BAYESIAN STATISTICS 2, pp. 329-360
J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith (Eds.)

Highly Informative Priors

E. T. Jaynes
Washington University, St. Louis

SUMMARY

After discussing the role of prior information in statistical inference, historically and in current problems, we analyze the problem of seasonal adjustment in economics. Litterman (1980) has shown how informative priors for autoregressive coefficients can improve economic forecasts. We find that in seasonal adjustment informative priors can have a much greater effect on our conclusions. In our model, even the dimensionality of the joint posterior distribution of the irregulars depends on prior information about the seasonal component; and some functions of the irregulars can be determined more accurately than in sampling theory.

Keywords: BRUTE STACK; FOURIER SERIES; JEFFREYS PRIOR; LEAST SQUARES ESTIMATES; NECESSARY VIEW; NONINFORMATIVE PRIORS; PRIOR INFORMATION; SAMPLING THEORY METHODS; SEASONAL ADJUSTMENT; STATISTICAL MECHANICS; SUBJECTIVITY; WEIGHTED AVERAGE.

1. INTRODUCTION

The statistical problems envisaged in our pedagogy are almost always ones in which we acquire new data D that give evidence concerning some hypotheses $H, H', \ldots$ (this includes parameter estimation, since H might be the statement that a parameter lies in a certain interval); and we make inferences about them solely from the data. Indeed, Fisher's maxim, "Let the data speak for themselves" seems to imply that it would be wrong – a violation of "scientific objectivity" – to allow ourselves to be influenced by other considerations such as prior knowledge about H.

Yet the very act of choosing a model (*i.e.* a sampling distribution conditional on H) is a means of expressing some kind of prior knowledge about the existence and nature of H, and its observable effects.

This was noted by John Tukey (1978), who observed that sampling theory is in the curious position of holding it decent to use judgment in deciding which parameters should be present in a model; but then indecent to use judgment in estimating their values. He saw the Bayesian method as something which "allows one to do the indecent thing while modestly concealed behind a formal apparatus".

Here we do this indecent thing, and note how it changes the problem of seasonal adjustment, an unusual problem in that all the parameters are nuisance parameters. But it is just for this reason that our formal apparatus enables us take into account things beyond the technical means, and even the concepts, of sampling theory. Jimmie Savage advocated noninformative priors on the grounds that they didn't make much difference; we advocate informative priors on the grounds that they do make a very important difference. As Litterman (1980) found in a similar problem, taking into account cogent information not contained in the sampling distribution can improve the accuracy and reliability of our conclusions.

In the following two Sections we digress to comment on the strange history of the prior information issue, with its controversies still not entirely resolved, and to note some other new applications of informative priors. Our seasonal adjustment calculation starts in Section 4.

2. NECESSARIANS

There is a surprisingly wide range of philosophical views about the role of prior information. Since the 1930's a common view has been that it is just plain wrong to take prior information into account. As a student in the late 1940's, the writer was strongly indoctrinated with this view. Indeed, in most orthodox works the term "prior information" does not appear at all. Perhaps it was felt that since prior knowledge is hard to document, the user would be under the temptation to slip in prior opinions, in the guise of prior knowledge – a terrible sin in the view of some (van Dantzig, 1957).

But then there was a seemingly violent swing to the opposite extreme view of the so–called "subjective Bayesians". Jimmie Savage (1954) proclaimed it his intention to incorporate prior opinions – not prior knowledge – into scientific inference (or at least into the reasoning of an idealized being called "the person", thought of as a normative model for scientific inference).

He rejected formal principles (symmetry, maximum entropy) by which prior probabilities might express, and be determined by, prior knowledge; and accused those of us who advocated such principles of holding "necessary" views of probability and of claiming to get something for nothing (Savage, 1981, p. 731). Indeed, he took it as a fundamental tenet (Savage, 1954, p. 3) that two "persons" with the same prior knowledge might assign different prior probabilities without either being unreasonable.

These philosophical differences leave a practical scientist, concerned with problems of the real world, with the uncomfortable feeling of being caught in the middle. From his viewpoint, it then appeared that there were two active camps, holding opposite extreme positions equally unreasonable and inapplicable to his problems of inference.

An earlier discussion (Jaynes, 1968) stressed the need to find a safe middle ground between these extremes, which recognizes the relevance of cogent prior knowledge and the need to take it into account (in seeming disagreement with Fisher); but also recognizes the claims of logical consistency (in seeming disagreement with Savage). That is, we took it as our fundamental "Desideratum of Consistency" that in two problems where we have the same relevant prior information, we should assign the same prior probabilities; and showed that this desideratum is already sufficient to determine priors in some cases. Savage (1981, p. 736) proceeded to dismiss this as "an unusual necessarians position".

It appears to us, however, that this position and desideratum are neither unusual nor "necessarian" as Savage defined that term. Zellner (1982) has also noted the contrast between Savage's definition of "necessary views" and the actual position of Jeffreys, whom he accused of holding them. Indeed, it would be hard to cite anyone, in the entire history of probability theory, who has ever held a "necessary" view (Keynes perhaps came closest).

Reliable judgments on these matters cannot, however, be made merely by examining a writer's philosophical remarks. Doubtless all of us have had the experience of writing some interpretive statement, while our minds were preoccupied with one context; and later seeing it in print and wishing that we had made a different choice of words, since the statement as published was misleading if taken in a different context.

Any author who has written extensively over many years can be made to appear to hold almost any position one wishes to impute to him, by a carefully selective quotation of his philosophical remarks, made in many different contexts. The comments of Kass (1982) on Jeffreys demonstrate this very nicely. Even Karl Popper (1959) can be made to seem a Jeffreys type Bayesian if one quotes only from his earlier chapters and not the later ones.

For this reason we think that the issue of being or not being a "philosophical necessarian" is slippery, probably undecidable, and therefore not very relevant. What is relevant and decidable is the actual content of one's calculations; whether he can be termed a "functional necessarian".

In the early 1960's I had tried, in correspondence and conversation with Jimmie Savage, to persuade him that Laplace and Jeffreys were not necessarians, and in fact not only their philosophical remarks taken as a whole – but far more importantly, their methods of calculation – show the opposite of the position that he called "necessary".

Their use of probability theory, far from supposing that probability measures the extent to which one proposition, out of logical necessity, confirms the truth of another, clearly denies this. For the numerical value of $p(A|B)$ does not depend only on A and B; it depends also on the sample space, or hypothesis space, in which A is embedded.

Even though we do not change the propositions A and B, when we change the set of alternatives against which A is being compared, we clearly – and rightly – change $p(A|B)$. This "anti–necessary" dependence is always present implicitly, and sometimes appears explicitly, in the calculations of Laplace and Jeffreys. It seems, then, that the only functional necessarians are the sampling theorists who use significance tests that make no reference to alternatives. But Jimmie's only reaction to my arguments was to include me in his list of necessarians! More specifically (Savage, 1981, p. 542), I became "a latter-day necessarian".

Jimmie Savage thus poses a curious problem to us: so much of what he said was absolutely correct, deeply insightful, and of timely importance to statistics, that his failure to appreciate the work of Jeffreys stands out as an inexplicable puzzle. Fifteen years before Savage, Jeffreys (1939) had not only enunciated the same Bayesian principles and anticipated Savage's generalities about prior probabilities; but he also constructed explicit priors and demonstrated their successful application to real problems with a thoroughness that Savage never approached. Yet in what must have been his final judgment, Savage (1981, p. 727) still clung to his original position of 1954; in his eyes it was Jeffreys' theory that was seriously incomplete, and he saw no cogency in even its "ostensible beginnings".

My own view – then and now – has been that Savage's theory is seriously incomplete for real applications, just because of its failure to deal seriously with prior knowledge. Jeffreys had at least made a good start on remedying this.

Recently (Jaynes, 1983) I made a strenuous effort to resolve this puzzle and reconcile our different positions. There seem to be two possible explanations. Firstly, it may be that Savage gave Jeffreys a cursory reading, through the eyes of one indoctrinated by the "orthodox" teaching of the time, before his own independent thinking on the subject had matured; and just never overcame a wrong first impression.

As a second possible explanation, we note that his "necessarian" charges were directed mainly at physicists. I now conjecture that the problems Savage had in mind were quite different from the ones physicists faced. Part of the blame for this is mine, for in 1963 I had ample opportunity to point out to him, in conversation, certain technical details (existence of a deeper hypothesis space and prior knowledge about it) that gave many of our problems – and also many of engineering and economics – more structure than those of the statistics textbooks; but failed to do so.

As a result, it may be that Jimmie Savage never needed to go beyond the "diffuse prior" mentality, because he never faced a problem in which there was specific prior information that needed to be taken into account. So it appeared to him that those of us who were seeking a formal apparatus by which one can construct informative priors, were "necessarians" – doing things that seemed to him unnecessary.

3. NEW APPLICATIONS

Coming back to the present, it seems to us that if Bayesian theory is ever to lay any claims of full logical consistency, a high priority research problem must be the development of the formal apparatus that can realize the aforementioned desideratum by converting specific prior information into specific prior probability assignments, in a wider variety of problems. We face this need not only for logical reasons, but also for pragmatic ones. Today, there are important new applications where informative priors are not just window dressing; but required for the application to succeed at all.

To date we have made substantial progress in this direction, in the principles of Group Invariance, Maximum Entropy, Marginalization (Jaynes, 1980), and Coding Theory (Rissanen, 1983). As noted below, many current problems are now being solved routinely and successfully by these principles. But it is basically an open-ended program and much remains to be done.

These new applications are very recent; for the most part they have come into their own, in the sense of general recognition and acceptance, only since our last meeting here four years ago. Before then, our "necessarian" strivings did not have much effect on the treatment of real problems, because in the one problem (Statistical Mechanics) where a formal principle for determining priors was most needed, we had it already from J. Willard Gibbs – only masquerading unrecognized under a different name. In most of the other problems then being considered it was only illogical and inefficient, not fatal, to ignore prior information.

Indeed, once a model has been set up, in the "classical" problems of inference studied in the past the data D were so much more informative than our prior information I about the values of the parameters that it would have made little difference whether we used I or not. Then from a pragmatic standpoint "orthodox" or sampling theory methods were satisfactory unless there were technical problems – nuisance parameters, lack of sufficient or ancillary statistics, a rectangular likelihood function, *etc.* – that sampling theory has not learned to deal with in a satisfactory way.

But from the standpoint of logic and principle, problems of inference are basically ill-posed if prior information is not considered. Even if our model is not freely chosen but imposed on us from above, the answer to the query: "What do you know about H after seeing the data D?" depends "necessarily" on this: "What did you know about H before seeing D?"

All of us recognize this in our everyday inferences; in trying to guess whether it will rain today we take into account not only how the sky looks now, but also what the weather map showed yesterday. A medical diagnostician could be accused of malpractice if he failed to take into account the available information about a patient's medical history as well as his present symptoms. In many real situations, it would be foolhardy to "let the data speak for themselves".

Nevertheless, most current Bayesian practice tends to imitate sampling theory in that one incorporates little or no prior information beyond the choice of the model, and so seeks "noninformative" priors. The Bayesian formal apparatus can then be expected to out-perform sampling theory only when the latter faces some technical problem of the aforementioned kind. But the great potential advantage of Bayesian methods lies in exploiting this unused capability of taking prior information into account with informative priors.

Probably the most impressive example of the power of prior information is the Statistical Mechanics of J. Willard Gibbs (1902). For many years this was presented in a language and conceptual framework so different that most authors saw it as an unfinished, and only partly satisfactory, attempt to apply the laws of physics; and did not recognize it as a problem of inference at all. Only recently has Gibbs' work been seen in its simplicity and generality, as a method of inference in which prior information about multiplicity factors W converts a vague, ill-posed problem into a well-posed, accurately solvable one.

The Gibbs formalism is based on the maximization of entropy $S = \log W$ subject to the constraints of our data. This is, from our present standpoint, not an application of a law of physics, but simply locating the peak of a distribution that contains W (the "size" of the deeper microscopic hypothesis space that I failed to communicate to Jimmie Savage) as a factor. The accuracy of our predictions (sharpness of that peak) is due to the fact that W is an enormously rapidly varying function of the macroscopic quantities – and of course, that we know the laws of physics well enough to calculate it correctly.

Once this had been recognized, it was evident that the reasoning was equally applicable in other problems than thermodynamics. The recent advances in the techniques for Spectrum Analysis (Burg, 1975; Childers, 1978; Currie, 1981; Jaynes, 1982), Image Reconstruction (Gull & Daniell, 1978; Frieden, 1980; Gull & Skilling, 1980), the determination of crystallographic and biological macromolecular structure from X-ray scattering data (Bricogne, 1982; Wilkins, et al, 1983; Bryan, et al, 1983), and estimating mathematical functions from a few moments (Mead & Papanicoleau, 1983) have resulted from this recognition of the Maximum Entropy principle. In effect, it is a rule for constructing informative priors when we have partial prior information that restricts the possibilities significantly but not completely.

Also among Statisticians and Economists, several recent Bayesian works have recognized the importance of prior information and the growing need for general methods for constructing informative priors. Much thought and effort has gone into techniques for elicitation of such priors from subject-matter experts; see, for example, Kadane (1980), Winkler (1980) and references therein. Arnold Zellner's presentation at this meeting has an impressive survey of recent uses of informative priors in Econometrics.

As already noted, Litterman (1980) showed that economic forecasts using an autoregressive model can be improved by using informative priors that express common-sense judgments about the autoregressive coefficients (*i.e.* they surely fall off rapidly with increasing lag). We apply here the same idea to seasonal adjustment, showing how similar common-sense judgments about the harmonic content of the seasonal component can improve our estimates of the irregular component. It appears that in the seasonal adjustment case the effect may be greater.

4. BAYESIAN SEASONAL ADJUSTMENT

We have a discrete time y series of length N; think of it as a monthly economic report over $N/12$ years:

$$y_t = s_t + e_t, \quad 1 \le t \le N \tag{1}$$

composed of a periodic seasonal component: $s_t = s_{t+12}$, and the part e_t, variously termed "irregular", "error", or "noise". The seasonal component is represented by a finite Fourier series, containing 12 parameters, $(A_0 \ldots A_6, B_1 \ldots B_5)$:

$$s_t = A_0 + \sum_{k=1}^{6} [A_k C(kt) + B_k S(kt)] \tag{2}$$

where $C(kt) = \cos(2kt/12), S(kt) = \sin(2kt/12)$. Define, for uniform summation limits, $B_0 = B_6 = 0$. The inversions

$$A_0 = (1/N)\sum_{t=1}^{N} s_t, \tag{3a}$$

$$A_k = (2/N)\sum_t s_t C(kt) \quad 1 \le k \le 5 \tag{3b}$$

$$A_6 = (1/N)\sum_t (-)^t s_t \tag{3c}$$

$$B_k = (2/N)\sum_t s_t S(kt) \quad 0 \le k \le 6 \tag{3d}$$

are exact if N is a multiple of 12, as we suppose here.

There is still another parameter, the standard deviation σ of our prior density for the irregulars, $p(e_1 \dots e_N|I)$, where I denotes the prior information. With no creative imagination, we simply follow custom by assigning the iid Gaussian prior:

$$e_t \sim N(0, \sigma), 1 \le t \le N. \tag{4}$$

The rationale by which this custom can be justified is a rather lengthy topic; we think it is far better justified than is usually realized. More comments about this are in Appendix A.

Like any other model, this one can be extended endlessly; in particular we could combine seasonal adjustment with detrending by adding to (1) terms linear, quadratic, etc. in t. We keep our model as simple as possible so as not to obscure the point to be made; and so to heed Arnold Zeller's wise advise about "sophicatedly simple" models. Further reasons include the accuracy of our estimates and the need for diagnostic checks from limited data, discussed in Appendix C.

As noted in the Introduction, this problem is unusual in that all of these parameters are nuisance parameters. Our goal, just the opposite of the usual Bayesian goal, is to estimate the "noise" instead of the "signal".

The calculation may be organized as follows. Using the abbreviations:

$$\begin{aligned} I &= \text{prior information} \\ y &= (y_1 \dots y_N), \text{ data} \\ e &= (e_1 \dots e_N), \text{ irregulars} \\ s &= (s_1 \dots s_N), \text{ seasonal values} \\ A &= (A_0 \dots B_5), \text{ seasonal parameters,} \end{aligned}$$

we want the joint posterior density of $(e_1 \dots e_N)$ conditional on the data and the prior information

$$p(e|yI) = \int\int p(e|\sigma AyI)p(\sigma A|yI)d\sigma dA. \tag{5}$$

But if y and A are given, then e is known; so σ is irrelevant in the first factor: $p(e|\sigma yAI) = p(e|yAI)$. Then σ integrates out of the second factor, leaving,

$$p(e|yI) = \int p(e|yAI)p(A|yI)dA. \tag{6}$$

Direct evaluation of (6) would be tedious because as a function of A, $p(e|yAI)$ is nonzero only on a complicated set of points. But one can avoid going into all these intricate details by calculating first the N-fold Fourier transform of (6): using the notation $r \cdot a = \sum_t r_t a_t$,

$$E(e^{ir\cdot e}|yI) = e^{ir\cdot y} E(e^{-ir\cdot s}|yI) = e^{ir\cdot y} f(r). \tag{7}$$

So in general the calculation could proceed in three steps:

(A) Evaluate the joint posterior density $p(A|yI)$ of the seasonal parameters.

(B) Calculate, with s given by (2), the characteristic function

$$f(r) = \int e^{-ir\cdot s} p(A|yI) dA. \tag{8}$$

(C) Invert $f(r)$, translating by the data y:

$$p(e|yI) = (2\pi)^{-N} \int f(r) e^{ir\cdot(y-e)} d^N r. \tag{9}$$

Part A is the conventional Bayesian exercise. The joint likelihood of all the parameters is proportional to $p(y|AI)$:

$$L(A_0 \ldots B_5; \sigma) = \sigma^{-N} \exp\left(-\frac{Q_L}{2}\right) \tag{10}$$

where Q_L is the quadratic form

$$Q_L(A_0 \ldots B_5) = \sigma^{-2} \sum_{t=1}^{N} \left\{ y_t - \sum_{k=0}^{6} [A_k C(kt) + B_k S(kt)] \right\}^2 \tag{11}$$

from which we find that the joint maximum likelihood estimates of $(A_0 \ldots B_5)$ are given by (3) with s_t replaced by y_t (this being exact if N is a multiple of 12 as supposed; otherwise new small terms of relative order N^{-1} would be present).

We assign independent Gaussian priors for our seasonal parameters:

$$A_k \sim N(a_k, \sigma_k), \quad 0 \le k \le 6 \tag{12a}$$

$$B_k \sim N(b_k, \sigma_k), \quad 1 \le k \le 5 \tag{12b}$$

and define for formal reasons, $b_0 = b_6 = 0$. Then their joint prior distribution has another quadratic form:

$$p(A_0 \ldots B_5|I) \propto \exp(-\frac{Q_P}{2}) \tag{13}$$

where

$$Q_P = \sum_{k=0}^{6} \sigma_k^{-2} \left[(A_k - a_k)^2 + (B_k - b_k)^2 \right]. \tag{14}$$

In general the joint posterior density of the seasonal parameters will be

$$p(A|yI) \propto p(A|I) p(y|AI) = p(A|I) p(y|A\sigma I) p(\sigma|AI) d\sigma. \tag{15}$$

But if we assign independent priors to σ and A, we have $p(\sigma|AI) = p(\sigma|I)$, and our result is

$$p(A_0 \dots B_5|yI) \propto \int p(\sigma|I)\sigma^{-N} \exp\left[-\frac{(Q_L + Q_P)}{2}\right] d\sigma. \tag{16}$$

If σ is supposed known in advance, then $p(\sigma|I)$ is a delta function concentrated on a single point and in (16) we need only keep the exponential term. If σ were initially completely unknown, then the Jeffreys prior $p(\sigma|I) = 1/\sigma$ would be appropriate, and (16) would be a multivariate t-distribution with the same quadratic forms. Realistic prior information is presumably intermediate between these extremes. The choice of $p(\sigma|I)$ is discussed further in Appendix B.

For present purposes (to illustrate an entirely different point: the effect of prior information about the seasonal component), even these extremes do not lead us to very different conclusions unless we have a large amount of data that sharply contradicts the informative prior. But in that case a diagnostic check would lead us to doubt the correctness of the model or the prior information. The way in which Bayesian theory automatically provides the needed diagnostic check is explained in Appendix C. In the following we shall, therefore, suppose σ known.

Alternatively, one could say that we are calculating only $p(A|\sigma yI)$, and an integration of our final result with respect to any $p(\sigma|I)$ can still be performed.

This shortens the calculation, enabling us to bypass steps (B) and (C) above; for it is obvious that $p(e|yI)$ must be a multivariate Gaussian, determined by its first and second moments. Evaluating the Fourier transforms (which amounts to calculating all moments) is not needed. The distribution (16) reduces to

$$p(A|\sigma yI) \propto \exp\left[-\frac{(Q_L + Q_P)}{2}\right]. \tag{17}$$

Expanding these merged quadratic forms we have, to within an irrelevant additive constant,

$$Q_L + Q_P = \sum_{k=0}^{6} M_k \left[(A_k - \hat{A}_k)^2 + (B_k - \hat{B}_k)^2\right] \tag{18}$$

in which, as before, $B_0 = B_6 = b_0 = b_6 = 0$; and in consequence $\hat{B}_0 = \hat{B}_6 = 0$ from (20d) below. The reciprocal variances (exact if N is a multiple of 6) are

$$M_k = \frac{N}{\sigma^2} + \frac{1}{\sigma_k^2}, \qquad k = 0, 6 \tag{19a}$$

$$M_k = \frac{N}{2\sigma^2} + \frac{1}{\sigma_k^2}, \qquad 1 \le k \le 5 \tag{19b}$$

and the mean values (the optimal estimates with a symmetric loss function):

$$\hat{A}_0 = M_0^{-1}\left[\frac{N}{\sigma^2} \cdot \frac{1}{N}\sum_t y_t + \frac{a_0}{\sigma_0^2}\right], \tag{20a}$$

$$\hat{A}_k = M_k^{-1}\left[\frac{N}{2\sigma^2} \cdot \frac{2}{N}\sum_t y_t C(kt) + \frac{a_k}{\sigma_k^2}\right], \qquad 1 \le k \le 5 \tag{20b}$$

$$\hat{A}_6 = M_6^{-1}\left[\frac{N}{\sigma^2} \cdot \frac{1}{N}\sum_t (-)^t y_t + \frac{a_6}{\sigma_6^2}\right], \tag{20c}$$

$$\hat{B}_k = M_k^{-1}\left[\frac{N}{2\sigma^2} \cdot \frac{2}{N}\sum_t y_t S(kt) + \frac{b_k}{\sigma_k^2}\right]. \qquad 0 \le k \le 6 \tag{20d}$$

The Bayes estimates (20) are weighted averages of the prior estimates a_k, b_k, and the maximum likelihood estimates; a rather old result. In his *Essai Philosophique* (1814) Laplace discusses a similar problem where the "prior" distribution (14) is – as it could well be in our problem – actually the posterior distribution from a different set of data, and records his pleasure at finding this rule by calling it "une analogie remarquable de ce poids, avec ceux des corps compares a leur centre commun de gravite". This is possibly the origin of the term "weighted average".

Thanks to what we shall term (in conformity with Stigler's Law of Eponymy; see Appendix A) the Gaussianity of $p(A|yI)$, we need now only find the posterior expectations and covariance matrix for the irregulars:

$$\hat{e}_t = E(e_t|\sigma yI) \tag{21}$$

$$R_{tr} = E(e_t e_r|\sigma yI) - \hat{e}_t\hat{e}_r. \tag{22}$$

We find for the former

$$\hat{e}_t = y_t - g_t - \sigma^{-2}\sum_{r=1}^{N} R_{tr}y_r \tag{23}$$

where

$$g_t \equiv \sum_{k=0}^{6} \frac{[a_k C(kt) + b_k S(kt)]}{M_k\sigma_k^2} \tag{24}$$

is a kind of shrunken prior estimate of the seasonal component s_t, in which different harmonics are weighted according to their prior variances. (23) may also be written as $\hat{e}_t = y_t - \hat{s}_t$ where $\hat{s}_t$ is Laplace's "plus avantageux" weighted average estimate of s_t.

The covariance matrix R is found to be

$$R_{tr} = \sum_{k=0}^{6} M_k^{-1}\cos\frac{2\pi k(t-r)}{12} \tag{25}$$

which, like any covariance matrix, must be positive semidefinite; a direct proof of this which also determines the rank of R is given below.

The joint posterior distribution of the irregulars is therefore

$$p(e|yI) = \exp\left[-\frac{1}{2}(e-\hat{e})' R^{-1}(e-\hat{e})\right]. \tag{26}$$

This solution reveals a great deal of interesting (and to the writer unexpected) insight into the seasonal adjustment problem. To see the kind of results that are in (23) - (26) the next Sections examine the effect of different kinds of prior information $(I_1, I_2, \ldots)$, starting with very simple special cases. Even the seemingly trivial cases are instructive.

How would sampling theory deal with this problem? One answer is given by Tukey *et al* (1980). They would also subtract from the data an estimate of the seasonal as in (23); but say "few would argue" that the proper way to estimate that seasonal is by the monthly averages, which are the least squares estimates:

$$(\hat{s}_t)_{ST} = (12/N)\sum_m y_{t+12k}. \tag{27}$$

This is doubtless the most obvious thing to do. In similar problems it is much used also (to reduce the amount of data to be analyzed) by geophysicists, who call it a "Brute Stack". But we seem to be among those few; the Bayesian estimates in (23) appear so totally different from (27) that it is not clear whether there is any case in which they would agree. We shall try to understand this difference, which arises entirely from prior information that brute stacking ignores.

5. SIMPLE PRIOR INFORMATION - EXAMPLE 1

Example 1. Let the prior information be: I_1 = "There is no oscillating seasonal component, but there may be a DC offset A_0". Although this seems too trivial to be worth analyzing, let us do it anyway. Mathematically, from the general solution (23), (25) we are to pass to the limit

$$a_k \to 0, \quad b_k \to 0, \quad \sigma_k \to 0, \quad 1 \le k \le 6$$

Then $M_k^{-1} \to 0$, $1 \le k \le 6$ and (23), (25) reduce to

$$R_{tr}^{(1)} = M_0^{-1} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}, \quad 1 \le t, r \le N \tag{28}$$

$$\hat{e}_t^{(1)} = y_t - \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left(\frac{a_0}{\sigma_0^2} + \frac{N\overline{y}}{\sigma^2}\right) \tag{29}$$

with $y = N^{-1} \sum y_t$, the sample mean. The solution corrects for the unknown offset A_0 by subtracting from the datum y_t our "best" weighted average (20a) estimate of A_0.

If now $\sigma_0 \to 0$ (we know in advance that $A_0 = a_0$) this reduces, as it should, to

$$\hat{e}_t = y_t - a_0 \tag{30}$$

and in the opposite limit $\sigma_0 \to \infty$ (we have no prior knowledge of A_0) it becomes

$$\hat{e}_t = y_t - \overline{y} \tag{31}$$

and we must "let the data speak for themselves", having nothing else to rely on.

Out of all the Bayesian results that are in Case 1, Eq. (31) appears to be the only one that could have been found also from sampling theory. Of course, the prior estimate a_0 and weighting factors of (29) can hardly emerge from a theory which does not admit prior distributions. But more important are the correlations of the different e_t in their joint posterior distribution expressed by the matrix R. Clearly, if e_1 and e_2 are negatively correlated in this distribution, then we can estimate $(e_1 + e_2)$ more accurately than $(e_1 - e_2)$, while the opposite is true if they are positively correlated, as is the case here.

If our goal is to estimate some function $f(e_1 \dots e_N)$, and not just the e_t individually, these correlations in the posterior distribution can greatly affect both our estimate of f and its accuracy. Yet it appears to us that sampling theory, far from being able to take this into account, lacks even the conceptual basis and vocabulary in which one could *state* that such logical connections exist.

The reason for this is clear when we note from (25) that these correlations come entirely from the prior information; the posterior covariance matrix R can be known before one has the data y. Faced with this result, a sampling theorist will tend first to question the cogency and trustworthiness of the information contained in R. For that reason, we have pointed out the phenomenon in a simple, intuitive case.

A mechanical application of the sampling theory result (27) would estimate the sum or difference of e_1, e_2 by

$$(\hat{e}_1 \pm \hat{e}_2)_{ST} = [y_1 - (\hat{s}_1)_{ST}] \pm [y_2 - (\hat{s}_2)_{ST}] \tag{32}$$

and since the sampling distributions of y_1 and y_2 are independent, would ascribe to both the sum and difference estimates the same accuracy, given by their sampling standard deviation of $(24\sigma^2/N)^{\frac{1}{2}}$. Of course, a sampling theorist would perceive at once that in this case it would be better to use all his data in the estimator (31), reducing the error. But if he tried to judge the errors in the estimates by the sampling variances of the estimators, he would find that

$$E([y_1 + y_2]^2 | A_0) = 4A_0^2 + 2\sigma^2/N \tag{33a}$$

$$E([y_1 - y_2]^2 | A_0) = 2\sigma^2/N \tag{33b}$$

and thus conclude that the sum and difference estimates have equal probable error $(2\sigma^2/N)^{\frac{1}{2}}$.

In contrast to this the Bayesian, looking instead at the posterior distributions, would find from (28), (31)

$$E([e_1 + e_2]^2 | yI_1) = (\hat{e}_1 + \hat{e}_2)^2 + 4\sigma^2/N \tag{34a}$$

$$E([e_1 - e_2]^2 | yI_1) = (\hat{e}_1 - \hat{e}_2)^2 \tag{34b}$$

and conclude that the probable error in estimating $(e_1 + e_2)$ is larger by a factor $\sqrt{2}$ than indicated by (33); but that $(e_1 - e_2)$ can be estimated with perfect accuracy.

But this is obviously the case; for if we know that there is no oscillating seasonal component, then however poorly we may know the offset A_0, we evidently do know the difference

$$e_1 - e_2 = (y_1 - A_0) - (y_2 - A_0)$$

exactly. Indeed, all differences $(e_m - e_n)$ are known exactly.

Presumably, an alert sampling theorist will see this also, and will decide not to use sampling theory to estimate $(e_1 - e_2)$. Would he still use sampling theory for $(e_1 + e_2)$? Perhaps; if so, the Bayesian will agree with his estimate, but will say that (33a) is overoptimistic about its accuracy, by that factor $\sqrt{2}$.

But again the Bayesian result is obviously correct, for the sampling distribution variances $(2\sigma^2/N)$ in (33) are entirely irrelevant. In the sampling theorist's scenario we are to think of repeating all this many times with A_0. fixed; then our estimates would indeed vary according to (33). But the true values of the e_t would vary along with them by the same amount, so the accuracy of our estimates would have nothing to do with (33). It is determined, rather, by the accuracy of our estimate of A_0. In estimating $(e_1 - e_2)$, whatever error may be in it cancels out, as noted. But in estimating $(e_1 + e_2)$ the identical error occurs twice; whatever errors we are making in our estimates of e_1 and e_2 separately are not independent, which would lead to a variance $(2\sigma^2/N)$ as in (33); but they are perfectly correlated, leading instead to $(4\sigma^2/N)$ as in (34).

The point of this trivial example is to make it obvious that the prior information contained in R, far from lacking in cogency and trustworthiness, has restored both the agreement with deductive reasoning, and the recognition of the perfectly correlated errors, that a mechanical application of sampling theory would miss.

Of course, we do not suggest that a competent sampling theorist would fail to see these points in such a simple case; we think that common sense would force him to agree with the Bayesian results. But he would be hard put to give a sampling theory justification for them, since that common sense is using prior information that formalism of sampling theory does not recognize. In the more subtle cases to be considered next these effects of prior information are still present and just as cogent; but they are no longer obvious.

6. SIMPLE PRIOR INFORMATION - EXAMPLE 2

Example 2. Now consider the prior information: I_2 = "The seasonal component is purely sinusoidal of period 12, with no DC offset". We are to pass to the limit $a_0 \to 0$, $\sigma_0 \to 0$, and

$$a_k \to 0, \quad b_k \to 0, \quad \sigma_k \to 0. \quad 2 \le k \le 6 \tag{35}$$

Now the covariance matrix (25) reduces to

$$R^{(2)}_{tr} = M_1^{-1} \cos \frac{2\pi(t-r)}{12} \tag{36}$$

and the estimate of irregulars to

$$\hat{e}^{(2)}_t = y_t - \hat{A}_1 C(t) - \hat{B}_1 S(t) \tag{37}$$

where $\hat{A}_1$, $\hat{B}_1$ are given by (20) and as in (2), $C(kt) = \cos(2kt/12)$, $S(kt) = \sin(2kt/12)$ are the k'th harmonic seasonal sinusoids. The solution now subtracts from the data our "best" weighted average estimate of the first harmonic seasonal part. Again, in the limits $\sigma_1 \to 0$, $\sigma_1 \to \infty$ this goes into what our common sense tells us it should.

But again, the interesting result is in what the covariance matrix (36) tells us. As we shall prove in the next Section, R is now of rank 2; so our solution tells us that there are (N-2) algebraically independent functions of the irregulars:

$$f_m(e_1 \dots e_N), \quad 1 \le m \le (N-2) \tag{38}$$

that can be estimated exactly. To the writer and several others, this result was at first glance very far from obvious indeed. It first appeared in our attempt to solve this case by the Fourier transform method (8), (9), in which puzzling divergent integrals appeared in what was thought to be a straightforward, highly convergent Gaussian calculation. Some study was required before we were convinced that it is, after all, correct. But once understood, this result also can be made to seem "obvious" by the following argument.

Factor the joint posterior distribution (26) into a two-point distribution and a conditional distribution:

$$p(e_1 \dots e_N|yI) = p(e_3 \dots e_N|e_1 e_2 yI) p(e_1, e_2|yI). \tag{39}$$

Now if we know that there is only a first harmonic seasonal component, then the model equations (1), (2) reduce to

$$y_t = A_1 C(t) + B_1 S(t) + e_t. \tag{40}$$

But if e_1, e_2 are given in additional to the data y_t, then we can solve (40) for the two unknowns A_1 and B_1. Then all the subsequent values $(e_3 \dots e_N)$ are also known. In other words, the possible vectors (e) compatible with our information do not vary over a manifold of dimension N; the posterior probability $p(e_1, e_2|yI)$ on a two-dimensional manifold already contains full information and the conditional probability in (39) is a product of delta-functions.

This is why the Fourier transforms diverged, when we tried to jump directly into the limit $\sigma_k = 0$ at the beginning of the calculation. The difficulty is avoided if we first work out the general solution in the safe territory where all $\sigma_k > 0$, and then approach various limits from it.

Given e_1 and e_2, the extrapolation to all t is just

$$e_t = y_t + (y_1 - e_1)\frac{S(t-2)}{S(1)} - (y_2 - e_2)\frac{S(t-1)}{S(1)}, \quad 1 \le t \le N \tag{41}$$

for (41) has the required form of y_t plus a first harmonic and is obviously true for $t = 1$, $t = 2$; so it must be true for all t. Likewise, given any two values e_m, e_n the interpolation and/or extrapolation determining the others is

$$e_t = y_t + (y_m - e_m)\frac{S(t-n)}{S(n-m)} - (y_n - e_n)\frac{S(t-m)}{S(n-m)}, \quad 1 \le t- \le N \tag{42}$$

This is evidently most stable when $|S(n-m)| = 1$; *i.e.* when $(n-m)$ is 6 months, 18 months, etc. It fails when $(n-m)$ is a multiple of 12; for given e_n we know that $e_{n+12} = y_{n+12} - (y_n - e_n)$ and only one piece of information has been given.

If our prior information had been that there can be only two harmonics, for example $k = 1$, $k = 2$ present in the seasonal component, then this argument would work if we factor $p(e|yI)$ into $p(e_1 e_2 e_3 e_4|yI)$ and a probability conditional on $(e_1 \dots e_4)$. Then, given any four non-redundant values of the irregular, the others would be known exactly. The joint posterior density of $(e_1 \dots e_N)$ is nonzero only on a manifold of dimension 4; and so on.

We have here a kind of "anti-collinearity" phenomenon. Singularity of a covariance matrix R does not mean that some components of the vector e cannot be estimated from the data; it means just the opposite. That is, if R is of rank r, then the N-dimensional space S_N of its eigenvectors has a subspace S_r of dimension r, spanned by those eigenvectors of R with nonzero eigenvalues, in which all the posterior probability is concentrated. Vectors with non-zero projections into the complementary $(N-r)$-dimensional subspace S_{N-r} have zero posterior probability. But to keep them out of S_{N-r} requires $(N-r)$ algebraically independent conditions on $(e_1 \dots e_N)$; hence $(N-r)$ independent functions of $(e_1 \dots e_N)$ are determined exactly. Collinearity is not bad, but good!

There is a mathematical lesson to be learned from this example; from a casual glance at the posterior distribution (26) one would at first say that if R becomes singular, then the quadratic form

$$Q = (e - \hat{e})' R^{-1} (e - \hat{e})$$

blows up and (26) becomes meaningless. But the quantity of interest is not Q, but $\exp(-Q/2)$, which does not blow up. As an eigenvalue of R tends to zero, the "thickness" of the basic support set of the distribution (26) in the direction of the corresponding eigenvector goes smoothly and continuously to zero. In the limit the support set lies on a manifold of smaller dimensionality. Instead of blowing up, we therefore have a mathematically well-behaved and useful solution; (26) is nonzero only when $(e - \hat{e})$ lies entirely in the subspace S_r.

It seemed worth while to stress this point, because some working on the lore of improper priors have become entangled in "paradoxes" much less subtle than this. In working with any kind of singular mathematics we recommend very strongly the procedure that theoretical physicists have learned, over many years, to follow: (I) start from safe territory where everything is finite, convergent, and well-behaved and there is no question about what is the correct solution; (II) approach the singular cases cautiously, as limits from this.

The limit of a sequence of "good" solutions may or may not be a "good" solution in itself. A mathematically well-behaved limit is one wherein certain quantities just become smaller and smaller and eventually disappear, leaving behind a simpler analytical expression. If the limit is not well-behaved in this way (but instead, for example, blows up or oscillates forever), then the limit cannot be interpreted as a valid solution to the problem, and any attempt to find a solution by jumping directly into that limit would have led to nonsense, *the cause of which cannot be seen by looking only at the limit.* We think that most of the recent paradoxing in statistics could have been averted by following this "cautious approach" policy.

7. EFFECT OF NEW PARAMETERS

We have seen the cases in which the prior information I_1. tells us that only the DC offset A_0 is present, and I_2. that only the first harmonic component is present. Suppose now $I_3 =$ "Both are present". How are our results changed?

There is a common "folk-theorem" in the statistical literature to the effect that adding more unknown parameters in a problem must lead to a deterioration in the accuracy with which we can estimate the old parameters; and so one should not do this unless the data clearly call for it. Carrying this further, several authors state that it is fundamentally impossible to estimate more than N parameters from N data points; and practically impossible to estimate more than a small fraction of that number.

Although we feel intuitively that there must be some element of truth in these folk-theorems, we have never seen a proof of either. But from our general solution we can learn something about their validity and unstated qualifications.

First, note that if we are estimating the seasonal parameters $(A_0 \ldots B_5)$, we have from (18) that their posterior distributions are independent. Therefore our estimate of A_1 and its accuracy are the same whether or not A_0 is considered known, or whether it is also being estimated. The folk-theorem is simply wrong in this case.

There is a fairly general class of situations that include seasonal adjustment and many other problems, in which these effects are easy to understand. Consider two different problems; in problem (a), we are estimating an "old" parameter β, with prior information I_a that specifies a model containing only β. In problem (b) we are still estimating β but there is a "new" parameter ϕ, also unknown. The class of situations considered is that in which model (a) corresponds to setting $\phi = 0$. The the posterior distribution of β in the two problems are:

$$p(\beta|D, I_a) = p(\beta|D, I_b, \phi = 0) \tag{40}$$

$$p(\beta|D, I_b) = \int p(\beta|D, I_b, \phi)p(\phi|D, I_b)d\phi \tag{41}$$

respectively. The model (a) result appears, in the context of model (b), as a conditional distribution conditional on $\phi = 0$; while the model (b) result is a weighted average of conditional distributions conditional on all possible values of ϕ.

Evidently, then, considering ϕ unknown will in general make our estimate of β worse, in agreement with that intuitive folk-theorem. But the folk-theorem can also be false, in some cases, since it is possible for a weighted average of distributions to be more sharply peaked than some particular one of those distributions. In this model, in order to cause appreciable deterioration in our estimate of β, two conditions must be present: our estimate of β from the conditional posterior distribution $p(\beta|DI_b\phi)$ must depend appreciably on ϕ; and ϕ must itself be not well determined by the prior information I_b and the data D.

Once stated, this seems so obvious that intuition should have seen it long ago; but we can point to no record indicating that it actually did. It would be hard for sampling theory to state such a result because the deterioration is caused by the old parameter becoming correlated, in the posterior distribution, with an unknown quantity; and sampling theory does not recognize such a notion.

Of course, the situation may be different if introducing the new parameter causes a drastic change in the model, not just adding a new dimension to the parameter space. The two problems might be so different that there is no meaningful comparison at all.

In our seasonal adjustment problem, there is no such drastic change, and adding A_0 to the Example 2 problem affects our conclusions thus: our estimate of the oscillating seasonal component is not changed at all, because the added constant term is orthogonal to the seasonal sinusoids $C(kt)$, $S(kt)$. But our estimates of the irregulars $(e_1 \dots e_N)$ are shifted by an amount proportional to our estimate of A_0, and their posterior correlations are increased by an amount corresponding to the posterior error of our estimate of A_0. But actually A_0 can be estimated quite accurately from the data of a few years, and so the last factor in (41) is sharply peaked and there is very little effect on the accuracy of our results.

The rules (40), (41) also enable us to judge how various extensions of our seasonal adjustment model will affect our conclusions. In particular, detrending would have to be included in many real problems by adding to the model equations (1), (2) a term Ct where C is a new trend rate parameter to be estimated from the data. But the Bayesian detrended seasonal adjustment will differ from what is commonly done on intuitive grounds. Our new joint posterior distribution of $(A_0 \dots B_5)$ will not be an estimate from the detrended data; but rather as indicated by (41) we should take a weighted average of the joint distributions conditional on all possible values of C. The Bayes estimates would approach an estimate from detrended data if C were itself accurately determined by the data.

We do not go into this analysis here, but it has been carried out and we find astonishingly little change in our seasonal adjustment conclusions. The reason is that the new linear term Ct is nearly orthogonal to the seasonal sinusoids. The estimates of $(A_1 \dots B_5)$ are slightly changed by amounts proportional to our estimate of the trend rate C, their joint posterior distribution is no longer quite independent, and their probable errors are slightly increased, by a factor of $(1 - 6N^{-2})^{-1}$, which is only 4% for $N = 12$, and quite negligible if we have data for two or more years.

Thus Bayesian seasonal adjustment – contrary to what a naive application of that folk-theorem might suggest – accommodates detrending easily. Of course, estimates of the irregulars are corrected if there is evidence for a strong trend; but there is very little change in their posterior correlations or accuracy.

8. RANK OF THE COVARIANCE MATRIX

The $(N \times N)$ matrix R defined in (25) is real and symmetric; therefore it has a full set of N orthonormal eigenvectors $\mathbf{h_i} = (h_{1i} \dots h_{Ni})$ and eigenvalues β_i:

$$\sum_{r=1}^{N} R_{tr} h_{ri} = \beta_i h_{ti}, \quad 1 \le t, i \le N \tag{42}$$

Let $\mathbf{y}$ be any real $(N \times 1)$ vector with components $(y_1 \dots y_N)$, not all zero and denote scalar products of real vectors by $(x, y) = \sum x_t y_t$. Since $\mathbf{y}$ has an expansion $\mathbf{y} = \sum_i (h_i, y)\mathbf{h_i}$, the quadratic form

$$F(\mathbf{y}) = (y' R y) = \sum_{t,r=1}^{N} R_{tr} y_t y_r \tag{43}$$

can be decomposed as

$$F(\mathbf{y}) = \sum_{i=1}^{N} |(h_i, y)|^2 \beta_i. \tag{44}$$

But in our case, from (25) this is also equal to

$$F(\mathbf{y}) = \sum_{k=0}^{N} M_k^{-1} \left| \sum_t y_t \exp\left(\frac{i\pi kt}{6}\right) \right|^2 \geq 0 \tag{45}$$

therefore R is positive semidefinite. But then if $F(\mathbf{y}) = 0$, from (44) $\mathbf{y}$ must be orthogonal to all $\mathbf{h_i}$ with $\beta_i > 0$; *i.e.* $\mathbf{y}$ is itself an eigenvector with zero eigenvalue; call it a *zector* for short. The number of linearly independent zectors is $(N - r)$, where r is the rank of R.

Consider, then, the prior information I_1 of Sec. 5, that there is no oscillating seasonal component but there may be a DC offset A_0, which led in (28) to $R_{tr} = M_0^{-1}$, $1 \leq t, r \leq N$. This gives

$$F(\mathbf{y}) = M_0^{-1} \left| \sum_t y_t \right|^2 = M_0^{-1}(u_0, y)^2 \tag{46}$$

where the vector $\mathbf{u_0}$ with components $(\mathbf{1} \ldots \mathbf{1})$ is an eigenvector of R with eigenvalue $\beta = N/M_0 > 0$. But from (46) every vector $\mathbf{z}$ orthogonal to $\mathbf{u_0}$ yields $F(\mathbf{z}) = 0$, and is therefore a zector. In an N-dimensional space there are $N-1$ linearly independent vectors orthogonal to $\mathbf{u_0}$, therefore the positive eigenvalue is nondegenerate, and the rank of R is one, as stated in Sec. 5.

The prior information I_2 of Sec. 6, that only a fundamental seasonal component, of period 12, is present, led to the covariance matrix (36) in which only the term in M_1^{-1} is present. For this we have

$$\begin{aligned} F(\mathbf{y}) &= M_1^{-1} \left| \sum_t y_t \exp(\frac{i\pi t}{6}) \right|^2 = M_1^{-1} \left| (u_1 + iv_1, y) \right|^2 \\ &= M_1^{-1} \left[(u_1, y)^2 + (v_1, y)^2 \right] \end{aligned} \tag{47}$$

where $\mathbf{u_k}$, $\mathbf{v_k}$ denote the linearly independent vectors

$$\begin{aligned} \mathbf{u_k} &\text{ with components } (u_{tk} = \cos\tfrac{\pi kt}{6}, \quad 1 \leq t \leq N) \\ \mathbf{v_k} &\text{ with components } (v_{tk} = \sin\tfrac{\pi kt}{6}, \quad 1 \leq t \leq N) \end{aligned} \tag{48}$$

Evidently, $F(\mathbf{u_1}) > 0$ and $F(\mathbf{v_1}) > 0$; and any vector $\mathbf{z}$ that is orthogonal to both $\mathbf{u_1}$ and $\mathbf{v_1}$ is a zector. There are $N-2$ linearly independent zectors, so $R^{(2)}$ is of rank 2.

For the prior information I_3 which allowed the possibility of both the DC offset and the first harmonic component, R yields the quadratic form

$$F(\mathbf{y}) = M_0^{-1}(u_0, y)^2 + M_1^{-1} \left[(u_1, y)^2 + (v_1, y)^2) \right] \tag{49}$$

from which it is now evident that R is of rank 3, since every vector orthogonal to $\mathbf{u_0}$, $\mathbf{u_1}$, and $\mathbf{v_1}$ is a zector.

In general, then, we have

$$F(\mathbf{y}) = \sum_{k=0}^{6} M_k^{-1} \left[(u_k, y)^2 + (v_k, y)^2 \right] \tag{50}$$

and the rank of R is the number of scalar products appearing. Note, however, that the cases $k = 0$, $k = 6$ are special, since $v_0 = v_6 = 0$. Every nonzero M_k^{-1} contributes 1 or 2 to the rank, and when all are nonzero the maximum possible rank of R is 12.

Stated differently, every seasonal parameter in the set $(A_0 \ldots B_5)$ that is initially unknown contributes one to the rank of R.

9. CONCLUSION

Our analysis has shown how prior information about the seasonal parameters can have a major effect on our estimates of the irregulars or functions of them. In the extreme case where we know that a particular harmonic component is zero, the dimensionality of the posterior distribution is reduced. A real application will probably never be so utopian, but if we know that a particular harmonic component must be very small, R will approach a matrix of reduced rank, and as a result it will be possible to estimate some functions of the irregulars much more accurately than one would have thought from the sampling theory result (27).

Consider, briefly, what kind of prior information one might have in the case of a real economic time series. From prior knowledge familiar to all of us, we expect that department store sales will peak rather sharply in December, while sales of beer and ice cream will peak more broadly in July, bank loans to individuals may peak just before April 15, agricultural employment will peak at harvest time, etc.

In all these cases the fundamental seasonal component is clearly the major one. Indeed, it is hard to think of any case where we believe there is a repetitive mechanism tending to generate a second harmonic (*i.e.* a reason why anything should go from maximum to minimum in alternate quarters), much less any higher harmonic. That is, virtually every economic time series surely has a "driving mechanism" with a basic period of one year, the appearance of harmonics being due only to the nonsinusoidal nature of the variation, rather than to any influence that encourages repetition after a shorter period.

There are two basic kinds of periodic but nonsinusoidal behavior: (a) high-low asymmetry making, for example, sharp peaks but broad troughs, as we conjecture to be the case for department store sales; and (b) up-down asymmetry tending, for example, to make the falling portion of a curve steeper than the rising portion. It seems plausible that sales of bathing suits might rise slowly throughout the Spring and Summer, but drop precipitously in the Fall.

Type (a) behavior is represented by a Fourier series with only even order harmonics, while type (b) has both odd and even. Doubtless, both effects are present to some extent in most time series; but it seems highly unlikely that any economic quantity would exhibit only odd harmonics, making the peaks and troughs mirror images of each other (although that is the usual case for the electrical engineer's seasonal adjustment problem, elimination of complex and changing hum interference waveforms from sensitive circuits). It appears that in most cases the economist may expect the sinusoidal components of order $k = 1, 2, 4$ to predominate, while $k = 3, 4, 6$ should be much weaker.

We suggest that putting this information into our priors may make a noticeable improvement in seasonal adjustment, just as Litterman's use of priors that express our common-sense judgment that high order autoregressive coefficients are small, improves the forecasting of time series.

Clearly, what is needed now is to put these ideas to the test by analysis of real data for which conventional seasonal adjustments have been made by X-11, SABL, or some other current program. The writer will undertake to do this computation but, not being a professional economist, feels the need of help in choosing samples of data that appear to be promising for this purpose. Cases where hindsight was able to make a significant correction of the first seasonal adjustment would be particularly valuable.

APPENDIX A: WHY A GAUSSIAN ERROR DISTRIBUTION?

Stigler's Law of Eponymy (1980), illustrated by its name, states that "No scientific discovery is named after its original discoverer". Thus we find that the distribution $f(x) = \exp(-x^2/2)$ was used by Laplace (1774) three years before Gauss was born, and by de Moivre (1733) sixteen years before Laplace was born. So we are well within the Letter of the Law if we continue to call $f(x)$ a "Gaussian distribution".

The usual rationale for assigning iid Gaussian prior probabilities to "random errors" is that if the real errors are the resultant of many small independent contributions, then by the Central Limit Theorem the total error will have nearly a Gaussian frequency distribution whatever the distributions of the individual small components. Such an argument, although of course correct as far as it goes, does not take note of other reasons that may be equally cogent or more so.

Our use of the term "prior probabilities" in this context may seem unusual; conventionally, $p(e|I)$ would be called a sampling distribution. Note, however, that "sampling distributions" are from a Bayesian standpoint simply the prior probabilities we assign to the errors, or "noise", not different in logical status from prior probabilities assigned to parameters or hypotheses. Indeed, for seasonal adjustment the term seems particularly called for, since our aim is just to convert the prior probability $p(e|I)$ assigned to the noise, into its posterior distribution $p(e|yI)$.

In Bayesian inference, a sampling distribution is no more required to be a frequency distribution than is any other prior. Our aim in writing a prior distribution for a parameter is to represent our state of prior knowledge about the range of possible values that parameter may have in the specific case at hand; and this is not necessarily a frequency in any real or conceivable set of repetitions of the problem. Indeed, it will not be a frequency except in the special case (of which no example is known to us) where our prior information about the parameters consists solely of frequencies with which various values have occurred in other cases.

But in almost every real problem there are special circumstances, which make the present instance unique and not comparable to others. Then whether our parameter would or would not have the same value in some other problem that we are not reasoning about, is irrelevant for our problem.

Likewise, in writing a sampling distribution we are representing our state of prior knowledge about the possible errors that may occur in the specific case at hand – which, depending on special circumstances, may or may not be related to the frequencies of those errors in some class of other imaginary cases that we are not reasoning about.

In fact, there is almost no real problem in which we actually have prior knowledge of the frequency distribution of errors; and if we did have such knowledge, it would not in general suffice to determine a reasonable Bayesian sampling distribution. For in the case of errors as well as that of parameters, it is typical of real problems that we have prior information which does not happen to consist of frequencies; but is none the less cogent.

A rational prior error distribution, or sampling distribution $p(e|I)$ should incorporate all our prior information about the errors; not just the part of it that happens to refer to frequencies. Indeed, the frequency part is usually missing altogether or incomplete, consisting of only one or two moments of the error distribution. An experimental physicist or electrical engineer usually knows the average power level of his noise, less often something about its spectral distribution; and seldom anything else. An economist making use of past experience about the magnitude of the irregular fluctuations; but either lacking, or mistrusting the relevance of, past frequency data, would be in much the same position.

As we have argued extensively elsewhere, the prior distribution that most honestly represents our state of knowledge (*i.e.* that agrees with what we know but does not assume anything beyond that) is the one with maximum entropy subject to the constraints imposed by what we know. If the prior information fixes (or can be reasonably thought of as fixing) the first two moments, this principle will lead to the Gaussian prior distribution, as has been observed countless times.

However, this is not the whole story. Another of the common "folk-theorems" of statistics is that if we use an iid Gaussian prior for the errors, but the errors do not in fact have an iid Gaussian frequency distribution, then something terrible will happen to us and we shall be led to draw all manner of wrong and misleading conclusions. Like the other folk-theorems mentioned above, this one is in need of more careful statement.

We learn from the Asymptotic Equipartition Theorem of Information Theory (Feinstein, 1958) that the entropy of a distribution is essentially a measure of its "size"; *i.e.* over how large a volume W of the sample space is the probability density of the errors, or "noise", reasonably large? As $N \to \infty$ the iid Gaussian distribution is the one that occupies, asymptotically, the greatest possible volume of N-dimensional sample space for given first and second moments. As it is usually put in the literature of Information Theory, entropy is an asymptotic measure of the size of the "basic support set" of the distribution.

From this we see that the Central Limit Theorem is, in a very fundamental sense, a special case of the principle of maximum entropy. For first and second moments, being additive under convolution, constrain the possible distributions that can be reached by convolution. The CLT thus tells us that, baring very unusual circumstances, no other constraints exist; a distribution with finite first and second moments will expand under repeated convolution until it fills up the entire volume allowed by those two constraints. In a class of generalized CLT's, distributions could be merged in other ways than convolution, and would expand asymptotically into the one with maximum entropy subject to whatever constraints are imposed by the new method of merging.

If we assign an iid Gaussian prior $N(0, \sigma)$ to our noise sequence $(e_1 \dots e_N)$, then the high-probability volume, or basic support set, is an N-dimensional sphere R of radius about $\sigma N^{\frac{1}{2}}$. Any systematic effect that one is trying to detect by a significance test will be effectively obscured by this noise (*i.e.* we will not be able to distinguish it from the noise) if the effect is so small that the sample values remain inside the noise sphere R. It will appear to be statistically significant if and only if its effect is large enough to carry the sample values outside R.

Thus the iid Gaussian prior assignment is not a "physical assumption" that might lead us to errors if wrong; quite the contrary, it is the safest, most conservative assignment we can make if we know only the first two moments, because its high-probability region R takes into account every possible noise vector that is allowed by that information. To use any other prior assignment in this state of knowledge would amount either to contradicting our prior knowledge (if our prior had different moments); or to make additional assumptions, not warranted by our information, that contract R to some arbitrary subset $R' \subset R$; and thus invite erroneous conclusions. For if the noise vector happens to lie in a complementary set R - R', there will appear to be a real, statistically significant effect that is actually only an artifact of our particular choice of R'.

Put differently, if we assign an iid Gaussian error distribution but the frequency distribution of errors in N measurements is not, in fact, iid Gaussian, then for a given mean square error the result will not be that we shall see spurious effects that are not there; but only that our discriminating power to see small effects is not as sharp as it could be, and the accuracy of our parameter estimates is not as high as it could be.

If we had additional prior information, beyond the mean square error, about the specific way in which the noise values depart from iid Gaussian, then we could use that information to define a subset $R'' \subset R$ within which we know that the noise almost certainly lies; and then any sample that falls in $R - R''$ will be statistically significant. Thus additional prior knowledge about the noise can be crucial for deciding whether a given data set does or does not give significant evidence for a real effect, and in determining the accuracy of our estimates.

But this information need not consist of frequencies. Any information that constrains the possible N-dimensional noise vector to a subset $R'' \subset R$ will have this effect of increasing the discriminating power of our significance tests and the accuracy of our estimates. In particular, information about non-independence (the noise is not "white") makes the noise sequence in part predictable, and is thus highly valuable for extracting systematic "signals" form the noise.

Maximum-entropy spectrum analysis (Burg, 1975; Childers, 1978; Jaynes, 1982) is just the means by which we exploit the increased predictability of a time series that results from information about a few values of its autocovariance function. The maximum entropy formalism is the analytical means for locating the subset R'' defined by this information.

Likewise, maximum-entropy reconstruction of images, crystallographic, or molecular structure is the analytical means for locating the high-probability region R'', in the space of possible true states of nature, defined by the incomplete data of a blurred image. We think that future advances in pattern recognition will come from a similar taking into account of information that is ignored by sampling theory because it does not consist of frequency data.

There is no reason other than historical precedent why these methods should be confined to physics and engineering; they should find use in any application where there is cogent information that sampling theory finds indigestible.

Thus we applaud the custom of assigning iid Gaussian priors. Nothing better could have been done by the Bayesian or anyone else, unless he had additional, quite specific prior information beyond the first two moments of the noise. Whatever the "true" frequency distribution of the noise, if it is unknown then we cannot make use of it for inferences; and if it is known, additional information may be equally cogent.

But what if our prior information is so meager that we do not even know the second moment σ^2? This is perhaps the most common situation outside of physics. We do not claim it as the optimal thing to do in every case, but it is useful and computationally feasible – a kind of Bayesian jackknife – to reason as follows: if we did know σ, the Gaussian assignment would be indicated, so σ is a relevant hyperparameter and the problem reduces to assigning a prior to σ.

APPENDIX B: THE PRIOR FOR SIGMA

In discussing our choice of prior $p(\sigma|I)$, we can answer some common misgivings by noting that the Bayesian formalism automatically provides a diagnostic check on our priors and model. Of course, their adequacy cannot be tested by Bayesian or any other methods if we have only a small amount of data. But with enough data this becomes possible; for then the posterior density of σ from the Jeffreys prior becomes sharply peaked, the data alone pointing to a well-defined value of σ. A highly informative prior sharply peaked at a very different value thus stands in conflict with the evidence of the data; intuitively, we would be led to doubt the validity of our prior information or model.

This observation leads to the formal significance tests of Jeffreys, used recently by Zellner and coworkers, if we convert that intuitive feeling into a well-posed question. In our view, then, the "diagnostic phase" is indeed an essential part of inference; but it requires no departure from Bayesian methods. Quite the contrary, Bayesian methods are required for a full treatment of the diagnostic phase; nonBayesian significance tests are only approximate and/or incomplete substitutes for a full Bayesian test. These points are discussed further in Appendix C.

If we have very little data, then of course our prior distributions can matter a great deal for the conclusions we are able to draw, since our prior and posterior states of knowledge are not very different. But a t-distribution with many degrees of freedom goes asymptotically into a Gaussian, and so if we have a reasonably large amount of data, even the aforementioned extremes in the prior $p(\sigma|I)$ cannot lead us to very different conclusions about seasonal adjustment unless the data sharply contradict our informative prior, in which case we should start over again anyway. So we made the simplest choice in the text.

APPENDIX C: THE "DIAGNOSTIC PHASE" OF INFERENCE

G. E. P. Box (1982) noted the Bayesian significance test for comparing different models, but criticized it on the grounds that it could lead us to misleading conclusions if the class of alternatives did not happen to include the true one. At first glance it may appear that a test that does not refer to any specific class of alternatives is free from this objection. Such tests are indeed useful, and we do not mean to argue against using them as easy approximations, often good enough for the purpose. However, we note three reasons why a full,, well-posed diagnostic test must be Bayesian.

In the first place, not specifying a class of alternatives does not mean that the alternatives are not there; it means only that the test has not been fully defined. Any choice of test statistic, whatever rationale is given for it, is necessarily an implicit assumption about some class of alternatives. That is, given any null hypothesis H_0, data D, test statistic $t(D)$, and rule such as "reject H_0 if $t > t_0$", we are judging H_0 against a class of alternatives for which one expects large values of t.

If one fails to specify what that class is, he is not thereby prevented from applying the test; but having done so, he does not know what the test has accomplished, or what to do if H_0 is rejected. As Jeffreys (1939) put it, there is not the slightest use in rejecting any hypothesis unless we can do it in favor of some alternative known to be better.

Of course, not all "frequentist" significance tests have a Bayesian interpretation, for there are still frequentists who do not believe in the likelihood principle. For example, suppose that the point at issue is the value of some parameter β. If D and D' denote two different data sets that give the same likelihood function $L(\beta)$, but D is in the "accept" region and D' in the "reject" region, then in two situations where we have the same state of knowledge about β we are drawing different conclusions, a violation of our basic "Desideratum of Consistency". Such an irrational test cannot have – nor would we wish it to have – any Bayesian interpretation.

But a test that respects the likelihood principle (data sets that give the same likelihood function also lead to the same decision) necessarily partitions the class of possible posterior distributions $p(\beta|DI)$ into "accept" and "reject" subclasses. Such a test can be interpreted – and defined – in Bayesian terms.

Thus for the most common significance tests, Chi-squared, or the one-sided t-test and F-test, it is straightforward mathematics to construct a Bayesian significance test that leads to the same test statistic and the same procedure; but does refer to a specific class of alternatives, and therefore does tell us what the test has accomplished.

Given a null hypothesis H_0 and some class C of alternative hypotheses $(H_1, H_1, \ldots)$, a usable test, that goes at least part way toward a full Bayesian posterior odds ratio test, is to search out class C for the best alternative to H_0 by the likelihood criterion; *i.e.* given the data D, calculate the test statistic

$$t(D;C) = \max_{H \epsilon C} \log \left[p(D|H)/p(D|H_0) \right] \qquad (C1)$$

which tells us how much the data could support an alternative in C, relative to H_0.

For example, the Chi-squared test with n categories is commonly cited as a test without alternatives; yet Chi-squared is readily interpreted in Bayesian terms, as the statistic that searches out the class B_n of Bernoulli alternatives (*i.e.* n possible results at each trial, independence of different trials). Chi-squared is a two-term Taylor series approximation to $2t(D;B_n)$. Thus the numerical value of Chi-squared has a definite meaning: it tells us how much improvement in fit could be obtained within the class B_n of alternatives.

This brings up the second reason for using a Bayesian test, the logic of the decision criterion, or choice of the critical value t_0. In the traditional Chi-squared test the decision is based, not on the numerical value of Chi-squared, but on tail areas of the Chi-squared distribution conditional on H_0. But the illogical nature of any test that tries to decide solely on grounds of probabilities conditional on the null hypothesis, while simply ignoring the probabilities conditional on the alternatives, is too obvious to dwell on.

The situation is particularly embarrassing if H_0 is rejected; for surely, if we reject H_0, then we must also reject probabilities conditional on H_0; but then if no other probabilities have been used, what was the justification for the decision? Orthodox logic seems to saw off its own limb. A Bayesian test is free of this dilemma, since it decides on grounds of all the probabilities involved.

Another difficulty with interpreting the Chi-squared test as concerned with tail areas but not with alternatives, is that there is then no reason why one tail should be better than another; one ought to reject H_0 just as readily if Chi-squared turns out to be much smaller than expected. But then it would be clearly illogical to reject H_0 in favor of any alternative in B_n for H_0 is supported by the data more than any such alternative. An unexpectedly small Chi-squared could only be grounds for rejecting H_0 in favor of an alternative for which Chi-squared is expected to be small (such as one with negative correlations or some mechanism that constrains the data).

But this brings home to us the third reason why a fully satisfactory diagnostic test must be Bayesian; whether we should actually reject our model cannot be judged reasonably until we take into consideration not only the class of alternatives in some quantity like $t(D;C)$, but also their prior probabilities. Usually, one is willing to reject H_0 when Chi-squared is large, because the indicated class of alternative in B_n is judged to have a reasonably high prior probability. But one would seldom reject H_0 when Chi-squared is unexpectedly small, because the alternatives for which one expects a small Chi-squared have very low prior probability. Tests that ignore alternatives and their prior probabilities could be – and we think have been – very misleading.

A Bayesian test with a set of alternatives so poorly chosen that the true one is not even in it could indeed be misleading, in the sense that it could not lead us to the true one; but would a test that ignores alternatives leave us in any better positions? At least, the Bayesian would know what class of alternatives had been searched out, and the most likely one in that class. Surely, an antiBayesian who does not know even that much is not in less danger of falling into error; but more.

It would be very nice to have a formal apparatus that gives us some "optimal" way of recognizing unusual phenomena and inventing new classes of hypotheses that are most likely to contain the true one; but this remains an art for the creative human mind. In trying to practice this art, the Bayesian has the advantage because his formal apparatus already developed gives him a clearer picture of what to expect; and therefore a sharper perception for recognizing the unexpected. To one who expects nothing in particular, nothing can be unexpected. This applies especially to methods of data analysis that decline to use any formal apparatus at all.

REFERENCES

Box, G. E. P. (1982), "An Apology for Ecumenism in Statistics", Technical Report #**2408**, Mathematics Research Center, University of Wisconsin.

Bricogne, G. (1982). In *Computational Crystallography,* D. Sayre, Ed., Oxford University Press, pp. 258-264. Also Maximum Entropy and the Foundations of Direct Methods (1983), Dep. of Biochemistry Report, Columbia University, New York.

Bryan, R. K., Bansanl, M,. Folkhard, W., Nave, C. and Marvin, D. A. (1983). "Maximum Entropy Calculation of the electron density at 4A resolution of PF1 filamentous bacteriophage", *Proc. Nat. Acad. Sci.* USA, **80**, pp. 4728-4731.

Burg, J. P. (1975) "Maximum Entropy Spectral Analysis" Stanford University Thesis.

Childers, D., Ed. (1978), "Modern Spectral Analysis", IEEE Press, New York.

Currie, R. G., (1981), "Solar Cycle Signal in Earth Rotation", Science, 211, 386-389; See also "Evidence for 18.6 year M_N Signal in Temperature and Drought Conditions in North America Since AD 1800", J. Geophys. Res. **86**, 11055-11064.

Feinstein, A. (1958). *Information Theory*, New Your: Wiley.

Frieden, B. R. (1980), "Statistical Models for the Image Restoration Problem", Computer Graphics and Image Processing, **12**, 40-59.

Gibbs J. Willard (1902) Elementary Principles in Statistical Mechanics, reprinted in The Collected Works of J. Willard Gibbs, Vol. 2, by Yale University Press, New Haven, Conn, 1948 and by Dover Publications, Inc., New York, 1960.

Gull, S. F. and G. J. Daniell (1978), "Image Reconstruction from Incomplete and Noisy Data", Nature, **272**, 686-690; See also "The Maximum Entropy Algorithm Applied to Image Enhancement", Proc. IEEE, **5**, 170-173 (1980).

Jaynes, E. T. (1968), "Prior Probabilities" IEEE Trans. on Systems Science and Cybernetics, **SSC 4**, 227-241.

Jaynes E. T. (1981), "Marginalization and Prior Probabilities", in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner, Editor, North-Holland Publishing Co., Amsterdam.

Jaynes E. T. (1982), "On the Rationale of Maximum-Entropy Methods" Proc. IEEE, **70**, 939-952.

Jaynes, E. T. (1983), *Papers on Probability, Statistics and Statistical Physics*, D. Reidel Publishing Co., Dordrecht-Holland

Jaynes, E. T. (1984), "The Intuitive Inadequacy of Classical Statistics", in *Proceedings of the International Convention on Fundamentals of Probability and Statistics*, Luino, Italy, Sept. 17-19, 1981; D. Costantini, Editor (in press).

Jeffreys H. (1939), *Theory of Probability*, Oxford University Press, London.

Kadane J. B. (1980), "Predictive and Structural Methods for Eliciting Prior Distributions", in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner, Editor, North-Holland Publishing Company, Amsterdam.

Kass R. (1982), "A Comment on 'Is Jeffreys a Necessarist?"', Am. Stat. **36**, 390-392.

Laplace, P. S. (1814). *Essai Philosophique sur les Probabilities,* Courcier Imprimeur, Paris; reprints of this work and of Laplace's much larger *Theorie Analytic des Probabilities* are available from Editions Culture et Civilisation, 115 Ave. Cabriel Lebron, 1160 Brussels, Belgium.

Litterman R. (1980), "A Bayesian Procedure for Forecasting with Vector Autoregression", Manuscript dated September 1980 and Ph.D. Thesis, University of Minnesota, 1979.

Mead, L. R. and Papanicolaou, N. (1984). "Maximum Entropy in the Problem of Moments," *J. Math. Phys.* **25**, 2404-2417.

Popper K. R. (1959), *The Logic of Scientific Discovery*, Basic Books, Inc., New York.

Rissanen J. (1983), "A Universal Prior for the Integers", Ann. Stat., July 1983.

Savage L. J. (1954), *The Foundations of Statistics*, J. Wiley & Sons, Inc., New York.

Stigler, S. M. (1980). "Stigler's Law of Eponymy," *Trans. New York Acada. Sci.* **39**, 147-159.

Tukey, J. W. (1978). Granger on Seasonality, in *Seasonal Analysis of Time Series,* U. S. Government Printing Office: A. Zellner, Editor, Washington.

Van Dantzig, D. (1957). "Statistical Priestcraft: Savage on Personal Probabilities," *Statistica Neerlandica,* **11** 1-16.

Wilkins, S. W., Varghese, J. N. and Lehmann, M. S. (1983). "Statistical Geometry I. A Self-Consistent Approach to the Crystallographic Inversion Problem Based on Information Theory," *Acta Cryst.* **A39**, pp. 47-60.

Winkler, R. L. (1980). "Prior Information, Predictive Distributions, and Bayesian Model Building," in *Bayesian Analysis in Econometrics and Statistics,* North-Holland, Amsterdam: A. Zellner, Editor.

Zellner, A. (1982). "Is Jeffreys a Necessarist?" *Am. Stat.* **36,** 28-30.

DISCUSSION

J. J. DEELY *(University of Canterbury, NZ)*

Firstly, I must apologize to Professor Jaynes for not having enough time to verify the details of his paper. I had to leave New Zealand early in August and have been on the go ever since. I should also point out to him that his paper is the only one presented during this conference in which *two* discussants were devoted solely to one paper. Perhaps the fact that his first two initials are "E. T." had some bearing.

Professor Jaynes is to be commended for his labour of love in reminding us again that the Maximum Entropy Principle (MEP) provides us with a method which converts prior information into highly informative priors. This paper provides a new orchestration for that old song. He uses a simple forecasting model:

$$y_t = s_t + \epsilon_t, \quad 1 \leq t \leq N$$

where

$$s_t = A_0 + \sum_{k=0}^{6} \left(A_k \cos \frac{2\pi kt}{12} + B_k \sin \frac{2\pi kt}{12} \right)$$

with $B_6 = 0$. This implies a seasonal model with $s_t = s_{t+12}$ for every t. The coefficients are assumed to have prior distributions, $A_k \sim N(a_k, \sigma_k)$, $B_k \sim N(b_k, \sigma_k)$ and the errors ϵ_t are independent and $N(0, \sigma)$. It is the choice of these prior distributions which causes me concern. Professor Jaynes has stated "... a high priority research problem must be the development of the formal apparatus that can realize the aforementioned desideratum, *i.e.* to have a theory which deals seriously with prior knowledge by converting specific prior information into specific prior probability assignments, in a wider variety of problems," and "... problems of inference are basically ill-posed if prior information is not considered". he then adds "... it (MEP) is a rule for constructing informative priors when we have partial prior information that restricts the possibilities significantly but not completely". I am in complete agreement with these statements, but what has not been done is to convince me that MEP is a *good* way to deal with prior information. This probably sounds like heresy but I feel the subjective atmosphere here is so health that diverse opinions can be openly expressed. (I hope that future Bayesian Symposiums will zealously guard this attitude). The reason for my saying that I am not convinced about MEP begins with a lecture in my undergraduate days at Georgia Tech. who introduced the subject of Entropy in a thermodynamics course, by saying, "you will never use this again" and I believed him! More importantly my statistical training brought me through classical procedures and least favorable prior distributions, and to use Bayes only for admissibility. Hence it is not surprising that a paper, Deely, Zimmer and Tierney (1970) appeared, which showed that MEP does not correspond to least favorable priors, not even in the restricted minimax sense.

I give this bit of history to show how far off the MEP track I am. To reach me and others like me (if there are any?) a decision theoretic structure will have to be used or something similar using beautiful mathematics – even hand waving – but not repeating over and over again that MEP is *good*. Jim Berger has already suggested in his paper that using the decision theory structure as a check for various procedures may not be a bad conversion this week and that word is "neutral". In the paper he describes it in the following way, "As we have argued extensively elsewhere, the prior distributions that most honestly represents our state of knowledge (*i.e.* that agrees with what we know but does not assume anything beyond that) is the one with maximum entropy subject to the constraints imposed by what we know". Thus all that remains to do is to give a nice intuitive decision theoretic definition of "neutral" or "honestly represents" and then derive necessary and sufficient conditions for MEP to be neutral.

Now considering the specific applications in this paper, I'm not surprised at the nice results. However as Professor Jaynes admits, his examples and results for them are not realistic. "Clearly, what is needed now is to put these ideas to the test by analysis of real data ..." I certainly concur with this and would like to point out one definite area where I think his model is unrealistic. To allow for changes in the seasonal vectors over time, it seems to me that the coefficients A_k and B_k $(k = 1, \ldots 6)$ should be allowed to change. One way to do this is to imagine the coefficients being drawn repeatedly each 12 months from the same distribution. Thus given data for $N/12 = n$ years, forecasting the future requires updating the prior information using the past data for $n - 1$ years. In this context I believe MEP would be incoherent, and possibly I can prove this by the next Symposium.

A. F. SMITH *(University of Nottingham)*:

I believe that Jaynes' paper contain a very important message but I'm not sure that it has anything to do with Time Series or Entropy. However, it does concern a topic which Jaynes emphasized at our previous conference: namely, the notation of "question posing" as a dual activity to "inference" (formalized, to some extent, in the work of R. T. Cox). In particular, the business of deciding which questions can be well-answered by a particular data set (given a parametric model and a prior specification) can be thought of as corresponding to a principal component analysis in parameter space, performed on the posterior covariance matrix (perhaps following various exploratory non-linear transformations of individual parameters). Part of Jaynes' analysis seems close in spirit to this idea, with the added refinement that – with all other ingredients fixed – one can give an operational meaning to "highly informative priors" as those which lead to a simplifying principal component analysis.

J. R. M. AMEEN *(University of Warwick)*:

It is appealing that prior probabilities should be considered as highly informative relative to the hypotheses under which the data is collected, when the data have almost nothing to tell on the basis of some discrimination measure between prior and posterior probabilities. Is this the procedure under which Professor Jaynes' prior specification turn out to be highly informative?

Regarding model specification, although the simplicity and restrictedness seem to be for the sake of argument, the Dynamic Linear Models of Harrison and Stevens (1976) may have been more justified.

Prior probabilities need not be completely ruined by the occurrence of a sharp change in the data as it is stated at the end of appendix B of the paper. This point is emphasized with practical examples in Ameen and Harrison (1983).

J. BERGER *(Purdue University)*:

This was a very enjoyable paper to read, not only because of the interesting Bayesian analysis of seasonal adjustment, but also because of the many philosophical meanderings that are sprinkled throughout. With only two of these meanderings did I not feel immediate agreement.

One was in Appendix C, in the discussion of significance testing "sawing off its own limb". the "justification" for the decision to reject H_0 if the tail area probability of the significance test is small, is surely just that data inconsistent with the hypothesis casts doubt on the hypothesis. I fully agree, however, that any attempt to treat a tail area significance level as a measure of doubt in any quantifiable sense is generally meaningless. It is perhaps this which Professor Jaynes is calling "limb-sawing" illogic: how can a significance level provide any *absolute* measure of doubt for the hypothesis if it is calculated under the hypothesis and the hypothesis seems wrong. Of course, the comments about the need to consider alternatives raise the main concern with significance testing.

The only real disagreement I had was with the justification in Appendix A of the Gaussian error distribution. I view maximum entropy as a wonderful tool but, as with all wonderful tools, fear the tendency to make the problem fit the tool. In particular, maximum entropy, as here, works best when one assumes that prior knowledge provides moments of the distribution. Now I can conceive of many situations where such may be the case but, in the majority of situations that I see, prior knowledge is more likely to be in the form of, say, medians and quartiles of a distribution. Medians and quartiles can be assessed by consideration of sets with large probability, while moments depend on delicate features of a distribution (such as its tail behavior) which are hard to assess. It may be a mistake to assume that moments even exist (*i.e.,* the error distribution could have a fat enough tail that, as an approximation, it would better to proceed as if the moments were infinite). The rather unsatisfactory behavior of maximum entropy distributions for specified medians and quartiles (and continuous distributions) is then a source of concern. The justification of normal error distributions (or, for that matter, normal prior distributions) by maximum entropy thus leaves me somewhat uneasy.

W. H. DuMOUCHEL *(M. I. T.)*:

I agree strongly with Professor Jaynes that the real opportunities for Bayesians lie in the use of informative priors. How else could we hope to do better than frequentists? To use noninformative priors is, basically, to play on their turf.

The author's discussion of the special problems involved in seasonal adjustment is interesting, but, in spite of his apology, I regret the absence of real data in his presentation. One wants to see the methods in action – which harmonics re used? Why? What values of σ and b are used? Quite possibly Professor Jaynes might develop more sympathy for those who allow arbitrary harmonics to enter their models if he had experience with a wide variety of real series, only a moderate proportion of which follow idealized seasonal models. For example, there may be a spike in which one month stands out from its adjacent months, or there may be more than one harvest time, several months apart, etc. If simple fits to trigonometric series work consistently will, frequentists should be able to profit from them as well as Bayesians. Show us the data!

The assumption that the prior distributions of the size of the first and second harmonics are independent doesn't seem reasonable to me. If one switches to polar coordinates, so that R_1^2 and R_2^2 are $(A_1^2 + B_1^2)$ and $(A_2^2 + B_2^2)$, respectively, then I wold be more willing to believe that R_1 is independent of R_2/R_1.

Finally, I am afraid I don't see the point of the author's elaborate discussion of the singular posterior distributions of the $\{e_i\}$, at least in the context of the Bayes/non-Bayes controversy. Classical statisticians at least since Fisher have known that the residuals and the fitted values form a Regression model each having singular distributions. Isn't that all you are saying?

I. J. GOOD *(Virginia Polytechnic Inst. and Statue Univ.)*:

Regarding minimax entropy I'd like to draw attention to my comment on the paper by Bernardo and Bermudez.

A previous discussant said that the principle of maximum entropy is not a minimax principle with the usual loss functions. But it seems to me that he minimization of expected weight of evidence (minimum discrimination information), which is a generalization of the maximization of entropy, must be a minimax procedure if weight of evidence is regarded as a utility or quasi-utility (Good, 1969). This is by virtue of Wald's theorem that the least favorable prior distribution is minimax. The minimax property by itself would not be of much interest if weight of evidence were not a reasonable quasi-utility having the property of invariance under transformations of the independent variable for continuous distributions, and having the analogous "splitative" property for discrete distributions. That is, weight of evidence, in the discrete case, is unchanged if categories are split, such as tossing a coin (Good, 1973).

REPLY TO THE DISCUSSION

It is I who must apologize to John Deely for failure to get my complete manuscript to him before this meeting. Desperate attempts to meet three publication deadlines simultaneously led also to omission of some of the results (particularly numerical analysis of real data) that I had hoped to present.

But all heard my vow not to speak at any more meetings until I could present such an analysis. This should go a long way toward meeting the vary valid criticism of Deely and others, that one does need to judge performance in the arena of real applications, and not just philosophy and theorems. Nobody feels that need more strongly than I, and only the pressure of other prior commitments has deterred me.

However, we are not operating in a complete vacuum: my paper gives several reference to recent work where useful numerical results, not duplicated by other statistical methods, have been obtained with MEP. Applications in several different fields are now growing so rapidly, the number of workers appearing to double every year, that it is no longer possible for one person to keep up with what is being done.

And this growth owes nothing to philosophy of theorems. In every field of application, nobody will believe that the method will work (the philosophy is attacked and the theorems are ignored) until somebody actually applies it and shows that it does work. It is the computer printouts – the sharp detail, uncluttered by the spurious artifacts that were generated by previous methods of data analysis like lag windows and inverse filters – that have produced the converts. A method that gives such results would be used just as much if it had no theoretical justification at all.

Today, we are rather far beyond the stage of "repeating over and over again the MEP is good" which would have been a valid criticism in 1965. Even at the time of Deely's 1970 criticism of MEP, Buge's Maximum-Entropy Spectrum Analysis (MESA) had been available in the literature for three years. By 1978 the literature had grown to the point where the IEEE issued a special volume (Childers, 1978) of reprints on MESA. The September 1982 IEEE Proceeding is a special issued devoted largely to it. But there is still a serious shortage of such numerical results outside of physics and engineering. This can and will be corrected.

In a recent talk George Tias noted that: "In the marketplace of statistical ideas, there are many sellers but few buyers" – a profoundly true observation in spite of the seeming paradox that before one can become a seller he must first have been a buyer. The explanation is that one must be a seller for a very long time, because of the difficulty – more acute in statistics than in any other field – of clarifying to potential byers exactly what it is that one is trying to sell.

Nothing could be further from our intention than to construct a "least favorable" prior. We seek rather a prior that deals most honestly with our prior information by representing the whole truth and nothing but the truth, thus enabling one to separate the truth from the artifacts. The philosophy, the theorems, and the computer printouts all support the view that MEP is accomplishing this. On the other hand, "least favorable" priors ignore – and therefore in general contradict – our prior information, and could be disastrous in these problems.

I cannot imagine what one could mean by a decision theoretic justification for MEP. Wald's decision theory makes no contact with any principles for assigning priors. Various theoretical justifications for MEP are in the literature, but they appeal rather to requirements of logical consistency, which are neutral toward the value judgments underlying decision theory. This neutrality appears to me the very essence of "scientific objectivity".

Similarly, it was not my intention to discuss all the real problems of seasonal adjustment. My paper formulated and solved one specific problem in order to demonstrate one point, the effect that prior information can have in seasonal adjustment. Believing that point to be new, we did not want to obscure it with unnecessary details. In various situations the model chosen may indeed be "not realistic" in may different ways; and so we noted the possibility, and the ultimate necessity, of extending it. But, having seen this one solution, the extensions are straightforward exercises that John Deely could assign to his students as homework.

In reply to Adrian Smith:

My presentation had, indeed, very little to do with entropy which was, after all, invoked only to justify what everybody has been doing all along. It is surprising that some other comments are so preoccupied with entropy, since it the same Gaussian prior had been called simply "assumptions" nobody would have questioned them. It seems that new rational are disturbing even when they support previous practice, as Harold Jeffreys and Jimmie Savage found also.

There was, however, something to do with time series. Just how much will be clearer when those promised but still undelivered analyses of real data are at hand.

I am in full – indeed, enthusiastic – agreement with Adrian Smith on the importance, for practice and future theoretical development, of recognized that the "design of models" or "design of hypotheses" cannot be separate from the "design of experiments". Now that we are happily beyond the stage of mysticism (was the specific Latin square generated by a *truly* random selection process?), a new attack on these problems is possible, in which we give just as much attention to: "Which questions can be well-answered by a given kind of data?" as to the converse: "Which kind of data can best answer a given question?"'

The instincts of good scientists have always told them that it is idle to invent hypotheses which cannot be tested by the data that it is feasible to get. Progress is always made by asking the questions that are answerable at the time. For Isaac Newton it would have been foolish to ask questions which were not foolish for Erwin Schödinger 250 years later. For Gregor Mendel it would have been foolish to ask questions which were not foolish for Francis Crick today. The same point was made by Arnold Zellner's call for "sophisticatedly simple" models in econometrics.

With the good beginnings made by R. T. Cox, it should be possible to formalize this intuitive wisdom in a new theory of scientific inference, which includes the "diagnostic phase" and hypothesis formulation automatically, and will prove to be a far more powerful tool than our intuition; just as our present Bayesian principles formalize and strengthen our intuition about plausible reasoning from given hypotheses. In both, prior information will be of crucial importance. The principle components analysis in parameter space suggested by Adrian Smith may be a very good starting point for this quest.

In reply to Jack Good:

Jack Good's remarks illustrate an important point that deserves to be emphasized; in statistics it is the rule that a given procedure may be interpreted and justified in various ways.

The maximum entropy *principle*, as a rational for assigning priors, was questioned by John Deely who wanted to see instead a minimax decision theory type argument; and by Jim Berger, who left us wondering what he would like instead. Deely's goal seems to me dubious because priors and decisions lie at opposite ends of the inference-action chain.

In contrast, the maximum entropy *procedure* exists independently of any rational' and my comments on Professor Csiszar's paper noted one conceivable way of interpreting it in decision-theory terms. Jack Good's comments seem to offer a rational for thinking of entropy as related to utility rather than priors, and so tend to support this view.

But we are not compelled to choose one view and abandon the others; perhaps they are both appropriate n different circumstances, and this shows only that he procedure solves more than one problem. Such a situation is not new in statistics.

Indeed, as John Tukey has noted, a procedure does not have hypotheses, and really needs no rational at all; its justification lies in the results it gives. In this connection, nothing in the procedure requires the quantity whose entropy is maximized to be even a probability.

As far as the mathematics is concerned, "any" non-negative integrable function $f(x)$ has an entropy $H = -\int f(x)\log f(x)dx$, whose maximization is a well-posed problem under very general conditions. The solution will be a purely mathematical result, independent of whatever conceptual meaning you or I might attach to $f(x)$.

Some of the current applications take advantage of this flexibility; the cited work of Papanicolaou and Meed finds the maximum entropy procedure a surprisingly powerful method for approximating functions $f(x)$ about which only a few moments are know, wit advantages in both accuracy and stability over other common schemes such as Padé approximants.

Other interesting examples of the procedure escaping from the confines of its original rational (such as filling gaps in incomplete contingency tables) are in the articles cited and in others shortly to appear.

Jim Berger's comments came as welcome relief because, unlike most of my recent discussants and commentators, he does not ask me to defend statements that I did not make. Instead, I am able to note that I did say, in different words, the things that the mentioned. But his remarks end abruptly just at the most interesting point; I hope he will end the suspense by continuing his lie of thought elsewhere.

On the trivial "limb-sawing" matter, surely we agree that probabilities conditional on an hypothesis that we have rejected stand in a rather peculiar logical position. Would you use them for future predictions? In principle, it must be the probability of the hypothesis, conditional on the data, that justifies our decision to reject it. Pragmatically, it doesn't matter because the two probabilities are related mathematically through Bayes' theorem. But orthodox statistics cannot say it that way, and so puts up a logically puzzling rationale.

The important matter here is the status of iid Gaussian sampling distributions and of maximum entropy as a principle for assigning probability distributions. As Jim correctly notes, the former is only a special case of the latter. I too stated that this rationale for the Gaussian assignment holds when the prior information consists of the first two moments (or just the second moment) of the noise; and that when our prior information is different, a different prior distribution will be appropriate. So we are in agreement here.

But Jim and I seem to work on different kinds of problems; my experience is that in the great majority of those arising in physics and engineering the first two moments are exactly the prior information we have, because they are connected directly to the noise energy. I have never seen a real problem in which we had prior information about percentiles, and would need to know more about them (are percentiles the only information?) before deciding what to recommend.

Now we come to the serious point that calls for more explanation. He suggests that if the prior information does not consist of moments, then not only will the id Gaussian distribution not be appropriate, but also that the maximum entropy principle may not be appropriate to find the new distribution. This is tantalizing because it seems to imply either that eh proposes to leave us with no principle, or that some other principle will then be more appropriate.

All I can say to this is that, if Jim Berger or anyone else is in position of a principle for assigning priors that can, while using the same prior information, give more satisfactory results than does maximum entropy in even one cases, then I and a few thousand others are breathlessly eager to hear what it is. If I knew of a better principle, I would be expounding it and using it in my current research.

In reply to Dr. Ameen:

The point of may work was to show the surprisingly large effect that prior information can have in seasonal adjustment. to do that, one naturally chooses one specific model to analyze, and keeps it as simple as possible. Anyone who wishes to analyze a different model is perfectly free to do so. Presumably, the same sensitivity to prior information will be found in any model of seasonal adjustment – or indeed any problems where we want to estimate "noise" – if it is given a full Bayesian analysis.

Finally, in response to Prof. DuMouchel:

Professor DuMouchel calls on me to defend statements that I did not make. As in other replays, I must emphasize that the purpose of my presentation was to demonstrate, by analyzing a simple case, the large effect that prior information can have in seasonal adjustment. It is indeed true that in the real world one can find other cases than the one I chose. Therefore extensions of my model, in a dozen different ways, are also of interest; by all means, let us study them.

Let me assure everybody that, being one of them, I have the greatest sympathy for "those who allow arbitrary harmonics to enter their models". Nothing in my argument forbids one to model whatever harmonics he pleases. But in the model I chose to analyze (known period of one year, monthly data) there are no higher harmonics than the sixth.

Likewise, nothing forbids one to consider different prior probabilities than the ones I chose. I one has evidence supporting a non-independent prior distribution for the first and second harmonics, by all means use it. Presumable, still more interesting things will then be found. I would be very interested to learn how these different kinds of prior information interact; perhaps Professor DuMouchel might tell us about this at a future meeting.

Having statisticians since Fisher known about the singular posterior distributions for the $\{e_i\}$ that I pointed out? How could they even recognize the existence of correlations in any posterior distributions, without becoming Bayesians?

REFERENCES IN THE DISCUSSION

Ameen, J. R. M. and Harrison, P. J. (t984). "Normal discount Bayesian models." This volume.

Deely, J. J., Zimmer, W. J. and Tierney, M. S. (1970). "On the usefulness of the maximum entropy principle in the Bayesian estimation of reliability." *IEEE Trans. Re. R*-**19,** 110-115.

Good, I. J. (1969). "What is the use of a distribution." *Multivariate Analysis II,* P. R. Krisnaiah (ed). New York: Academic Press, 183-203.

Good, I. J. (1973). "Information, reward, and quasi-utilities." *Science, Decision, and Value.* J. J. Leach, R. Butts & G. Pearce (eds.). Dordrecht: D. Reidel, 115-127.

Harrison, P. J. and Stevens, C. S. (1976). "Bayesian forecasting." *J. Roy. Stat. Soc.* **B, 38,** 205-247 (with discussion).