Computers Math. Applic. Vol. 24, No. 8/9, pp. 277–288, 1992
Printed in Great Britain.

0097-4943/92 $5.00 + 0.00

REPLY TO E.T. JAYNES' AND A. ZELLNER'S COMMENTS ON MY TWO ARTICLES

CORNELIS A. LOS
NMB Bank, 135 East 57th St., New York, NY 10022, U.S.A.

PREFACE

The Bayesian physicist E.T. Jaynes has written a very detailed and intelligent commentary. He clearly understands the scientific essence of my two articles. The essence of my two articles is that a system of linear relationships has to be identified *from the data alone*, that the conventional (statistical) identification schemes introduce *ad hoc* prejudices, and that there currently exists serious research of prejudice-free alternatives. The Bayesian econometrician A. Zellner provides some very brief remarks.

A detailed reply is required since, unfortunately, both commentaries include statements that factually misrepresent, and therefore confuse, the contents of my two articles and of my collaborative research with Kalman, e.g., I never propose a "Los Solution", as Jaynes claims. The particular scheme he refers to is the Frisch scheme, which he calls a "radical new approach." Radical maybe, but new? Frisch proposed it already in 1934 [1]. Moreover, Jaynes' "Simple Solution" is the Principal Components scheme, thoroughly criticized in the first of my two articles. Both schemes I showed to be prejudiced.

In this reply, I will: (i) correct similar erroneous representations; (ii) clarify issues that necessarily remained unresolved at the time I wrote the articles; (iii) add some commentary on the nonscientific use of Kolmogorov's probability theory; and, most importantly, (iv) present the application of Kalman's new (1991) objective method of linear identification to my original data for comparison.

Despite the considerable difference in viewpoints between the two sides in this debate, two major positive conclusions emerge: (1) both sides agree that it is essential to identify the number of linear relations from the data. Jaynes and I both conclude that my data are modelled by *two* independent linear relationships—and *not one* relationship, as the original researchers at the Federal Reserve Bank of New York presumed—plus some relatively small noise. Moreover, (2) even Jaynes shows in his commentary that probability theory plays no role whatsoever to reach this conclusion. The data covariance matrix and linear algebra are all that is needed. The presumption of stochasticity of the data is superfluous. This must be a clear warning for statisticians, who indiscriminatingly use probability theory.

Algebraic data analysis in general and linear identification in particular involve both "deduction" and "induction". The problem is that the exact isomorphism between exact data and the minimal model (= UNIQUENESS PRINCIPLE) breaks down as soon as uncertainty is introduced. We investigate why and how this happens. Statisticians have provided a pseudo-solution to this problem by adding suitable prejudices, like "all data are samples from populations" to create uniqueness of the resulting models, i.e., to create certainty out of thin air.

According to the UNCERTAINTY PRINCIPLE [2, Lecture 1, p. 2]: "Uncertainty (noise) in the data is inherited as uncertainty in the model for the data." Engineers know that uncertainty, or noise, can persist even for infinite samples. This should not surprise us. Noise is not necessarily stochastic or probabilistic, but can be deterministic, as the theory of deterministic chaos (based

I thank the Guest Editor for providing for this debate a *forum*, for his efforts to eliminate noise and for insisting on *decorum*.

on nonlinear equations) teaches us. Noise, uncertainty, is what is unknown and unexplainable within the adopted (linear) framework. To assume certainty, even if that is the certainty of a probability distribution, where there is not, may be acceptable in the occult, but is not in science.

For the reader to understand this debate, it is essential to know that throughout their commentaries Jaynes and Zellner use the term "unknown errors", while we use the term noise. In information and system theory the term noise is well defined: it is that part of the data that is not the exact signal. Furthermore, Jaynes uses the term "solution", where I prefer the term scheme. The term scheme indicates that prejudices are involved. Newton's technical concept of *prejudice* is defined as an *a priori* assumption, not obtained from the data, or tested against the data. This scientific concept is very helpful in dissecting some embarrassing problems in statistics.

REBUTTALS

1. First things first: the data. The data in Appendix A [3, pp. 1301–1302] are not time series data, as Zellner assumes, but cross-sectional data, as I explicitly state: "These data are from 32 large bank holding companies followed regularly by Salomon Brothers, Inc., in their published statistical report for the year 1985." [3, p. 1301]. Not every T stands for time: T = 32 large banking holding companies.

These *data* were clearly *given* to me: "The data were kindly provided by the Banking Studies and Analysis Functions of the Federal Reserve Bank of New York" [3, p. 1301]. I was not concerned how they were collected or chosen, because that was irrelevant for my purpose. I am concerned about the prejudices inherent in the currently used linear identification schemes to analyze this and similar empirical data sets. Article [4] contains the precise mathematical formulation of these concerns. Article [3] illustrates, and compares these concerns, on the basis of these simple data.

2. In [3], I never intend to prove that the 32 banking holding companies behave today or tomorrow as in 1985. Zellner's inference that "Los would have us believe" so, is false. I only prove that this cross-sectional data set of 1985 is completely described by two independent linear relationships, plus some relatively small noise. Not such a startling conclusion, maybe, *but it is derived from the data alone.* That should be startling. Only if we collect a similar three-variable data set from these same 32 banks in another year and we prove that this new data set is also described by the same two linear relationships, we would make a small, but progressive step on the long and arduous road of *scientific* investigation. Because then the profitability behavior of these banks would show integrity.

3. Contrary to Jaynes' assertion, nowhere in my articles I recommend that all conventional schemes for estimating linear relations from noise-contaminated data be scrapped. I state that all conventional schemes are severely prejudiced and I mathematically show how in [4]. These conventional schemes do not produce objective, scientific results. They produce biased, prejudiced results. Our task is to find out exactly what their inherent prejudices are and how we can eliminate or neutralize these prejudices.

4. In both papers, which were written in 1986 and published in 1989, I state clearly that at that time the noisy identification problem was still not objectively solved [3, p. 1285; and 4, p. 1282]. However, I am happy to report that *now there does exist an exact, objective mathematical method for identification of a linear system from noisy data.* This method uses the Least Squares scheme as its computational tool and uses the data covariance matrix as its only input. (*All* existing linear identification schemes use the data covariance matrix as an input). The Least Squares scheme is prejudiced and the effect of its prejudices has to be neutralized. Therefore, the required scientific objectivity is only obtained by applying the Least Squares scheme in a *comprehensive* fashion. This new identification *method* could not be described in my two papers, since Kalman only formulated it in the fall of 1989 and presented it in [5]. That is scientific progress.

5. Jaynes claims that "methods for dealing with the problem which are optimal in all cases have been known for 20 years." But the "solution" presented in [6], for which Jaynes is a proponent, is clearly not "the correct general solution". It is not, since it keeps the prejudices, inherent in the conventional identification schemes, implicit and does not neutralize them. In addition, the

Table 1. Objective identification of q.

Least Squares Comprehensively Applied

$$\Sigma = \begin{bmatrix} 0.7022 & 6.9040 & -10.6826 \\ 6.9040 & 99.0556 & -114.7687 \\ -10.6826 & -114.7687 & 259.2516 \end{bmatrix} \qquad \Sigma^{-1} = \begin{bmatrix} 6.0992 & -0.2749 & 0.1296 \\ -0.2749 & 0.0331 & 0.0033 \\ 0.1296 & 0.0033 & 0.0107 \end{bmatrix}$$

The 3 LS (3,1) normalized system coefficients: 1, 2/			
	a_{12}	a_{13}	where $A' =$
LS1:	-0.0451	0.0213	$[1 \quad a_{12} \quad a_{13}]$
LS2:	-0.1205	-0.0121	
LS3:	0.0257	0.0823	
(1) No sign consistency when $q = 1$			
The 3 LS(3,2) normalized system coefficients:			
	a_{13}	a_{23}	where $A' =$
LS1:	0.0412	0.4427	$\begin{bmatrix} 1 & 0 & a_{13} \\ 0 & 1 & a_{23} \end{bmatrix}$
LS2:	0.0657	0.6463	
LS3:	0.0602	0.8631	
(2) Sign consistency when $q = 2$			

Notes:
(1) LS(n, q) = Least Squares scheme applied to n variables with q independent linear relationships
(2) LSi = the computational result from the i^{th} projection

theory of the British clergyman T. Bayes accepts stochastic subjectivity as its *raison d'etre*, a flagrant violation of the tenets of exact scientific objectivity.

6. To clarify the issues, let me once more state the Problem: if we have T observations on an n component vector, thus a $(T \times n)$ matrix X of data values, the objective is not to see whether there is evidence for q linear relations between the n data values of the form $XB = 0$, as Jaynes restates it, but whether there is evidence in the data X for q linear relations among the n exact signals, of the form $\hat{X}A = 0$, where $\hat{X}$ is the exact part of the data X and A an $(n \times q)$ matrix of unknown rank q, with $0 < q < n$. (q cannot be equal to 0 or to n). The objective is to determine q from the data alone, and in the process to separate $\hat{X}$ from X, where $X = \hat{X} + \tilde{X}$, with $\hat{X}$ orthogonal to $\tilde{X}$. (*All* existing linear identification schemes accept the additivity and orthogonality of the signal and noise components of the data.) $\hat{X}$ is the (multi-channel) signal, while $\tilde{X}$ is the noise component of the data.

7. I challenge Jaynes and Zellner to present to the *scientific* community a published example where, "in the conventional approaches", a researcher has taken a particular value of q only as a *"tentative working hypothesis"* (Jaynes' emphasis). After extensive research, I only find that even seasoned researchers choose subjectively a particular q, and keep that value of q as a "maintained hypothesis", when they attempt to determine the value of the coefficients of A. They never go back to check their chosen q against the data.

Kalman has now a Linear Identification Theorem [5, Theorem 9.6] which has the following operational interpretation. If you choose q *a priori* and $q(chosen) < q(true)$, then the computed values of the coefficients of A do not cluster closely around a point but lie on a hyperplane, frequently in different orthants. (Surprisingly, this is something that can be checked easily by plotting the computed values). The values of the coefficients of A are likely to be sign-inconsistent, so there is not even the possibility of a "best fit" [4, pp. 1274–1275, Example 4]. Therefore, I can state unconditionally that "significance testing" will not accomplish the checking of q. If $q(chosen) > q(true)$, then the computed coefficients of A do also not cluster closely around a point, but this time they are truly randomly "scattered" in a totally unpredictable fashion, throughout the hyperspace. Only if $q(chosen) = q(true)$, the computed coefficients of A cluster closely around a point and are thus sign-consistent.

These results only hold true when the noise is relatively low. No method can extract reliable information from noisy data when the noise levels are too high. However, since the new method places absolute boundaries on the computed coefficients of A, based on the data, it provides a clear indication of the accuracy of the results.

8. Let me now present the application of Kalman's new linear identification method to the data in [3, p. 1302]. The simplicity of the method is rather surprising: it relies only on the row vectors of the adjoint of the data covariance matrix Σ and on combinatorics. But that is its strength, since modern computers are extremely good in processing such combinatorial applications, even when the number of variables n is very large. Table 1 summarizes all the relevant results for the objective determination of q from the data. The inescapable conclusion is that $q = 2$, since only then all possible Least Squares results are sign-consistent. (A plot can be found in [3, p. 1291].)

Note that in Tables 1, 2 and 3 the A matrices are normalized for the purpose of comparison only. This normalization is arbitrary. All possible normalizations might be tried to neutralize the prejudice of a particular choice of normalization. Such a procedure produces another combinatorial computation. The particular normalization in these tables consists of the inversion of a submatrix of the adjoint of Σ, which may introduce numerical computer errors because of fixed memory bit lengths. There are ways to avoid these numerical errors, for example, by using exact symbolic algebra.

In Table 2, I compare the computational results of this identification method with the best results from the two major prejudiced schemes discussed in [3]. It is striking how close the computed coefficients of A cluster together, for all applications, when $q = 2$. The conclusion is that the noise in the data is not overwhelming the signals and that the $q(\textit{chosen}) = q(\textit{true}) = 2$.

Notice that the computational results of Jaynes' "Simple Solution", which is the Principal Components scheme, lie between those boundaries, after Jaynes has somehow convinced himself that $q = 2$. In contrast, the new method *computes* the value of q strictly from only the data. For all cases I provide in the Tables a summarizing indication of the signal/noise ratios by computing $tr\hat{\Sigma}/tr\tilde{\Sigma}$ and the related percentages of the total data variance explained by the $q = 2$ system.

In [3], in response to earlier commentary by Zellner in Tempe, Arizona, in March, 1988, I said I hope that the contents of Tables 1 and 2 would be convincing and would function as standards of comparison for any further discussion of these issues. Since Jaynes and Zellner are still not convinced, I challenge them to provide me with a set of static data that, in their opinion, would be more convincing. In the true spirit of scientific competition, I ask them to analyze these data applying Bayesian Inference, of course within the limits of Linear Identification, i.e., only applying "linear modelling". I will analyze the same data. Let us then compare the results. I request a set of static data (say, cross-sectional), since the problems of objectively analyzing dynamic data are of a much higher order than discussed here and are still not solved by anyone in the scientific community despite appearances.

9. The data covariance matrix Σ contains all the information needed for determining q and A. Of course, it is trivially true that for the reconstruction of the original X, the means of the data need to be added back in. Using the data covariance matrix as input we obviously operate on the deviations from the means, *in all existing linear identification schemes.*

The number of data points T is relevant only for questions of the homogeneity of the data. If I have three million data points, as Jaynes suggests, I would like to do windowing with the new scheme, to establish the homogeneity of the data and, *ergo*, the integrity of the linear system. It is not necessarily true, as Jaynes implies, that more data points lead somehow to greater accuracy of the computational results. To apply such a "Law of Large Numbers" often uncheckable assumptions (prejudices?) are conventionally made, such as: "the data are homogeneous" and "the datapoints are m-independent." Observing raw data, I am not prepared to make such facilitating, and often fallacious, assumptions. I examine the issue by applying (overlapping, non-overlapping, etc.) windows, i.e., by creating subsets of data and comparing the results of all subsets with each other. In other words, not I but statisticians impose prejudicial restrictions on the data, by supposing to know their characteristics *a priori*.

10. Jaynes' misrepresentation is most obvious when he computes the spectral decomposition of Σ and then states: "Los demands that we choose our linear relations solely from [this spectral

Table 2. Comparison of computational results for $q = 2$.

I. Comprehensive Least Squares (Objective)				
LS1: $A' = \begin{bmatrix} 1 & 0 & 0.0412 \\ 0 & 1 & 0.4427 \end{bmatrix}$	$\hat{\Sigma} = \begin{bmatrix} 0.4402 & 4.7291 & -10.6826 \\ 4.7291 & 50.8072 & -114.7686 \\ -10.6826 & -114.7686 & 259.2516 \end{bmatrix}$	$\tilde{\Sigma} = \begin{bmatrix} 0.2620 & 2.1749 & 0 \\ 2.1749 & 48.2484 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	Signal/Noise Ratio 1/ 6.40	Data Explained (%) 2/ 86.49
LS2: $A' = \begin{bmatrix} 1 & 0 & 0.0657 \\ 0 & 1 & 0.6463 \end{bmatrix}$	$\hat{\Sigma} = \begin{bmatrix} 0.7022 & 6.9040 & -10.6826 \\ 6.9040 & 67.8820 & -105.0347 \\ -10.6826 & -105.0347 & 162.5216 \end{bmatrix}$	$\tilde{\Sigma} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 31.1736 & -9.7340 \\ 0 & -9.7340 & 96.7300 \end{bmatrix}$	Signal/Noise Ratio 1.81	Data Explained(%) 64.37
LS3: $A' = \begin{bmatrix} 1 & 0 & 0.0602 \\ 0 & 1 & 0.8631 \end{bmatrix}$	$\hat{\Sigma} = \begin{bmatrix} 0.4812 & 6.9040 & -7.9992 \\ 6.9040 & 99.0556 & -114.7686 \\ -7.9992 & -114.7686 & 132.9743 \end{bmatrix}$	$\tilde{\Sigma} = \begin{bmatrix} 0.2210 & 0 & -2.6835 \\ 0 & 0 & 0 \\ -2.6835 & 0 & 126.2773 \end{bmatrix}$	Signal/Noise Ratio 1.84	Data Explained(%) 64.76
II. Frisch (Prejudiced) 3/				
$A' = \begin{bmatrix} 1 & 0 & 0.0602 \\ 0 & 1 & 0.6463 \end{bmatrix}$	$\hat{\Sigma} = \begin{bmatrix} 0.6426 & 6.9040 & -10.6846 \\ 6.9040 & 74.1729 & -114.7686 \\ -10.6826 & -114.7686 & 177.5831 \end{bmatrix}$	$\tilde{\Sigma} = \begin{bmatrix} 0.0590 & 0 & 0 \\ 0 & 24.8828 & 0 \\ 0 & 0 & 81.6684 \end{bmatrix}$	Signal/Noise Ratio 2.37	Data Explained(%) 70.30
III. Principal Components or Statistical Common Factor Scheme With Covariance Matrix As Input (Prejudiced) 4, 5/				
$A' = \begin{bmatrix} 1 & 0 & 0.0448 \\ 0 & 1 & 0.5218 \end{bmatrix}$	$\hat{\Sigma} = \begin{bmatrix} 0.5033 & 5.8621 & -11.2352 \\ 5.8621 & 68.2835 & -130.8710 \\ -11.2352 & -130.8710 & 250.8252 \end{bmatrix}$	$\tilde{\Sigma} = \begin{bmatrix} 0.1989 & 1.0419 & 0.5525 \\ 1.0419 & 30.7721 & 16.1024 \\ 0.5525 & 16.1024 & 8.4264 \end{bmatrix}$	Signal/Noise Ratio 8.11 (max possible)	Data Explained(%) 89.03 (max possible)

Notes:

(1) Signal/Noise Ratio = trace($\hat{\Sigma}$)/trace($\tilde{\Sigma}$). The signal/noise criterion works well only if you know q.

(2) Data Explained = 100*Signal/Data = trace($\hat{\Sigma}$)/trace($\tilde{\Sigma}$) = 100/[1 + 1/(Signal/Noise)]

(3) Frisch (3, 2) has a unique solution since Wilson's $d(3, 2) = 0$

(4) Uniqueness is achieved because of arbitrary retention of largest eigenvalue

(5) This is Jaynes' Simple Solution, as can be verified by applying the formulas of Theorem 2 in [4, p. 1277]

decomposition] (2) without making use of what we may know about the measurement errors in the three directions, or other relevant evidence." (If we really have such information there is more data than Los assumes.) Los [4, pp. 1276–1277] states that such a spectral decomposition (on which the Principal Components scheme is based) is trivial and does nowhere make such a demand. However, the conventional normalization, used by Jaynes, is a prejudice, rendering the decomposition unique.

Confusion may have arisen, because the Frisch scheme looks similar to the spectral decomposition of the data covariance matrix. But it clearly isn't, because the eigenvalues (and eigenvectors) of the spectral decomposition contain signal plus noise, since they are both based on Σ. The diagonal elements of the Frisch noise matrix $\tilde{\Sigma}$ are variances of the Frisch noise only. The spectral decomposition works with the data (= signal + noise), but does not by itself decompose the data in signal and noise. The prejudices of the Principal Components or of the Common Factor Scheme are necessary to accomplish this, by prejudging how many (= $n - q$) eigenvalues to retain [4, pp. 1276–1278). Or as Jaynes prefers it: how many eigenvectors to retain.

The results of Table 2 also show that knowledge about the noise is irrelevant for system identification. Notice that identified system coefficients in matrices A are identical for LS2, LS3 and for Frisch, but the corresponding noise matrices $\tilde{\Sigma}$ differ.

11. It is revealing that Jaynes uses phrases like: "If we *want* to find two relations" (i.e., $q = 2$) and "... if one *believes* that these data give good evidence for the presence of two linear relations" (my emphasis). Science has no place for "wants" or "beliefs": science is about data analysis, about identifying the system, about finding out from the data if there are one, or more linear relations among the exact components of the data series.

The Frisch scheme does not *a priori* determine the number of relations q. But the Frisch scheme is prejudiced, as I noted: "But the Frisch scheme apparently has its own prejudice" [3, p. 1296]. Why? Precisely because of the presupposed diagonality of its noise matrix, i.e., the presupposed orthogonality of the noise vectors. Not only did I clearly note that the Frisch scheme is prejudiced, but I made it clear that it illustrates some of the subtle issues involved in such prejudices. For example, the results of the Frisch scheme do not exist in general, but only in a limited set of cases, because of Wilson's inequality [3, pp. 1295–1296, and 4, p. 1280], which makes this peculiar scheme non-trivial.

It is remarkable that in all statistics textbooks, the sign of this inequality is incorrectly reversed, virtually from the very moment after it was published by Wilson in 1929 [7]. It was Wilson's inequality that finally convinced Kalman and me that the Frisch scheme was prejudiced [3, p. 1296, and 4, pp. 1280–1281]. Worse, the Frisch prejudice cannot be neutralized, unlike the Least Squares prejudices.

12. I wonder: what can one know about the "measurement errors" of these cross-sectional banking data when the only data available is contained in X? If one does know more (from these 32 large banking holding companies ?), then one has already extended the data set X by some extraneous information. I do not have this extraneous information, nor did, apparently, the Banking Studies and Analysis Function of the Federal Reserve Bank of New York.

Jaynes asserts: "There can be no defensible solution until the data are supplemented by additional information about the nature of the problem." Obtaining additional data is always a good idea, if possible, but often we cannot obtain it. The *data* are the only *given*. "Prior information" is additional data, and then only if it is obtained by objective measurement, not by subjective assertion. It is not in the interest of science to enlarge our data sets by beliefs (in particular when strongly held on the basis of "authority").

13. Jaynes asserts that the spectral decomposition tells him that $q = 2$. But he leaves his audience in the dark how he arrives at this judgement, and why he retains, subjectively, the largest eigenvalue. He calls his solution the "Simple Solution", which he compares with the Frisch Solution. There are many examples in major computer software manuals that show that, on the basis of a plot of the eigenvalues, a researcher can often not discriminate between what is signal and what is noise [4, p. 1280].

Since the Frisch solution is the exact equivalent of the factor analysis scheme with the data covariance matrix as input, it does not surprise me that the Frisch solution is only a small

rotation away from Jaynes' "Simple Solution" or spectral decomposition (which is equivalent to the Principal Components scheme). The rotation by 5.35 degrees achieves diagonality of the noise matrix, i.e., the prejudice of the Frisch scheme.

14. Jaynes discusses extensively the impact of a change in units on his "Simple Solution". It is educational to analyze in detail what he says. The relevant computational results are all in Table 3. First, the change in units of measurement changes the eigenvalues and eigenvectors. That is correct, since they contain both signal and noise components. The impact of a 100-fold increase in the first data series x_1, is dispersed over all three eigenvalues and their corresponding eigenvectors. (That is exactly why *any* choice of q in this scheme remains a prejudice, no matter how arrived at.) One cannot determine where the impact will be the greatest. Magnitudes and directions change: it is not clear *a priori* where the noise resides and how large it is. However, I will show that it does not matter much for any linear identification scheme, not even for the "Simple Solution".

It is easy to see this mathematically: premultiply X by a positive definite matrix F (in case of Jaynes' change in units of measurement this would be the diagonal matrix $F = \mathrm{diag}(1/100, 1, 1)$, then

$$\begin{aligned} A'F^{-1}F\Sigma F' &= A'F^{-1}F\hat{\Sigma}F' + A'F^{-1}F\tilde{\Sigma}F' = A'\hat{\Sigma}F' + A'\tilde{\Sigma}F' \\ &= A'\Sigma F' = A'\tilde{\Sigma}F'. \end{aligned}$$

The system equation has not changed,

$$A'\hat{\Sigma} = A'F^{-1}F\hat{\Sigma}F' = 0,$$

nor the rank q of A. Notice that the transformed system matrix is $A'F^{-1}$. For Jaynes' example, the a_{11} coefficient is now 100 times smaller, but since we normalized $a_{11} = 1$, this is equivalent to a_{13} to be 100 times larger in this relationship.

I do not understand why Jaynes states that the new "Simple Solution" is nearly orthogonal to the original one. As observed in Table 3, the original and new U matrices of eigenvector are not orthogonal to each other, and why should they? (Note that the original U matrix in Jaynes' Example is inconsistent *qua* vector signs, undoubtedly a typing error. I computed the correct original U in Table 3.) What I do see is that the new (normalized) system matrix "$A'F^{-1}$", which represents the linear dependencies, has a coefficient a_{13} in the first row that is a factor of about 100 larger than the earlier A' of the "Simple Solution", as one would expect. The first row and column of Σ are both increased by a factor of 100, and thus one expects that those of $\hat{\Sigma}$ and of $\tilde{\Sigma}$ would increase by about the same factor. (Notice how the signal/noise ratios are affected by this transformation.)

15. Jaynes himself admits that the change in measurement units does not matter for the Frisch scheme, when he states: "However, we must concede a surprising point: Los does manage to achieve invariance under a change of units, in spite of the fact that he takes no note of the need for it". Where is the need? My mathematical explanation in the preceding Rebuttal clearly shows that there exists no such need for the Frisch scheme, nor for any other linear identification scheme that does not refract the noise arbitrarily. Premultiplication of the data X by a positive definite matrix has no substantial effect on the identification results. It does not change the q, and it does change the original A matrix as expected: $A'F^{-1}$. The problem is not "ill-posed" or "mathematically undetermined", as Jaynes claims. The identification problem remains the same: determine the unique minimal number of relations q from the data Σ.

Jaynes' concession also shows that the Frisch scheme is rational (as is the Least Squares scheme), while his "Simple Solution" is not. The change in unit of measurement has no impact on the computational results of the Frisch scheme or of all applications of the Least Squares scheme, but the results of his "Simple Solution" are affected. It is precisely for this reason that Kalman and I use the Least Squares scheme as computational tool in our new identification method. The Least Squares scheme does not suffer from the Frisch prejudice nor from the prejudices of the "Simple Solution", since it operates *linearly* on the submatrices of the transposed

Table 3. Comparison of Jaynes' Original Simple Solution and this solution after a change in measurement units.

I. Jaynes' Original Simple Solution 1/

$$A' = \begin{bmatrix} 1 & 0 & 0.0448 \\ 0 & 1 & 0.5218 \end{bmatrix} \qquad \hat{\Sigma} = \begin{bmatrix} 0.5033 & 5.8621 & -11.2352 \\ 5.8621 & 68.2835 & -130.8710 \\ -11.2352 & -130.8710 & 250.8252 \end{bmatrix} \qquad \tilde{\Sigma} = \begin{bmatrix} 0.1989 & 1.0419 & 0.5525 \\ 1.0419 & 30.7721 & 16.1024 \\ 0.5525 & 16.1024 & 8.4264 \end{bmatrix}$$

Signal/Noise Ratio 1/	Data Explained (%)
8.11	89.03

$$\textit{Matrix of Eigenvalues} = \begin{bmatrix} 0.1636 & 0 & 0 \\ 0 & 39.2338 & 0 \\ 0 & 0 & 319.6120 \end{bmatrix} \qquad \textit{Matrix of Eigenvectors } U_1 = \begin{bmatrix} -0.9988 & -0.0302 & -0.0397 \\ 0.0451 & -0.8856 & -0.4622 \\ -0.0212 & -0.4634 & 0.8859 \end{bmatrix}$$

II. Jaynes' Simple Solution With Change in Unit of Measurement of $x_1$3/

$$\text{“}A'F^{-1}\text{”} = \begin{bmatrix} 1 & 0 & 6.4798 \\ 0 & 1 & 0.6412 \end{bmatrix} \qquad \text{“}F\hat{\Sigma}F'\text{”} = \begin{bmatrix} 7019.0241 & 694.5690 & -1083.2244 \\ 694.5690 & 68.7299 & -107.1898 \\ -1083.2244 & -107.1898 & 167.1630 \end{bmatrix} \qquad \text{“}F\tilde{\Sigma}F'\text{”} = \begin{bmatrix} 2.7217 & -4.1705 & 14.9617 \\ -4.1705 & 30.3257 & -7.5789 \\ 14.9617 & -7.5789 & 92.0886 \end{bmatrix}$$

Signal/Noise Ratio	*Data Explained* (%)
57.98	98.30

$$\textit{Matrix of Eigenvalues} = \begin{bmatrix} 29.5890 & 0 & 0 \\ 0 & 95.5470 & \\ 0 & & 7255.1700 \end{bmatrix} \qquad \textit{Matrix of Eigenvectors } U_2 = \begin{bmatrix} -0.0764 & 0.1633 & -0.9836 \\ 0.9875 & 0.1242 & -0.0973 \\ 0.1380 & -0.9787 & 0.1518 \end{bmatrix}$$

Or, After Linearly Transforming Back to the Original Unit of Measurement for x_1

$$A' = \begin{bmatrix} 1 & 0 & 0.0650 \\ 0 & 1 & 0.6412 \end{bmatrix} \qquad \hat{\Sigma} = \begin{bmatrix} 0.7019 & 6.9457 & -10.8322 \\ 6.9457 & 68.7299 & -107.1898 \\ -10.8322 & -107.1898 & 167.1630 \end{bmatrix} \qquad \tilde{\Sigma} = \begin{bmatrix} 0.0003 & -0.0417 & 0.1496 \\ -0.0417 & 30.3257 & -7.5789 \\ 0.1496 & -7.5789 & 92.0886 \end{bmatrix}$$

Signal/Noise Ratio	*Data Explained* (%)
1.83	65.90

CONCLUSION 1: Jaynes' Simple Solution (= Principal Components Scheme) is variant with respect to changes in the unit of measurement.

$$U_1'U_2 = \begin{bmatrix} 0.1180 & 0.1895 & 0.9748 \\ -0.9362 & 0.3485 & 0.0456 \\ -0.3311 & -0.9179 & 0.2185 \end{bmatrix} \qquad U_2'U_1 = \begin{bmatrix} 0.1203 & 0.5958 & -0.7940 \\ -0.9861 & -0.0204 & -0.1647 \\ -0.1143 & 0.8028 & 0.5851 \end{bmatrix}$$

CONCLUSION 2: The two matrices of eigenvectors are not orthogonal

Notes:

(1) Signal/Noise Ratio = trace $(\hat{\Sigma})$/ trace $(\tilde{\Sigma})$

(2) Data Explained = 100* Signal/Data = trace $(\hat{\Sigma})$/ trace $(\Sigma) = 100/[1 + 1/(\text{Signal/Noise})]$

(3) Data series x_1 is multiplied by 100. Noise difraction leads to slightly inexact results

data covariance matrix. To paraphrase Jaynes, any rational method of identification from noisy data ought to lead to the same results, if the noise is relatively small, irrespective of what we believe the noise to be. His "Simple Solution" does not.

This mathematical analysis demonstrates also that Jaynes' statement: "Common sense might tell us that it is not the absolute size of the data covariance, but the data covariance *in comparison with the sampling variances* that is relevant for inference" is devoid of any meaning.

16. I fail to understand how "Conventional methods of inference achieve this necessary invariance by using sampling distributions," as Jaynes claims. The issue is not sampling, as we proved! It is significant, though, that Jaynes invokes the *deus ex machina* of "sampling distributions", since that raises other questions. How does he know that (1) the data are sampled? Or, that (2) the data are stochastic? We only know that the data are inexact = uncertain = can-not-be-directly-described-by-a-linear-relationship. And, (3) even if the data were stochastic, why would they adhere to an *a priori* assumed distribution? In the case of the banking data, I even reject the assertion that the data are a sample, *since the data represent the available universe.* No sampling is involved.

But, in analogy to the exact situation, I know how such a change in measurement unit would be reflected in my computational results, if the noise is relatively small. Contrary to Jaynes' assertion, there is no arbitrariness involved. There is only uncertainty in the data, that gets reflected in the computed a_{13}, in any linear identification scheme.

17. Correlations, being dimensionless, are indeed independent of the system of units. However, since the data correlations are ratios of data covariances and data variances, which both contain noise, I fail to see any improvement in using ratios that contain noise divided by noise, as inputs for identification. The scaling is done by elements that contain noise and thus distorts the information available in the data covariance matrix [4, p. 1278].

The scaling by variances, *c.q.* standard deviations, is different from the change in units of measurement, where the transformation is accomplished by noise-free constants that do not distort the information in the data for system identification. Most certainly, the use of correlations will not assist in "correcting the arbitrary distortions in covariance functions" as Jaynes believes. Since the *data covariances* are the *given*, I do not understand how one can even assert to know that there are "distortions". There is only one relevant form of invariance: and that is of q, the (minimal) number of independent linear relations among the exact data. The q remains invariant under multiplicative transformations, such as a change of measurement, as shown in Rebuttal 14.

18. Jaynes correctly states: "Even the criterion of smallness of the residuals leads to arbitrary results." Nature does not have a criterion for smallness of the residuals, only statisticians have. "Least Squares", for example, is a misnomer: it is easily proved, but, regrettably, not well known, that all least squares results can be directly read from the adjoint of the data covariance matrix Σ. Therefore, the Least Squares (projection) scheme can be characterized by minimization, but need not be defined in that way. Kalman proves, though, *that for a given matrix* A, the Least Squares scheme *implies* the "smallest" noise covariance matrix of all the linear identification schemes [2, Lecture 6]. This result is far more general than the conventional "best"-ness criterion of the econometric BLUEs.

19. In our formulation, the objective is to decompose the data matrix (and not the reduced data matrix!) into an "exact" or "explained" component $\hat{X}$, and a "noise" or "unexplained" component $\tilde{X}$. However, Jaynes blissfully calls $\hat{X}$ the "data vectors", although they are the vectors of the exact components of the data. Because Jaynes fails to see this crucial distinction (he writes, for example, $\hat{X} = (y|X_2)$, while he should have written $\hat{X} = (\hat{y}|\hat{X}_2)$). Because $A\hat{X} = 0$, we have $AX = A\tilde{X}$.

The condition (8) in Jaynes' Commentary is implied by $A'\hat{\Sigma} = 0$ and is, indeed, a tautology, or an Organizing Principle, as I called it [3, p. 1287]. It does not make reference to the data (It is based on $\hat{\Sigma}$, the exact part of the data covariance matrix, not the data covariance matrix Σ). The only interest in introducing (8) is to show that *all* linear identification schemes (conventional and unconventional) are structured by this Organizing Principle. A fact that no statistician before has discovered. This Organizing Principle defines a set of linear schemes that is much wider than the set of linear identification schemes considered in conventional statistics.

20. The "linearity" does impose restricting conditions on the universe of all possible identification schemes. But while linear algebra is well defined, nonlinear algebra (whatever that may be) is not. So I think the scientific community will forgive us for using linear identification schemes for the time being. In addition, one is, of course, allowed to non-linearly transform the data, if the transformation is one-to-one. Linear identification is about linearity in the coefficients. The objective is to reduce the noise in the transformed data set. After all, the noise incorporates the non-linearities. Once the linear system of the transformed data is established, one can go directly back to the original data by reversing the one-to-one transformation. Such one-to-one transformations of the raw data are what scientific observational instruments are all about. The creativity and technology of finding a useful one-to-one transformation of the data that reduces the noise, i.e., that leads to the unique determination of the system, is what drives scientific progress.

The great discovery of Kalman is that there are new linear identification schemes to be discovered, i.e., new scientific instruments for the observational sciences. He opened up a new world. In [3], I discussed the comparative application of such a scheme, that has not been researched by statisticians. The Frisch scheme was researched in some detail by the mathematician E. B. Wilson in the 1930s. It continues to live an obscure and very limited existence in econometrics as in a variant of the "errors-in-variables" model, and in psychometrics as the exact equivalent of the common factor scheme.

21. To call $q = n - 1$ a general solution, as Jaynes does, for almost all real data with large n, strikes me as highly arbitrary. It is the one common factor "solution" of Spearman [8], which is, by all standards, wrong. However, in our data case of $n = 3$, indeed $q = 2$. While the Frisch scheme fits equation (8), the Frisch scheme is not a projection operator, contrary to what Jaynes asserts, as can be checked easily using the algebraic formulas in [4].

One of Kalman's fundamental discoveries is that not all linear identification schemes (i.e., schemes that fit the system equation (5), respectively (8) in Jaynes' Commentary) are projection operators. The mathematical interest in the Frisch scheme derives from the fact that it not a projection operator, although it belongs to the set of linear identification schemes. Both the Least Squares and Principal Components schemes, e.g., Jaynes' "Simple Solution", are projection operators [4, pp. 1272 and 1277].

The diagonality of the Frisch noise matrix superimposed on the data is a prejudice *par excellence*, and "almost certainly false", as Jaynes correctly states. The Frisch prejudice, however, is remarkable by its explicitness. Something that cannot be said about all the other conventional statistical identification schemes, where the prejudices are often implicit and, worse, not acknowledged by researchers.

22. Jaynes correctly observes that to turn the linear identification equation (8) into a well-defined mathematical problem, $n(n+1)/2 - q(q+1)/2$ exact conditions must be imposed on the noise (*"on those measurement errors"*, as he calls it). He also correctly states: "However this is done, it requires to assume a great deal of information which we do not possess; how could one ever justify any such assumption?" However, *all* existing linear identification schemes do impose such conditions, one way or another, prejudicially.

23. As a general principle: if the noise is relatively small, all linear identification schemes will lead to the same consistent results. What if the noise levels are relatively large? Then no linear identification scheme can lead to conclusive results. Statisticians are so hung up on probability theory that they always expect to find point estimates. (This subjectivity in statistics is most clearly expressed in the use of terminology like "significance levels" and "degrees of confidence". Different fields of science appear to use different, but arbitrary, calibration levels for these notions, like 90%, 95% or 99% confidence). But since we cannot know if the data are stochastic, the best we can do is establishing absolute limits on the possible coefficient values of A. If the intervals between these boundaries are close together, then that is *proof* that the noise levels are relatively low and that we can be confident of our findings. For example, looking at Table 2, I am impressed to see how close the absolute boundaries on coefficient values of A lie together.

24. Jaynes states correctly, "Surely, the fundamental basis of scientific inference—almost a principle of morality—is that we should

(1) take into account all the relevant information we have, of whatever type,
(2) carefully avoid assuming information that we do not have."

I could not agree more, but a casual reading of the statistical literature provides evidence that statisticians violate both rules. For example, anybody who runs only one regression using n variables is clearly violating rule (1), since he/she ignores most of the information available in the $n \times n$ data covariance matrix. This has led to extensive literature on the discussion of an artificial problem—the "multicollinearity problem"—which is in fact not a problem. The collinearity of the data expresses in coded form the information about the system we try to identify.

Next, rule (2) is clearly violated when one states that all empirical data are samples from probability distributions, or even only that all empirical data are stochastic, since, in general, we don't know and cannot know that.

Contrary to Jaynes' assertion, I did not commit "egregious violations of both these rules" by discussing the prejudices of the available linear identification schemes, conventional or unconventional. *Au contraire*, Kalman and I pride ourselves in serving science by pointing out these prejudices.

25. Not unrelated to Jaynes' concern about the "morality" of scientific research, I find it disturbing that J. C. Bailar III plans to write ethics books on the Scientific Method [9, p. 5]. In particular since Bailar plans to base his books on the (false) tenets that

(i) "Inference is the heart of the scientific method." The substance of our debate shows it isn't;
(ii) "Statistical thinking is the heart of inference." The results of Kalman's and my articles make this, at least, doubtful, since simple algebra suffices; and
(iii) "Anything that damages the process of inference is unethical."

I must repeat a harsh truth that scientists have known since about four centuries: *Scientific claims are not ethical or unethical: science is true or false.*

26. I now address a series of erroneous assertions at the end of Jaynes' commentary, which I cite here between quotation marks and then comment on:

(i) "Least squares is a special case of maximum likelihood." Least squares is an algebraic scheme of identification that presumes the number of relations and attributes zero noise to a subset of the data variables, the so-called "regressors". Maximum likelihood introduces the additional (restrictive) notions of probability and even of (Gaussian) probability distribution within the framework of algebraic identification schemes. Within the framework of linear identification schemes, maximum likelihood can be demonstrated to be a very special case indeed. It produces conflicting results, as demonstrated by Solari [10], depending on the assumption if the signal is stochastic or not, and is impotent in identification, as demonstrated by Kalman [5;2, Lecture 7].
(ii) "One may have highly cogent prior information about the likely linear relations." "Stochastic information" is no information, but a subjective guess. I agree that if one has exact prior information, like the existence of identities between variables, one should use this information by eliminating as much as possible of the noise in the data set. But in the observational sciences, we seldom have such priceless information.
(iii) "We need some algorithm to update our estimated relations when new data become available", when one has a sequence of data sets. One does not "update relations", i.e., update systems. One updates forecasts. If one has a sequence of data sets, one should apply the windowing approach and analyze each data set as if it were a stand-alone set. If all sets produce the same results one has found something to publish in a scientific journal: the data sets are homogeneous and the system shows integrity. If a particular data set is isomorph to a different system, i.e., a different q, the conclusion is that the system has no integrity.

A Kalman filter can be interpreted as a sequential Least Squares estimator. Therefore one has to know the system, *i.e.* q for it. One knows q in most engineering situations, but

not in most observational sciences. In those sciences one has to determine the system, i.e., q first, before one can apply the Kalman filter to sequentially update the exact X. This is precisely the problem we have been discussing here. Kalman started this research into the identification of linear systems from empirical data ten years ago, because he became concerned about the abuse of his filter in the observational sciences.

(iv) "The real problem may have "nuisance parameters" of no interest to us." The problem of parametrization is trivial. Linear identification results in "reduced form" systems, that can be parametrized, i.e., "interpreted" in expert language, in an infinite number of ways [11, p. 280]; by changing, for example, the units of measurement. But q remains invariant under all these transformations. The concept of "nuisance parameters" is therefore just that: a nuisance that the modeler invents to annoy himself.

(v) "One needs also a way to judge the relative merits of different models, in the light of the data, so that cumulative improvements can take place over long time." For one particular data set there exist only one particular model with one particular q. That is what science is all about: the uniqueness of the model to be identified. The story of the discovery of the structure of the DNA molecule by J. D. Watson and F. H. C. Crick is an example. There was no cumulative improvement of models. There was a unique geometric constellation of atomic elements that fitted the observations. When Watson and Crick found the correct double-helix structure, it was immediately obvious to all involved in the research that it was true. No additional data were required.

There can be uncertainty about the actual coefficient values of a linear system, because of the noise in the data, but not about q, the minimum number of linear relationships minimally describing the data. "Improvement" only occurs, when new data show less noise and the same system. If additional data show different q's, there is heterogeneity of data: the systems describing the different data do not show integrity and no scientific progress is guaranteed.

References

1. R. Frisch, Statistical confluence analysis by means of complete regression systems, Publication No. 5, University of Oslo Economic Institute, Oslo, Norway, (1934).
2. R.E. Kalman, *Nine Lectures on Identification*, Springer-Verlag, New York, (1992) (to appear).
3. C.A. Los, Identification of a linear system from inexact data: a three-variable example, *Computers Math. Applic.* **17** (8-9), 1285–1304 (1989).
4. C.A. Los, The prejudices of least squares, principal components and common factor schemes, *Computers Math. Applic.* **17** (8-9), 1269–1283 (1989).
5. R.E. Kalman, A theory for the identification of linear relations, In *Frontiers in Pure and Applied Mathematics* (Edited by R. Dautray), pages 117–132, North-Holland, Amsterdam, (1991).
6. A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York, (1971).
7. E.B. Wilson, Book review of Crossroads in the mind of man: A study of differentiable mental abilities (Edited by T.L. Kelly), *J. Gen. Psychol.* **2**, 153–169 (1929).
8. C. Spearman, General intelligence, objectively determined and measured, *Am. J. Psychol.* **15**, 201–293 (1904).
9. John C. Bailar III becomes MacArthur fellow, Editorial in *Amstat News*, page 5 (November, 1990).
10. M.E. Solari, The "maximum likelihood solution" of the problem of estimating a linear functional relationship, *J. Royal Stat. Soc.*, Series B **31**, 372–375 (1969).
11. P.J. Dhrymes, *Introductory Econometrics*, Springer-Verlag, New York, (1978).