Revidovaný!
111

SEPARATUM

Problems of Control and Information Theory, Vol. 13 (5), pp. 303–319

GNOSTICAL THEORY OF SMALL SAMPLES OF REAL DATA

P. KOVANIC
(*Prague*)

(Received September 10, 1983)

Theory of small samples of real data is based on the gnostical theory of individual data exposed in a foregoing paper of the author. A simple data composition axiom is assumed from which gnostical characteristics of data samples are derived. These characteristics approach the statistical moments of the first and second order when the effect of uncertainty is weak. For strong effects of uncertainties the gnostical characteristics differ from the classical statistical ones. They are more robust with respect to outlying or inlying data. Practically applicable estimators are derived the sensitivity of which can be chosen. Gnostical formulae are given for a direct estimatior

AKADÉMIAI KIADÓ, BUDAPEST
PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
VERLAG DER UNGARISCHEN AKADEMIE DER WISSENSCHAFTEN
MAISON D'EDITIONS DE L'ACADEMIE DES SCIENCES DE HONGRIE
ИЗДАТЕЛЬСТВО АКАДЕМИИ НАУК ВЕНГРИИ

Problems of Control and Information Theory, Vol. 13 (5), pp. 303–319

GNOSTICAL THEORY OF SMALL SAMPLES OF REAL DATA

P. KOVANIC
(*Prague*)

(Received September 10, 1983)

Theory of small samples of real data is based on the gnostical theory of individual data exposed in a foregoing paper of the author. A simple data composition axiom is assumed from which gnostical characteristics of data samples are derived. These characteristics approach the statistical moments of the first and second order when the effect of uncertainty is weak. For strong effects of uncertainties the gnostical characteristics differ from the classical statistical ones. They are more robust with respect to outlying or inlying data. Practically applicable estimators are derived the robustness or sensitivity of which can be chosen. Gnostical formulae are given for a direct estimation of the probability density from small data samples. Examples of practical applications are shown.

1. Introduction and summary of previous results

A new approach to the problem of uncertainty of real data has been introduced in [1]. For each particular datum an *ideal gnostical cycle* exists including three phases: quantification, estimation and attenuation. *Quantification* (measuring of real quantities or counting of real objects) is the way of obtaining a datum which is a numerical image of a real quantity. This image is unprecise because of uncertainty. Under ideal conditions with no influence of uncertainty the result of quantification would be z_0. Actual results of quantification involving uncertainty (*real data*) are z_i $(i=1, \ldots, n)$. By z a possible result of quantification will be denoted. *Ideal estimation* is an optimal transformation of a datum which together with the attenuation yields an *estimate* $\tilde{z}_0$ which coincides with the quantity z_0. An ideal gnostical cycle is optimal in the sense that it minimizes the loss of information. Such a loss has been shown to be unavoidable with an arbitrary closed gnostical cycle. The following results of [1] will be used here:

It results from the model of a possible result of quantification

$$z=z_0e^{\Omega} \qquad (z_0 \in R_+)(\Omega \in R_1) \tag{1}$$

(which is taken as Axiom 1 of the gnostical theory) that the mathematical model of quantification is

$$\mathbf{u}' = \mathbf{K}_q(\Omega)\mathbf{u}_0 \tag{2}$$

where

$$\mathbf{u}'_i := \begin{pmatrix} z_0 \operatorname{ch} \Omega \\ z_0 \operatorname{sh} \Omega \end{pmatrix} := \begin{pmatrix} x \\ y \end{pmatrix} \tag{3}$$

$$\mathbf{K}_q(\Omega) := \begin{pmatrix} \operatorname{ch} \Omega & \operatorname{sh} \Omega \\ \operatorname{sh} \Omega & \operatorname{ch} \Omega \end{pmatrix} \tag{4}$$

$$\mathbf{u}_0 := \begin{pmatrix} z_0 \\ 0 \end{pmatrix} \tag{5}$$

and where the parameter Ω is determined by the contribution of uncertainty. For a datum z_i parametrized by Ω_i the ideal estimating transformation can be written in the form

$$\mathbf{u}''_i = \mathbf{K}_e(\omega_i)\mathbf{u}'_i = \mathbf{K}_e(\omega_i)\mathbf{K}_q(\Omega_i)\mathbf{u}_0 \tag{6}$$

where

$$\mathbf{u}''_i = \begin{pmatrix} r_i \\ 0 \end{pmatrix} \tag{7}$$

$$r_i := \sqrt{x_i^2 + y_i^2} \tag{8}$$

$$\mathbf{K}_e(\omega_i) := \begin{pmatrix} \cos \omega_i & -\sin \omega_i \\ \sin \omega_i & \cos \omega_i \end{pmatrix} \tag{9}$$

and where the relation

$$\operatorname{tg} \omega_i = -\operatorname{th} \Omega_i \tag{10}$$

holds. The attenuating transformation which closes the ideal gnostical cycle determined by the datum z_i is

$$\mathbf{u}_0 = \mathbf{K}_a(k_i)\mathbf{u}''_i \tag{11}$$

where

$$\mathbf{K}_a(k_i) := \begin{pmatrix} k_i & 0 \\ 0 & k_i \end{pmatrix} \tag{12}$$

and

$$k_i := z_0 / r_i. \tag{13}$$

Some important quantities $\mathbf{K}_q^2(\Omega) \equiv \mathbf{K}_q(2\Omega)$ and $\mathbf{K}_e^2(\omega) \equiv \mathbf{K}_e(2\omega)$ have been obtained in [1] as special cases of characteristics of dissimilarity between vectors. They have the form

$$\mathbf{K}_q^2(\Omega) = \begin{pmatrix} 1/f & h_q \\ h_q & 1/f \end{pmatrix} \qquad \mathbf{K}_e^2(\omega) = \begin{pmatrix} f & -h_e \\ h_e & f \end{pmatrix} \tag{14}$$

where for the case $\operatorname{tg} \omega = -\operatorname{th} \Omega$ the following relations hold:

$$f := \frac{x^2 - y^2}{x^2 + y^2} = \frac{2}{\xi^2 + \xi^{-2}} = \frac{1}{\operatorname{ch} 2\Omega} = \cos 2\omega \qquad \text{(“fidelity”)} \tag{15}$$

$$h_q := \frac{2xy}{x^2 - y^2} = \frac{\xi^2 - \xi^{-2}}{2} = \operatorname{sh} 2\Omega = -\operatorname{tg} 2\omega = \frac{h_e}{f} \tag{16}$$

(*“quantifying irrelevance”*)

$$h_e := \frac{-2xy}{x^2 + y^2} = -\frac{\xi^2 - \xi^{-2}}{\xi^2 + \xi^{-2}} = -\operatorname{th} 2\Omega = \sin 2\omega = -h_q f \tag{17}$$

(*“estimating irrelevance”*)

where

$$\xi := z / z_0. \tag{18}$$

It has been shown also that the quantities

$$I_q := \int_0^{h_q} 2\omega \, dh_q \qquad I_e := \int_{h_e}^{0} 2\Omega \, dh_e \tag{19}$$

called the *quantifying and estimating change of information*, respectively, may be written in the form

$$\begin{aligned} I_q &= 2H'(1/2) - H'(p_q) - H'(1 - p_q) \\ I_e &= 2H'(1/2) - H'(p_e) - H'(1 - p_e) \end{aligned} \tag{20}$$

where

$$H'(p) := -p \ln(p) \tag{21}$$

and

$$p_q := (1 + ih_q)/2 \quad (i = \sqrt{-1}) \qquad p_e := (1 + h_e)/2. \tag{22}$$

The main theorem of [1] states that the overall change of information within an ideal gnostical cycle is negative and that the loss of information of each other closed gnostical cycle defined by the same datum exceeds the loss of information within the ideal gnostical cycle. It has been also proved that the square fidelity (f^2) is proportional to the source of field of I_q over the interval of quantifying irrelevance h_q and the inverse square fidelity (f^{-2}) is also proportional to the source of field of I_e over the interval of estimating irrelevance h_e.

The aim of this paper is to make use of the gnostical theory of individual data for an attempt to develop a gnostical theory of data samples and its application in solutions of fundamental tasks of data treatment.

2. Data composition

A crucial point of treatment of uncertain data is the way of their composition which should suppress the uncertainty as much as possible.

Definitions. A data sample, denoted $Z(z_0, n)$ or shortly Z, is an n-tuple of real data $z_1, \ldots, z_n$ the ideal value of which is z_0. A function of all $z_i \in Z(z_0, n)\,(i=1, \ldots, n)$ and of z_0 will be called a *characteristic* of the sample Z. A composite vector of the data sample Z is a quantity $u_{cq}^T=(z_0 \operatorname{ch} \Omega_c, z_0 \operatorname{sh} \Omega_c)$ or $\mathbf{u}_{ce}^T=(r_c \cos \omega_c, r_c \sin \omega_c)$ where Ω_c, ω_c and r_c are characteristics of Z.

Axiom 2 (composition rule). Let $Z(z_0, n)$ be a data sample. Then ∎

$$\mathbf{K}_q^2(\Omega_c) = \frac{1}{w_q^2}\sum_1^n \mathbf{K}_q^2(\Omega_i)$$
$$\mathbf{K}_e^2(\omega_c) = \frac{1}{w_e^2}\sum_1^n \mathbf{K}_e^2(\omega_i) \tag{23}$$

where

$$w_q^2 = \operatorname{Det}\left\{\sum_1^n \mathbf{K}_q^2(\Omega_i)\right\}$$
$$w_e^2 = \operatorname{Det}\left\{\sum_1^n \mathbf{K}_e^2(\omega_i)\right\}. \tag{24}$$
∎

Normalizing weights w_q and w_e are thus also characteristics of the data sample Z as well as both matrices $\mathbf{K}_q^2(\Omega_c)$ and $\mathbf{K}_e^2(\omega_c)$ with their components which will be denoted by $1/f_c$, h_{qc}, f_c and h_{ec}.

Theorem 1. Let Z be a data sample and Ω_c, ω_c, w_q and w_e its characteristics (23) and (24). Then

$$2\Omega_c = \operatorname{arcth}\frac{\sum_1^n h_{qi}}{\sum_1^n f_i^{-1}} \qquad 2\omega_c = \operatorname{arctg}\frac{\sum_1^n h_{ei}}{\sum_1^n f_i} \tag{25}$$

$$w_q = n\sqrt{1+\frac{1}{n^2}\sum_{i>j}^n\left(\frac{z_i}{z_j}-\frac{z_j}{z_i}\right)^2}$$
$$w_e = n\sqrt{1-\frac{1}{n^2}\sum_{i>j}^n f_i f_j\left(\frac{z_i}{z_j}-\frac{z_j}{z_i}\right)^2} \tag{26}$$

where $f_i = f(2\omega_i)$, $h_{qi} = h_q(2\Omega_i)$ and $h_{ei} = h_e(2\omega_i)$. ∎

Proof. By substitution of (14)–(18) into (23) and (24). ∎

Corollary 1.1. Let $n=2$. Then

$$\Omega_c = (\Omega_1+\Omega_2)/2 \qquad \omega_c = (\omega_1+\omega_2)/2 \tag{27}$$

$$w_q = \sqrt{2+2\operatorname{ch}(\Omega_1-\Omega_2)} \qquad w_e = \sqrt{2+2\cos(\omega_1-\omega_2)}. \tag{28}$$
∎

Corollary 1.2. Let $n>1$. Then

$$w_q = \sqrt{\sum_{i,j}^n \operatorname{ch} 2(\Omega_i-\Omega_j)} \qquad w_e = \sqrt{\sum_{i,j}^n \cos 2(\omega_i-\omega_j)}. \tag{28}$$
∎

It has been shown in [1] that the sums and differences of parameters Ω_i and Ω_j (ω_i and ω_j) are characteristics of dissimilarity of two data z_i and z_j (belonging to the same z_0). Relations (25)–(28) demonstrate that they together with z_0 fully determine the mentioned characteristics of a data sample.

Corollary 1.3. Let $Z(z_0, n) := Z(z_0, n') * Z(z_0, n'')$ be a concatenation of data samples Z' and Z'' $(n=n'+n'')$. Let $\Omega_c, \omega_c, w_q, w_e, \Omega_c', \omega_c', w_q', w_e'$ and $\Omega_c'', \omega_c'', w_q'', w_e''$ are characteristics of the data samples Z, Z' and Z'', respectively. Then

$$\mathbf{K}_q^2(\Omega_c) = \frac{w_q'}{w_q}\mathbf{K}_q^2(\Omega_c') + \frac{w_q''}{w_q}\mathbf{K}_q^2(\Omega_c'')$$
$$\mathbf{K}_e^2(\omega_c) = \frac{w_e'}{w_e}\mathbf{K}_e^2(\omega_c') + \frac{w_e''}{w_e}\mathbf{K}_e^2(\omega_c'') \tag{29}$$

and

$$w_q = \sqrt{w_q'^2 + w_q''^2 + 2w_q'w_q''\operatorname{ch} 2(\Omega'-\Omega'')}$$
$$w_e = \sqrt{w_e'^2 + w_e''^2 + 2w_e'w_e''\cos 2(\omega'-\omega'')}. \tag{30}$$

Corollary 1.4. Let the assumptions of Corollary 1.3 hold. Then

$$\sqrt{w_e'^2 + w_e''^2} \leq w_e \leq w_e' + w_e'' \leq n' + n'' = n. \tag{31}$$

So each (even an outlying) real datum or a sample of such data is useful in the sense that it increases the weight w_e of the concatenated sample. But the maximum possible increase of the weight is limited by the increase of the total number of data. Such an increase may be obtained only in the case when all data are identical.

Corollary 1.5. Let

$$\bar{q} := \frac{1}{n}\sum_{i=1}^n q_i \tag{32}$$

denote the arithmetical mean of some quantities q_i.

Let $f_c := f(2\omega_c) = \cos 2\,\omega_c$, let $1/f_c' := \mathrm{ch}\,2\,\Omega_c$.

Then

$$1/f_c' \leqq \overline{1/f} \tag{33}$$

and

$$f_c \geqq \bar{f}. \quad ▮ \tag{34}$$

The composition rule (23) is thus "better" than the arithmetical mean of the composed quantities in the sense that it yields greater fidelity.

Corollary 1.6. Let $\overline{f^{-1}}$, $\overline{h_q}$, $\bar{f}$, $\overline{h_e}$, $\overline{h_q^2}$ and $\overline{h_e^2}$ be arithmetical means like (32). Let

$$C_q(k) := \frac{1}{n-k}\sum_{i=1}^{n-k} h_q(2\Omega_i)h_q(2\Omega_{i+k})$$
$$C_e(k) := \frac{1}{n-k}\sum_{i=1}^{n-k} h_e(2\omega_i)h_e(2\omega_{i+k}). \tag{35}$$

Then

$$w_q = n\sqrt{(\overline{f^{-1}})^2 - (\overline{h_q})^2} = n\sqrt{(\overline{f^{-1}})^2 - \frac{1}{n}\overline{h_q^2} - \frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)C_q(k)} \tag{36}$$

$$w_e = n\sqrt{\bar{f}^2 + (\overline{h_e})^2} = n\sqrt{\bar{f}^2 + \frac{1}{n}\overline{h_e^2} + \frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)C_e(k)}. \quad ▮ \tag{37}$$

The sense of quantities determining both weights in (36) and (37) is worth to be discussed below. All characteristics of data samples introduced here will be called *gnostical* characteristics.

3. Correspondence between gnostical and statistical characteristic

It is interesting to demonstrate a correspondence of gnostical characteristics to statistical parameters of data samples and to show the special conditions under which such a correspondence takes place. We of course deal with a correspondence of numerical characteristics of data samples and not of mathematical models which stay to be quite different.

Definitions. Let us denote

$$d_i = z_i/z_0 - 1 \qquad (i = 1, \ldots, n) \tag{38}$$

and

$$\varepsilon = \max_{z_i \in Z} |d_i|. \tag{39}$$

In the case of a small ε we shall speak of *the case of weak uncertainties.* ▮

Theorem 2. Let (39) hold. Then

$$\left.\begin{matrix} \Omega_c \\ -\omega_c \\ \overline{h_{qc}}/2 \\ -\overline{h_{ec}}/2 \end{matrix}\right\} = \bar{d} + O(\varepsilon^2) \tag{40}$$

$$\left.\begin{matrix} (\overline{f^{-1}} - 1)/2 \\ (1 - \bar{f})/2 \\ \overline{h_q^2}/4 \\ \overline{h_e^2}/4 \end{matrix}\right\} = \overline{d^2} + O(\varepsilon^3) \tag{41}$$

$$\left.\begin{matrix} \frac{1}{4}C_q(k) \\ \frac{1}{4}C_e(k) \end{matrix}\right\} = \frac{1}{n-k}\sum_{i=1}^{n-k} d_i d_{i+k} + O(\varepsilon^3) \tag{42}$$

$$w_q = n(1 + 2(\overline{d^2} - (\bar{d})^2) + O(\varepsilon^3) \tag{43}$$
$$w_e = n(1 - 2(\overline{d^2} - (\bar{d})^2) + O(\varepsilon^3). \quad ▮$$

Proof. By Taylor's expansion of gnostical characteristics. ▮

Under condition of weak uncertainties all characteristics Ω_c, $-\omega_c$, $\overline{h_{qc}}/2$ and $-\overline{h_{ec}}/2$ approach thus the mean relative error of the data, all characteristics $(\overline{f^{-1}} - 1)/2$, $(1 - \bar{f})/2$, $\overline{h_q^2}/4$ and $\overline{h_e^2}/4$ approach the mean square relative error of the data, the quantities $C_q(k)/4$ and $C_e(k)/4$ the relative covariance of a part of the data ordered into $n-k$ pairs, the weights w_q and w_e approaching the number n of the data. In the case of weak uncertainties all analyzed gnostical characteristics have thus a close connection to the basic statistical characteristics of the data sample. We shall use these characteristics to obtain estimates of the ideal quantity z_0. We may therefore expect that in the case of weak uncertainties the gnostical estimates of the quantity z_0 will approach the arithmetical mean $\bar{z}$ of the data.

However, if uncertainty is not weak then the gnostical characteristics differ substantially from the statistical ones and one from another.

4. Sensitivity and robustness of gnostical characteristics of a data sample

Definition. Let $Z(z_0, n-1)$ be a data sample of data $z_1, \ldots, z_{n-1}$ and $Z(z_0, n)$ a data sample obtained from the former one by concatenating of it with a new datum z_n. Let g_{n-1} and g_n denote a gnostical characteristic of both samples. The characteristic g_n will be said to exhibit the *sensitivity* α, β with respect to the datum z_n if

$$\lim_{z_n \to 0} \frac{g_n - g_{n-1}}{z_n^{-\alpha}} = \text{const.} \qquad \lim_{z_n \to \infty}{}' \frac{g_n - g_{n-1}}{z_n^{\beta}} = \text{const}'. \quad \blacksquare \quad (44)$$

The negative value of the parameters α or β will thus characterize a feature inverse to sensitivity, the robustness of the characteristic.

For characteristics symmetrical with respect to the quantities z_i and z_i^{-1}, both parameters α and β naturally coincide.

Theorem 3. Sensitivity of gnostical characteristics of data sample $Z(z_0, n)$ is given by Table 1:

Table 1. Sensitivity of some gnostical characteristics of a data sample with respect to a datum z_n

Characteristic	$\overline{f^{-2}}$	$\overline{h_q^2}$	$\overline{f^{-1}}$	$\overline{h_q}$	w_q	C_q	C_e	w_e	$\overline{h_e}$	$\overline{h_e^2}$	$\bar{f}$	$\overline{f^2}$
Sensitivity ($\alpha=\beta$)	4	4	2	2	2	2	0	0	0	0	−2	−4

Proof. By verification of (44) using formulae (15)–(18), (24) and (35). ∎

There exists thus a large scale of sensitivity of gnostical characteristics of a data sample. It make it possible to choose a proper characteristic for a given task: Value $\alpha=\beta=4$ means the highest sensitivity to outlying data and the lowest relative sensitivity to inlyers, with $\alpha=\beta=-4$ we obtain an opposite case.

5. Actual estimation

Gnostical characteristics of data samples are functions of the unknown ideal value z_0 which is the object of estimation. We know already the ideal estimating procedure but to realize the ideal gnostical cycle we would need also the quantity z_0. But it is possible to estimate this quantity by an extremalization of a gnostical characteristic. The estimate would then have a feature approaching that of the ideal gnostical cycle of individual data or a new extremal feature connected with mutual relations between data.

We shall consider only eight types of estimates taking into account the equivalences

$$(\overline{h_q}=0) \Leftrightarrow (\mathrm{K}_q^2(\Omega_c)=1) \Leftrightarrow (\Omega_c=0) \Leftrightarrow (h_{qc}=0) \Leftrightarrow (1/f_c'=1) \quad (45)$$

$$(\overline{h_e}=0) \Leftrightarrow (\mathrm{K}_e^2(\omega_c)=1) \Leftrightarrow (\omega_c=0) \Leftrightarrow (h_{ec}=0) \Leftrightarrow (f_c=1). \quad (46)$$

They characterize two types of symmetry of a data sample. These symmetries are multiplicative: So — for example — the numbers 1/2 and 2 are symmetrically positioned with respect to 1.

Definitions. Let $Z(z_0, n)$ be a data sample. Then the estimate of the ideal value z_0 of the type J ($J = qI, qC, qF, qS, eC, eS, eF, eI$) will be denoted by z_J. The estimates will be evaluated to satisfy conditions specified in Table 2:

Table 2. Optimality conditions for the actual estimation of the ideal value z_0

Type of the estimate	Condition of the optimality	Required feature
$\bar{z}_0 = z_{qI}$	$\frac{d\overline{h_q^2}}{dz_0} = 0$	Minimal changes of information due to quantification
$\bar{z}_0 = z_{qC}$	$\frac{d}{dz_0}\left((\overline{h_q})^2 - \frac{1}{n}\overline{h_q^2}\right) = 0$	Minimum of the absolute value of the sum of covariances
$\bar{z}_0 = z_{qF}$	$\frac{d\overline{f^{-1}}}{dz_0} = 0$	Minimal mean inverse fidelity of the sample
$\bar{z}_0 = z_{qS}$	$\overline{h_q} = 0$	Symmetry of the data sample
$\bar{z}_0 = z_{eS}$	$\overline{h_e} = 0$	Symmetry of the data sample
$\bar{z}_0 = z_{eC}$	$\frac{d}{dz_0}\left((\overline{h_e})^2 - \frac{1}{n}\overline{h_e^2}\right) = 0$	Minimum of the absolute value of the sum of covariances
$\bar{z}_0 = z_{eF}$	$\frac{d\bar{f}}{dz_0} = 0$	Maximal mean fidelity of the sample
$\bar{z}_0 = z_{eI}$	$\frac{d\overline{h_e^2}}{dz_0} = 0$	Maximal changes of information due to estimation

∎

Theorem 4. Let $Z(z_0, n)$ be a data sample. Then the estimate z_J ($J = qI, qC, qF, qS, eS, eC, eF, eI$) of the ideal value z_0 optimal in the sense defined by Table 2 is given by a solution of the equations

$$A\colon\ z_J = \sqrt[4]{\frac{\sum_i^n f_i^m z_i^2}{\sum_i^n f_i^m z_i^{-2}}} \qquad \text{or} \qquad B\colon\ z_J = \sqrt[4]{\frac{\sum_{i \neq j}^n f_j^k f_i^m z_i^2}{\sum_{i \neq j}^n f_j^k f_i^m z_i^{-2}}} \tag{47}$$

(where $f_i = 2/((z_i/z_J)^2 + (z_J/z_i)^2)$) specified by Table 3:

Table 3. Specifications of the equation of gnostical estimates of z_0

Type of the estimate z_J	Equation A/B	m	k	Sensitivity of z_J^4: Numerator α	Numerator β	Denominator α	Denominator β
z_{qI}	A	−1	–	0	4	4	0
z_{qC}	B	0	−1	−2	2	2	−2
z_{qF}	A	0	–	−2	2	2	−2
z_{qS}	A	0	–	−2	2	2	−2
z_{eS}	A	1	–	−4	0	0	−4
z_{eC}	B	1	2	−4	0	0	−4
z_{eF}	A	2	–	−6	−2	−2	−6
z_{eI}	A	3	–	−8	−4	−4	−8

∎

Proof. Equations A for z_{qS} and z_{eS} result directly from (16) and (17) substituted into the condition equalling the arithmetical mean of quantifying or estimating irrelevance to zero. The equations for the estimates z_{qI}, z_{qF}, z_{eF} and z_{eI} are equivalent to the equation

$$\frac{d}{dz_J} \sum_i^n f^{m-1}(z_i/z_J) = 0 \tag{48}$$

where $m = 0, 2, 3$ and 4, respectively, as follows from the definitions of extremalized quantities (Table 2). Equations for z_{qC} and z_{eC} may be obtained also from the condition given by Table 2 using the equivalences

$$(\overline{h_q})^2 - \frac{1}{n}\overline{h_q^2} = \sum_{i \neq j}^n h_{qi} h_{qj}/n^2 \qquad (\overline{h_e})^2 - \frac{1}{n}\overline{h_e^2} = \sum_{i \neq j}^n h_{ei} h_{ej}/n^2 \tag{49}$$

resulting from (35)–(37). ∎

Corollary 4.1. It holds

$$z_J = \bar{z} + O(\overline{d^2}) \tag{50}$$

for all $J = qI, qC, qF, qS, eS, eC, eF$, and eI. ∎

In the case of weak uncertainties all gnostical estimates z_0 approach thus the arithmetical mean $\bar{z}$. But if the uncertainty is not weak then their properties are different as shown in Table 3. This enables us to choose the sensitivity or robustness of the estimate z_J with respect to outliers or inliers to match the requirements of each particular task of data treatment.

The gnostical estimates z_J are not necessarily unique. If a data sample consists of several more or less separate "clusters" then each of them may have its own "location parameter" z_J.

6. Distribution and density functions of a data sample

The quantities p_e and $1 - p_e$ (22) appear as parameters of the estimating change of information I_e (20). Thus they play a role analogous to probability, although we do not consider a probabilistic model.

Theorem 5. Let z_i be a datum. Let the quantities $z \in R_+$ and $\Omega_i \in R_1$ take such values that

$$z_i = z e^{\Omega_i} = \text{const} \tag{51}$$

Then the quantity $1 - p_{ei}(\Omega_i) = 1/(1 + e^{-4\Omega_i})$ is a distribution function of the quantity Ω_i on R_1. The quantity $p_{ei}(z) = z^4/(z_i^4 + z^4)$ is a distribution function of the quantity z on R_+. The corresponding density functions are

$$\frac{d(1 - p_{ei})}{d\Omega_i} = f_i^2 \qquad \frac{dp_{ei}}{dz} = \frac{1}{z} f_i^2 \tag{52}$$

where

$$f_i = 1/\text{ch}2\,\Omega_i \equiv 2/(z_i^2 z^{-2} + z^2 z_i^{-2}). \qquad ∎ \tag{53}$$

Proof. Both functions $1 - p_{ei}$ and p_{ei} change on their definition intervals from 0_+ to 1_- monotonously. Their explicit form results from (53) and from the formulae of the estimating irrelevance (17). ∎

Corollary 5.1. Let $B \in R_+$ be an interval

$$B := \{z : z_1 \leq z \leq z_2\} \tag{54}$$

for each pair $z_2 \geq z_1 > 0$.

Then the quantity

$$P_i(B) = p_{ei}(z_2) - p_{ei}(z_1) \tag{55}$$

induces a finite measure on Borel sets of R_+ (given z_i). ∎

Corollary 5.2. Let $B' := \{z' : 0 < z' \leq z\}$ and $B'' := \{z'' : z \leq z'' < \infty\}$. Then

$$p_{ei}(z) = p_i(B') \quad \text{and} \quad 1 - p_{ei} = p_i(B''). \qquad ∎ \tag{56}$$

If $z=z_i$ then $P_i(B')=P_i(B'')=1/2$. After a single result of quantification equalling to $z_i \in Z(z_0, n)$ has been obtained, we may guess with the same degree of confidence that the unknown quantity z_0 satisfies $z_0 \geqq z_i$ as $z_0 \leqq z_i$. The confidence that the ratio z_0/z_i takes a particular value may be quantified by the quantity $I_{ei}=H'(p_{ei}(z_0))+ + H'(1-p_{ei}(z_0))$.

Corollary 5.3. Let B be an interval (54) and $P_i(B)$ its measure (55). Then

$$P_i(B) = \frac{z_2-z_1}{z_i} + 0\left(\max_{k=1,2}\left(\frac{z_k-z_i}{z_i}\right)^2\right). \quad ▮ \quad (57)$$

For a couple of quantities z_1 and z_2 sufficiently close to the datum z_i the measure P_i of the interval B approaches thus the Lebesque's measure of B divided by z_i.

Corollary 5.4. Let $Z(z_0, n)$ be a data sample. Let (51) hold for quantities z and Ω'_c. Let z_{eI} be the solution of equation (47A) with $m=3$. Then the estimate z_{eI} maximizes the function

$$f^2(2\Omega'_c) = \frac{1}{n}\sum_{1}^{n} f_i^2 \quad (58)$$

of a variable z (51). ~~where~~

~~$$\Omega'_c = -\operatorname{arcth}(\operatorname{tg} \omega_c) \quad (59)$$~~

~~and the quantity ω_c is determined by composition rule (23) after substitution of (14), (15) and (17) with $\xi_i = z_i/z$.~~ ▮

Definition. The function $f^2(2\Omega'_c)$ (58) is *the density function of the data sample Z.* ▮

7. Correspondence between gnostical theory of data samples and the information theory

Corollary 5.5. Let ρ be a binary random quantity which took, in an experiment consisting of N_1+N_0 trials, N_1 times the value "1" and N_0 times the value N_0. Let p_1 be the probability of the result "1" and $\tilde{p}_1 = N_1/(N_1+N_0)$ the frequency estimate of this probability. Then

$$\tilde{p}_1 = p_{ec} \quad (60)$$

where $p_{ec} = 1/(1+e^{4\Omega_c})$ and $\Omega_c = \frac{1}{4}\ln(N_1/N_0)$. ▮

In this case the quantity p_{ec} may thus be interpreted as an estimate of probability and the quantity I_e (20) as an estimate of the Shannon's information obtained by the experiment.

A correspondence exists also with the entropy H_2 which is a special kind of generalized entropies introduced by Rényi [2] (1960). For the case considered above this entropy is

$$H_{2e} = cp(1-p) \quad (61)$$

where c is a constant. This type of entropy has been studied by Onicescu [3], Perez [4] and by Vajda [5] who pointed out interesting properties of this entropy similar to Shannon's entropy. Let us substitute the estimate p_{ec} (60) instead of probability p into H_2. We obtain

$$H_{2e} = \frac{c}{4} f^2(2\Omega_c). \quad (62)$$

It has been shown in [1] that the square f^2 of the fidelity is proportional to a source of field of changes of information I_q (20) over the interval of quantifying irrelevance h_q. We have seen above that $f^2(2\Omega_c)$ may be interpreted as a density function of a data sample. There exist thus two gnostical interpretations of the entropy H_2. Moreover, we obtained practically applicable estimating formulae for the amount of entropy H_{2e} for a data sample or even for a single datum $z_i = z_0 \exp(\Omega_i)$:

$$H_{2ei} = \frac{c}{4} f^2(2\Omega_i). \quad (63)$$

Properties of the square fidelity f^2 (alias source of information I_q, alias density function of a data sample, alias entropy H_{2e}) may be demonstrated by the following practical examples.

8. Examples

In the following an extended data model

$$z_i = z_0 \exp(\Omega_i | s) \quad (64)$$

will be used. The quantity s (the "scale parameter") characterizes the width of data sample. It can be estimated directly from data.

Example 1. Systolic blood pressure of 24 healthy women of a fertile age are summarized in Table 4. They are distributed randomly into two data samples A and B containing 12 data.

Table 4. Real data for Example 1

Data sample	Systolic blood pressure z_i (mmHg)											
A	115	120	145	120	150	135	125	120	125	120	120	140
B	130	130	120	145	135	140	110	120	135	120	110	115

Empirical distribution functions of both samples differ substantially although both samples have (randomly) the same arithmetical mean. Table 5 presents some values of density functions $\overline{f^2}$ of both data samples.

The gnostical estimates z_{el} equal to the pressure maximizing the density functions. It has been numerically obtained as $z_{elA}=122.5$ and $z_{elB}=123.8$. An interesting question might be what will happen with the estimates z_{el} when another (the 13th) datum will be added to the samples. The dependence of both estimates on the value of a 13th datum z_{13} is shown by Table 6. For a comparison, the values of the arithmetical means of all 13 data are also given in Table 6.

Table 5. Density functions of both data samples of Example 1

Systolic blood pressure (mmHg)	Mean square of fidelity f^2		
	$\overline{f_A^2}$	$\overline{f_{A\cup B}^2}$	$\overline{f_B^2}$
80	0.003	0.004	0.002
90	0.048	0.046	0.024
100	0.504	0.410	0.243
110	2.155	0.824	1.508
120	3.369	3.411	3.575
130	3.291	3.114	3.028
140	2.109	2.123	2.160
150	0.733	1.116	1.434
160	0.187	0.429	0.619
170	0.046	0.134	0.198
180	0.012	0.040	0.059
190	0.003	0.013	0.018
200	0.001	0.004	0.006

Table 6. Dependence of the estimates z_{elA} and z_{elB} of an ideal quantity z_0 on a thirteenth additional datum z_{13}

Additional datum z_{13}	Gnostical estimates		Arithmetical means
	z_{elA}	z_{elB}	$z_A = z_B$
60	124.6	124.8	121.2
70	124.0	124.2	121.9
80	123.6	123.7	122.7
90	123.4	123.6	123.5
100	123.7	123.8	124.2
110	124.3	124.5	125.0
120	125.1	125.3	125.8
130	125.9	126.2	126.5
140	126.7	126.9	127.3
150	127.2	127.5	128.1
160	127.6	127.9	128.8
170	127.8	128.1	129.6
180	127.8	128.1	130.4
190	127.7	128.0	131.2
200	127.6	127.9	131.9

Because of the nonlinearity of the estimates z_{el} the influence of outlying values z_{13} on them is suppressed. For $z_{13}\to 0$ as well as for $z_{13}\to\infty$ the estimates z_{el} may be shown to reach their values which correspond to the original samples of 12 data. The extremal values of data are thus fully ignored.

In Table 6 two interesting points appear, those points where

$$\frac{dz_{el}}{dz_{13}}=0$$

holds. Let us denote these "critical" points (z'_{13}, z'_{el}) and (z''_{13}, z''_{el}). Their numerical values have been evaluated and summarized in Table 7.

Table 7. Critical points of both data samples of Example 1

Data Sample	Estimate z_{el}	Critical points (mmHg)			
		z'_{13}	z'_{el}	z''_{el}	z''_{13}
A	122.5	114.4	121.6	123.4	130.8
B	123.8	114.3	120.9	127.9	134.9

These figures may be used to consider the sensitivity of the estimate z_{el} with respect to a new single datum having an arbitrary value. The estimate z_{el} cannot appear to be outside the interval (z'_{el}, z''_{el}) which can be taken as a tolerance interval. The quantities z'_{13} and z''_{13} have also an interesting function, they separate an interval (z'_{13}, z''_{13}) where the reaction of z_{el} to an increase of the added value z_{13} is "natural" (rising), from intervals $(0, z'_{13})$ and (z''_{13}, ∞) with an "unnatural" reaction of the characteristic z_{el} (falling). It makes it possible to test the "membership" of z_{13} to the data sample.

Comparison of gnostical characteristics of both data samples shows their agreements in spite of the difference of empirical distributions of both samples.

Example 2. In a group of randomly collected men the following weights have been observed:

Table 8. Data sample for Example 2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Weights (kg)	89	84	78	78	66	90	78	87	74	75	87	91	65	75	135

This sample contains an "outlier" z_{15}. It is interesting to demonstrate its influence on the gnostical characteristics of the data sample. Let us denote the data sample of all 15 data Z and the sample obtained from only 14 data without the z_{15} as Z'. The density functions of both samples are shown in Table 9:

Table 9. Density functions of data samples Z and Z' of Example 2 for $s_{Z'} = 0.228$

Weight (kg)	Mean square of fidelity f^2	
	f_Z^2	$f_{Z'}^2$
40	0.007	0.000
50	0.096	0.027
60	0.605	0.490
70	1.545	1.717
80	2.086	2.653
90	1.706	2.135
100	0.915	0.751
110	0.422	0.169
120	0.266	0.038
130	0.245	0.009
140	0.213	0.003
150	0.147	0.001
160	0.087	0.000
170	0.047	0.000

Influence of the "outlier" z_{15} on the gnostical characteristics is demonstrated also by Table 10:

Table 10. Gnostical characteristics of both data samples Z and Z' of Example 2

Data sample	Gnostical estimates		Arithmetical means	Critical points			
	s	z_{eI}	$\bar{z}$	z'_{n+1}	z'_{eI}	z''_{eI}	z''_{n+1}
Z	0.335	80.6	83.5	71.1	79.5	81.7	91.3
Z'	0.228	80.3	79.8	72.6	78.5	83.2	89.7

The density function of the "censored" data sample Z' appeared to be sharper than that of the original complete sample Z. In spite of this the gnostical characteristics in Table 10 changed in a less degree than the arithmetical mean did. It is interesting that the quantity z_{15} is far behind the critical point z''_{n+1} in *both* cases, it is an outlier even from the "point of view" of the data sample Z which contains it. Two small data (z_5 and z_{13}) appeared to be under the critical point z'_{n+1}, they are also "not typical".

Acknowledgements

It is a pleasant duty of the author to express his sincere gratefulness to J. Šindelář for his efficient help which enabled to prepare this publication.

References

1. *Kovanic, P.*, Gnostical theory of individual data, Problems of Control and Information Theory **13**, No. 4 (1984).
2. *Rényi, A.*, On measures of entropy and information, Proc. 4-th Berkeley Symp. on Math. Stat., Part I, Berkeley 1961.
3. *Onicescu, O.*, Energie informationnelle, C. R. Acad. Sci. Paris, série A, 28 nov. 1966, **263**, 841–842.
4. *Perez, A.*, Sur l'energie informationnelle de M. Octav Onicescu, Rev. Roum. Math. Pures et Appl., **XII** (1967), No. 9, 1341–1347.
5. *Vajda, I.*, Оценки минимальной вероятности ошибки при проверке конечного или счетного числа гипотез. Проблемы передачи информации, **4** (1968), 9.

Гностическая теория малых наборов действительных данных

П. КОВАНИЦ

(Прага)

Теория малых наборов действительных данных строится на основе теории отдельных данных, изложенной в предшествующей статье автора. Для складывания отдельных данных принимается простая аксиома, из которой выводятся гностические характеристики набора данных. При слабом влиянии неопределенностей на данные эти характеристики сходятся к статистическим моментам первого и второго порядков. При сильных неопределенностях они от них существенно отличаются, обладая повышенной или пониженной чувствительностью к выделяющимся данным. Выводятся формулы для практического оценивания гностических характеристик, степень чувстви-тельности или робастности которых можно задавать. Указаны гностические формулы для непосредственного оценивания плотности вероятности по данным из малого набора и даны примеры практического применения.

P. Kovanic
Institute of Information Theory and Automation
182 08 Prague 8
Pod vodárenskou věží 4
Czechoslovakia