GALTON S ERROR AND THE UNDER-REPRESENTATION OF SYSTEMATIC RISK

CORNELIS A. LOS, PH.D.

Centre for Research in Financial Services, Working Paper No. 97-01

ABSTRACT. Science progresses by improving its measurement apparatus. This holds true in nance too. Our new methodology of complete identi cation, using simple algebraic geometry, throws new light on the continued commitment of Galton s Error in nance and economics and the resulting misinformation of investors. Mutual funds conventionally advertise their relative systematic market risk, or betas, to potential investors based on incomplete measurement by unidirectional bivariate projections: they commit Galton s Error by under-representing their systematic risk. Consequently, far too many mutual funds are marketed as defensive and too few as aggressive. Using our new methodology we found that, out of a total of 3,217 mutual funds, 2,047 funds (63.7%) claimed to be defensive based on the current industry standard methodology, but only 608 (18.9%) actually are. This under-representation of systematic risk leads to inefficiencies in the capital allocation process, since biased betas lead to mis-pricing of mutual funds. Our complete bivariate projection produces a correct representation of the epistemic uncertainty inherent in the bivariate measurement of relative market risk. Our conclusions have also serious consequences for the proper bench-marking and recent regulatory proposals for the mutual funds industry. Extension of the new methodology to multivariate systematic risk measurement by Asset Pricing Theory is suggested.

1. INTRODUCTION

The scienti c measurement of systematic risk has become an important feature of the global nancial markets. Accuracy of return and risk measurement and their nancial analysis is at a premium, now that the American nancial markets are once again hovering in the stratosphere [1]. There is now even a global price for the determination of covariance risk [35]. Unfortunately, the scienti c measurement of covariance risk appears to be not well understood by either nancial analysts,

Date: January 1, 1997.

The author is Associate Professor in Banking and Finance. This paper has been presented in the Advanced Research Seminar on Probability, System Theory and Econometrics at INREA, Sophia Antipolis, France, June 3-5, 1997; in the Economics Seminar at the University of Maastricht, The Netherlands, June 12, 1997; at the 14th International Conference in Finance (AFFI97) in Grenoble, France, June 23 - 25, 1997; and at Quantitative Methods in Finance Conference (QMF97) in Cairns, Australia, August 25 - 29, 1997. With special thanks to Professors Rudolf Kalman (ETH), Yves Rouchaleau (INREA), Gerrit Meijer (Maastricht), Jürgen Backhaus (Maastricht), Patrice Fontaine (Grenoble) and Carl Chiarella (New South Wales) for their encouragement, comments, critique and constructive suggestions.

[1] *The Economist*, July 6, 1996, p. 21. Partially in reaction to professional demands, in the Fall of 1996 a new journal appeared, *The Journal of Performance Measurement*, devoted exclusively to issues of portfolio return and risk measurement.

economists or statisticians, even though the study of risk has already a long, and remarkable, history [14]. In this Introduction we provide some background for the scienti c measurement of systematic risk and for the understanding of Galton s Error in nancial economics, in addition to an outline of the paper.

1.1. **Scienti c Measurement.** The measurement of covariance risk is a scienti c measurement problem. The problem of *scienti c measurement* is how to identify, or realize, a model, or system, from empirical data, i.e., data that are inexact and contain noise. Many disciplines, including economics and nance, use *unidirectional* projections for the process of system identi cation, combined with statistical hypothesis testing based on assumed probability [62] [2]. But the results of such statistical modeling turn out to be very biased, unreliable and disputable. The crucial question is: why? Our short answer is: because of the introduction of uncorroborated presumptions extraneous to the data [40], [54] [3].

In contrast, mathematical modeling schemes, which use *complete* projections and rely on the algebraic geometric structure of the empirical data and the exact mathematical laws of mapping, have booked reliable successes. Compare, for example, the success of crystallography and magnetic resonance imaging in DNA and protein structure recognition [66]. In 1952 this 3D DNA crystallographic research used fuzzy 2D X-ray projection pictures of proteins as its raw empirical data. Its multi-dimensional projection success culminated in the mid-1980s in the 15-year Human Genome recognition project. In the past 44 years, the protein crystallography has produced a burgeoning, billion-dollar bioengineering industry with great promise for the 21st century [2]. No such clear-cut success can be ascribed to the unidirectional statistical modeling and probabilistic hypothesis testing approaches in economics or nance. But it is not immediately obvious why that should be the case.

After all, other examples of computerized multidimensional mathematical maps currently drawn by scientists to visualize the complex world surrounding us, and to navigate and nd solutions to crucial problems, have begun to abound in every eld of science. A virtual Renaissance of observational sciences is underway, thanks to more sophisticated mathematics and very inexpensive computing power. The scope of these maps has increased dramatically: from remotely sensed, falsely colored Landsat maps of the earth, used to predict the size of harvests, to computer modeled paleoclimatologic maps giving insights in earlier vegetations and to maps of gravity anomalies in the earth s core to correct modern compasses; from spectrometers measuring ozone layers, raising concerns about the production and use of aerosols, or measuring galactic molecular clouds and far away rather esoteric black holes, to the three-dimensional computed tomography of human brains and bodies used

[2]The projections we study here are the conventional regressions, to which principal component and factor analysis schemes can be shown to be equivalent [54].

[3]Our Socratian exposure of the prejudices of the conventional statistical methodology in the late 1980s was a necessary preliminary to true understanding and knowledge of the problem of system identi cation from inexact data.

for virtual surgery [4]; from scanning probes mapping atomic surfaces [5], to the cartography of subatomic particle detectors [29].

These new navigational charts are being drawn, all based on the accurate measurement and complex algebraic geometry of the data. But this is not happening in the so-called social or soft sciences, which include economics and nance. This difference in the maps and in the reliability of results in the hard versus the soft sciences should give pause for a serious reassessment of the research paths of the statistical disciplines. These statistical disciplines rely on incomplete unidirectional projections, assumed probability theory and statistical hypothesis testing [17] [6]. It should also give pause to those statisticians, who, like us, follow a geometric approach, but are unwilling to discard particular statistical conventions [69], [76].

Our professional concern is raised, because the unreliability and bias of the statistical results in economics and nance is likely to cause serious misallocations in the global process of investing the billions of retirement funds. It is also doubtful that such a deplorable situation must persist, now that new, more reliable research methods have become available.

To improve the analytic research methodology in economics and nance, we propose to use the complete algebraic geometric modeling from inexact empirical data, the so-called **super lter** methodology. This methodology uses the well - understood characteristics of linear covariance systems, that can easily be implemented by the social sciences in general and by nance and economics in particular. In fact, the super lter methodology has a close historical affinity to the linear modeling based on covariance matrices as practiced by econometricians and nancial

[4]For the latest somewhat lugubrious, but very convincing example of the success of 3D mapping, see the digital humans now inexpensively available on CD-ROM [3]. Scienti cally these modern 3D pictures surpass the 2D woodcut prints of Titian s studio in Vesalius *De Humani Corporis Fabrica* (*On the Structure of the Human Body*;1543). This masterfully illustrated treatise of pioneering anatomy of the 16th century Flemish physician Andreas Vesalius helped establish modern observational science. These digital humans are examples of transforming 2D data - the color pictures of millimeter thin slices of two deep-frozen human cadavers - into 3D data by mathematical projection and by combination with 3D Nuclear Magnetic Resonance data of their bone structure. As Hall [29] explains, computer graphics have led to a renaissance in cartography that affects now all scienti c disciplines.

[5]A particularly beautiful recent example of the usefulness of computerized molecular maps is the scanning electron micrograph of a cell infected by AIDS viruses, which helped with the crucial identi cation of the structure of an HIV virus to nd effective inhibitors [26].

[6]It is not that there are no sporadic attempts to enhance the scienti c content of the statistical sciences. For example, there exist now colored mathematical maps, called the lakes of Wada elucidating the paradoxical laws of chaos, which are used for studying electromagnetic elds [29] pp. 265 - 281. Currently there emerges even more serious empirical research, also in nance, on the issues of stationarity, independence, and randomness [72], Chapters 1 - 3 and [43]. Some have argued that efforts like the binomial and Black-Scholes option pricing models are accurate maps of empirical nancial data. They aren t, since these nominal valuation models are based on conventions, postulates and prescriptions, religiously followed by the nancial services industry, but not on actual identi cation from empirical market data. In fact, empirical research shows considerable discrepancies between the measured results from actual option pricing processes and these conventional postulate models, giving rise to pro table arbitrage opportunities.

analysts [7], but differs crucially from the conventional approach by taking account of all covariances simultaneously from all directions [41], [42], [54], [53], [57], [59].

We give in this paper an introductory and rather didactic account of system identi cation from inexact data by the super lter method, by concentrating on a nancially highly relevant and timely bivariate example, although we also show the general case.

In our methodology we explicitly adopt some minor restrictions. In particular, we adopt two scienti c premises, rst, that the systematic variation in the observable variables is generated by linear systems and, second, that the underlying systems are static. The linearity restriction is not as restrictive as it seems. By a linear system is meant that there is linearity in the model coefficients and not necessarily in the variables, which may be uniquely transformed and scaled by exact relationships, like exponentials and logarithms. For example, the data to be analyzed by such linear system identi cation could be $y_t = Ce^{dz_t^a}, t = 1, ..., T$, where z_t are the original data and C, d and a are *known constants*. Consequently, the remaining unsystematic variation contains the nonlinearities in the transformed data, after possible unique transformations of the set of the original raw data. When we transform the original data by a unique mapping, as in the example, we also transform the noise in the original data. Such preparatory transformation of the data is like the focusing of a camera. To bring an object in focus one adjusts the lens, which is a unique mathematical transformation of the light rays re ected by the object. Furthermore, the restriction to work with linear, or linearized, models is already accepted practice in economics and nance.

Secondly, in this paper, we only look at static models and static systematic correlations and not at dynamic models or correlations, i.e., correlations over time. Thus the appearance of randomness in the remaining uncertainty, after the restriction of our special camera - the linear model - is imposed, can also be caused by unaccountable time dependence, or by time dependence in combination with the remaining nonlinearities. This is an area of important current theoretical research related to chaos theory, i.e., the theory of *deterministic* nonlinear differential equations, and of empirical research [43], [45], [44], but that is outside the scope of this paper.

In this linear modeling context, the paper uses Kalman s pragmatic de nition of model uncertainty (Cf. [42], Lecture 1):

model uncertainty = inexactness = non-uniqueness

This de nition of model uncertainty is appropriate for linear measurement models. In addition, we introduce a slightly more general de nition:

epistemic uncertainty = uncertainty in our knowledge from modeling

[7] The designations super lter and data-microscope for our new methodology were rst used by the famous mathematician Professor Rouchaleau of the Ecole des Mines, Paris, at a recent advanced econometrics seminar at I.N.R.E.A. in Sophia Antipolis in Southern France in May 1996 (according to a personal e-mail message from Kalman, March 26, 1996). *Super lter* is the preferred designation, since it indicates that it is a step up from the original Kalman lter, because it determines the system s invariant - the corank - from the data, instead of relying on engineering presumptions, as we suggested in Beijing for a $(n, q) = (3, 2)$ system [60].

These simple epistemic de nitions appear all that is required for the scienti c identi cation of linear systems from empirical data, as this paper will demonstrate [8]. Surprisingly, no important additional assumptions are required.

1.2. **Galton s Error.** The immediate motivation for this paper is derived from a monumental scienti c error made more than hundred years ago by Sir Francis Galton, the inventor of the omnipresent regression method. His error of omission is still being committed, as evidenced by scores of papers in learned scienti c journals and respected treatises.

In Section 3. of this paper we will discuss the simplest bivariate example of Galton s Error in the general context of linear identi cation from inexact data. The example we use is that of the determination of the relative risk - the so-called beta - of mutual funds. It will elucidate Galton s Error in the context of modern nance, in particular, of that of the familiar bivariate Capital Asset Pricing Model (CAPM).

What was Galton s Error? As part of serious anthropological research, Francis Galton, a cousin of Charles Darwin, proposed in 1886:

> ...to express by formulae the relation that subsists between the statures of speci ed men and those of their kinsmen in any given degree, and to explain the processes through which family peculiarities of stature gradually diminish, until in every remote degree of kinship the group of kinsmen becomes indistinguishable from a group selected out of the general population at random [25], p. 42.

Most scientists now acknowledge that it was a serious scienti c error of Galton to accept his downward biased regression results as conclusive evidence for his asserted hereditary process of regression towards the mean of the stature, or height, of the human race. Because, had Galton correctly interpreted the computational results of what we now call reverse regressions (which he did run in both his 1885 and 1886 papers, [24], [25]), he could possibly have derived the opposite conclusion: that historically the stature, or height, of the human race becomes more dispersed. But such an acknowledgment of Galton s Error doesn t imply understanding it.

What is not well known is that the opposite conclusion would also have been erroneous, since the properties Galton thought he observed in the data were generated by his. They were generated by his measurement apparatus, his unidirectional statistical projection camera. They were not properties inherent in his data. It is surprising that even statisticians who acknowledge Galton s Error and who are sympathetic towards our visual, geometric approach to scienti c measurement, appear to be blind to this essence of Galton s Error. See, for example, the articles by [37], [17], [67], and, more recently, [16].

The essence of Galton s Error is that there is no scienti c basis in the data for the conventional *a priori* differentiation between regressands and regressors, between explained and explanatory variables. The statisticians conventional notation differentiating between y s and x s has no scienti c basis in the data. In contrast, in our new methodology, all data are considered equal and we don t

[8] The Greek *epist$\overline{e}$m$\overline{e}$* = knowledge. Our fundamental research is essentially epistemological, since it investigates the origin, nature, methods and limits of knowledge obtained by linear system measurements. With thanks to the economists, methodologists and philosophers of science at the University of Groningen in The Netherlands and at the London School of Economics, who rst inspired me, as a student in the 1970s, to do methodological research in economics and nance.

differentiate between y s and x s. Our super lter is a linear system camera which views the complete data set from all directions, similar to what is done in X-ray diffraction research of proteins, where a goniometer is used to make such a multi-directional projection possible [9].

In addition, Galton made several other scienti c errors, all less important than this basic one. But, surprisingly, all his errors are still regularly committed by reputable researchers in various disciplines of learning. For example, Galton asserted, but did not provide scienti c evidence for, the assumed stationarity and homogeneity of his data sets. He also presumed that his data were random and even probabilistic, although that was (and is) irrelevant for his (biased) conclusions with respect to the relationship studied. Because these historical errors have remained prevalent, Appendix I to this paper discusses in detail several of them as exempli ed in Galton s own historical anthropological papers of 1885 and 1886.

Based on our new super lter methodology for the identi cation of complex systems from inexact empirical data, we can now unambiguously conclude that Galton incompletely researched and misunderstood his covariance data. He did not understand the geometric structure of his data and the mathematics of multidimensional covariance. Consequently, he did also not understand the geometry of his epistemic uncertainty or inexactness. His lack of understanding was not innocent or without serious consequences, since it provided ample room for the many statistical prejudices introduced by his successors. These prejudices were introduced in erroneous attempts to reduce or eradicate the irreducible epistemic uncertainty inherent in the empirical data. We fear that these same errors cause now serious misallocation in the nancial markets.

From a scienti c point of view, the uncertainty of Galton s data should not have allowed him to draw his biased conclusion that the stature of men is diminishing over time, *since his data evidence was too uncertain to be factually conclusive.* The variation in Galton s data was 77.8% unsystematic, or uncertain, and only 22.2% systematic, as can be checked using Galton s own published results (Cf. Appendix I of this paper for a detailed account). Thus Galton s ignorance, i.e., his epistemic uncertainty, was about three times larger than his knowledge, i.e., his model of the data.

Galton s Error also persists in nance, as we will demonstrate in Section 4. with the published, recommended and marketed, but severely biased classi cation of mutual funds into aggressive , neutral, or market-index like, and defensive funds. This classi cation of investment alternatives is based on the nancial industry s standard practice of the computation of unidirectionally projected empirical betas, i.e., their relative return volatilities, with respect to a particular market index, or their relative attributions of their systematic variation, which were proposed by the 1990 (joint) Nobel Prize winner William Sharpe [71].

Again, this seemingly innocent practice is not without serious consequences. There is currently an alarming, and misdirected, regulatory interest in a single risk measure to classify mutual funds [32]. Sharpe s beta has been proposed by many analysts as such a measure. This official interest in a single risk measure is just as

[9]The Greek *gonia* = an angle, corner. A goniometer is an instrument for measuring angles, especially of solid bodies. The only difference between our super lter and a real camera, which takes 2-dimensional pictures of 3-dimensional objects, is that the super lter is a universal camera which can be applied to relational, array type data sets with many more than three dimensions. It could therefore become a useful analytic tool for particle research [58].

misdirected as the development of a single measure for intelligence - the infamous Intelligence Quotient. The I.Q. was based on factor analysis at the beginning of this century by psychometricians [54]. Since then it has been exposed as nonsense by several scientists and non-scientists alike, in particular by [27], pp. 234 - 320). Furthermore, in the context of nancial derivatives, Sharpe s betas are now used in the cross market pricing of commodity futures, e.g., gold and silver, or copper and aluminum futures. Again, such delicate pricing assumes a greater degree of accuracy than is warranted by the market data or is considered acceptable by the nancial industry.

These consequential errors are just as easily exhibited in bivariate models like Galton s family regression, as in the bivariate Capital Asset Pricing Model [10]. For three dimensions, we can graphically demonstrate the scienti c errors committed, as we will see in Section 5. All these errors are less easily demonstrated in model situations with dimensionality higher than three, as, for example, in the currently popular multivariate single equation Asset Pricing Model of Ross [18], or the Credit Scoring and Bankruptcy Prediction Model of Altman [8], both of which include often more than three variables. However, the inconsistency of these models with the multivariate covariance data can still be demonstrated by analyzing their information matrices, i.e., the inverses of their data covariance matrices. For an example, see [55], where such data inconsistency is demonstrated in a popular ve variable economic forecasting model [11]. In Section 5. we show a simple graphical extension of the concept of epistemic uncertainty to the trivariate case, which provides the bridge to the general case with more than three variables, to be discussed in a companion paper.

One essential problem of the trivariate case does not occur in the bivariate case discussed here. That is the *identi cation of the invariant dimensionality of the data structure*, or, in technical terms, the identi cation of the corank of the (Grassmanian) system representing the systematic part of the data s total variation. This particular problem, recognized already in 1934 by the 1996 (joint) Nobel Prize winner Ragnar Frisch [23], was not solved in economics or nance. It was papered over by the Cowles Commission in the 1940s and 1950s. But this problem is now be solved thanks to the discovery and development of some important identi cation

[10]Not surprisingly, we nd articles and treatises on the economics of the family by the 1992 Nobel Prize winner Gary Becker, who, like Alfred Marshall almost a century before him before him, is still erroneously impressed by Galtonian family regression [10]. But what to think of the following? On March 1, 1995, in response to an earlier draft of this paper, we received a letter from Dr. Mico Loretan, Economist in the Division of International Finance of the Board of Governors of the U.S. Federal Reserve System, claiming that Galton s Error has long been recognized as an error. He confused the distinct concepts of mean regression and (least squares) projection. On the basis of this incorrect assessment, our paper was rejected as an entry for the Joint Bank Conference on Stress-Testing of Risk Management Models and Systemic Risk in November 1995. As this paper will demonstrate, algebraically and by reference to Galton s own paper, the asserted confusion was NOT Galton s Error and the distinction made by Loretan is empty. With all due respect to the Federal Reserve, Galton s Error is still committed by all Economists of the Federal Reserve System who use conventional regression analysis and similar methods, unless somebody falsi es that factual observation. Reading the publications of the Federal Reserve, the author is certain that little has changed since he was an Economist of the Federal Reserve Bank of New York in 1981 - 1987.

[11]Some did not nd our demonstrations of the $(n, q) = (3, 2)$ and $(n, q) = (5, 4)$ cases convincing. For some critical commentary by a Bayesian physicist and two conventional econometricians, see, respectively [38] (rebutted in [57]) and [13] (rebutted in [56].

Theorems by the 1985 Kyoto Prize winner Rudolf Kalman in 1990/91 [41], [57], [59]. This means that the problems of unidirectional multivariate single equation models like Stephen Ross 1976 Asset Pricing Model [68], which is notorious for its coefficient instability, can nally be satisfactorily resolved in a scienti c fashion.

The complexity of the algebraic geometric analysis remains at all times tractable in our recommended linear matrix notation, as we will demonstrate. The volume of the epistemic, or model uncertainty can always be found from the adjoint of the data covariance matrix. The ratio of this epistemic uncertainty volume relative to that of the relevant data orthant is the multidimensional extension of the inverse of the conventional signal/noise ratios of engineers discussed in Section 3.

In the Conclusion, based on results from our new research methodology, we sound a clear warning for the nancial services industry, in particular the mutual funds industry and its regulators, to distrust its conventional risk measure, Sharpe s beta, and to not base its capital costing on this prejudiced and biased measure. For example, of $3,215$ regularly monitored funds in the U.S., which contain measurable systematic risk, $1,488$, or 46.3%, can not even be unambiguously categorized as defensive, aggressive or market index using Sharpe s beta, given the amount of epistemic uncertainty or inexactness in the nancial data. Still, this hasn t stop the mutual funds industry from categorizing them as defensive. This implies that of the 63.7% of the funds claimed to be defensive, only 18.9% actually are. This amounts to substantial falsehood in advertising in the mutual funds industry in the U.S.

By extension, we caution the readers when confronted by unidirectional projections of any kind, which provide an incomplete picture of the multidimensional covariance data. Having done extensive surveys, e.g., of the *American Economic Review,* the *Journal of Finance*, and similar journals, we have no longer any doubt that similar methodological de ciencies can be culled from existing published research in the economics and nance literature. Moreover, we saw our decade old doubts corroborated and our worst fears about the persistent lack of scienti c integrity reinforced, when in 1992 regression towards mediocrity in economic stature became a topic for serious debate in a leading economic journal [79].

2. SYSTEM IDENTIFICATION FROM INEXACT DATA

In this Section we rst introduce some simple matrix notation to facilitate the following discussion. Primarily for educational purposes, and to connect to existing statistical conventions, we present a cookbook recipe for the bivariate linear modeling from empirical, inexact data. Econometricians, nancial analysts and other statistical researchers conventionally begin by postulating some theory and then nd the data to corroborate that theory. In contrast, we begin with the data set to be explained and try to nd what system can have produced the observed covariance structure of the complete data set.

2.1. **Data.**

The rst and second moments of the original, or raw data series, i.e., the expected value (average, mean) and the variance and covariances, respectively, can *always* be computed. Let $\mathbf{y}$ be the vector of T data points or observations of

order $T \times 1$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_T \end{bmatrix}$$

with elements y_t for the integer t, $1 \square t \square T$.

The expected value, or mean of the elements of vector $\mathbf{y}$ is the scalar

$$\overline{y} = E(y_t) = \frac{\sum_{t=1}^{T} y_t}{T}$$

The deviations from the mean $\overline{y}$ form the $T \times 1$ vector

$$\mathbf{x}_1 = \begin{bmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ ... \\ y_T - \overline{y} \end{bmatrix} = \mathbf{y} - \boldsymbol{\iota}\overline{y}$$

where $\boldsymbol{\iota}$ is the $T \times 1$ unit vector $\boldsymbol{\iota} = \begin{bmatrix} 1 \\ 1 \\ ... \\ 1 \end{bmatrix}$

Notice that, *vice versa*, the original data vector is $\mathbf{y} = \mathbf{x}_1 + \boldsymbol{\iota}\overline{y}$, i.e., the data form the sum of their mean and their deviations from the mean. Therefore, we can always equivalently analyze the deviations instead of the original data and reconstruct the original data from the deviations and the mean. Also, the expected value of the deviations always equals a $T \times 1$ vector of zeros: $E(\mathbf{x}_1) = \mathbf{0}$. Geometrically, by taking deviations from the mean, we have only laterally shifted the frame of data reference through its origin, but not changed the data structure, i.e., its information content.

The variance of the elements of the vector $\mathbf{y}$ is the scalar

$$\sigma_{yy} = \frac{\sum_{t=1}^{T}(y_t - \overline{y})^2}{T} = \frac{\mathbf{x}_1'\mathbf{x}_1}{T}$$

where $\mathbf{x}_1'$ represents the $1 \times T$ *transpose* of the $T \times 1$ vector $\mathbf{x}_1$, i.e.,

$$\mathbf{x}_1' = \begin{bmatrix} y_1 - \overline{y} & y_2 - \overline{y} & ... & y_T - \overline{y} \end{bmatrix}$$

Thus the variance of a series of data can be computed as a scalar product of deviations from its mean [12].

For our bivariate problem there is a set of T observations on $n = 2$ variables. After subtracting the means, we have the matrix of deviations

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} y_{11} - \overline{y_1} & y_{12} - \overline{y_2} \\ y_{21} - \overline{y_1} & y_{22} - \overline{y_2} \\ ... & ... \\ y_{T1} - \overline{y_1} & y_{T2} - \overline{y_2} \end{bmatrix}$$

where $\mathbf{x}$ is a $T \times 2$ matrix, so that the expectation $E(\mathbf{x}) = \mathbf{0}$, a $T \times 2$ matrix of zeros.

[12] Notice that we use the notation σ_{yy} instead of σ_y^2, which is the conventional notation for the variance. Our notation σ_{yy} clearly indicates which two variables are involved in the computation, as becomes clearer when we introduce the covariances.

Our simple scienti c question is thus: how does $\mathbf{x}_1$ form a linear system with $\mathbf{x}_2$, or, equivalently, how does $\mathbf{x}_1$ and $\mathbf{x}_2$ linearly depend on each other?

2.2. **Data Covariance Matrix.** The data covariance matrix of these two data series is the 2×2 symmetric covariance matrix Σ of averaged products of deviations from the respective means. The diagonal elements of this covariance matrix, σ_{ii}, are variances, while its off-diagonal elements, σ_{ij} for $i \neq j$, are covariances. Each off-diagonal element of the data covariance matrix provides a bivariate picture of the covariation two data series. There are $\frac{n(n-1)}{2}$ such independent bivariate covariance pictures. Thus the bivariate case

$$\Sigma = \frac{\mathbf{x}'\mathbf{x}}{T} = \begin{bmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 \\ \mathbf{x}_2'\mathbf{x}_1 & \mathbf{x}_2'\mathbf{x}_2 \end{bmatrix} /T = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

provides in this simplest case only $\frac{2.1}{2} = 1$ independent bivariate covariance, although it can be looked at from different projection directions!

Data analysis for the identi cation of an inexact (linear) model is quite different from the realization of an exact model. For exact data the covariance matrix Σ is singular, because of the exact linear dependencies among the variables. But for inexact empirical data the symmetric data covariance matrix is generically non-singular, or positive de nite, and, consequently, invertible, so that the information matrix or inverse of the covariance matrix, Σ^{-1}, exists. This information matrix is the complete collection of our photographic plates. Despite the empirical noise , it contains all the 2D covariance information about the multi-dimensional structural geometry of the system that generated the data.

As we will see, the conventional division by T (or $T-1$, based on degrees of freedom reasoning), which I have maintained in this presentation, is strictly irrelevant for model identi cation, since it cancels out in the identi cation procedure, as we will see. Thus it doesn t do any harm to include or to exclude it. In contrast, conventional statistics pays much attention to the number of observations T and implies that the more observations the better, because of the in nity limit arguments it uses. In our methodology, T is irrelevant, since no in nity limit arguments are used. However, the positive de niteness of the data covariance matrix is relevant, since the positive de niteness of the data covariance matrix represents crucial information about the epistemic uncertainty of the data.

2.3. **Linear System or Model.** By a linear model we mean a model linear in its coordinates, or coefficients. Thus it does not mean linear in its variables, since the model s variables can be unique (nonlinear) transformations of the original data. The linear model is generically de ned by the expression

$$\mathbf{A}\widehat{\mathbf{x}}' = \mathbf{0}$$

with $\widehat{\mathbf{x}} = \mathbf{x} - \widetilde{\mathbf{x}}$ such that $\widehat{\mathbf{x}} \perp \widetilde{\mathbf{x}}$, since what is known of the data, i.e. explained by the linear model (= linear dependency), $\widehat{\mathbf{x}}$, is orthogonal (logically disjointed) to what is unknown, $\widetilde{\mathbf{x}}$. For the bivariate case

$$\mathbf{A}\widehat{\mathbf{x}}' = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{x}}_1 \\ \widehat{\mathbf{x}}_2 \end{bmatrix} = \mathbf{0}$$

Consequently, also $\mathbf{A}\widehat{\Sigma} = \mathbf{0}$, where the systematic covariance matrix (= matrix of linear dependencies) $\widehat{\Sigma} = \Sigma - \widetilde{\Sigma}$, where $\widetilde{\Sigma}$ is the unsystematic covariance matrix

(= noise matrix), with both $\widehat{\Sigma} \geq \mathbf{0}$ and $\widetilde{\Sigma} \geq \mathbf{0}$, positive semi-de nite, i.e., not necessarily invertible.

Technically, $\mathbf{A}$ is the $q \times n$ matrix containing the computed dual Grassman coordinates, conventionally known by econometricians as the model coefficients. The system invariant or corank q (= number of independent linear relations in the exact model) has to be determined from the inexact data for $n > 2$, where $n =$ number of variables, since generically $1 < q < n$. For our simple bivariate example, $n = 2$, thus $q = 1$.

Such a corank invariant analysis of the data is usually not done by statistical analysts, who conventionally presume or prejudge this number of relations q on philosophically reasoned or theoretical grounds. But such a research approach begs the question: do we explain theory by data, or data by theory ? Our methodology takes the data as the given...., as it classically should [13]. Fortunately, it is easy to show by simple examples that if the q presumed by the analyst is different from the q dictated by the data, the resulting model coefficients are totally random in the truest sense of that term [54].

In contrast to the conventional view in statistics, we conclude that strict Popperian model falsi cation is possible, when a presumed model is confronted with the *complete* covariance data. When the geometric linear structure of the model is not conform the geometric linear structure of the data, the data will reject the model algebraically, when all the covariance data are analyzed in a complete fashion. Indeed, the empirical econometric and nancial analysis literature is full of references to unstable, unreliable, or even chaotic model coefficients. Compare, for example, the historical debate following the rst presentation of the rst empirically estimated production function of Cobb and Douglas [19] or of the Monetarists money demand equation [48], or, more recently in the nancial literature, of Ross Asset Pricing Models. Often economic and nancial researchers have attributed the observed instabilities to the underlying economic and nancial systems generating the data [14]. But the observed instabilities can now be shown to result from the de cient conventional research methodologies.

For our bivariate model example with $n = 2$, of course, there can be no other q than $q = 1$, since there can not be more than one linear dependency between two variables. In Section 5. where we show an example with $n = 3$, q could be 1or 2, *a priori*. There can be one or two independent linear dependencies between three variables. The empirical data dictate $q = 2$, two independent linear dependencies, *a posteriori).*

2.4. **Complete Least Squares Projections.** Galton committed the serious scienti c omission not to research his covariance data completely, so that he drew a scienti cally erroneous conclusion. In this context, completely means that *all* the covariance data are used to compute the Grassman coordinates and not a prejudiced selection of the covariance data. An unidirectional projection, such as ordinary least squares (OLS) can be easily shown to select sections of the covariance matrix in a prejudiced fashion. The prejudicial choice of the regressand predetermines the projection direction.

[13] The Latin *data* = the given.

[14] Also *mea culpa*, since in my doctoral dissertation I made the same fundamental, but uncorroborated, assumption [50].

The original idea of complete regression systems hails from Frisch [23], although he, and his intellectual successors, like the 1989 Nobel Prize winner Trygve Haavelmo [28] and the 1984 Nobel Prize winner Sir Richard Stone [75], [74], did not completely understand their properties. Kalman proved in the winter of 1990/91 that Complete Least Squares (CLS) projections always exist (i.e., can always be computed) and that they are the best of all linear projections to compute Grassman coordinates [41]. They are the best projections in the classical sense that any other linear projection is worse since it produces the same noise matrix as the CLS projector plus an unspeci ed positive nonde nite matrix.

This concept of best complete projector is a much more general concept than the best linear unbiased estimator (or b.l.u.e.) in the conventional statistical literature. The conventional b.l.u.e. is (a) prejudiced with respect to the data selection, while the complete projector is not, and (b) the conventional b.l.u.e. relies on the narrow concept of only the trace of a noise matrix, while the complete projector does not [15]. The complete projector takes the whole noise matrix into account. Using the methodology of this paper it is easy to show why the conventional b.l.u.e. is severely biased in a non - statistical, algebraic and even common sense.

2.4.1. *De nition of Least Squares Projection.* In matrix notation, the de nition of this CLS projector is deceptively simple [16]. The most general de nition of a Least Squares (LS) Projection is that it is a projection $\widetilde{\mathbf{P}}$ (with the de ning characteristic $\widetilde{\mathbf{P}}=\widetilde{\mathbf{P}}^2$, as can easily be checked!), projecting the noise, residuals, or unsystematic variation $\widetilde{\mathbf{x}}=\widetilde{\mathbf{P}}\mathbf{x}$, with an unsystematic noise matrix $\widetilde{\Sigma}$ such that

$$\widetilde{\Sigma}=\widetilde{\Sigma}\Sigma^{-1}\widetilde{\Sigma}$$

Any other linear projection will give a larger noise matrix, such that $\widetilde{\Sigma}\geq\widetilde{\Sigma}\Sigma^{-1}\widetilde{\Sigma}$. This bound for positive semi-de nite (partitioned) matrices was found earlier by Bekker [11], [12]. The de ning equality provides the true meaning of least squares, being the best projection resulting in least residual noise . The LS noise projection is given by the matrix $\widetilde{\mathbf{P}}=\widetilde{\Sigma}\Sigma^{-1}$(Check that $\widetilde{\mathbf{P}}=\widetilde{\mathbf{P}}^2$). Consequently, the systematic projection $\widehat{\mathbf{P}}$ (Check again that $\widehat{\mathbf{P}}=\widehat{\mathbf{P}}^2$), projecting the signal, or systematic variation $\widehat{\mathbf{x}}=\widehat{\mathbf{P}}\mathbf{x}$, is $\widehat{\mathbf{P}}=\widehat{\Sigma}\Sigma^{-1}$ with an systematic covariance matrix $\widehat{\Sigma}=\Sigma-\widetilde{\Sigma}$.

2.4.2. *Theorem for Computing Complete Least Squares.* The following Theorem instructs how to compute the complete *set* of these LS projections.

Theorem 2.1. (Complete Least Squares) *For* all *linear models* $\mathbf{A}\widehat{\mathbf{x}}=\mathbf{0}$ *with* $rank(\mathbf{A})=q$*, which are identi able from the data covariance matrix* $\Sigma>\mathbf{0}$ *and, which by de nition, must satisfy the orthogonality requirement* $\widehat{\mathbf{x}}\perp\widetilde{\mathbf{x}}$*, it is true that the LS noise covariance matrix*

$$\widetilde{\Sigma}^{LS}\equiv\Sigma\mathbf{A}'(\mathbf{A}\Sigma\mathbf{A}')^{-1}\mathbf{A}\Sigma$$

[15]The trace of a square matrix is the sum of its diagonal elements.

[16]Indeed, so deceptively simple that many commentators ask us the question: What s new? But that s a question familiar to anybody who has introduced new concepts into academic debate, starting with the Greek philosopher Socrates in Plato s *Meno*, and it should not deter us from innovation.

is the best, most efficient, or smallest in the sense that any other noise matrix is larger, so that for any noise matrix $\widetilde{\Sigma}=\widetilde{\Sigma}^{LS}+\mathbf{Q}$*, where* $\mathbf{Q} \geq \mathbf{0}$*, a positive semi-de nite matrix.*

Proof. Cf. [41]. ∎

This LS noise covariance matrix does satisfy the de nition of LS Projection, since

$$\widetilde{\Sigma}\Sigma^{-1}\widetilde{\Sigma} = \Sigma\mathbf{A}'(\mathbf{A}\Sigma\mathbf{A}')^{-1}\mathbf{A}\Sigma\Sigma^{-1}\Sigma\mathbf{A}'(\mathbf{A}\Sigma\mathbf{A}')^{-1}\mathbf{A}\Sigma = \Sigma\mathbf{A}'(\mathbf{A}\Sigma\mathbf{A}')^{-1}\mathbf{A}\Sigma = \widetilde{\Sigma}$$

Furthermore, it satis es the Linear Model requirement, since

$$\mathbf{A}\widehat{\Sigma}^{LS} = \mathbf{A}\Sigma - \mathbf{A}\widetilde{\Sigma}^{LS} = \mathbf{A}\Sigma - \mathbf{A}\Sigma\mathbf{A}'(\mathbf{A}\Sigma\mathbf{A}')^{-1}\mathbf{A}\Sigma = \mathbf{A}\Sigma - \mathbf{A}\Sigma = \mathbf{0}$$

Because the exact values of the model s projection coefficients $\mathbf{A}$ remain essentially undetermined (only the value of the structural invariant q can be identi ed, while in nitely different projection directions can be chosen, in principle), CLS noise remains essentially undetermined too. However, CLS noise must be nite, since the data covariance matrix is nite.

Which coefficient matrix should be chosen? Answer: any $q \times n$ matrix $\mathbf{A}$ will do! The Theorem only states that for each projection and each corresponding coefficient matrix $\mathbf{A}$, there exists a corresponding LS noise matrix $\widetilde{\Sigma}^{LS}$. Since there are in nite projections possible based on linear combinations of the n orthogonal projections, there are in nite model projection coefficients $\mathbf{A}$, and corresponding noise matrices, compatible with the data. Thus the projection coefficients of $\mathbf{A}$ don t have a unique interpretation, as is erroneously assumed in virtually all statistical disciplines of learning which use Galton s regression. Transformation of the data $\mathbf{x}$ by pre-multiplication with a positive de nite matrix $\mathbf{F}$

$$\mathbf{F}\mathbf{x} = \mathbf{F}\widehat{\mathbf{x}} + \mathbf{F}\widetilde{\mathbf{x}} = \widehat{\mathbf{z}} + \widetilde{\mathbf{z}}, \text{ where } \widehat{\mathbf{z}} = \mathbf{F}\widehat{\mathbf{x}} \text{ and } \widetilde{\mathbf{z}} = \mathbf{F}\widetilde{\mathbf{x}}$$

does not essentially alter the data structure. It only rotates the frame of systematic data reference, as can be easily observed from the following expression

$$\mathbf{A}\mathbf{F}^{-1}\mathbf{F}\widehat{\mathbf{x}} = \mathbf{B}\widehat{\mathbf{z}} = \mathbf{0}, \text{ where } \mathbf{B} = \mathbf{A}\mathbf{F}^{-1}$$

so that we can uniquely retrieve $\mathbf{A} = \mathbf{B}\mathbf{F}$. Thus the lateral shifts discussed in Section 2.1 and these rotations of the data frame of reference don t alter the structural information in the data [17]. Consequently, LS projections don t identify the model. They only compute the Grassmanian coefficients. Thus a LS projector is a true, unidirectional scienti c measurement tool. It is a tool to make a picture in one particular direction. Many pictures can and should be taken of a multi-dimensional object in many directions and translated into linear models with computed coefficients. But only a particular set of these LS pictures provides a consistent picture,

[17] The point about positive semi-de nite rotations was not understood by even an eminent physicist as E. T. Jaynes [38], [57], who states in his Book References on the Internet about the author: This astonishing economist condemns not only our Bayesian analysis, but virtually every useful thing ever done in data analysis, going back to Gauss. (Cf. his URL http://omega.albany.edu:8008/JaynesBook.html). Some classical econometricians came close, but they did also not understand it [11]. However, the point was apparently not lost on the 1990 (joint) Nobel Prize winner Harry Markowitz, cf.. the Appendix to his book [64]

and scienti c insight when combined with some physical knowledge. Our super l-ter is thus a method of generating a complete set of pictures which enables the identi cation of the system.

This discussion brings us to the educational historical example of the recognition of the helical DNA structure, mentioned in the Introduction, as an illustration of our scienti c approach. Rosalind Franklin and the (joint) 1962 Nobel Prize winner Maurice Wilkins at Kings College took many fuzzy X-ray diffraction (= correlation) pictures of DNA. Diffractometry of a protein crystal put on goniometer is done in circles in all three Euclidean dimensions. Each noisy diffractograph is uniquely translated to a noisy set of many atomic distance pictures by the exact formula (= lens) of Bragg s Law. The genius of (joint) 1962 Nobel Prize winners James Watson and Francis Crick was to combine a (lucky) set of atomic distance pictures with some physical knowledge about possible chemical bonding of the known atoms of DNA to produce DNA s helical structure. The importance of the discovery of the helical structure is that this structure explains why DNA can reproduce itself exactly in cell-division. This exact reproduction is essential for the combinatorial transmission law of the hereditary characteristics (like Galton s stature) discovered in 1865 by Mendel by non-probabilistic, combinatorial breeding procedures [18]. Despite claims by statisticians in the 1910s and 1920s, e.g., by R. A. Fisher, no probability is involved neither in the hereditary processes, nor in the identi - cation of these hereditary processes, even though the hereditary results may look probabilistic [19].

[18]Mendel s revolutionary insights remained unnoticed by the scienti c community for several decades, because his discovery was premature. Nothing new here. In the middle of the 19th century Mendel s mathematical modeling methodology, by means of which he interpreted his results, was foreign to the biologists way of thinking in the middle of the 19th century, just as our CLS methodology apparently remains foreign to economists and nancial analysts, who continue to prefer unidirectional projections and probabilistic research procedures. Like Mendel, Kalman and the author have been disappointed that no one has undertaken our experiments, even after an explicit challenge by the author [53], p. 1285. But there is some hope. After all, Mendel s 1865 paper was rediscovered in 1901, at the turn of the twentieth century, when the application of mathematics to biology had become common place.

[19]The famous misunderstanding of this crucial issue is now attributed to the mathematician Hardy [33]. It was codi ed into the mistaken Hardy-Weinberg Law in population genetics. Hardy took Mendel s 3:1 transmission ratio of the dominant characteristic as the underlying relative frequency, or probability, to be found expressed in the observed population statistics. The 3:1 ratio (actually 2.98:1. Cf. Galton s similar statistical approximation fudging discussed in the rst Appendix) found in the rst generation of hybrids does NOT imply that, in the long run, there will be three times the number of dominant forms as recessive observed in the population, because that is not how Markovian transmission works. Based on Hardy s misunderstanding, statisticians tried in the 1910s - 1920s to go the reverse route and to discover from observed population statistics what the underlying genetic transmission ratio should be. In particular, based on Hardy s misunderstanding of Mendel s methodology Fisher developed his vacuous Maximum Likelihood method [21], [22]. Cf. [41]. But Fisher s statistical inference method is NOT what Mendel did with his hybridization experiments, or how he found his genetic law. Why do we even discuss Mendelian hybridization experiments in a paper on nancial risk? Well, there may be an important lesson in Mendel s careful non-random, combinatorial research of plant hybridization for option pricing specialists, who study the binomial pricing of options and try to infer from the observed prices what the underlying option pricing law is, like the postulated Black-Scholes Law, which has clearly not been scienti cally corroborated. Although the resulting market transaction prices may look probabilistic, this doesn t imply that they are. In fact, it is very difficult to establish from observations that a process is random [45], [44], [72]. A more detailed discussion of such and other implications for nancial market research must await another paper. For Mendel s

Similarly, the CLS Theorem translates the $n(n-1)/2$ bivariate correlations of the data covariance matrix Σ into a multidimensional re ection, the Grassman coefficient matrix $\mathbf{A}$, with minimum noise covariance matrix $\widetilde{\Sigma}^{LS}$. Like any picture, the coefficient values of $\mathbf{A}$ depend on the direction in which the picture was taken, i.e., the projection direction. However, only pictures with the same system invariant q consistent with the system invariant of the data will provide the required consistency within the complete set of all possible pictures. If the $q^{model} \neq q^{data}$ the coefficient values of the various $\mathbf{A}$ matrices will be inconsistent and unreliable. Small changes in the data set will cause such inconsistent coefficient values to vary wildly and even ip- op in sign! [54]

How does the camera of the CLS Theorem lter the signals from the data? For a given $rank(\mathbf{A}) = q$, where q represents the number of independent simultaneous equations of the model, as determined from the information matrix, we can always compute the corresponding LS noise matrix and thus the LS projection noise, or residuals, using the projection

$$\widetilde{\mathbf{x}}^{LS} \equiv \widetilde{\Sigma}^{LS}\Sigma^{-1}\mathbf{x}$$

But this means that the exact LS signal is given by

$$\widehat{\mathbf{x}}^{LS} \equiv \widehat{\Sigma}^{LS}\Sigma^{-1}\mathbf{x} = (\Sigma - \widetilde{\Sigma}^{LS})\Sigma^{-1}\mathbf{x} = (\mathbf{I} - \widetilde{\Sigma}^{LS}\Sigma^{-1})\mathbf{x}$$

since $\widehat{\Sigma} = \Sigma - \widetilde{\Sigma}$, or, signal = data - noise. Consequently, the systematic *LS projector* is

$$\mathbf{P}^{LS} \equiv (\mathbf{I} - \widetilde{\Sigma}^{LS}\Sigma^{-1})$$

Let s elaborate a bit the radical implications of the CLS Theorem for a better understanding. The CLS Theorem states that no matter how you decompose the data covariance matrix Σ into a systematic covariance matrix $\widehat{\Sigma}$, so that $\mathbf{A}\widehat{\Sigma} = \mathbf{0}$, and an unsystematic (noise) covariance matrix $\widetilde{\Sigma}$, so that $\Sigma = \widehat{\Sigma} + \widetilde{\Sigma}$, the LS procedure can *always* compute the Grassman coordinates in $\mathbf{A}$, for *any* $\mathbf{A}$. This implies that we have to compute *all* possible $\mathbf{A}$, and not just the one a researcher happens to fancy. We must investigate the whole range of LS projections allowed by the data, to obtain a complete research picture, since any selection of $\mathbf{A}$ would be prejudiced. One particular selection of $\mathbf{A}$ does *not* determine the values of the true underlying system that generated the data. Thus even in the simplest, bivariate case one needs to present at least the two extreme orthogonal projection results to establish the whole projection range allowed by the epistemic uncertainty in the data [20].

How does the CLS projector differ from the conventionally de ned ordinary (OLS) and general (GLS) least squares projections ? By taking account of the *complete* set of covariances among the available data! Both the OLS and GLS projections are unidirectional projections and don t represent all the available data. They both present only one particular and very incomplete picture of the data by selecting one particular partition of the data covariance matrix. The scienti c

crucial 1866 paper, access the MendelWeb, which was written up in *The Sciences* magazine of The New York Academy of Sciences [65]

[20] Anybody who is familiar with the economic and nancial literature knows rsthand that this has not been done (yet). Interestingly, some Bayesians, like E. T. Jaynes, came close in their methodology, but a true understanding of the issues was prevented by their unnecessary probability assumptions. Similarly for some classical econometricians, like [49], [47], [11].

error - Galton s Error - of the classical LS projections is to exclude some essential data covariance evidence, by not analyzing the (co-)variances of the so-called regressors. Historically, these regressor covariances have been considered a nuisance under the label multi-collinearity problem. This prejudiced selection of a particular data covariance partition becomes very clear when we analyze *all* possible classical orthogonal (or perpendicular) projections [62]. One nds that all other possible projections form linear combinations (a cone) of these orthogonal projections. We will now illustrate our general discussion with the speci c results for the simple bivariate case.

2.5. **Classical Orthogonal Projections.** It is crucial for the understanding of our new methodology to note that two variables imply two orthogonal LS projections, or in general, that n variables imply n orthogonal LS projections. Let s focus rst to the bivariate case and compute symbolically its *two* extreme noise and signal covariance matrices, assuming rst no noise in variable 1, $\widetilde{\sigma}_{11} = 0$, followed by no noise in variable 2, $\widetilde{\sigma}_{22} = 0$. The rst orthogonal projection gives $\mathbf{A}_1 = \left[\begin{array}{cc} 1 & -\frac{\sigma_{11}}{\sigma_{12}} \end{array} \right]$, the second $\mathbf{A}_2 = \left[\begin{array}{cc} 1 & -\frac{\sigma_{12}}{\sigma_{22}} \end{array} \right]$. Using *Theorem 2.* to compute the two corresponding extreme LS noise matrices $\widetilde{\Sigma}_1^{LS}$ and $\widetilde{\Sigma}_2^{LS}$,we can now nd that the LS noise resulting from the corresponding projections is

$$\widetilde{\sigma}_{11} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}, \text{ when } \widetilde{\sigma}_{22} = 0 \text{ (= the conventional case)}$$
$$\text{and}$$
$$\widetilde{\sigma}_{22} = \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}, \text{ when } \widetilde{\sigma}_{11} = 0 \text{ (= the reverse case)}$$

This implies that the *percentage* of epistemic uncertainty of the data is independent of the projection direction, since

$$\frac{\widetilde{\sigma}_{11}}{\sigma_{11}} = \frac{\widetilde{\sigma}_{22}}{\sigma_{22}} = 1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} = 1 - \frac{\beta_2}{\beta_1} = 1 - \rho_{12}^2$$

In the bivariate case one needs to present at least the two extreme orthogonal projection results to establish the complete projection range allowed by the uncertain data. Conventional statisticians compute only one of these projections, namely the vertical projection (normalized on $\mathbf{x}_1$, with $\widetilde{\sigma}_{22} = 0$), thus the conventional noise matrix looks like:

$$\widetilde{\Sigma} = \left[\begin{array}{cc} \widetilde{\sigma}_{11} & 0 \\ 0 & 0 \end{array} \right]$$

where the model uncertainty variance is assumed to reside in the rst variable $\widetilde{\sigma}_{11} = \widetilde{\sigma}_{11} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$, since the noise variance of the variable on which is projected is assumed to be $\widetilde{\sigma}_{22} = 0$ and the model coefficients are

$$\mathbf{A}_2 = \left[\begin{array}{cc} 1 & -\beta_2 \end{array} \right] = \left[\begin{array}{cc} 1 & -\frac{\sigma_{12}}{\sigma_{22}} \end{array} \right]$$

Or, in more familiar notation, $\mathbf{A}_2\mathbf{x}' = \widehat{x}_1 - \widehat{x}_2\beta_2 = 0$, so that $\widehat{x}_1 = \widehat{x}_2\beta_2$, with the coefficient $\beta_2 = -\frac{\sigma_{12}}{\sigma_{22}}$. This lower, vertical projection is the only one presented in the economics and nancial literature for bivariate data sets. But, of course, to be complete, we have, similarly, for the upper or horizontal projection (similarly normalized on $\mathbf{x}_1$), which is classically known as the inverse regression, the noise matrix:

$$\widetilde{\Sigma} = \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{\sigma}_{22} \end{bmatrix}$$

where now all model uncertainty variance is assumed to reside in the second variable $\widetilde{\sigma}_{22} = \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}$, since the noise variance of the variable on which we project is assumed to be $\widetilde{\sigma}_{11} = 0$ and

$$\mathbf{A}_1 = \begin{bmatrix} 1 & -\beta_1 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{\sigma_{11}}{\sigma_{12}} \end{bmatrix}$$

Or, in more familiar notation, $\mathbf{A}_1\mathbf{x}' = \widehat{x}_1 - \widehat{x}_2\beta_1 = 0$, so that $\widehat{x}_1 = \widehat{x}_2\beta_1$, with the coefficient $\beta_1 = -\frac{\sigma_{11}}{\sigma_{12}}$. To emphasize this point, notice how each particular projection result (= picture) depends on the projection direction, which decides where the residual noise will reside.

There is an interesting contrast between this scienti c approach to modeling based on nite data and the speculative approach based on the presumption of in nite theoretical probability. The computed slope coefficient β remains uncertain, because the data are uncertain and not linearly exact. We can compute many possible values of β, all uniquely and exactly computed from the data between the two extreme nite measurement boundaries established by the data orthogonal projections: $\beta_2 < \beta < \beta_1$, because the data and their (co-) variances are uncertain but nite. In contrast, the (theoretical) probability approach presumes that there can be in nite empirical data and thus, a priori, an in nite slope coefficients β. Consider, for example, the conventional statistical presumption of a normal distribution of β with in nite tails. But this presumption is not corroborated by any empirical experience, since nobody has ever observed in nite empirical observations, only nite ones.

It is epistemologically not clear how the in nite tails of presumed continuous probability distributions relate to the nite empirical observations via a unique direct mapping, without some act of faith. Still, statisticians customarily postulate such a mapping, by assuming that the observations are drawn from an in nite and continuous universe. Thus the valid question can been raised if such a speculative approach can be called scienti c, i.e., a method which relates valid conclusions uniquely to the empirical data? The answer must be negative, since even cosmologists acknowledge the nite boundary and the energy granularity of the physical Universe.

Having provided all the ingredients for linear identi cation from empirical data, we can now discuss the geometric uncertainty relationship for bivariate systems and actually observe why we use the metaphor of a camera for scienti c measurement.

3. BIVARIATE GEOMETRIC UNCERTAINTY RELATIONSHIP

Simple trigonometry shows that for bivariate systems the degree of identi cation or model determination is given by the conventional coefficient of determination. Using the preceding discussion, it is easy to prove that the coefficient of determination

$$\rho_{12}^2 = \frac{\beta_2}{\beta_1} = \frac{tan(\theta_1)}{tan(\theta_1 + \theta_2)}$$

where the angles$\theta_1 = a\tan(\beta_2)$ and $\theta_3 = a\tan(1/\beta_1)$, with $\theta_1 + \theta_2 + \theta_3 = \frac{\pi}{2}$ radians, as in Fig.1.

Notice how the angle θ_2 between the cone formed by the two systematic slope lines of the elementary LS measures the nite modeling uncertainty. The true systematic slope coefficient β lies in between these two extreme slopes and is uncertain, i.e., inexact, although it is uniquely computed determined by a particular LS projection. In principle, there may exist an in nite number of LS projections between the two extreme elementary LS projections. Each of these projections must be a linear combination of these two extremes. The closer the slopes of the two extreme elementary projections are together, the more certain we can be of the model coefficient β.

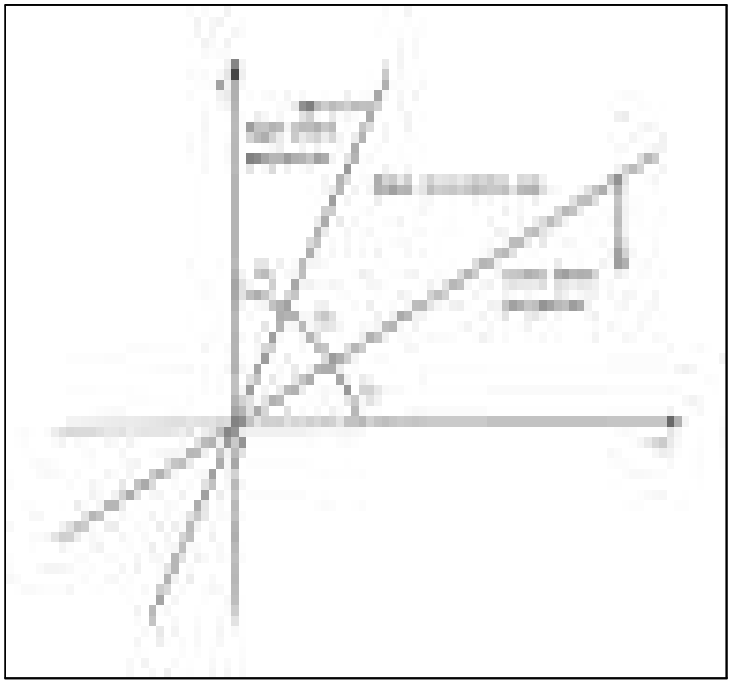

FIGURE 1

One of the many intermediate LS projections is special, and therefore also prejudiced with respect to the data: the so-called principal components (PC) projection. The PC projection projects all the data orthogonally toward a unique slope line and xes thereby a unique slope coefficient β^{PC}. But the uncertainty of the data does not allow to uniquely x β as representative of the data, since the model must express the data uncertainty to be representative of all the available information, including the quality of that information. Scienti cally it is not allowed to substitute certainty for where there can be only uncertainty, otherwise the model does not express the uncertainty characteristics of the data. A model, or map, must be an honest and unique expression of the observed data to be called scienti c [21].

The camera of linear modeling produces only projections and we can choose freely the direction of such projections. Therefore, it is crucial to provide a complete

[21]Interestingly, many geographic maps are not scienti c by this rigorous criterion but approximating abstractions, as Hall explicitly acknowledges [29], Introduction and Ch. 1.

picture of the covariance of the data set and not a prejudiced selection, otherwise our conclusions may be misleading. A unidirectional picture is a very incomplete and arti cially certain representation or mapping of reality.

3.1. **Noise/Signal Ratio.** The information engineering concept of the signal/noise ratio, respectively its inverse noise/signal ratio, has a clear geometric interpretation in the bivariate model. The following two equivalent noise/signal ratio representations follow immediately from the preceding uncertainty relationship by the usual sinusoidal relationships. (Appendix 2. provides the elementary proofs). The noise/signal ratio is:

$$\frac{N}{S} = (1 - \rho_{12}^2)/\rho_{12}^2 = \frac{\sin\theta_2}{\sin\theta_1 \sin\theta_3}$$

Perhaps it is not so remarkable that this noise/signal ratio, measuring the relative lack of exact information, is expressed in wave - like terms. The amount of noise, uncertainty or unsystematic variation is measured in sinusoidal terms relative to the product of two expressions measuring the extremes of the systematic variation, i.e., the signal in the data, also in sinusoidal terms.

Sinusoidal expressions are periodic. Their values repeat themselves for every 2π rad the frame of data reference rotates through (Actually, because of symmetry of the covariances, the periodicity is π rad). This inherent periodicity of the noise/signal ratio also suggests that by an appropriate rotation of the translated frame of data reference, i.e., by an appropriate affine transformation (= translation & rotation) of the original raw data variables $\mathbf{y}$, the noise can be reduced relative to the signal, so that the signal can be received with less distortion. This affine transformation of the data is similar to the ne tuning of the lens in a camera or the rotation of the dial in a radio receiver.

3.2. **Modeling Uncertainty, Inexactness, or Non-Uniqueness.** Using the preceding uncertainty relationship and the noise/signal ratio, we have now, at least, ve *equivalent* ways of presenting bivariate modeling certainty and uncertainty:

(1) Bivariate modeling uncertainty

(i) $|\Sigma| > 0$, the data covariance matrix is positive de nite, i.e., its determinant is positive;

(ii) $0 < \rho_{12}^2 < 1$, the coefficient of bivariate determination shows less than complete explanation or inexact determination;

(iii) $\beta_2 < \beta < \beta_1$, the upper and lower projection slopes do not coincide;

(iv) $0 < \theta_2 < \frac{\pi}{2}$, there exists an uncertainty gap in between the orthogonal frames of data reference;

(v) $\frac{N}{S} > 0$, the noise/signal ratio is positive, since the inexact data contain some noise, together with the signal.

(2) Bivariate modeling certainty

(i) $|\Sigma| = 0$, the data covariance matrix is singular, i.e., its determinant equals zero;

(ii) $\rho_{12}^2 = 1$, the coefficient of bivariate determination shows complete explanation or exact determination;

(iii) $\beta_2 = \beta = \beta_1$, the upper and lower projection slopes do coincide;

(iv) $\theta_2 = 0$, there exists no uncertainty gap in between the orthogonal frames of data reference;

(v) $\frac{N}{S} = 0$, the noise/signal ratio is zero, since the exact data consist only of the signal.

Although one would think that the case of modeling certainty is the limit of the case of modeling uncertainty, with uncertainty approaching zero, this is *not* true [45]. The case of certainty is a *theoretical abstraction* that can be approached in empirical science by improved measurement, but that can empirically never be reached. Therefore in empirical science the determinant of the data covariance matrix $|\Sigma| > 0$, always [22].

3.3. **Relationship to conventional t-statistic.** The relationship between the bivariate Noise/Signal ratio and the conventional t-statistic is straightforward, via a direct mapping. This mapping throws a surprising light on the interpretation of the t-statistic. For the bivariate case, the conventional t-statistic does contain the same information already provided in our bivariate data analysis based on the 2×2 data covariance matrix and it doesn t add anything new.

The t-statistic is conventionally used to test if the computed projection (slope) coefficient differs from zero based on presumed Gaussian probability. For the bivariate case the t-statistic is de ned by *two* equivalent expressions, using the two extreme LS projection values of β, namely β_1 and β_2:

$$t = \frac{\beta_2}{\sqrt{\frac{\widetilde{\sigma}_{11}}{\sigma_{22}}}} = \frac{\beta_1}{\sqrt{\frac{\sigma_{22}}{\widetilde{\sigma}_{11}}}}$$

This means that, in the bivariate case, the t-statistic is truly directionless : it is independent of the projection direction. This not so well-known fact becomes even more obvious when, after a few simple substitutions by the previously derived expressions for β_1, β_2 and $\widetilde{\sigma}_{11}$, respectively $\widetilde{\sigma}_{22}$, we nd that the t-statistic is also equivalent to

$$t = \frac{\sigma_{12}}{\sqrt{|\Sigma|}}$$

Thus the t-statistic tests if the measured covariance between two data series, σ_{12}, is larger than the volume of the epistemic uncertainty in the data, as measured by the magnitude or determinant of the data covariance matrix, $|\Sigma|$ (See [30], for a detailed explanation of the geometric volume of determinants).

Finally, from the preceding expression it immediately follows that the t-statistic is also equivalent to the following expression

$$t = \sqrt{\frac{\rho_{12}^2}{1 - \rho_{12}^2}} = \sqrt{\frac{S}{N}}$$

Thus the bivariate t-statistic is the square root of the Signal/Ratio and can be computed from the coefficient of bivariate determination!

[22] Often the limited storage capacity of numerical computer registers give the false impression that some $|\Sigma| \neq 0$, although it should be exactly $|\Sigma| = 0$. For a humorous discussion of this problem with Russian missile tracking based on unidirectional regression projection, see [41], Section 10 Prejudice in Mathematics.

In the bivariate case, the t-statistic is essentially a test to see if the signal is larger than the noise in the data, be they stochastic, or deterministic, normally distributed, or not adhering to any probability law! The presumed probability distributions are immaterial for the information obtained and are only introduced to play the scienti cally irrelevant game of signi cance testing. [23]

For example, on the basis of the conventional Student s t-distribution, derived from an assumed Gaussian probability distribution, and using subjective signi - cance preferences, statistics often requires that $t \geq 3.078$ for 10% signi cance testing. Using our formulas, this implies that the coefficient of determination must be $\rho_{12}^2 \geq 0.9045$, corresponding with a signal/noise ratio $\frac{S}{N} \geq 9.4741$, or, equivalently, an upper limit on the noise/signal ratio $\frac{N}{S}$ □ 0.1056. Similar but stricter standards of measurement accuracy are included in Table 1. These conventional standards of statistical measurement accuracy are obviously subjective and ad hoc and nd no basis in the data [24].

Table 1.	Conventional	Standards	of	Accuracy
Indicator	$\mathbf{t} \geq$	$\boldsymbol{\rho}_{12}^2 \geq$	$\frac{S}{N}$	$\frac{N}{S}$
10% signi cance	3.078	90.45%	9.4741	10.56%
5% signi cance	6.314	97.55%	39.8665	2.51%

These are conventional maximum boundary standards for the accuracy of scienti c measurement equivalent to those of genetic research, surprisingly very much stricter than those of atomic particle research, but much lower than those of bacterial research, as can be observed in Table 2.

Table 2.	Epistemic	Uncertainty	in	Science
Research Object	**Approximate Mass**	**Approximate Diameter**	**Uncertainty of Position**	**Noise/Signal Ratio**
Units	(kg)	(meters) (1)	(meters) (1)	[=(2)/(1)]
Homo Sapiens	$9.0E+01$	$1.6E-01$	$5.6E-36$	$3.5E-35$
Amoeba	$4.0E-09$	$1.6E-04$	$1.1E-13$	$7.2E-10$
Bacterium	$1.0E-15$	$1.0E-06$	$2.3E-10$	$2.3E-04$
Gene	$4.0E-20$	$3.4E-08$	$3.6E-08$	$1.1E+00$
Uranium Atom	$4.0E-25$	$1.0E-10$	$1.1E-05$	$1.1E+05$
Proton	$1.7E-27$	$1.0E-15$	$1.8E-04$	$1.8E+11$
Electron	$9.1E-31$	$1.0E-15$	$7.6E-03$	$7.6E+12$
Bivariate Beta	by statistical	convention		$1.0E-02$

In Table 2. the lower uncertainty bounds and corresponding Noise/Signal ratios are computed by applying Heisenberg s Uncertainty Principle using Planck s fundamental quantum constant (Cf. Appendix III. for the details). From Table 2. it is clear that social scientists cannot claim that they experience too much uncertainty

[23] This term game for something which is considered a serious professional enterprise is not the author s, but of the Dutch econometrician Keuzenkamp s [46], p. ix.

[24] Signal processing and communications engineers would nd these nowadays still unacceptably high noise/signal limits. Try to listen to a radio or telephone or watch a television screen with a 10.56% noise/signal ratio!

in the data, since the particle physicists experience much higher levels of uncertainty. The uncertainty levels of the social scientists correspond more with those experienced and found acceptable in bacteriological and genetic research. Therefore, social scientists, like economists and nancial analysts, should be required to adhere to similar rigorous standards of scienti c measurement precision. The universe imposes a physical minimum boundary for the accuracy of scienti c measurement, which can be improved until this absolute minimum boundary is reached [25]. Adherence to irrational *conventional* boundaries of con dence, signi cance, or, in general, acceptability, does not further science.

4. EMPIRICAL EXAMPLE: MPT MUTUAL FUNDS SELECTION BASED ON BETA

Despite the early recognition of Galton s error, the statistical literature, including the economic and the nancial literature, still reports exclusively the lower projection slope β_2 and the bivariate coefficient of determination ρ^2_{12} $(= R^2)$, but not the upper projection slope β_1. Also, it doesn t report the noise/signal ratios, i.e., ratio of the unsystematic risk to the systematic risk. In other words, it reports only the downward biased computational result of β, often, but not always, together with an indication of the model uncertainty ρ^2_{12} ,but it does not provide the complete picture. This de ciency is even more pronounced for the cases with more than two variables, where it is never reported how the system invariant q is determined, otherwise than from theory. In almost all cases it is (incorrectly) assumed that $q = 1$, the model consists of a single linear equation.

We contend that this selective and biased model presentation of the data has led to persistent and expensive misunderstanding of the concepts of modeling uncertainty and risk in the nancial industry, as the following simple empirical example illustrates.

Current nancial industry presentation standards recommend to select mutual funds by their funds by their risk/return pro le. The risk is measured by the relative rate of return volatility, i.e., as measured relative to that of a benchmark market index, and the return by some average return over a appropriate period [15]. This relative risk measure is called Sharpe s beta [70], [71] [26]. When a fund s beta, β, is below unity, the fund is categorized as defensive, because the volatility of its investment returns is lower than that of the market as a whole. With a β greater than unity, a fund is categorized as aggressive. Finally, with a β equal to unity, the fund is categorized as a neutral or a market index like fund, because it behaves similarly to the selected market index (usually with a considerably smaller number of different assets).

Regrettably, Sharpe s beta is computed and presented by the nancial industry as the lower projection β_2, as recommended, for example, by *The AIMR Performance Presentation Standards* [1], pp. 34 - 35, and [5], pp. 92 - 95, which

[25]The computations for Table 2., based on the approximating computations in Appendix 3. Epistemic Uncertainty in Science, were inspired by the work of the late great science popularizer Isaac Asimov [9].

[26]William Sharpe shared the 1990 Nobel Prize in Economics for his contribution to nancial economics, in particular for his beta concept, which allowed the unique pricing of capital assets. Unfortunately, this paper makes clear that inexact empirical data cannot provide such uniqueness. The CAPM controversy is not new, as a recent volume made clear [34], although our explanation for the beta s uncertainty is.

are adopted as part of the (new) AIMR s Standard of Professional Conduct V.B concerning Performance Presentation. The de cient, but official recommendation concerning the computation and presentation of the beta is now promoted to become a global standard [27]. But these simple computations have led to a severe under-representation of the empirically observed systematic risks of the selected funds by the nancial industry. Therefore the question can be rased if the current recommendations by the AIMR are consistent with its own Standard of Professional Conduct IV, the Relationships with and Responsibilities to Clients and Prospects, in particular Standard IV.A.2 concerning Research Reports and Standard IV.A.3 concerning Independence and Objectivity [28].

This under-representation of systematic investment risk can be demonstrated by looking at how many mutual funds are ranked aggressive, defensive, or equivalent to the market index by Sharpe s beta and how many are truly aggressive defensive or neutral, when taking account of all the modeling uncertainty implied by the data.

For the data we use the computed betas and coefficients of determination in Morningstar s convenient (Windows based) *Principia for Mutual Funds* of July 1995, as released on computer diskettes to the public on December 31, 1995 [29].

	TABLE 3. SYSTEMATIC RISK OF MUTUAL FUNDS	#	%
1.	Morningstar s Principia for Mutual Funds universe, 12/31/95	7,051	
2.	Together with the condition $0 < \rho_{12}^2 \square 1$	3,227	
3.	And with 3-year (Sharpe s) beta $0 < \beta_2$	3,215	
4.	**AIMR Performance Presentation Standards, 1993:**		
	(i) Defensive funds: $0 < \beta_2 < 1$	2,047	63.7
	(ii) Neutral, market index funds: $\beta_2 = 1$	67	2.1
	(iii) Aggressive funds: $1 < \beta_2$	1,101	34.2
	Total funds with measurable systematic market risk	3,215	100.0
5.	**Kalman-Los analysis, 1989:**		
	(i) Defensive funds: $0 < \beta_2 \square \beta_1 < 1$	608	18.9
	(ii) Neutral, market index funds: $\beta_2 = \beta_1 = 1$	18	0.6
	(iii) Aggressive funds: $1 < \beta_2 \square \beta_1$	1,101	34.2
	(iv) Undecided: $0 < \beta_2 < 1 < \beta_1$	1,488	46.3
	Total funds with measurable systematic market risk	3,215	100.0

[27]The original AIMR Performance Presentation Standards [1], which took effect on January 1, 1993, were amended and restated on September 13, 1996 to include some international concerns [5]. This restatement did not amend the incomplete computation of Sharpe s beta. The AIMR Performance Presentation Standards form part of the AIMR s Code of Ethics and Standards of Professional Conduct [6].

[28]New knowledge is not always appreciated. When the author, who is a member of the AIMR, raised these difficult issues in letters of August 2, 1994 and January 3, 1995, respectively, to two successive Directors of Research of the AIMR, his proposals for amendments were twice officially and rmly rejected in writing.

[29]These data diskettes are available, at cost, from Morningstar, Inc., 225 West Wacker Drive, Chicago, Illinois 60606, and are updated quarterly. Morningstar is a respected mutual funds monitor with an excellent reputation that computes the betas and corresponding coefficients of determination of the mutual funds strictly according to the accepted industry standards. According to Morningstar s OnFloppy User s Guide (p.22): Morningstar bases alpha, beta, and R-squared on a least squares regression of the fund s excess return over T-bills compared with the excess returns of the fund s benchmark index. These calculations are computed for the trailing 36-month period.

First, we notice in Table 3. that only 3, 227 out of a total universe of 7, 051 funds have measurable risk, as indicated by a computed coefficient of determination larger than zero, or 45.8% of the total universe. The other funds are younger than 3 years and don t have a 3-year record to base such computations on. However for 12 of these 3, 227 funds the lower beta β_2 equals zero in the two published digits beyond the decimal point. Thus only 3, 215 funds have measured systematic market risk as de ned by the CAPM, or 45.6% of the total universe.

If we accept Sharpe s criterion for selecting funds by their relative volatility or systematic market risk characteristic, then the number of defensive funds selected by correctly implementing Sharpe s beta is 25.6% of the 2, 047 claimed to be defensive by the current industry standards. In addition, the number of actual market index funds is only 26.9% of the 67 funds claimed to be market index funds in this representative data universe. Finally, of the 3, 215 funds for which the appropriate data were available 954, or 45%, could not be categorized as defensive, aggressive or market index, in spite of the claims of the nancial industry.

In addition, we may want to apply the criterion of accuracy of the measurement of this systematic risk as in Table 4.

Table 4. Systematic Risk and Accuracy

Kalman-Los analysis, 1989:	$\frac{N}{S}$ **10.56%**	%	$\frac{N}{S}$ **2.51%**	%
(i) Defensive funds: $0 < \beta_2$ $\beta_1 < 1$	182	40.4	23	28.4
(ii) Neutral, market index funds: $\beta_2 = \beta_1 = 1$	18	4.0	18	22.2
(iii) Aggressive funds: $1 < \beta_2$ β_1	171	38.0	27	33.3
(iv) Undecided: $0 < \beta_2 < 1 < \beta_1$	79	17.6	12	14.8
Funds with measurable systematic market risk	450	100.0	81	100.0

Table 4. shows that accuracy of measurement of the systematic market risk is an important criterion when one insists on truth in advertising. Based on the reasonable criterion of a noise/signal ratio of less or equal to 10.56% - corresponding with $\rho^2_{12} = 0.90$, i.e., 90% con dence in the parlance of conventional statistics - only 450 out of a total universe of 3, 215 funds with measurable market risk pass the test. That is an astonishingly low 14.0% of the total universe! When we increase the measurement accuracy only a bit further to a noise/signal ratio of less or equal to 2.51% - corresponding with $\rho^2_{12} = 0.975$, or 97.5% con dence), only 81 funds, or 2.5% of our universe, pass this simple accuracy test. Based on these nontrivial results of the exceedingly low risk measurement accuracy, professional nancial economists should express a note of concern about the exaggerated advertising claims of the mutual funds industry.

To gain an impression of some of the investment magnitudes involved, look at the following gures. The mutual fund industry in the United States grew from $US\$95$ billion in assets in 1979, to nearly $US\$2$ trillion by the end of 1994, an increase of over 20 times. Even after taking account of consumer price in ation and the resulting loss of purchasing power in the U.S. of more than 90% over the same period, that is still a very sizeable increase in real assets of eleven times in fourteen years.

Most of this increase has actually occurred in the last three years. American investors poured a net $US\$377$ billion into equity mutual funds alone in $1993 - 95$. Since the end of 1994 until the middle of 1996, the Dow Jones industrial average

climbed by nearly 50% and the broader $S\&P500$ index by 46%, increasing America s nancial wealth by $US\$2.4$ trillion, more than the entire annual output of Germany [30].

Compare now these market sizes with the magnitudes of the universes we analyzed. By September 1993 there existed 4,347 open-ended mutual funds. The following year Morningstar monitored about 79% of them. Its *Mutual Funds On-Floppy* universe contained 3,434 funds with an average median market capitalization of $US\$0.5$ billion in net assets by the end 1994. Its updated successor universe, Morningstar s *Principia for Mutual Funds,* used in our analysis, contained already more than double this number at the end of 1995: 7,051 funds. Because of the fast growth in the number of new funds, there were now many more smaller funds include, since the average median market capitalization of this universe is $US\$264.9$ million in net assets. But the more restricted universe of 3,215 funds, on which the conclusions of Table 3. are based, has an comparable average median capitalization of $US\$514.6$ million in net assets, while the universes of 450 funds and of 81 funds have average median market capitalizations of $US\$510.5$ million, respectively $US\$510.4$ million in net assets.

Since this increasingly massive process of mutual fund selection and pricing is biased by the under-representation of market risk, as our analysis suggests, very serious misallocation between the investment alternatives could result, based on their currently presented biased relative risk and return pro les. Also, since a substantial amount of this investment may be hot, these market allocations are not likely to be patient or secure. Indeed, *The Economist* [31] refers to the argument that many mutual-fund investors do not understand what they are doing; and that, when they realize what the risks are, they will ee. There is no reason for panic, however, according to the same article, because of the apparent maturity of the modern investors. The younger investors not only say they accept the risk involved - in a recent survey by American nancial regulators, 94% of investors said they knew they could lose money in share dealings as well as gain it - they also seem, in practice, to respond calmly when prices fall.

The biased published betas do not only raise macro concerns relevant for national policy makers or global asset allocators, but also micro concerns relevant for individual portfolio managers. Since the downward biased beta s are used in the computation of cross hedging ratios, when portfolio positions are hedged by derivatives, like futures, to reduce the systematic risk exposure of these positions, serious doubts should be raised about the effectiveness of such hedges. In our opinion, there is more uncertainty about the systematic risk than current portfolio managers, regulators and the educators of nancial analysts recognize..

A scienti c debate on the issue of the adequacy of a single risk measure for mutual funds, like the beta, is therefore timely. The Securities and Exchange Commission (S.E.C.), in reaction to recent sharp price drops for several supposedly low-risk mutual funds, has asked fund managers to look more carefully at their risk management controls that track derivative positions [31]. The S.E.C. is trying to condense the myriad risks of mutual funds into a single measure that would convey these risks to investors [32]. In 1995 the S.E.C. oated a Concept Release (= White Paper) on the issue, requesting comments on or before July 7, 1995

[30] *The Economist*, July 6, 1996, pp. 18 and 21.
[31] *The Economist*, July 6, 1996, p. 18.

[4]. The comments in this paper, should forewarn the S.E.C. that its quest for a single measure for multi - faceted investment risk is likely to be just as quixotic and fruitless as the quest for a single I.Q. measure when fundamental principles of science are ignored [27].

A complete representation of the empirical systematic uncertainty and risk is required [32]. Thus for the bivariate CAPM two measures must be published: the correlation coefficient ρ_{12} (or, equivalently in the bivariate case, the coefficient of determination ρ_{12}^2) together with the β_2, since all other bivariate measurements can be derived from these two. Next, one must educate the investors about the uncertainty range for β, about $\beta_2 \leqslant \beta \leqslant \beta_1 = \frac{\beta_2}{\rho_{12}^2}$. It was because of the recommendable practice of Morningstar to publish both β_2 and ρ_{12}^2 that we were able to properly categorize the mutual funds, while still using the accepted CAPM categorization.

5. TRIVARIATE SYSTEM IDENTIFICATION FROM INEXACT DATA

Having brought the theory of modeling or system identi cation down to earth (or at least to Wall Street, and perhaps to Washington, D.C.), we can now raise the following question. Does the preceding protocol for our modeling methodology extend to the multivariate case with more than two variables, for example, to the credit scoring models of [8], [7] used in the commercial pricing of distressed sovereign and bank debt? The correct pricing of such distressed debt has become an important global phenomenon, after the banking crises in the United States, Japan, China, Thailand, etc.[33]. And does it extend to Multi-Index Models used in portfolio analysis, as presented, for example, by [20]? Elton and Gruber s simplest Industry Index Model includes three variables: the rates of return of a rm, a market index and one industry index, e.g., the steel industry ([20], p. 164). Our answer is: yes, it does, as is illustrated by the following two gures with the pictures (projection maps) for the trivariate case.

Figure 2. illustrates what is meant with multidimensional modeling *certainty*, when $n = 3$ and $q = 2$. dots in the center of Figure 2 are the observations in the 3D frame of data reference [34]. The origin of this frame of reference is in the center

[32]Of course, an investor can reduce the risks of his portfolio further by appropriate diversi cation, as Markowitz demonstrated in 1952 [63], but that is the topic of another paper with a critique of J. P.. Morgan s RiskMetrics, also based on the fundamental concept of epistemic uncertainty. In this paper, I only add that, while Sharpe s erroneous beta compares with Galton s error of regressing towards the mean, the current practice of factor, or principal components analysis of investment portfolios, based on Asset Pricing Theory (APT) compares more closely to the erroneous practice of I.Q. testing [27]. In this paper we follow deliberately Sharpe s 1963/64 Capital Asset Pricing Model - CAPM - approach to mutual fund selection [70], [71], since that is still the most widely accepted and recommended standard in the nancial industry. We are NOT recommending this bivariate presentation of systematic market risk practice as the only or best presentation of systematic risk. We only propose that the current practice is severely biased and misleading.

[33]In 1991 - 93 the author was a Chief Economist and Economic Advisor for ING Bank in New York. ING Bank (now ING Capital) in New York was then a market maker in distressed sovereign debt of emerging markets, in particular in Latin America, and of distressed corporate debt in North America.

[34]This trivariate case was discussed earlier in the *Eastern Economic Journal* [55], but, regrettably, the Editor forgot to publish the crucial gure. In 1993, the Editor sent an apologetic letter and promised to provide another occasion to publish that gure. His promise never came true, but the gure was nally published in 1994 [58], and 1995 [59] in sources usually not accessed by nancial analysts.

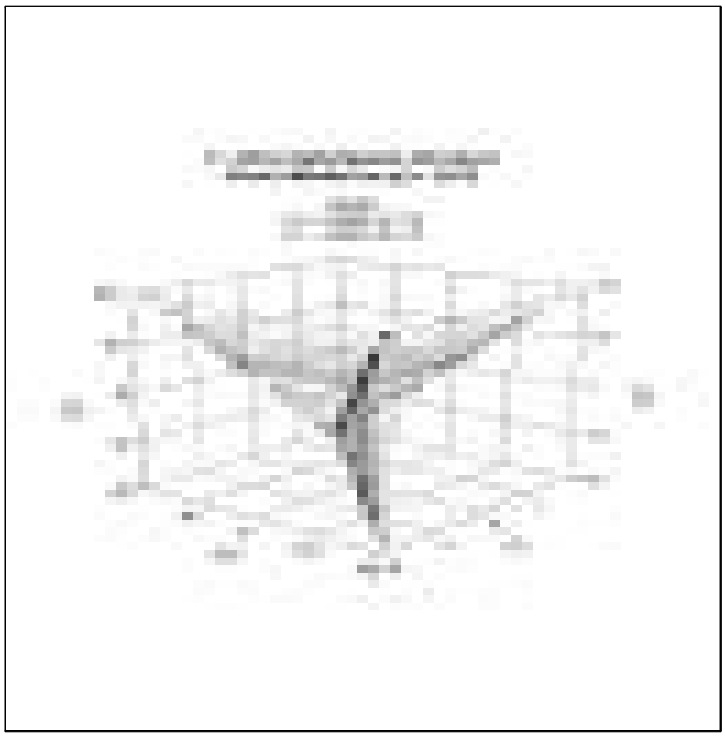

FIGURE 2

of the gure. The observations lie on a straight line, or ray. This ray is the model abstracted from the data. Such a ray model must be described by $q = 2$ linearly independent equations, since a single three - variable equation would describe a plane and leave the Grassman coordinates, or model coefficients, undetermined. In the context of Figure 3. this seems to be trivial. However, considering that statisticians, and in particular econometricians refuse to understand that their problem of multicollinearity is not a problem at all, but the indication that the empirical covariance data exhibit a linear system, this comment is NOT trivial [39].

Notice the projections of the observations and the model on the 2D frames of data reference on the three sides. There we nd in each of the three 2D frames $\rho_{ij}^2 = 1$. Thus there is complete system certainty. The system invariant q is easily identi ed from the exact data.

Let s now turn to the empirical, inexact or noisy case in Fig. 3., which illustrates what is meant with multidimensional modeling *uncertainty*, when $n = 3$ and $q = 2$. The blocks in the center of the gure represent again the cloud of now uncertain empirical data observations, with the origin in the center of the gure. The elongated cigar of the cloud of observations dictates again that the linear model should be a ray through the origin and not a plane. Thus the data dictate that the model abstracted from the data should contain two independent equations, $q^{presumed} = q^{true} = 2$, and not a single independent equation, among the three variables. All too often a single equation, $q^{presumed} = 1$, which describes algebraically a plane, is automatically presumed by the statistical, economic and

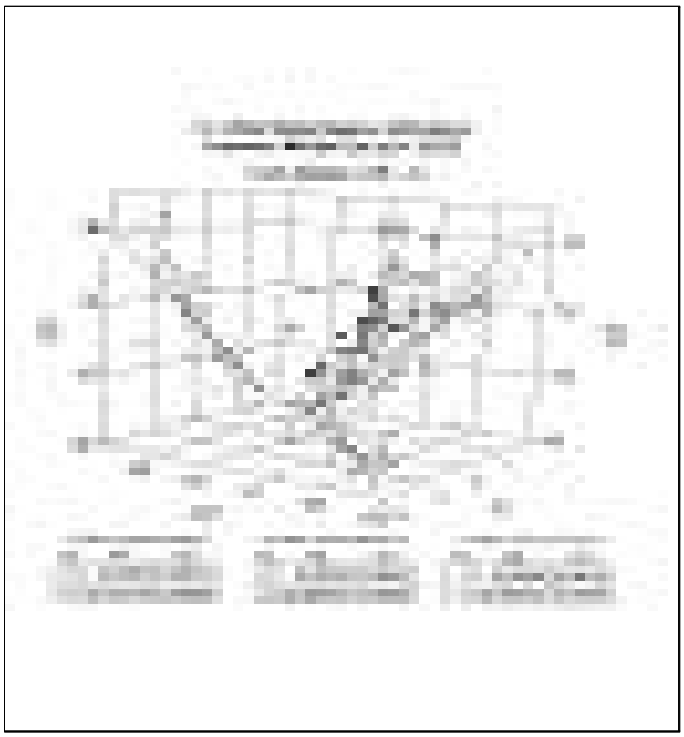

FIGURE 3

nancial analysts. But there exists now an unambiguous Popperian falsi cation test: the Grassman coordinates will remain undetermined (i.e., truly random) if the presumed model consisted of only one relation, $q^{presumed} = 1$, describing a plane and not two relations describing a ray, as reality requires, $q^{true} = 2$. To obtain reliability and not unstable, ip- opping coefficients, the linear model structure, $q^{presumed}$ must equal the data structure, q^{true} [35].

The three vertices, LS1, LS2 and LS3, are the three complete LS boundary projections, which map the data in three different projection directions, depending on where the system uncertainty is allocated. The cone-shaped area between them represents the complete three - dimensional system uncertainty inherent in the data and is computed from the adjoint of the data covariance matrix [73]. Although the true Grassmanian system coordinates within this cone are uncertain, they are clearly bounded by nite boundaries formed by the three vertices, as in the bivariate case.

The ratio of the volume of this cone spanned by the three vertices of the extreme LS projections relative to the volume of the data orthant is the multidimensional extension of the classical bivariate coefficient of determination.

Notice again the projections of the observations and the three boundary projections in the 2 - dimensional frames of data reference. There we nd in each of the

[35]Interestingly, the author nds that mathematicians and engineers have little difficulty understanding this point about the necessary equivalence between model and data structure, but statisticians do.

three 2 - dimensional frames

$$0 < (1 - \rho_{ij}^2)/\rho_{ij}^2 = \frac{\sin\theta_2}{\sin\theta_1 \sin\theta_3} < 1$$

Thus there is, indeed, complete, but limited system uncertainty.

When there is complete system certainty, as in Figure 2, which is clearly an arti cial abstraction, we nd that the uncertainty cone collapses along a ray and in each of the 2 - dimensional frames of data reference $\rho_{ij}^2 = 1$. The extension to still higher dimensional data sets is thus straightforward. For example, in 1991 we published a simple economic forecasting model with $(n, q) = (5, 4)$ [55] [36].

6. CONCLUSIONS AND NEW RESEARCH DIRECTIONS

6.1. **Conclusions.** The new methodology of super ltering by complete system identi cation using simple algebraic geometry, as developed over the past decade by Kalman and Los, is used in this paper to focus on the commitment of Galton s Error in nance, in particular in the bivariate, CAPM based, measurement of systematic risk. Galton s Error is committed when one accepts prejudiced and biased results of incomplete, unidirectional measurement as scienti c facts. This paper demonstrates how to avoid Galton s Error by *complete* scienti c measurement, in particular in nance and economics.

First, we discuss the history and the mathematics of Galton s Error in nance, as exempli ed by the conventional CAPM classi cation of defensive, aggressive and neutral investments. Next, the paper illustrates its mathematical ndings with empirical statistics from mutual funds. Mutual funds conventionally advertise their quantitative systematic market risk as measured by betas. The published betas are incompletely measured by unidirectional bivariate (least squares) projections. Because these projections are unidirectional and incomplete, they ignore vital covariance information and, *a fortiori*, they induce Galton s Error in the systematic risk measurements. Far too many mutual funds are currently marketed as defensive (with low systematic risk) and too few as aggressive (with high systematic risk). Furthermore, the conventional CAPM classi cation doesn t include a category for undecided (with indeterminable systematic risk), although the epistemic uncertainty in the data clearly requires such a scienti c category.

Using the new methodology, a revised classi cation for mutual funds is proposed, which includes the new category undecided, and which makes the conventional bivariate categories defensive, neutral, market-index like and aggressive mathematically precise and scienti cally complete. Using *Morningstar s Principia* database of 7, 051 mutual funds of December 31, 1995, we nd that of a universe of 3, 215 mutual funds with measured bivariate systematic risk, 2, 047 funds, or 63.7%, claim to be defensive, according to the industry standard, conventional CAPM classi cation. But using our new, revised CAPM classi cation based on complete bivariate covariance measurement, only 608, or 18.9% are actually defensive. In other words, of all the mutual funds claimed to be defensive by the nancial industry, only 29.7% are truly defensive. Of the 67 funds (2.1%) which claim to be neutral market-index like funds, only 18 (0.6%) are neutral, market-index like. Again, of the total claim for neutral funds, only 26.9% ts that billing. Finally, only 53.7% of the total number of mutual funds that are classi ed by the published beta

[36] Since $r = n - q = 1$, there was only one major economic factor in the presented ve variable economy, commonly known as the business cycle.

standard as defensive, market index like, or aggressive, can be so classi ed. The remaining 46.3% of the funds are unclassi ed, since the data are too uncertain to draw a rm conclusion.

The substantial under-representation of systematic risk in mutual funds caused by the biased unidirectional projections necessarily must lead to inefficiencies in the capital allocation process, since it distorts the quantitative return/risk pro les of the funds. These quantitative return/risk pro les are used, for example, in computerized fund manager universe searches by pension fund consultants. The conventional downward-biased betas lead to mispricing of mutual funds and other securities. Funds which are conventionally classi ed and priced as defensive are actually aggressive.

The under-representation of systematic risk by the mutual fund industry is in serious con ict with the rules and regulations of the Securities and Exchange Commission (SEC) concerning truth in advertising by investment advisers (e.g., the Investment Advisers Act of 1940, with its Amendments for improved enforcement of the 1980s), or the AIMR s Performance Presentation Standards adopted in 1993 (and amended in 1996), which require performance information to be fair, accurate and complete. The paper demonstrates that the conventional CAPM based unidirectional projections are mathematically incomplete, biased and don t indicate the amount of inaccuracy in the data.

In contrast, our new, complete bivariate projections produce a correct representation of the epistemic uncertainty inherent in the bivariate CAPM measurement of systematic risk. They measure the amount of inaccuracy inherent in the data and produces a complete and unbiased (fair) CAPM classi cation of mutual funds. Some funds are simply not classi able by the bivariate CAPM measurements because of too much data uncertainty.

In a nutshell, our conclusions can have serious consequences for (1) the current trend of bench-marking for style-based asset allocation, since most existent bench-mark classi cations will have to be revised; for (2) the concomitant remuneration schemes of fund managers proposed by sponsors, since fund managers take more risks than conventionally indicated; and for (3) the recent (summer 1995) regulatory proposals of the SEC for the mutual fund investment industry, which may become global standards, and which therefore should be scienti cally correct.

6.2. **New Research Directions.** Because of the fundamental insights obtained in our research, we foresee other revealing applications of our research methodology by revisiting older, or existing published data sets in nancial economics. For example, we are reviewing the original Cobb and Douglas [19] production function data. The crucial scienti c questions are if there actually exists a production function between the three variables, output, capital input and labor input in their data set, and if it is a system with one equation ($q = 1$) or a system with two equations ($q = 2$). The second question was not even raised by Cobb and Douglas. Estimated production functions form the basis for the determination of the growth potentials of countries [77], as pioneered by the 1987 Nobel Prize winner Robert Solow, and of concurrent pension fund requirements [78], as pioneered by the 1985 Nobel Prize winner Franco Modigliani.

Another current research project is to scienti cally review the politically charged debate about the money demand equation, given its widely reported instability [48]. That research project is important since many emerging markets, e.g., mainland

China, are now adopting a monetary policy apparatus modeled on that of the Federal Reserve System of the U.S. However, the relationships between interest rates, GDP and the monetary aggregates (credit) is extremely uncertain and unreliable. Instead of one money demand equation used for the projection of the demanded money supply, the data suggest that a *system* of at least two independent equations is needed, one for the relationship between GDP and money and one between money and interest rates (the third equation is automatically implied [51], [52].

Currently, a global competition in applied econometrics with 35 contenders is underway to enhance the failing credibility of conventional econometrics [61]. But this paper demonstrates that conventional econometrics, so heavily imbued with probability theory, has little to say about *science = analytic data measurement* and *modeling = system identi cation from inexact data*, since it presumes to know, a priori, the invariant q of the data structure, instead of deriving it from the data. This leads to models inconsistent with the data and to unreliable coefficient values. Finally, in nance we will closely scrutinize the current crop of APT multi-index models [20] and the scoring models for the pricing of distressed securities and the determination of the likelihood of corporate bankruptcies [7].

The results of such serious scienti c reviews will be of importance for the scienti c veracity of current growth and development theory, monetary policy recommendations and recommendations for nancial risk measurement in the investment industry.

Various statisticians have asked us how to judge the signi cance of the systems identi ed, when we no longer accept the game of signi cance testing based on presumed probability. The scienti c way of determining the permanence, or inertia, of a system s observed existence, i.e., the data generating system would be to create separate, non - overlapping data sets or independent windows and then to compare the system invariants identi ed from these non-overlapping data sets to determine if they are generated by the same system [37]. The crucial test is if the structural system invariant q is the same in all data sets, indicating system stationarity. If not, the data set s homogeneity and stationarity, respectively the system s integrity, i.e., the value of q, is in doubt. When it is, the system s coefficients, as we observed, may be uncertain, they lie in de nite ranges limited by the projection cones of the data sets. The uncertainty cones identi ed in the windowed data sets should at least overlap to have the possibility of system integrity.

In the past decade, we have surveyed a new territory of scienti c nancial economic research and cut some new trails. Now we sincerely hope that the new generations of students will have the audacity to travel along these still unbeaten trails to correct the errors of the past instead of blindly following the conventional statistical textbooks. Perhaps, these students can use the new research methodology of this paper to review existing economic and nancial data sets and come to exhilaratingly different conclusions from the ones that have been published. Some such effort is already under way in Singapore and Zurich.

[37] Sherry arrived at the same approach when he wanted to test the stationarity of the information processing of both nervous systems and stock markets [72]

7. ACKNOWLEDGMENTS

Since I left Columbia University, as a professional Economist and Senior Economist at the Federal Reserve Bank of New York and Nomura Research Institute, and as Economic Advisor for ING Bank confronted with applied econometric practice in a nancial environment, I developed serious epistemological doubts about the established practices of econometrics, nancial analysis and similar statistical analyses. These doubts culminated in a series of Federal Reserve Research Papers in 1985, 1986 and 1987, which were nally published in refereed journals in 1989, 1991 and 1992. My fundamental methodological doubts were, once again, reinforced by the lecture on Stochastic Modeling Without Probability by Dr. Rudolf E. Kalman on May 3, 1993 at the Sixth International Symposium on Applied Stochastic Models and Data Analysis at the University in Chania, Crete, Greece. The presumption of probability is irrelevant for system identi cation, which requires a geometric algebraic approach. Since then Kalman has proved that there is very little, if any, scienti c basis for the presumption of the empirical existence of Kolmogorovian probability.

I presented my 3D maps of Section V. on March 11, 1992 in a very well attended invited lecture at The New York Academy of Sciences (invited by the 1980 Nobel Prize winner Lawrence Klein); on March 16, 1992 in an invited Engineering and Statistics Seminar at M.I.T. (with some help of the 1970 Nobel Prize winner Paul Samuelson); on April 3, 1992 in an invited presentation on the identi cation of complex empirical systems at the Symposium on The Interpretation of Quantum Theory: Where Do We Stand? at the Italian Academy for Advanced Studies in America, Columbia University, New York City, NY, April 1-4, at the International Conference (by invitation only; invited by the 1972 (joint) Nobel Prize winner Kenneth Arrow) on New Research on Identi cation in Econometrics , Department of Economics and Operations Research of Stanford University, Palo Alto, CA, November 4 - 6 as well as at the aforementioned International Symposium in Chania on May 4, 1993. Section 2. of this paper was presented on March 18, 1994 at the 20th Annual Convention of the Eastern Economic Association in Boston. I thank Dr. Nancy Wulwick for her invitation to this convention and her editorial comments and I thank my students and colleagues at the Nanyang Technological University in Singapore for their constructive critique.

8. APPENDIX I: GALTON S REGRESSION TOWARDS MEDIOCRITY

Galton s 1886 paper, far from obscure, was presented on January 21, 1886 to the Royal Society of London, with Professor Stokes presiding.[25] It extended and complemented earlier remarks of Galton in his 1985 Presidential Address to the Anthropological Section of the British Association, and in his often cited 1985 paper in the Journal of the Anthropological Institute Regression towards Mediocrity in Hereditary Stature [24]. From these two papers statisticians inherited the very concept of regression, in particular, of regression towards the mean. Both papers are revelations of prejudiced Victorian science.

Anticipating our comments in Section 3. and our Figure 1., Galton s 1986 article contains on page 55 two gures, Figs. 5 and 6, with his regression lines, i.e., his downward biased LS projections towards the mean. His Figure 5 is labeled Mean Stature of Children of Mid-Parents of Various Heights, (For Galton s lial regression), for which Galton computed a ratio of regression, or what now is

called a slope coefficient of $w = \tan 3 = 2/3$, i.e., smaller than unity. His Figure 6 is labeled Mean Stature of Brothers of Men of Various Heights (for Galton s fraternal regression), for which he computed the same slope coefficient of $w = 2/3$. On pages 56 and 57 of his article Galton discussed converse ratios of regression, or slope coefficients of the reverse regressions, for some of which he computed the value of 1/3. Galton was not very precise about the computed values, since he considered the approximations, for which he often preferred to substitute his *a priori* beliefs about their true values.

The protocol of our research paradigm in Section 3 of our paper leads us to conclude that the coefficient of determination in Galton s regressions had the value of only $2/3x1/3 = 2/9$ or 22.2%. Thus Galton s so - called universal law of regression towards mediocrity explained only 22.2% of the already small variation in the heights of the adults in his data set. The other 77.8% of the variation is unsystematic or uncertain. In short, his ignorance was three times larger than his knowledge . Why did Galton not notice this glaring de ciency in his conclusions ?

The solution to this question is found in the Appendix to Galton s 1886 article, which was prepared by his collaborator J. D.. Hamilton Dickson, Fellow and Tutor of St. Peter s College, Cambridge, England. The Appendix shows, rst, that both Galton and Dickson erred in the mathematics of covariance (Cf. [25], pp. 50 and 57); also expressions (1) and (2) for the exponent in the bivariate normal distribution on p. 63 in Dickson s Appendix, are incorrect). Secondly, Galton and Dickson erred with the mathematics of (upward biased) reverse regressions, since they did not normalize them on the same variables as the original (downward biased) regressions. In other words, they compared apples and oranges!

As explained in Section 3, Galton should have compared the slope coefficient of 2/3 of the lower bound projection with the inverse of the slope coefficient of the reverse regression, i.e., $1/(1/3) = 3$, of the upper bound projection. The gap between the values of 3 and 2/3 is caused by the uncertainty of the data. Instead, they compared the slope of 2/3 with that of 1/3 and took them for being similar. This simple mathematical misunderstanding still prevails among statisticians.

Galton and Dickson focused too much on the smoothed, continuous surface of the abstracted bivariate frequency distribution and were thus clearly misled by their probability considerations. Consequently, Galton and Dickson could not understand the mathematical concept of uncertainty, or inexactness, as was some decades later expressed by the correlation coefficient between two variables and by its square, the coefficient of determination. Of course, current statisticians no longer have this excuse for their confusion.

Galton collected two distinct data sets of the statures of 783 brothers and of 205 couples of parents and their 930 adult children of both sexes. He found the mean, and median, of his combined data set of 2,123 adults to be 68.3 inches. Galton measured the dispersion around the mean by a quartile deviate of 1.7 inches, which he called the *probable error*. (Our emphasis). This is now conventionally replaced by the so-called standard error. Thus the mean/variance ratio of his data was about 23.6 times and thus one must wonder about the relevance of Galton s regression research, since the variation in the data is exceedingly small. The heights in his restricted data set of 2,123 adults varied between 64.9 inches ($= 5'4.9''$) and 71.7 inches ($= 5'11.7''$).

In addition, the quality of the observations in the second set was acknowledged to be bad. As Galton self formulated it: There is in many cases considerable doubt whether the measurements refer to the height with the shoes on or off; many entries, I fear, only estimates, and the heights are commonly given only to the nearest inch (Galton, 1886, p. 52).

In addition, Galton smoothed the data, without scienti c justi cation, by discarding the outliers in his combined data set or, as he called them, the irregularities, and stated bluntly: These are unimportant in the present inquiry and I disregard them. (Galton, 1886, p. 43). Galton massaged his data further to create greater homogeneity and to make his data t the exponential law of frequency of error (p. 46), now conventionally known as the Gaussian or normal distribution. In true Victorian fashion, he transmuted all the female heights to their male equivalents by multiplying them by a constant coefficient, which as regards my data is 1.08, he wrote (on p. 46; again later p. 52), without any speci c justi cation

A close reading of his article reveals other prejudices Galton harbored with respect to his data. When Galton computed a slope coefficient that he considered too large, he unfailingly discarded the result. Thus when he regressed the statures of all children on the statures of men of the same height, he stated: They yield a ratio of regression of 0.40 instead of 0.33 as above. I disregard it, and adopt the latter, namely $w = 1/3$ (p. 55).

Despite all these de ciencies in his scienti c research, Galton still asserted that his regression towards the mean was a universal anthropological law and he, again, surreptitiously introduced probability theory, when he stated on (p. 50):

> It is a universal rule that the unknown kinsman in any degree of any speci ed man, is *probably* more mediocre than he. Let the relationship be what it may, it is safe to *wager* that the unknown kinsman of a person whose stature is 68 $1/4 + x$ inches, is of some height 68 $1/4 + x'$ inches, where x' is less than x. (Our emphasis).

He stated as the reason for this universal rule two causes : (1) statistical constancy, or what now is called stationarity, and (2) the reasonable *presumption* of similarity between a sample of the original population and a sample of their kinsmen in any speci ed degree, or what is now called homogeneity of the data set. (Our emphasis).

We emphasize that Galton s statements were all unfounded assertions: he did not scienti cally establish the stationarity, the homogeneity of his data set, or the independence of his observations. This practice of presuming or asserting stationarity, homogeneity and independence still prevails among statisticians and statistical analysts. But homogeneity was clearly in doubt, since Galton s data consist of two sets of practically independent observations, i.e., two distinct data sets, which were collected in completely different fashion at different times, according to Galton s own descriptions (p. 52; on p. 59 he even calls the second set less trustworthy). In fact, the asserted stationarity and homogeneity of his combined data set is still an open scienti c question (Cf. [72], Chapters 2 and 3).

9. APPENDIX II: ALGEBRAIC GEOMETRIC MEASUREMENTS FOR $n = 2$

First we establish the bivariate uncertainty relationship, or noise/signal ratio, between the coefficient of determination ρ_{12}^2 and the slope angles θ_1, θ_2 and θ_3.

Proposition 9.1.

$$\frac{N}{S} = \frac{1-\rho_{12}^2}{\rho_{12}^2} = \frac{\sin\theta_2}{\sin\theta_1 \sin\theta_3}$$

Proof. For the bivariate case, we have the relationship between the coefficient of determination ρ_{12}^2 and the slope coefficients β_2 and β_1 (Cf. Fig.1.)

$$\rho_{12}^2 \equiv \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} = \left(\frac{\sigma_{12}}{\sigma_{11}}\right)\left(\frac{\sigma_{12}}{\sigma_{22}}\right) = \tan\theta_3 \tan\theta_1 = \frac{\tan\theta_1}{\tan(\theta_1+\theta_2)} = \frac{\beta_2}{\beta_1}$$

By expanding the tangens expressions into sinuses and cosinuses expressions, we nd

$$\rho_{12}^2 = \frac{\sin\theta_1 \cos(\theta_1+\theta_2)}{\cos\theta_1 \sin(\theta_1+\theta_2)}$$

Substituting this expression into the bivariate noise/signal ratio and expanding the result using some well known trigonometric expressions and canceling terms, we have

$$\begin{aligned}
\frac{N}{S} &= \frac{1-\rho_{12}^2}{\rho_{12}^2} = \frac{1}{\rho_{12}^2} - 1 \\
&= \frac{\cos\theta_1 \sin(\theta_1+\theta_2) - \sin\theta_1 \cos(\theta_1+\theta_2)}{\sin\theta_1 \cos(\theta_1+\theta_2)} \\
&= \frac{\cos\theta_1 \left[\sin\theta_1 \cos\theta_2 + \cos\theta_1 \sin\theta_2\right] - \sin\theta_1 \left[\cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2\right]}{\sin\theta_1 \cos(\theta_1+\theta_2)} \\
&= \frac{\sin^2\theta_1 \sin\theta_2 + \sin^2\theta_1 \sin\theta_2}{\sin\theta_1 \cos(\theta_1+\theta_2)} \\
&= \frac{\sin\theta_2}{\sin\theta_1 \cos(\frac{\pi}{2}-\theta_3)} \\
&= \frac{\sin\theta_2}{\sin\theta_1 \sin\theta_3}
\end{aligned}$$

∎

Next, we establish the relationship between the conventional t-statistic, the epistemic uncertainty in the data, as measured by the determinant of the data covariance matrix $|\Sigma|$ and the noise/signal ratio for the bivariate case $\frac{N}{S}$.

Proposition 9.2.

$$t \equiv \frac{\beta_2}{\sqrt{\frac{\tilde{\sigma}_{11}}{\sigma_{22}}}} = \frac{1/\beta_1}{\sqrt{\frac{\tilde{\sigma}_{22}}{\sigma_{11}}}} = \frac{\sigma_{12}}{\sqrt{|\Sigma|}} = \sqrt{\frac{S}{N}}$$

Proof. Let s rst look at the t-statistic for β_2, for which

$$t \equiv \frac{\beta_2}{\sqrt{\frac{\tilde{\sigma}_{11}}{\sigma_{22}}}}$$

Substituting the appropriate expressions in for β_2 and $\widetilde{\sigma}_{11}$ we nd

$$\begin{aligned} t &= \frac{\frac{\sigma_{12}}{\sigma_{22}}}{\sqrt{\frac{\left(\sigma_{11}-\frac{\sigma_{12}^2}{\sigma_{22}}\right)}{\sigma_{22}}}} = \frac{\frac{\sigma_{12}}{\sigma_{22}}}{\sqrt{\frac{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}{\sigma_{22}^2}}} \\ &= \frac{\sigma_{12}}{\sqrt{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}} = \frac{\sigma_{12}}{\sqrt{|\Sigma|}} \end{aligned}$$

our rst expression.. But then also

$$t = \frac{\sigma_{12}}{\sqrt{|\Sigma|}} = \frac{\sqrt{\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}}}{\sqrt{(1-\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}})}} = \sqrt{\frac{\rho_{12}^2}{(1-\rho_{12}^2)}} = \sqrt{\frac{S}{N}}$$

Similarly

$$t \equiv \frac{1/\beta_1}{\sqrt{\frac{\widetilde{\sigma}_{22}}{\sigma_{11}}}}$$

Again, substituting the appropriate expressions in for β_1 and $\widetilde{\sigma}_{22}$ we nd

$$\begin{aligned} t &= \frac{\frac{\sigma_{12}}{\sigma_{11}}}{\sqrt{\frac{\left(\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}\right)}{\sigma_{11}}}} = \frac{\frac{\sigma_{12}}{\sigma_{11}}}{\sqrt{\frac{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}{\sigma_{11}^2}}} \\ &= \frac{\sigma_{12}}{\sqrt{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}} = \frac{\sigma_{12}}{\sqrt{|\Sigma|}} \end{aligned}$$

∎

This clearly demonstrates that the bivariate t-statistic is independent on the direction of projection, since it depends only on the value of the observed data covariance σ_{12} and the magnitude of the determinant $|\Sigma|$,which is a measure of the magnitude of the model uncertainty, i.e., the lack of linear dependency in the data.

10. APPENDIX III: EPISTEMIC UNCERTAINTY IN SCIENCE

The following exposition is a condensation of the essential points of [9], to demonstrate the only true physical limits on numerical measurement. Expressing physical measurement uncertainty by the symbol Δ, Heisenberg s Uncertainty Principle, or Principle of Indeterminism [36], is formalized by the inequality

$$\Delta p.\Delta mv \geq \frac{h}{4\pi} J.s$$

Measuring position p in meters, mass m in kilograms, and velocity v in meters per second, the fundamental and exceedingly small energy graininess of the Universe is measured by Planck s constant $h = 6.6256 \times 10^{-34} J.s$. (The unit of measurement $Js = Joule.second = \frac{kg.meter^2}{\sec}$). The well-known ratio of the circumference of a circle to its diameter $\pi = 3.1416$. Thus Heisenberg s Uncertainty Principle states that the product of uncertainty in position and uncertainty in momentum (= mass $\times$ velocity) is greater than or, *at best*, equal to

$$\frac{h}{4\pi} J.s = 5.2725 \times 10^{-35} J.s$$

This is the *physical lower bound for the precision of empirical measurement.* For relatively low velocities, mass is independent of velocity, i.e., the relativity relationship can be ignored, so that the product $\Delta mv = m\Delta v$ and

$$\Delta p.\Delta v \geq \frac{5.2725 \times 10^{-35}}{m} \frac{meter^2}{\text{sec}}$$

Let now the uncertainty in velocity be numerically equal to the uncertainty in position (velocity v is measured in meters per second and position p in meters), then

$$(\Delta p)^2 \geq \frac{5.2725 \times 10^{-35}}{m} meter^2$$

so that the uncertainty in position is

$$\Delta p \geq \frac{7.2612 \times 10^{-18}}{\sqrt{m}} meter$$

With this equation the minimal possible uncertainty in position of an object can be determined, given its mass. Using the approximate diameter of an object as unit of measurement of its position, the uncertainty in its position can be expressed as a non-dimensional number, or percentage, and presented as the following Noise/Signal ratio

$$\frac{N}{S} = \frac{\Delta p}{p} \geq \frac{7.2612 \times 10^{-18}}{p \times \sqrt{m}}$$

The minimal numerical value of this Noise/Signal ratio for some objects of scienti c investigation are given in Table 2 in the main text of this paper. It is important to note that there is nothing about this physical lower bound on the precision of empirical measurement that relates to probability. This boundary is imposed by the energy architecture of the Universe in which we live.

References

[1] *Performance Presentation Standards.* Association for Investment Management and Research, Charlottesville, VA, 1993.

[2] Atomic, molecular and optical science: An investment in the future. Washington, D.C., 1994.

[3] *Digital Humans: A Multimedia Tour of the Visible Human Project.* Multimedia Medical Systems, Charlottesville, VA, 1995.

[4] Improving descriptions of risk by mutual funds and other investment companies. Technical Report 17 CFR Parts 239, 270 and 274 [Release Nos. 33-7153; 34-35546; IC-20974; File No. S7-10-95], Securities and Exchange Commission, March 29 1995.

[5] *AIMR Performance Presentation Standards Handbook 1997.* Association for Investment Management and Research, 2nd edition, 1996.

[6] *Standards of Practice Handbook.* Association for Investment Management and Research, 7th edition, 1996.

[7] E. I. Altman. Credit-scoring models and the valuation of xed-income securities and commercial loans. CREFS Seminar presentation at SAB/NBS Nanyang Technological University, June 21 1996.

[8] E. I. Altman, R. G. Haldeman, and P. Narayanan. Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 1(1):29 54, 1977.

[9] I. Asimov. *The Certainty of Uncertainty.* Avon Books, New York, NY, 1976.

[10] G. S. Becker. *A Treatise on the Family.* Harvard University Press, Cambridge, MA, 1981.

[11] P. Bekker, A. Kapteyn, and T. Wansbeek. *Measurement Error and Endogeneity in Regression: Bounds for ML and IV Estimates*, pages 85 103. Springer Verlag, Berlin, 1984.

[12] P. A. Bekker. *Essays on Identi cation in Linear Models with Latent Variables.* PhD thesis, Katholieke Hogeschool (Catholic University) Tilburg, 1986.

[13] C. Benzing and K. Dunleavy. A comment on "scienti c" economic analysis. *Eastern Economic Journal*, 17:523 525, 1992.
[14] P. L. Bernstein. *Against the Gods: The Remarkable Story of Risk*. John Wiley and Sons, Inc., New York, NY, 1996.
[15] F. Black. Return and beta. *The Journal of Portfolio Management*, 45:8 18, 1993.
[16] J. Bring. A geometric approach to compare variables in a regression model. *The American Statistician*, 50(1):57 62, 1996.
[17] P. Bryant. Geometry, statistics, probability: Variations on a common theme. *The American Statistician*, 38:38 48, 1984.
[18] E. Burmeister, R. Roll, and S. A. Ross. *A Practitioner s Guide to Arbitrage Pricing Theory*. The Research Foundation of The Institute of Chartered Financial Analysts, Charlottesville, VA, 1994.
[19] C. W. Cobb and P. H. Douglas. A theory of production. *American Economic Review*, 18:139 165, 1928. Supplement.
[20] E. J. Elton and M. J. Gruber. *Modern Portfolio Theory and Investment Analysis*, chapter 8, The Correlation Structure of Security Returns: Multi-Index Models and Grouping Techniques, pages 160 180. John Wiley and Sons, Inc., 5th edition, 1995.
[21] R. A. Fisher. On an absolute criterion for tting frequency curves. *Messenger of Mathematics*, 41:155 160, 1912.
[22] R. A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700 725, 1925.
[23] R. Frisch. *Statistical Con uence Analysis by Means of Complete Regression Systems*, volume 5. University of Oslo Economic Institute, 1934.
[24] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15:246 263, 1885.
[25] F. Galton. Family likeness in stature. *Proceedings of the Royal Society of London*, 40:42 63, 1886.
[26] C. Gorman. Man of the year: The disease detective. *Time*, 148(27):24 31, December 30, 1996 - January 6, 1997 1996.
[27] S. J. Gould. *The Mismeasure of Man*. W. W. Norton and Company, New York, NY, 1981.
[28] T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:Supplement, 1944.
[29] S. S. Hall. *Mapping the Next Millenium: How Computer-Driven Cartography is Revolutionizing the Face of Science*. Vintage Random House, New York, NY, 1993.
[30] J. Hannah. A geometric approach to determinants. *The American Mathematical Monthly*, 103(5):401 409, 1996.
[31] S. Hansell. S.e.c. asked to study derivatives in funds. *The New York Times*, June 16:D8, 1994.
[32] S. Hansell. U.s. seeking mutual fund risk gauge: Wants a simple system to inform investors. *The New York Times*, June 20:D1 D2, 1994.
[33] G. H. Hardy. Mendelian proportions in a mixed population. *Science*, 28:49 50, 1908.
[34] D. R. Harrington and R. A. Korjczyk, editors. *The CAPM Controversy: Policy and Strategy Implications for Investment Management*. Association for Investment Management and Research, 1993.
[35] C. R. Harvey. The world price of covariance risk. *The Journal of Finance*, 46:111 157, 1991.
[36] W. Heisenberg. The physical content of quantum kinematics and mechanics. In J. A. Wheeler and W. H. Zurek, editors, *Quantum Theory and Measurement*, pages 62 84. Princeton University Press, Princeton, NJ, 1983. English translation from original German publication with the title "Uber den anschaulichen Inhalt der quantumtheoretischen Kinematik und Mechanik, Zeitschrift fur Physik, 43 (1927) pp. 172 - 198.
[37] D. G. Herr. On the history of the use of geometry in the general linear model. *The American Statistician*, 34:43 47, 1980.
[38] E. T. Jaynes. Commentary on two articles by c. a. los. *Computers and Mathematics With Applications*, 24:265 275, 1992.
[39] R. E. Kalman. We can do something about multicollinearity. *Communications in Statistics - Theory and Methods*, 13:115 125, 1984.
[40] R. E. Kalman. Lecture at kyoto prize celebration. 1985.
[41] R. E. Kalman. *A Theory for the Identi cation of Linear Relations*, pages 117 132. North - Holland Publihing Co., Amsterdam, 1991.

[42] R. E. Kalman. *Nine Lectures on Identi cation*. Springer Lecture Notes on Economics and Mathematical Systems. Springer Verlag, Berlin, 1993.
[43] R. E. Kalman. Randomness reexamined. *Modeling, Identi cation and Control*, 15(3):141 151, 1994.
[44] R. E. Kalman. Addendum to "randomness and probability". *Mathematica Japonica*, 41(2):463, 1995.
[45] R. E. Kalman. Randomness and probability. *Mathematica Japonica*, 41(1):41 58, 1995.
[46] H. A. Keuzenkamp. *Probability, Econometrics and Truth: A Treatise on the Foundations of Econometric Inference*. PhD thesis, Catholic University Brabant, 1994.
[47] S. Klepper and E. E. Leamer. Consistent sets of estimates for regressions with errors in all variables. *Econometrica*, 52:163 183, 1984.
[48] D. E. Laidler. *The Demand for Money: Theories and Evidence*. Dun-Donnelley, Inc., New York, NY, 1977.
[49] E. E. Leamer. *Speci cation Searches, Ad Hoc Inference with Nonexperimental Data*. John Wiley and Sons, New York, NY, 1978.
[50] C. A. Los. *Econometrics of Models with Evolutionary Parameter Structures*. PhD thesis, Columbia University, 1984.
[51] C. A. Los. Collinearity analysis of a simple money demand equation. Technical Report Research Paper No. 8604, Federal Reserve Bank of New York, 1986.
[52] C. A. Los. Why there is still no empirical evidence for a money demand equation! Technical Report Research Paper No. 8614, Federal Reserve Bank of New York, 1986.
[53] C. A. Los. Identi cation of a linear system from inexact data: A three variable example. *Computers and Mathematics with Applications*, 17:1285 1304, 1989.
[54] C. A. Los. The prejudices of least squares, principal components and common factor schemes. *Computers and Mathematics with Applications*, 17:1269 1283, 1989.
[55] C. A. Los. A scienti c view of data analysis. *Eastern Economic Journal*, 17:61 71, 1991.
[56] C. A. Los. Reply to benzing s and dunleavy s comments on "a scienti c view of economic data analysis". *Eastern Economic Journal*, 17:526 531, 1992. The Editor forgot to print two essential Figures, which nally appeared in Los, 1992.
[57] C. A. Los. Reply to e. t. jaynes and a. zellner s comments on my two articles. *Computers and Mathematics With Applications*, 24:277 288, 1992.
[58] C. A. Los. The measurement of complex empirical systems. In L. Accardi, editor, *The Interpretation of Quantum Theory: Where Do We Stand?*, pages 243 256. Instituto della Enciclopedia Italiana fondata da G. Treccani (distributed by Fordham University Press), Roma, 1994. This volume represents the proceedings of the Symposium on "The Interpretation of Quantum Theory: Where Do We Stand?" of the Italian Academy for Advanced Studies in America at Columbia University, New York, April 1 - 4, 1992.
[59] C. A. Los. A scienti c view of economic and nancial data analysis. In C. H. S. Janssen, J. and C. Zopounidis, editors, *Advances in Stochastic Modelling and Data Analysis*, pages 111 127. Kluwer Academic Publishers, Dordrecht, 1995.
[60] C. A. Los and C. M. Kell. How to determine the corank and noise level of a system? In *Identi cation and System Parameter Estimation*, pages 599 606, Oxford, UK, 1988. International Federation of Automatic Control, Pergamon Press. Selected Papers from the 8th IFAC/IFORS Symposium, Beijing, PRC, 27 - 31 August 1988.
[61] J. R. Magnus and M. S. Morgan. An experiment in applied econometrics. *Journal of Applied Econometrics*, pages 213 216, 1995.
[62] M. S. Margolis. Perpendicular projections and elementary statistics. *The American Statistician*, 33:131 135, 1979.
[63] H. M. Markowitz. Portfolio selection. *The Journal of Finance*, 7:77 91, 1952.
[64] H. M. Markowitz. *Mean - Variance Analysis in Portfolio Choice and Capital Markets*, chapter Appendix: Elements of Matrix Algebra and Vector Spaces, pages 347 368. Blackwell, Oxford, 1987.
[65] G. Mendel. Experiments in plant hybridization. *Journal of Heredity*, 42(1), 1951. English translation from Mendel s original German lecture (1865) and publication with the title "Versuche uber P anzen-Hybriden," Verhandlungen des naturforchenden Vereins (Proceedings of the Natural History Society in Brunn), 4 (1866), and now, in corrected form, available on the World Wide Web via URLhttp://www.netscape.org/MendelWeb.
[66] G. Rhodes. *Crystallography Made Crystal Clear*. Academic Press, Inc., New York, NY, 1993.

[67] J. L. Rodgers, W. A. Nicewander, and L. Toothaker. Linearly independent, orthogonal, and uncorrelated variables. *The American Statistician*, 38:133 134, 1984.
[68] S. A. Ross. The abitrage theory of capital asset pricing. *Journal of Economic Theory*, 13:341 360, 1976.
[69] D. J. Saville and G. R. Wood. A method for teaching statistics using n-dimensional geometry. *The American Statistician*, 40:205 214, 1986.
[70] W. F. Sharpe. A simpli ed model for portfolio analysis. *Management Science*, 9:277 293, 1963.
[71] W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19:425 442, 1964.
[72] C. J. Sherry. *The Mathematics of Technical Analysis: Applying Statistics to Trading Stocks, Options and Futures*. Probus Publishing Co., Chicago, IL, 1992.
[73] R. Sinkhorn. The range of the adjugate map. *Mathematics Magzine*, 66(2):109 113, 1993.
[74] R. Stone. Linear expensiture systems and demand analysis, an application to the pattern of british demand. *Economic Journal*, 64:511 527, 1954.
[75] R. Stone. *The Measurement of Consumer s Expenditure and Behavior in the United Kingdom 1920 - 38*, volume 1. Cambridge University Press, Cambridge, UK, 1954.
[76] G. Thomas and J. O Quigley. A geometric interpretation of partial correlation using spherical triangles. *The American Statistician*, 47:30 32, 1993.
[77] Worldbank. *The East Asian Miracle*. Oxford University Press, Oxford, 1993.
[78] Worldbank. *Averting the Old Age Crisis*. Oxford University Press, Oxford, 1994.
[79] D. J. Zimmerman. Regression toward mediocrity in economic stature. *The American Economic Review*, 82:409 429, 1992.

Nanyang Technological University, Singapore 639798
E-mail address: ACALOS@ntu.edu.sg