~~182~~
179

# GNOSTICAL MODELLING OF UNCERTAINTY

Pavel Kovanic[1]

*The gnostical theory of uncertain data is an alternative to statistics that is applicable to the treatment of small samples of strongly disturbed data. Gnostical procedures are therefore efficient tools which are suitable for the analyses of "bad" data having no statistical model as well as for data for which statistical modelling is not reasonable at all (e.g. microeconomic data). The principal axioms and theoretical results of gnostics are briefly exposed.*

## Structure of Uncertain Data

The gnostical theory of uncertain data, exposed in more detail in [1], is proposed as an alternative to statistics. In gnostics, data error bears no relationship to a random process. Each change in the value of a datum has its cause, and it could be rationally explained, if only there were sufficient information. *Data uncertainty is thus a lack of information* and it is therefore highly *subjective*. This theory is based on elementary axioms:

- data structure – Cartesian product of two Abel groups,
- uncertainties – analytical operators over the data structure,
- data composition law – additive with respect to data entropy.

The first axiom can be shown to summarize the well-known axiomatics of the measurement theory. It can be therefore reformulated in a more popular way: "Both true data value and its disturbance are (**practically!**) measurable." This axiom together with the second one (which makes the model unique) enable the derivation of the *gnostical theory of individual uncertainty*. The third axiom makes use of additivity of the thermodynamic entropy which is introduced here by a plausible "Gedankenexperiment". The additivity of the entropy makes the data composition non-additive with respect to data. This - together with nonlinearity of the error and data weight characteristics - results in the remarkable robustness of gnostical estimates.

[1]Institute of Information Theory and Automation of the Czech Academy of Sciences, P.O. Box 18, 182 08 Prague, Czech Republic

## Principal Gnostical Results on Individual Uncertainty

Let $Z_0$ denotes the true (unknown) data value and $Z_i$ its $i-$th observation. Define an auxilliary quantity

$$q_i(Z_0, s) = (Z_i/Z_0)^{2/s} \tag{1}$$

(where $s$ is a positive scale parameter) for use in the calculation of the *fidelity*

$$f_i(Z_0, s) = 2/(1/q_i(Z_0, s) + q_i(Z_0, s)) \tag{2}$$

and the *irrelevance*

$$h_i(Z_0, s) = (1/q_i(Z_0, s) - q_i(Z_0, s))/(1/q_i(Z_0, s) + q_i(Z_0, s)). \tag{3}$$

Within the framework of the gnostical theory, irrelevance plays the role of the distance between $Z_0$ and $Z_i$ (the "observation error") and the fidelity is the weight (or "trustworthiness") of the datum $Z_i$. The *distribution function* which describes the uncertainty of the individual datum $Z_i$ is then

$$p_i(Z_0, s) = (1 + h_i(Z_0, s))/2. \tag{4}$$

Now for a real $p \quad (0 < p < 1)$ define the following real functions:

$$H(p) = -p * \ln(p) - (1 - p) * \ln(1 - p) \tag{5}$$

$$I(p) = H(1/2) - H(p). \tag{6}$$

The quantity $I(p_i(Z_0, s))$, in gnostical theory is used to evaluate the *information loss* due to the contamination that caused the datum value $Z_i$ be observed instead of the true value $Z_0$. The form of $H(p)$ is reminiscent of Shannon's information of a binary system with probabilities $p$ and $1 - p$. However, we have *not* assumed a probabilistic model. We are considering only one datum and the possible values which it could assume on observation. We shall therefore speak of the *expectation* rather than of the probability of a particular value $Z_0$. Having observed $Z_i$, we evaluate our expectation of the event "the true value is $Z_0$" by $p_i(Z_0, s)$. (The unknown parameter, $s$ , is estimated by gnostical procedures for each data sample).

Using the classification of error magnitudes as set out in Table 1, some of the principal features of individual uncertainty are illustrated in Table 2 and explained as follows:

- The *first* column shows that under conditions of very small errors, the observation error is evaluated by the linear function of the difference between true and observed data and all data are given the same weight (1.0). Having observed $Z_i$ we expect that 50% of the future observations of $Z_0$ will be less than and that 50% will be greater than $Z_i$ and that the distribution will be the step function. Neither the entropy nor the information is

| Error magnitude | Symbol | Condition |
|---|---|---|
| Very small | VS | $0.99 \leq Z_i/Z_0 \leq 1.01$ |
| Small | S | $0.97 \leq Z_i/Z_0 \leq 1.03$ |
| Big | B | $0 < Z_i/Z_0 < \infty$ |
| Limit | L | $Z_i/Z_0 \to 0$ or $Z_i/Z_0 \to \infty$ |

Table 1: Classification of Relative Data Errors.

| Quantity name | Error Magnitude | | | |
|---|---|---|---|---|
| | VS | S | B | L |
| Error | $2 * \frac{Z_i - Z_0}{Z_0}$ | $2 * \frac{Z_i - Z_0}{Z_0}$ | $h_i(Z_0, s)$ (15) | $\to \pm 1$ |
| Weight | 1 | $1 - 2 * (\frac{Z_i - Z_0}{Z_0})^2$ | $f_i(Z_0, s)$ (14) | $\to 0_+$ |
| Entropy ch. | 0 | $-2 * (\frac{Z_i - Z_0}{Z_0})^2$ | $f_i(Z_0, s) - 1$ | $\to -1_+$ |
| Expectation | 1/2 | $1/2 - \frac{Z_i - Z_0}{Z_0}$ | $p_i(Z_0, s)$ (16) | $\to 0_+$ or $1_-$ |
| Inform. loss | 0 | $2 * (\frac{Z_i - Z_0}{Z_0})^2$ | $I(p_i)$ (19) | $\to \ln 2$ |

Table 2: Main Gnostical Characteristics of Data Uncertainty.

affected by the data contamination. The result is that in the case of very small relative contamination of data, the outcome of gnostical procedures will approach those obtained by statistical methodology.

- The *second* column, describing conditions where small errors exist, presents a linear approximation to the distribution function of expectation and a quadratic dependence of the data weight on the data error. There are non-zero, quadratic approximations for the entropy change and information loss, but their sum is zero. This means that the two changes are offsetting. (An analogy exists in information theory, where the change of Shannon's information differs from the change in Boltzmann's entropy only by the sign. However, the gnostical formula of entropy evaluates the change of thermodynamic entropy, not of the statistical but of the Clausius type.) This column shows why the least squares method frequently yields good results: by minimizing squared errors, we minimize information losses. On the other hand, this holds *only for small errors*.

- The *third* column displays the general gnostical formulae which are valid for the estimation of arbitrarily contaminated data.

- The lesson resulting from the *fourth* column has also both theoretical and practical importance. Unlike the statistical characteristics of data errors, all gnostical characteristics are

bounded with respect to limit changes of the data error. This is why gnostical procedures are remarkably robust with respect to outliers.

The formulae which have been presented are not 'ad hoc' definitions. They were derived by consistent mathematical reasoning based on the two algebraic gnostical axioms. The theoretical results of individual data contamination can be summarized in the following way. They present:

- New formulae for evaluation of data error, entropy increase, and for information loss caused by contamination,
- A new formula for computing data weight,
- An entropy $\Longleftrightarrow$ information conversion law according to which the compensation of changes in both quantities takes place on the level of their second derivatives,
- A special form for the distribution function of individual uncertainty,
- Variation theorems for geodesic lines (circular paths constituting the ideal gnostical cycle) proving its optimality.

### Applications of Gnostics

Gnostical theory was developed to meet requirements of practice for the software suitable for tasks where a statistical model is not known or even not reasonable. This is why in parallel with the development of the theory, many experimental applications have been testing. It appeared, that gnostical programs are worth attention in reliability and survival studies, in analyses of geological, chemical, medical, technological data and in all other fields where the availability of data is limited and where the data uncertainty is not negligible. One of the most prospective field for gnostics is the economy [2].

## References


[1] Pavel Kovanic, A New Theoretical and Algorithmical Basis for Estimation, Identification and Control, *Automatica*, 22:657–674.

[2] Kovanic P., Humber M.B., A New Paradigm for Econometrics, The Third International Workshop on Artificial Intelligence in Economics and Management, Portland, Oregon, U.S.A., August 25–27, 1993