Reply to Benzing's and Dunleavy's Comments on "a Scientific View of Economic Data Analysis"

Cornelis A. Los *

Cynthia Benzing and Kevin Dunleavy (B&D) provide an erroneous, but not uncommon, interpretation of the scientific problem of modeling from inexact data and of its solution. To correct their error and clarify the concepts of exact and inexact modeling, I begin by asking the following: What model can be identified from data such as those in Figure 1? Specifically:

1. Are these data series exact or inexact and are they stochastic or not ?
2. By how many linear relationships can these data be minimally described ? As Einstein once said: "A scientific explanation must be as simple as possible, but not simpler."
3. What do these relationships look like ?

In this simple case the answers to the three questions are given at once by plotting the data in their 3 - dimensional data space. The black dots in the center of Figure 2. are the data points. The arced dots in the three basis planes of (x1,x2), (x2,x3) and (x1,x3), respectively, are the projections of these 32 observations. These projections are connected to the data points by drop lines. Note that:

(1) The data are obviously exact: they lie along a ray, as is also shown by their 2 - dim. projections on the basis planes.
(2) The data can be minimally described by two independent linear equations.
(3) This two - equation model can be written in convenient matrix notation as $A'x = 0$, where A' is the coefficient matrix of order $(q \times n) = (2 \times 3)$ and x the matrix of 32 observations on 3 variables, i.e. of order $(n \times T) = (3 \times T)$.

There is no scientific justification for the statistician's distinction between "explained" and "explanatory" variables on the basis of only the data, since *a linear model is nothing but the minimal linear combination of exact data series*. Note from Figure 2. that x1 is positively related to x2 and simultaneously negatively to x3. Consequently, x2 is also negatively related to x3.

I constructed this example to explain what is meant by exact data and exact linear modeling. Series x1 happens to be an *empirical* data series, measured as deviates from their mean, while x2 and x3 are constructed from x1 using the model. The model realizes all the data and the data are realized by the model. The model is thus an isomorphism of the exact data. Note that the concept of exactness does not say anything about stochasticity or probability. Series x1, and thus x2 and x3, could be stochastic or deterministic. That question is irrelevant for identification.

The following example of inexact data contains empirical economic data previously used in two technical articles [Los, 1989b, 1992]. Since one picture is worth a thousand words, it suffices to summarize that discussion by using pictures analogous to the ones above. (These pictures are not in the referred articles).

The cloud of 32 black empirical data dots are plotted in their 3 - dim. data space in Figure 3. The arced dots are their projections on the respective 2 - dim. basis planes. This time only the drop lines for the projection on the (x2, x3) plane are drawn to not clutter the picture too much.

The projections show positive correlation between x1 and x2, negative correlation between x1 and x3 and negative correlation between x2 and x3, as is evident in the data covariance matrix in the first North-West quadrant of Figure 4.

* NMB Bank 135 E. 57th Street New York, NY 10022

According to a very useful Theorem of Kalman, each row of the inverse (or rather, the adjoint) of the data covariance matrix in the second, North - East quadrant of Figure 4. is, an elementary, $q = 1$, single equation least squares (LS) regression. This can be demonstrated by normalizing on the diagonal elements in the South-East quadrant of Figure 4. For example, by dividing the first row by the first diagonal element we have the elementary regression of x1 on x2 and x3 (as can be replicated by running the conventional regression software on the data series in Los, 1989b, pp. 1301 - 1302). Similarly, the second row is the regression of x2 on x1 and x3 and the third row of x3 on x1 and x2. (Note that we use deviates from the means of the respective series. Inclusion of the means in the model is straightforward).

To compare these regression results in the South-West quadrant, I normalize, arbitrarily, on the first elements of the rows, by dividing each row by its first element. Other normalizations lead to equivalent results. Note that the system coefficients b1 and b2 of the system coefficient matrix A' = [1 b1 b2] are sign inconsistent with each other. For example, in normalized regressions A and B, b1 is negative, but in regression C it is positive. The b - coefficients of these three regressions lie in different orthants. Ergo, such a one - equation model can not realize the required signs consistent with the data covariance matrix.

Contrary to B&D these elementary problems and their solutions are nowhere explained in the statistical literature; also not by the alleged expert on "reverse regressions," who maintains the prejudice of a priori knowing the number of equations without understanding that precisely this prejudice leads to the observed inconsistency with the complete data. Indeed B&D state, "...reverse regression does not specify an exact system of equations to be used nor does it tell us the interrelationships of the variables in the system." Wrong. It does, as I will now demonstrate.

Let's increase the rank q of the system coefficient matrix A' from 1 to 2. There are three ways of selecting two rows out of the three available in the inverted data covariance matrix. By arbitrary rotation, that is by pre - multiplication by a (2×2) positive definite matrix, I have created some ones and zeros, to reduce the number of valued coefficients and to facilitate comparison of the results.

Thus, analogously to the exact data case, I computed the three normalized least squares simultaneous two - equation systems (H, E), (F, G) and (D, I). Notice from the three South-East quadrants in Figure 5. that now complete sign-consistency results and that, despite the inexactness of the data, the range of values of the relevant A' coefficients is very narrow.

Note that values of the coefficients in matrix A' remain arbitrary up to the multiplication by any positive definite matrix. That is why we state that the only invariant of a linear system is q, the minimal rank of A'. This equivalency result of parametrization is well - known in econometric "simultaneous equation estimation" and in factor analysis. The system coefficient values of A' have no intrinsic value. The "frame of reference" (i.e. the orthants in Figures 2. and 3.) can always be rotated to gain an interpretation that is meaningful to the economic researcher.

Figure 6. summarizes our 3 - dim. results. The three computed LS two - equation models are represented by the arced rays through the black data points in the center of Figure 6. They capture all systematic variation consistent with the data, given the three possible projection choices of assumed exact variables. The corresponding three signal/noise ratios show that the two-equation model explains between 64.8% and 86.5% of the variation of these empirical cross - sectional data. On the basis planes I have projected these three LS rays to show how well they represent the exact components of the covariations between x1 and x2, between x2 and x3, and between x1 and x3.

The rays deviate slightly from each other because of the inexactness of the data or "lack of linear fit." If the data had been less inexact, the black dots would cluster closer along a ray through the cone formed by the three least squares regression results LS1, LS2 and LS3. (The origin of the data is located in the center of the picture). Moreover, these three LS rays would cluster closer together. Of course, thanks to continuity, in the limit the black data points and the three LS rays would coincide as in the exact data case.

A single equation model would represent an indetermined plane through the data ray in the center of Picture 6., which could satisfy only two of the three relevant correlation signs at a time, but not all three simultaneously. The coefficients of a single equation model are indetermined when the data are are realized by a two - equation model. If the prejudice is $q = 1$, then $A' = [1 \;\; b1 \;\; b2] = \lambda. \, [1 \;\; 0 \;\; a13] + (1 - \lambda).[1 \;\; a22 \;\; 0]$, with l anywhere. Thus in case of exact data: the coefficients of the prejudiced one - equation model are indetermined linear combinations of the coefficients of the true exact two - equation model. Most empirical

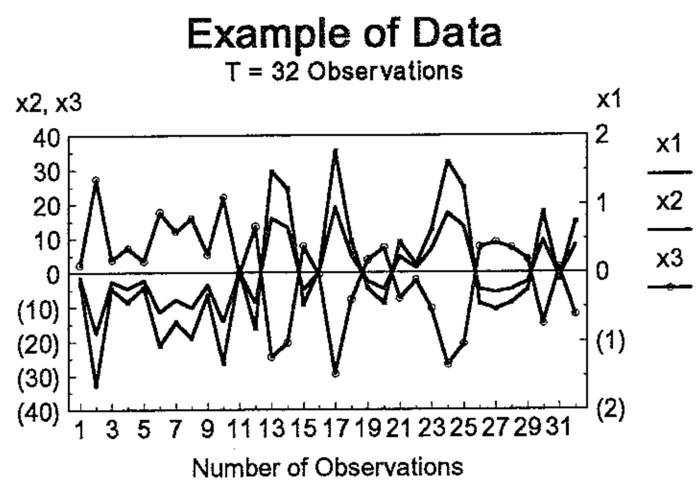

Fig. 1

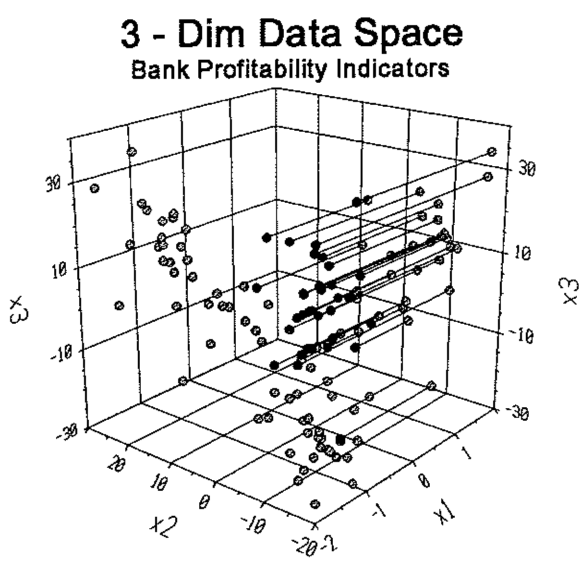

Fig. 3

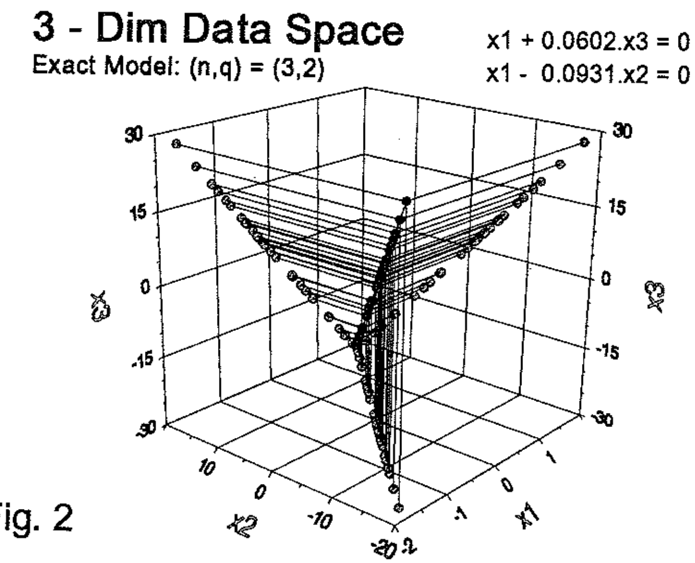

Fig. 2

Covariance Analysis To Identify q

Data Covariance Matrix Σ

	x1	x2	x3
x1	0.7022	6.9040	(10.6826)
x2	6.9040	99.0556	(114.7687)
x3	(10.6826)	(114.7687)	259.2516

Inverted Data Covariance Matrix

	x1	x2	x3
x1	6.0992	(0.2749)	0.1296
x2	(0.2749)	0.0331	0.0033
x3	0.1296	0.0033	0.0107

Normalization on First Column

	x1	x2	x3
A	1	(0.0451)	0.0213
B	1	(0.1205)	(0.0121)
C	1	0.0257	0.0823

Normalization on Diagonal

	x1	x2	x3
x1	1.0000	(0.0451)	0.0213
x2	(8.3013)	1.0000	0.1006
x3	12.1435	0.3123	1.0000

Fig. 4 A' = [1 b1 b2]

economic researchers have observed such "instability of coefficients" resulting from the misspecification of the rank of their model A', like for example in the case of the "money demand equation," the "Phillips curve," or the logarithmic "Cobb Douglas production function". But none of these researchers understood that it was their own prejudice of $q = 1$ that caused such instability and that each of these three examples should have been modeled by q 1 models.

Ours is an attempt to transform economics from subjective (mathematized) philosophy into objective exact science. It is no longer a question if economic and financial data can be analyzed scientifically and economics can become a science. Economics can be "subjected to the same scientific rigor as physics and the other so-called "hard" or "exact" sciences," despite the inexactness of the data. But our critique is that this has not been accomplished by the conventional "soft" or "inexact" probability methods originally proposed by Haavelmo in 1944. Our empirical method has not "already been used in economics to deal with noisy data," as B&D claim, since even B&D correctly acknowledge that we try "to leave no room for subjective analysis."

Probability plays no material role in scientific identification. It is remarkable that economists and statisticians continue to mistake stochasticity for inexactness. What is inexact is not necessarily stochastic, only inexact. Contrary to B&D's beliefs, in quantum physics it is still not established that electrons move probabilistically or that it is our lack of exact measurement that makes us assume they do. After 70 years of debate, it is also likely that also conventional quantum theory is not science but (mathematized) philosophy, a point that Albert Einstein would have appreciated ! [Accardi, Beltrametti, Cartwright, Lahti, Namiki, Van Fraassen and Van der Merwe, 1992].

The point is that Science is measurement. For example, in 1953 James Watson and Francis Crick directly realized the exact 3 - dim. atomic model structure of the DNA molecule from the inexact, or "noisy", 2 - dim. X - ray diffraction pictures taken by Rosalind Franklin and Maurice Wilkins late in 1952. To achieve this model realization, Watson and Crick applied Bragg's exact mathematical Law $2.d.\sin(\theta) = n\lambda$. They directly translated the measured distances between reflections in that picture, expressed by the wavelength of the X-ray radiation l and the angle of incidence θ, to the spacing d between the planes of the atom lattices in the DNA molecule. This knowledge of the exact spacing d identified the atomic bondings and the precise double helical structure (with a diameter of 2 nanometers and repetition every 3.4 nanometers).

In contrast, Haavelmo's 1944 approach of identifing models via their noise characteristics, is misdirected after Kalmans (1989) show first two Theorems: that a particular linear model A' can be fitted to the data in a consistent fashion with infinite many different noise patterns or residuals. Thus noise is undefinable, as it should be. Second, the generalized LS scheme can always produce the residuals required to let any linear model A' of any rank q fit the data. A characteristic of linear LS is that the covariance matrix of the inexact components of the data, is the smallest in a certain sense of all the noise covariance matrices that can be computed for the linear scheme A' and the data. But surprisingly, this is not the defining characteristic of LS; the rank q of model A'is.

The (truly) generalized LS scheme remains prejudiced, but its prejudice of the rank q can be neutralized by computing the complete regression models, as Frisch proposed almost 60 years ago [Frisch, 1934], and Kalman proved [1990]. As I demonstrated, all columns of the data covariance matrix together reveal the true rank q of the model.

It can not be emphasized enough that the givens in scientific research are the data. Data covariances can always be computed. Our method identifies the allowable narrow value ranges of the coefficients of A', up to an equivalency relationship, from the inexact data and separates the exact model from the noise. It is even capable of distinguishing outliers and other extraneous aberrations. There are no "true" or "non - true" data. Just data. The degree of inexactness depends on how close the data cluster around the exact linear model. (Note that we use linearity in the coefficients only. One - to - one transformations of the data series to finetune the pictures, so to say, is always allowed).

With our paradigm results are true or false. The hypotheses that the empirical data of the three-variable example above, can be described by a single-equation model. Similarly, I have demonstrated that I can falsify theoretical hypotheses about positivity or negativity of relationships.

Nowhere do I contend that "if enough data were available, economists could precisely predict our future." Our paradigm is about identification from the data, not prediction. Prediction implies extrapolation beyond the available data based on the belief in the continuing integrity of the identified model. Such a

prediction will never be exact. Prediction is never science, i.e. hard knowledge, since it involves a meta-physical assumption. To enhance the credibility of this assumption, however, one can apply "windowing," i.e. dividing the set of historical data into subsets and establishing that the exactly identified model of one subset is the same in all subsets.

REFERENCES

L. Accardi, E. Beltrametti, N. Cartwright, P.J. Lahti, N. Namiki, B. van Fraassen, A. Van der Merwe [1992], The interpretation of Quantum Theory: where do we stand ?, *Symposium of the Italian Academy for Advanced Studies in America at Columbia University*, New York, 1 - 4 April.

R. Frisch [1934], Statistical confluence analysis by means of complete regression systems. Publication No. 5, University of Oslo Economic Institute, Oslo, Norway.

R.E. Kalman [1990], A theory for the identification of linear relations, *Proceedings Colloque* LIONS (Edited by H. Brezis, P.G. Ciarlet), Paris.

C. A. Los [1989a], The prejudices of least squares, principal components and common factor schemes, *Computers and Mathematics With Applications*, 17 (8 - 9), 1269 - 1283.

C.A. Los [1989b], Identification of a linear system from inexact data: a three - variable example, *Computers and Mathematics With Applications*, 17 (8 - 9), 1285 - 1304.

C.A. Los [1992], Reply to E.T. Jaynes' and A. Zellner's Comments on my two articles, *Computers and Mathematics With Applications*, 20, (galleyproved, to be published in May).