AIEM3 - The Third Intern. Workshop on Artificial Intelligence in Economics and Management, Aug. 25-27, 1993 Portland, Oregon

A NEW PARADIGM FOR ECONOMETRICS

Pavel Kovanic* Marcel B. Humber†

June 7, 1993

Abstract

Key Words: Econometrics, Gnostics, Non–Euclidean Geometry, Finance

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 analysis of economic data. The principal axioms of gnostics are briefly exposed and simple examples using financial data from three industries are presented to demonstrate the efficiency of this new methodology.

Do Econometricians Need an Alternative to Statistics?

Econometrics relies on statistics to collect data, but the use of mathematical statistics as the technology of choice for extracting information from these data may be a questionable procedure. Reference to "statistical methods" in this paper addresses only this latter function.

The historical achievements of statistics, especially in physics, warrant consideration of the methodology in the analysis of economic phenomena. Theories of statistical thermodynamics, chain fission reaction, and neutron slow-down and diffusion yield precise engineering calculations for nuclear reactors. This constitutes one 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 to economics? Because economic processes are substantially different from from physical ones, it is not likely. Benjamin Graham, the father of "fundamental" investment analysis stated [1]:

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

*Institute of Information Theory and Automation of the Czech Academy of Sciences, P.O. Box 18, 182 08 Prague, Czech Republic, Tel. (42) (2) 815-1111.

†Visiting Assistant Professor of Finance, George Washington University, (Department of Finance, Suite 101, Lisner Hall, 2323 G Street, Washington, D.C., 20052, Tel. (202) 994-8205.

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 a 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 others, pertinent critiques of the statistical approach to economic problems include Los [2]:

...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 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 pseudo–scientific 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 [3]:

...Statistics is not science but a kind of prescience, a pseudoscience, a "gedankenscience".[1] Perhaps it's best called an "ersatzscience".[2]

...Uncertainty in nature cannot be modeled (and therefore must not be modeled) by conventional, Kolmogorov[3] probability schemes, because no such scheme may be identified from real data.

...The trouble is that probabilities are not identifiable.

We would not reject statistics because it is a "gedankenscience". The power of mathematics results from that fact that it is a "gedankenscience," due to its independence from the facts of real life. However,

[1] *Der Gedanke...* the thought (in German). Appears frequently in natural sciences in the word *der Gedankenexperiment* – 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.

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

[3] A. N. Kolmogorov – Russian mathematician, who developed in the 1930's the most commonly accepted version of probability theory.

the practical applicability of mathematical or statistical models goes outside the borders of a "gedanken-science". Many processes studied in physics are modelable by "gedanken–experiments" because useful models of their behavior are simple enough to be formulated by humans. We can accurately describe 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. 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 inpropriety of statistical applications to many economic propositions 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 economic events. *Identical distribution* refers to stationarity and repeatability which is also a doubtful characteristic of economic 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.

Mathematical statistics have 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, but taken as a whole they do not go a very long way in solving the dilemma noted by Graham. There may exist a more radical solution: to leave the statistical environment completely, and to try something entirely new.

The Gnostical Paradigm

Introduction

The gnostical theory of uncertain data, exposed in more detail in [4], is proposed as an alternative to statistics. The theory is general, and was developed without relation to a particular field of application. It is a mathematical theory which originated at the intersection of several abstract scientific disciplines and, therefore, it is not very easily explainable (particularly with a limited reference to mathematics). (We recall A. Einstein's requirement that, *an explanation should be as simple as possible but not simpler*). However, the good results obtained with the application of gnostics to economic problems motivates this attempt to call the attention of economists to this useful new tool.

Commutativity: $e_i * e_j = e_j * e_i$ holds for all e's. (The impacts of individual factors do not depend of their order.)

Neutral element: The "observed" value $Z_i = 1$ (100%) is also a possible element of the structure (e.g. as an accurate observation of the true value Z_0 which equals 1).

Invertibility: We also accept $1/e$ as a possible contaminating factor for each e. (The information channel can "amplify" as well as "attenuate", we accept the existence not only of profit but also of losses.).

Using mathematical language we can summarize these natural assumptions as the first gnostical axiom: *The structure of contaminated data is isomorphic with the multiplicative group.*

Effect of Individual Uncertainty

There are two components, e.g. Z_0 and e_i in (1), which together determine the value of the "observed" datum. A report from the agency C to its client B on Mr. A's balance should also consist of two components, the estimate of the balance and an estimate of the error of the report. The theoretical data model must then also have two components. To get a second equation dual to (1) we apply the assumption of invertibility: There exists Z_i' for each Z_i in (1) such that

$$Z_i' = Z_0/e_i \tag{2}$$

holds. We have created the second component by multiplying by the inverse of the contamination factor. Now an important coordinate transformation is applied to demonstrate one of the most striking results of gnostical theory. Introducing the notation

$$e_i = \exp(\Omega_i) \tag{3}$$

and using standard hyperbolic functions $\cosh(\Omega)$ and $\sinh(\Omega)$, we define

$$x_i = Z_0 * \cosh(\Omega_i), \tag{4}$$

$$y_i = Z_0 * \sinh(\Omega_i). \tag{5}$$

We can now find the relationship between the attempts to quantify A's bank balance by C and D:

$$x_j = x_i * \cosh(\Omega_k) + y_i * \sinh(\Omega_k), \tag{6}$$

$$y_j = x_i * \sinh(\Omega_k) + y_i * \cosh(\Omega_k), \tag{7}$$

where

$$\Omega_k = \Omega_j - \Omega_i. \tag{8}$$

It can be easily verified that

$$\sqrt{x_i^2 - y_i^2} = \sqrt{x_j^2 - y_j^2} = Z_0. \tag{9}$$

The startling result is that in spite of the errors induced in the quantification processes of agencies C and D, a simple two coordinate function provides not only the same result, but also an outcome which

In gnostics, data error bears no relationship to a random process. Each change in the value of a data value 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***. Mr. A's bank balance is exact from his perspective, and he can explain all its changes. From the point of view of Mr. B, who has no information concerning Mr. A, the status of the account is uncertain. Since Mr. B has no knowledge of the true value of Mr. A's bank balance, any attempt to estimate it will be inaccurate (contaminated). The amount of contamination can be decreased by hiring an agency, C, which has collected information about Mr. A's professional and commercial activities.

Two terms are introduced here which have a special meaning in gnostical theory. *Quantification* is the (contaminating, error producing) process of measuring a real quantity. *Estimation* is the reverse process over which the true value of the quantity is extracted from the measured (contaminated) data.

One objective of gnostics is to provide a realistic theory applicable to small data samples. It therefore treats finite samples *not* by using rules valid for samples of infinite size but by considering small samples as a composition of individual contaminated data with their highly developed theory.

Structure of Uncertain Economic Data

It will be shown that, under plausible conditions, each individual contamination is influenced by important regularities. This is the subject of the *gnostical theory of individual uncertain data*. Returning to the bank account of Mr. A, whose balance is denoted by Z_0, Mr. B obtains from agency C an "information channel" producing an (inexact) image, Z, of the state of A's account. (Z is thus the "observed" value of the true value Z_0). By treating assets and liabilities separately, both Z_0 and Z are always strictly positive. The error introduced through the information channel is evaluated as a proportional change to Z_0 (therefore multiplicative by application of the positive quantity e.) This is a natural approach since it parallel's the periodic changes to the value of assets which are exposed to interest rates, a profit margin, inflation, taxes, etc. Thus for an i--th contaminated "observation":

$$Z_i = Z_0 * e_i. \tag{1}$$

To further increase the accuracy of his estimate, Mr. B could also have approached another agency, D, which would quantify the account Z_0 as Z_j. Moreover, there can be more than one source of error affecting each agency's observations, each one having a multiplicative effect: e.g. $e_i = e_{i1} * e_{i2}$. Therefore a whole set of N factors, e_i, could exist. We can thus consider not only one, but a set of possible "observed" data – each datum representing a possible value for the asset. Each of the Z_i's is an element of this set. So also are all the contaminating factors e_i, because they can be interpreted as "observed" values Z_i of the asset Z_0 which itself equals 1 (100%). We accept the operation of arithmetical multiplication as defined over this set. A portion of these multiplicative changes can be explained (and are certain), the rest are unexplained. For this simple structure other mathematical properties can be assumed:

Closedness: Each "observed" value of the asset is the asset, each product of contaminating factors e is the contaminating factor.

Associativity: $(e_i * e_j) * e_k = e_i * (e_j * e_k)$ holds for all e's.

equals the unknown true value Z_0 which is seen to be *invariant* to the quantification procedure. This is true because both agencies attempted to quantify *the same* unknown Z_0. It might appear that equation (9) is not very useful, because e.g. agency C knows only the sum

$$x_i + y_i = Z_0 * \exp(\Omega_i) = Z_i \tag{10}$$

but is not able to decompose it into Z_0 and Ω_i; it will be shown how to resolve this problem approximately by using a data sample. There is an unexpected theoretical significance in (4), (5),(6) and (9): *The quantification process can be modelled using the rules of Minkowskian geometry.* Formula (9) evaluates the norm of the vectors (x_i, y_i) and (x_j, y_j) and states that this value is invariant, independent of the data contamination. Under Minkowskian geometry then, formulae (6) and (7) can be interpreted geometrically as an orthogonal rotation of the vector (x_i, y_i). These types of rotation have their own famous name, *Lorentz transformations*. This is the same transformation law that is used in relativistic mechanics to characterize the (nonlinear) dependence of observations on the velocity of the coordinate system.

From this, we conclude that there exists a close *geometrical* similarity between two apparently very different processes: the contamination of economic data, and the laws of movement of relativistic particles. As shown below, this similarity is of an even more profound nature.

The Estimation Process and the Ideal Gnostical Cycle

In the following development, it is important to accept the notion of an arbitrary "observed" output, Z_q, of the quantification process represented by the point $\langle Z_0 * \cosh(\Omega_q), Z_0 * \sinh(\Omega_q)\rangle$ on the plane (x, y) "moving" (being driven by the changing contamination Ω_q.) Its starting point is the unknown errorless point $\langle Z_0, 0\rangle$ and the endpoint is the point $\langle x_i, y_i\rangle$ which in our example corresponds to the report of agency C. The form of the path is identified as an arc of the Minkowskian circle having the radius Z_0. This arc will be called *the quantification path.*

The quantification process thus has its path and the invariant Z_0 is the true value. We can imagine that we are playing a game against Nature: we want to get her secret (Z_0), but she is not willing to give it up free of charge and is penalizing us by contaminating the data, rotating it along a Minkowskian circle. The quantification is the move of Nature; we are given its result, e.g. the value Z_i, theoretically represented by the point $\langle x_i, y_i\rangle$. We now have to choose our move against Nature to minimize the penalty. It is therefore necessary to find *the best possible estimating path* by which we can return to the starting point Z_0. It is natural, hence, to choose that path which will cause this value to become the invariant quantity of the transformation. Any arbitrary point $\langle x_e, y_e\rangle$ on the estimating path satisfies the relation

$$\sqrt{x_e^2 + y_e^2} = Z_i. \tag{11}$$

We thus choose the equation of an "ordinary" (Euclidian) circle as the model of the estimating process.

The goal is to reach the true value Z_0 by the end of the estimation process, closing the quantification loop through estimation. This is possible – at least theoretically – as shown in Figure 1. Nature "moves" the point of interest from $\langle Z_0, 0\rangle$ to $\langle x_i, y_i\rangle$ along the path of a Minkowskian circle: this is the quantification process. Our response, estimation, consists of two parts: starting with the endpoint $\langle x_i, y_i\rangle$ which represents the observation to its mirror image $\langle x_i, -y_i\rangle$ along the Euclidian circle path, and then from

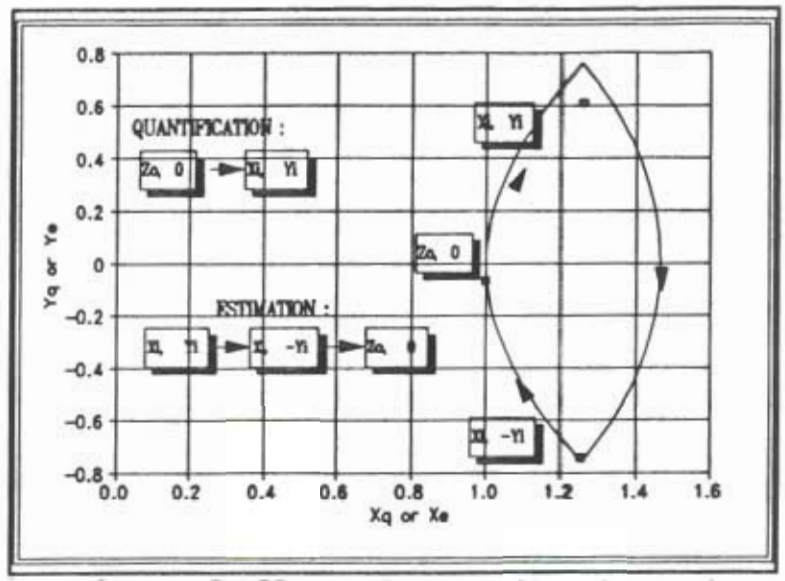

Figure 1: The Ideal Gnostical Cycle.

there to $\langle Z_0, 0\rangle$, the true value. This transformation can never be achieved because we never know precisely how to decompose the "observed" datum Z into its components x and y. This is not the only reason why this cycle is "ideal": it is also the theoretical model of the *best possible manner by which to process the data.* As proved by gnostical theory, the quantifying "move of Nature," using the Minkowskian path *maximizes* the contamination measured both by the entropy increase and the information loss. Using the Euclidian estimating path, we *minimize* the overall entropy increase and information loss of the whole gnostical cycle. Both these fundamental measures of uncertainty are functions of the Minkowskian angle $2 * \Omega_i$ which corresponds in Figure 1 to the rotation of the radius-vector necessary to trace the movement between the two related points $\langle x_i, -y_i\rangle$ and $\langle x_i, y_i\rangle$. Its Euclidian equivalent $2 * \omega_i$ can be determined using the obvious relation

$$\tanh(\Omega) = \tan(\omega) \tag{12}$$

holding for all points of the Minkowskian path. The idea of the entropy of an individual datum is introduced in gnostical theory using a *Gedankenexperiment* and thermodynamic notions. Applying consistent mathematical operations to entropy, we then derive not only the information but also the distribution function of the individual datum. As a result of fundamental importance, the entropy $\Longleftrightarrow$ information conversion law is derived using these principles.

Principal Results: Gnostical Theory of Individual Uncertainty

The relationships which permit the definition of uncertainty in terms of the data will now be defined. In order to apply the theory to the data, we introduce the *scale parameter,* a positive number, s, which is related to the choice of measurement units for the rotation angles. The choice of value for the scale parameter is closely connected to the dispersion of the data in the sample. The following relations are discussed in more detail in [4].

Define an auxilliary quantity

$$q_i(Z_0, s) = (Z_i/Z_0)^{2/s} \tag{13}$$

for use in the calculation of the *fidelity*

$$\cos(2 * \omega_i) = f_i(Z_0, s) = 2/(1/q_i(Z_0, s) + q_i(Z_0, s)) \tag{14}$$

and the *irrelevance*

$$\sin(2 * \omega_i) = h_i(Z_0, s) = (1/q_i(Z_0, s) - q_i(Z_0, s))/(1/q_i(Z_0, s) + q_i(Z_0, s)). \tag{15}$$

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

$$p_i(Z_0, s) = (1 + h_i(Z_0, s))/2 \tag{16}$$

having the density

$$d_i(Z_0, s) = \frac{d(L_i(Z_0, s))}{dZ_0} = f_i^2(Z_0, s)/(Z_0 * s). \tag{17}$$

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

$$H(p) = -p * \ln(p) - (1 - p) * \ln(1 - p) \tag{18}$$

$$I(p) = H(1/2) - H(p). \tag{19}$$

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. Noting the formal coincidence of (18) with Shannon's formula of classical information theory, we surprisingly find that the quantity $p(Z_0, s)$ plays the role analogous to the probability of Z_0 given the observed datum Z_i. 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* instead 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:

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.

- 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 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.

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)	$\rightarrow \pm 1$
Weight	1	$1 - 2 * (\frac{Z_i - Z_0}{Z_0})^2$	$f_i(Z_0, s)$ (14)	$\rightarrow 0_+$
Entropy ch.	0	$-2 * (\frac{Z_i - Z_0}{Z_0})^2$	$f_i(Z_0, s) - 1$	$\rightarrow -1_+$
Expectation	1/2	$1/2 - \frac{Z_i - Z_0}{Z_0}$	$p_i(Z_0, s)$ (16)	$\rightarrow 0_+$ or 1_-
Inform. loss	0	$2 * (\frac{Z_i - Z_0}{Z_0})^2$	$I(p_i)$ (19)	$\rightarrow \ln 2$

Table 2: Error Dependence, Main Gnostical Characteristics of Data Uncertainty.

- 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 is useful, but it 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.

Composition Law for Contaminated Data

In addition to the already noted similarity between gnostical and relativistic events, another important (Lorentz–invariant) relationship exists: a linear mapping between the structure of vectors of the type $\langle \cosh(2 * \Omega), \sinh(2 * \Omega) \rangle$ and the structure of vectors $\langle energy, momentum \rangle$ of free relativistic particles. This mapping motivates a composition law in the following way: the energy-momentum conservation law of relativistic mechanics states that the total energy (total momentum) of a system of particles is given by the sum of the energies (momenta) of all particles. This law has been accepted as experimentally verified. Each contaminated datum has its "relativistic partner," a particle, and this mapping associates its *quantifying weight* $\cosh(\Omega)$ with the particle's energy and its *quantifying irrelevance* $\sinh(\Omega)$ with its momentum. It is therefore natural to require the same mapping for the vectors $\langle total\ weight, total\ irrelevance \rangle$ (of the data sample) and $\langle total\ energy, total\ momentum \rangle$ (of the particle's system). This is the idea behind the *third* gnostical axiom. It assumes that the quantifying weight and the irrelevance are to be composed additively and extends this requirement to the estimating weight and irrelevance of the vectors $\langle \cos(2 * \omega), \sin(2 * \omega) \rangle$. This composition law (which is nonlinear with respect to either the data or to the squares of their linear errors) leads to another striking property of gnostical procedures, their high robustness.

Economics and Relativistic Mechanics ... a Connection?

The implication given by gnostics that there exist commonalities between economics and relativistic mechanics may confuse and frustrate some practitioners. However, they do apply without hesitation or confusion the operations of the arithmetical mean and standard deviation and use least squares regression models inherently based on covariance matrices. These fundamental notions of classical statistics were brought from Newtonian mechanics. The arithmetical mean of data is the analogue of the coordinate of the center of gravity; the sum of data square errors are calculated in the same way as are diagonal components of the energy-momentum tensor of a system of mass points. Covariances are the images of nondiagonal components of this tensor. The additive composition law adopted in statistics both for data and square errors is thus motivated by the Newtonian version of the energy-momentum conservation law. As noted in Table 2, this is also valid from the gnostical point of view; but only under the condition of small relative data errors. Newtonian mechanics viewed by relativistic eyes is a special case of relativistic mechanics under the conditions of small (with respect to light speed) relative velocities. This velocity corresponds to relative data errors in gnostics. Gnostics has developed formulae valid for relative errors of any arbitrary size. These formulae are akin to recent developments in physics as classical statistical formulae are to the medieval mechanics. Economic data are often highly contaminated. The message for economists is therefore clear: *do not use medieval statistical formulae when treating highly contaminated data!* These cannot provide reliable results. Instead, apply modern gnostical data treatment technology!

Examples of Gnostics using Financial Statement Relationships

Body temperature is one of basic indicators of human health; there is no argument as to what a "normal" value should be. No comparable measure exists to judge the economic health of a firm. It is necessary to rely on data analysis, and to make comparisons with values obtained from similar economic entities. The set of specifically comparable units is usually very small; this situation is well suited to the application of gnostics: small samples of highly dispersed data which have no theoretically justified statistical model. Two approaches will be illustrated.

Robust Regression Analysis

A profit model was tested using 1989 data for companies in the food, beverage, and tobacco industries. Both classical and gnostic methodology was employed to examine the following relationship:

$$Profit = TA + CL + LTD + NW + Sales \tag{20}$$

The estimated parameters are available from the authors, however, of primary interest here is a comparison of the effectiveness of the two techniques. These results are presented in Table 3.

Industry	Modeling Method	Statistics				
		RS	ME	MAE	$WMAE$	$WMSQE$
Food	Least Squares	0.803	0.0	34.7	34.7	60.6
	Gnostical	0.953	8.9	33.0	12.7	22.5
Beverages	Least Squares	0.934	0.0	81.2	81.2	87.7
	Gnostical	0.975	-16.3	70.4	32.2	4.7
Tobacco	Least Squares	0.996	0.0	39.2	39.2	56.1
	Gnostical	0.999	18.8	26.2	6.5	17.8

Table 3: Effectiveness Measures: Least Squares vs Gnostical Regressions.

Each of the statistics shown, even the "method neutral" measure of Mean Absolute Error (MAE), using *fixed* weights that equal 1 in both cases, indicate an improvement in the information content of the gnostical model. When the variable (gnostical) weights are applied, the statistical quality measures RS, $WMAE$, and $WMSQE$ all significantly improve, but at a small sacrifice to the unbiasness of the LS formulation (where $ME = 0$) by construction).

In order to visually demonstrate the effect of using the gnostical weights, a simple bivariate model, regressing profit on net worth, is shown in Figure 2.

In particular, the gnostical model has suppressed the influence of Seagram Co. and the line passes closer to the more "average" performers, Molson, Brown Ferris, Anheuser-Busch, etc.; (might one be tempted to throw Seagram out of the LS regression as an atypical outlier?). The validity of these results is of course constrained by the hypothesis of the *linearity* of the regression function. A more detailed gnostical analysis shows that a nonlinear model is in much better correspondence with the data. Such a

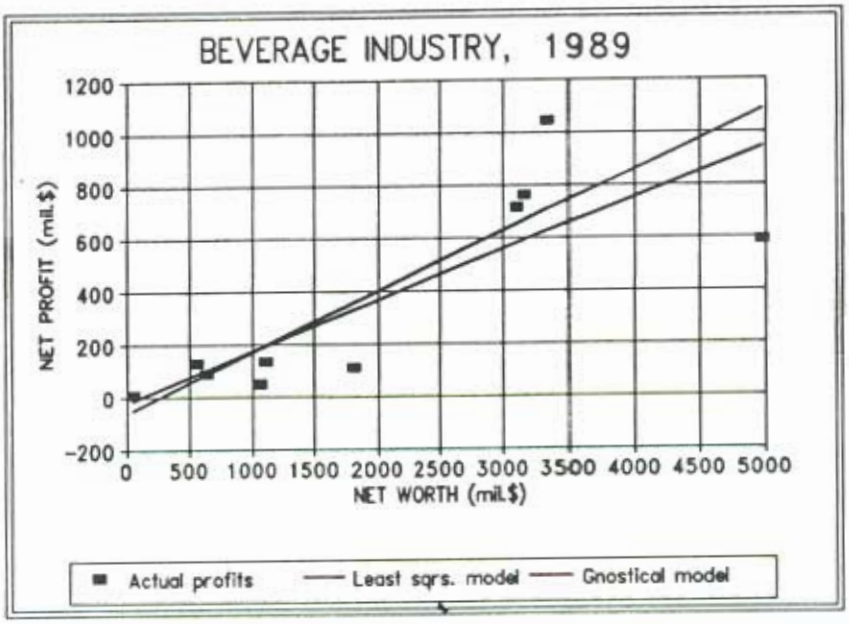

Figure 2: Net Profit vs. Net Worth.

model suggests that Seagram does not belong to the relatively homogenous cluster formed by the other members of the data set. These examples clearly demonstrate the superiority of gnostical models in an economic framework.

Gnostical Distribution Functions

Preliminaries

There are two kinds of distribution functions (d.f.) supported by gnostical theory: 'local' and 'global'. Both are obtained by further development of the idea underlying the composition law. The local function is not constrained while the global function assumes that the data sample contains only one homogeneous cluster. To describe both d.f.'s we use the functions previously defined and the notation

$$\bar{f}(z,s) = \sum_{i=1}^{N} f_i(z,s)/N \tag{21}$$

applies to N fidelities. The data sample weight is introduced as

$$w(z,s) = \sqrt{(\bar{f}(z,s))^2 + (\bar{h}(z,s))^2}. \tag{22}$$

The *local d.f.* $L(z,s)$ is simply the arithmetical mean of the d.f.'s of individual data:

$$L(z,s) = \sum_{i=1}^{N} L_i(z,s)/N. \tag{23}$$

The *global d.f.* $G(z,s)$ is obtained by applying the weight w, to the same sum:

$$G(z,s) = \sum_{i=1}^{N} L_i(z,s)/w(z,s). \tag{24}$$

With weak contamination (small data errors), the two functions differ only slightly. However, their behavior is quite different under gross errors.

In the sense of being a monotonic function of an arbitrary data sample, the local d.f. always exists. Were a statistical model of the data to exist, the d.f. $L(z,s)$ could be interpreted as a nonparametric estimate of the probability d.f. and the function $d_i(z,s)$ (17) as a proper kernel of the Parzen's type [5]. In such a case, gnostical theory is used only as a background which motivates the choice of the special kernel (17). Under these circumstances (but not necessarily limited to these), gnostical theory generates remarkably smooth density curves even with small data samples. The asymptotic features of the estimate can then be examined by established statistical methods.

In the more general case of not having a statistical model for the data, equation (23) is still useful as a continuous model of the data sample's distribution function and as an estimate of the expectation of an another datum having the same origin, Z_0. The steep descent of the gnostical kernel (17), manifested by the peakedness of the curve, has an important consequence: the individual subclusters of data influence each other only weakly. This enables the local details of the data sample to be characterized and provides an efficient method for cluster analysis.

Unlike the local d.f., the global function $G(z,s)$ only has theoretical support for homogenous data samples, i.e. for data with a unimodal density function. When applied to multimodal cases, this function may lose the fundamental feature of a d.f., its monotonic nature. This permits the hypothesis of homogeneity of the data sample to be tested. The global d.f. has no known statistical analogy. Its practical importance is related to its remarkable robustness with respect both to outlying data and to outlying subclusters of data. When estimating an expectation ("probability") for extremal quantiles, the 'central', or 'main' part of the data sample plays the dominant role. The global d.f. thus establishes the overall distribution law for the data. The utility of this methodology includes the d.f.'s excellent performance when applied to small samples and its applicability to samples generated by different distribution laws (logistic, normal, Weibull, etc.) as documented in [6].

The local and global d.f.'s differ substantially in their dependence on the scale parameter (s). Let $F(N)$ be the 'empirical' distribution function of the data sample. $F(N)$ then has the known form of an irregular staircase. The local d.f. $L(z,s)$ of the same sample can be made to approach $F(N)$ as closely as required by choosing a sufficiently small positive value for the scale parameter. In contrast, for the global function, there is a value of $\tilde{s}$ which minimizes the maximum distance between d.f.'s $G(z,\tilde{s})$ and $F(N)$. This value of $\tilde{s}$ is a robust estimate of the scale parameter which then causes the d.f. $G(z,\tilde{s})$ to approach the empirical distribution function $F(N)$ as closesly as is possible.

Gnostical Distribution Functions of the Return on Equity

There were 8 large firms in the Tobacco Industry quoted on U.S. stock Exchanges in 1989. The local d.f. (23) and density distribution (17) for the ROE were both computed using an estimate of the scale parameter set at $s = 0.8$. The left hand graph in Figure 3 displays these data and reveals a large main

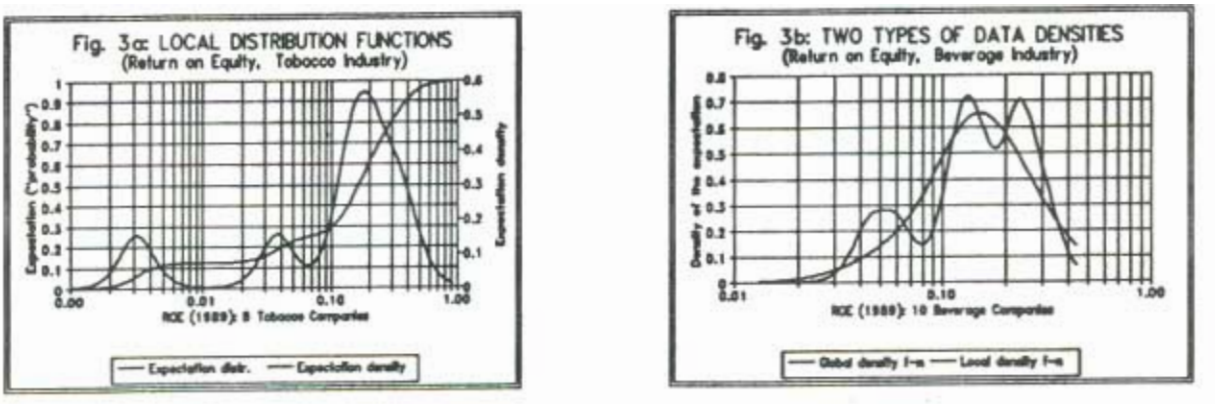

Figure 3: Examples of Local and Global Distribution Functions.

cluster and two smaller ones which represent two outliers: Culbro Corp. ($ROE = 0.00318$) and Std. Commercial ($ROE = 0.0374$). The principal cluster is comprised of 6 firms, the ROE of which ranges from 0.140 (Universal Corp.) to 0.395 (UST Inc.). The maximal density of the outlying clusters coincides with their ROE values. The most frequently expected value within the firms in the main cluster is the maximum at $ROE = 0.180$. (The closest firm to this "typical" value is Dibrell Bros. at 0.194). The graph implies that an erroneous conclusion would be drawn by taking a "normal" approach, using the arithmetical mean of ROE's of the group, to rank or estimate the performance of one of its members. The arithmetical mean of all 8 tobacco firms is 0.176. This single number does not speak to the actual structure of data, which is very far from being of Gaussian form. If one were to exclude the two outliers (have we the right not to take them into account at all?), and calculate the arithmetical mean of the firms in the main cluster, the result is 0.227. Is this better information? The maximal density of this cluster is at 0.180. The high value of the arithmetical mean is due to the asymetry of the cluster. Its falling branch is less steep than the rising one. Were one concerned with evaluating a firm in the group, the ranking provided by Figure 3a would provide more useful information than that gained from only using a point estimate.

Now returning to the beverage companies, and applying the same gnostical technique, Figure 3b is obtained. Note however that there is an important difference between the sample of tobacco firms and that composed of beverage companies. No global distribution function exists for the former because of the non–homogeniety of the data group. In the latter case we can compute and apply (as shown) both d.f.'s. The global function is the best possible representation of all 10 ROEs formed by viewing the data sample as a whole. The maximum density is at 0.151. The advantage of this global representation is that it enables one to reliably estimate the expectation of extreme ROE values.

With respect to the local d.f., there again appear three clusters but with a different composition. The lowest cluster (max. density, 0.0529) represents two firms (Coors) and Coca-Cola Enterprises), at 0.0416 and 0.0602 respectively. The central clusters correspond to, on the left, the ROEs of 4 firms (Seagram, Labatt, Molson and A&W Brands) which ranged from 0.119 to 0.152. The maximum density is at $ROE = 0.133$. The right hand cluster contains the 4 best performers (Brown-Forman, Anheuser-

Busch, Pepsico Inc. and Coca-Cola) whose ROE range between 0.229 to 0.312. The maximum density here is found at $ROE = 0.235$.

The additional information provided by the local distribution function resolves the conflict as to whether Seagram's extreme outlier status justifies its retention in the regression. The two points of view are complementary speaking to different aspects of the data. The regression is influenced by the extreme value of Seagram's net worth, while the distribution function demonstrates that the firm's performance falls within one of the major clusters. This once again cautions against reliance on "simple" analyses.

Conclusions

Size limitations have prevented a more detailed explanation of the mathematical origins of gnostical theory, and the interested reader is referred to [4] for a more complete exposition. The simple examples used to illustrate the theory, however, clearly demonstrate the better insight that can be obtained on the structure of data, and the attendant advantages which can accrue to economic analysis by utilizing gnostic procedures.

References

[1] Benjamin Graham, *The Intelligent Investor,* Harper & Brothers, 1959, New York, N.Y. 10022, p. xv.

[2] Cornelis A. Los, A Scientific View of Economic Data Analysis, *Eastern Economic Journal,* XVII, 1 (1991) 61–70.

[3] R. E. Kalman, 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.*

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

[5] E. Parzen, On estimation of a probability density function and mode, *Ann. Math. Stat.*, 35(1962), 1065-1076.

[6] R. H. Baran, Comments on "A New Theoretical and Algorithmical Basis for Estimation, Identification and Control" by P. Kovanic, *Automatica* 24(1988), 283-287.