Industry is illustrated in Fig. 25.1 in a manner similar to that used in Fig. 23.8.

The quarterly data of 36–38 companies make up the data set and the time period covered spans 90Q2–01Q2[1]. Time is on the horizontal axis of Fig. 25.1, while the vertical axis measures the probability of values of total returns TR expressed in percent, ie as $100 * Probability(\%)$. The lines drawn on the graphs connect points/quantiles, which correspond to one of nine levels of probability: the lowest magenta line thus traces values of TR not exceeded by 2% of the companies in the US Chemical Industry

[1] Symbol 90Q2 again stands for the "Year 1990, Quarter 2."

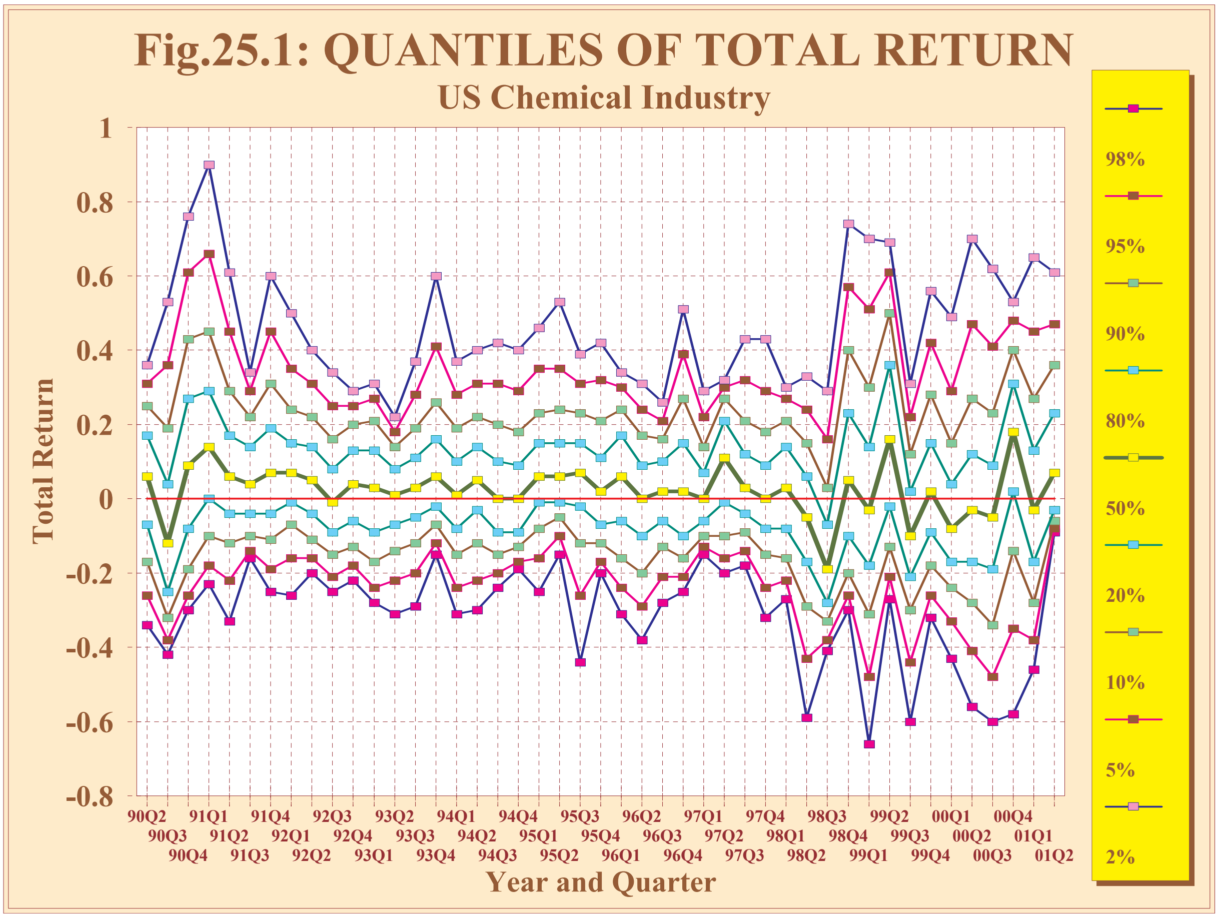


Fig.25.1: QUANTILES OF TOTAL RETURN

US Chemical Industry

during any quarter. The bold green line corresponds to the medians of the distribution functions, ie it connects the quantiles of probability 0.5. The vertical distance between quantile lines measures the interval between values of TR (their 'spread' in percentage form). So eg the vertical distance between the 98% and the 2% quantile lines represents the interval of TR, which is covered by 96% of the companies in the Industry. A glance at this family of quantile lines provides a general view on the development of the TR over the time period 1990Q2–2001Q2: the median of the TR traces a slow and small oscillation about its long-term arithmetic average of 0.071 (the long-term median is 0.082). The marks on the graph's lines for each quarter show 9 values of the probability distribution function representing the time at Q. The following general characteristics are observed:

Stationarity: The process cannot be taken as stationary: all the quantile lines have too much variability; relatively quiet periods such as 94Q1–96Q1 are followed by turbulent periods similar to 90Q2–92Q3 or 98Q1–01Q2.

Changing distributions: The forms of the cross-section distribution functions can change both smoothly (91Q4–93Q2) and discretely, or suddenly (98Q3–98Q4, 99Q2–99Q3 and 01Q1–01Q2).

Symmetry: The cross-section distributions can be nearly symmetrical (98Q1) as well as strongly asymmetric (91Q1, 93Q4, 01Q2).

The complex variability of the process does not permit the volatility to be described by a single number such as variance.

It also follows, that the process cannot be characterized by using an a priori 'prescribed' and easily parameterized distribution function of a standard type or—in the even worse case—by numerical statistics such as means, variance or others. Graphs in the form of Fig. 25.1 are more universal, and they bear much more information.

25.1.2 Comparison of Industries by Cross-section Analysis

The analytic results obtained in the preceding sections came about through the application of gnostic algorithms to data representing the US Chemical Industry. And, it was shown there, that not all companies in this industry are comparable to each other (see section 23.2.7). The industry was made up of several clusters, each of which was formed from firms, that behave in accordance to the same multidimensional model, which binds the firms' financial ratios. The models of the individual clusters differ; this means, that companies belonging to different clusters behave differently in the sense, that they do not respond in the same way to identical changes in the same financial ratios. Intra-cluster comparability thus exists, while inter-cluster comparability generally does not. The intra-industrial comparability of firms becomes a no trifling problem, which calls for the careful analysis considered above. This raises an important and an interesting question: Does inter-industry comparability even exist?

To examine this issue, data characterizing different industries, oriented toward high technology, and therefore substantially different from the Chemical Industry were taken over the same period. The sample was composed of 89–120 firms[2], mainly involved in the consolidation of the crude or natural gas extraction, refining or transmission process, electrical services,

[2] Sample sizes (number of firms with a complete set of available data for a quarter) varied nearly each quarter.

computers, and various aspects of wireless technology.

A graphical presentation will be used for the first stage of the cross-section analysis. An important feature of such a summary for numeric data is, that it is more easily perceived than a numerical representation: people are 'not designed' to view digits like computers. However, people are endowed with high-capacity, high-resolution, broad-band optical information channels and a brain to analyze an image in the blink of an eye. A careful observer can therefore easily discern both similarities and differences between the two industries by comparing Fig. 25.1 with the analogous graphs of the US High-Tech Industry, Fig. 25.2:

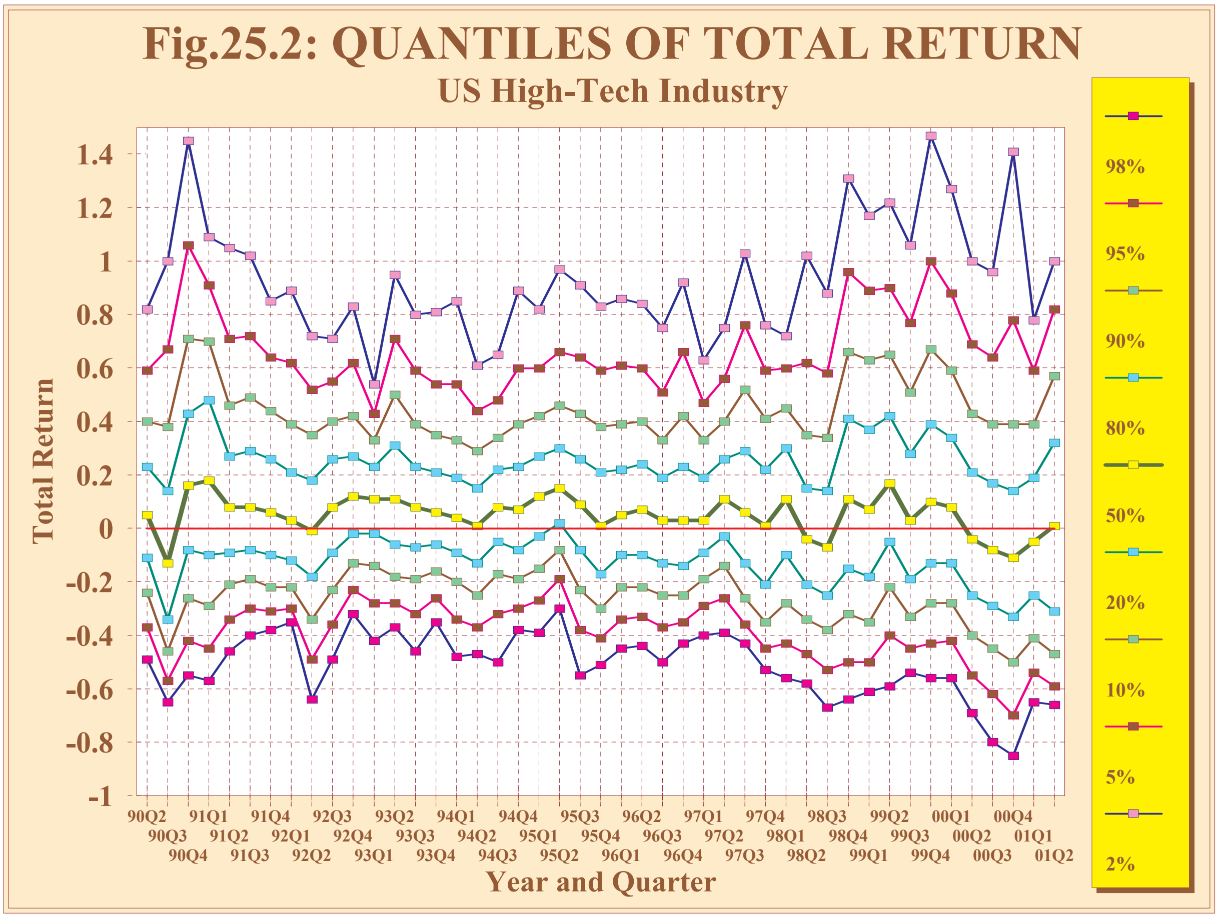


Fig.25.2: QUANTILES OF TOTAL RETURN
US High-Tech Industry

Volatility: The variability of the TR for Hi-Tech is much stronger.

Performance: The bold green lines of the medians show (at least over the period 90Q4–98Q1) a systematically better level of TR in High-Tech than for Chemistry (the long-term average over the whole interval 90Q2–01Q2 for the two groups of firms is respectively 0.0774 and

0.0309).

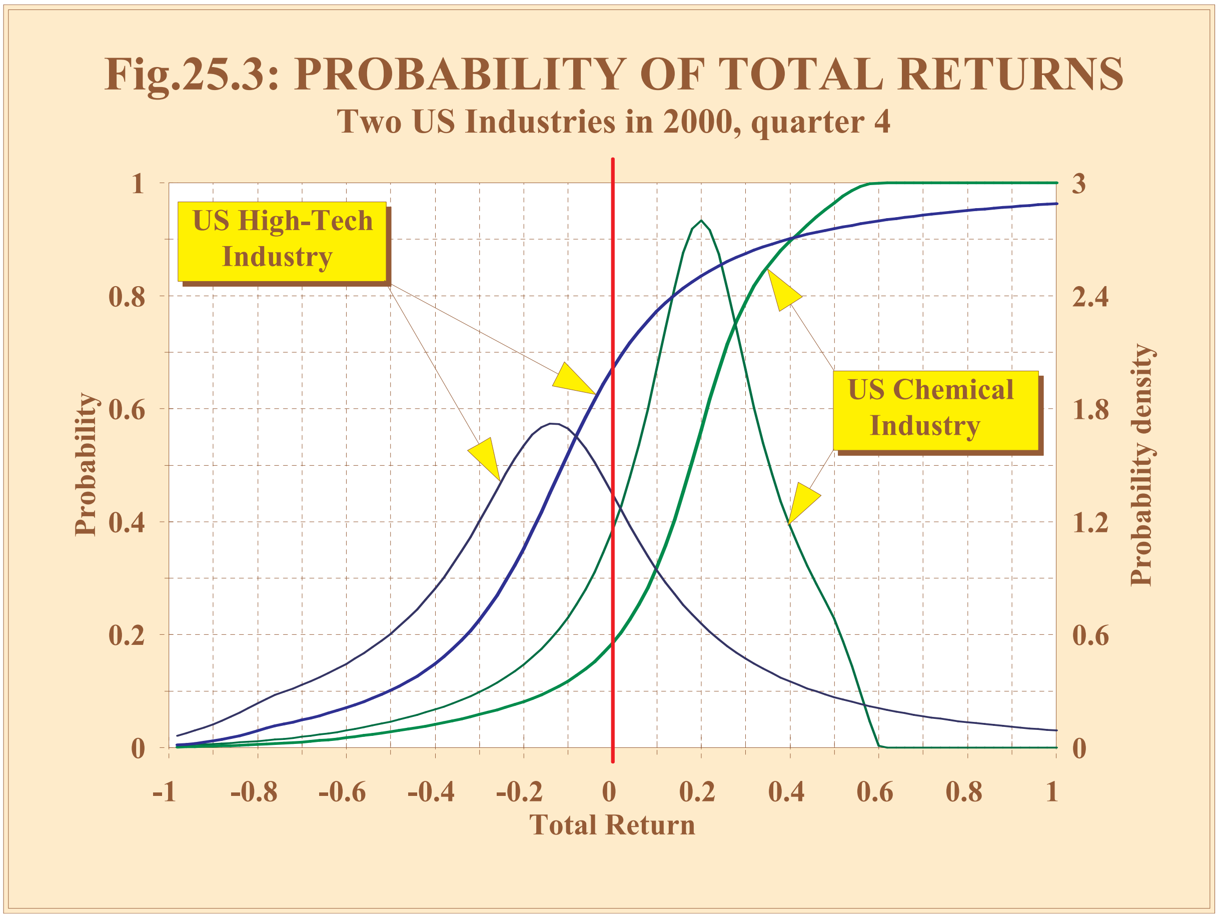
Common external impacts: In spite of large differences in economic and technological orientation, both industries are obviously subjected to strong impulses initiated by the overall economic system. This is manifested by nearly synchronous reactions in both graphs (90Q2–90Q4, 98Q3–98Q4, 99Q3–99Q4) to the same external impulses as well as by the striking similarity of the forms of median lines over 90Q2–99Q4.

Phase shift: The reactions of the Chemical Industry are not always strictly synchronous to those of the High-Tech Industry: the former lagged the latter about a year and a quarter over the whole period 91Q3–95Q2 attaining a synchronous movement over 95Q3–99Q1, but with an amplified sensitivity, manifested especially by steeper losses in the case of Chemistry (98Q2–98Q3 and 99Q3) then reaching the level of High-Tech in 99Q2. The comparative behavior changed further after 00Q1, when High-Tech began its long lasting slump, while Chemistry behaved in a relatively relaxed manner and then rose to a record level in 00Q4. It is also worth noting, that in 01Q2 a much larger portion of the Chemical Industry had positive TRs, while gains and losses of the High-Tech Ind. were still well out of balance.

A more detailed comparison of both industries using distribution and density functions of the total returns for 2000 is in Fig. 25.3.

This example shows, that a 'black-and-white' conclusion of the type 'Industry A performed better than Industry B in 2000' cannot be drawn:

1. Chemistry's results had a steeper probability distribution and a sharper and more concentrated density function with a maximum at $TR = 0.199$, while the same location parameter for the High-Tech was -0.132. Hence, Chemistry's expected performance was much better and less volatile than that of High-Tech.
2. The probability of falling into the 'red numbers' (TR negative or zero) was about 0.18 for Chemistry and 0.67 for High-Tech. Hence, it is expected, that Chemistry will turn in a better performance.
3. Chemistry's probability density falls to zero at about $TR = 0.6$, while there is no such a cut-off in the case of High-Tech. The probability of exceeding $TR = 0.6$ is zero for Chemistry with about 7% of High-Tech's firms having a chance to reach these much higher returns. Hence, in this respect, Chemistry was worse.



The drastic change in behavior of the two industries, which took place in 98Q1, can also be seen in Fig. 25.4: the robust means of stock prices—although on differing levels—ran parallel to each other during the long period 90Q2–98Q1.

The fast fall of the High-Tech Industry in 98Q1–98Q2 was matched by only a modest decline in Chemistry's prices, after which the prices movements diverged: over (98Q3–00Q1) chemical stock prices rose rapidly, while the technical firms continued to decline. A further deterioration of High-Tech's prices in 99Q4–00Q3 was followed by Chemistry, but lagged by a quarter: the rapid return of technological equities to their former highs in 00Q3–01Q1 was not imitated by the chemical group.

These examples show how useful the modern tools of analysis can be in announcing, that **something** indeed did happen, and **what** and **how** it happened in processes documented by data. The first portion of the analytic effort can be made efficient in this manner. However, there remains

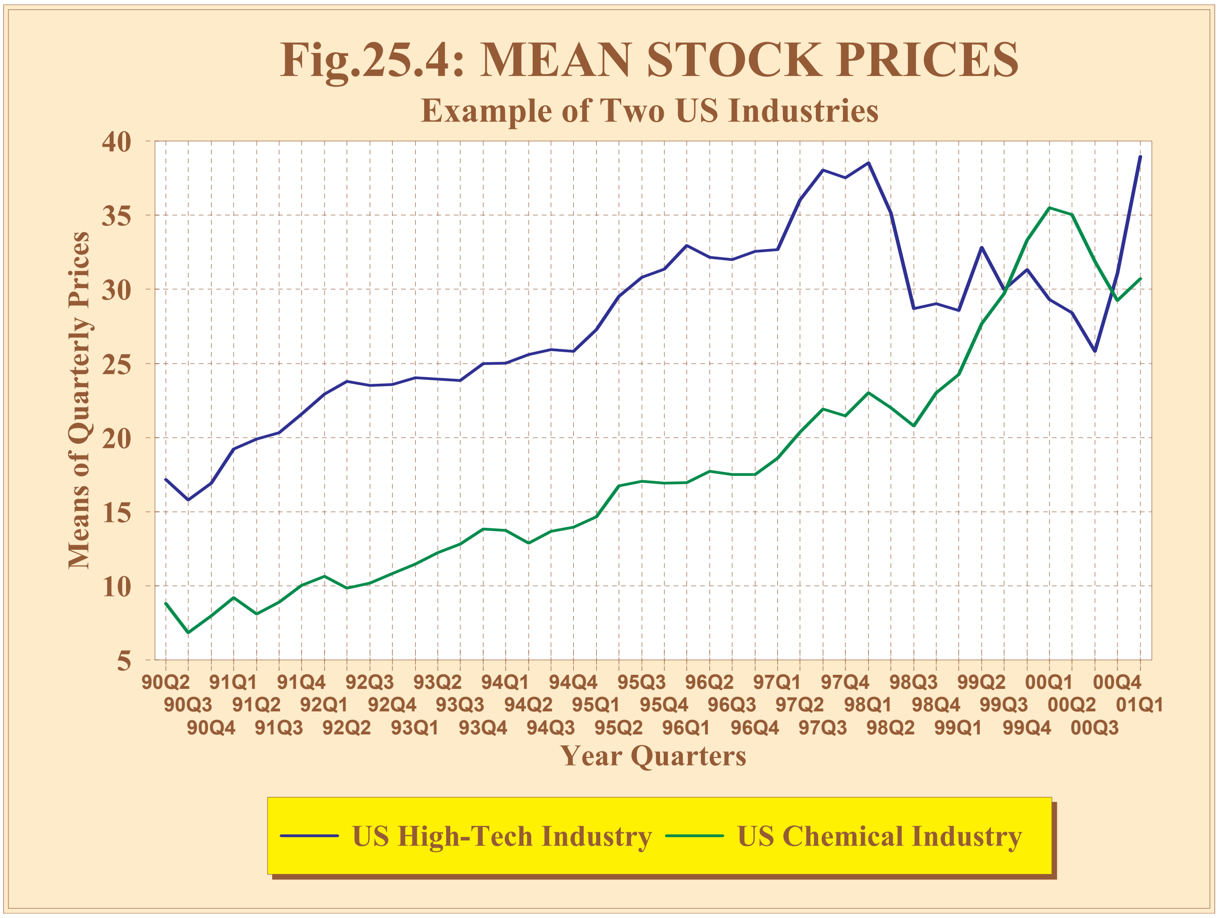


Fig.25.4: MEAN STOCK PRICES

Example of Two US Industries

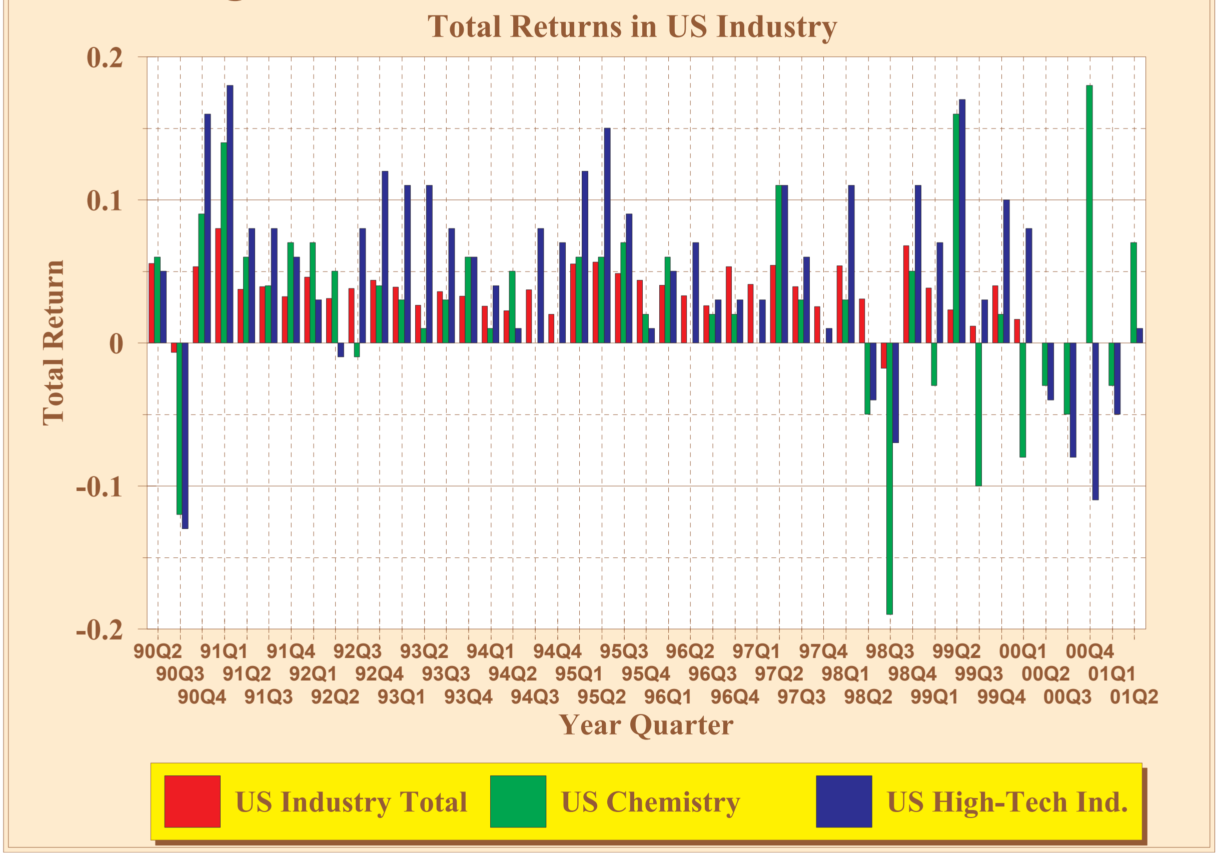
another task, to **explain why** an event occurred. To do this, further effort is necessary and it requires both additional data and advanced analytic tools.

25.1.3 The Marginal Comparability

To demonstrate some of the differences between the Chemical and Hi-Tech Industries, Fig. 25.5 compares the evolution of the quarterly total returns in these sectors together with the total of all US industry (the red columns).

The total is naturally the most stable, while both industries fluctuate significantly. The High-Tech Industry strongly outperformed the US industrial total in 9 out of 13 quarters, while showing better results than Chemistry in all 13 periods. Chemistry bettered the industrial total in only 4 cases. It can also be observed in Fig. 25.5, that both the High-Tech and the Chemical Industries followed and reacted more drastically to the

Fig.25.5: COMPARISON OF INDUSTRIES



negative development of the US industrial sector (see especially the period 98Q1 through 98Q3 in Fig. 25.5 and the corresponding fall of the quantiles in Figs. 25.1 and 25.2). However, the main point of Fig. 25.5 is, that Chemistry behaved differently from the High-Tech industry and both of them differed from the industrial composite.

This conclusion is further supported by Tab. 25.1, which compares characteristic values of the five fundamental financial ratios. The results from both industries, considered separately, are shown together with the combined ratios. There were 40 companies in US Chemistry and 110 firms in Hi-Tech. The combination was subjected to tests of (marginal, ie unidimensional) homogeneity. Twelve out of 15 samples appeared to be homogeneous. The only samples with outliers concerned the ROA. (Note the different sample sizes in the last column of Tab. 25.1, which presents sample sizes of homogeneous samples and subsamples.)

Symbols:
RWC. . . working capital divided by total assets,

$TATO$... total asset turnover (sales divided by total assets),
TL/TA... total liabilities divided by total assets,
ROA... return on assets (earnings before tax divided by total assets),
TR... total return (relative change in the stock price plus dividend yield),
LB... the lower bound of the data support estimated by EGDF,
Min... minimum of the data sample,
LSB... the lower bound of the sample estimated by EGDF,
$RMed$... robust median (quantile of probability 0.5),
LP... location parameter of EGDF (location of maximum density),
USB... the upper bound of the sample estimated by EGDF,
Max... maximum of the data sample,
UB... the upper bound of the data support estimated by EGDF.
Size... the number of data in the homogeneous sample.

Ratio	**Ind.**	LB	Min	LSB	$RMed$	LP	USB	Max	UB	**Size**
RWC	Ch.	-4.887	-0.240	-2.297	0.136	0.144	0.490	0.485	0.658	40
RWC	HiT.	-0.264	-0.090	-0.264	0.238	0.231	0.888	0.715	0.893	110
RWC	Both	-0.444	-0.240	-0.443	0.207	0.194	0.748	0.715	4.212	150
$TATO$	Ch.	0.000	0.076	0.001	0.257	0.273	0.451	0.444	0.676	40
$TATO$	HiT.	0.021	0.076	0.021	0.257	0.260	2.912	1.554	2.990	110
$TATO$	Both	0.003	0.076	0.004	0.256	0.264	2.955	1.554	2.960	150
TL/TA	Ch.	0.000	0.161	0.001	0.626	0.653	1.488	0.970	1.491	40
TL/TA	HiT.	0.028	0.126	0.126	0.496	0.516	2.974	1.696	2.977	110
TL/TA	Both	0.000	0.161	0.000	0.535	0.564	2.849	1.696	2.951	150
ROA	Ch.	-0.031	-0.014	-0.014	0.021	0.020	0.236	0.053	0.285	39
ROA	HiT.	-0.074	-0.042	-0.074	0.021	0.020	0.137	0.106	0.138	107
ROA	Both	-1.729	-0.142	-0.865	0.021	0.023	0.155	0.106	0.159	149
TR	Ch.	-8.206	-0.634	-3.985	-0.023	-0.011	0.829	0.548	0.830	40
TR	HiT.	-1.022	-0.580	-1.021	0.054	0.008	7.150	1.460	8.724	110
TR	Both	-1.080	-0.634	-1.079	0.024	-0.009	7.195	1.460	8.434	150

Tab. 25.1: Comparison of fundamental financial ratios of two US industries and of their combination, based on data for 1999.
(Ch. ... Chemical Industry, HiT. ... High Technology Ind.)

The most noticeable differences between the two industries are seen in the performance of RWC and TR: all the characteristics of the Chemical Industry are significantly lower than those of the Hi-Tech group. For TR, this finding corresponds to Fig. 25.1, but the pattern in Tab. 25.1 allows a conclusion as to the form of the probability distribution to be made: Neither of the distributions of TR are symmetrical: very low values for the Chemistry and very high ones for the Hi-Tech Industry can be expected.

Very low values of RWC are also more probable in the Chemical Industry, while there are high probabilities of attaining large values of $TATO$ and TL/TA in the Hi-Tech Industry. There are also interesting differences in the ROA: although the means ($RMed$ and LP) of both industries coincide, high returns are more probable for Chemistry. Note, that here is no contradiction between the vividly opposite relationship between the ROA and TR of the two industries: ROA is an objective (accountancy) parameter independent of the market's evaluation, while TR is strongly dependent on the market's 'point of view', which is a product not only of expectations of future performance, but also of the reigning mass psychology, investment 'fashion' and other subjective factors[3]. It is to be emphasized, that this conclusion is valid only for 1999; these relationships are fluid and do change over each year and quarter.

A brief comment with respect to the results obtained for the combination of both industries: they fall in between the two industries, obscuring any distinguishing features of the specific industries together with the opportunity to distinguish their individual identities.

At this point, a tentative conclusion can be put forward, that membership in an industry may be an important specific feature, which cannot be neglected in the marginal analysis of financial ratios, for it can establish a set of bounds for these measures, that apply to that group of firms. From the point of view of marginal analysis, enterprises belonging to different industries are not comparable to each other[4], a specific analysis of the relationship of the financial ratios of each industry is necessary.

25.1.4 Comparison by Correlations

The marginal incomparability of firms belonging to different industries manifested by the substantial differences in the distributions of individual financial ratios as seen in the previous section does not necessary imply multidimensional incomparability. Enterprises are comparable, when they behave in a similar way and they behave in a similar way, if they are subjected to the same multidimensional model of relations between their fundamental financial ratios. When examining different cars, one finds, that the relations between such parameters as the car's weight, the power of its engine and acceleration are similar for different types of cars. This

[3] Recent developments in the market shows, that the figures in Tab. 25.1 can be interpreted as a serious warning, that the market was overestimating the real potential of the Hi-Tech Industry.

[4] At least for the group of industries treated here.

Correlations in US Chemical Industry					
Ratio	RWC	$TATO$	TL/TA	ROA	TR
RWC	1.000	0.237	-0.734	0.360	0.007
$TATO$	0.237	1.000	-0.138	0.098	-0.001
TL/TA	-0.734	-0.138	1.000	-0.532	-0.330
ROA	0.360	0.098	-0.532	1.000	0.139
TR	0.007	-0.001	-0.330	0.139	1.000

Correlations in US Hi-Tech Industry					
Ratio	RWC	$TATO$	TL/TA	ROA	TR
RWC	1.000	0.227	-0.659	0.337	-0.035
$TATO$	0.227	1.000	0.055	0.204	-0.257
TL/TA	-0.659	0.055	1.000	-0.432	-0.105
ROA	0.337	0.204	-0.432	1.000	0.092
TR	-0.035	-0.257	-0.105	0.092	1.000

Tab. 25.2 Robust correlation matrices of fundamental ratios for two US industries, data for 1999)

is a natural consequence of the Laws of Mechanics, the validity of which is universal. Such strict regularities have not been discovered in economics, but this does not mean, that firms can behave quite arbitrarily, without respecting any constrains. This point can be demonstrated by examining the robust correlation coefficients of several financial ratios for firms belonging to the two different industries considered above:

The symbols for the ratios are identical with those in Tab. 25.1. Both similarities and differences are observable in Tab. 25.2. Interdependencies of (RWC, $TATO$), (RWC, TL/TA), (RWC, ROA) and (ROA, TL/TA) can be considered as revealing an analogous mechanism in both industries. On the other hand, correlations (ROA, $TATO$) differ, while interactions (TL/TA, $TATO$) have the opposite sign. However, the behavior of TR differs substantially. The impact of $TATO$ on TR is negligible in Chemistry, but strongly negative for the Hi-Tech Industry. The negative effect of TL/TA on TR is three times stronger in Chemistry than in the Hi-Tech Industry. Therefore, it can be expected, that the multidimensional models, which include TR and the bulk of each of the industries will be parameterized differently. A multidimensional cluster analysis applied jointly to the two industries will be useful in clarifying this point.

25.1.5 Comparison Using MD-models

The problems of multidimensional models were examined in sections 24.3 through 24.5. It was shown, that within an industry, there can be several clusters of firms, which are comparable within their cluster, but not comparable with firms in other clusters. On the other hand, comparable firms operating according to the same MD-model can be found in different industries. This situation does not allow one to speak of the MD-model of an industry nor to compare separate industries by "their" MD-models. It also was shown, that the means of ratios in different clusters can have very broad ranges. It is not the values of the ratios themselves, which are decisive for the inclusion of a firm into a cluster, but the pattern of the whole set of ratios.

These findings are important in financial statement analysis, because to make a judgment as to "good" or "bad" financial ratios, comparability is decided not by membership in an industry, but by belonging to a cluster of firms potentially formed from different industries.

An attractive task, the solution of which could vastly simplify routine procedures of both financial statement analysis and financial management, can now be designed. The idea is to split the process into two stages:

Assembling a catalogue: A general 'stocktaking' of multidimensional models for all the industries by multivariate cluster analysis, which should provide

1. the formulae for the multidimensional models of all the clusters found,
2. the parameters of distribution functions of the models' residuals obtained by application of the models to all members of the cluster,
3. comments on the specifics of the cluster's model and on the performance of each cluster's membership.

This job could be performed by teams of skilled analysts in specialized firms (eg such as rating agencies). The catalogue would take the form of a data base and since all these relations are dynamic, would include a program for its routine periodic revision.

Use of the catalogue: A judgment as to the financial position or with respect to a necessary control action related to a particular firm would consist of following steps:

1. Preparing the data on the firm.
2. Running the program to query the data base to find a cluster of

the most comparable firms:

(a) Step by step substitution of the firm's parameters into the clusters' models to compute the modeling error (residual).

(b) Probability evaluation of the error by using the distribution function of the cluster's residuals.

(c) Determination of the most suitable model, ie the model, for which the tested firm's probability of error would differ the least from 0.5.

It is obvious, that these steps can be automated.

3. Collecting and examining the comments on the specific features of the cluster's firms in the database.
4. Making a decision based on the information gathered.

This stage would not require any special analytical/mathematical skill from the user. Nevertheless, the decisions taken would likely lead to successful outcomes because of the principles of advanced analysis implicitly applied.

The main message of this section is: firm-to-firm as well as industry-to-industry comparability exists and can be effectively used to make judgements on the financial position of member firms, but the real picture is much more colorful than what is seen, when only trivial financial statement analysis is applied.

25.1.6 Comparison Using Interval Analysis

A profound look at the behavior of industries over time can be obtained by using Interval Analysis. Such an example was discussed in subsection 24.4.6 and is illustrated in Fig. 25.6.:

1. High stability of bounds of typical data: intervals of typical data are narrow for all the ratios considered. Moreover, their width does not change significantly with time, it only shifts to sensitively reflect overall changes in the ratio values. Therefore, these bounds can be used as reliable numerical characteristics for the classification of data ranges.
2. Changes over time within industries: the relative total liability $TLdA$ is the most stable indicator in both industries, while total returns change in the most sensitive manner.
3. A strong asymmetry is caused by data, which are outside of the typical intervals. This is especially true for the most volatile ratios ROA, TR and RWC.

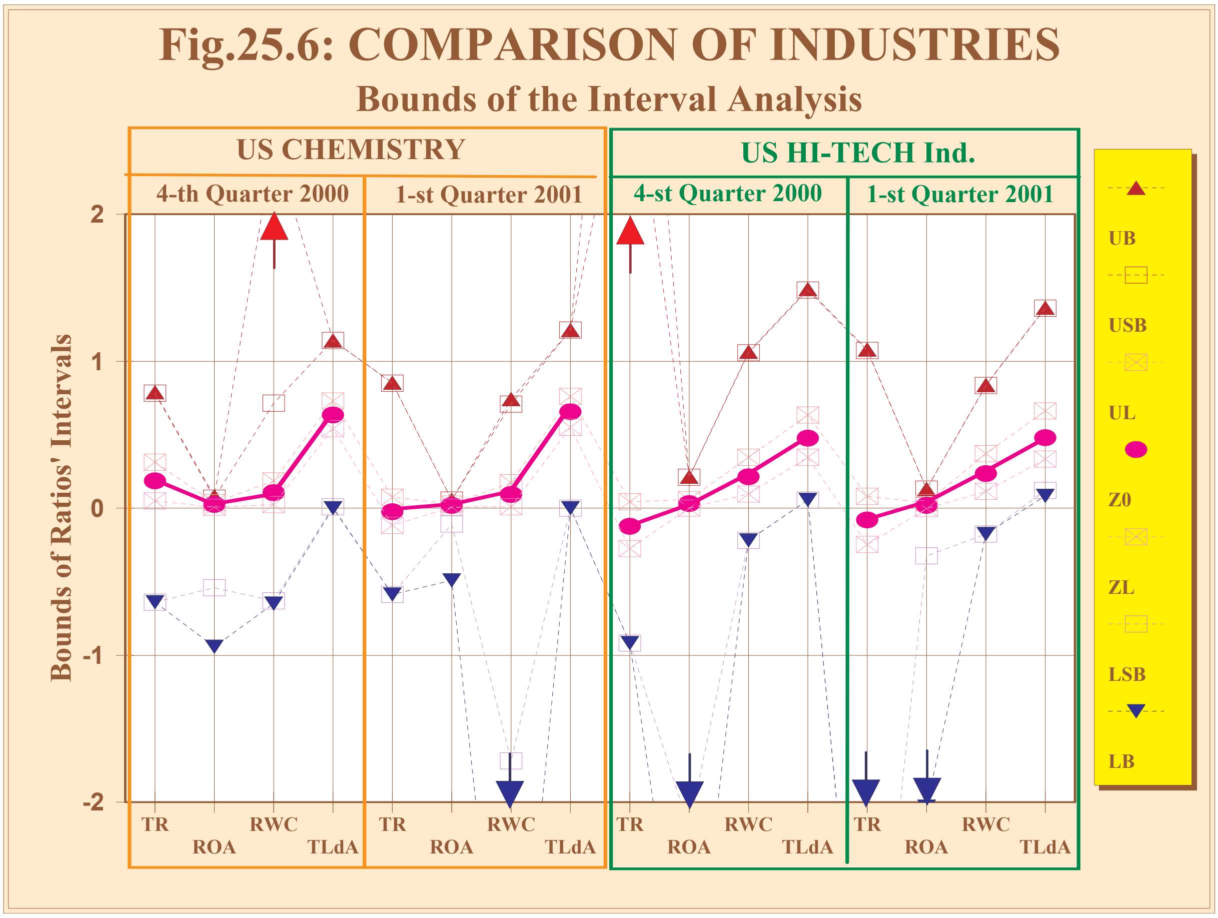


Fig.25.6: COMPARISON OF INDUSTRIES

Bounds of the Interval Analysis

4. The structure of intervals (especially those of typical data) for both industries are similar.

It can be concluded, that interval analysis can be useful to reveal fine differences between industries reliably and to judge the significance of the impacts of common ‘external’ factors, that influence the industries.

25.2 Stock Market Analysis

The history of mankind is inseparable from the historical development of markets. History has its milestones ordinarily arising from the great inventions of the human mind. The creation of money allowed market history to be subdivided into two periods: the barter age, and intermediation. Another leap in the development of markets resulted from the invention of shares, the ‘moneyless money’ that has gradually taken over the reins of the world economy. The dominant role of markets has become a permanent fixture in all facets of human development; it is no surprise, then, that the market mechanism is a permanent subject of analyzes, the special focus of which is on two important topics:

1. The ability of the market to generate economic stability: 'the role of the invisible hand.'
2. The efficiency of the real markets: the setting of prices and the allocation of resources.

Without getting too deeply embroiled in economic theory, a brief examination of these points of view can shed some light on current thinking in these areas.

25.2.1 The Invisible Hand Revisited

It has been popular in recent times to cite Adam Smith in support of whatever economic policy is being espoused, regardless of the orientation of the proposal. While many of the principles, which he developed in *Wealth of Nations [102]*, are as applicable today as they were in 1776, when the work was first published, the words that are bandied about are all too often plucked off of pages, which have little bearing on the subject being debated. Commentaries ranging across the economic continuum use this vast work to undergird their broadly diverging propositions. Given these varied referrals, it seems appropriate at this juncture to review several of Smiths's ideas and to put them into their initial context. The most often quoted phrase, referring to the *invisible hand* appears only once in this work of over 1000 pages. Its setting is a discussion of foreign trade:

> ... *As every individual, therefore, endeavors as much as he can both to employ his capital in support of domestic industry, and to direct that industry that its produce may be of greatest value; every individual necessarily labours to render the annual revenue of the society as great as he can. He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an* **invisible hand** *(bold face supplied) to promote an end which was no part of his intention. Nor is it always worse for the society that it was no part of it. By pursuing*

> *his own interest he frequently promotes that of the society more effectually than when he really intends to promote it*[5] ... [102].

The thrust here was for support of domestic markets and production over the importation of goods because of the inability to supervise and control invested capital far from home. The *hand* in guiding the investment and productivity of capital, would increase the domestic output and thereby increase the welfare of the society. With the existence of rapid communications, this is no longer the case today, when securities or goods may be traded instantaneously almost anywhere in the world. Further, he advocated the purchase of foreign goods, when their prices (and their costs of production) were less than would be the case were they produced at home.

> *When the produce of any particular branch of industry exceeds what the demand of the country requires, the surplus must be sent abroad and exchanged for something for which there is a demand at home. Without such exportation, a part of the productive labour of the country and the value of its annual produce diminish*[6].

Smith's economics, and the setting of prices for commodities are based on perfectly competitive markets, not on the oligopolistic structure which has developed in the industrialized world. Hence:

> *These ordinary or average rates [for the employment of labor and stock] may be called the natural rate of wages, profit and rent, at the time and place in which they commonly prevail.*
> *When the price of any commodity is neither more nor less than what is sufficient to pay the rent of the land, the wages of the labour, and the profits of the stock employed in raising, preparing and bringing it to market, according to their natural rates, the commodity is then sold for what may be called its natural price....*
> *The actual price at which any commodity is commonly sold is called its market price. It may be above or below its natural price.*
> *The market price of every particular commodity is regulated by the proportion between the quantity which is actually brought to market, and the demand of those who are willing to pay the natural price of the commodity*[7]...

[5] Book 4, Chapter 2, pp 485

[6] Book 2, Chapter 5, pp.463

[7] Book 1, Chapter 7, pp 62-63

Under these circumstances, efficient allocation of resources leads to lower prices, and growth. On the other hand, market power leads to the erosion of competition and eventually to monopolistic domination.

Variations in supply and demand will affect the costs of the factors of production and the profit earned so as to keep the price of the commodity above or below the ‘natural’ price, but only if there is freedom for the market to react to developments and adjust to change. With free entry, increases in competition lead to a reduction in prices and profit as competitors attempt to enter the market:

> *But though the market price of every particular commodity is in this manner continually gravitating, if one may say so, towards the natural price, yet sometimes particular accidents, sometimes natural causes, and sometimes particular regulations of police [policy], may in many commodities, keep up the market price, for a long time together, a good deal above the natural price*[8].

Therefore some of the parties to the economic contract may have an interest in dampening these fluctuations:

> *It is to prevent this reduction of price, and consequently of wages and profit, by restraining that free competition which would most certainly occasion it, that all corporations, and the greater part of corporation laws, have been established. . . . But this prerogative of the crown [to control and regulate incorporation] seems to have been reserved rather for extorting money from the subject, than for the defense of the common liberty against such aggressive monopolies*[9].

However, at some point in the development of an economic system, a civil framework of law and regulation becomes necessary. This would occur, when the accumulation of property exceeds the immediate needs of the consumer.

> *Wherever there is great property, there is great inequality. For every rich man there must be at least five hundred poor and the affluence of the few supposes the indigence of the many*[10].
>
> *. . . The rich, in particular, are necessarily interested to support that order of things, which can alone secure them in the possession of their own advantages. Men of inferior wealth combine to*

[8] Book I, Chapter 7, pp 67-68

[9] Book I, Chapter 10, pp 142-143

[10] Book V, Chapter 1, pp 766

> *defend those of superior wealth in the possession of their property, in order that men of superior wealth may combine to defend them in the possession of theirs.*
> *. . . Civil government, so far as it is instituted for the security of property, is in reality instituted for the defense of the rich against the poor, or of those who have some property against those who have none at all*[11].

However, acquisition of market or monopoly power acts against the individual's ability to pursue his independent agenda.

> *A monopoly granted either to an individual or to a trading company has the same effect as a secret in trade or manufactures. . . . The one [monopoly price] is upon every occasion the highest which can be squeezed out of the buyers, or which, it is supposed, they will consent to give; the other [natural price] is the lowest which the sellers can commonly afford to take, and at the same time continue their business*[12].

> *To give monopoly of the home-market to the produce of domestic industry in any particular art or manufacture, is in some measure to direct private people in what manner they ought to employ their capitals, and must, in almost all cases, be either useless or a hurtful regulation*[13].

Those, who lean on Smith to support greater regulatory control speak to his condemnation of the demeaning nature of the repetitive tasks imposed by the specialization of labor,

> *. . . His dexterity at his own particular trade, seems in this manner to be acquired at the expense of his intellectual, social, and martial virtues. But in every improved and civilized society this is the state into which the laboring poor, this is, the great body of the people must necessarily fall, unless government takes some pains to prevent it*[14].

or the restrictions on the mobility of labor and the free choice of employment in one trade or another.

> *Though men of reflection too have sometimes complained of the law of settlements as a public grievance; yet it has never been the*

[11] Book V, Chapter 1, pp 771

[12] Book I, Chapter 7, pp 69

[13] Book IV, Chapter II, pp 485

[14] Book V, Chapter 1, pp 840

> *object of any general public clamor, such as that against general warrants ... There is scarce a poor man in England of forty years of age, I will venture to say, who has not in some part of his life felt himself most cruelly oppressed by this ill contrived law of settlements*[15].
>
> *Let the same natural liberty of exercising what species of industry they please [for displaced workers], be restored to all his majesty's subjects, in the same manner as to soldiers and seamen [who could choose any employment once discharged from active service]; that is, break down the exclusive privileges of corporations, and repeal the statute of apprenticeship, both which are real encroachments upon natural liberty, and add to these the repeal of the law of settlements, so that a poor workman, when thrown out of employment either in one trade or in one place, may seek for it in another trade, or in another place, without the fear either of a prosecution or of a removal*[16]...

Smith was very critical of regulation and government intervention in the marketplace, which is where the proponents of open markets obtain their ammunition. Yet, as we have seen, he was very conscious of the distortions brought about by the application of market power, and thus the need for oversight and control of these activities.

> *It is thus that every system which endeavours, either, by extraordinary encouragements, to draw towards a particular species of industry a greater share of the capital of the society than what would naturally go to it; or by extraordinary restraints, to force from a particular species of industry some share of the capital which would ordinarily be employed in it; is in reality subversive of the great purpose which it means to promote. It retards instead of accelerating, the progress of the society towards real wealth and greatness; and diminishes, instead of increasing, the real value of the annual produce of its land and labour.*
> *... the sovereign has only three duties to attend to; three duties of great importance, indeed, but plain and intelligible to common understandings; first, the duty of protecting the society from the violence and the invasion of other independent societies; secondly the duty of protecting, as far as possible, every member of the society from the injustice of or oppression of every other member*

[15] Book I, Chapter 10, pp 162-163

[16] Book IV Chapter 3, pp501

> *of it, or the duty of establishing an exact administration of justice; and, thirdly, the duty of erecting and maintaining certain public works and certain public institutions, which can never be for the interest of any individual, or small number of individuals, to erect and maintain; because the profit could never repay the expense to any individual or small number of individuals, though it may frequently do much more than repay it to a great society*[17].

Therefore the greatest benefit is derived by letting the normal course of events take its own path:

> ... *The natural effort of every individual to better his own condition, when suffered to exert itself with freedom and security, is so powerful a principle, that it is alone, and without any assistance, not only capable of carrying on the society to wealth and prosperity, but of surmounting a hundred impertinent obstructions with which the folly of human laws too often incumbers its operations*[18].

Much of the present argument to let the market determine the true price of an asset glosses over the requirement for 'perfect markets' and the costless flow of information available to all participants. Indeed, these imperfections are at the root of the recent turmoil. Among others:

1. the vast amount of and rapidity of the flow of information coupled to the limited ability to analyze and digest its implications in a timely manner,
2. rationing information: special briefings for large investors, etc.,
3. biased analysts' reports and recommendations,
4. symbiotic relationship between banking and investment,
5. mechanical trading rules for large blocks of securities (program trading) automatically triggered at predetermined price levels,
6. the 'kinship' syndrome, news affecting one firm tarring all the others in the same industry segment,
7. the misapplication of the 'true and fair' principle in financial reporting.

Adam Smith's *invisible hand* could be alive and well today, but only if the environment were to permit the individual to freely pursue his own interests. Unfortunately, the glut of data, the paucity of good analysis, and the monolithic bureaucratic structure of big business and government impede the exercise of the simple self-serving actions, which would allow

[17] Book IV, Chapter 9, pp 745

[18] Book IV, Chapter 5, pp 581

it to function properly.

The history of the past several centuries has shown, that Adam Smith's ideas on the inherent power of the free market are capable of developing national or regional economies in an impressive manner. The invisible hand of the market really exists and it works in principle, but only if it has proper information as raw material. This thought suggests a comparison to Darwin's thesis on the development of biological species: What those processes were that went missing were only partially exploited by nature, and could have led to ... better adapted, more useful, more highly specialized ... living forms, or what detours could have been taken before final outcomes were achieved, will never be known. On the other hand, the development of the world's economies has also had its failures and (verbatim!) dead ends. But unlike the case of natural history, it allows questions about missing or at least underestimated elements to be asked as well as answered: **information**. The invisible hand of the market is blind without information and the decisions based on incomplete, fuzzy or even false information are chaotic, counterproductive and even dangerous. Theories dealing with market efficiency emphasize the role of information, however this does not automatically mean, that the market's recent performance has been due to the influence of the best quality of information available.

Returning to Fig. 24.8, it is possible to state, that the market indicators really behaved in a chaotic manner, while the tools of advanced analysis revealed the true development of the financial position of a real economic object. This statement was strongly supported by Fig. 25.4, which was able to explain the historical performance of the object.

25.2.2 Is the Stock Market Really Efficient?

In recent years, market theory has been dominated by versions of the *efficient-market hypotheses* developed in part from a paper written by Kendall [43]. Market efficiency relates to the extent, that information is costlessly and timely available to market participants, and already imbedded in stock prices. The idea, that prices moved in a random walk had been proposed years earlier by Bachelier [3] as well as popularized more recently by Malkiel [73]. If all information about a company is reflected in the firm's stock price, then only *new* information can cause prices to change. Since good news is equally likely as bad news, then prices have an equal probability of rising or falling. Typically, two independent and diametrically opposed approaches are taken to exploit this information set.

The fundamentalists try to uncover new information about the company, its markets, products and its expected profitability. In contrast, the technicians study past prices and patterns and speculate, that these movements will be repeated over time. The fundamental analysts feel confident, that all relevant information is available to investors, while the technical analysts (also dubbed "elves") guarantee, that the current price reflects all past sequences in pricing. For a complete discussion on this subject see Roberts [92].

Another condition difficult to meet in practice, even if accurate and timely information is possessed is the ability to cover trading costs. Small traders, in particular, are unable to easily negotiate commissions, and in general, successful outcomes are more likely to be the consequence of frequent small trades rather than a few 'killer strikes.'

If—as commonly expected—all publicly available information is taken into account in a timely manner, then the *semi–strong* version of efficiency holds; therefore:

- A diligent search to develop new information can reward the patient investor. But information is costly both in the time or in the knowledge necessary to develop it, or from the direct cost of purchasing it from professional analysts.
- Beating the market consistently is difficult, and most mutual fund managers have found, that they can barely cover the cost of hiring and using superior analysts, who can develop timely investment recommendations.
- This has lead to the proliferation of 'index funds'—if you cannot beat it, then join it!

But the major assumption of the theory, the universal and relatively costless availability of all information to all participants, is not very realistic. There will always be new developments in computers and other communication hardware not easily exploitable by all investors; therefore the technical level of information to consumers will not be the same and they will not be provided with the same data at the same time. There is also another important problem: **to have the same data does not mean to have the same information.** Data bear information, but information must be extracted from the data by software. Software is influenced by rapid development as well as by theories, on which data treatment methods are based. It is obvious, that all analysts do not use (nor do not probably understand) the most recent state of the art in analytical methods. These comments notwithstanding, nothing above infers, that the ethical and legal relationships of market participants are in question; a conclusion, which given recent events invokes further discussion.

The belief in market efficiency (or—as journalists say—in the almighty invisible hand of the market) is deeply rooted. Every finance textbook instills the idea, that stock price is deemed *efficient* if it equals the present value of the expected payoff divided by 1 plus required rate of return ([83]). Natural questions arise in connection with this notion: Even if accurate estimates of future pay offs were obtainable, whose expectations and whose required return are assumed? Expectations and requirements are highly subjective and as such, they are hard put to provide an objective price estimate. The label linked to the price suggests, that the market's expectations and the market's requirements should be used, but this would work properly only if the market were really efficient, in which case, then, efficiency would play an objective role.

In a truly efficient market, absent new information, assets would trade for their 'natural price' ± the cost of carry. Given the wide price differentials, which are observed, it is more probable, that the (real, market) price is not efficient and perhaps a more plausible explanation for the excessive volatility is, that many investment decisions are based on little if any or, false, or badly interpreted information.

The importance of information in decision making as highlighted in [83] infers a similar conclusion:

> *...a forecast requires analysis of the firm's earning ability. Future prices and dividends cannot be forecast in a vacuum, without understanding the ability of the firm to deliver value. Future prices, in particular, will depend on how well the firm performs in its operations. So your ability to gain arbitrage opportunities*[19] *will come from developing better forecasts than the market of what the firm will deliver in its operations in the future. As better forecasts come from better information and better analysis of information, we could say, that arbitrage opportunities arise from having good information and being able to see the implications of that information.*

It is thus recommended to an analyst to obtain better information so as to find a market that is not in balance (with his new information) and to take that opportunity to make a risk-less profit greater than what would be

[19] Arbitrage opportunity refers to the ability to buy an asset in one market and sell it at a profit in another, where the information set or requirements are different.

possible in equilibrium, where the required return is the only ‘legal’ profit opportunity.

All of this can be interpreted as a call for effective advanced analysis. It was shown in previous chapters, how important results can be obtained by advanced financial statement analysis. Robust multidimensional models were used to characterize dependencies between the principal characteristics of the building blocks of a firm’s financial condition. Models, that include such dependencies may be used to extract additional information, which is not readily available to all participants in the market. The above demonstration using the two US industries, Chemistry and High-Tech, are an example of the utility, that can be gained from the exploitation of these ideas.

25.2.3 The Reappraisal of Shares

Investment decisions depend on expectations of the future cash flow generated by particular shares and this depends on forecasts of earnings to implement any particular strategy. There are two levels of prediction: the more difficult is to predict future total returns, while the simpler task is to predict at least the sign of the trend of the returns’ movement. The latter problem can be defined as follows:

> Given the past and more recent basic accounting data of a group of (in a way) comparable firms and the market’s estimate of their financial positions as reflected by current values of the total returns; provide, as reliably as possible, a response as to whether the next total return for each individual firm will be more or less than it is at present.

A change in total returns results from either a stock price movement (external factor) or a dividend policy decision (internal factor) by the firm’s board of directors. The dividend’s input reflects the firm’s financial position and policies as well as its future performance expectations, while the change in stock price results from the market’s appraisal of these policies, taking into account not only the firm’s financial data, but also additional information from outside sources. These may include an appraisal of uncertainty, the interpretation of which is most likely not very accurate. By accepting such a point of view, supported by his experience with the market’s behavior, a potential investor can propose the following audacious hypotheses:

1. The market's mechanism for the appraisal of shares is imperfect.
2. To find the 'true' value of a share and to establish its fair price, the market needs a not negligible portion of time.
3. Imperfections in the results of the market's appraisal are at least partially caused by inadequate treatment of available quantitative information.

If these hypotheses are indeed valid, then a successful **reappraisal** of shares is possible based on an improved treatment of the available data. To accomplish such a task, and to estimate the reliability of its results, the appraisal procedure must be described and the definition of **success** must be set out. The following procedure will be implemented:

- Select M variables $R_{m,q,k}$ (financial ratios based on entries in financial statements of K firms for a series of Q consecutive quarterly periods), which are suitable as explanatory variables in the cross-section models of total returns

$$TR_{q,k} = K_{0,q} + \sum_{m=1}^{M} Km, q * R_{m,q,k} \qquad (k = 1, \dots, K, q = 1, \dots, Q). \tag{25.1}$$

- Estimate the model coefficients $K_{m,q}$ $(m = 0, \dots, M)$ for all Q quarters using a robust method.
- Estimate the quarterly total returns $\widetilde{TR}_{q,k}$ of each (k-th) firm for each q-th quarter using these models. Accept these estimates as the reappraised values of total returns imbedded in the actual market appraisal $TR_{q,k}$.
- Take such a result as a **successful** reappraisal, when one of the following relations hold:

$$(\widetilde{TR}_{q,k} > TR_{q,k}) \wedge (TR_{q+1,k} > TR_{q,k}) \tag{25.2}$$

or

$$(\widetilde{TR}_{q,k} < TR_{q,k}) \wedge (TR_{q+1,k} < TR_{q,k}). \tag{25.3}$$

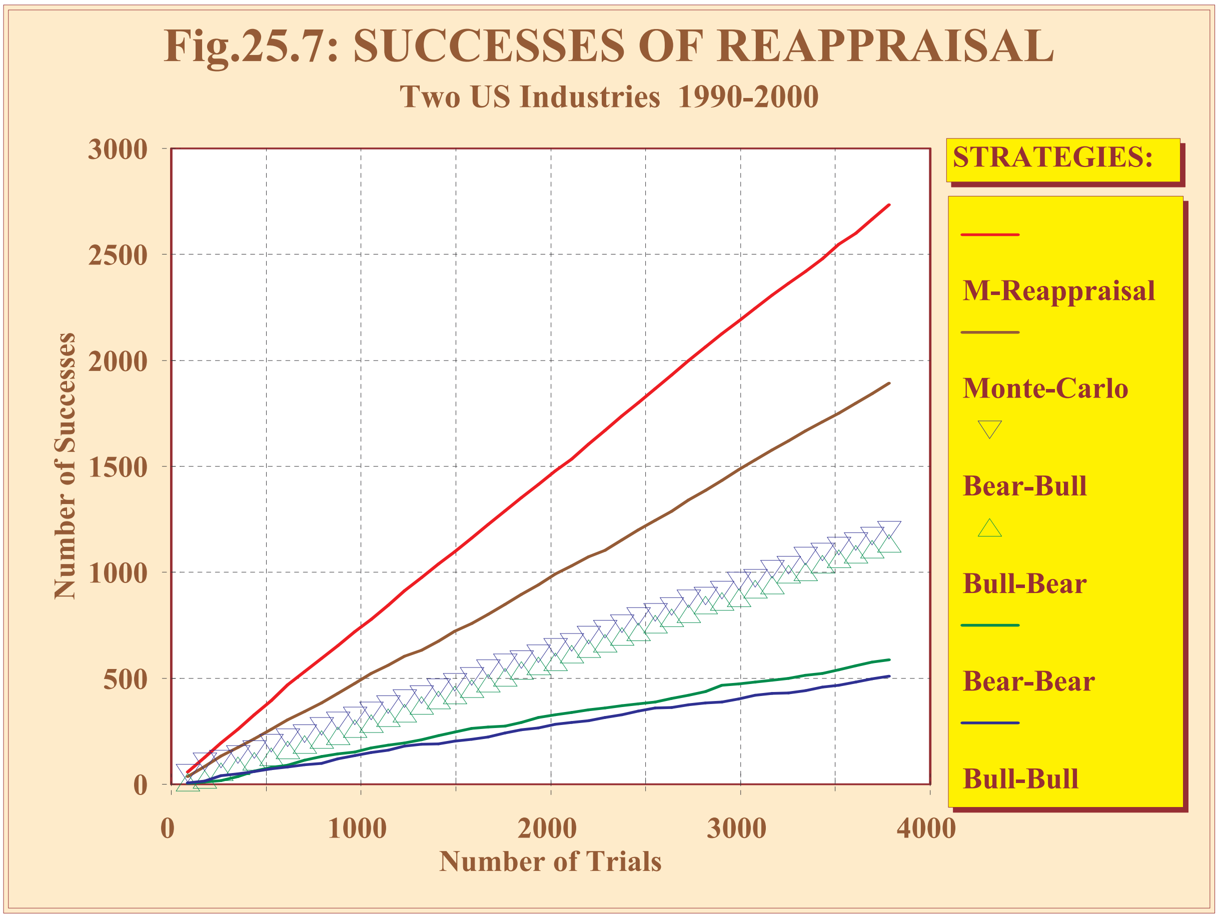
- Using all the data, estimate the probability of success as the number of successes divided by the number of trials.

The idea behind this experiment can be simply interpreted: the relation $\widetilde{TR}_{q,k} > TR_{q,k}$ says "the k-th firm's shares were underestimated by the market", while the validity of the relation $TR_{q+1,k} > TR_{q,k}$ means "the market recognized and corrected its mistake in the next quarter in the direction shown by the reappraisal." The interpretation of 25.3 is analogous, but the sign of the error is in the opposite direction.

Before presenting numerical results, the basic trivial/naive trend-predicting strategies for the next quarter of a year are set out below:

Bull-Bull: Rising TR will continue to rise.
Bear-Bear: Falling TR will continue to fall.
Bull-Bear: Rising TR will fall.
Bear-Bull: Falling TR will rise.
Monte-Carlo: Decision by tossing a coin.

The data used to test the above reappraisal strategy were again taken from US Chemical Industry (30 firms) and US High-Tech Industry (58 firms) for 43 quarters from the 2-nd quarter of 1990 through the 4-th quarter of 2000, which resulted altogether in 88×43=3784 trials realized. Curves showing a gradual increase in the number of successes with a rising number of trials are depicted in Fig. 25.7.



(The brown line of the Monte-Carlo approach was obtained by simulating the outcome of tossing a coin by a pseudorandom generator. This line is important, because it describes a purely random change in the bull/bear market posture ('the Brownian motion'). The mean probability of successes for 3784 trials was 0.723 for the gnostic reappraisal, against the 0.5 theoretical and 0.501 actually realized probability obtained by the Monte-Carlo approach. The trivial strategies yielded 0.301 for Bear-Bull, 0.318 for Bull-Bear, 0.155 for Bear-Bear and 0.135 for Bull-Bull guesses. In view of the large number of trials, the significance of the results of the reappraisal refutes any suspicion, that the results were random. The hypotheses of imperfection in the market are therefore supporting the meaning, that **the market can be beaten by using a mathematical technique,** at least on the level of the less difficult task of predicting quarterly trends. The alternative interpretations of these results are: **the mathematical reappraisal is better than the market's appraisal of stocks** and **the stock market is not efficient.**

25.3 Example of Real-time Foreign Exchange Trading

Internet trading as well as Internet banking has become an everyday activity. While these networks were established many years ago, initially, only the top banks, which were principally engaged in foreign exchange trading had access to such electronic facilities; today there are many participants. In 1994, an international scientific congress was organized to review the processes employed. Real time data for such studies were made available to the participants in this conference by the sponsoring institution[20], which monitors the network and organizes its database. These data were made available to the authors, and are used here to illustrate the type of results, which could be obtained by the application of gnostic filtering methodology to the trading mechanism. The data set predated the adoption of the Euro, and the Deutsche Mark still existed as an independent currency.

Over the course of a day, many billions of dollars of value are transferred over very short, almost instantaneous, time periods. Fig. 25.8. models trading over about 15 minutes, showing both the bid and the ask prices for the DM/US$ exchange rate for trades which were consumated.

The events occur randomly and to treat and depict them individually would be difficult; their frequency is too large. The processes are therefore condensed by averaging their number over intervals of ten seconds.

[20] Olsen & Associates, Research Institute for Applied Economics. Seefeldstrasse 233, CH-8008 Zurich, Switzerland. E-mail: hfdf@olsen.ch.

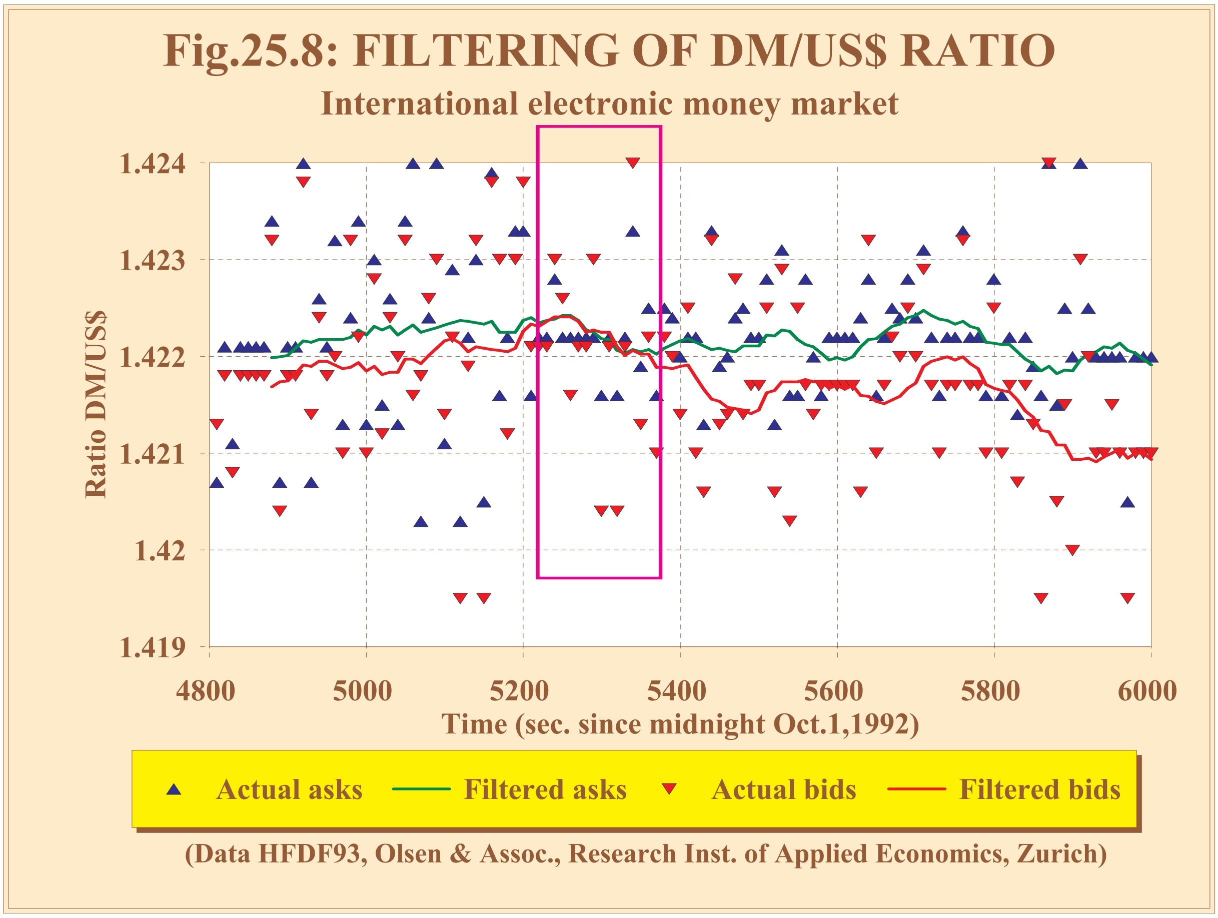


Fig.25.8: FILTERING OF DM/US$ RATIO

International electronic money market

(Data HFDF93, Olsen & Assoc., Research Inst. of Applied Economics, Zurich)

The fluctuations in the DM/US$ rate are strong even after the averaging. Therefore the application of a gnostic robust filter would facilitate the decision-making, ie it would show, that the mean short-term spread (the difference between asks and bids) was zero between 5200 to 5350 seconds (this time interval is emphasized by the magenta frame). This does not necessarily mean, that individual favorable transactions could not be made but the information on this temporary equilibrium must be taken in account by the trader. As shown by the graph, it was better to wait a few seconds for a substantial favorable recovery to take place in the spread than to force a trade during this hiatus.

A filter for smoothing data in this application must be both robust to outliers and sensitive to changes in the short-term mean of the processes. The importance of robustness is emphasized in Fig. 25.9, where the dynamics of the gnostic filter are compared with those of a popular robust statistical filter 53H of the L-type based on moving medians [74]

$$\tilde{y}_k = \frac{S(5,k-2)}{4} + \frac{S(5,k-1)}{2} + \frac{S(5,k)}{4}, \tag{25.4}$$

where y_k are elements of the time series and

$$S(v,m) = \mathrm{median}(y_{m-(v-1)/2}, \ldots, y_m, \ldots, y_{m+(v-1)/2}). \tag{25.5}$$

Both filters are applied in Fig. 25.9 to determine the short-term means of the spread minus the long-term mean.

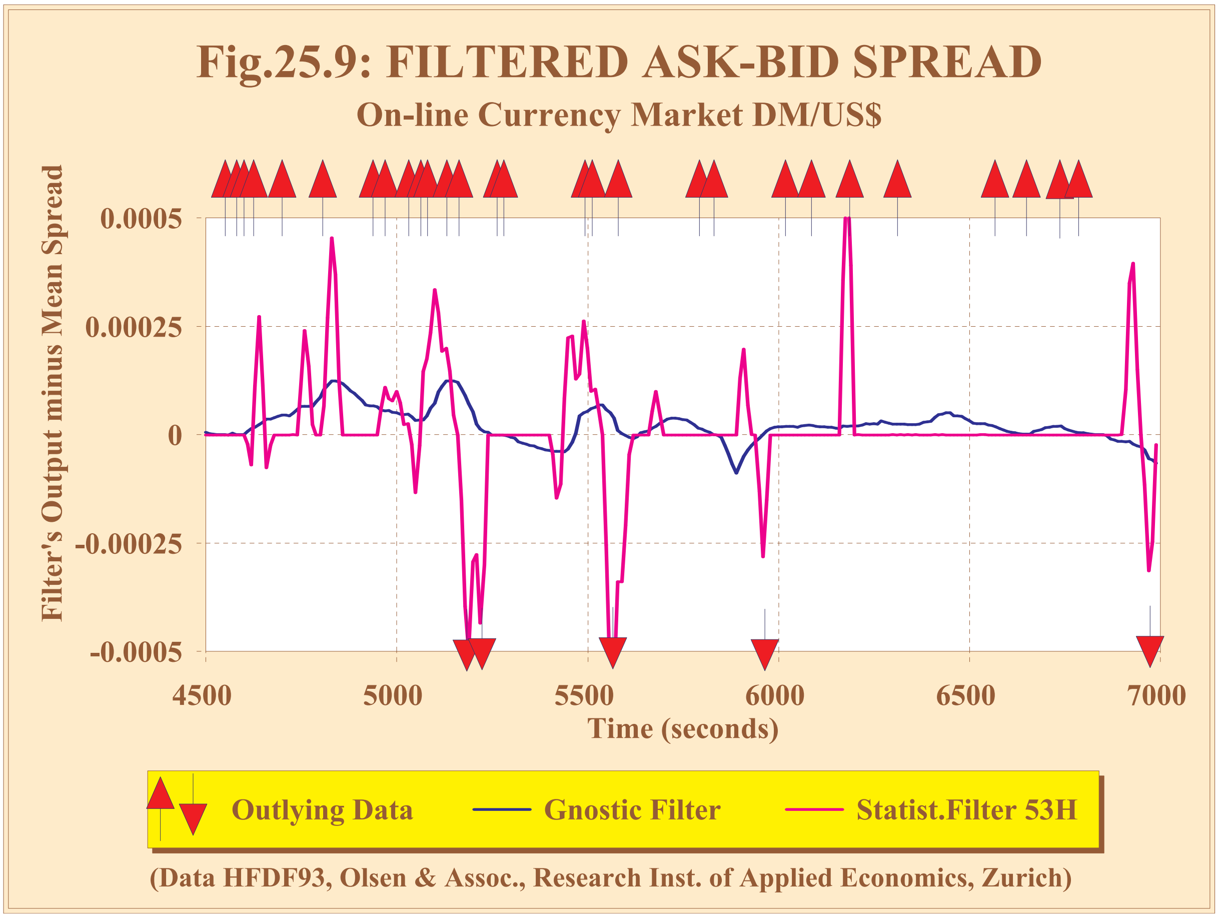


Fig.25.9: FILTERED ASK-BID SPREAD

On-line Currency Market DM/US$

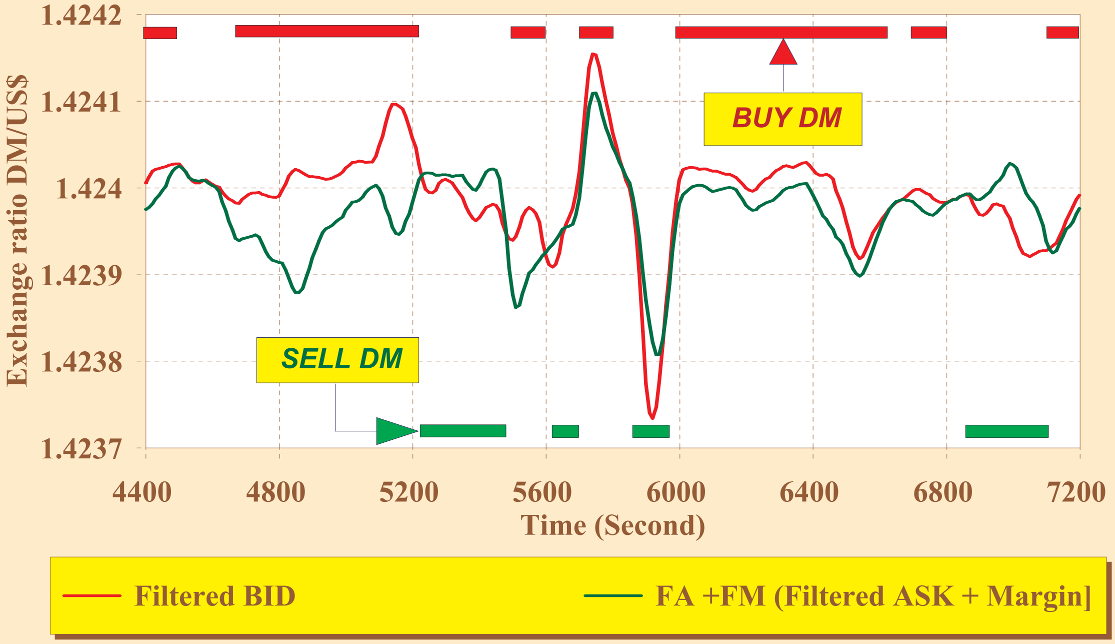
(Data HFDF93, Olsen & Assoc., Research Inst. of Applied Economics, Zurich)

The deviations of the filter's outputs from the long-term value signal a temporary imbalance between the asks and the bids, which can be exploited for buy/sell decision making. The arrows in Fig. 25.9 signify outlying data. The filter's robustness should protect the decision process from hasty reactions to outliers, while sensitively delimiting the periods of favorable actions. The graph shows a better performance from the gnostic filter. Its output is smooth enough to be used even for the predictions.

Fig. 25.10 replicates a trading scenario: the red line shows the predicted time series of bids, while the green line corresponds to the sum of predicted asks and long-term spread.

Fig.25.10: BUY/SELL DECISIONS (DM/US$)

Gnostic predictor for DM/US$ exchange

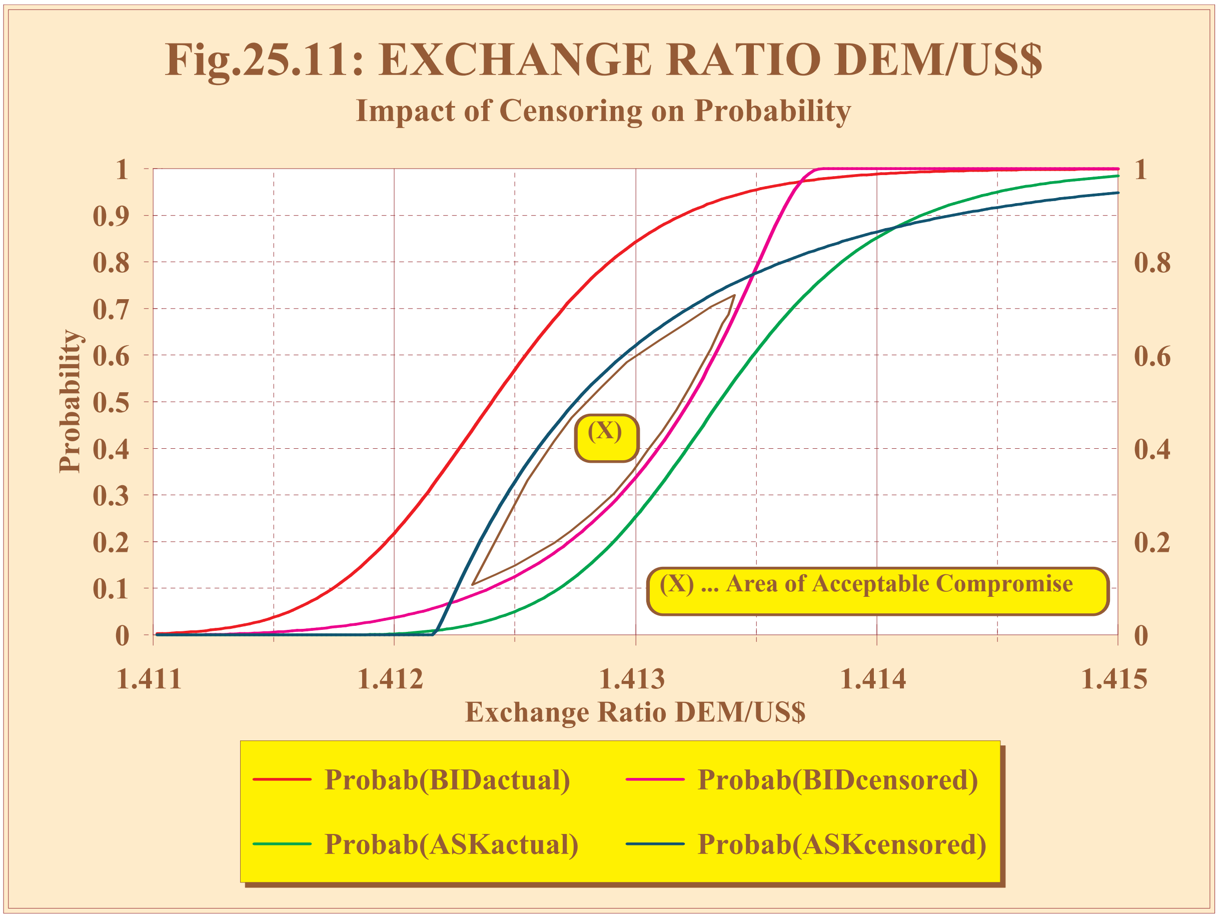


FA . . . Filtered ASK series FB . . . Filtered BID series FM . . . Filtered ASK minus BID series

(Data HFDF93, Olsen & Assoc., Research Inst. of Applied Economics, Zurich)

The time intervals of bids that exceed the points on the green line are favorable for selling the German Mark, while the opposite situation calls for a buy decision.

The idea of censored data was discussed in Chapter 22 (section 22.2.11). Bargaining can also be viewed as a manipulation with censored data: the selling party aims to get the maximum for its goods, pushing the price as high over the (hidden, mostly unknown) true value as possible. The interest of the buying party is the opposite. The problem can be formulated as a probabilistic one: there is no reliable information on the true value of the good in question. The bid price is the minimum estimated (and required) by the buyer, but the true value could be “anywhere” (with the same probability) between this minimum and the upper bound of the prices: the assumption of a uniformly distributed true value over this interval corresponds to the complete lack of information. But, this is the model of right-censored data. Analogously, the asks can be viewed as left-censored data. Fig. 25.11 compares the distribution functions of asks and bids with those obtained under the assumption of censoring.



A possible occurrence of true values exceeding the required bid shifts the distribution of bids to the right, while the possible existence of true values lower than asks results in a left shift of the ask's distribution. The area (X) between both censored distributions can be interpreted as a region of compromises acceptable by both sides. As the graph shows, the compromise area is reasonable for exchange ratios in the narrow interval 1.4122 through 1.4134. The probability of occurrence of conditions favorable for such compromise can be roughly estimated as 0.77 minus 0.06, ie by 0.71.

These results are based on extreme assumptions (a complete lack information as to the true value of the "goods"). The availability of additional information would likely permit a less uncertain decision to be made.

25.4 A Case study: The French Car Market in 2000

25.4.1 The Role of the Market Analysis

Success in marketing is a reward for having properly solved a long chain of related and difficult tasks: the identification of a need/demand for a new product ⇒ its conception and initial design ⇒ finding a niche in the market that will accommodate a certain quality product, which can be sold for a competitive price ⇒ the establishment of both its technical and economic parameters ⇒ the development, design and production of the product so as to satisfy the demand of a specific class of consumers ⇒ setting both technical and economic parameters of the product for designers ⇒ the final development and design of the product.

The automobile is an everyday tool of modern life, a mass produced commodity, which has a significant impact on the world economy. As such, then, the car market can serve as a useful application field for advanced analysis. Since 'everybody understands the problem,' the results of such an analysis are easily understood using common sense. Those in the market for an automobile generally are faced with a set of questions, which are resolved more or less objectively and include, among others:

- the real needs to be satisfied and the main purpose to be served,
- the principal user/users,
- cargo, if any to be carried,
- the character and habits of the driver and his driving style,
- the desired safety level,
- reliability,
- economic requirements related to the
  - purchase price,
  - fuel consumption,
  - maintenance, service, insurance and projected repair costs,
- availability of a network of dealerships, service and repair shops,
- availability of spare parts,
- comfort level required,
- personal preferences as to taste, style and fashion,
- prestige aspects.

This—surely incomplete—list of factors allows several conclusions to be drawn:

1. the choice of a car is highly individualized and manifold,
2. individual preferences must be reflected by prices,

3. the diversity of the demand results in diversity in the supply of cars.

The objective of such an analysis of automobile market data should be to provide to manufacturers an objective view of both the technical (engineering) and economic parameters of cars on the market so as to identify the market segments, in which they desire to compete. An alternative goal could be to show to a car buyer 'what fulfillment of his individual wishes cost' and 'what he has to pay for.'

25.4.2 French Car Market Data

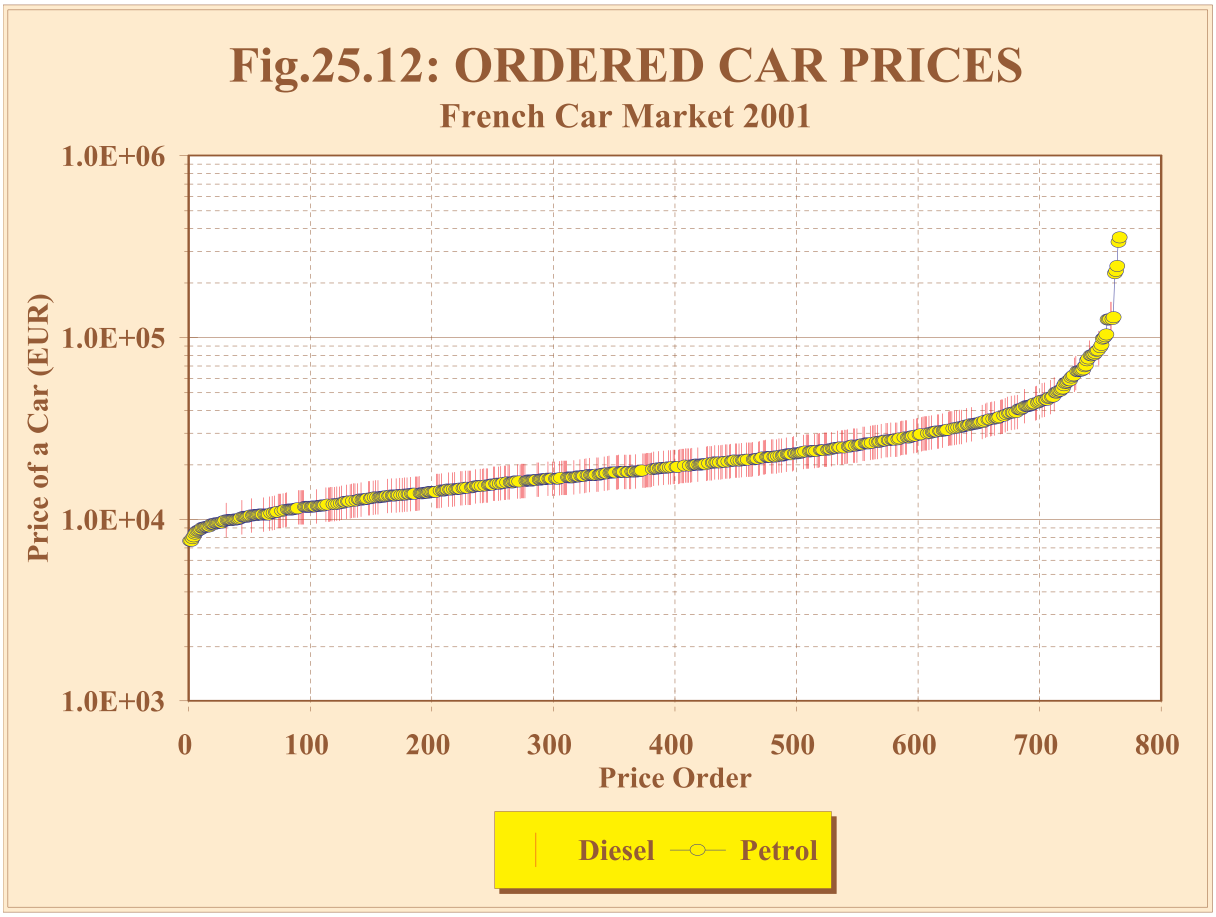
The preparation and maintenance of a complete and up-to-date data base covering a substantial sector of a car market is undoubtedly a task for a professional agency. However, this study only aims to demonstrate that by using advanced analysis useful conclusions can be drawn even from basic data available to the general public. Specifically, these data were taken from the French journal, L'Automobile Magazine for May 2001, which summarized the following basic parameters of cars offered on the French market:

1. Manufacturer,
2. model of the car,
3. type of engine (Diesel, gasoline),
4. French fiscal category for the model,
5. size (volume) of the engine (liters),
6. power (HP),
7. maximum speed (km/h),
8. fuel consumption (liter/100 km),
9. price (Euro).

The Journal lists 262 different diesel cars and 504 gasoline models. An overall view of the market is in Fig. 25.12, where all 766 car types are presented ordered by their price distribution.

The prices of gasoline models cover a broad range from the cheapest (price of Euro 7607) through the most expensive ones, (Euro 357073). The price range of diesel types was narrower, from Euro 9907 through Euro 126838. The differences in price distributions of the two basic classes can be seen in Fig. 25.13.

The ranges of the technical parameters also were very broad. A true 'King of the Road' among Diesels was the car with top characteristics: drive volume 4.9 liter, power 400 HP, maximum speed 250 km/h and fuel



consumption 15 l/100 km. The smallest Diesel engine was that declaring the smallest fuel consumption (4.5 l/100 km). The lowest power Diesel drive was 58 HP and the lower bound for maximum speed 129 km/h.

These same parameters for cars with spark ignition engines were spread over even broader intervals: the largest engine volume was 8 liters, which had a voracious appetite: 30 l/100 km), the highest power was 426 HP, greatest maximum speed 307 km/h. On the small end: volume, 0.8 l, power 50 HP, the lowest maximum speed 139 km/h and the most economical consumption 6.6 l/100 km.

The fiscal category of the car type relates to taxes; it reflected the official French evaluation of the compromise between practical needs and the luxury of the particular car type (the higher fiscal category, the higher both the imputed luxury and therefore the tax). The fiscal categories of Diesels were spread from 3 through 32, while for gasoline cars they ranged from 4 (for many of the smallest cars from nearly all manufacturers)

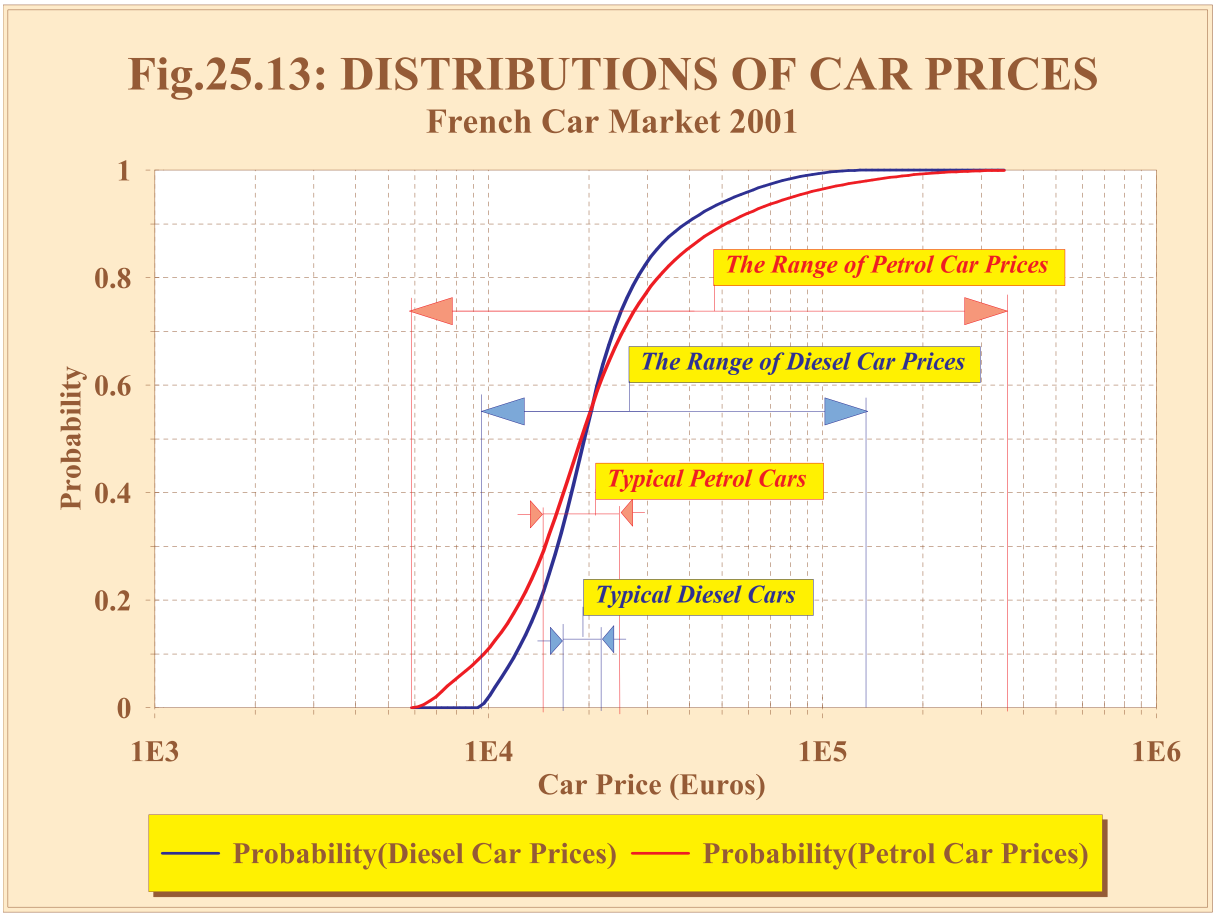


Fig.25.13: DISTRIBUTIONS OF CAR PRICES

through 36.

It is obvious, that these sets of parameters could neither completely describe the car nor reliably compare each of them with others. Important information was missing, which strongly influenced both the price and safety of the car (number of air bags, electronic safety devices such as ABS—automatic brake system, MBA—power brakes, ASR and MSR—transmission controllers, EDS—traction control, ESB—electronic stabilization system, CAN—electronic data bus connecting all the electronic devices, various computer chips, brake servos, power steering, xenon lights, automatic fuel cut off in case of an accident, automatic activation of the windshield wiper, a rear view TV set, TV for passengers, electronic map with a satellite orientation, electronic car protection and/or immobilization, satellite tracing, radio, magnetic tape or CD player, leather upholstery, adjustable and heated seats, remotely controlled and heated rear view mirrors, electrically driven windows, wood trim and so on). Experience shows, that some customers may base their buy decision not only on "serious" aspects, but also on such trifles like the number of cup holders.

Another drawback was lack of data on the total numbers of cars of individual types, which were sold. Such data could help in making a judgement as to the popularity of the makes and types. The incompleteness of the data subjected to analysis represented the main limitation of the analysis. Nevertheless, there were factors, which could partially compensate for this lack of detail. The principal of these is the competition between various models and makes, which forces them to include a certain "commonly accepted" standard of safety and luxury to various classes and prices of cars. The popularity of cars and their producers is also reflected by prices, which can be associated with both the producer's goodwill as well as with a measure of prestige for the car owner; it is important, that these factors can be quantitatively evaluated after the true car value has been provided by the multidimensional model. These additional facts can enhance the trustworthiness of the mathematical analysis of the limited data set available.

25.4.3 Steps of the Analysis

The analytic tools used in Chapters 23 and 24 will be employed here:

- marginal analysis of individual variables by means of distribution functions EGDF (section 15.2.5) and ELDF (15.2.4) and interval analysis (16.4),
- robust multidimensional analysis by means of non-traditional methods (17.3) including
  - explicit MD-regression in probabilities,
  - implicit MD-regression in probabilities.

It is hypothesized, that the broad scale of car prices observed in Fig. 25.12 and 25.13 correspond to a wide range of car quality that is reflected—at least partially—by the technical parameters. The robust multidimensional modeling technique will be used to evaluate the relationship between the parameters and prices. The explicit probabilistic price model applicable to the multidimensional ordering and pricing of cars takes on the following form:

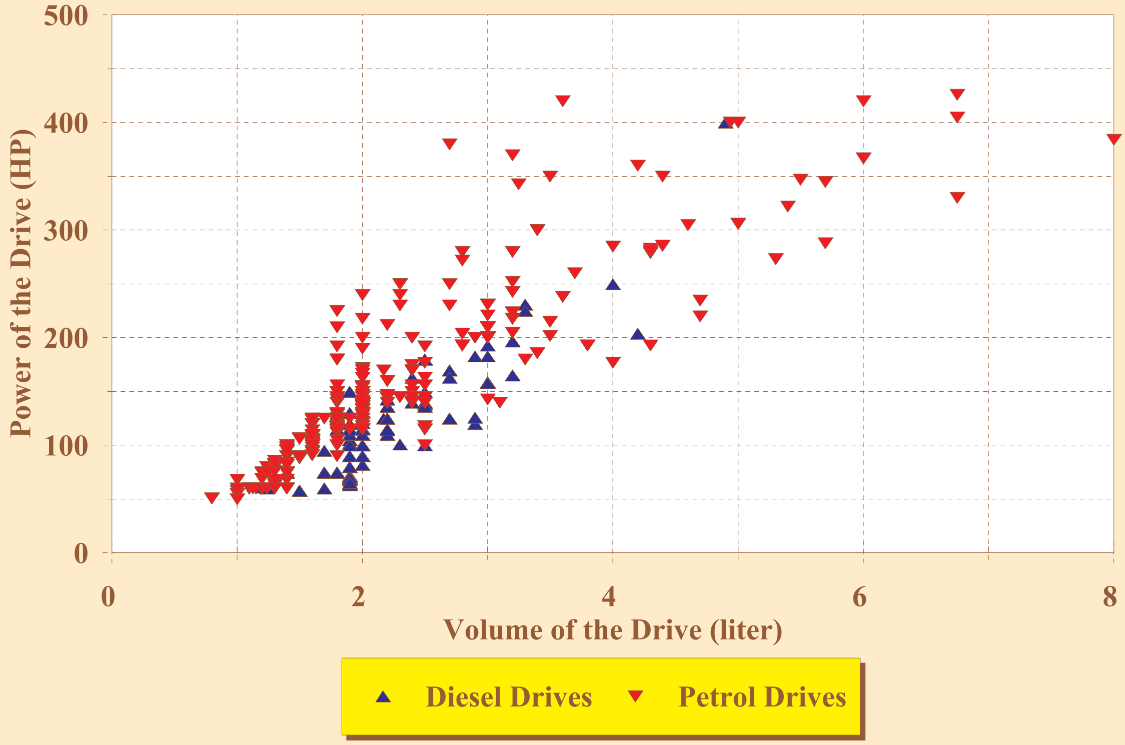
$$\begin{aligned} Pr\{Pri_k\} = C_0 + C_1 * Pr\{Vol_k\} + C_2 * Pr\{HP_k\} \\ +C_3 * Pr\{MxSp_k\} + C_4 * Pr\{Cons_k\}, \end{aligned} \tag{25.6}$$

(where $k = 1, \ldots, K$) with the following notation:
K …the total number of automobile types,
$Pr\{x\}$ …an estimate of the probability of the variable x,
Pri …purchase price (Euro),
$C_0, \ldots C_4$ …model parameters,
Vol …engine size (volume in liters),
HP …engine power (HP),
$Cons$ …consumption of fuel (liters/(100 km)).

An objection to these models can be expected based on the belief, that there is a proportionality factor between volume and power in automobile engines. If so, then this interdependence spoils the models. However, this idea is not justified. If there were such a relationship, the relation $HP(Vol)$ would be represented by a straight line—which, as seen in Fig. 25.14, is not the case. Therefore these two variables are not linearly dependent and using both of them in the regressions is reasonable.

Fig.25.14: RELATION DRIVE VOLUME-POWER

Cars on French Market 2001



In addition, the relations between the technical parameters and prices can be viewed differently depending on whose point of view is chosen: for the producer to raise the parameters, the **costs** of development and production will increase. These same changes are viewed by a potential owner as additional *utility* received for his money: the better the parameter values for a lower price, the more utility provided. To obtain such a measure of utility, parameters Vol, HP and $MxSp$ will be divided by the purchase price of the car. The impact of fuel consumption on the utility is negative: the higher consumption, the lower utility. This is why the notion of *kilometrage*[21] (measured in kilometers the car runs while burning 100 liters of fuel) will be used instead of a raw consumption figure. The following implicit multidimensional regression in probabilities is used to describe a car's utility:

$$\begin{aligned} c_1 * Pr\{Vol_k/Pri_k\} + c_2 * Pr\{HP_k/Pri_k\} \\ +c_3 * Pr\{MxSp_k/Pri_k\} + c_4 * Pr\{Kmge_k/Pri_k\} = 1, \end{aligned} \tag{25.7}$$

where Pri again stands for the purchase price and $Kmge$ is the kilometrage. The implicit form of the equation is chosen to escape of the necessity of designating one of the variables as 'dependent'.

The analysis takes on the following steps:

1. The marginal analysis of all variables:
   (a) Estimate robust distribution functions of the EGDF type for all five variables,
   (b) verify the homogeneity of the analyzed data samples and (in the case of inhomogeneity) find the EGDF of the main homogeneous cluster of each sample along with its bounds of data support LB and UB and scale parameter S,
   (c) use the parameters of the EGDFs with distributions ELDF to apply gnostic interval analysis (finding the interval bounds of typical data ZL and ZU along with the location parameter $Z0$),
   (d) apply the parameters to decompose cars into three classes (below typical, typical, above typical),
   (e) summarize the mean parameters of classes separately for Diesel and gasoline cars.
2. The multidimensional analysis:
   (a) using the distribution functions EGDF, calculate all the values of

[21] This can be viewed as the European analogy to the American notion of *miles per gallon.*

probability to be substituted into 25.7 and 25.8 (see 25.4.6 for the latter),

(b) solve both 25.7 and 25.8 separately for Diesel and gasoline cars,

(c) Apply the MD-models to

i. estimate the ‘true and fair’ values of cars,

ii. order car types with respect to their relation to the pricing model,

iii. estimate the ‘premiums’ connected with purchasing each car type,

iv. evaluate the cost of prestige and the goodwill of car producers,

v. order car types with respect to their utility,

vi. order makes of cars from the point of view of utility.

25.4.4 The Classes of Typical Cars

The results of the robust interval analysis are presented in Tab. 25.3.

Interval	Drive type	
Bound	Diesel (Euro)	Gasoline (Euro)
LB	9325	5868
ZL	16522	13698
$Z0$	20041	17745
ZU	24983	23744
UB	127481	358979

Tab. 25.3: Bounds of Price Intervals

The notation is the same as in section 16.4:
LB . . . the lower bound of the data support,
ZL . . . the lower bound of the typical data,
$Z0$ . . . the location parameter (the mode of the ELDF’s density),
UL . . . the upper bound of the typical data,
UB . . . the upper bound of the data support.

The interval bounds of typical prices of Diesels do not differ significantly from those of gasoline cars, but the spread of the prices of these cars is much broader than that of the Diesels. This was shown in Fig. 25.13.

The bounds indicated in Tab. 25.3 were obtained by using the distribution functions although the data were non-random. This approach is frequently applied in gnostic analysis, the methods of which are founded on a non-statistical conception of data uncertainty: instead of random events, the uncertainty is considered to be a lack of information as to events and their character. From this point of view, the gnostic distribution functions do not describe statistical probability. Their real sense is a quantification of inferences drawn from data as to 'what can be expected.' This means—in application to bounds LB and UB of data support—that 'prices of new cars of less than LB and over UB were not to be expected on the French market.'

The interval bounds of Tab. 25.3 can be used to define three classes of cars. The mean parameters of these are summarized in Tab. 25.4 for cars with Diesel engine and in Tab. 25.5 for those using gasoline. The symbol $\overline{X}$ again denotes the arithmetical mean of the data vector X. Price classes are symbolized by **BT** (below typical), **T** (typical) and **AT** (above typical). The fiscal category is denoted $FisC$.

Class	**Range (Euro)**	%	$\overline{FisC}$	$\overline{Vol}$	$\overline{HP}$	$\overline{MxSp}$	$\overline{Cons}$	$\overline{Pri}$
BT	9325 ÷ 16522	27.9	4.9	1.80	75.0	163.9	6.68	13496
T	16522 ÷ 24983	39.7	6.4	2.00	104.8	181.9	7.50	20176
AT	24983 ÷127481	32.4	9.4	2.42	146.8	194.3	9.18	35595

Tab. 25.4: Characterization of Diesel Cars

The technical parameters, as expected, rise with prices, although not as fast. The mean prices of the "economical class" **BT** are only about one third cheaper than those of the typical cars, but the prices of the "luxury" cars are much higher. Using the parameters of the gasoline classes allow a comparison with the Diesels to be made (Tab. 25.5):

Both typical and economical (lower than typical) gasoline cars are significantly cheaper than Diesels, but the extremely broad range of prices of the more luxurious gasoline cars raises their mean to nearly five times that of the economical cars. Compared to the Diesels, both the 'economic' and 'typical' gasoline cars are even cheaper, but the highest gasoline class has much higher prices due to both greater expected luxury and higher technical parameters. Typical class shares in both cases are approximately

Class	**Range (Euro)**	%	$\overline{FisC}$	$\overline{Vol}$	$\overline{HP}$	$\overline{MxSp}$	$\overline{Cons}$	$\overline{Pri}$
BT	5868 ÷ 13698	33.9	5.0	1.30	74.5	163.4	7.90	11143
T	13698 ÷ 23744	39.9	7.4	1.72	113.9	186.4	9.20	18007
AT	23744 ÷358979	26.2	15.2	3.00	213.7	218.7	12.62	50927

Tab. 25.5: Characterization of Gasoline Cars

the same, about 40%. The mean values of the 'tax variable' $FisC$ do not contradict the tendencies of the parameters derived from the interval analysis.

25.4.5 Ordering of Cars by the MD-Models of Prices

The model 25.7 enables car prices and their components to be ordered in multidimensional space. Indeed, this MD-model represents the relation $Probability(prices(parameters))$ as 'viewed' by the whole market. It is a hyperplane in the MD-space, where each car price's probability is depicted by two spatial points: the modeled price's probability (determined by the right-hand linear form of 25.7), which lies on the hyperplane and the actual price's probability $ProbPri_k$, which is located in a general case somewhere else. The probabilities of some prices (those of cheaper cars) fall below the hyperplane, while the more expensive ones appear above it. It is not necessary to determine multidimensional distances between the two points, because they correspond to differences between sides of 25.7.

These are of dimension one and can thus be ordered, when their polarity is taken in account. The order number ('*score*') can be used as an indicator of the position of the car on the 'ladder' of car types. Let the smallest score be assigned to cars, the modeled price of which is at the lowest point below the market price. In other words, the lower the score, the better price of the car given its technical parameters from the point of view of the car owner. The results of this ordering are summarized in Tab. 25.6 grouped by manufacturers. The score is expressed as a percent of the range, so eg score 125 for a Diesel will be shown as 100∗125/262=47.4%, while the same order number for a gasoline car would be 100∗125/504=24.8%. This is done, so that a total score (for both Diesel and gasoline drives) can be provided using the sums of scores weighted by the number of models offered.

The main message of these results is, that MD-ordering of cars based on the considered parameters leads to a conclusion, that the best relations between car parameters and price are obtained with the least expensive cars.

Recall, that a position in this table is not determined only by the price but also by the technical parameters. From the point of view of a manufacturer, a higher score (and a higher price) for his model rather than for a technically comparable competitive car is a success. It is also a (marketing) success to sell a car with inferior parameters, when better parameters

Car	Diesel Engine		Gasoline Engine		Total
Manufacturer	Types	Score%	Types	Score%	Score%
1	7	32.1	7	22.2	27.1
2	0	—-	2	29.5	29.5
3	2	42.0	9	27.7	30.3
4	7	31.6	4	32.2	31.8
5	28	35.6	42	31.9	33.4
6	16	32.1	18	34.7	33.5
7	27	46.8	59	34.4	38.3
8	35	43.0	58	39.3	40.7
9	4	46.6	8	39.1	41.6
10	21	41.6	24	41.8	41.7
11	18	43.4	27	40.7	41.7
12	17	45.5	24	39.3	41.8
13	6	48.3	17	43.6	44.8
14	2	55.9	10	43.0	45.1
15	10	61.9	15	37.3	47.1
16	4	74.8	5	30.2	50.0
17	0	—-	1	51.0	51.0
18	2	60.1	6	50.9	53.2
19	5	52.1	8	55.2	54.0
20	2	68.1	6	52.4	56.3
21	0	—-	13	59.2	59.2
22	1	81.3	2	49.0	59.8
23	5	69.6	4	64.6	67.4
24	13	68.1	22	74.2	72.0
25	0	—-	6	74.3	74.3
26	0	—-	6	75.8	75.8
27	4	79.2	17	75.1	75.9
28	0	—-	6	77.1	77.1
29	3	91.0	2	57.2	77.5
30	2	80.3	5	77.9	78.6
31	11	82.7	22	79.7	80.7
32	8	91.0	18	84.1	86.2
33	0	—-	5	86.9	86.9
34	0	—-	8	88.5	88.5
35	2	93.7	1	90.1	92.5
36	0	—-	2	92.6	92.6
37	0	—-	5	94.7	94.7
38	0	—-	2	94.9	94.9
39	0	—-	1	96.0	96.0
40	0	—-	1	97.0	97.0
41	0	—-	2	97.5	97.5
42	0	—-	3	99.1	99.1
43	0	—-	1	99.6	99.6

Tab. 25.6: Multidimensional Ordering of Car Manufacturers by Price Scores

are offered for the same price by a competitor. The buyer's decision in such cases must have been motivated by other aspects, eg perhaps by the prestige of the make.

The variable 'fiscal category' used in France for tax purposes can be interpreted as an official measure of the car's imputed luxury. There is no detailed description available of the mechanism, which defines the specific assignment of this parameter, but, as seen in Fig. 25.15, there is no serious contradiction between the multidimensional ordering (score) and the fiscal category attached to specific models.

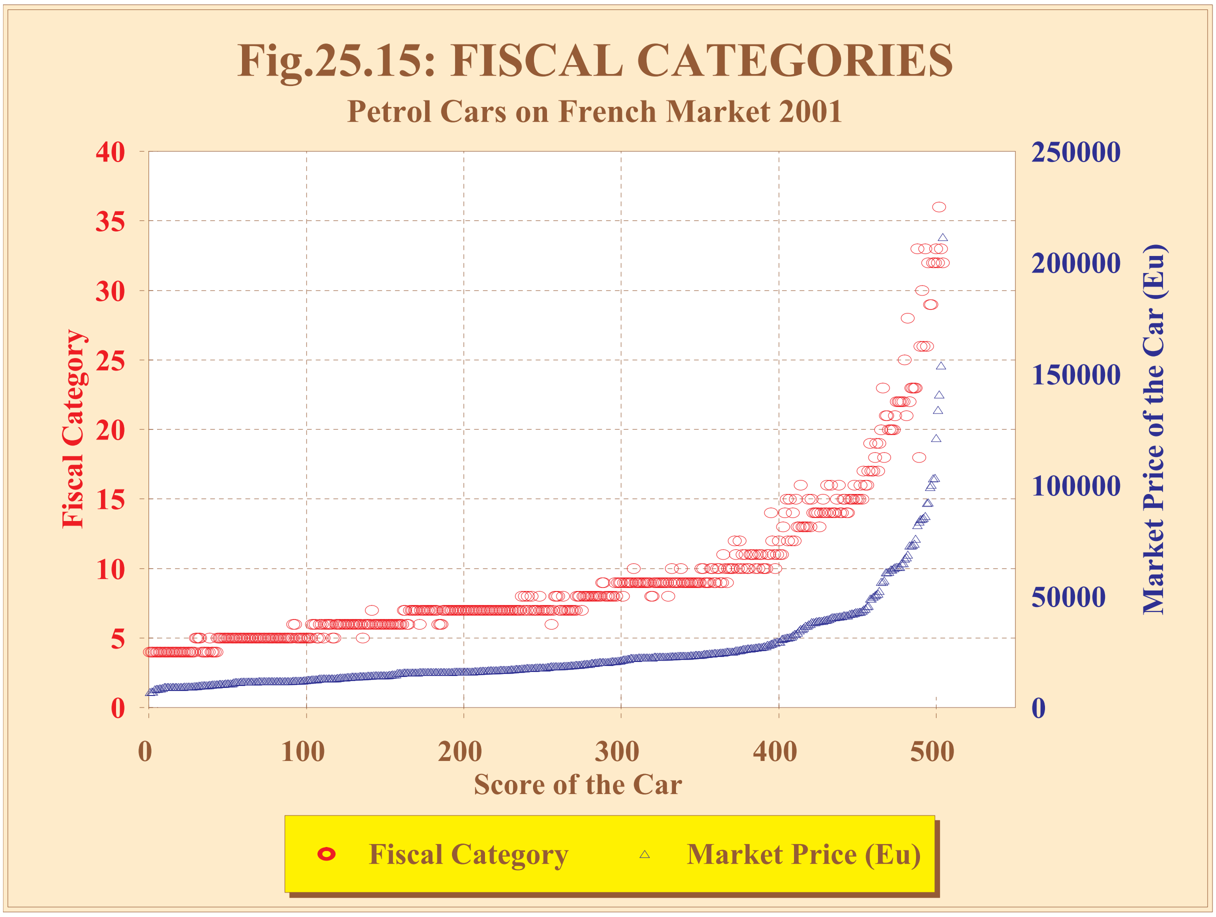


Fig.25.15: FISCAL CATEGORIES

Petrol Cars on French Market 2001

However, the uncertainty in the assignment of fiscal category to the various types of cars is greater than for the scores or prices.

It is interesting to note the relations between scores of Diesels and gasoline cars of the same make. There is a close correspondence not only for cheaper cars, but also in cases of more expensive ones. However, for some manufacturers, there are substantially different evaluations for their Diesels and gasoline cars. In these cases the position of Diesel cars is more favorable than for the gasoline versions. An explanation of this effect would require additional analysis: there are at least two thoughts in this regard, perhaps some extra perceived qualitative factor attributed to Diesels, but maybe also a different pricing strategy.

25.4.6 Impacts of Parameters on the Car Price

Equation 25.7 describes the dependence of the probability distribution of price on the probability distributions of the parameters, whose impacts are quantified by the model coefficients C_0 through C_4. The mean contribution of a k-th car parameter v_k to the mean probability $\overline{Pr\{Pri\}}$ of prices is $C_k * \overline{Pr\{v_k\}}$, where $\overline{Pr\{v_k\}}$ is the mean probability of the variable v_k. Expression

$$\rho_k := \frac{C_k * \overline{Pr\{v_k\}}}{C_0 + \Sigma_{m=1}^{4}(C_m * \overline{Pr\{v_m\}})} \tag{25.8}$$

can therefore be used to quantify the mean share of v_k on $\overline{Pr\{Pri\}}$. After all the distribution functions have been prepared and the model coefficients have been estimated by solving equation system 25.7, the ordering can proceed. After the mean values of the probabilities have been calculated, the impacts ρ_k can be determined and are reviewed in Fig. 25.16.

They can be interpreted as measures of 'what people want when buying cars.' Indeed, although the data characterize the supply side of car market, they provide insight into the demand side as well, because the market is more or less in equilibrium. The quantified impacts thus reflect not only the structure of the supply, but also the weighted interest of the buying public attributable to the individual technical parameters. The results shown in Fig. 25.16 suggest, that the following conclusions can be drawn:

1. The requirements for Diesel cars differ from those related to gasoline cars.
2. There are only two really significant factors, that influence the price of Diesels, power and—to a smaller extent—fuel consumption.
3. The impacts on prices of gasoline cars are less concentrated on the power, volume of the engine plays a role too.
4. Fuel consumption has roughly the same weight with either type of drive.
5. Maximum speed plays a relatively negligible role in both classes.

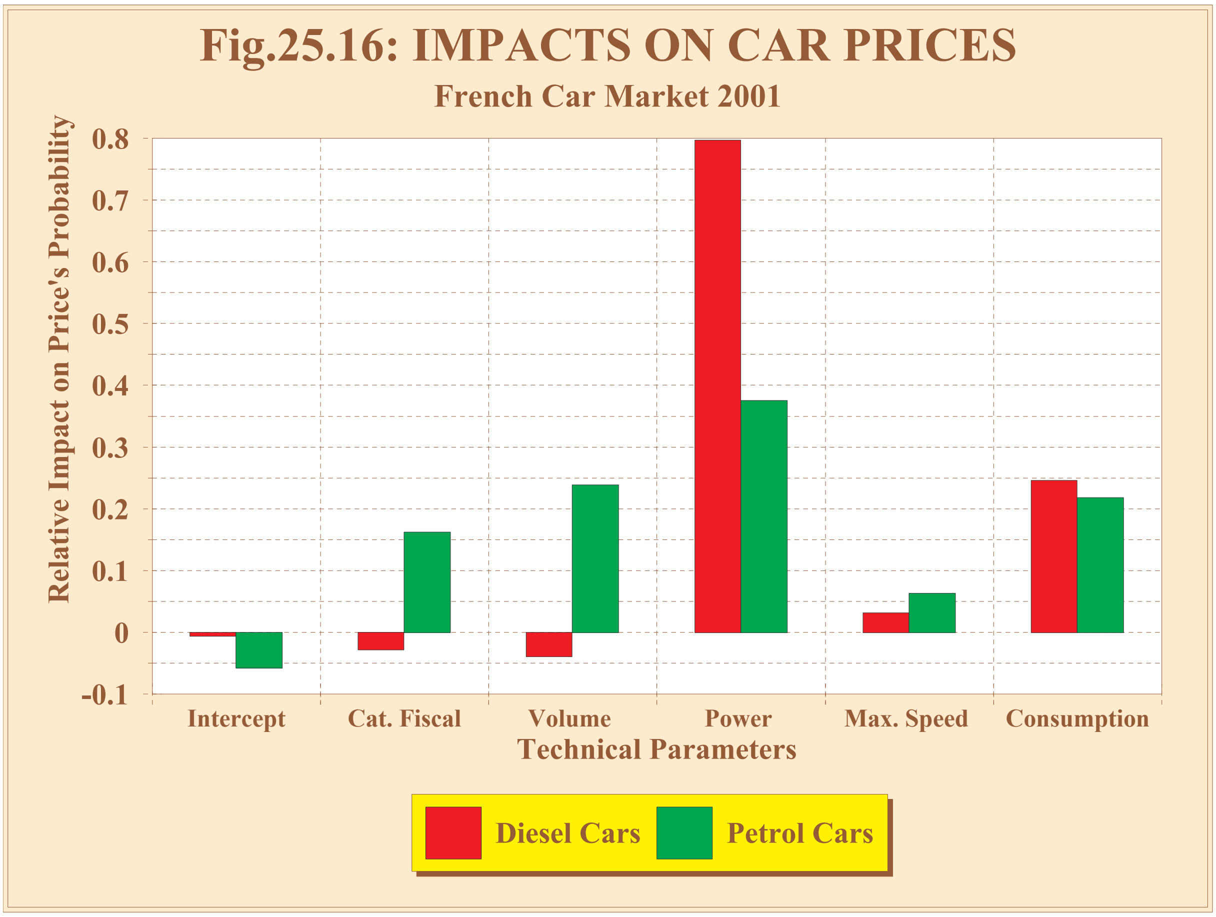


Fig.25.16: IMPACTS ON CAR PRICES

French Car Market 2001

6. The relatively small contribution of the intercept infers, that the model's parameters sufficiently explain the data, ie the chosen linear model is suitable.

The weak impact of maximum speed is a bit surprising, but it is probably due to the fact, that increasing maximum speed does not cost the manufacturer as much as increasing power. On the other hand, an emphasis on high power can be explained by reference to safety aspects: high power enables the driver to escape dangerous situations by rapid acceleration (eg when overtaking). However, a non-trivial question remains unanswered: is power exceeding 400 HP (which could move a military tank) really necessary for a personal automobile?

The impacts of the parameters will be discussed in more detail in connection with the multidimensional cluster analysis to follow.

25.4.7 Prestige, Goodwill & Premium Components of Car Prices

It is well-known from everyday experience, that two products of equivalent quality can be sold for different prices. There are at least three main factors, which lead to this market segmentation:

1. the goodwill of the manufacturer and of his trademark (brand loyalty),
2. the prestige felt by or imputed to the owner,
3. fashion.

Goodwill can be defined [9] as an intangible asset, that reflects value above that generally recognized in the tangible assets of the firm. It arises from the reputation of the name of the business and/or of its trademark or of the premises, where it is conducted or of the person, who has been carrying on the business [25]. Important roles are played by long-term tradition, generally shared and accepted positive experience and intensive and skillful marketing.

Prestige connected with purchasing and owning a product (and especially an automobile) is closely related to the reputation of the product's make. Frequently, the make of one's automobile is a symbol of the owner's social status. There are also countries, where the opportunity to obtain financing is measured by the estimated price of client's car. There are also places, where it is not socially acceptable to have a car of a class higher than that of one's boss. The classical proverbs "Fine feathers make a fine bird" or "The tailor makes a man" can be brought up to date by "Fine cars make fine men." And, of course, expensive cars are frequently used as a public demonstration of the personal wealth of the owner. All these factors help to raise the price of the most expensive models in an explosive manner as is seen in Fig. 25.12 and 25.15.

Neither is the **fashion** effect negligible. From the time of the giant American 'ships of desert', which accented aerodynamics through the cyclical change from angular to round lines of automobile design to the current SUV craze, this attribute has been a driving force in the car market. Fashion is sometimes considered an element of culture as well as a means of entertainment, but its first aim is business. Ladies cut their skirts so as to be subsequently forced to buy new long skirts. Men abandon their obsolete square-edged cars to buy modern round ones (and to return to the boxy shapes once again in the near future). Some component, or colors can be fashionable, even technical innovations to the engines or suspension systems, automatic devices, etc. Fashionable things of course raise the feeling of comfort (and price) and add to prestige.

Neither goodwill nor prestige are provided free of charge and it is interesting to establish their price. This can be done, because the hidden "fair and true" value of cars is estimated by quantiles of the probability distribution of the right hand side of the model 25.7. The difference between this value and the purchase price is a *premium*, which can be positive (gain to the buyer) as well as negative (gain to the seller). These premiums averaged by car makes are summarized in Tab. 25.7.

25.4.8 Ordering of Cars by Utility

The equation system 25.8 can be used to evaluate the utility of different car types. The implicit regression model in probabilities is obtained by robustly solving this system of equations. A score can then be attached to each car with respect to the difference between the sides of 25.8 as the equation's order number in the ordered series of these differences.

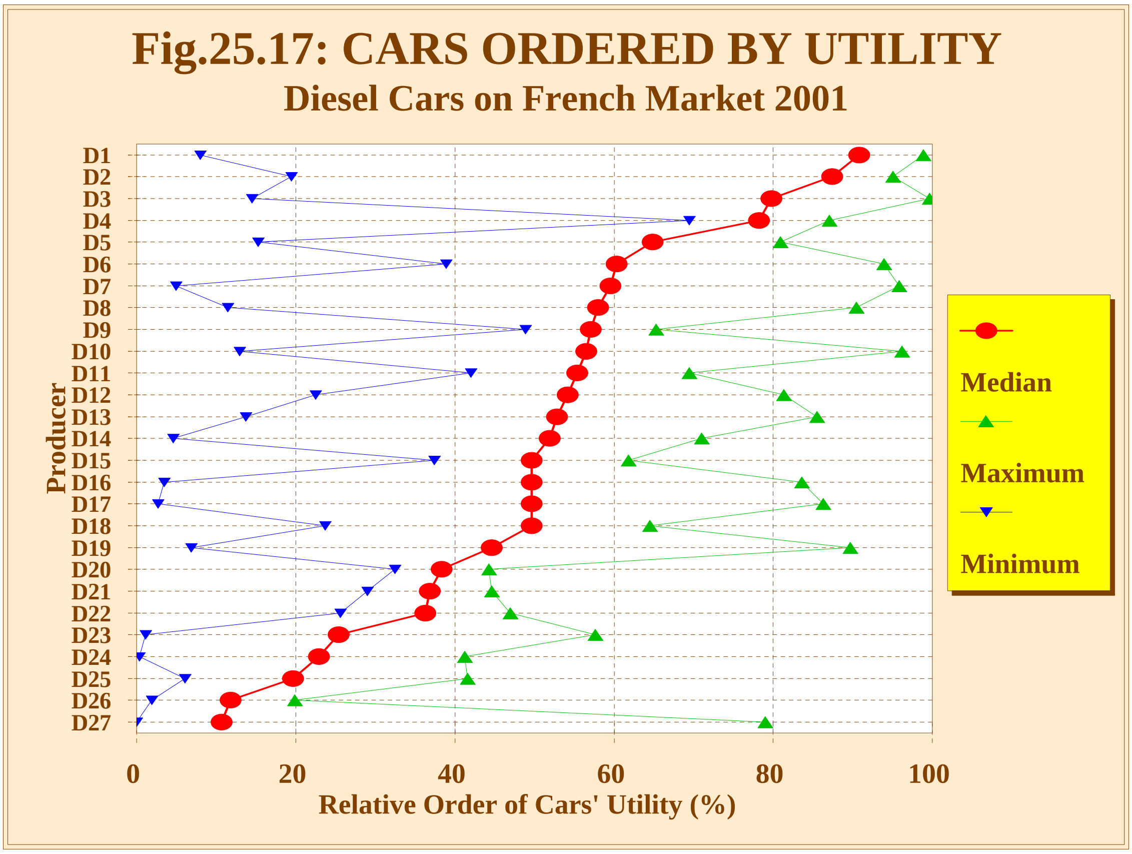
So as to be able to compare gasoline cars with Diesels, the relative order measure is expressed in percent and applied as $Rn = 100 * N/Nmax$, where N is the order number of the car and $Nmax$ the total number of the car types, ie 262 for Diesels and 504 for gasoline types. Most manufacturers offer several types, therefore to characterize the producer, the median, minimum and maximum values of the relative order are determined and shown in Figs. 25.17R and 25.18R:

The scale is chosen so as to attach the value of 100% to the largest utility; the smaller the value the less utility obtained for the purchase price. The main conclusion is in correspondence with the analysis of prices: the more prestigious and luxurious the car is, less utility is measured by the technical parameters divided by the price.

The results for the "economic" cars deserve a comment: not only do they occupy favorable positions in Tab.25.6, Tab. 25.7, Fig. 25.17 and Fig. 25.18, but they allow a broad and flexible compromise to be made between serviceability/economy and luxury using the wide interval between minimum and maximum utility in Fig. 25.17 and Fig. 25.18. In this way, a large portion of the market is addressed. There are large bands for the utility intervals offered by other large car producers, but not by all: the maximum utilities in both graphs fall with the medians: the manufacturers of the most luxurious cars have little interest in sharing the market for cheap cars.

Car	Diesel Engine		Gasoline Engine		Total	
Producer	$\overline{Pri}$ Euro	$\overline{Pre}$ Euro	$\overline{Pri}$ Euro	$\overline{Pre}$ Euro	$\overline{Pre}$ Euro	$\overline{Pre}$ %
A	——	—-	302606	-215465	-215465	-71.2
B	——	—-	265391	-121511	-121511	-45.8
C	——	—-	22089	-3473	-3473	-15.7
D	40778	-7655	59187	-9691	-9012	-15.2
E	35782	-6485	39988	-5796	-6052	-15.1
F	——	—-	83712	-11662	-11662	-13.9
G	——	—-	52988	-6679	-6679	-12.6
H	——	—-	28974	-3552	-3552	-12.3
I	28500	-2477	32784	-1891	-2003	-6.1
J	23224	-1404	20485	-1102	-1181	-5.8
K	19615	40	16977	-1312	-888	-5.2
L	27297	-800	27739	-1959	-1315	-4.7
M	46802	-3277	51274	-1850	-2289	-4.5
N	19460	-125	17611	-923	-657	-3.7
O	19704	-803	18102	-500	-641	-3.5
P	——	—-	73251	-2482	-2482	-3.4
Q	18560	-1409	15182	126	-488	-3.2
R	23355	3120	32758	-2069	-587	-1.8
S	20040	-1099	18034	428	-147	-0.8
T	23019	-1836	20618	418	-146	-0.7
U	23271	215	19061	56	96	0.5
V	——	—-	31278	321	321	1.0
W	19769	-935	16912	971	209	1.2
X	——	—-	28315	783	783	2.8
Y	——	—-	54645	1641	1641	3.0
Z	16386	487	12869	514	501	3.9
AA	——	—-	14033	660	660	4.7
AB	21161	1935	15947	453	1046	6.6
AC	15784	1007	14677	1066	1038	7.1
AD	17766	2667	19999	950	1611	8.1
AE	18176	684	15702	1700	1279	8.1
AF	20794	8	16663	1627	1358	8.1
AG	——	—-	129581	11235	11235	8.7
AH	——	—-	67078	5907	5907	8.8
AI	15201	1083	13644	1569	1260	9.2
AJ	26861	7665	43143	995	5442	12.6
AK	16754	1883	12839	1691	1726	13.4
AL	21327	3879	12668	1559	2591	20.4
AM	——	—-	39243	8561	8561	21.8
AN	24178	2902	13598	5504	4637	34.1
AO	——	—-	62235	21406	21406	34.4
AP	——	—-	36319	15530	15530	42.8

Tab. 25.7: Multidimensional Ordering of Car Manufacturers by Purchase Premiums

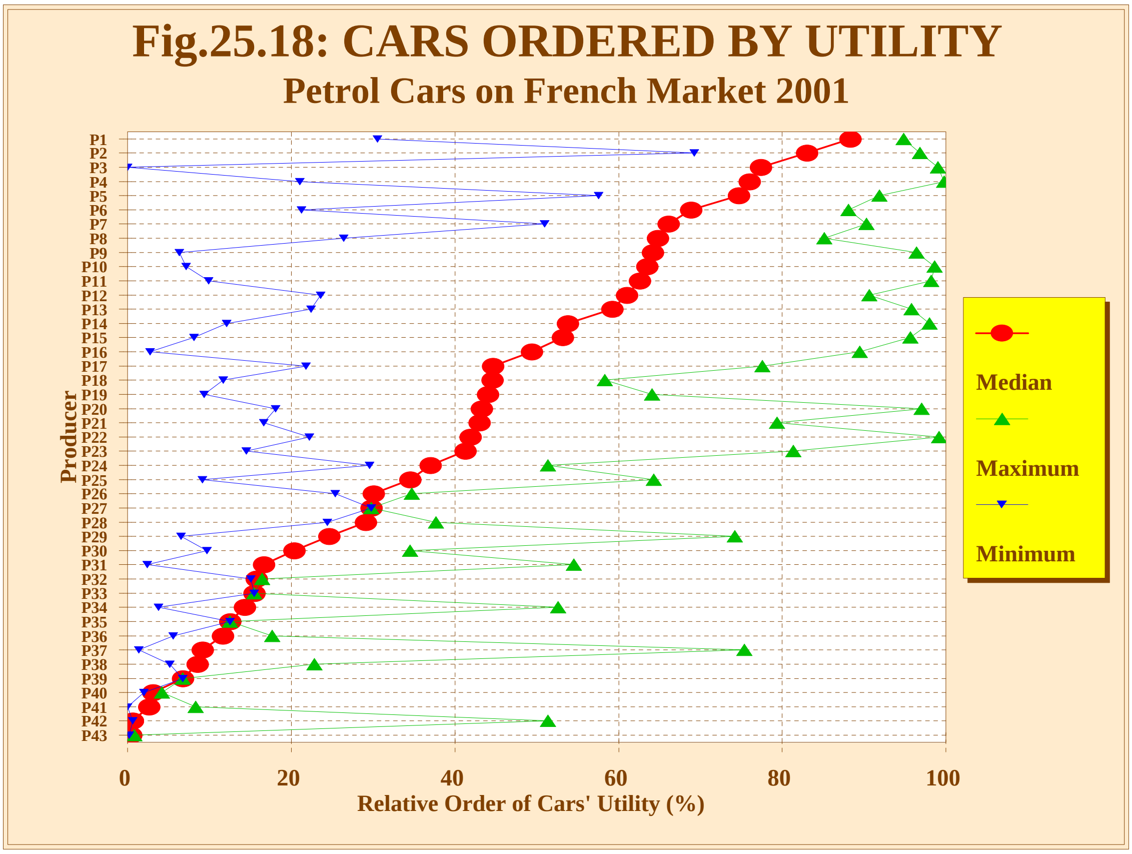


25.4.9 Structure of the Market

It was shown in Fig. 25.14, that there is no direct proportionality between such parameters as engine volume and power. To consider the variety of relations between parameters of cars, multidimensional cluster analysis can be applied to the equation system 25.8. These results are presented in Fig. 25.19 and Fig. 25.20.

The concept of comparability of multidimensional objects (financial positions of firms) has been introduced in previous chapters: multidimensional objects are comparable if their functions can be reproduced by the same mathematical model. The same concept can also be applied to equation system 25.8. The distribution function of the residuals of its robust solution reveals the inhomogeneity of the data.

To achieve a true comparability of car types, the whole set of types has to be decomposed into homogeneous clusters, each of which has a perfect model (in the sense of classical statistical variance analysis). This was successfully pursued to obtain six homogeneous clusters of Diesel types denoted CL1 through CL6 and formed by 139, 60, 35, 16, 6 and 5 car



types. One car appeared to be different from all the clusters. The analysis of gasoline cars resulted in eight homogeneous clusters CL1 through CL8 of 266, 105, 96, 12, 7, 6, 6, and 5 car types with one car, which had no peers and was outside of all the clusters. Both the ‘outliers’ are shown as ‘Rest’ in Figure 25.20.

The intra-cluster utility parameters along with the mean prices of the car clusters are given in Fig. 25.19 for Diesels and in Fig. 25.20 for gasoline cars. To be able to use the same scale for the parameters, some unusual currency units were applied: deciEuro . . . dEuro (Euro/10), Eurocent . . . Euroc (Euro/100), and miliEuro . . . mEuro (Euro/1000). The right-hand vertical axis in both graphs depicts the utility parameters using these units: volume in cm^3/dEuro, power in HP/Euroc, maximum speed in (km/h)/Euc, kilometrage in 100 km/liter/mEuro. The left-hand vertical axis shows the car price expressed in Euros. The general tendency of utility parameters to fall with rising price is distinctive, especially in the case of gasoline cars and for clusters of the most expensive cars. The relative patterns of clusters seem to be similar, but they have different models.

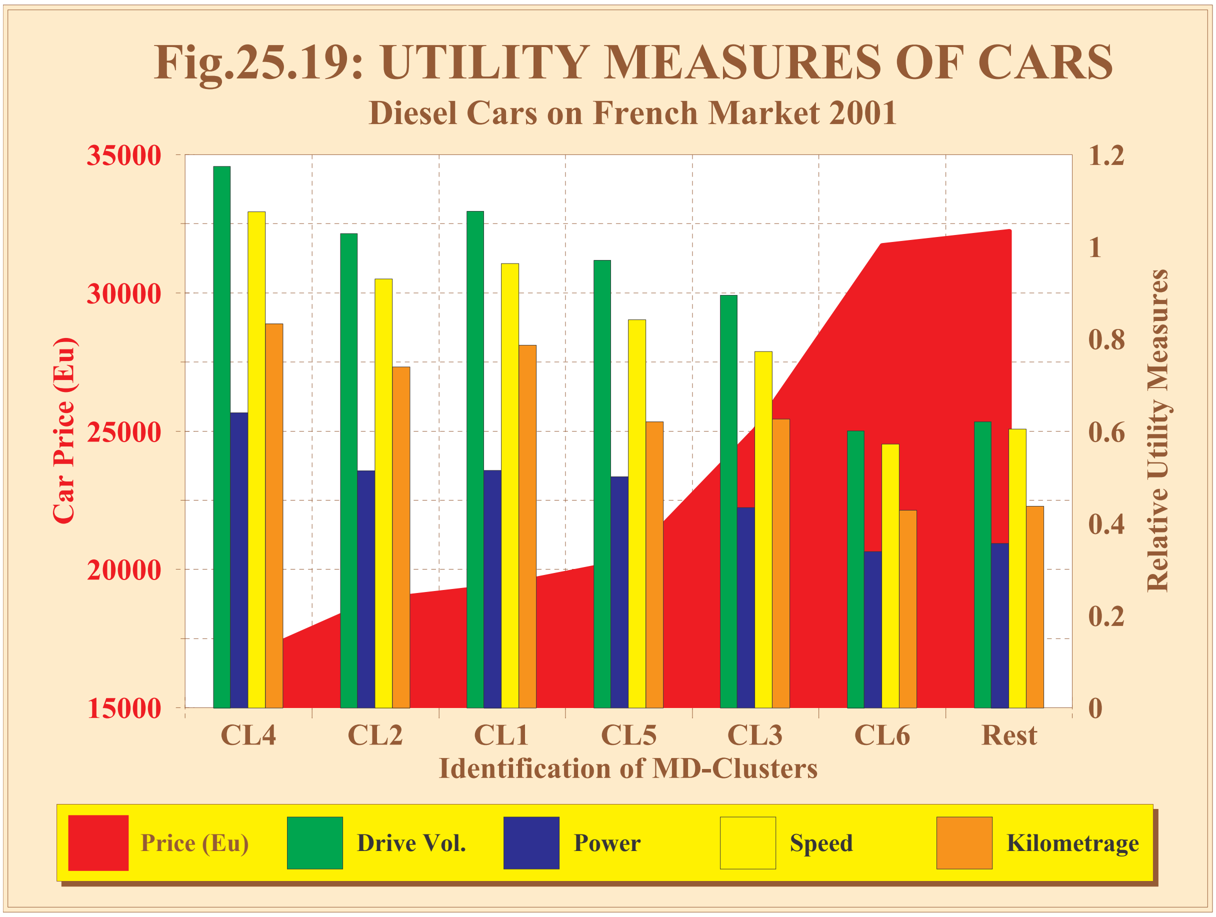


Fig.25.19: UTILITY MEASURES OF CARS

Diesel Cars on French Market 2001

The differences between clusters are emphasized, when the weighted impacts of the individual utilities in 25.8 are used instead of the mean parameters themselves.

This is shown in Fig. 25.21 and Fig. 25.22, where the product of the model coefficient and the mean value of the parameter represents the mean ‘contribution’ to the sum (equal to 1) of addends (coefficients multiplied by the means) in 25.8, the weight or mean impact.

The patterns of these weights differ in both graphs at first sight. Recall, that cluster CL1 of Diesels is formed by 139 of 262 cars and CL1 representing the gasoline models includes 266 of the 504 types, ie more than half for either of the classes. If the supply’s structure answers to demand, then the patterns of CL1 testify to the wishes of the prevailing part of the drivers (and of their families): Figure 25.21, CL1 shows, that the leading role in the decision to purchase a Diesel car is primarily linked to maximum

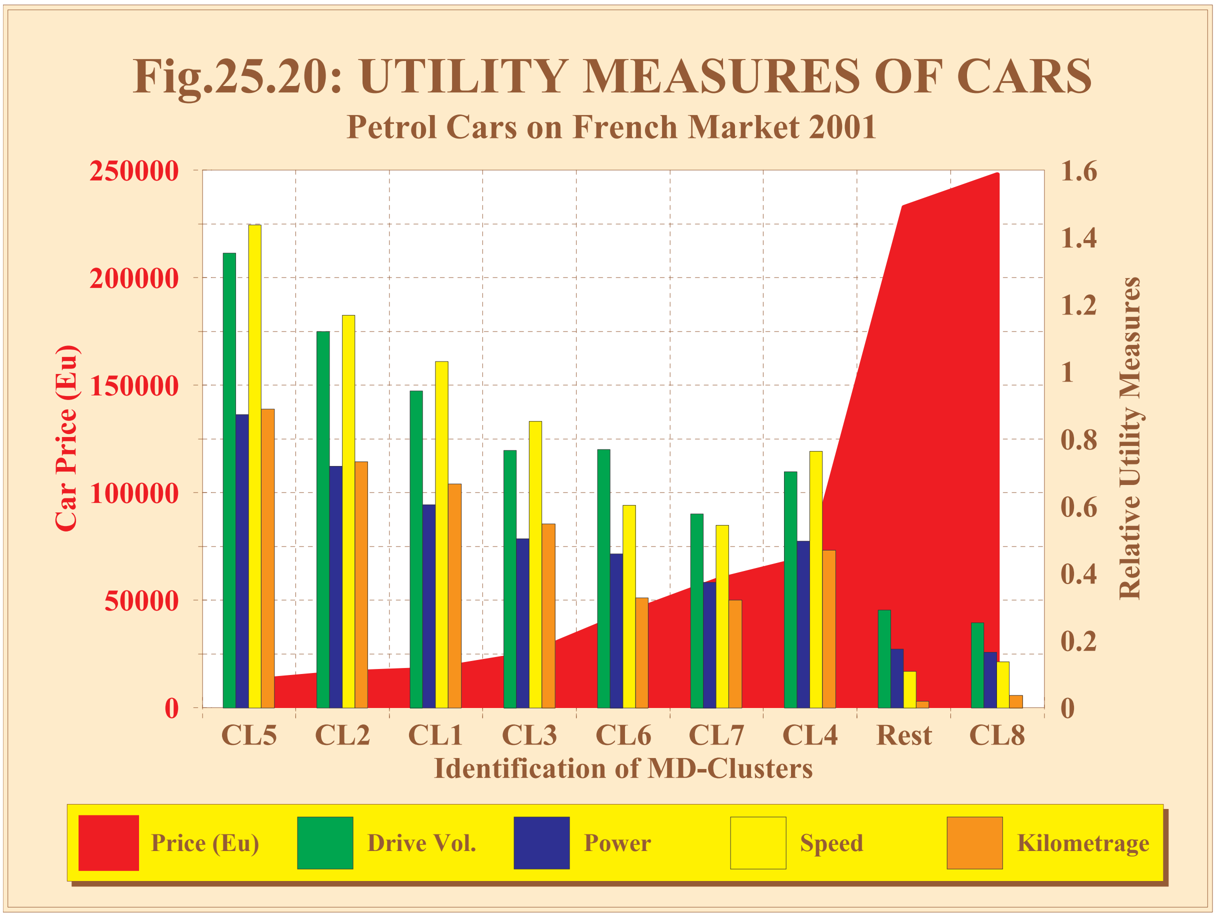


Fig.25.20: UTILITY MEASURES OF CARS

Petrol Cars on French Market 2001

speed[22] followed closely by the power aspect. Engine volume (the required power given) is not as important.

This is different from the largest cluster of gasoline cars (Fig. 25.22, CL1): the power and volume aspects play an equal role, dominating the maximum speed and kilometrage. A peculiar collection of gasoline car types is identified by CL6 in Fig. 25.22: high power and small engine volume. Some features are common to all clusters:

- high power requirements (acceleration) are universal,
- the inverse roles of maximum speed and kilometrage—the higher speed, the less kilometrage (the higher fuel consumption),
- in 10 of 14 clusters, the high power impact is accompanied by a strong negative impact for maximum speed,
- cluster CL5 in Fig. 25.21 has a structure similar to that of CL4 in

[22] This finding is interesting in comparison with Fig. 25.16, which shows, that maximum speed has only a weak impact on the Diesel car price.

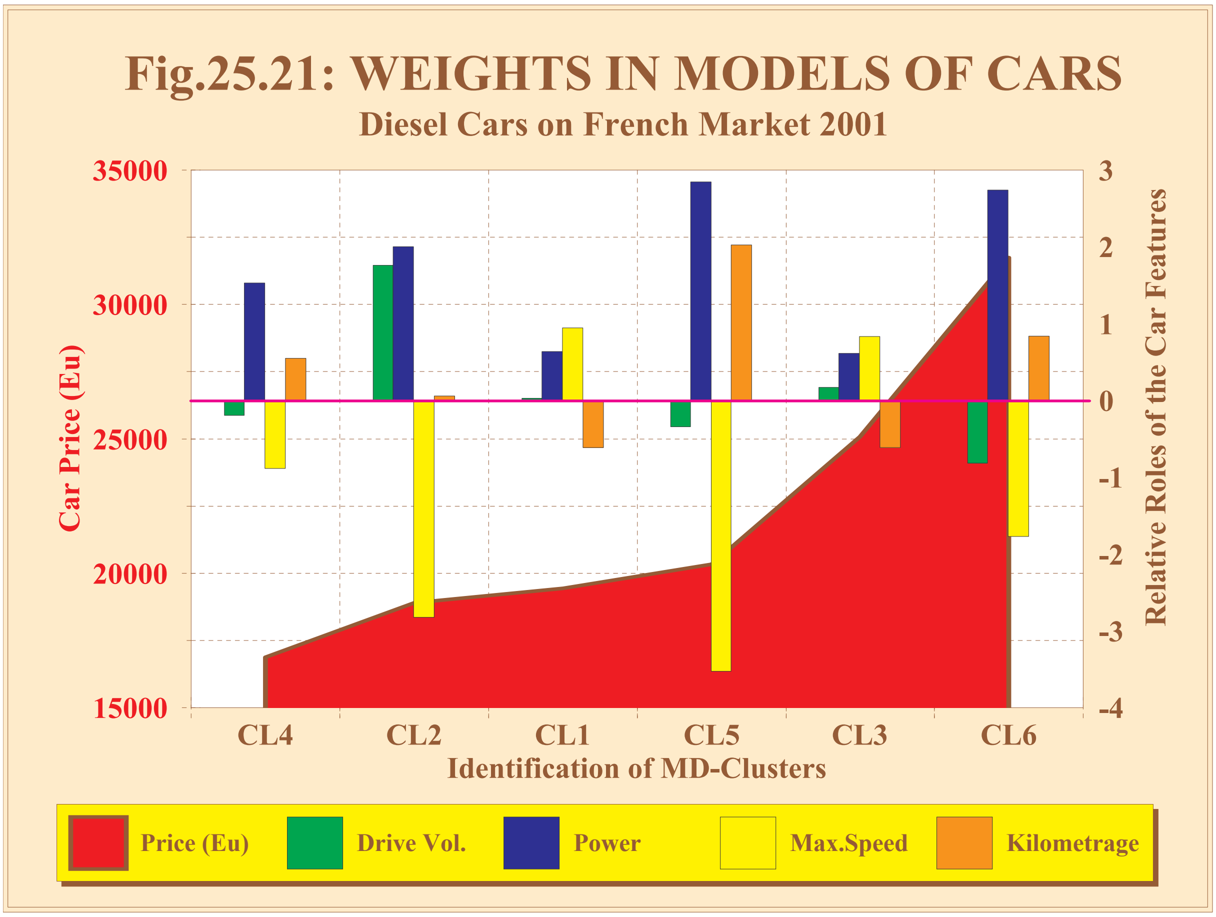


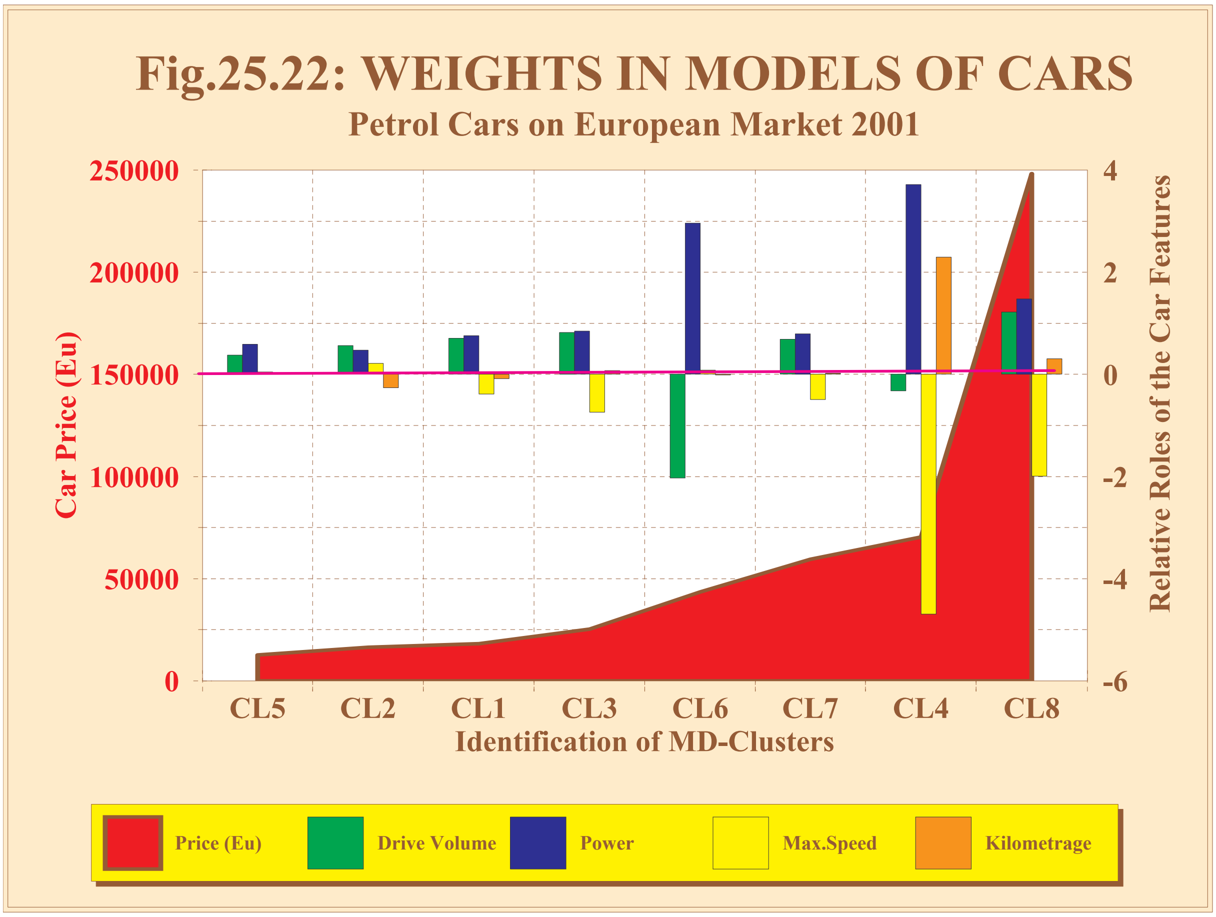
Fig.25.21: WEIGHTS IN MODELS OF CARS

Diesel Cars on French Market 2001

Fig. 25.22; in both cases, the mean prices are at the cross-over between rapidly increasing prices and the more economical models. In spite of the relatively high purchase price of these cars, fuel consumption is an important factor.

Cluster CL4 is the exception among the other gasoline cars, where fuel consumption does not play an important role. In contrast, for Diesel models, Fig. 25.21 shows, that in five of the six clusters kilometrage/consumption plays a measurable role in the composition of the car features.

The results show, that the advanced analysis put into a market framework is capable of producing information useful for both managerial and personal decision making.



25.5 Summary

The examples of applications of the advanced economic analysis to markets show, that for many problems, the solutions justify a more thorough analytical effort than that ordinarily applied:

- Beneficial information, useable for diagnosing the financial condition of firms can be obtained by the analysis of similarities and dissimilarities between different industries by means of
  – robust multi-marginal (repeated uni-variate) analysis of financial ratios,
  – analysis of robustly estimated variances and covariances of financial indicators,
  – robust multidimensional cross-section analysis,
  – comparison of robust multidimensional models,
  – robust multidimensional cluster analysis.

- A substantial improvement in the pricing of assets can be achieved as was demonstrated by using real examples:
  - unlike recent investment decisions based only on market valuations (external with respect to firms), which induce harmful and unnecessary volatility in share prices, a thorough (internal, based on firms' financial statements) analysis can reveal the true and fair financial position of firms and more reliably estimate their shares' value,
  - robust multidimensional models based on financial statements of homogeneous clusters of comparable firms enable a successful reappraisal of market share prices to be made.
- Robust filtering of time-series can lead to improvements in on-line trading as shown by the use of an example based on the foreign exchange market.
- The tools of advanced analysis can offer a deep insight into a specific commodity market as shown by example of a large segment of the European automobile market:
  - a robust multidimensional pricing model for cars can be established, which reflects both technical and economic parameters,
  - a multidimensional evaluation of the utility attributable to various car models and the ordering of car types by their utility can be performed,
  - the impacts of technical parameters on car prices can be quantified,
  - car manufacturers can be compared and ordered by the technical and economic aspects of their cars,
  - the role of manufacturers' goodwill and their makes as reflected in car prices can be evaluated,
  - the structure of the market can be described by means of multidimensional cluster analysis. This can provide an orientation for designers on the requirements and tastes of potential customers.

Chapter 26

Miscellaneous Applications

26.1 The Role of Economics of Information

26.1.1 Universality of the Problem

A number of examples of the application of gnostics to problems of both micro- and macroeconomic were introduced in Chapters 23 through 25. They document the usefulness of advanced analytical methods in economics. However, a straightforward and unbiased view of these examples reveals, that they are solutions of special cases of more general problems, which are faced universally in many branches of science, technology, medicine, biology, etc. These situations are noted especially, when the reality of an event and its processes are to be recognized and quantitatively evaluated by data analysis. The important elements of the problem are:

1. Wherever and however obtained, data are expensive.
2. The higher the data quality, the higher its cost.
3. The number of data available for analysis can be limited, not only by cost, but also by other restrictions of a more fundamental nature.
4. Real measurements are always disturbed by undesirable effects.
5. The desired information is not generally given directly by data, but it must by extracted from them through analysis.

In all application fields obtaining quantitative information bears a cost and it is the outcome of some special technology. The acquisition of information and processing it economically is therefore an important aspect in the choice of data and its processing methods. The advantage of gnostics over other approaches is its ability to measure the amount of obtained information under much more universal conditions, restricted only by a few realistic and plausible assumptions. This special technology allows the maximum amount of information to be gathered from data by using gnostic methods.

The ultimate significance of taking into account the economic aspects of information can be measured by considering the utility of the information, which is acquired and reflected by the gains in knowledge or the losses resulting from not having had the best data, on which to base a decision. A correct and timely diagnosis can save the life of a patient. The culling out of defective products or the recognition of factors, which lower the useful life of goods result in increasing the producer's goodwill and in a justification for higher prices. An emergency signal extracted from noisy real-time measurements can prevent the dangerous development of a process, when the data treatment method is sensitive enough to real changes in the process, while it minimizes false alarms. Recognition of the true nature of objects and processes along with a capability to create realistic models permits these real processes to be efficiently controlled. The economics of everyday life is thus conditioned by the economics of information processing.

26.1.2 Information, Normality, Geometry

Information is the goal of the game, assumptions such as 'normality' are tools to assist in reaching this goal. The statistical notion of normality is tied to the special case of a Gaussian probability distribution and was discussed in section 20.3.2. Reliance on the unquestioned applicability of an assumption of this type has serious negative consequences:

1. Tests of the normality of processes and of membership in data samples can generate false information.
2. This notion of normality can be unacceptably narrow.
3. Prejudices as to the Gaussian character of events can distort information as to the nature of the process being observed.

A decision of the type 'normal/not normal' leads to specific diagnostic and/or discriminative information. Consider typical examples of such a diagnosis:

1. (a) "The object behaves in accordance with the Gaussian probability distribution. Hence, it is normal."
   (b) "The object is normal, therefore its probability distribution is Gaussian."
2. "The value of an event is normal, if and only if it belongs to the interval

bounded by prescribed quantiles of the Gaussian distribution."

The information inferred from such a diagnosis can have far reaching or even critical consequences: it can miss a natural, economic, technological or even a military disaster as well as trigger a very expensive false alarm. Recall a period of the cold war, when the false recognition of objects in space could have ignited a nuclear war. Not only incidents, which could have repercussions for all of society fall into such categories; it is far more likely, that events dealing with industrial production and control or medical trials or studies would be subjects for examination.

A careful consideration of the roots of these problems inevitably leads to geometry. Indeed, the question "how far is this distribution from that (eg 'normal') one" concerns the determination of the **distance**. Analogously, the question of "how far a datum is allowed to be from other data in the sample to be still accepted as a member" is again a function of distance. Both are thus problems of geometry similar to those visited in sections 6.2.1 and 9.2.2. It will now be useful to go a step further.

Assuming, that it is the true distribution of data, let $P(x)$ be a distribution function attached to the data sample $\mathcal{X} := \langle x_1, \ldots, x_N \rangle$. The probability, that x will exceed a value x_U is

$$P\{x > x_U\} = P(x_U) \tag{26.1}$$

and its differential is

$$dP(x) = g(x)\mathrm{d}x, \tag{26.2}$$

where $g(x) := \frac{\mathrm{d}P}{\mathrm{d}x}$ is the probability density. Such an expression was identified in 9.2.2 as a univariate formula for distance measured in a special type of Riemannian geometry defined by the weighting function $g(x)$, which is the corresponding univariate metric tensor. When this function is declared to have the Gaussian form of probability density,

$$g(x) = \frac{\sigma\sqrt{2\Pi}}{\exp}(-\frac{(x-\mu)^2}{2\sigma^2}), \tag{26.3}$$

then this causes the **special geometry to be chosen**. The Central Limit Theorem justifies convergence to this density function for a sufficiently large number of arbitrarily distributed independent random variables, which have a mean μ and variance σ^2. The special geometry is therefore well founded in this instance, but not in every general case. The existence of a mean and of a variance cannot be taken for granted. The non-zero probability of a large outlier can jeopardize the validity of the normality conditions especially if there is an insufficient number of central data, which can override the effect of the extreme datum.

It is a common practice to mathematically transform the real data so as 'to make them normal.' There are places, where official norms of quality assessment methods not only allow such practices, but even recommend them. Riemann's view, that the choice of geometry is not a task for mathematicians, but of Nature, has already been cited. Therefore the arbitrary choice of a Gaussian distribution for non-normal data is an unjustified a priori designation, which distorts both the data and the results of any analysis conducted on them.

26.2 Research Methodology

This section briefly reviews several engineering/industrial problems, which were resolved through the use of gnostic methods.

26.2.1 A Geological Survey

Geological materials are standardized using traditional internationally established procedures:

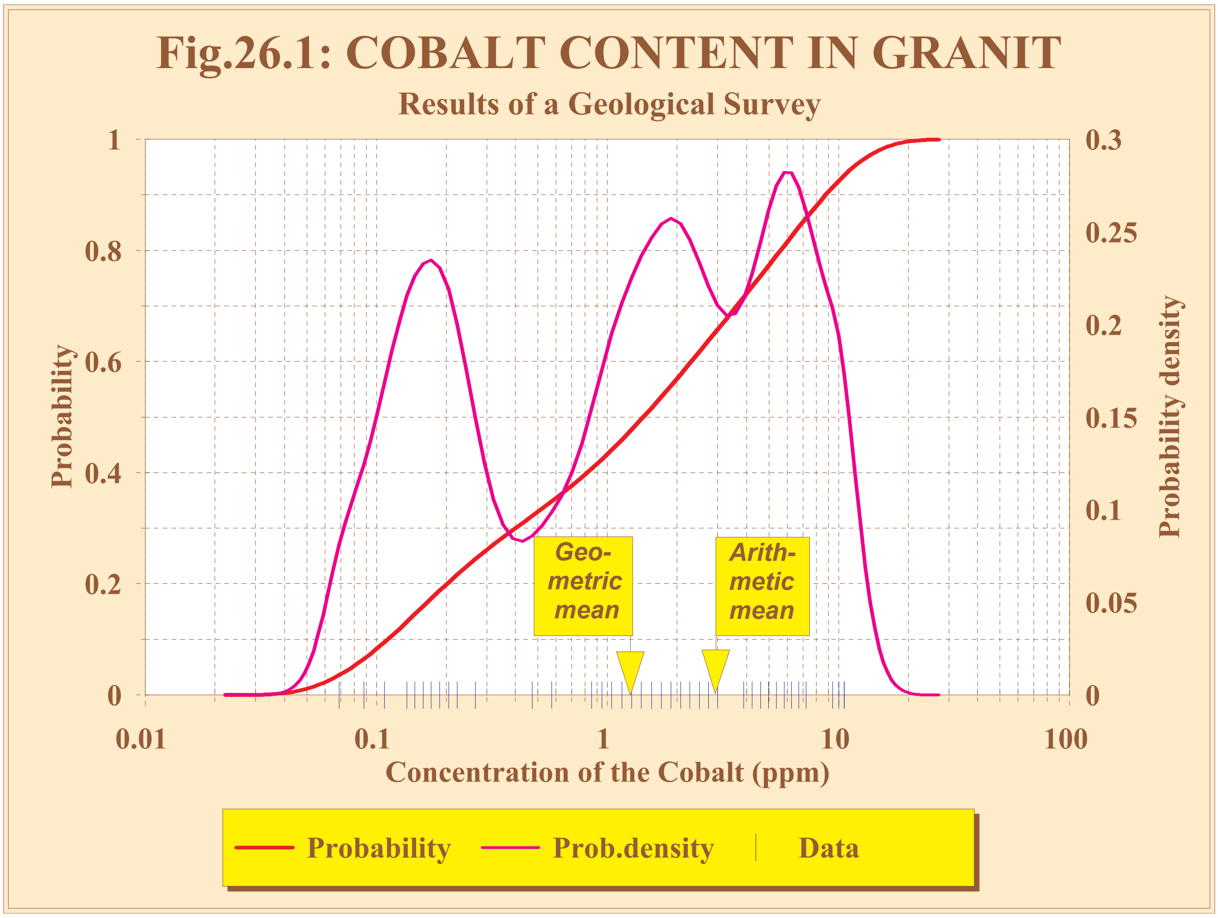
1. The geological raw material to be standardized is defined.
2. The location of the deposit of the cleanest and most typical material of this kind is identified.
3. A qualified national institution in that country is selected to conduct a specific survey.
4. An appropriate amount of the raw material is extracted.
5. It is pulverized, homogenized and divided into a large number of portions/samples.
6. Portions of the samples are distributed among leading geochemical and geophysical laboratories the world over.
7. The reports of all these laboratories are accumulated by the responsible national institutes.
8. The foregoing results are given a final review and a closing survey report is prepared along with a specification, which contains a summary of the analyzes.
9. The samples with their specification are made available on a worldwide basis so as to serve as standard for the material.

In the late 1980's interest was focused on the storage of spent nuclear fuel rods, which emit a strong flux of fast neutrons as they decay. The half-life of such radioactive residue is extremely long therefore these radioactive wastes must be contained in safe repositories. For this reason, granite, and especially its cobalt content, became a point of interest, because cobalt has a very high cross-section with respect to thermal neutrons, ie it is capable of very efficiently capturing them. The construction of repositories in large granite deposits is suitable for storage purposes because of the high chemical, geological and mechanical stability of the material and its nuclear characteristics. The fast neutrons emitted by the radioactive wastes are slowed down to thermal energy levels by the hydrogen nuclei of water and—with a high probability—captured by the cobalt.

A deposit of granite suitable for consideration was found in the Czech Republic and the Czech Geological Survey was charged with carrying out the survey. At the time that the worldwide analytical report of the search for suitable granite deposits was being prepared, an experimental gnostic analyzer was made available to the Czech team. An analysis, using this technology, led to the publication of [82], which represented a marked departure from the usual results summarized by the sample mean plus minus the sample estimate of the standard deviation. The conclusions of the summary report of the survey are reproduced in Fig. 26.1.

Instead of the expected nice single bell-form of a Gaussian density curve, three distinctly separated peaks with widely diverging maxima appeared. Initial disbelief in this outcome was resolved, when a more thorough examination of the laboratory reports revealed the true origin of the inhomogeneity: each of the three clusters resulted from a different class of chemical or physical methods, meaning, that systematic errors were made in the application of the techniques used. The individual laboratories did not catch the errors, because their measurements were taken using the methodology, in which the laboratory was specialized and the results had an acceptable spread.

Both the arithmetic and the geometric mean, marked in Fig. 26.1, are respectively equal to 2.73 and 1.07. It is difficult to interpret, what these numbers (and the standard deviations of 2.96 and 1.93) reveal. It suffices to say, that had they been taken as the survey's conclusions, information of fundamental importance on the inhomogeneity of the results and the inconsistency of measuring methods would have been lost.



26.2.2 Incompatibility of Measuring Methods

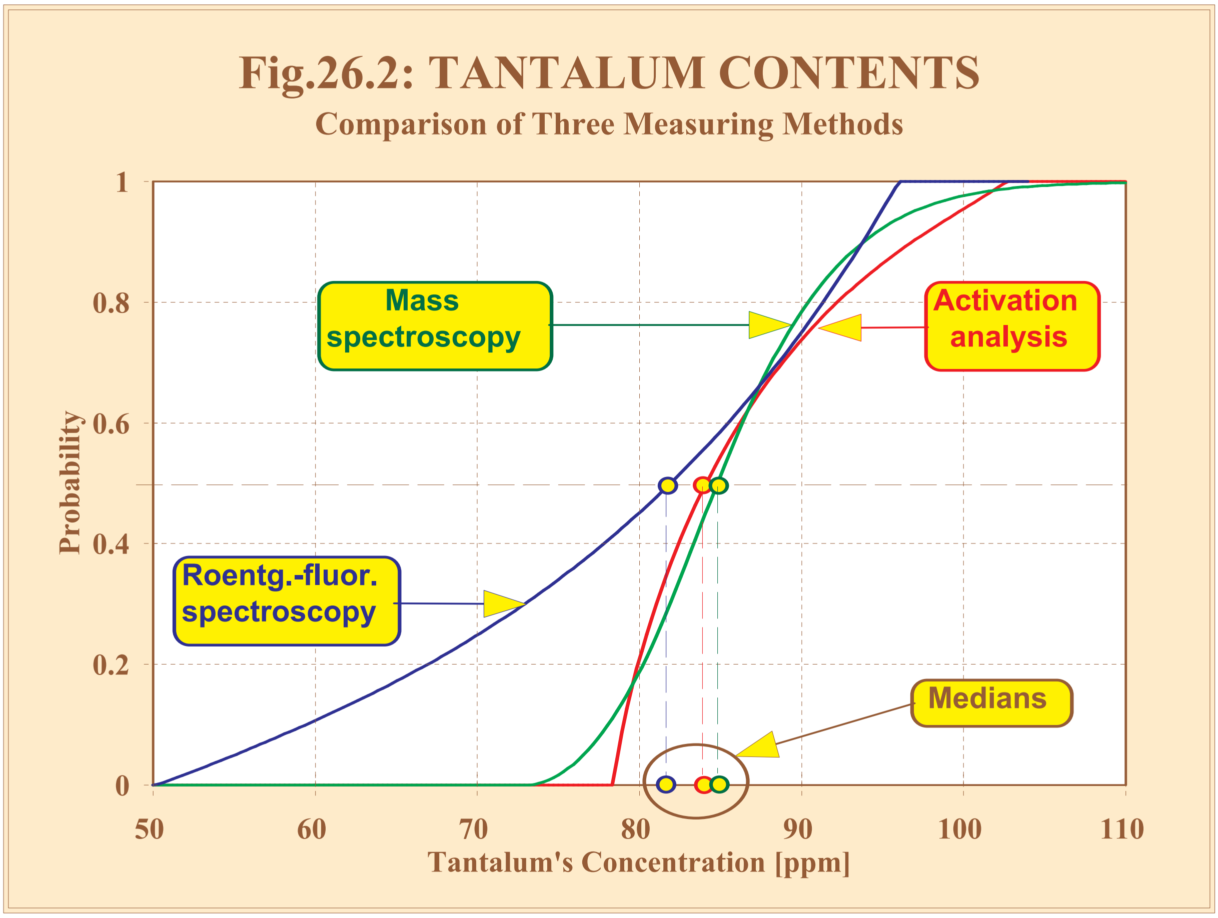
A further example shows, how—without having knowledge of the distribution functions—the use of only the point statistics of a sample or an ordinary regression can lead to a false interpretation of information from the data. In addition, this example also shows how results can vary with different measuring methods. Data cited from [76] are in Tab. 26.1 and are taken from three methods used to measure the tantalum content of a sample:

NM . . . nuclear method (activation analysis),
MS . . . mass spectroscopy,
RF . . . roentgen-fluorescence method.

None of the customary classical statistical tests to measure the differences between the three methods, also cited in [76], *were statistically significant (at the level of 0.05)*. The application of a gnostic distribution function (EGDF) to this data leads to Fig. 26.2.

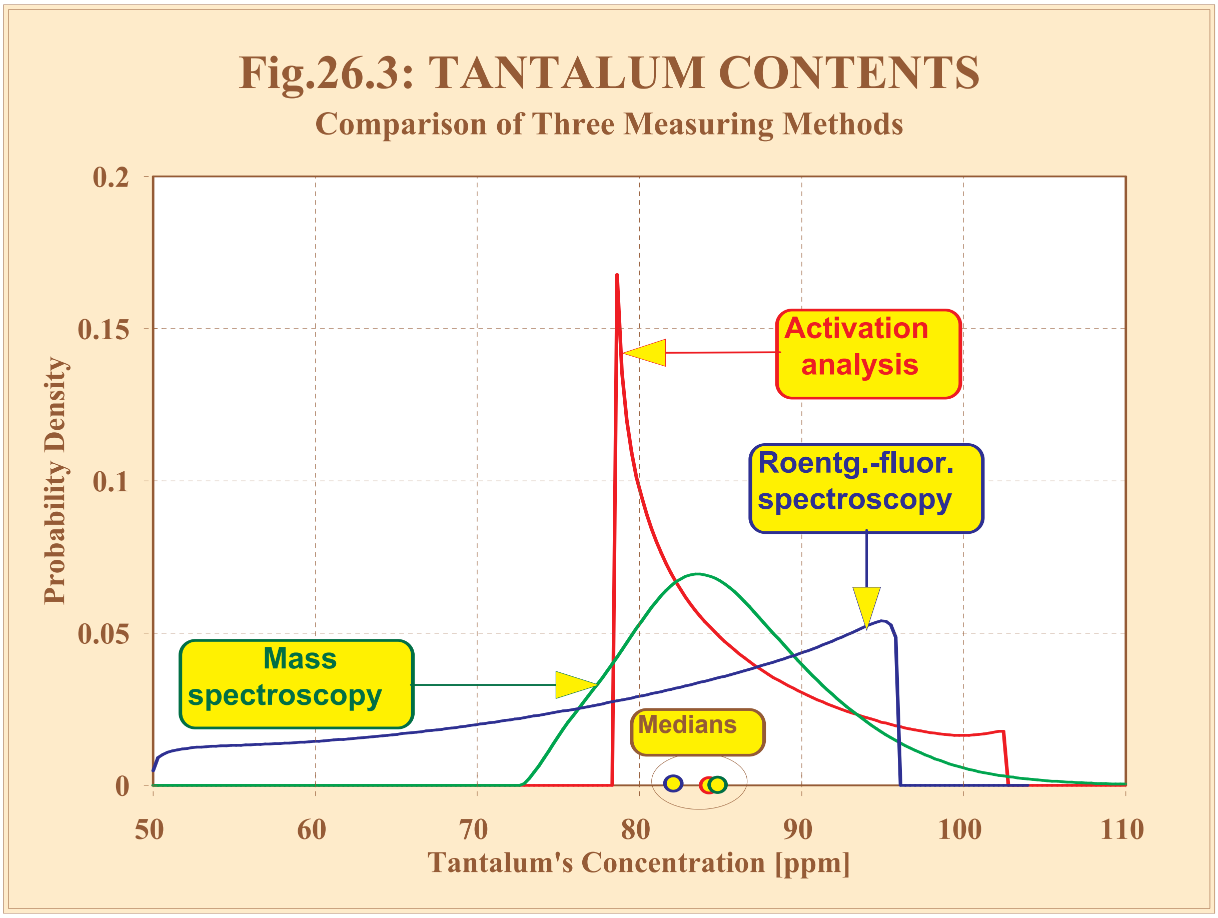
Method	Data									
NM	68.00	78.72	79.79	80.90	81.43	85.30	86.50	89.00	92.70	100.00
MS	32.50	77.16	82.00	83.00	85.63	90.59	93.00	—	—	—
RF	56.00	65.00	74.00	82.00	82.33	86.17	—	—	—	—

Tab. 26.1: Data from the analysis of tantalum content



The EGDF's probability densities are shown in Fig. 26.3. The following conclusions can be drawn from these graphs:

1. None of the distributions can be taken either as normal or as lognormal, because the data supports for all three distributions are bounded from below by a large positive bound. Moreover, two of them are also bounded from above.
2. The medians of the three distributions, determined as quantiles of probability 0.5, are close to each other.
3. The forms of the three distributions are not at all similar; this is seen especially in Fig. 26.3.



It is vividly evident, that while the medians are very close to each other, when the gnostic distribution functions are introduced into the analysis, the wide differences between the three measurement techniques can be clearly seen and their equivalence cannot be justified.

The above illustration is also suitable to demonstrate another important task: the identification of an outlier.

26.2.3 Membership of an Extreme Datum

The most frequently used approach to quality assessment consists of testing a product's quality parameter q to see whether it satisfies the inequality $|q - \mu| < K * \sigma$, where μ is the arithmetical mean of the data sample and σ is the sample estimate of the standard deviation. The constant K is determined so as to ensure the required statistical significance for the test.

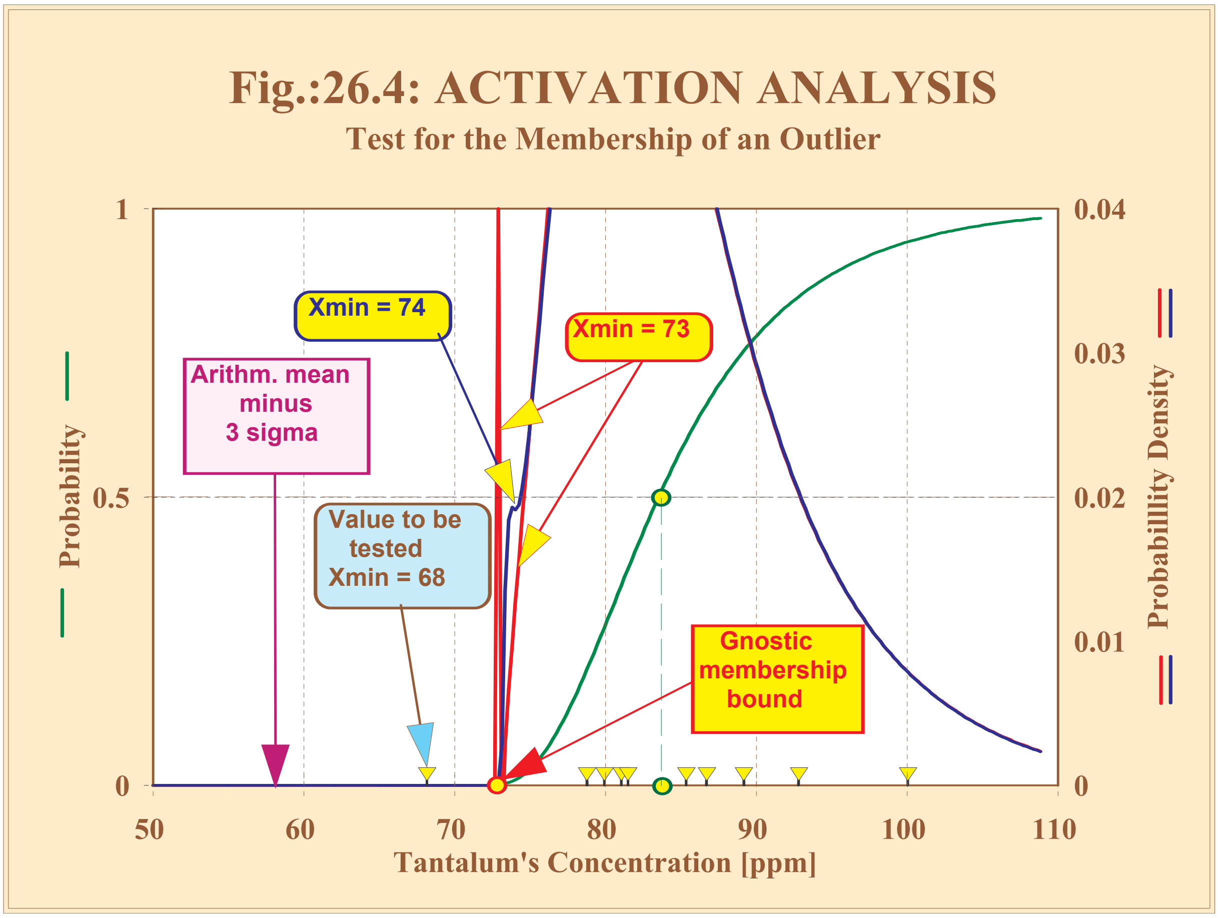
It is evident, that such a test can be justified if and only if the statistics μ and σ are sufficient to identify the actual distribution of data. For a normal distribution, their sufficiency is warranted, but even then at least four serious objections arise:

1. The test is based on the assumption, that even when there is a datum exceeding the 'tolerance,' the actual distribution remains unchanged and normal. What evidence is there to support this assumption? The presence of a significant outlier can also be interpreted as a change in the form of the distribution.
2. Both statistics μ and σ are unrobust, ie they are sensitive to outliers. When their 'old' values from a previous 'good period of the process' are applied, then there is no control as to their actual recent values and the test is unreliable. When the recent/current estimate of these statistics is used, they are distorted by the outlier and the test is also unreliable.
3. The choice of the constant K is mostly based on a subjective decision. The test is then subjective.
4. The tested inequality is related to the determination of the distance between the extreme datum and the mean of the sample. Therefore, the applied geometry corresponds to the (subjectively and a priori assumed) distribution. The test is once more subjective.

The tantalum example discussed above allows the gnostic approach to this problem to be examined together with the usual statistical solution. The **MS** (mass spectroscopic) data from Tab. 26.1 are depicted on the horizontal axis in Fig. 26.4 by triangular marks.

There are six data spread over an interval from 77.17 through 93.00 and one 'suspect point', 32.5. A careful analyst should test such an extreme data value for 'membership' in the sample. For $K = 3$ (which is the frequently used value) the arithmetic mean minus three standard deviations equals 15.6. From the point of view of this statistical test the extreme value 32.5 cannot be rejected as an outlier and it would be accepted as an orderly member of the data sample.

In contrast to the above, the gnostic approach to the bounds of 'membership interval' of a data sample (which was described in section 15.3.7) is based on the fact, that the probability density of the global distribution function (EGDF) has only one maximum in the case of a homogeneous data sample. This distribution function is continuous and all its derivatives exist. When a data sample's characteristics change and tend toward inhomogeneity and so to a second maximum in the EGDF's density, there is a critical transient moment at that point called the *point of inflection.* This is seen in Fig. 26.4: the six valid members of the data set are fixed and



the seventh one (the suspect) is moved upward from 32.5. The blue and red lines respectively depict the density function, when the tested datum reaches 52 and 53 respectively (using an enlarged scale). The red curve shows, that for $x = 51$ (and fixed other data) the point of inflection is reached, while for $x = 52$ the density still rises monotonously. It is therefore logical to accept $x = 52$ as the lower bound of the samples membership interval. It is not necessary to manipulate the density function in this way to determine this bound. Instead, equations 15.40 through 15.42 can be solved simultaneously. It is obvious, that by doing this, no subjective intervention is required, the solution is unique and it is determined only by data ('speaking for themselves').

The results of the gnostic test has an opposite outcome with respect to the arbitrary 'three sigma' approach: the extreme value 32.5 cannot be considered as belonging to the sample formed by the six other data.

Figure 26.4 provides a similar outcome for the **NM** (nuclear method, activation analysis) method.

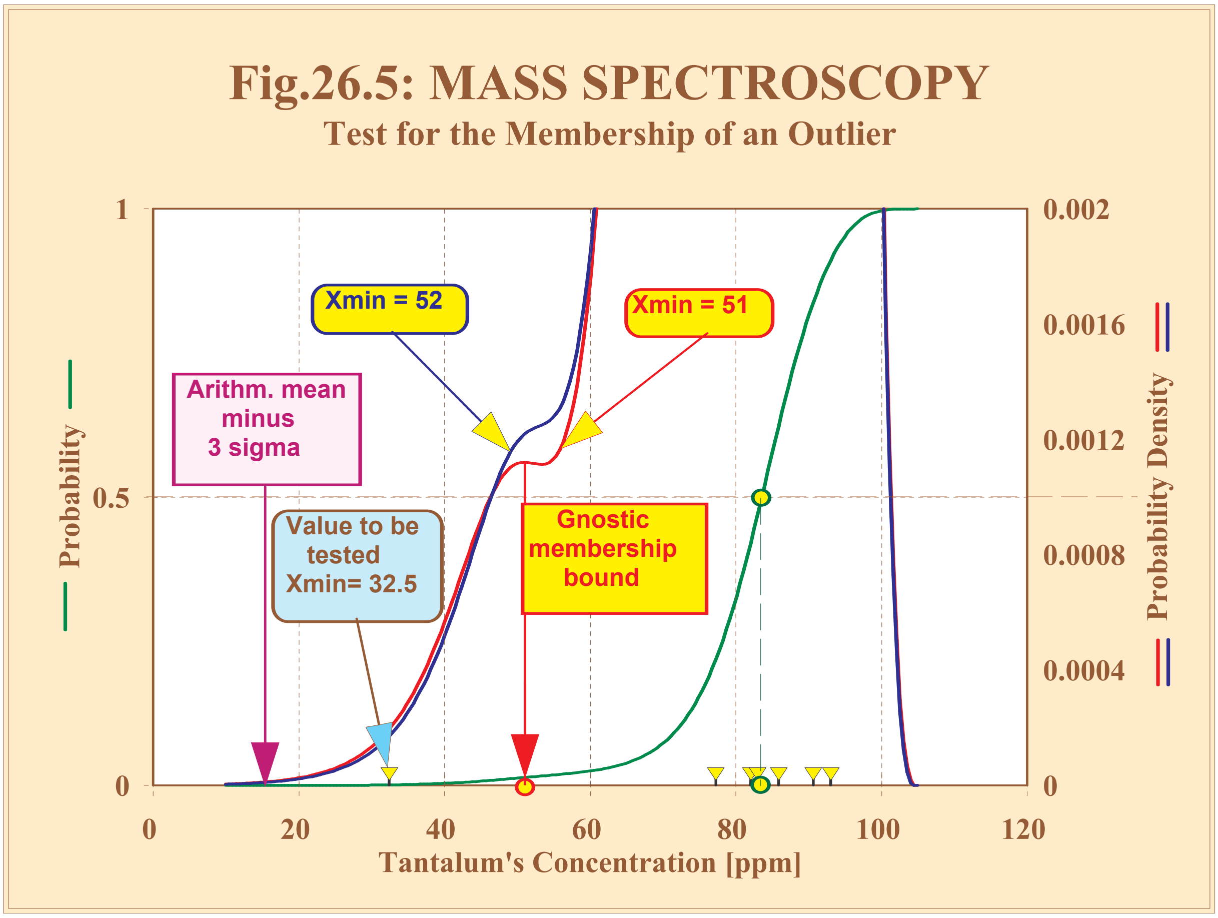


Fig.26.5: MASS SPECTROSCOPY

Test for the Membership of an Outlier

The smallest value, 68, located well down in the interval allowed by the statistical test is rejected by the gnostic membership test, which declares 73 to be the lower bound of the membership interval.

26.2.4 What Kind of World Do We Live In?

“Leben wir in einer Volterra-Welt?”[1] is the title of [85], which summarizes the results of a mathematical investigation of natural evolution dynamics. The leitmotif of this study is the capability of nonlinear equations of the Volterra-Lotka type to successfully model many evolutionary processes, which take place in biology, ecology, sociology, demography, economics and—among others—in technology. This idea may seem strange at first sight: what can be common between the mechanisms of extending stress or fatigue cracks in an automobile drive shaft and the growth dynamics of a tree or an animal, in the development of a business, etc? The explanation

[1] Do we live in a Volterra-World? (In German).

can be found in the original purpose of this type of equation: to describe the dynamics of conflict between two or more participants in a competition for survival. A simple classical example is that of interactions between two populations, rabbits and foxes [60]. An increasing number of rabbits augments the foxes' chances of survival and ability to multiply, but at the same time, the rising population of foxes decimates that of the rabbits, which in turn decreases the foxes' survival opportunity. Similar competitive forces can be found in other evolutionary processes. If this scenario is indeed valid, then the idea of a Volterra-World could be interpreted as the promotion of competition to a more prominent role as one of decisive Laws of Nature.

It is not difficult to see the role of 'randomness' in such processes:

- A fox can catch a rabbit only with a certain probability.
- A scratch on a rotating shaft's surface can give rise to a fatigue crack and can be viewed as a random event (although its existence has some entirely deterministic causes).
- There are positive as well as negative factors associated with the growth of a plant, including uncertainties in their occurrence.
- A nuclear reactor functions only if there is a balance in the competitive processes of the neutrons nascency (by fission of the fuel kernels) with their loss from non-productive capture and escape from the process. All of the individual reactions of neutrons with the kernels can be viewed as random processes.

This leads to the conclusion, that we really do live in a Volterra-World ruled by countless modifications of the Law of Competition with a substantial contribution by uncertainty.

The utilization of the expression 26.2, which was interpreted as a definition of the *metric* of the space of uncertainties and the application of the Gaussian model of a distribution function implies 'Gaussian metric' in a 'Gaussian world of uncertainty.' People prefer—if they are allowed to by the circumstances—simplicity. It would be pleasing to view a single model of the world, say a Volterra-Gaussian one. But this is impossible. The explosion of the Cernobyl nuclear reactor surely represented a specific modification of Volterrian conflict of competitive factors, but the particular dynamics differed from those of the pandemic spread of AIDS or of the rate of introduction of mobile telephones in developed countries. On the other hand, it was shown, that many different distribution functions can be inferred from data, that the 'normal' distribution has no monopoly. 'Normal' should not be accepted as a universal synonym for 'Gaussian.' To call something 'normal' is to say, that for a specific event or set of events, the behavior of the data *in those circumstances* is as would be expected. That is to say, they behave normally under the conditions of their observation. Thus, before any definition for normality can be ascribed, the investigator must make the necessary effort to determine the model of that specific set of 'normal conditions.'

The world we live in is too complex to be characterized by a single word. It can take many forms, sometimes it might be Volterrian and its uncertainty is sometimes Gaussian, but very frequently it is non-Gaussian. In further clarification, it can be added, that neither the Gaussian nor the non-Gaussian world of uncertainty is Euclidean. The probability density $g(x)$ in 26.2 would imply a Euclidean metric only if it were constant, independent of x which is rarely true.

26.3 Analysis of Historical Coins

The idea of revealing interesting historical and economical facts from the Middle Ages by advanced analytical methods can seem strange at first sight. Indeed, historical studies frequently base their conclusions on archeological findings. Among such findings, an important role is played by coins, because:

1. Historical coins were minted from precious metals. Therefore, they did not lose their value as time passed. They remained unlimitedly liquid and they could be used not only as a trading medium, but also for the accumulation of wealth/treasure.
2. Their physical/chemical composition allowed them to survive for centuries.
3. Coins describe historical facts and reflect the economic life of the time of their mintage.

The last statement is supported by the results of the investigation, which was carried out.

26.3.1 Grossi Pragenses

The name for these coins is, of course, Latin and it is derived from Prague Grosh[2]. Its use became popular in Europe because of the economic power

[2] From the Czech 'groš' and German Grosch.

Mintage		**Weight (g)**		**Purity (rel.unit)**	
Type	Time	Range (g)	Pieces	Range (1)	Pieces
I.	1346-1348	3.235-3.645	14	0.843-0.875	3
II.	1348-1355	3.204-3.490	3	0.875	1
III.	1350-1358	3.125-3.520	13	0.847-0.870	7
IV.	1358-1378	2.430-3.586	13	0.838-0.865	9
V.	1370-1378	2.480-3.787	24	0.749-0.767	7

Tab. 26.2: Number of Coins Analyzed for Weights and Purity and Ranges of the Values

of the Czech kingdom and its adoption as national currency by several countries. The coin, minted in Czech silver was embossed with the Latin name of the Czech king, Vienceslaus Secundus (Václav II.), on one side and with the Czech crown and double tailed lion on the other. The coin was introduced in 1300 during an economic reform instigated by King Václav II (1278-1305) and the coinage was called 'eternal and holy' probably because of its stability and universal liquidity stemming from its broad use in Europe. It is estimated, that annual production of silver from only the Kutná Hora[3] mines was around 20 thousand kilograms. The coin maintained its value primarily due to its silver content and mint marks, which guaranteed its authenticity. This contributed to its continued employment in international trade. Due to its silver content, it circulated for several centuries until the quality of the coins deteriorated. Collections of these coins, which exist today, mostly stem from archeological sources, and unfortunately, only a few of them have been quantitatively described so as to conduct a thorough analysis. It is something of a paradox, that numismatists are ready to spend large sums of money to complete their collection, but consider the costs of measuring and analyzing their coins too high. The objective of the study [86] was to show, that worthwhile historical information can be obtained by analyzing old coins.

26.3.2 Available Facts

The coins made available for analysis dated from the reign of the Czech King and Holy Roman Emperor Charles IV, (1346-1378). The designation of 'types' and subsequent identification of the issue dates had been numismatically determined previously by comparing minor differences in the images, that had been struck. Table 26.2 reports the results of the analysis.

[3] A Czech city listed and protected by UNESCO.

Mintage		Bounds Estimate		Median
Type	Time	*LB* (g)	*UB* (g)	*Med* (g)
I.	1346-1348	3.233	6.369	3.478
II.	1348-1355	1.852	3.556	3.375
III.	1350-1358	1.458	3.590	3.402
IV.	1358-1378	0.201	3.723	3.297
V.	1370-1378	2.475	12.73	2.901

Tab. 26.3: Data Support Bounds and Medians for Coin Weights

Mintage		Bounds Estimate		Median
Type	Time	*LB*	*UB*	*Med*
I.-II.	1346-1355	0.843	0.936	0.848
III.	1350-1358	0.844	0.870	0.865
IV.	1358-1378	0.837	0.865	0.863
V.	1370-1378	0.643	0.767	0.763

Tab. 26.4: Data Support Bounds and Medians for Purity

Robust estimates of lower and upper bounds of the data support (LB and UB) and median determined as quantile of probability 0.5 are reviewed in Tab. 26.3 for analysis of weights and in Tab. 26.4 for purity.

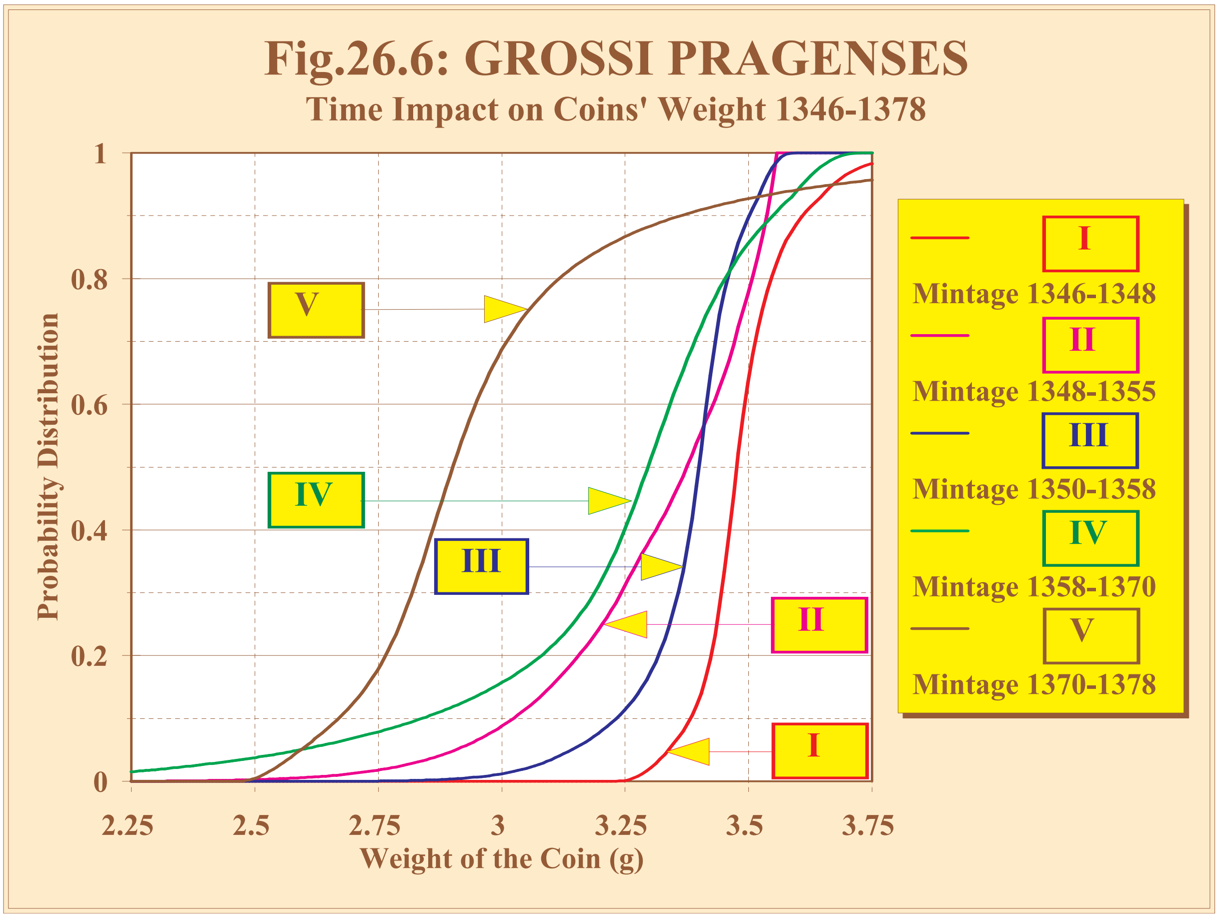
While modern purity measurement methods do not damage the coins, the traditional/historical metallurgical methodology for analysis was destructive, therefore the number of data, that could be used in the cited study was relatively small. This explains the preference of collectors for untested and undamaged specimens over those of known purity, but which have suffered under essay. Even so, modern analytical tests are felt to be expensive, given the results provided. However, the information provided in Table 26.4 could lead to a reconsideration of this point.

26.3.3 Distributions of Coins' Quality

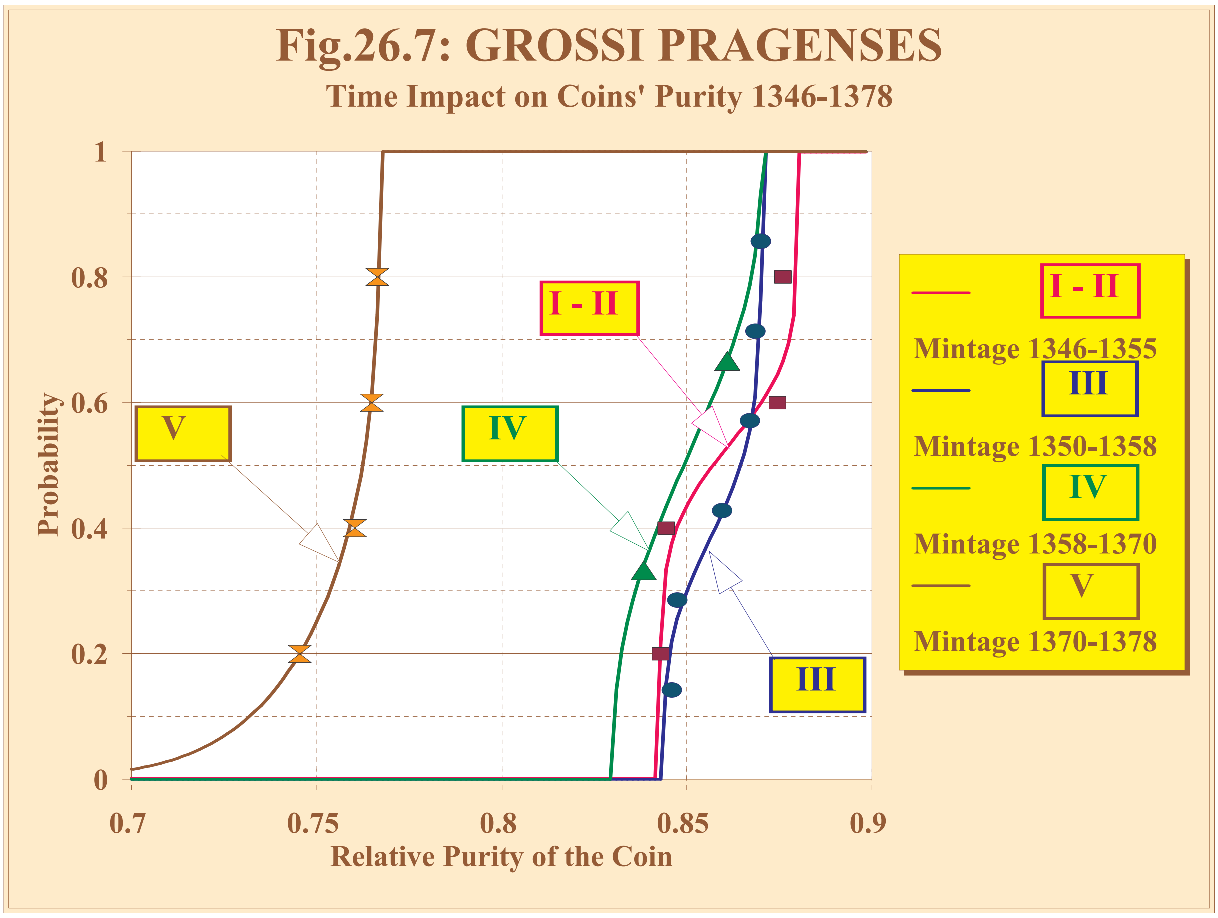
The manufacturing technology for coins then was much the same, but cruder, than what is common today. Purity was controlled by the amount of base metals included in the alloy, and a sheet was hammered out and rolled to a desired thickness by a mill. The coins were then stamped out using an engraved die. The master of the mint was responsible to the King for controlling these two parameters; this was necessary for wide acceptance and trust in the coinage of the realm.

The relative spread for purity in Tab. 26.4 is much smaller than that of the weights. This can be easily explained by the fact, that most of the coins were found in buried treasure and corrosion had a greater impact on the coins' weights than on their chemical composition. A second likely reason is, that the ability to maintain standard weights over several centuries would have been more difficult than adhering to a purity norm.

The results show a gradual but substantial degradation in the value of the money. This conclusion can be supported and a more detailed insight into the process can be obtained by considering the distribution functions EGDF of weights in Fig. 26.6 and of the purity in Fig. 26.7.



The form of the distribution functions of the weight in Fig. 26.6 and the values of the upper bound of data support (UB) allow several conclusions to be voiced as to the development of the quality of production. The distribution of the oldest coins (type I) gives evidence of a very wide tolerance in the thickness of the silver sheet, which results in nearly random behavior of the function, and in high probabilities of exceeding the standard with a very high value of UB (6.369 g). The curve for the type II shows a substantial improvement.



The standard (initial) weight was fixed and accurately maintained with UB=3.556 g, to which the distribution function sharply rises.

Recall, that the steeper the probability curve rises, the more certain the quantiles are, and the less the ‘local’ volatility. The curve for type III in Fig. 26.6 attracts attention from this point of view: it is the steepest one and—like the type II—it distinctly marks the upper break of the function close to the established standard (which can be estimated by the UB of 3.590 g). It is interesting to note, that according to historical records, there was a new and strict master of the mint appointed at that time. However, after 1358 the quality started to deteriorate with the UB of the type IV coins increasing to 3.373 g followed by a type V UB of 12.73 g: the ‘old good order of things’ returned signifying the decline of the society. This process can also be traced by values of the robust median, which was 3.402 g (type III) to fall to 3.297 g (type IV) and to 2.901 g (type V).

In the lower part of the distribution functions in Fig. 26.6 the spread of LB values is much larger than that of the UB. This is caused by deterioration through corrosion rather then by poor control over the weight of the coins.

Since the smallest values of the lower bounds belong to Types IV and V, this supports the comment on poor minting quality. This is further supported by Figure 26.7, which shows, that in the case of type V the purity was poor.

The probability distribution functions of the coins' purity shown in Fig. 26.7 support the conclusions inferred from the weights. The bounds set out in Tab. 26.4 for types I, II, III and IV document good control for the coins' purity. The change in quality over time can also be observed: The initial upper bound UB was the highest for the types I-II (0.936). However, then it was decided to be more economical in the use of silver and to decrease the upper limit to 0.870 for type III, while keeping its lower limit unchanged (LB=0.844), so as to maintain a solid foundation of precious metal. These results infer, that a decision was made to lower the silver content, yet retaining quality to ensure, that the value of the coinage was maintained. The form of the functions here is even more unusual than in Fig. 26.6: instead of an S-form, the very tight bounds on purity force them to rise steeply and cut off sharply on both-sides of the bounds. Care in the quality of production seemed to be approximately maintained for type IV at the upper bound, but the lower bound LB was allowed to fall to 0.837. After 1370, all pretense for coin quality was abandoned completely: the upper bound UB of purity was drastically decreased to 0.767 and the respect for the reputation of the money obviously declined judging by the fall of the lower bound to LB=0.643.

This brief look at the manufacture of coinage over a short period of history can provide some interesting insights in the development of society even though a significant period of time has passed, and the physical quantity of evidence is sparse.

26.4 Production Quality

26.4.1 Certification of Quality

The range of the impact of product quality on the market was vividly shown in Chapter 25: the greater the quality, the more goodwill accrued to the producer and the reputation of his brand as well as to the popularity of his product, market share and—last but not least—his ability to command higher prices. Certain products (eg medications, food, military supplies and others) have especially strict norms of quality and there are complex procedures to protect consumers. But the globalization of the marketplace has forced not only producers in various countries to pay attention to the quality of their goods, but also has given rise to the establishment of national quality standards, and governmental intervention into quality control is gradually expanding to all markets. Producers have little chance of exporting goods without an official certification. This process is controlled by special regulations and only designated institutions with skilled and experienced personnel are licensed to control the merchandise and issue certificates of quality. This process is expensive, consuming both time and investment and has become a special occupation. Unfortunately, the state of art does not really meet the needs of the industry nor exploit the level of available knowledge. The problem stems from both the inadequate traditional methodologies, that are used, as well as the inertia contributed by the ingrained bureaucracy that is in place to administer the system. This conclusion begs a more detailed justification. A firm, that desires to compete in the international marketplace, should establish procedures, which include the following elements:

1. The creation of a computerized information system based on its local network, which is capable of registering the basic production data:
   (a) the identification of the raw materials and the various component parts used in production,
   (b) the results of quality tests on the inputs,
   (c) a description of all production operations,
   (d) the results of interim control operations,
   (e) the results of final tests of quality of the products,
   (f) details concerning inventory storage, maintenance and shipment,
   (g) a list of personnel responsible for accepting inputs, for the manufacturing operation, and for the quality control tests.
2. The preparation of a detailed plan of the production process.
3. The elaboration of a detailed plan for the control operations including the prescribed tolerance limits and emergency thresholds.
4. The preparation of a plan to cover extraordinary and emergency situations.

The production quality subsystem is a part of the complete business information system used to share economic data with other company nodes. The real sense of the words illuminate the problem. A *computer* originally was a device meant for computing. Today, its functions are used for computing only a small part of the time. The original purpose of a computerized *information system* was to provide information to the user. In reality, the make up of information systems justifies only that they be called *data collecting and storage* rather than *information* systems. What they can deliver to the user is only data, but rarely information, because the latter can be obtained from the data only by computer analysis. Unfortunately it is a fact, that many business information systems have no on-line or even in-line analytical modules. If there are analytical tasks to be performed, it is the generally accepted the point of view of the designers of these systems, that the user should employ a professional statistical analytic package for that purpose. Such an approach simplifies the design of the information systems and releases the designers of the need to solve complex data treatment tasks. But this approach abdicates responsibility and is akin 'throwing the baby out with the bath water.'

Indeed, the real purpose of a business information system is to provide the means to direct the business. The most efficient control is that of a feedback system: the results of control actions feeding the input are compared on the output side with the required quantities and the deviations are returned to the input after the optimization of a corrective action. The inclusion of an off-line analytical system, that requires personal intervention, complicates the communication effort and introduces undesirable time delays. A closed-loop system should be provided with an analytic module using real-time automated software. This requirement relates to both the quality and business control functions of the system.

A change in mentality is required: sophisticated information systems are installed and maintained to serve as databases without satisfying their main purpose—the provision of information. Emphasis on administrative and bureaucratic functions as an end, such that the certificate itself rather than the quality of the product is the desired objective, subverts the purpose of the system.

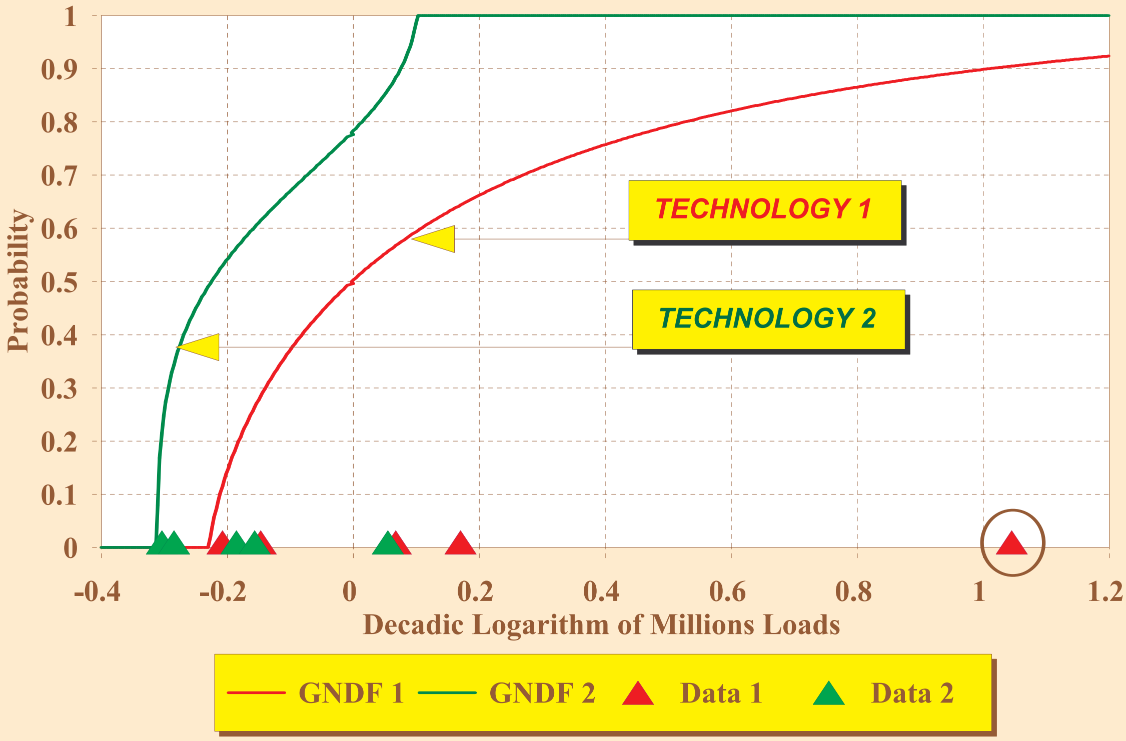
However, the mere installation of an in-line analytical module is insufficient without the inclusion of advanced software. Programs incorporating the usual prejudices toward 'normality,' independence, stationarity, homogeneity, homoscedascity, Euclidean measure of errors, etc. are not capable of providing the information needed to reliably govern a real process.

26.4.2 Comparison of Technologies

The following is an illustration of the ability of gnostics to compare production technologies. It took place in the Czech firm TATRA, which produces unique air-cooled trucks made famous by gaining multiple wins in the difficult Paris-Dakar race. These vehicles are designed to serve under the most difficult off-road conditions. Their reliability depends—among other things—on the life of the suspension springs. The main problem is prevention of fatigue cracks and this depends on the characteristics of the material and on the technology used to produce the springs. Testing for durability is expensive and time consuming, because it requires hundreds of thousands of cycles of loading and unloading operations. In addition, the test is destructive, therefore the number of completed tests (which ordinarily end by breaking the spring) is small. The results of tests of two technologies (5 cases for each method) are shown in Fig. 26.8. The unit of measurement used was millions of loading cycles until the spring was broken.

Fig.26.8: COMPARISON OF TECHNOLOGIES

Two Technologies of a Truck's Springs



The nature of the data is obviously multiplicative, therefore the use of a logarithmic horizontal scale, on which the common logarithms of millions loads are shown. Taking the customary approach with point statistics and assuming a lognormal distribution of data, the mean logarithm of four tests for Technology 1 mu_1, is -0.0289 and its standard deviation, σ_1, is 0.154. The critical value $\mu_1 + 3 * \sigma_1$ is thus 0.491. The fifth outcome, the largest value (circled) 1.046 therefore far exceeds the 'critical bound' and should be discarded as a sure outlier. The analogous test for technology 2's data gives $\mu_2 + 3 * \sigma_2 = 0.514$. The largest value is 0.055 and it is not for being an outlier. Student's test is unable to refute the hypothesis, that there is no difference between the two technologies.

Shifting the approach so as to use distribution functions (EGDF) rather than point estimates leads to an opposite conclusion. The seeming outlier 1.046 (ie this spring survived 10^1.046 millions loads!) becomes important, because it forces the distribution function to the concave form even for high values of the spring's life. This value identifies the best spring by far, the one which really survived the record number of loads. It would be illogical to discard this value. On the other hand, the 'strange' form of the (green) distribution function of technology 2 is quite natural, when it is examined from the point of view of the mechanism of fatigue scratches: during the initial interval, the repeated changes of load cause no measurable effects, but the material 'registers' them, accumulating both the number and the intensity of the stresses. But once a certain limit is exceeded, a crack starts from a disturbance on the material's surface and gradually increases with the increasing number of cyclic loads. The resistance of the material increases slowing down the growth of the incipient crack until the moment, when the growth rate increases. From this point, the curve changes its form from concave to convex and the crack extends with increasing speed until the break occurs. A comparison of the distribution functions (DF1 and DF2) allows the following conclusions to be drawn:

1. The crack in DF1 begins, when DF2 reaches a value of about 0.5, ie when about half of the springs of technology 2 are broken.
2. The change of DF2's form from concave to convex occurs, when DF1 reaches 0.5, ie when about half of the springs of technology 1 survive (when 80% of T2 are expected to have failed.)
3. DF2 reaches a value of 1, when the DF1's value is about 0.6, ie when all springs of technology 2 are broken and about 40% of the springs of technology 1 still survive.
4. The behavior of DF1 close to the right end of the graph is nearly

linear, ie the change from concave to convex character of DF1 can be expected only at a still longer lifetime.

Using the distribution functions allows the superiority of technology 1 to be proved.

26.4.3 The Quality of Caprolactam

Caprolactam is an important chemical product used as a raw material in the textile industry. Some years ago, the quality control department of a producer of caprolactam initiated a study to identify the factors of production that have a greater influence on the quality of this product. The behavioral relationships between the variables traditionally used to measure quality and the main quality indicator, *light absorbance,* were not known with sufficient confidence to justify the setting of control thresholds for important variables. Caprolactam is produced in batches and measurements of nine quality indicators from 39 batches were provided as inputs to the study. The variables are listed in Tab. 26.5.

Only variables 1 through 5 are continuous; the rest are dichotomous: only two levels can be reliably distinguished: 'high' or 'low.' Limitations as to permissible values for absorbance, AB, are set by the end user, and the permanganate number, PN, plays the next most important role in the manufacturing process. The analysis is intended to explore the impact of these individual indicators on the absorbance.

Preliminary information can be obtained by examining the probability distribution of absorbance over the set of all batches. The EGDF of the 39 values of absorbance results in the probability and density functions shown in Fig. 26.9.

No.	Symbol	Name	Data type
1	AB	Absorbance	Continuous
2	PN	Permanganate number	Continuous
3	VA	Volatile alkalis	Continuous
4	CH	Color by Hazen	Continuous
5	AL	Alkalescence	Continuous
6	MC	Moisture content	Continuous
7	SP	Solidification point	Dichotomous
8	AR	Annealing rest	Dichotomous
9	MI	Mechanical impurities	Dichotomous

Tab. 26.5 Quality parameters for caprolactam

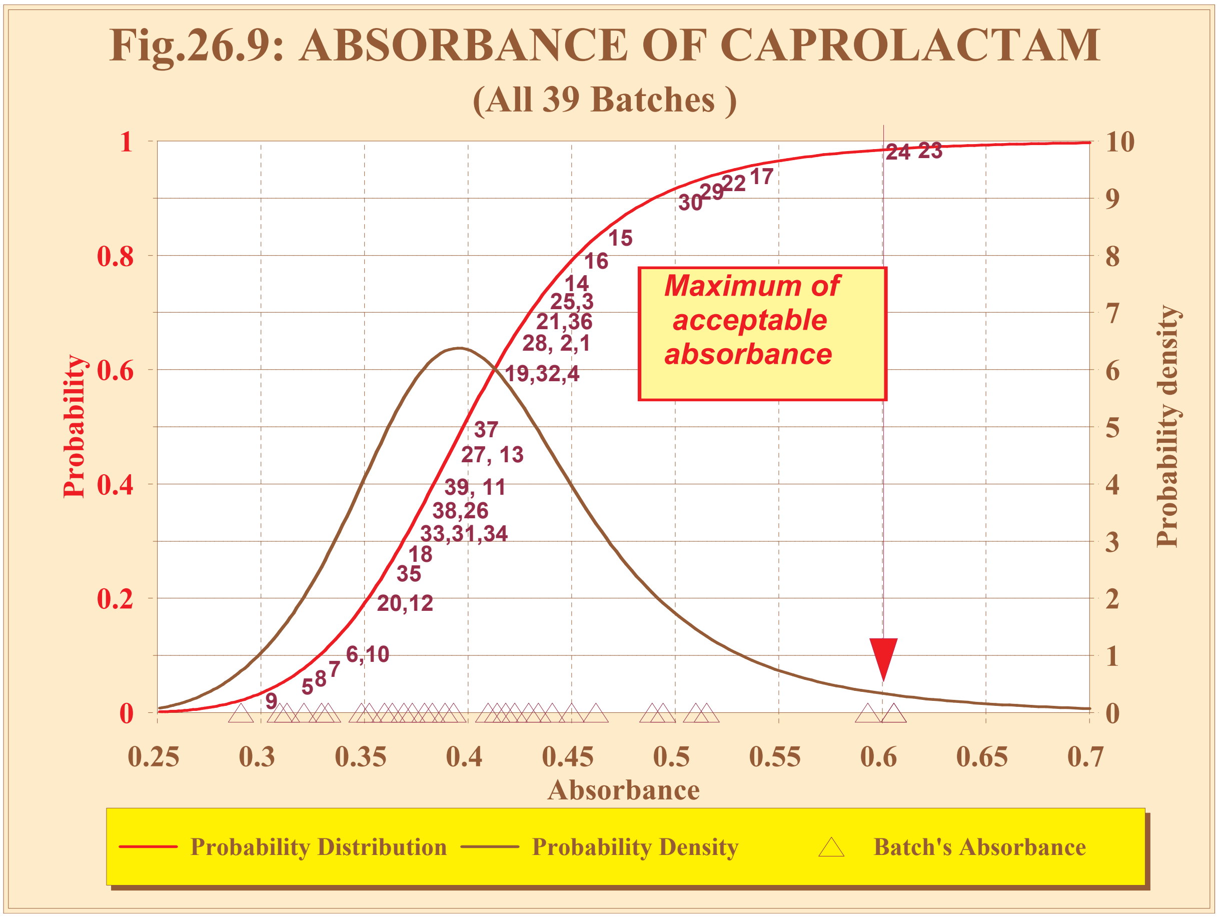


Fig.26.9: ABSORBANCE OF CAPROLACTAM

(All 39 Batches)

Maximum of acceptable absorbance

Data are depicted as triangles on the X-axis. Labels along the probability distribution identify each individual batch. The estimated bounds of the data support are very broad ($LB = 0.191$ and $UB = 17.07$), which along with the bell-like form of the density curve might suggest normality or logarithmic normality. This would mean, that values exceeding the maximum acceptable absorbance level (0.6) are possible and can be expected. Indeed, two batches (No.23 and 24) reach the limit and should not be marketed. The requirement, that probability at the desired limiting point of 0.6 be equal to 1 implies, that neither normal nor lognormal probability distributions can be used, because such distributions would reach 1 only at infinity.

The above shows, that the determination of what causes high absorbance is really needed. The first hypothesis to be tested is, that the absorbance AB can be explained by the permanganate number PN. However, the linear regression $AB = C_0 + C_1 * PN$ leads to a coefficient of determination of $R^2 = 0.177$, ie this model can explain only 17.7% of absorbance's variance and should thus be rejected. A better outcome may be obtained by extending the number of explanatory variables.

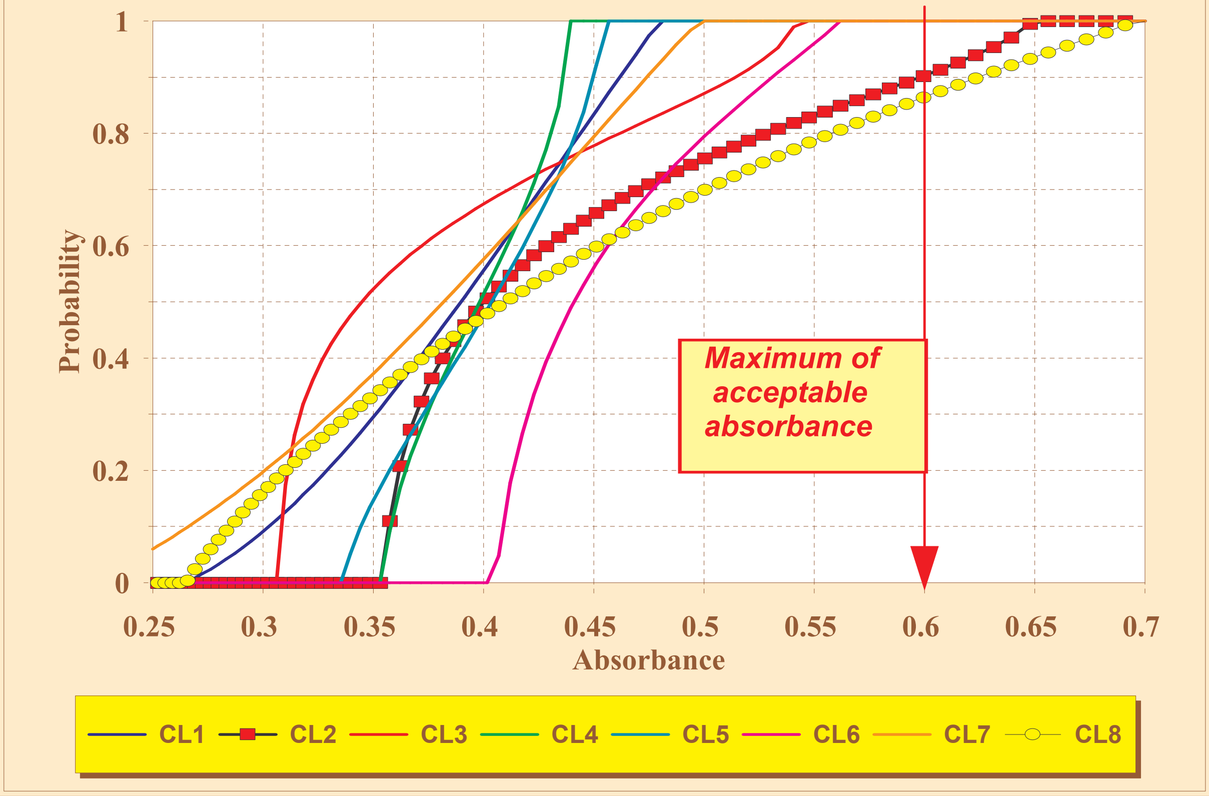
The four other continuous variables can be used as explanatory variables in a multidimensional regression model of the type

$$AB_k = C_0 + C_1 * PN_k + C_2 * VA_k + C_3 * CH_k + C_4 * AL_k + C_5 * MC_k, \quad (26.4)$$

where k=1,...,39.

Unfortunately, this model does not do a much better job of explaining a greater portion of the variance, because its R^2 is only 0.494. The distribution function of residuals of this model reveals substantial inhomogeneities in the data. Moreover, variables AL and MC can be left out, because they have only a negligible effect on the solution.

Fig.26.10: CLUSTERS OF CAPROLACTAM



The next step is therefore to institute a robust multidimensional cluster analysis (see section 21.7), which decomposes the set of batches into 7 clusters and an unclassified remainder of 4 batches (which is insufficient to build a 5-dimensional model). Each cluster contains 5 batches. There are 4 coefficient to be determined, so the models have one degree of freedom. All seven models are acceptable from the point of view of statistics, because their R^2 are all greater than 0.985 and the standard fitting errors of the models are all less than 0.0166. The decomposition is thus successful.

An insight into the clustering can be obtained from Fig. 26.10, where probability distributions of the absorbance for individual clusters are shown (the left over four batches are included as CL8).

The distributions of six 'good' clusters reach 1 for absorbance well before 0.6, while the probability of exceeding this value is high for CL2 and even larger for CL8.

After a robust solution of the system 26.4 has been obtained, a review of the mean impacts of the explanatory variables on the absorbance can be calculated (eg for the variable PN as $C_1 * \overline{PN}$, where $\overline{PN}$ is again the arithmetic mean). The mean impacts within the clusters are depicted in Fig. 26.11.

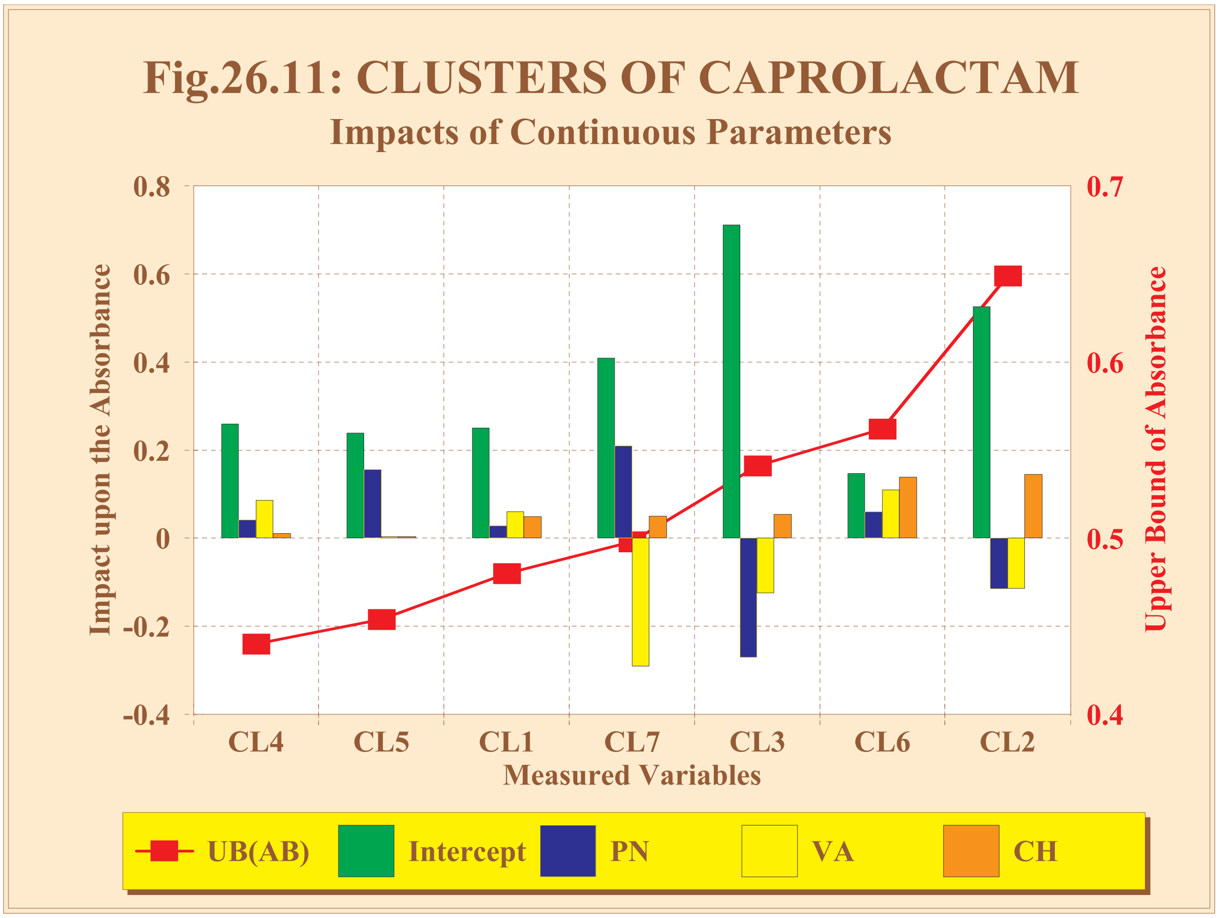


Fig.26.11: CLUSTERS OF CAPROLACTAM

Impacts of Continuous Parameters

The upper bound of the data support (of the absorbance) $UB(AB)$ determined for each of clusters is shown by the red line. There really are six clusters (CL4, CL5, CL1, CL7, CL3 and CL6) of ‘passing’ batches, for which the absorbance limit is not reached. Their $UB(AB)$ is 0.440, 0.454, 0.480, 0.498, 0.541 and 0.562, respectively. However, in the case of the cluster CL2 the probability of exceeding the acceptable level 0.6 is surely less than 1, because its $UB(AB)$ is 0.696. The patterns of the impacts differ widely and only one of the continuous variables (CH) seems to be well correlated with absorbance and can be identified as showing a contribution to higher absorbance values. The patterns of the cluster's composition shown in Fig. 26.11 show, that the three parameters PN, VA and CH are interdependent; their effects on the absorbance in some combinations can compensate for, as well as amplify each other. The large impacts of the intercept suggest, that there are still missing variables, which indicates, that the impact of the dichotomous variables should be examined.

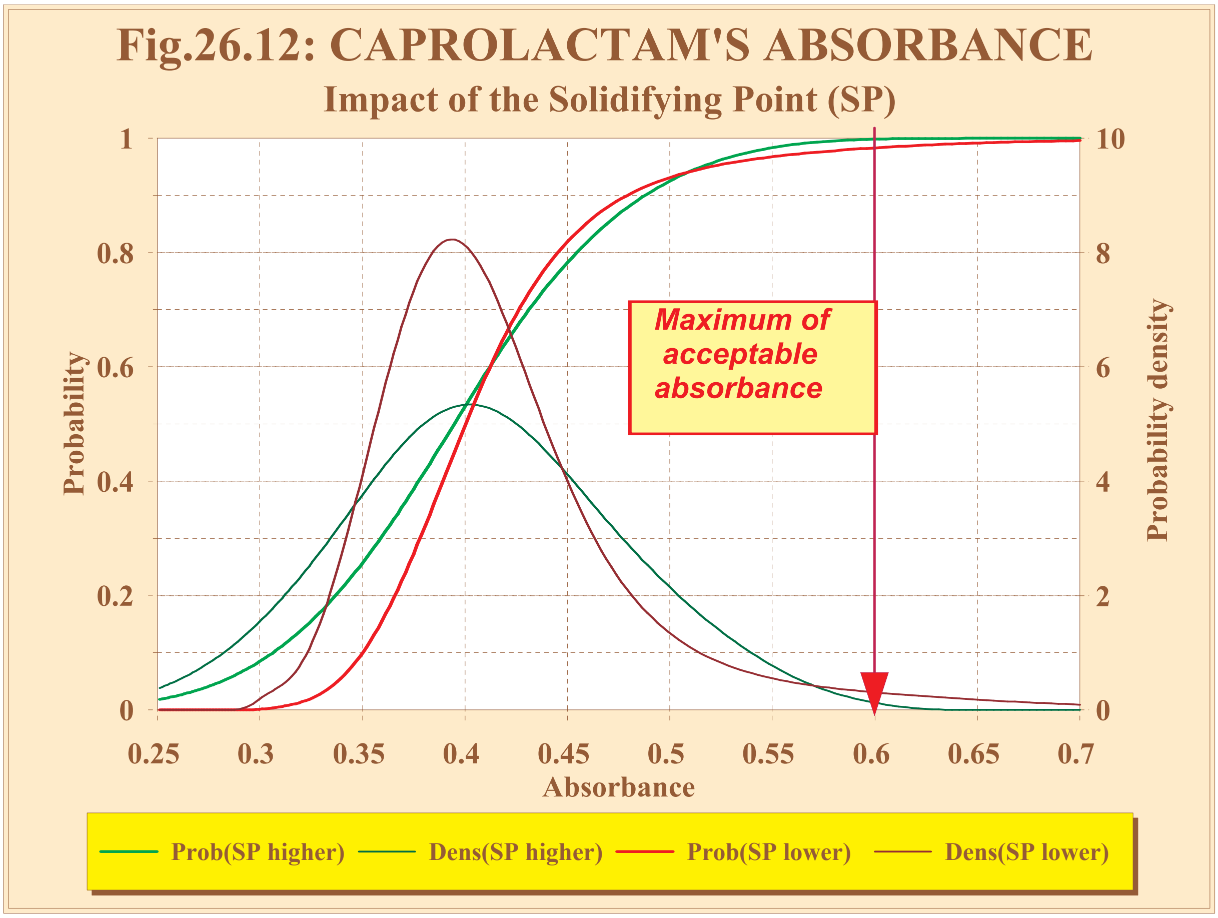
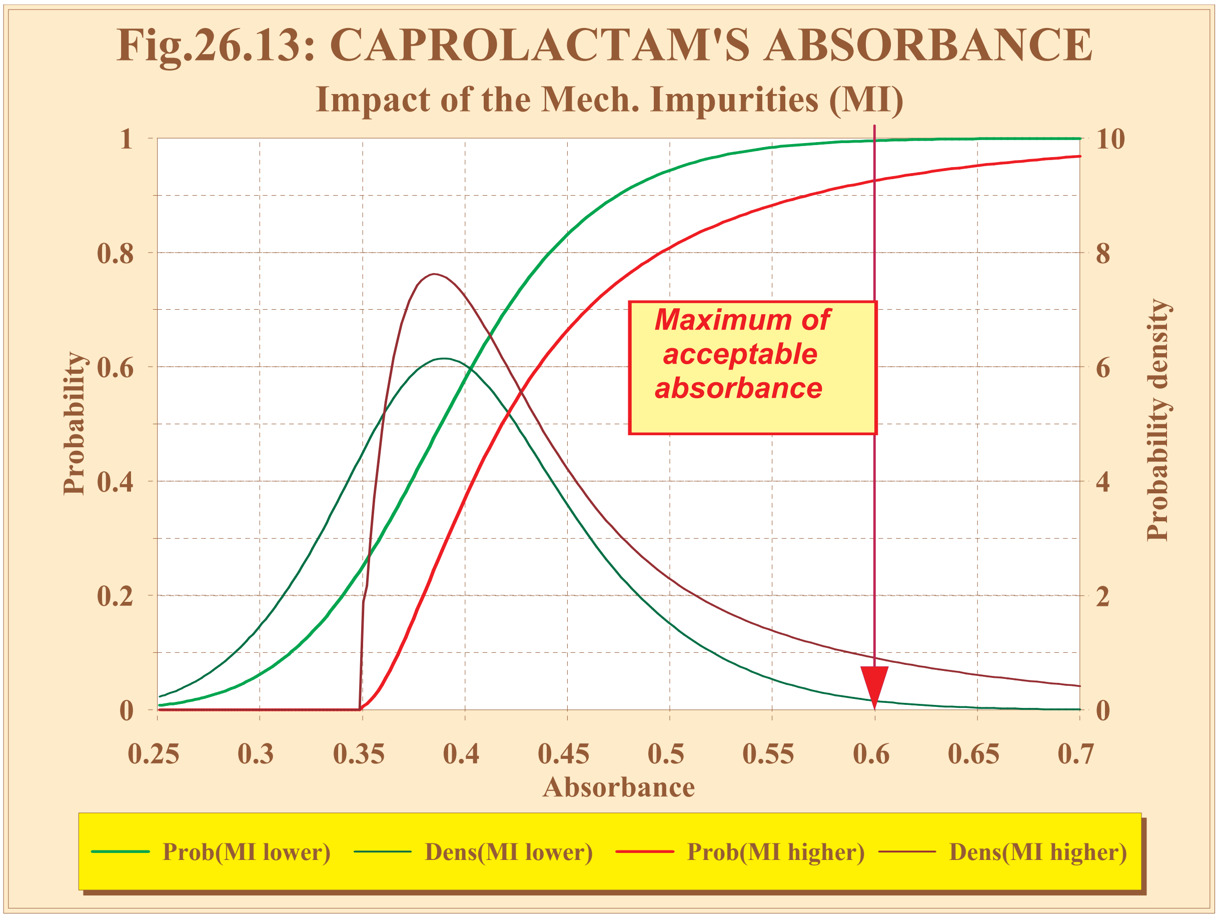


Fig.26.12: CAPROLACTAM'S ABSORBANCE

Impact of the Solidifying Point (SP)

Maximum of acceptable absorbance

A firm interested in streamlining its operations and improving its overall quality control should look for commonalities within batches in its production cycle. Such items as common conditions existing ‘within’ clusters and for differences ‘between’ clusters can serve as important performance indicators. Factors such as the composition of crews (especially shift supervisors), the time factor (day/night), calendar effect (Monday/Wednesday/Friday[4]), etc, are likely to provide clues to differences, which are found over time.



However, there are still unused data in the considered case: the three binary-valued factors SP, MI and AR. Their impact on the absorbance can be easily analyzed using the natural advantage of the non-parametric estimation technique based on an estimation of the *conditional probability distribution functions*. Indeed, to obtain the distribution function $Pr\{AB \leq x|(SP = low)\}$ it is sufficient to separate the batches for, which the condition $SP = low$ holds and to find the distribution function of these batches. Analogously, the alternative $SP = high$ can be explored. The

[4] Recall the Hailey’s novel, Wheels.

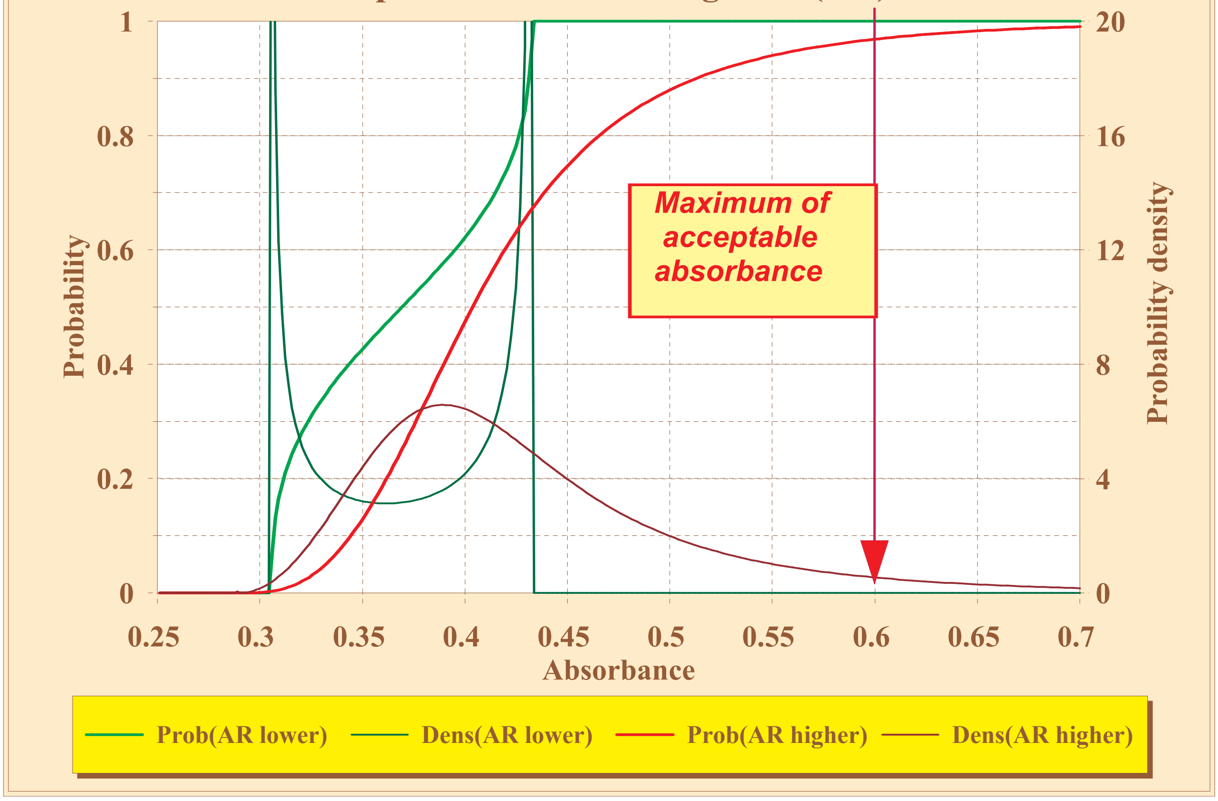
resulting conditioned distributions for both alternatives of the solidifying point SP are shown in Fig. 26.12.

The difference between the alternatives is especially evident in the density functions: although the case $SP = high$ has a broader density function with a lower maximum (ie the volatility is higher), the density's integral (probability) is higher at the point 0.6. This means, that a higher SP is better for quality. This example demonstrates the power of the analysis based on the distribution functions: reliance on the point estimate of the variance would lead to an opposite conclusion ('the higher variance, the higher risk of exceeding the limit'). The result is, that increasing SP has a small but positive effect on the quality.

The conditioned distributions for the factor MI are in Fig. 26.13.

Fig.26.14: CAPROLACTAM'S ABSORBANCE

Impact of the Annealing Rest (AR)



The effect on the quality is much stronger here than what was seen for SP and the conclusion is opposite: to get a low absorbance, keep the mechanical impurities as low as possible. The effect of AR shown in

Fig. 26.14 is really striking: the difference between alternatives $AR = low$ and $AR = high$ is not only quantitative, but qualitative.

Instead of a bell-form, the density of absorbance for batches with a low AR has the U-form. The probability of exceeding a value 0.43 is zero and reaching the limit 0.6 is not expected. The annealing rest has been shown to be the strongest factor in the determination of caprolactam's quality.

The analysis suggests the following conclusions:

1. The strongest impact on the absorbance of caprolactam is the annealing rest AR. To ensure a high caprolactam quality the quantity AR must be kept at a low level.
2. The second strongest factor is the amount of mechanical impurities MI. A high quality requires MI be low.
3. The continuous variables PN, VA and CH taken individually have no remarkable effect on the quality, but they are interdependent and affect the quality in certain combinations.
4. Variables AL (alkalescence) and MI (moisture content) have no significant effect and can be left out of the control procedures.

26.4.4 The Cleanroom Problem

For some manufacturing processes, the quality of the product depends on many factors, among which an important role is played by the environmental conditions that prevail, where the work is performed. Many procedures require extreme cleanliness of the air in the workplace. This is necessary not only in operating rooms, food processing plants or pharmaceutical production, but also in the high-tech industry. Experience has shown, that the reliability of microchips depends on the concentration of airborne microcontamination particles in the rooms, where uncovered chips are manipulated. Such a room is called a *cleanroom*. These special needs are incorporated in the building's plans and are similar to those used, when working with radioactive or biological materials, but with opposite effect: instead of preventing radioactivity or biohazards from leaking outside, the cleanroom is protected against the infusion of contaminated air from the outside. The requirements for perfect filtration of the air before it enters the cleanroom include anterooms for personnel to change their clothes, and cleaning all the materials and devices needed in the process. There are also requirements for a precise measuring and information system. The needs of such a facility are set out in [18] as follows:

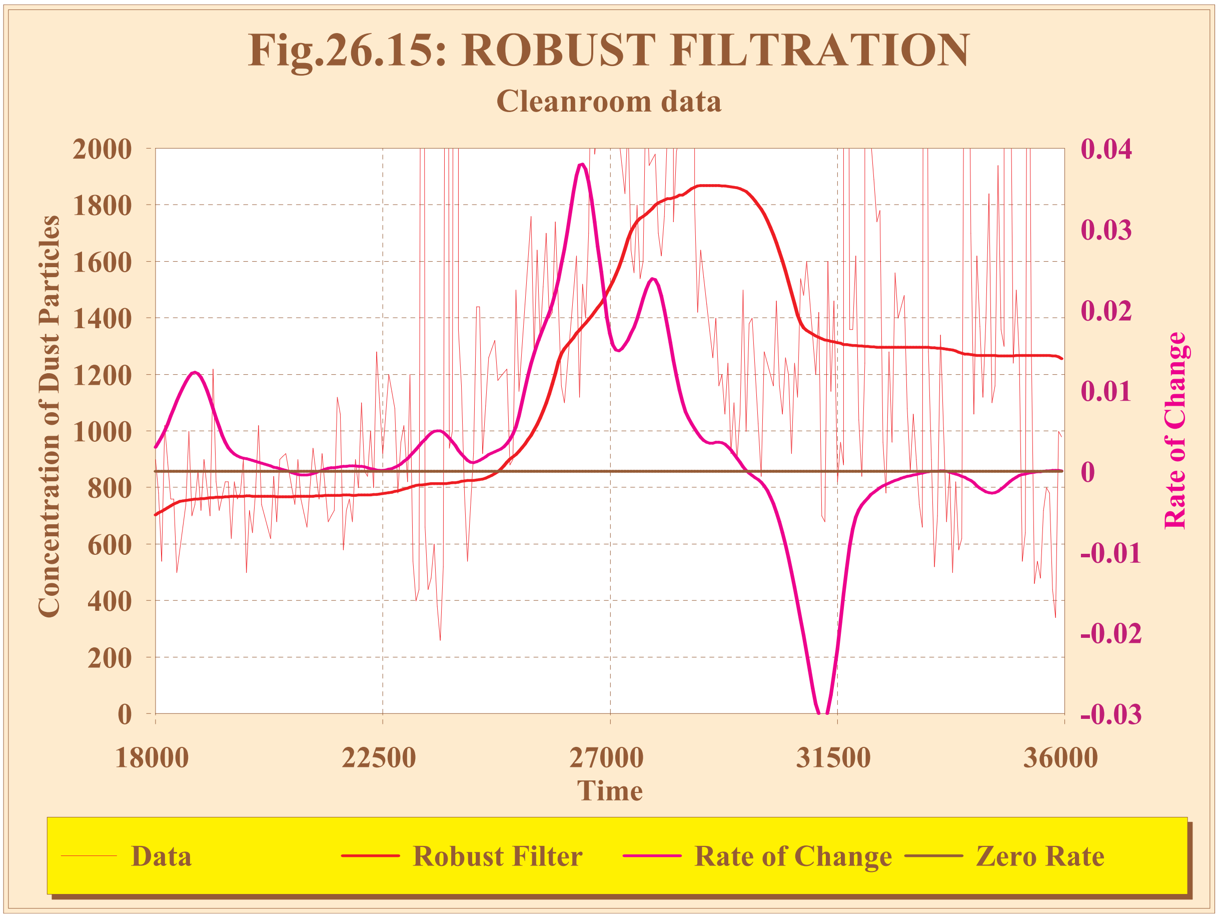
T1 to confirm cleanroom class conformity,
T2 to provide timely information about particle movements,
T3 to provide diagnostic information about contamination, whether it be personnel, equipment, or facility-generated,
T4 to quickly identify particle count excursions and air handling failures before they have an impact on the yield,
T5 to identify locations and operations contributing to high concentrations of particles,
T6 to confirm proper operation of the cleanroom and equipment appropriate for semiconductor research and development,
T7 to establish a database for correlation of particle counts with process yield,
T8 to identify particle monitoring strategies for future fabrications.

However, an important function for each automatic monitoring system does not appear in [18]: to provide autodiagnostic information on the proper functioning of the monitoring system and on its readiness to fulfill all its tasks.

The very complex behavior of the airborne microcontamination particles poses difficulties in the design of the system: the more or less stationary 'noise' generated by many small particles is disturbed by sudden very strong excursions probably caused by a clump of particles clustered together. These bursts occur randomly and then cease without affecting the background level. The difficult issue to be confronted is to decide, whether these are a very short term temporary disturbances or the start of a dangerous rise in the particles' density. The article [18] describes attempts to solve the problem by an approach based on fractal methodology. In parallel with this effort, a research project was run[5] based on gnostic methodology. A decision was made to solve the main tasks **T1** through **T6** by using a fast, but robust gnostic recursive filter running in-line. An example of its activity is reproduced in Fig. 26.15.

The thin red line shows the unfiltered output of a particle counter, while the thick red line is the output of the gnostic filter. Note, that the filter smoothes even large temporal excursions, but is sensitive to and reacts in a timely manner to changes in the actual level of the process. This provides reliable information sufficient to satisfy tasks **T1** and **T3**. A solution to tasks **5** and **6** can also be based on such filters to read signals from local detectors (a cleanroom is fitted with a large number of particle detectors

[5] Sponsored by the DEC (Digital Equipment Corporation), Vienna



both of the laser type and suitable versions of Wilson chambers).

The magenta line in Fig. 26.15 represents the rate of change of the robust filter (its first derivative) obtained by application of the differentiating linear operator (see sections 18.4.1 and 18.4.4). This output is both robust and sensitive enough to satisfy tasks **T2** and **T4**. The filtered signal together with its derivative stored in a database are suitable to fulfill tasks **T7** and **T8**. Should their signals fall below a threshold, the autodiagnostic signal indicating a failure of the monitor would be triggered.

26.5 Application to Psychology

26.5.1 Job Stress Survey

The significant impact of stress in the workplace on employee health, well-being, and effectiveness has been increasingly recognized in the last several decades. Stress and strain in work settings are generally attributed to the interactions between an individual and that person's occupational environment. Work stress results primarily from an incompatible person-environment fit that produces psychological strain and stress-related physical disorders. The methodology of the Job Stress Survey (JSS) as set out in [39] and [40] was applied in the Czech Republic. The authors of the Czech version of the JSS were Helena Knotková, František Man and Charles D.Spielberger. It was decided to analyze the results not only by classical statistical methods as in the original JSS, but also—at least in part—by gnostic methods.

The JSS was administered to 209 university and corporate employees. Differences in the perceived severity, frequency of occurrence and overall level of occupational stress were evaluated for individuals working in these settings. The format for responding to the JSS Severity Scale was the same as in American JSS. Subjects first rated, on a 9-point scale, the relative amount (severity) of stress they perceive to be associated with each of the 30 JSS job stressors (eg 'Excessive paperwork', 'Working overtime') as compared to a standard stressor event, 'Assignment of disagreeable duties', which was assigned a value of '5'.

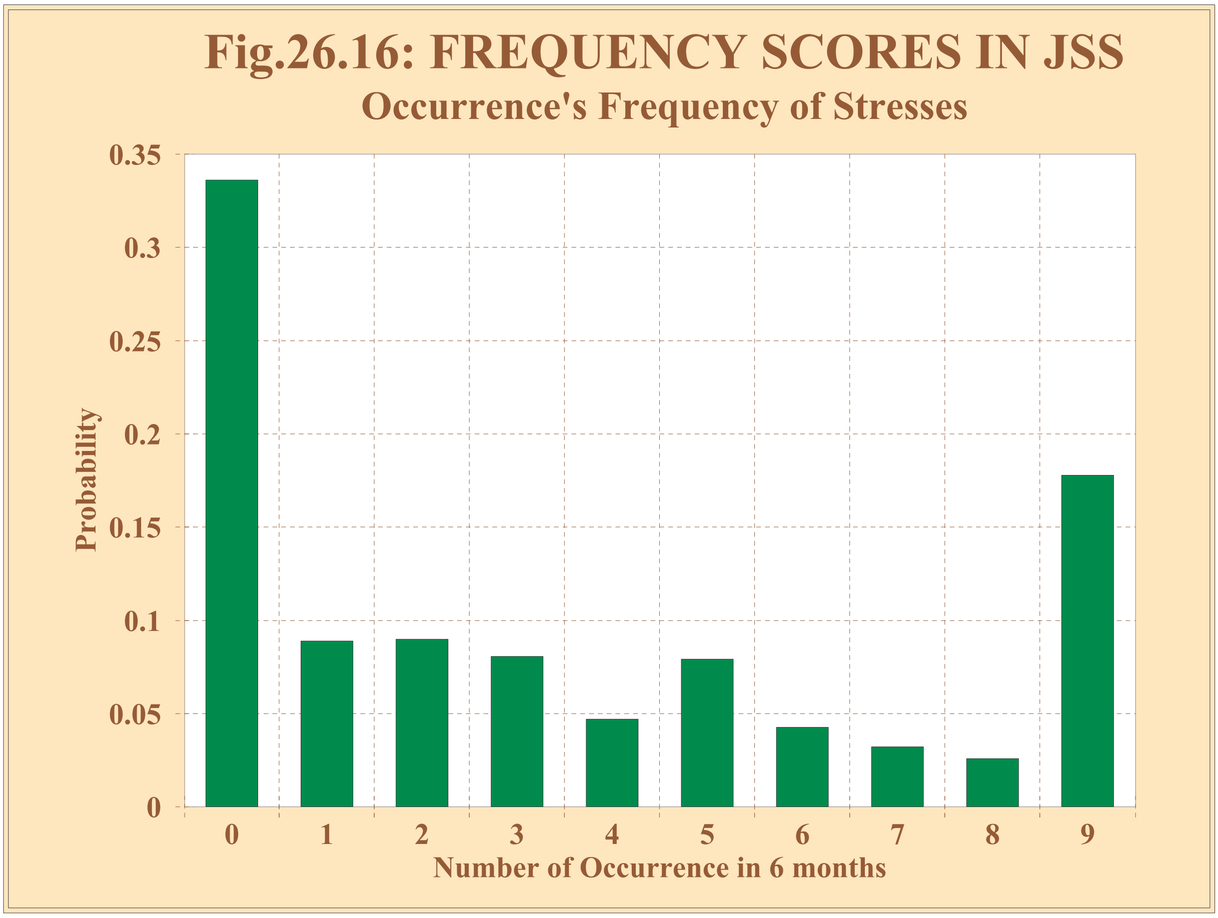
The JSS took into account the assessment of anxiety by asking respondents to indicate how frequently a stressor event had been experienced during the past 6 months using a scale from 0 through 9+ (nine or more) days. Scores of severity JSS-S and of frequency JSS-F were then obtained as the sum of all 30 ratings of the individual JSS items. The overall Job Stress Index (JSS-X) was calculated as the sum of cross-products of the severity and frequency scores.

26.5.2 Initial Analysis

The application of classical statistical methods (the testing of significance by standard deviations from the arithmetic mean, using analysis of variance, factor analysis and F-tests) raises certain questions. Are there really good reasons to presume, that the uncertainty in human decision making (in filling out a questionnaire about their stresses) is Gaussian? Recall, that the idea of a normal distribution consists of a superposition of many minor deviations from the (constant) mean with only rare outliers. Is this a true characteristic of human psychology? How do such things as prejudices, conservatism, momentary depression, and different triggering mechanisms fit in this scheme? A quotation from Humberto Eco's characterization of society in Paris during the Middle Ages ([21]) seems to support the idea of a ‘nonlinearity’:

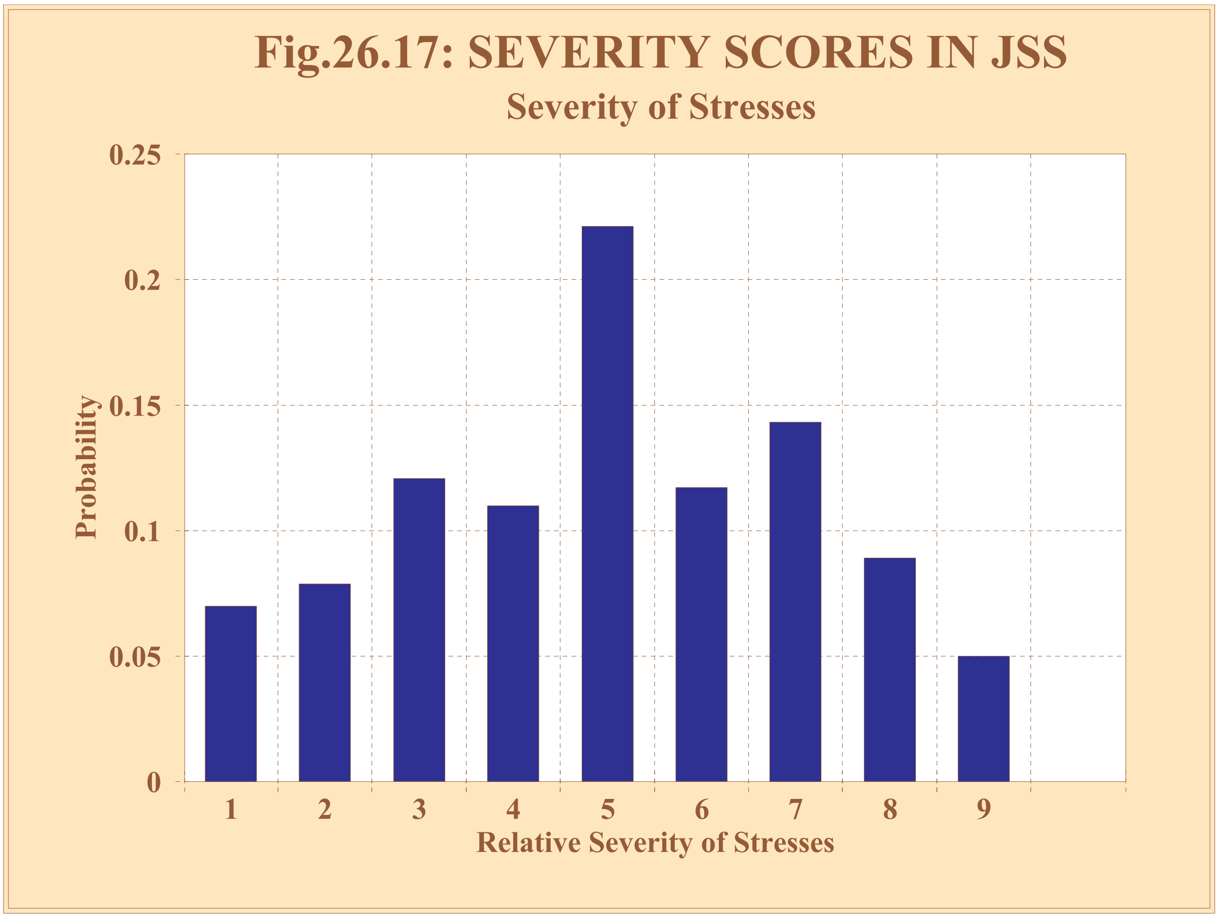
> *. . . as if the usual things, even if miraculous, no longer enlightened anyone, and only the unusually uncertain or the certainly unusual were able still to stimulate.*

Consider Fig. 26.16, which shows the frequency scores of stresses.



About a half of the respondents took extreme positions of 0 (‘never’/’unusually uncertain’) or 9 (’too much’/’certainly unusual’), while the other half were confident enough to quantify the frequency. If a more or less similar (‘average’) working environment in a society could be assumed, then a simple interpretation of the results can be proposed: about a third of the respondents had a ‘robust’ character insensitive to disturbance factors in the workplace, while about a sixth were oversensitive. If this is true, then their judgments can be taken as either prejudiced toward one extreme or the other, or at least, noncommittal.

Unlike these nearly dichotomous responses, Fig. 26.17, which reviews the extent of the stresses implies the existence of three clusters of responses, which could be expected if there were listed only three instead of nine stress levels: rather low, average and rather high.



In any case, neither the frequency nor the severity results can support the idea, that the character of the process is Gausian.

The questions which made up the questionnaire are not independent. Some are related to the same subject and are included intentionally to control the veracity of the respondent. The answers are therefore expected to be interdependent. Additional interdependence in the responses can exist due to specific personal character traits of the respondents.

An idea of the significance of individual stressors can be obtained by considering the overall Job Stress Index (JSS-X, the sum of the cross-products of the severity and frequency scores). The probability distribution and density functions of the reported values of Job Stress Index are in Fig. 26.18.

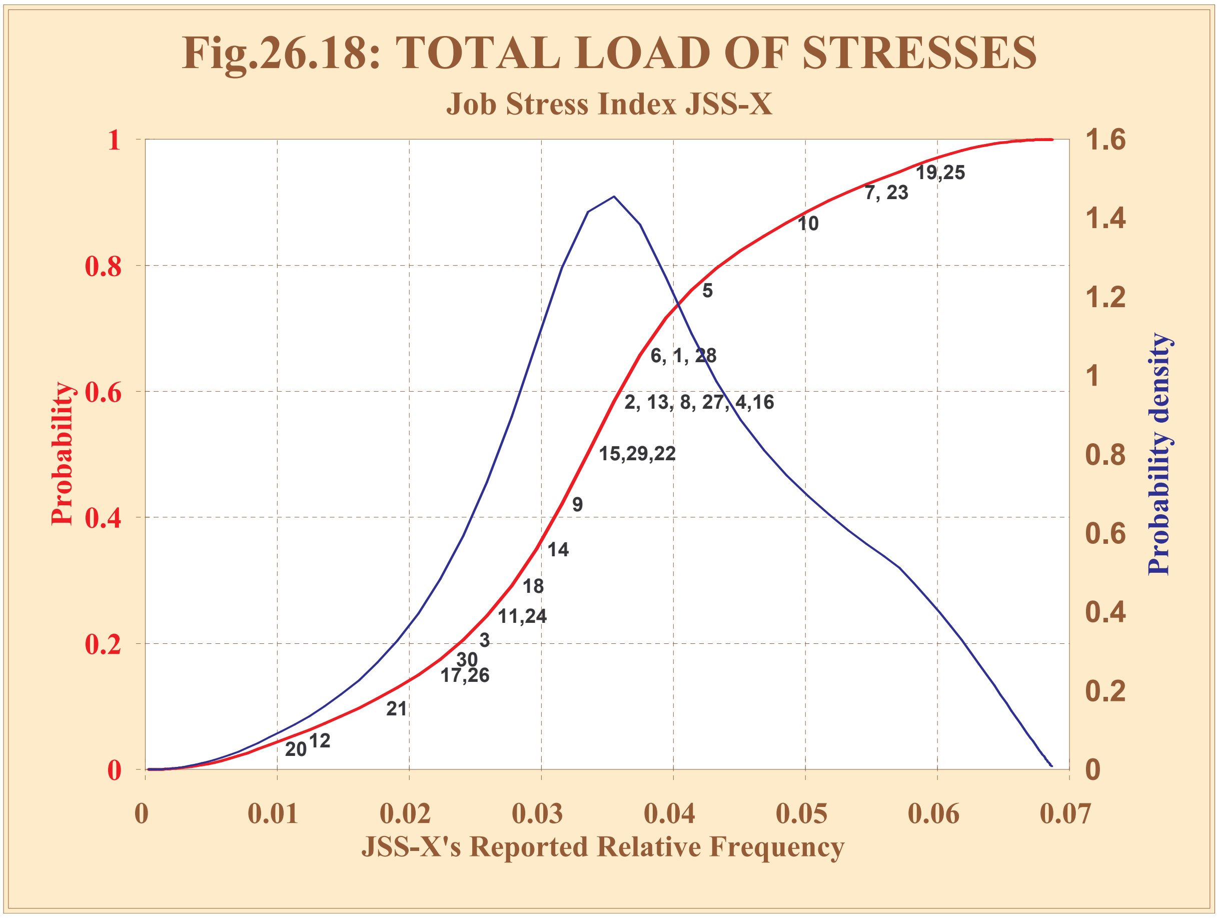


Fig.26.18: TOTAL LOAD OF STRESSES

Contrary to the normal and log-normal distributions, the data support for this distribution is bounded (the maximum Index's value is 0.0689). The identification numbers of the stressors are shown as labels on the probability function. The most significant stressors (those with probability exceeding 0.75) are 5, 10, 7, 23, 19 and 25, and the least significant ones (with probability below 0.25) are 20, 12, 21, 17, 26, 30 and 3. Table 26.6 identifies the nature of these stresses, both those bothering people the most (high probability) and the least bothersome ones (low probability). The relative frequency reported by the respondents (in per cent) is also shown in the table.

An overview of Table 26.6 suggests, that the majority of the respondents want to work, and that the most significant stresses are related to obstructions to their effort to work. Indeed, the strongest stressors 25, 23, 7, 10 a 5 transmit the same message: 'Let us do our work without meaningless interruptions!' On the other hand, personal interests are evidently dampened (stressors 20, 17, 30, 3). However, there is an exception,

The least bothersome situations		
Stressor		**Reported Occurrence**
#	**Type**	
20	Competition for careers	0.7%
12	Periods of inactivity	1.1%
21	Weak or inadequate control by the boss	1.6%
17	Personal offence	2.1%
26	Terms of meetings	2.1%
30	Conflicts with other departments	2.2%
3	Lack of opportunities for a career	2.4%
The most bothersome situations		
Stressor		**Reported Occurrence**
#	**Type**	
5	Co-workers not doing their jobs	4.1%
10	Inadequate or bad equipment	4.8%
7	Solving of crises	5.2%
23	Frequent disturbances	5.4%
19	Inadequate salary	5.5%
25	Excessive paperwork	5.6%

Tab.26.6: The most and least bothersome stressors evaluated by the Job Stress Index

a feeling that they are being inadequately paid for their efforts (19). It is interesting, that this stressor did not have as prominent a position in the original American survey. This difference can possibly be ascribed to the specifics of the postcommunist development of the Czech Republic.

The results of this analysis provides information on the point of view of the majority of the respondents and does not allow any inference about the homogeneity of the results nor about their structure. These aspects must be considered under a multidimensional cluster analysis.

26.5.3 Clusters in Job Stress Survey

To investigate the inner structure of the JSS' results, a multidimensional cluster analysis was performed using the robust implicit regression model (see section 24.4):

1. Severity vectors were constructed from the severity questionnaire responses $S_{m,n}$, where $m = 1, \ldots, 30$ are the (stressor numbers) and $n = 1, \ldots, 209$ (are the respondent's number).
2. Frequency vectors were made up from the frequency questionnaire answers $F_{m,n}$, where $m = 1, \ldots, 30$ (are the stressor numbers) and once again $n = 1, \ldots, 209$ (are the respondent's number).

3. 30-dimensional vectors of the overall stress index JSS-X were calculated as products $S_{m,n} \times F_{m,n}$ $(m = 1, \ldots, 30)$ for each $(n-\text{th})$ respondent and used as the explanatory variable.

$$\sum_{m=1}^{30} C_m * S_{m,n} * F_{m,n} = 1 \quad (n = 1, \ldots, 209). \tag{26.5}$$

4. The equation system 26.5 was solved by means of the gnostic robust method with respect to the model's parameters C_m.
5. The vector of residuals (equations' errors) was obtained by substituting all the values of C_ms and $S_{m,n} * F_{m,n}$s in the model.
6. Six homogeneous clusters of JSS-X vectors (of respondents) were extracted from the above by repetitively extracting the 'main' cluster.

The local distribution function ELDF of the residuals is shown in Fig. 26.19.

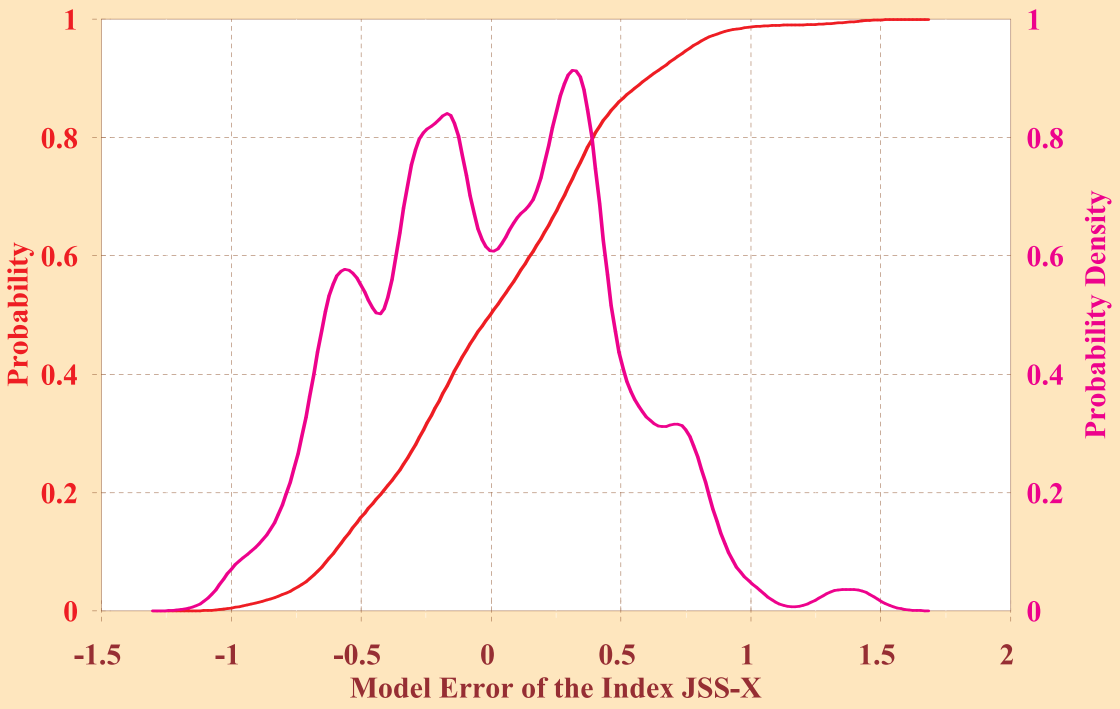


Fig.26.19: RESIDUES OF THE MODEL

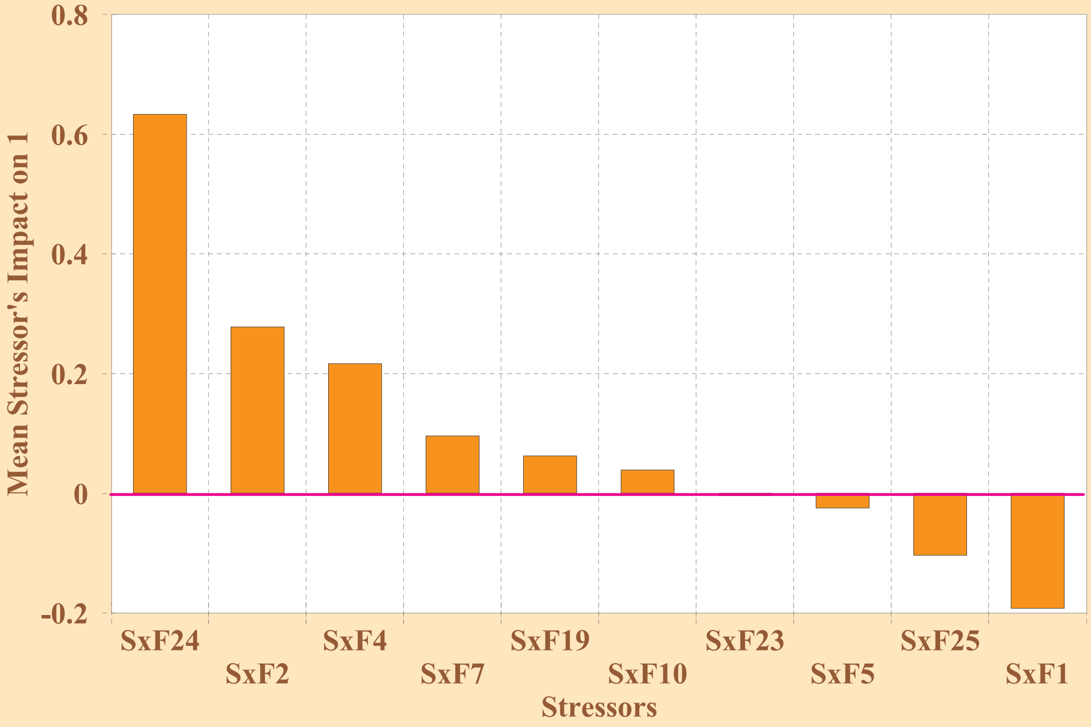
Implicit Regression Model of JSS-X

The inhomogeneity of the JSS-X's data is evident due to the appearance of several local maxima.

The structure of each cluster obtained by the multidimensional cluster analysis is specific, the roles of stressors differ from those found in the global analysis; they reflected only the overall tendencies. An illustration of a cluster's composition is shown in Fig. 26.20 for cluster 2.

Fig.26.20: STRUCTURE OF A CLUSTER

Mean Impacts of Stressors in Clust. #2



The columns are proportional to the mean impacts on the right hand value of 1. So, eg, the first column, belonging to stressor # 24, is obtained as $C_{24} * \overline{S_{24,n} * F_{24,n}}$, where n runs through the respondents' numbers that form the cluster, and where the overline denotes the arithmetic mean. The negative contributions to 1 can be interpreted as effects of mutual dependencies over the stressors represented in the cluster. For instance, stressor 24 characterizes the situation 'Frequent changes from boring to difficult tasks', while the (most negatively acting) stressor 1 relates to 'Assignment of unpleasant tasks.' They both speak to task assignment and are thus dependent on each other. Those respondents choosing # 24 (preferring it as the better formulation of the problem) increase its score, but by leaving # 1 out, they decrease it and vice versa.

The five strongest impacts in all six clusters are summarized in Tab. 26.7.

Although the respondents, who fall into a specific cluster, answer questions related to their working conditions, they do not characterize their workplace, because they come from different occupational environments. Rather, they give evidence of their own subjective sense of the environment. On the other hand, occupations differ in the stresses they produce. Therefore, the clustering reflects both objective and subjective factors. This can be seen in Fig. 26.21, where the respondents forming the clusters are distinguished by gender, profession and age.

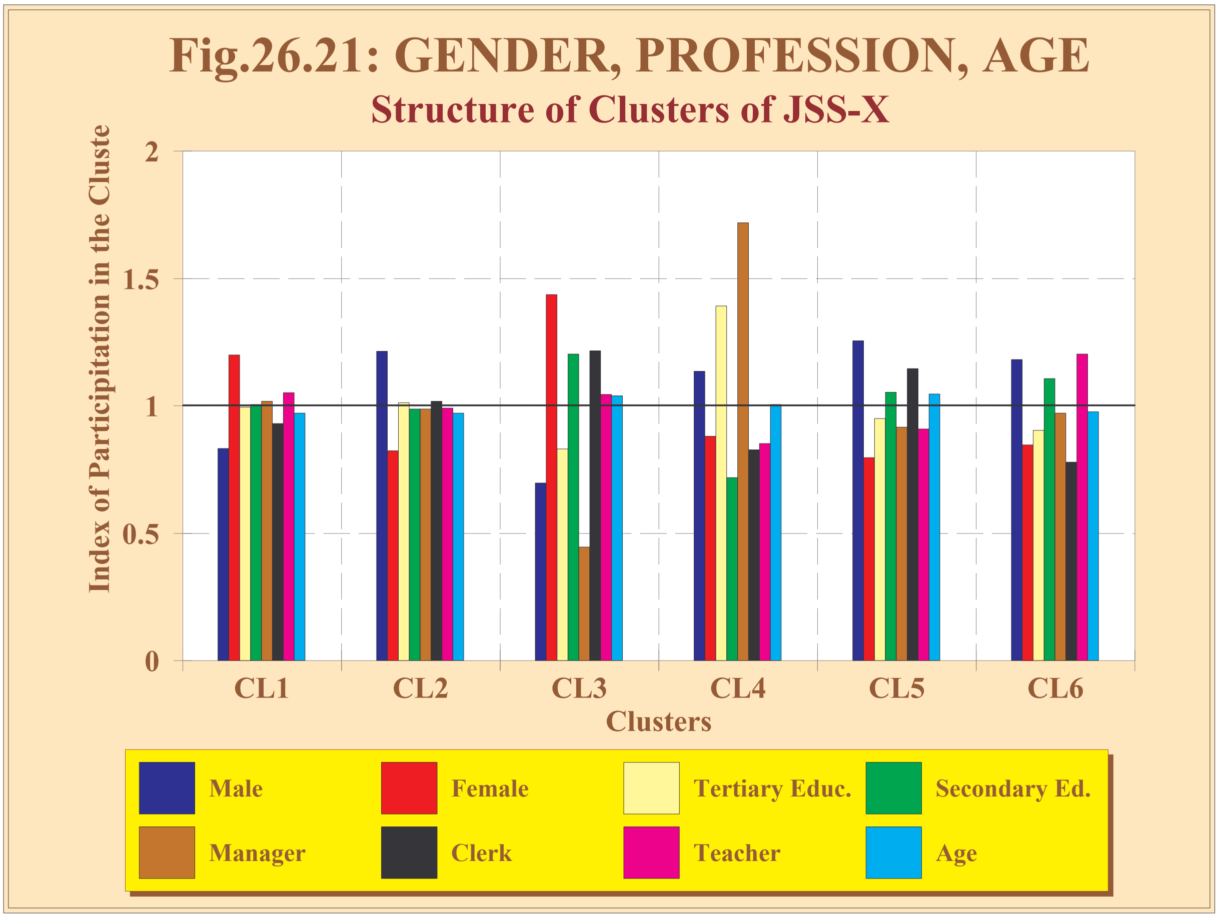


Fig.26.21: GENDER, PROFESSION, AGE

Structure of Clusters of JSS-X

The columns are proportional to the indexes calculated as the number of respondents in the cluster divided by the number of respondents with the same attributes, who participated in the survey.

The easiest trait to interpret is the age: it has no role, all age indexes are close to 1.

#	Stressors in Cluster 1	Impact
20	Competition for careers	0.55
28	Doing the work of a colleague	0.21
11	Assignment of a higher responsibility	0.06
12	Periods of inactivity	0.06
15	Insufficient staff to manage the tasks	0.05

#	Stressors in Cluster 2	Impact
24	Frequent changes from boring to difficult tasks	0.62
2	Working overtime	0.27
4	Assignment of new or unusual duties	0.22
7	Solving crises	0.096
19	Inadequate salary	0.063

#	Stressors in Cluster 3	Impact
15	Insufficient staff to fulfill the task	0.43
29	Inadequate motivation of co-workers	0.22
18	No participation in decision making	0.15
10	Inadequate or bad equipment	0.11
28	Doing the work of a colleague	0.08

#	Stressors in Cluster 4	Impact
19	Inadequate salary	0.28
3	Lack of opportunity for promotion	0.24
4	Assignment of new or unusual duties	0.12
29	Insufficient motivation of co-workers	0.11
2	Working overtime	0.07

#	Stressors in Cluster 5	Impact
6	Inadequate support from the boss	0.55
5	Co-workers not doing their jobs	0.27
22	Noisy workplace	0.13
11	Assignment of a higher responsibility	0.05
2	Working overtime	0.03

#	Stressors in Cluster 6	Impact
23	Frequent disturbances	0.52
3	Lack of chances for promotion	0.22
13	Difficult relations with the boss	0.17
4	Assignment of new or unusual duties	0.10
19	Inadequate salary	0.10

Tab.26.7: The main impacts in response clusters

An interesting comparison is offered by clusters 1 and 2, which differ only by gender: the first is dominated by women, the second by men, while all the other aspects are close to 1 and have no impact. Table 26.7 says, that the predominantly feminine personnel in cluster 1 tend to be affected by 'typically masculine' competitiveness, they prefer to be left alone to do their own work without being given excessive responsibility or overburdened. The mainly masculine makeup of cluster 2 prefer steady work at a reasonable salary with no unusual assignments.

In cluster 3 women prevail and this group are mainly clerks and personnel with a secondary education. They see the cause of their stresses as being overloaded, underestimated both as individuals and as a team as well being inadequately supported by their employer.

The main stress in cluster 4 is caused by the conflict between own ambitions and expectations of reward through promotion in view of the enhanced effort expended. This (largely masculine) cluster differs from the others in that it has a high proportion of managers and workers with a college education.

Cluster 5 is dominated by males and clerks, who express their dissatisfaction with relations with both their supervisors and colleagues, who are perceived to not be pulling their weight.

The (also largely masculine) cluster 6 has the largest participation from teachers: they develop high stress levels from being interrupted in their work or disturbed. Feelings of unsatisfied personal ambitions and difficult relations with supervisors seem to offset stresses developed from new assignments.

This examples demonstrates, that even data obtained by a very primitive and rather subjective 'measuring' technology of questionnaires can reveal useful facts, when the tools of advanced analysis are applied.

26.6 Medical and Genetic Diagnostics

Most of us know of someone, if not ourselves, who have been told by their doctors: 'You are overweight,' 'Your blood pressure is high,' 'Your cholesterol is at a dangerous level.' These pronouncements all say the same thing: 'The parameters that measure your physical well being are out of normal bounds.' These are all related to the problem of normality repeatedly discussed in this book. Physicians routinely must decide if the boundaries between healthy and sick have been crossed and if so then to diagnose the problem and recommend a cure. Each specific illness is defined by a complex set of features, which have to be compared with those, which describe good health. Many of these characteristics are expressed in numerical form, as some critical bounds, which have been established over the long-term. Some of these bounds are supported by research. An important problem to be examined is, whether this application field is sufficiently free from the prejudices of (Gaussian) 'normality' to accept open minded interpretations of experimental data. For a study to be accepted by a leading journal it is necessary that customary classical statistical tests 'demonstrating the statistical significance' of the results to be included. It must once again be noted, that such tests as variance or factor analysis, t- and F-tests and others are directly dependent on the acceptance of a priori assumptions as to the data's statistical model, which fact is—as a rule—not subject to verification.

26.6.1 Osteoporotic Data

With currently used techniques, in medical studies, there is a danger of misinterpreting analytical results if the distribution functions of the parameters in question are not Gaussian, and the further they extend from statistically normal characteristics, the worse the results. To obtain a sufficiently rich collection of distribution functions, data[6] were obtained from a large medical research project on osteoporosis in postmenopausal women. There has been long-term experience in the application of gnostic methods in this field [63]:

Calcium, one of the most important physiological regulators of parathyroid function, suppresses parathyroid hormone (PTH) serum levels. In postmenopausal women relative hyperparathyroidism was found manifested by a lower PTH suppressibility by exogenous hypercalcemia (Lo et al., 1988). This condition could have a negative impact on bone metabolism. Because it is not clear, whether this phenomenon is conditioned by a drop in estrogen levels, an experiment was designed to influence it by estradiol (E_2) administration, 100 μg/d for a period of three months. The PTH responses and attained serum calcium levels following calcium administration were assessed several times. PTH and calcium-induced values were estimated by measuring the changes in the areas under the graphs of both the PTH and the calcium measurements. The effect of E_2 was then determined by comparing the data for the same women before and after

[6] By courtesy of Doc.MUDr. Žofková, DrSc. of the Institute of Endocrinology, 116 94 Prague

the E_2 treatment. From estimates computed for 9 patients it ensued, that E_2 enhanced the PTH suppressibility, so that the decremental PTH area in eight of the nine women after the treatment was higher than before the treatment, despite the fact, that after the treatment, the same calcium load produced lower calcium levels than before.

From the point of view of a professional judgement, the experiment was undoubtedly successful, but it was not possible to reject the zero statistical hypothesis: the expectation of a positive effect of the treatment was not supported by the statistical test on a sufficiently significant level. Thus, a common sense judgement on the success of the experiment appeared to be in conflict with the statistical data analysis. When gnostic distribution functions were applied to these data, they demonstrated their superiority over traditional methods. The results of the new method were substantially less sensitive to possible data variations and more informative. The method enabled a reliable confirmation of the positive effect of estradiol on suppressibility of PTH by exogenous hypercalcemia to be made. It was possible to quantify the expectation of the success of the treatment and to point in the direction of a more detailed investigation.

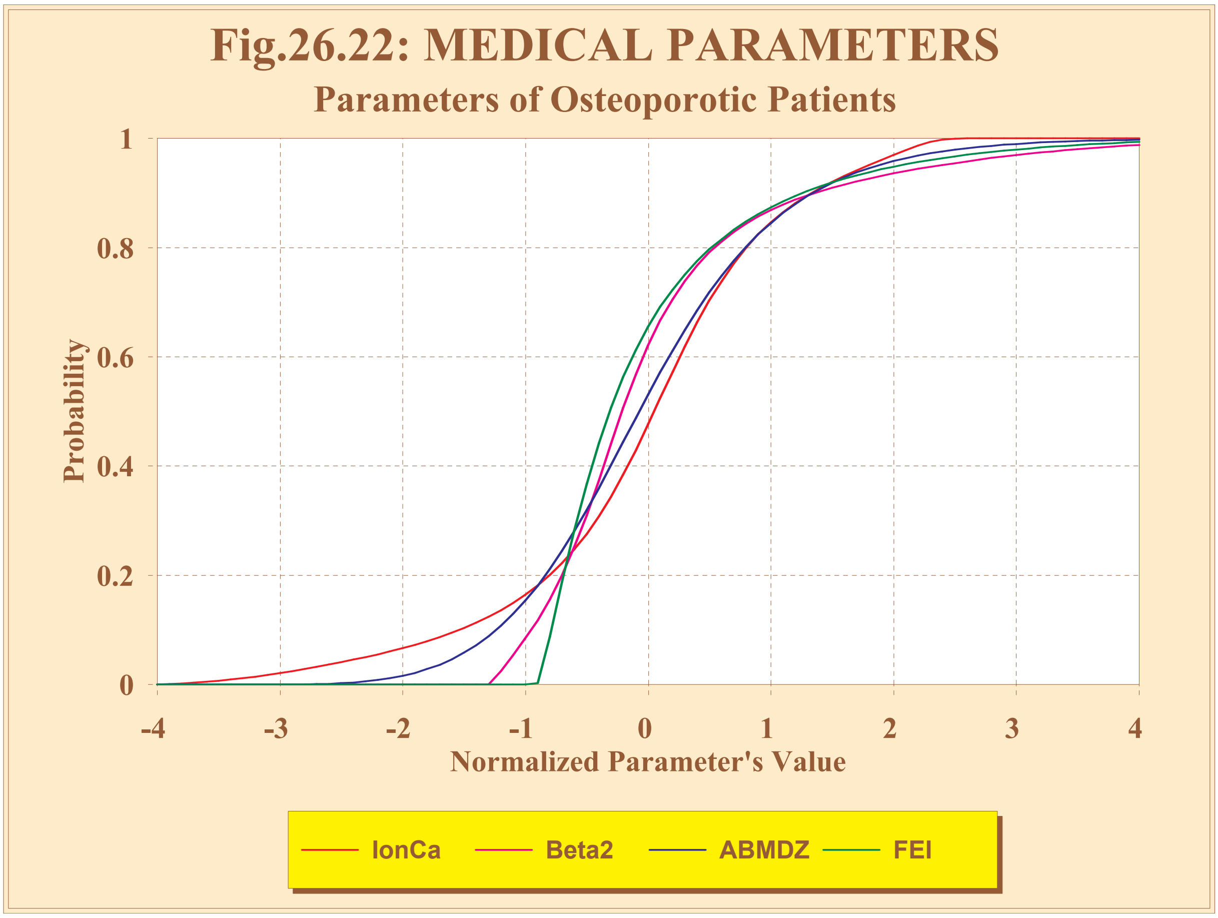
The research project has been expanded to result in a database of 114 female patients and has included the measurement of (up to) 39 medical parameters (Tab. 26.8) and 8 genes with their alleles (Tab. 26.9). The survey had a preventive character, the patients were not ill: their health was adequate for their ages, which were between 27 (the lower outlier) through 80 with a mean of 62.5 years.

The available data enabled the actual distribution functions of the medical parameters to be estimated and analyzed. To demonstrate the inappropriateness of the Gaussian assumption for most of the parameters, arithmetic means (AVG) and standard deviations (STD) were calculated for all the variables. The data (eg D_x) were than normalized by using the formula $N(D_x) = (D_x - AVG_x)/STD_x$, where $N(D_x)$ is the normalized value of D_x and AVG_x with STD_x its mean and standard deviation. Gnostic distribution functions of normalized data were then calculated, the homogeneous part of each data sample was determined together with its membership bounds LSB and USB. This allowed the lower and upper outliers (data out of the membership interval) to be determined. All these results are reviewed in Tab. 26.8.

Note, that statistics AVG and STD are expressed by using the variables' natural scale, while the bounds LSB and USB are normalized (centered and shown as multipliers of STD). This means, that if the nor-

Parameter	**Natural Scale**		**Normalized Scale**		**Description of the parameter**
	AVG	STD	LSB	USB	
Age	62.5	9.0	-1.86	1.97	Age
BMI	25.7	3.5	-1.94	3.46	Body Mass Index
YAMP	13.5	8.5	-1.72	2.92	Years after the menopause or OOX
IonCa	1.3	0.05	-4.11	2.52	Ionized calcium
AF	1.8	0.59	-1.95	17.6	AF total
YSM	13.4	8.35	-1.84	2.97	YSM
BMDWT	0.6	0.15	-2.05	3.74	BMD at the Wards Triangle
WardT	-2.3	1.34	-1.91	3.52	Wards T-score
WardZ	0.23	1.32	-1.67	2.68	Ward's Z-score
Beta2	2.3	0.79	-1.31	5.62	Beta2 microglobulin
IRI	11.3	6.6	-1.36	3.15	Insulin (IRI)
IGFI	153.6	59.2	-2.04	6.71	IGF - I
Estrad	0.08	0.10	-0.93	5.37	Estradiol
Calcit	14.8	16.7	-0.97	5.64	Calcitonin
Ostkalc	12.1	7.9	-1.56	3.60	Osteocalcin
Int.PTH	38.7	22.6	-1.62	18.9	Intact PTH - IRMA
Calcid	24.3	32.5	-0.78	8.42	Calcidiol
DPYD	7.1	3.9	-1.24	2.86	DPYD in urine/ 2 h.
ABMDts	0.86	0.15	-2.63	7.12	A - BMD Total Spine
PMBD	0.79	0.13	-1.55	3.50	P - BMD mean diameter of both extremities
ABMDT	-1.7	1.4	-2.43	5.78	A - BMD T-score
PBMDT	-1.5	1.1	-1.54	3.22	P - BMD T-score
ABMDZ	-0.12	1.40	-3.47	17.0	A - BMD Z-score
PBMDZ	-0.25	1.01	-1.93	7.85	P - BMD Z-score
CBUAT	-1.8	1.04	-1.72	3.81	Cuba BUA T-score
CBUAZ	0.02	0.89	-1.92	5.72	Cuba BUA Z-score
sPICP	125.7	38.1	-2.48	3.77	Serum PICP
CBUA	66.8	18.4	-1.69	3.89	Cuba BUA (PK+LK)/2
Dvit	44.5	19.2	-1.75	6.41	1.25 (OH) D vit. calcitriol
Testst	1.5	0.7	-2.19	3.05	Testosterone
DHEAS	2.3	1.6	-1.41	14.9	DHEAS
DHEAN	8.1	5.4	-1.49	14.1	DHEA unconjug.
Andrst	2.3	1.4	-1.45	20.5	Androstendion
SHBG	60.2	28.7	-1.87	3.11	HBG
Ca24	4.6	2.4	-1.90	2.70	Calcium in urine/24 h.
sICTP	3.1	1.0	-1.93	4.18	Serum ICTP
IGFBP	5.5	2.0	-1.63	3.31	IGFBP-3
FTI	3.3	2.5	-1.26	10.97	FTI = (free testosterone divided by SHBG)x100
FEI	0.23	0.28	-0.81	5.44	FEI = (free estradiol divided by SHBG)x100

Tab. 26.8: Statistics (arithmetic mean AVG and standard deviation STD) of osteoporotic patients in original scales and gnostic critical values (bounds of the membership intervals) in normalized scale $(X - AVG)/STD$



malized distributions of all variables were ‘the same’, all LSBs would be closely concentrated about a constant; analogous effects were observed for all USBs. The fact, that nothing like this appears in Tab. 26.8 is due to the widely differing form of the distributions: the lower bound LSB ranges from -4.11$*STD$ (Ionized Calcium) through -0.78$*STD$ (Calcidiol), while the upper bound USB can be as small as 2.70$*STD$ (Calcium in urine) and as large as 20.51$*STD$ (Androstendion).

Examples of the probability distribution functions are in Figs. 26.22 and of their densities in Fig. 26.23.

Only one of the distributions is reminiscent of the Gaussian form (ABMDZ), but as seen in Tab. 28.8 even in this case the membership interval is bounded (-3.47$*STD$, 17.0$*STD$). It can be seen, that the cut-offs of the distribution definition ranges causes their asymmetry. All lower and upper bounds of data support for all the analyzed medical parameters are finite and asymmetric.

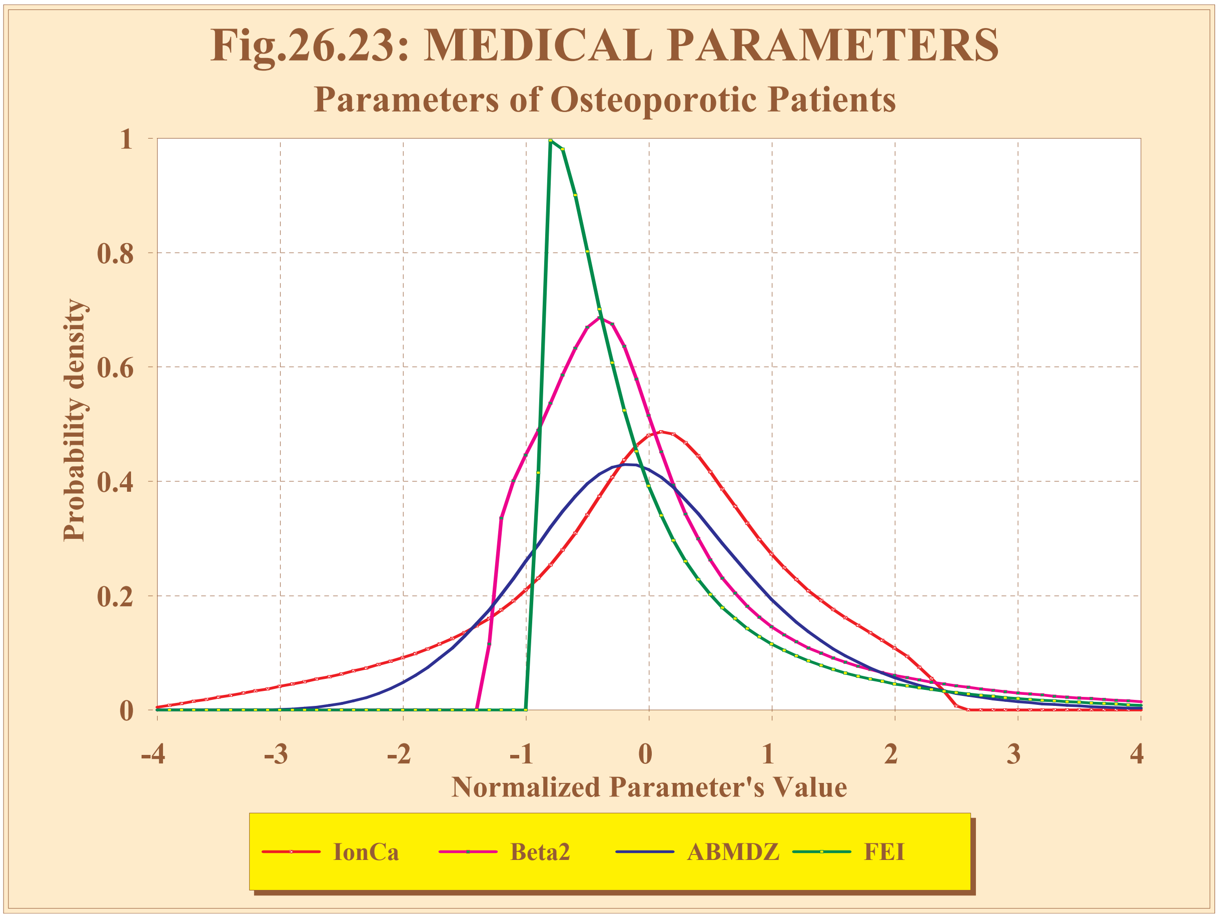


Fig.26.23: MEDICAL PARAMETERS

All of this supports the thesis on the necessity of not basing ones diagnostic judgments on assumptions/prejudices, but on the real specific form of the distribution of the parameter being considered. The diagnostic thresholds should be established individually for each of the medical parameters.

26.6.2 Genetic Impacts

An important aspect of the survey, which addresses osteoporosis, deals with the level of several hormones. However, the analysis of survey data found significant correlations only between hormones, but not between hormones and other medical parameters. The focus was then turned to the genetic characteristics of the participants.

Eight genes were identified within the framework of the survey and they are described in Tab. 26.9 together with their alleles[7].

[7] One of two or more different genes containing specific inheritable characteristics, that occupy corre-

Genes Description		
Symbol	Name	Symbols of Alleles in Tab. 26.10
G1	VDR/FOK I	A1...FF, A2...Ff, A3...ff
G2	VDR/Apa I	A1...AA, A2...Aa, A3...aa
G3	VDR/Taq I	A1...TT, A2...Tt, A3...tt
G4	VDR/Bsm I	A1...BB, A2...Bb, A3...bb
G5	ESR-1/Pvu II	A1...PP, A2...Pp, A3...pp
G6	ESR-1/Xba I	A1...XX, A2...Xx, A3...xx
G7	CALCR/Alu I	A1...TT, A2...TC, A3...CC

Tab. 26.9: Genes and their alleles identified in the patients

The objective here was to find possible relations between the genotypes and the other medical parameters. A useful result was obtained by using the gnostic concept of unidimensional cluster analysis: 'normal values d_x of a data sample X are those satisfying the conditions $LSB_x \leq d_x$ and $d_x \leq USB_x$', where LSB_x and USB_x are the bounds of the X's membership interval. This rule was applied to the osteoporotic survey to classify a patient as normal if all her's parameters fell within the bounds. This cluster contained 54 patients and was denoted M; the rest of the 114 respondents formed the lower or upper cluster (L or U). The relative incidence of each of the alleles in all of the clusters was counted and showed a strong nonuniformity. Therefore the ratios of these relative incidences (in 'normal' patients divided by the rest) was calculated for 21 genotypes. The results are in Fig. 26.24.

It can be shown, that in at least 6 cases of genotypes the relative incidence in normal patients versus the rest is statistically significant. Taking into account, that the identification of normal patients was realized by gnostic diagnostics, the result can be interpreted as statistical support for the gnostic method of identification of 'normal' patients.

The membership bounds in Tab. 26.8 were established by the univariate analysis. The incidences of genes in the 'M' cluster motivates a further step in the analysis to answer the question as to the impacts of genotypes on the probability distribution of hormones. Such an impact really exists as is shown in Fig. 26.25, where the probability distributions of parathyroid hormone (PTH) is presented for three alleles of the gene G1: the form of the distribution can be strongly dependent on the allele.

Even stronger impacts can be found in the bounds of membership intervals as can be seen in Tab. 26,10.

sponding positions on paired chromosomes.

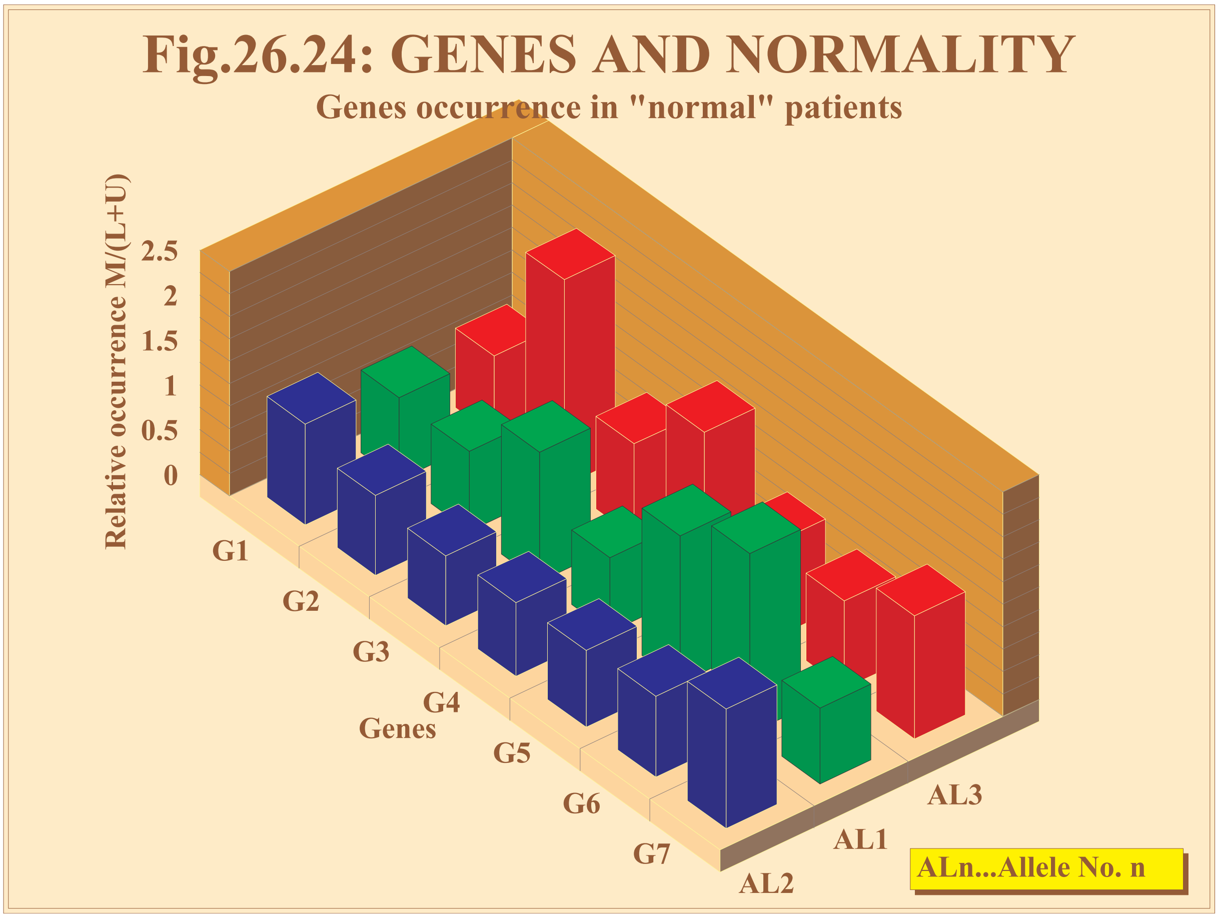
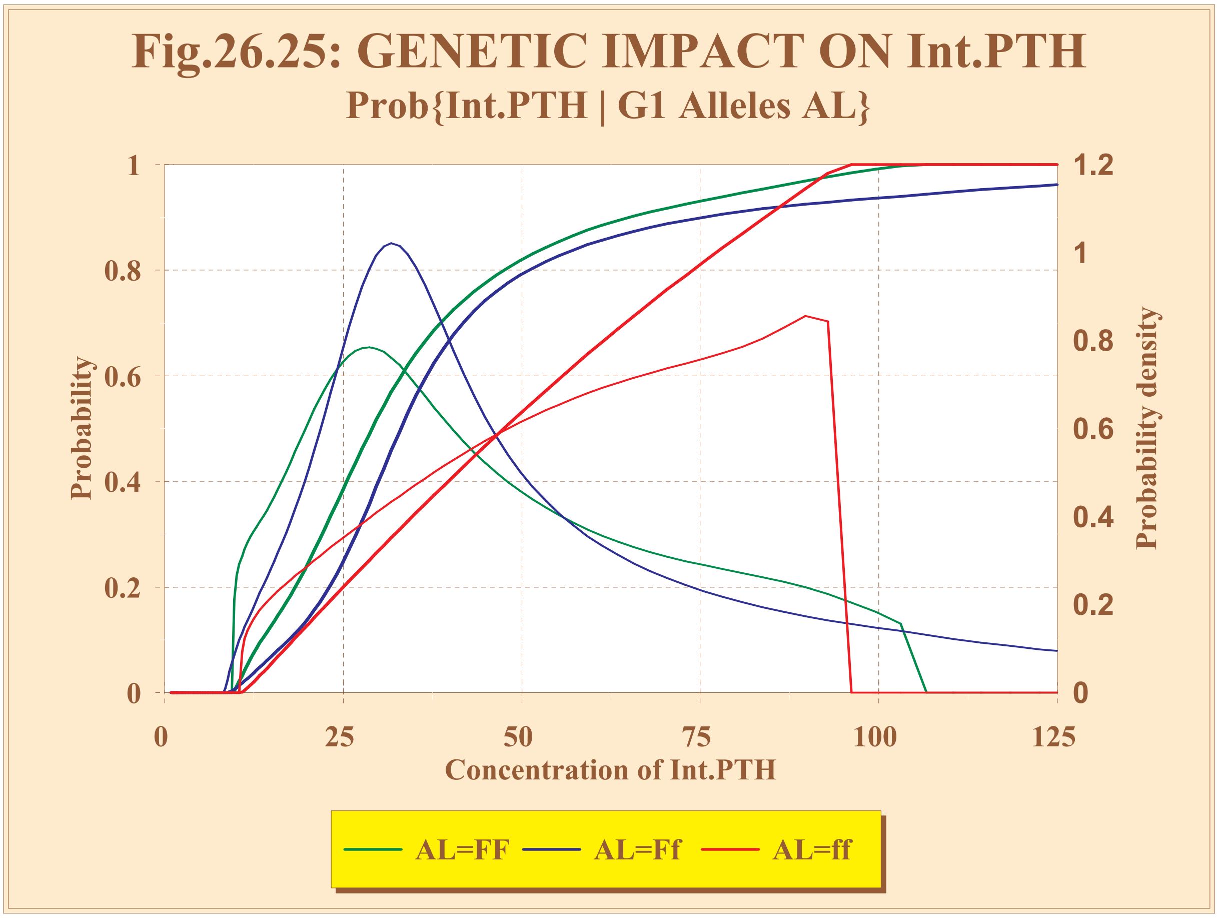


Fig.26.24: GENES AND NORMALITY

The above leads to the conclusion, that it would be impossible to establish diagnostic PTH thresholds for all patients without taking into account their genetic makeup.

	Gene's Alleles					
Gene's	A1		A2		A3	
Symbol	*LSB*	*USB*	*LSB*	*USB*	*LSB*	*USB*
G1	9.59	104.91	8.40	621.82	10.71	94.91
G2	12.01	92.19	5.92	198.31	9.92	122.27
G3	5.91	997.92	11.14	97.22	12.70	54.03
G4	12.55	88.38	9.02	94.68	4.24	4008.67
G5	17.07	118.47	16.68	111.74	7.95	112.47
G6	16.11	97.64	16.64	112.38	7.53	111.65
G7	5.92	116.16	12.90	120.02	0.70	44334.89
G8	19.76	124.38	6.00	138.08	26.14	1111.50

Tab. 26.10: Membership bounds LSB and USB for the parathyroid hormone in dependence on the genes and their alleles



26.7 Summary

All application fields, which use numerical data contaminated by uncertainty, require an economical methodology for data treatment. Given the cost of data, maximization of the information contained therein is a mandatory requirement. This chapter used a varied number of examples to demonstrate the importance of this task and to show how the use of advanced analytical methods can yield vastly superior results. The range of potential applications, which are suitable for the use of gnostics is very broad. This was shown by illustrations from such diverse fields as a geological survey, techniques of physical measurement, historical numismatics, production quality assessment in the chemical industry, mechanical engineering, computer chip production, psychology and medical genetics applications. Common problems exist in all these fields: the requirement for both unconditional and conditioned probability distribution functions completely determined by data, robust uni- and multidimensional cluster analysis, specific distribution-based diagnostic thresholds and other robust methods of mathematical gnostics, all of which maximize the information obtained by the results. The gnostic paradigm of uncertainty proves its viability by the efficiency of its applications.

Bibliography

Bibliography

[1] Altman E.I.: A Further Empirical Investigation of the Bankruptcy Cost Question, Journal of Banking and Finance, September 1984b, 1067-1089

[2] Smith A.: The Wealth of Nations, Random House, Inc., 1994, Modern Library Edition.

[3] Bachelier L.: Theorie de la Speculation, Gauthier-Villards, Paris, 1900

[4] Banerjee K.S., Carr R.N.: A comment on Ridge Regression. Biased Estimation for Nonorthogonal Problems. Technometrics **13**, No.4 (1971)

[5] Barker Joel A.: Paradigms, The Business of Discovering the Future. Harper Business, (1992).

[6] Bell David Arthur: Information Theory and its Engineering Applications, 2-nd Ed., Sir Isac Pitman and Sons Ltd., London (1956).

[7] Bell David Arthur: Entropy Change in Electrical Communication. American Scientist 40, (1952), p.682.

[8] Bloch A.: Murphy's Laws, Price Stern Sloan, Inc., Los Angeles (1993)

[9] Block S.B., Hirt G.A.: Foundations of Financial Management, Sixth Edition, Irwin (1992)

[10] Blum M., An Extension of the Minimum Mean-Square Prediction Theory for Sampled Input Signals, IRE Trans. **IT-2**, Sept. (1956)

[11] Bochner S.: Harmonic Analysis and the Theory of Probability, University of California Press (1955)

[12] Brillouin L.: Maxwell's Demon Cannot Operate: Information and Entropy. Journ.Appl.Phys. 22, (1951), p.334.

[13] Baeyer H.C. von: Maxwell's Demon (Why Warmth Disperses and Time Passes), Random House, New York, (1998)

[14] Chajdiak J., Analysis of Return—Pyramidal Models (in Slovak), STATIS, Bratislava, Slovak Republic, (1995)

[15] Chasteen L.G., Flaherty R.E. and O'Connor M.C., Intermediate Accounting, Fourth Edition, McGraw-Hill, Inc., New York, (1992)

[16] Cramer Harald, Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J., (1946).

[17] Helmholtz H. von, Zaehlen und Messen erkenntniss-theoretisch betrachtet, in *Philosophische Aufsaetze Eduard Zeller gewidmet*, Leipzig (1887), s.17-52

[18] Stephan Klement, William P. Acito, Karl W. Kratky and Johann Nittmann, Multifractal Analysis of Airborne Microcontamination Particles, Aerosol Science and Technology, Vol.20 (2), 1994

[19] Danos P., Imhoff Jr. E.A., Introduction to Financial Accounting, Irwin, Homewood (1991)

[20] David F.N., Probability Theory for Statistical Methods, Cambridge (1951)

[21] Eco U.: The Island of the Day Before, Vintage (1998)

[22] Fine Terrence L., Theories of Probability; an Examination of Foundations, Academic Press, New York and London, (1973).

[23] Alexander D., Nobes C.: Financial Accounting. An International Introduction, Pearson Education Limited, England (2001)

[24] Foster G.: Financial Statement Analysis, Prentice-Hall, Inc., Englewood Cliffs, New York (1986)

[25] French D., Dictionary of Accounting Terms, The Institute of Chartered Accountants in England and Walles, London (1985)

[26] Galton F., Family Likeness in Stature, Proceedings of Royal Society, London, vol.40 (1886), 42–72

[27] Gaynor P.E., Kirkpatrick R.C., Introduction to Time-series, Modeling and Forecasting in Business and Economics, McGraw-Hill, Inc., N.Y., 1994

[28] Gordon M., The Investment, Financing, and Valuation of the Corporation, Homewood, Ill., Richard D. Irwin, (1962)

[29] Gradstein I.S., Rizhik I.M.: Tabels of Integrals, Sums, Series and Products, GIFML, Moscow (1963) (in Russian)

[30] Graham, Benjamin: The Intelligent Investor, Fourth Revised Edition. Harper & Row, New York 1973

[31] Gujarati D.N., Basic Econometrics, McGraw-Hill Publ.Co., N.Y., 1988

[32] Horngren Ch.T., Sundem G.L. and Elliott J.A., Introduction to Financial Accounting, Fifth Edition,Prentice-Hall, Inc., London,...,(1993)

[33] Huber P.J, Robust Statistics, Wiley, New York, 1981

[34] Industrial Performance Analysis, ICC Business Publications Ltd., Field House, 72 Oldfield Road, Hampton, Middlesex TW12 2HQ, UK

[35] Interational Accounting Standards 2000, International Accounting Standards Committee, London (2000)

[36] International Financial Reporting Standards 2005. International Accounting Standards Board, London (2005)

[37] International Valuation Standards (2007), www.ivsc.org/pubs

[38] Johnson K.R., Optimum, Linear, Discrete Filtering of Signals Containing a Non-random Component, IRE Trans. **IT-2**, June (1956)

[39] Turnage J.J., Spielberger C.D.: Job stress in managers, professionals, and clerical workers, Work & Stress, **5** (1991), 165-176

[40] Spielberger C.D.: Professional Manual for the Job Stress Survey (JSS). Odessa. FL: Psychological Assessment Resources, Inc. (PAR)

[41] Kalman R.E., A New Approach to Linear Filtering and Prediction Problems, Trans. ASME, March (1950)

[42] Kalman R. E.: The Problem of Prejudices in Scientific Modeling. Final, written version of an invited lecture given on Sept. 4, 1986 at the European Econometric Meeting in Budapest, Hungary, with the title *Foundation Crisis in Econometrics within the Standard Statistical Paradigm*

[43] Kendall M.: The Analysis of Economic Time Series, Part I, Prices, Journal of the Royal Statistical Society, vol. 96, pp 11-25, 1953

[44] Kneale, D.: Into the Void: What Becomes of Data Sent Back From Space? Not a Lot as a Rule. The Wall Street Journal, January 12, 1988, pp.1 and 33.

[45] Kochin N. E., Vector Calculus and Principles of Tensor Calculus, Publishing House of Academy of Sciences, Moscow (1951)

[46] Kolmogorov A.N., Extrapolation and Interpolation of Stationary Random Sequences, (in Russian), Izvestiya Akademii Nauk SSSR (Reports of the Academy of Sciences of Soviet Union), math. series **5**, No.1 (1941)

[47] Kovanic P., Generalized Discrete Analogy of the Zadeh-Ragazzini's problem, (in Russian), Avtomatizacya i Telemechanizacya (Automation and Telemechanics), **XXVII**, No.2 (1966) 37–48

[48] Kovanic P., Static Programming of Data Handling, Nuclear Electronics, International Atomic Energy Agency, Vienna (1966), 559 – 574.

[49] Kovanic P., Identification of Operators in the Reactor Physics, (in German), Atomkernenergie **12**, No.11/12 (1967) 404-408

[50] Kovanic P., Optimum Digital Operators, IFIP World Congress (1968), Edinburgh, in Information Processing 68, North-Holland Publishing Company, Amsterdam (1969) 249–255

[51] Kovanic P., Rygl J., In-line Computers for a Reactor Safety System?, Proceedings "Performance of Nuclear Reactor Components", IAEA Vienna (1969), 259-272

[52] Kovanic P., Minimum Penalty Estimate, Kybernetika **8** No.5 (1972), 367–383

[53] Kovanic P., Generalized Linear Estimate of Functions of Random Matrix Arguments, Kybernetika **10** No.4 (1974),303–316

[54] Kovanic P., Votlucka J., Blecha K., Experimental Determination of Power Sensitivity of a Electrical Power Distributing System by Means of Periodical Impulses of Power (In Russian), Elektrotechnicky obzor 68 (1979), 3, 133-139

[55] Kovanic P., Votlucka J., Blecha K., Results of Measuring Power Sensitivity of the Joint Power Distributing System of Socialist Countries (In Czech), Elektrotechnicky obzor 68 (1979), 10, 614-619

[56] Kovanic P., Gnostical Theory of Individual Data, Problems of Control and Information Theory 13 (1984), 4, 259-274.

[57] Kovanic P., Gnostical Theory of Small Samples of Real Data, Problems of Control and Information Theory 13 (1984), 5, 303-319.

[58] Kovanic P., On Relations between Information and Physics, Problems of Control and Information Theory 13 (1984), 6, 383-399.

[59] Kovanic P., A New Theoretical and Algorithmical Basis for Estimation, Identification and Information, IX-th World Congress IFAC '84, Preprints, IFAC Budapest (1984), Vol.XI, 122-131.

[60] Kovanic P., A New Theoretical and Algorithmical Basis for Estimation, Identification and Control, Automatica V22, (1986), 6, 657-674.

[61] Kovanic P., Gnostical Theory of Uncertain Data, DrSc Thesis, Institute of Information Theory and Automation of Czech Academy of Sciences, 1990, 152 pp., 9 figs.

[62] Kovanic P., Novovicova J.: On Robust Estimators Worth to be Applied to Real Data. Research Report 1463 (1987), Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences, Prague

[63] Kovanic P., Žofková I.: Medical Experience with Small Data Samples Processing, Second European Congress on System Sciences, Prague, Oct. 5–8, 1993

[64] Kovanicova D., Kovanic P., Treasures Hidden in Accountancy, Part 2, Analysis of Financial Statements (in Czech), Polygon, Prague, Czech Republic, (1997)

[65] Kuhn Thomas, The Structure of Scientific Revolution, 2nd Edition, (Enlarged), The University of Chicago Press, Chicago, Ill., (1970).

[66] Larsen E.J., Modern Advanced Accounting, McGraw-Hill., Inc., New York, (1991)

[67] Leech, J.W.: Classical Mechanics, 2nd Ed. Methuen & Co., Ltd.., Frome and London, (Great Britain), (1965)

[68] Lees A.B., Interpolation and Extrapolation of Sampled Data, IRE Trans. **IT-2**, March (1956)

[69] Lewis T.O., Odell P.L., A Generalization of the Gauss-Markov Theorem, Amer. Stat. Ass. Journal (Dec,1966),1063–1066

[70] Los, A.L.: A Scientific View of Economic Data Analysis. Eastern Economic Journal, Vol.XVII, No.1, January-March 1991, pp.61-71.

[71] MacLane, Saunders & Birkhoff, Garrett, Algebra, 3rd Ed., Chelsea Publishing Company, New York, NY, 1993.

[72] The Mathematics of Markets, A Survey of the Frontiers of Finance. The Economist, 9 October 1993, pp.1-22 (special survey, after p. 60).

[73] Malkiel B.G., A Random Walk Down Wall Street, W.W. Norton, New York, 1975

[74] Mallows C.L., in: Time Series, North Holland, Amsterdam (1980)

[75] Mayer, M.: Markets. W.W. Norton & Company, New York and London, 1988, 303 pp.

[76] Meloun M., Militký J.: Statistické zpracování experimentálních dat. Sbírka úloh. (Statistical Treatment of Experimental Data. A Collection of Tasks. In Czech.) University of Pardubice (Czech Republic) (1996)

[77] Novovicova, J.: M-estimators and Gnostical Estimators of Location. Problems of Control and Information Theory, **18** (1989) pp.397-407

[78] Novovicova, J.: M-estimators and Gnostical Estimators of Location. Problems of Control and Information Theory, **19** (1990) pp.127-138

[79] Novovicova, J.: M-estimators and Gnostical Estimators for Identification of a Regression Model, Automatica, **26**, 3 (1990) pp.607-610

[80] Oxford Wordpower Dictionary, edited by Sally Weihmier, Oxford University Press, Oxford,..., Fifth impression (1994)

[81] Parzen, Emmanuel: On Estimation of a Probability Density Function and Mode, Ann. Math. Statist. 33 (1962) 1065-1076

[82] Paukert T., Rubeška I., Kovanic P., A New Look at Analytical Data Through the Gnostical Analyser, The Analyst, **118**, Febr. 1993, 145–148

[83] Penman S.H.: Financial Statement Analysis & Security Valuation, McGraw-Hill & Irwin, N.Y. (2001)

[84] Perez, A.: Mathematical Theory of Information, Application of Mathematics (in Czech) 3, 1 (1958) 1-21 and 2 (1958) 81-99

[85] Peschel, M., Mende, W., Leben wir in einer Volterra-Welt? (Do we live in a Volterra-World? In German). Akademie-Verlag Berlin, Band 14 (1983) Do We Live in a Volterra-World? Mathematische Forschung, Band 14, Akademie-Verlag Berlin 1983 (In German)

[86] Pinta V., Kovanic P.: Příspěvek k metrologii pražských grošøu Karla IV. (1346-1378). (Contribution to the metrology of Prague groshes of Charles the fourth). Numismatické listy (Numismatic Letters), National Museum Prague, Czech Numismatic Society, **LV**, No. 5/6 (2000) 142–148

[87] Luce R.D., Krantz D.H., Suppes P., Tversky A., *Foundations of Measurement*, Vol.III: Representation, axiomatization, and invariance, New York: Academic Press (1990).

[88] Rastrigin P.A., Markov V.A., Cybernetic models of recognition (In Russian), Zinatne, Riga (1976).

[89] Pringle R.M., Generalized Inverse Matrices with Applications to Statistics, Griffin, London (1971)

[90] Raymond R.C.: Communication, Entropy and Life. American Scientist 38, (1950), p.273.

[91] Raymond R.C.: The Well-informed Heat Engine. American Journal of Physics 19, (1951), p.109.

[92] Roberts H.V.: Statistical versus Clinical Prediction of the Stock Market, unpublished paper presented to the Seminar of the Analysis of Security Prices, University of Chicago, May 1967

[93] Rees B.: Financial Analysis, Prentice Hall International (UK) Ltd (1990)

[94] Rocke D.M., Downs G.W., Rocke A.J.: Are Robust Estimators Really Necessary? Technometrics 24 (1982), 2, 95–101

[95] Rosenblatt Murray: Remarks on some Non-parametric Estimates of a Density Function, Annals Math. Statist. 27 (1956) pp.832-837.

[96] Rozenfeld B.A., Multidimensional Spaces (In Russian), Nauka, Moscow (1953)

[97] Samuelson P.A., Nordhaus W.D.: Economics. Fourteenth Edition, McGraw-Hill,Inc., New York, (1992)

[98] Semyonov V.M., On Theory of Extrapolation of Random Processes, (in Russian), Sbornik naucznych trudov VVIA (Transactions of Scientific Papers of the Academy of Air Force), **1**, (1954)

[99] Sindelar J.: Models in Gnostical Theory. International Journal of General Systems. Vol.21, No.4 (1993), 365-378

[100] Sindelar J.: Measurement Theory and Gnostic Theory of Data (in Czech). Research report of the Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences No.1658, Prague (1990), 50 pp.

[101] Sindelar J.: Variational Theorems in Gnostical Theory of Uncertain Data, Kybernetika 31, 1 (1995), 65-82

[102] Smith A.: The Wealth of Nations. (First published in 1776). Pelican Books, Great Britain (1977)

[103] S-PLUS 4 Guide to Statistics, Data Analysis Products Division, MathSoft, Inc. Seattle, Washington 1997

[104] Mallows C.L.: Time Series. North Holland, Amsterdam (1980).

[105] Stevens Mark: Accounting Wars. Macmillan (1985).

[106] Stigler S.M. and discussants: Do Robust Estimators Work with Real Data? Annals of Stat., Vol.5 (1977), No. 6, 1055–1098

[107] Swerling P., Modern State Estimation Methods from the Viewpoint of the Method of Least Squares, IEEE Trans. on Automatic Control **AC-16**, 6 (1971),707–719

[108] Szilard L.: On Entropy-reduction in a Thermodynamic System through the Intervention of an Intelligent Being. Zeitschrift fuer Phys. 53 (1929), p.840.

[109] The Value Line Investment Survey, published weekly by Value Line Publishing, Inc., New York

[110] Triola M.F., Elementary Statistics, The Benjamin/Cummings Publishing Company, Inc., N.Y. (1989).

[111] Vajda, I.: Theory of Information and Statistical Decision (In Slovak), Alfa, Bratislava (1982)

[112] Webster's New World Dictionary of the American Language, Compact Desk Edition, The World Publishing Company, Cleveland and New York, 1963

[113] Weinberg Steven: Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity, John Wiley & Sonc, Inc. New York (1971)

[114] Wessel, D.: Fickle Forecasters. How Three Forecasters, After Crash, Revised Economic Predictions. The Wall Street Journal, December 31, 1987, pp.1 and 28.

[115] Whittle P.: On the Smoothing of Probability Density Functions, J. Roy. Statist. Soc., Ser. B 20 (1958) pp. 334-343.

[116] Wiener N.: Extrapolation, Interpolation and Smoothing of Stationary Time Series, John Wiley and Sons, N.Y. (1950)

[117] Wiener N.: Cybernetics. John Wiley, New York (1949).

[118] Yaglom I.M.: A Simple Non-euclidean Geometry and its Physical Basis, Springer-Verlag New York Inc. (1979)

[119] Zadeh L.A., Ragazzini J.R., An Extension of Wiener's Theory of Prediction, J.Appl.Phys., **21**, No.7, July (1950)

Index