chirp, 7/8/1996

BAYESIAN SPECTRUM AND CHIRP ANALYSIS[†]

E. T. Jaynes
Wayman Crow Professor of Physics
Washington University, St. Louis MO 63130

Abstract: We seek optimal methods of estimating power spectrum and chirp (frequency change) rate for the case that one has incomplete noisy data on values $y(t)$ of a time series. The Schuster periodogram turns out to be a "sufficient statistic" for the spectrum, a generalization playing the same role for chirped signals. However, the optimal processing is not a linear filtering operation like the Blackman–Tukey smoothing of the periodogram, but a nonlinear operation. While suppressing noise/side lobe artifacts it achieves the same kind of improved resolution that the Burg method did for noiseless data.

CONTENTS

[†] A revised and updated version of a paper presented at the Third Workshop on Maximum–Entropy and Bayesian Methods, Laramie Wyoming, August 1–4, 1983. Published in *Maximum–Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Editors, D. Reidel Publishing Co., 1987. This is the original investigation that evolved into the Bayesian Spectrum Analysis of Bretthorst (1988). The "plain vanilla" Bayesian analysis introduced here proved to be unexpectedly powerful in applications because of its flexibility, which allows it to accommodate itself to all kinds of complicating circumstances in a way determined uniquely by the rules of probability theory, with no need for any *ad hoc* devices or appeal to unreliable intuition.

1. INTRODUCTION

The Maximum Entropy solution found by Burg (1967, 1975) has been shown to give the optimal spectrum estimate – by a rather basic, inescapable criterion of optimality – in one well–defined problem (Jaynes, 1982). In that problem we estimate the spectrum of a time series $(y_1 \cdots y_N)$, from incomplete data consisting of a few autocovariances $(R_0 \cdots R_m)$, $m < N$, measured from the entire time series, and there is no noise.

This is the first example in spectrum analysis of an exact solution, which follows directly from first principles without *ad hoc* intuitive assumptions and devices. In particular, we found (Jaynes, 1982) that there was no need to assume that the time series was a realization of a "stationary Gaussian process". The Maximum Entropy principle automatically created the Gaussian form for us, out of the data. This indicated something that could not have been learned by assuming a distribution; namely that the Gaussian distribution is the one that can be realized by Nature in more ways than can any other that agrees with the given autocovariance data. In this sense it is the 'safest' probability assignment one could make from the given information. This classic solution will go down in history as the "hydrogen atom" of spectrum analysis theory.

But a much more common problem, also considered by Burg, is the one where our data consist, not of autocovariances, but the actual values of $(y_1 \cdots y_N)$, a subset of a presumably longer full time series, contaminated with noise. Experience has shown Burg's method to be very successful here also, if we first estimate m autocovariances from the data and then use them in the MAXENT calculation. The choice of m represents our judgment about the noise magnitude, values too large introducing noise artifacts, values too small losing resolution. For any m, the estimate we get would be the optimal one if (a) the estimated autocovariances were known to be the exact values; and (b) we had no other information beyond those m autocovariances.

Although the success of the method just described indicates that it is probably not far from optimal when used with good judgment about m, we have as yet no analytical theory proving this or indicating any preferred different procedure. One would think that a true optimal solution should (1) use all the information the data can give; *i.e.* estimate not just $m < N$ autocovariances from the data, but find our "best" estimate of all N of them and their probable errors; (2) then make allowance for the uncertainty of these estimates by progressively de–emphasizing the unreliable ones. There should not be any sharp break as in the procedure used now, which amounts to giving full credence to all autocovariance estimates up to lag m, zero credence to all beyond m.

In Jaynes (1982) we surveyed these matters very generally and concluded that much more analytical work needs to be done before we can know how close the present partly *ad hoc* methods are to optimal in problems with noisy data. The following is a sequel, reporting the first stage of an attempt to understand the theoretical situation better, by a direct Bayesian analysis of the noisy data problem. In effect, we are trying to advance from the "hydrogen atom" to the "helium atom" of spectrum analysis theory.

One might think that this had been done already, in the many papers that study autoregressive (AR) models for this problem. However, as we have noted before (Jaynes, 1982), introducing an AR model is not a step toward solving a spectrum analysis problem, only a detour through an alternative way of formulating the problem. An AR connection can always be made if one wishes to do so; for any power spectrum determines a covariance function, which in turn determines a Wiener prediction filter, whose coefficients can always be interpreted as the coefficients of an AR model. Conversely, given a set of AR coefficients, we can follow this sequence backwards and construct a unique power sectrum. Therefore, to ask, "What is the power spectrum?" is entirely equivalent to asking, "What are the AR coefficients?" A reversible mathematical transformation converts one formulation into the other. But while use of an AR representation is always possible, it may not be appropriate (just as representing the function $f(x) = \exp(-x^2)$ by an infinite series of Bessel functions is always possible but not always appropriate).

Indeed, learning that spectrum analysis problems can be formulated in AR terms amounts to little more than discovering the Mittag–Leffler theorem of complex variable theory (under rather general conditions an analytic function is determined by its poles and residues).

In this field there has been some contention over the relative merits of AR and other models such as the MA (moving average) one. Mathematicians never had theological disputes over the relative merits of the Mittag–Leffler expansion and the Taylor series expansion. We expect that the AR representation will be appropriate (*i.e.*, conveniently parsimonious) when a certain response function has only a few poles, which happen to be close to the unit circle; it may be very inappropriate otherwise.

Better understanding should come from an approach that emphasizes logical economy by going directly to the question of interest. Instead of invoking an AR model at the beginning, (which might bring in a lot of inappropriate and unnecessary detail, and also limits the scope of what can be done thereafter), let us start with a simpler, more flexible model that contains only the facts of data and noise, the specific quantities we want to estimate; and no other formal apparatus. If AR relations – or any other kind – are appropriate, then they ought to appear automatically, as a consequence of our analysis, rather than as arbitrary initial assumptions.

This is what did happen in Burg's problem; MAXENT based on autocovariance data led automatically to a spectrum estimator that could be expressed most concisely (and beautifully) in AR form, the Lagrange multipliers being convolutions of the AR coefficients: $\lambda_n = \sum_k a_k a_{n-k}$. The first reaction of some was to dismiss the whole MAXENT principle as "nothing but AR", thereby missing the point of Burg's result. What was important was not the particular analytical form of the solution; but rather the logic and generality of his method of finding it.

The reasoning will apply equally well, generating solutions of different analytical form, in other problems far beyond what any AR model could cope with. Indeed, Burg's method of extrapolating the autocovariance beyond the data was identical in rationale, formal relations, and technique, with the means by which modern statistical mechanics predicts the course of an irreversible process from incomplete macroscopic data.

This demonstration of the power and logical unity of a way of thinking, across the gulf of what appeared to be entirely different fields, was of vastly greater, and more permanent, scientific value than merely finding the solution to one particular technical problem.

Quickly, the point was made again just as strongly by applying the same reasoning to problems of image reconstruction, of which the work of Gull and Daniell (1978) is an outstandingly concise, readable example.

I think that, 200 years from now, scholars will still be reading these works, no longer for technical enlightenment – for by then this method of reasoning will be part of the familiar cultural background of everybody – but as classics of the History of Science, which opened up a new era in how scientists think.

The actual reasoning had, in fact, been given by Boltzmann and Gibbs long before; but it required explicit, successful applications outside the field of thermodynamics before either physicists or statisticians could perceive its power or its generality.

In the study reported here we have tried to profit by these lessons in logical economy; at the beginning it was decided not to put in that fancy entropy stuff until we had done an absolutely conventional, plain vanilla Bayesian analysis – just to see what was in it, unobscured by all the details that appear in AR analyses. It turned out that so much surprising new stuff was in it that we are still exploring the plain vanilla Bayesian solution and have not yet reached the entropic phase of the theory!

The new stuff reported here includes what we think is the first derivation of the Schuster periodogram directly from the principles of probability theory, and an extension of spectrum analysis to include chirp analysis (rate of change of frequency). Before we had progressed very far it became evident that estimation of chirp is theoretically no more difficult than "stationary" spectrum estimation. Of course, this takes us beyond the domain of AR models, and could never have been found within the confines of an AR analysis.

Our calculations and results are straightforward and elementary; in a communication between experienced Bayesians it could all be reported in five pages and the readers would understand perfectly well what we had done, why we had done it, and what the results mean for data processing. Such readers will doubtless find our style maddeningly verbose.

However, hoping that this work might also serve a tutorial function, we have scattered throughout the text and appendices many pages of detailed explanation of the reasons for what we do and the meaning of each new equation as it appears.

Also, since consideration of chirp has not figured very much in past spectrum analysis, and Bayesian analysis has not been very prominent either, the next three Sections survey briefly the history and nature of the chirp problem and review the Bayesian reasoning format. Our new calculation begins in Sec. 5.

2. CHIRP ANALYSIS

The detection and analysis of chirped signals in noise may be viewed as an extension of spectrum analysis to include a new parameter, the rate of change of frequency. We ask whether there exist principles for optimal data processing in such problems.

Chirped signals occur in many contexts; in quantum optics, (Jaynes, 1973; Nikolaus & Grischkowsky, 1983), the ionospheric "whistlers" of Helliwell (1965), human speech, the sounds of birds, bats, insects, and slide trombonists; radio altimeters, frequency–modulated radar, etc. Any transient signal, propagating through a dispersive medium, emerges as a chirped signal, as is observed in optics, ultrasonics, and oceanography; and presumably also in seismology, although the writer has not seen explicit mention of it. Thus in various fields the detection and/or analysis of chirped signals in noise is a potentially useful adjunct to spectrum analysis.

Quite aside from applications, the problem is a worthy intellectual challenge. Bats navigate skillfully in the dark, avoiding obstacles by a kind of acoustical chirped radar (Griffin, 1958). It appears that they can detect echoes routinely in conditions of signal/noise ratio where our best correlation filters would be helpless. How do they do it?

At first glance it seems that a chirped signal would be harder to detect than a monochromatic one. Why, then, did bats evolve the chirp technique? Is there some as yet unrecognized property of a chirped signal that makes it actually easier to detect than a monochromatic one?

Further evidence suggesting this is provided by the "Whistle language" developed by the Canary Islanders. We understand that by a system of chirped whistles they are able to communicate fairly detailed information between a mountaintop and the village below, in conditions of range (a few miles) and wind where the human voice would be useless.

In the case of the bats, one can conjecture three possible reasons for using chirp: (A) Since natural noise is usually generated by some "stationary" process, a weak chirped signal may resemble natural noise less than does a weak monochromatic signal. (B) Prior information about the chirp rate, possessed by the bat, may be essential; it helps to know what you are looking for. (C) Our whole conceptual outlook, based on years of non–chirp thinking, may be wrong; bats may simply be asking a smarter question than we are.

Of course, for both the bats and the Canary Islanders it may be that chirped signals are not actually easier to detect, only easier to recognize and interpret in strong noise.

After noting the existing "Spectral Snapshot" method of chirp analysis we return to our general Bayesian solution for a model that represents one possible real situation, then generalize it in various ways on the lookout for evidence for or against these conjectures.

In this problem we are already far beyond what Box and Tukey have called the "exploratory phase" of data analysis. We already know that the bats, airplanes, and Canary Islanders are there, that they are emitting purposeful signals which it would be ludicrous to call "random", and that those signals are being corrupted by additive noise that we are unable to control or predict. Yet we do have some cogent prior information about both the signals and the noise, and so our job is not to ask "What seems to go on here?" but rather to set up a model which expresses that prior information.

3. SPECTRAL SNAPSHOTS

A method of chirp analysis used at present, because it can be implemented with existing hardware, is to do a conventional Blackman–Tukey spectrum analysis of a run of data over an interval $(t_1 \pm T)$, then over a later interval $(t_2 \pm T)$, and so on. Any peak that appears to move steadily in this sequence of spectral snapshots is naturally interpreted as a chirped signal (the evanescent character of the phenomenon making the adjective "spectral" seem more appropriate than "spectrum").

The method does indeed work in some cases; and impressively in the feat of oceanographers (Barber & Ursell, 1948; Munk & Snodgrass, 1957) to correlate chirped ocean waves, with periods in the 10–20 second range, with storms thousands of miles away. Yet it is evident that spectral snapshots do not extract all the relevant information from the data; at the very least, evidence contained in correlations between data segments is lost.

The more serious and fundamental shortcoming of this method is that if one tries to analyze chirped data by algorithms appropriate to detect monochromatic signals, the chirped signal of interest will be weakened – possibly disastrously – through phase cancellation (the same mathematical phenomenon that physicists call Fresnel diffraction).

For this reason, if we adhere to conventional spectrum analysis algorithms, the cutting of the data into segments analyzed separately is not a correctible approximation. As shown in Appendix A, to detect a signal of chirp rate α – *i.e.* a sinusoid $\cos(\omega t + \alpha t^2)$ – with good sensitivity by that method, one must keep the data segments so short that $\alpha T^2 < 1$. If T is much longer than this, a large chirped signal – far above the noise level – can still be lost in the noise through phase cancellation. Appendix A also tries to correct some currently circulating misconceptions about the history of this method.

Our conclusion is that further progress beyond the spectral snapshot method is necessary and possible – and it must consist of finding new algorithms that (a) protect against phase cancellation; (b) extract more information from the data; and (c) make more use of prior information. But nobody's intuition has yet revealed the specific algorithm for this data analysis, so we turn for guidance to probability theory.

4. THE BASIC REASONING FORMAT

The principle we need is just the product rule of probability theory, $p(AB|C) = p(A|C)\,p(B|AC)$, which we note is symmetric in the propositions A and B. Therefore let

$$I = \text{prior information}, \qquad H = \text{any hypothesis to be tested}, \qquad D = \text{data}.$$

Then $p(HD|I) = p(D|I)\,p(H|DI) = p(H|I)\,p(D|HI)$ or, if $p(D|I) > 0$ (*i.e.* the data set is a possible one),

$$p(H|DI) = p(H|I)\,\frac{p(D|HI)}{p(D|I)} \qquad (0)$$

which is Bayes' theorem, showing how the prior probability $p(H|I)$ of H is updated to the posterior probability $p(H|DI)$ as a result of acquiring the new information D. Bayesian analysis consists of the repeated application of this rule.

Progress in scientific inference was held up for decades by a persistent belief – for which we can find no basis in the mathematics or the physical facts of the problem – that the equations of probability theory were only rules for calculating frequencies; and not for conducting inference. However, we now have many analyses (B. de Finetti, H. Jeffreys, R. T. Cox, A. Wald, L. J. Savage, D. V. Lindley, and others) showing from widely different viewpoints that these equations are also the uniquely "right" rules for conducting inference.

That is, it is a theorem that anyone who represents degrees of plausibility by real numbers – and then reasons in a way not reducible to these equations – is necessarily violating some very elementary qualitative desiderata of rationality (transitivity, strong domination, consistency, coherence, *etc.*). We are concerned simply with the logic of consistent plausible reasoning; there is no necessary connection with frequencies or random experiments (although of course, nothing forbids us to introduce such notions in a particular problem).

Put differently, sufficiently deep and careful intuitive thinking will, after all inconsistencies have been detected and removed, necessarily converge eventually to the Bayesian conclusions from the same information. Recognizing this only enables us to reach those conclusions more quickly – how much more quickly we shall see presently.

New demonstrations of the power of Bayesian inference in real problems – yielding in a few lines important results that decades of "frequentist" analysis or intuitive thinking had not found – have been appearing steadily for about twenty years, the present work providing another example.

However, before we can apply Eq. (0) quantitatively, our problem must have enough structure so that we can determine the term $p(D|HI)$. In its dependence on D for fixed H this is the "sampling distribution"; in its dependence on H for fixed D it is the "likelihood function". In the exploratory phase of a problem such structure may not be at hand.

Fortunately, our present problem is free of this difficulty. We shall apply (0) in which H stands typically for the statement that a multidimensional parameter lies in a certain specified region of the parameter space. Deploring, but nevertheless following, the present common custom, we use the same symbol $p(x|y)$ for a probability or a probability density; the distinction must be read from the context.

5. A SIMPLE BAYESIAN MODEL

In this section, which is possibly the first direct Bayesian analysis of the problem, attempts at conceptual innovation are out of order, and we wish only to learn the consequences of an absolutely standard kind of model. We follow slavishly the time–worn procedure of "assuming the data contaminated with additive white gaussian noise", which is time–worn just because it has so many merits. As has long been recognized, it is not only the most realistic choice one can make in most problems, the solutions will be analytically simple and not far from optimal in others.

But there is an even more cogent reason for choosing this probability assignment. In most real problems, the only prior information we have about the noise is its mean square value; often not even that. Then, as noted above, because it has maximum entropy for a given mean square noise level, the independent "white" Gaussian distribution will be the safest, most conservative one we can use; *i.e.* it protects us most strongly from drawing erroneous conclusions. In other words, it most honestly describes our *state of knowledge*. But since there are still many people who simply do not believe this in spite of all demonstrations, let us amplify the point.

The purpose of our assigning a prior distribution to the noise is to define the range of possible variations of the noise vector $e = \{e_1, \cdots, e_n\}$ – that we shall make allowance for in our inference. As is well known in the literature of Information Theory, the entropy of a distribution is an asymptotic measure of the size of the basic "support set" W of that distribution; in our case the n–dimensional "volume" occupied by the reasonably probable noise vectors. The MAXENT principle tells us – as does elementary common sense – that the distribution which most honestly represents what we know while avoiding assuming things that we do not know, is the one with the largest support set W_{max} permitted by our information.

Then what Bayesian inference does is that any data that lie well within W are judged to be almost certainly noise artifacts and are ignored; while data that lie well outside W are judged to be almost certainly real effects and are emphasized. There is necessarily a transition region, where the data lie near the boundary of W, where no sure conclusions can be drawn. Therefore, unless we have very specific prior information in addition to the mean square value, so that we know some particular way in which the noise departs from white Gaussian, it would be dangerous to use any other distribution in our inference. According to the MAXENT principle, to do so would necessarily be making an assumption, that restricts our considerations to some arbitrary subset $W \subset W_{max}$, in a way not justified by our prior information about the noise.

The price we would pay for this indiscretion is that if the true noise vector happened to lie in the complementary set $W' = W_{max} - W$, then we would be misled into interpreting what is only an artifact of the noise, as a real effect. And the chance of this happening is not small, as shown by the Entropy Concentration Theorem (Jaynes, 1982). To assign a distribution with entropy only slightly smaller than the maximum may contract the volume of W by an enormous factor, often more than 10^{10}. Then virtually every possible noise vector would lie in W', and we would be seeing things that are not there in almost every data set. Gratuitous assumptions in assigning a noise probability distribution can be very costly.

Indeed, unless we can identify and correctly understand some specific defect in this simplest independent Gaussian model, we are hardly in a position to invent a better one.

However, these arguments apply only to the noise, because the noise is completely unknown except for its mean square value. The signal of interest is of course something about which we know a great deal in advance (just the reason why it *is* of interest). It would be ludicrous to think of the signal as a sample from an "independent Gaussian distribution". This would amount to throwing away all the prior information we have about its structure (functional form), thus losing our best means of finding it in the noise.

Our only seemingly drastic simplification is that for the time being we suppose it known in advance that only a single signal can be present, of the form

$$f(t) = A\cos(\omega t + \alpha t^2 + \theta)\,, \tag{1}$$

and so the problem reduces to estimating its five parameters $(A, \omega, \alpha, \theta, \sigma)$ [although often we are interested only in (ω, α)]. However, this is about the most realistic assumption that could be made for the Barber–Ursell–Munk–Snodgrass oceanographic chirp problem discussed in Appendix A, where it is highly unlikely that two different signals would be present simultaneously. It is considerably more realistic than that supposition – which has actually been made in this problem – that the chirped signal is a "sample from a Gaussian random process".

In the end it will develop that for purposes of estimating the power spectrum, this assumption of a single signal is hardly an assumption at all; the resulting solution remains valid, with only a slight reinterpretation and no change in the actual algorithm, however many signals may be present. It is a restrictive assumption only when we ask more detailed questions than "What is your best estimate of the power spectrum?"

Other assumptions (constant amplitude and chirp rate, *etc.*) turn out also to be removable. In fact, once we understand the solution to this simplest problem, it will be evident that it can be generalized at once to optimal detection of any signal of known functional–parametric form, sampled at arbitrary times, in non–white noise.

But for the present our true signal (1) is contaminated with the aforementioned white gaussian noise $e(t)$, so the observable data are values of the function

$$y(t) = f(t) + e(t) \ . \tag{2}$$

In practice we shall have these data only at discrete times which we suppose for the moment to be equally spaced at integer values of t; and over a finite interval $2T$. Thus our data consist of $N = (2T + 1)$ values

$$D = \{y(t), \ -T \le t \le T\} \ , \tag{3}$$

and we assign the aforementioned independent gaussian joint probability distribution for the values of the noise $e(t)$ at the corresponding times, taking each $e(t) \sim N(0, \sigma)$ where the variance σ^2 is supposed known.

We have gone to some length to explain our basis for choosing this noise distribution because it is a matter of much confusion, different schools of thought holding diametrically opposite views as to whether this is or is not a restrictive assumption. In fact, this confusion is so great that the rationale of our choice still requires further discussion, continued in Appendix B.

Whatever school of thought one favors, our equations will be just the same; only our judgments of their range of validity will differ. Given any true signal $f(t)$, the probability (density) that we shall obtain the data set $D = \{y_i(t)\}$ is just the probability that the noise values $e(t)$ will make up the difference:

$$p(D|A, \omega, \alpha, \theta, \sigma) = \prod_{t=-T}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[y(t) - f(t)\right]^2\right\} \tag{4}$$

which is our sampling distribution. Conversely, given σ and the data D, the joint likelihood of the unknown parameters is

$$L(A, \omega, \alpha.\theta) \propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=-T}^{T} [y(t) - A\cos(\omega t + \alpha t^2 + \theta)]^2\right\} \tag{5}$$

In analysis of discrete time series, the mathematics has tended to get cluttered with minute details about "end effects" associated with the exact limits of summation. But this is only a notational problem; we can remove the clutter with no loss of precision if we adopt the convention (call it "infinite padding with zeroes" if you like):

$$y_t = y(t) = 0, \qquad |t| \ge T \ . \tag{6}$$

Then all our sums of functions $K(y_t)$ over time indices can take the form ΣK_t, understood to run over $(-\infty < t < \infty)$. In this notation, which we use henceforth, all those little details are present automatically, but kept out of sight.

Usually, the absolute phase θ is of no interest to us and we have no prior information about it; *i.e.* it is a "nuisance parameter" that we want to eliminate. We may integrate it out with respect to a uniform prior probability density, getting a marginal quasi–likelihood

$$L(A, \omega, \alpha) = \frac{1}{2\pi} \int_0^{2\pi} L(A, \omega, \alpha, \theta) d\theta \tag{7}$$

which represents the contribution from the data to the joint marginal posterior distribution of (A, ω, α). This is the course we shall pursue in the present work.

But an important exception occurs if our ultimate objective is not to estimate the parameters (A, ω, α) of the "regular" signal $f(t)$, but rather to estimate the particular "irregular" noise sequence $e(t)$ that occurred during our observation period. This is the problem of seasonal adjustment, where an estimate of θ is also needed, and our data processing algorithm will be quite different. The Bayesian theory of seasonal adjustment, to be given elsewhere[†] yields a new demonstration of the power of prior information to improve our estimates; we invite intuitionists to discover it without Bayesian methods.

6. THE PHASELESS LIKELIHOOD FUNCTION

We shall consider the exact relations later when we generalize to many signals,[‡] but for the moment we make an approximation which we believe to be generally accurate and harmless (although with obvious exceptions like $\omega = \alpha = 0$, or $\omega = \pi$, $\alpha = 0$:

$$\sum_t \cos^2(\omega t + \alpha t^2 + \theta) \simeq (2T + 1)/2 = N/2 \ . \tag{8}$$

Values of α differing by 2π are indistinguishable in this discrete sampled data; *i.e.* chirp aliasing, like frequency aliasing, confines us to the domain $(-\pi < \alpha \leq \pi)$.

With σ considered known, the joint likelihood of the four signal parameters is then

$$L(A, \omega, \alpha, \theta) = \exp\left\{ \frac{A}{\sigma^2} \sum_t y_t \, \cos(\omega t + \alpha t^2 + \theta) - \frac{NA^2}{4\sigma^2} \right\} \tag{9}$$

in which, since only the dependence on $(A, \omega, \alpha, \theta)$ matters, we may discard any factor not containing them.

The integration (7) over θ, carried out in Appendix C, yields the phaseless likelihood (or quasi–likelihood) function

$$L(A, \omega, \alpha) = \exp\left(-\frac{NA^2}{4\sigma^2} \right) I_0 \left(\frac{A\sqrt{NC(\omega, \alpha)}}{\sigma^2} \right) \tag{10}$$

where $I_0(x) = -iJ_0(ix)$ is a Bessel function[⋆] and

$$C(\omega, \alpha) \equiv N^{-1} \sum_{ts} y_t y_s \cos[\omega(t - s) + \alpha(t^2 - s^2)] \ . \tag{11}$$

[†] This appeared later, in Bernardo, *et al* (1985), pp. 329–360.

[‡] Given in detail in Bretthorst (1988).

[⋆] Bretthorst (1988) discovered that we avoid Bessel functions by considering instead the quadrature components $\cos(\omega t + \alpha t^2)$ and $\sin(\omega t + \alpha t^2)$ with independent amplitudes A_1, A_2. This amounts to a slightly different treatment of the prior information about amplitudes, in which the parameters A_1, A_2 take the place of our A, θ. In effect, our analysis assigns equal prior probability to equal ranges of A; Bretthorst's assigns equal probability to equal areas in the (A_1, A_2) plane. But by a well–known rule of thumb, this has about the same effect on the final conclusions as would having one less data point; he shows that his final conclusions are numerically indistinguishable from the ones obtained here.

The form of (10) already provides some (at least to the writer) unexpected insight. Given any sampling distribution, a likelihood or quasi–likelihood function, in its dependence on the parameters, contains all the information in the data that is relevant for any inference about those parameters – whether it be joint or individual point estimation, interval estimation, testing any hypotheses concerning them, *etc*. But the only data dependence in $L(A, \omega, \alpha)$ comes from the function $C(\omega, \alpha)$. Therefore, C plays the role of a "sufficient statistic"; this function summarizes all the information in the data that is relevant for inference about (A, ω, α).

Because of its fundamental importance, $C(\omega, \alpha)$ should be given a name. We shall call it the *chirpogram* of the data, for a reason that will appear in Eq. (13) below. It seems, then, that whatever specific question we seek to answer about a chirped signal, the first step of data analysis will be to determine the chirpogram of the data.

Of course, to answer a recent criticism (Tukey, 1984), in setting up a model the Bayesian – like any other theoretician – is only formulating a working hypothesis, to find out what its consequences would be. He has not thereby taken a vow of theological commitment to believe it forever in the face of all new evidence. Having got this far in our calculation, nothing in Bayesian principles forbids us to scan that chirpogram by eye, on the lookout for any unusual features (such as a peak stretched out diagonally, which suggests a change in chirp rate during the data run) that we had not anticipated when setting up our model.[†]

Indeed, we consider the need for such elementary precautions so obvious and trivial that it would never occur to us that anyone could fail to see it. If we do not stress this constantly in our writings, it is because we have more substantive things to say.

7. DISCUSSION – MEANING OF THE CHIRPOGRAM

The chirpogram appears less strange if we note that when $\alpha = 0$ it reduces to

$$C(\omega, 0) = N^{-1} \sum_{ts} y_t y_s \cos\omega(t-s) = N^{-1} \Big| \sum_t y_t e^{i\omega t} \Big|^2 = \sum_t R(t) \, \cos\omega t \tag{13}$$

where $R(t)$ is the data autocovariance:

$$R(t) = N^{-1} \sum_s y_s y_{s+t} \, . \tag{14}$$

Thus $C(\omega, 0)$ is just the periodogram of Schuster (1897).

For nearly a Century, therefore, the calculation of $C(\omega, 0)$ has seemed, intuitively, the thing to do in analyzing a stationary power spectrum. However, the results were never satisfactory. At first one tried to interpret $C(\omega, 0)$ as an estimate of the power spectrum of the sampled signal. But the periodograms of real data appeared to the eye as too wiggly to believe, and in some problems the details of those wiggles varied erratically from one data set to another.

Intuition then suggested that some kind of smoothing of the wiggles is called for. Blackman and Tukey (1958; hereafter denoted B–T) recognized the wiggles as in part spurious side–lobes, in part beating between "outliers" in the data, and showed that in some cases one can make an estimated power spectrum with a more pleasing appearance, which one therefore feels has more

[†] Likewise, in calculating within a given problem, the orthodoxian does not question the correctness of his sampling distribution, and we do not criticize him for this. But it is for him, as for us, only a tentative working hypothesis; having finished the calculation, his unhappiness at the result may lead him to consider a different sampling distribution. The Bayesian has an equal right to do this.

truth in it, by introducing a lag window function $W(t)$, which cuts off the contributions of large t in (13), giving the B–T spectrum estimate

$$\hat{P}(\omega)_{BT} = \sum_{t=-m}^{m} W(t) R(t) \cos\omega t \,, \tag{15}$$

in which we use only autocovariances determined from the data up to some lag m which may be a small fraction of the record length N.

That this estimate disagrees with the data [the measured autocovariance $R(t)$] at every lag t for which $W(t) \neq 1$, does not seem to have troubled anyone until Burg pointed it out 17 years later. He termed it a "willingness to falsify" the data, and advocated instead the Maximum Entropy estimate, which was forced by the constraints to agree with the data, the wiggles being removed by a totally different method, the "smoothest" extrapolation of $R(t)$ beyond the data. Others, including this writer, quickly echoed Burg's argument with enthusiasm; but as we shall see presently, this was not the end of the story.

In any event, a lag window $W(t)$ smoothly tapered to zero at $t = m$ does reduce the unwanted wiggles – at a price. It leads to a reasonable estimate in the case of a broad, featureless spectrum; but of course, in that case the contributions from large t were small anyway, and the window had little effect. But lag window smoothing (equivalent to a linear filtering operation that convolves the periodogram with the fourier transform of the lag window function, thus smearing out the wiggles sideways) necessarily loses resolution and makes it impossible to represent sharp spectrum lines correctly.

One wonders, then, why B–T stopped at this point; for other procedures were available. To put it with a cynicism that we shall correct later: once one has been willing to falsify the data in one way, then his virtue is lost, and he should have no objection to falsifying them also in other ways. Why, then, must we use the same window function at all frequencies? Why must we process $R(t)$ linearly? Why must we set all $R(t)$ beyond m equal to zero, when we know that this is surely wrong? There were dozens of other *ad hoc* procedures, which would have corrected the failure of (15) to deal with sharp lines without venturing into the forbidden realm of Bayesianity.

But history proceeded otherwise; and now finally, the first step of a Bayesian analysis has told us what a Century of intuitive *ad hockery* did not. The periodogram was introduced previously only as an intuitive spectrum estimate; but now that it has been derived from the principles of probability theory we see it in a very different light. Schuster's periodogram is indeed fundamental to spectrum analysis; but not because it is itself a satisfactory spectrum estimator, nor because any linear smoothing can convert it into one in our problem.

The importance of the periodogram lies rather in its information content; in the presence of white gaussian noise, it conveys all the *information* the data have to offer about the spectrum of $f(t)$. As noted, the chirpogram has the same property in our more general problem. This makes it clear that any smoothing or other tampering with the periodogram can only hurt, destroying some of the information in the data.

It will follow from (10) (Eq. 27 below), that the proper algorithm to convert $C(\omega, 0)$ into a power spectrum estimate is a nonlinear operation much like exponentiation followed by renormalization, an approximation being:[‡]

$$\hat{P}(\omega) \simeq A \exp\left[\frac{C(\omega,0)}{\sigma^2}\right]$$

[‡] In the slightly different treatment of Bretthorst (1988) this was found to be an exact relation.

This will suppress those spurious wiggles at the bottom of the periodogram as well as did the B–T linear smoothing; but it will do it by attenuation rather than smearing, and will therefore not lose any resolution. The Bayesian nonlinear processing of $C(\omega, 0)$ will also yield, when the data give evidence for them, arbitrarily sharp spectral line peaks from the top of the periodogram, that linear smoothing cannot give.

It is clear from (10) why a nonlinear processing of C is needed. The likelihood involves not just C, but C in comparison with the noise level. The procedure (15) smears out all parts of C equally, without considering where they stand relative to any noise. The Bayesian nonlinear processing takes the noise level into account; wiggles below the noise level are almost certainly artifacts of the noise and are suppressed, while peaks that rise above the noise level are believed and emphasized.

It may seem at this point surprising that intuition did not see the need for this long ago. Note, however, that Blackman and Tukey had in mind a very different problem than ours. For them the whole data $y(t)$ were a sample from a "stochastic process" with a multivariate Gaussian distribution. From the standpoint of our present problem we might interpret the B–T work as a preliminary study of the noise spectrum before the purposeful signal was added. So for them the notion of "C in comparison with the noise level" did not exist.

To emphasize this, note that B–T considered the periodogram to have a sampling distribution that was Chi–squared with two degrees of freedom, independently of the sample size. That would not be the case in the problem we are studying unless the signal $f(t)$ were absent.

From this observation there follows a point that has not been sufficiently stressed – or even noticed – in the literature: the B–T efforts were not directed at all toward the present problem of estimating the power spectrum of a signal $f(t)$ from data which are contaminated with noise.

Confusion over "What is the problem?" has been rampant here. We conceded, after a theoretical study (Jaynes, 1982) that pure Maximum Entropy is not optimal for estimating the spectrum of a signal in the in the presence of noise; but failed to see the point just noted. Immediately, Tukey & Brillinger (1982) proceeded to stress the *extreme* importance of noise in real problems and the necessity of taking it into account. But they failed to note that the B–T method does not take noise into account either, much less the robustness of Maximum Entropy with respect to noise (*i.e.*, its practical success in problems where noise is present). Finally, although one would expect Tukey to be the first to emphasize it, they did not note that a given procedure may solve more than one problem. In fact, the Maximum Entropy procedure is also the optimal solution to a Gaussian problem.

Therefore, although we agree with the need to take noise into account (as the present work demonstrates), we can hardly see that as an argument in favor of B–T methods in preference to Maximum Entropy in any problem. For we must distinguish between the B–T *problem* (spectrum of Gaussian noise) and the B–T *procedure* (15), which has not been derived from, or shown to have any logical connection at all to, that problem.*

Indeed, Burg's original derivation of the Maximum Entropy algorithm started from just that same B–T assumption that the data are Gaussian noise; and we showed (Jaynes, 1982) that pure Maximum Entropy from autocovariance data leads automatically to a Gaussian predictive distribution for future data. Thus it appears that the optimal solution to the B–T problem is not the B–T tapering procedure (15), but the Burg procedure!

But strangely enough, this enables us to take a more kindly view toward B–T methods. The procedure (15) cannot be claimed as the "best" one for any spectrum analysis problem; yet it has a

* This is one of the many difficulties with intuitive *ad hoc* procedures; not only are they invariably faulty on pragmatic grounds, whatever theoretical justification they may have lies hidden in the mind of the inventor. So they cannot be improved; only replaced.

place in our toolbox. As a *procedure* it is applicable to any data, and is dependent on no hypotheses of Gaussianity. Whatever the phenomenon, if nothing is known in advance about its spectrum, the tapering (15) is a quick and easy way to wash out the wiggles enough to let the eye get a preliminary view of the broad features of the spectrum, helpful in deciding whether a more sophisticated data analysis is called for. The most skilled precision machinist still has occasional use for a quick and dirty jackknife; and if in fact there are no sharp spectrum lines, the B–T procedure will lead to substantially the same results with less computation.

Nevertheless, any tapering clearly does falsify the data, throwing away usable information. But data contaminated with noise are themselves in part "false". The valid criticism from the standpoint of our present problem is that when the noise goes away the falsification in (15) remains; that is the case where Burg's pure Maximum Entropy solution was clearly the optimal one. But in the different problem envisaged by B–T, which led to (15), the noise cannot go away because the noise and the data are identical.

Thus from inspection of (10) we can see already some clarification of these muddy waters. The Bayesian method is going to give us, for spectrum estimation of a signal in the presence of noise, the same kind of improvement in resolution and removal of spurious features (relative to the periodogram) that the Maximum–Entropy formalism did in the absence of noise; and it will do this as well for chirped or non–chirped signals and for arbitrarily sharp lines. There is no meaningful comparison with B–T methods at all; for they do not address the same problem.

Realizing these things changed the direction of the present work. We started with the intention only of getting a quick, preliminary glimpse at what Bayesian theory has to say about monochromatic spectrum analysis in the presence of noise, before proceeding to entropy considerations. But as soon as the phaseless likelihood function (10) appeared, it was realized that (a) The status of B–T methods in this problem is very different from what we and others had supposed; (b) monochromatic spectrum estimation from noisy data must be radically revised by these results; and (c) given that revision, the extension to chirp is almost trivial in a Bayesian analysis (although impossible in an AR analysis).

Therefore, we now reconsider at some length the "old" problem of conventional pre–entropy spectrum estimation of a signal in noise, from this "new" viewpoint.

8. POWER SPECTRUM ESTIMATES

The terms "power spectrum" and "power spectrum estimator" can be defined in various ways. Also, we need to distinguish between the quite different goals of estimating a power spectral density, estimating the power in a spectrum line, and estimating the frequencies present.

In calling $\hat{P}(\omega)$ an estimate of the power spectral density we mean that

$$\int_a^b \hat{P}(\omega) d\omega \tag{16}$$

is the expectation, over the joint posterior distribution for all the unknown parameters, of the energy carried by the signal $f(t)$, (not the noise), in the frequency band $(a \leq \omega \leq b)$, in the observation time $N = 2T + 1$. The true total energy carried by the signal (1) is $NA^2/2$, and given data D we can write its expectation as

$$(N/2)E(A^2|D,I) = (N/2)\int_{-\pi}^{\pi} d\omega \int_0^{\infty} dA A^2 \int_{-\pi}^{\pi} d\alpha \, p(A,\omega,\alpha|D,I)\,. \tag{17}$$

For formal reasons it is convenient to define our spectrum as extending over both positive and negative frequencies; thus (17) should be equated to the integral (16) over $(-\pi,\pi)$. Therefore our power spectrum estimate is

$$\hat{P}(\omega) = (N/2) \int dA A^2 \int_{-\pi}^{\pi} d\omega \, p(A, \omega, \alpha|D, I) , \qquad (-\pi < \omega < \pi) \tag{18}$$

To define a power spectrum only over positive frequencies (as would be done experimentally), one should take instead $\hat{P}_+(\omega) = 2\hat{P}(\omega)$, $0 \le \omega < \pi$.

In (17), (18) $p(A, \omega, \alpha|D, I)$ is the joint posterior distribution

$$p(A, \omega, \alpha|D, I) = B \, p(A, \omega, \alpha|I) \, L(A, \omega, \alpha) \tag{19}$$

in which B is a normalization constant, I stands for prior information, and $p(A, \omega, \alpha|I)$ is the joint prior probability density function for the three parameters.

If we have any prior information about our parameters, this is the place to put it into our equations, in the prior probability factors; if we do not, then a noninformative, flat prior will express that fact and leave the decision to the evidence of the data in the likelihood function (thus realizing R. A. Fisher's goal of "letting the data speak for themselves").

Of course, it is not required that we actually have, or believe, a piece of prior information I or data D before putting it into these equations; we may wish to find out what would be the consequences of having or not having a certain kind of prior information or data, in order to decide whether it is worth the effort to get it.

Therefore, by considering various kinds of prior information and data in (18) we can analyze a multitude of different special situations, real or hypothetical, of which we can indicate only a few in the present study.

First, let us note the form of traditional spectrum analysis, which did not contemplate the existence of chirp, that is contained in this "formalism". In many situations we know in advance that our signals are not chirped, *i.e.* the prior probability in (19) is concentrated at $\alpha = 0$. Our equations will then be the same as if we had never introduced chirp at all; *i.e.* failure to note its possible existence is equivalent to asserting (unwittingly) the prior information: "$\alpha = 0$: there is no chirp".

[We add parenthetically that much of the confusion in statistics is caused by this phenomenon. In other areas of applied mathematics, failure to notice all the possibilities means only that not all possibilities will be studied. In probability theory, failure to notice some possibilities is mathematically equivalent to making unconscious – and often very strong – assumptions about them].

Setting $\alpha = 0$ in (10) gives the joint likelihood $L(A, \omega)$. Not using prior information (*i.e.* using flat prior densities for A, ω) our spectrum estimator is then

$$\hat{P}(\omega) = \frac{(N/2) \int_0^\infty dA \, A^2 \, L(A, \omega)}{\int_{-\pi}^{\pi} d\omega \int_0^\infty dA L(A, \omega)} \tag{20}$$

Note that this can be factored:

$$\hat{P}(\omega) = \left\{ \frac{(N/2) \int_0^\infty dA A^2 \, L(A, \omega)}{\int_0^\infty dA L(A, \omega)} \right\} \cdot \left\{ \frac{\int_0^\infty dA L(A, \omega)}{\int_{-\pi}^{\pi} d\omega \int_0^\infty dA L(A, \omega)} \right\} \tag{21}$$

or,

$$\hat{P}(\omega) = (N/2) E(A^2|\omega D I) \cdot p(\omega|DI) . \tag{22}$$

In words: $\hat{P}(\omega) =$ (conditional expectation of energy given ω) × (posterior probability density for ω given the data). Both factors may be of special interest so we evaluate them separately. In Appendix C we derive

$$(N/2)E(A^2|\omega DI) = \sigma^2 \left[1 + 2q + 2q\,\frac{I_1(q)}{I_0(q)}\right] \tag{23}$$

where

$$q(\omega) \equiv C(\omega, 0)/2\sigma^2 \, . \tag{24}$$

This has two limiting forms:

$$(N/2)E(A^2|DI) \rightarrow \left\{ \begin{array}{ll} 2C(\omega,0), & C(\omega,0) > 2\sigma^2 \\ \sigma^2 + C(\omega,0), & C(\omega,0) << \sigma^2 \end{array} \right\} . \tag{25}$$

The second case is unlikely to arise, because it would mean that the signal and noise have, by chance, nearly cancelled each other out. But if this should happen, Bayes' theorem recognizes automatically that the reasonable estimate of signal energy is no longer $2C(\omega, 0)$ but much greater; for now the mean square signal strength must be close to σ^2. This is something that intuition would be extremely unlikely to notice.

From Appendix C the second factor in (21) is

$$p(\omega|DI) = \frac{\exp(q) I_0(q)}{\int_{-\pi}^{\pi} \exp(q) I_0(q)\, d\omega} \tag{26}$$

This answers the question: "Conditional on the data, what is the probability that the signal frequency lies in the range $(\omega, \omega + d\omega)$ irrespective of its amplitude?" In many situations it is this question – and not the power spectrum density – that matters, for the signal amplitude may have been corrupted en route by all kinds of irrelevant circumstances and the frequency alone may be of interest.

Combining (23) and (26) we have the explicit Bayesian power spectrum estimate

$$P(\omega) = \sigma^2\, \frac{[(1 + 2q)I_0(q) + 2qI_1(q)] \exp(q)}{\int_{-\pi}^{\pi} \exp(q)\; I_0(q) d\omega} \tag{27}$$

A graph of the nonlinear processing function

$$f(q) \equiv [(1 + 2q)I_0(q) + 2qI_1(q)] \exp(q) \tag{28}$$

shows it close to the asymptotic form $(8/\pi q)^{1/2} \exp(2q)$ over most of the range likely to appear; in most cases we would not make a bad approximation if we replaced (27) by

$$\hat{P}(\omega) = \sigma^2\, \frac{4q \exp(2q)}{\int_{-\pi}^{\pi} \exp(2q) d\omega} \tag{29}$$

Whenever the data give evidence for a signal well above the noise level [*i.e.* $C(\omega, 0)$ reaches a global maximum $C_{max} = C(\hat{\omega}, 0) >> \sigma^2$], then $q_{max} = C_{max}/\sigma^2 >> 1$ and most of the contribution to the integral in the denominator of (27) will come from the neighborhood of this greatest peak. Expanding

$$q(\omega) = q_{max} - q''\,(\omega - \nu)^2/2 + \cdots \tag{30}$$

and using saddle–point integration we have then

$$\int_{-\pi}^{\pi} e^q I_0(q) d\omega \simeq 2(q_{max} q'')^{-1/2} \exp(2q_{max}) \tag{31}$$

the factor of 2 coming from the equal peak at $(-\nu)$. Near the greatest peak the positive frequency estimator reduces to

$$\hat{P}_+(\omega) \simeq 2C_{max} \frac{1}{\sqrt{2\pi(\delta\omega)^2}} \exp\left[-\frac{(\omega-\nu)^2}{2(\delta\omega)^2}\right] \tag{32}$$

where $\delta\omega = (q'')^{-1/2}$ is the accuracy with which the frequency can be estimated. Comparing with the factorization (22) we see that the Gaussian is the posterior distribution for ω, while the first factor, the estimated total energy in the line, is:

$$\int_0^\infty \hat{P}_+(\omega) d\omega = 2C_{max} \, . \tag{33}$$

in agreement with (25).

As a further check, suppose we have a pure sinusoid with very little noise:

$$y_t = A\cos\nu t + e_t, \qquad A >> \sigma \tag{34}$$

Then the autocovariance (14) is approximately

$$R(t) \simeq (A^2/2)\cos\nu t \tag{35}$$

and so from (13)

$$C_{max} = C(\nu, 0) = (A^2/2)\sum_t \cos^2 \nu t = NA^2/4 \tag{36}$$

so $2C_{max} = NA^2/2$ is indeed the correct total energy carried by the signal in the observation time. Likewise,

$$\sigma^2 q'' = \sum_t t^2 R(t)\cos\nu t \simeq N^3\, A^2/48 \tag{37}$$

which gives the width

$$\delta\omega \simeq \frac{\sigma}{n}\sqrt{\frac{12}{C_{max}}} \tag{38}$$

These relations illustrate what we stressed above; in this problem the periodogram $C(\omega, 0)$ is not even qualitatively an estimate of the power spectral density. Rather, when C reaches a peak above the noise level, thus indicating the probable presence of a spectrum line, its peak value $2C_{max}$ is an estimate of the total energy in the line.

9. EXTENSION TO CHIRP

Let $P(\omega, \alpha) d\omega d\alpha$ be the expectation of energy carried by the signal $f(t)$ in an element $d\omega d\alpha$ of the frequency–chirp plane. From (14) the frequency–chirp density estimator which does not use any prior information about (A, ω, α) is simply

$$\hat{P}(\omega, \alpha) = \frac{f(q)}{\int d\omega \int d\alpha e^q\, I_0(q)} \tag{39}$$

where $f(q)$ is the same nonlinear processing function (28), except that now in place of (24) we have

$$q \equiv q(\omega, \alpha) = C(\omega, \alpha)/2\sigma^2 \,. \tag{40}$$

As in (27), any peaks of $C(\omega, \alpha)$ that rise above the noise level will be strongly emphasized, indicating high probability of a signal.

If we have indeed no prior information about the frequencies and chirp rates to be expected, but need to be ready for all contingencies, then there seems to be no way of avoiding the computation in determining $C(\omega, \alpha)$ over the entire plane $(-\pi < \omega, \alpha < \pi)$. Note that, while $\hat{P}(\omega, 0) = \hat{P}(-\omega, 0)$ by symmetry, when chirp is present we have inversion symmetry, $\hat{P}(\omega, \alpha) = \hat{P}(-\omega, -\alpha)$ so half of the $(\omega - \alpha)$ plane needs to be searched if no prior information about the signal location is at hand.[†]

But noting this shows how much reduction in computation can be had with suitable prior information. That bat, knowing in advance that it has emitted a signal of parameters ω_0, α_0, and knowing also what frequency interval (determined by its flight speed v and the range of important targets) is of interest, does not need to scan the whole plane. It needs only to scan a portion of the line $\alpha \simeq \alpha_0$ extending from about $\omega_0(1 - v/c)$ to $\omega_0(1 + 4v/c)$, where c = velocity of sound, in order cover the important contingencies (a target that is approaching is a potential collision, one dead ahead approaching at the flight speed is a stationary object, a potential landing site, one moving away rapidly is uninteresting, one moving away slowly may be a moth, a potential meal).

10. MANY SIGNALS

In the above we made what seems a strong assumption, that only one signal $f(t) = A\cos(\omega t + \alpha t^2 + \theta)$ is present, and all our results were inferences as to where in the parameter space of (A, ω, α) that one signal might be. This is realistic in some problems (*i.e.* oceanographic chirp, or only one bat is in our cage, *etc.*), but not in most. In what way has this assumption affected our final results?

Suppose for the moment that the chirp rate $\alpha = 0$. Then the power spectrum estimate $\hat{P}(\omega)d\omega$ in (27) represents, as we noted, the answer to:

> *Question A*: What is your "best" estimate (expectation) of the energy carried by the signal $f(t)$ in the frequency band $d\omega$, in the interval of observation?

But had we asked a different question, such as:

> *Question B*: What is your estimate of the product of the energy transports in the two non–overlapping frequency bands $(a < \omega < b)$ and $(c < \omega < d)$?

the answer would be zero; a single fixed frequency ω cannot be in two different bands simultaneously. If our prior information I tells us that only one frequency can be present, then the joint posterior probability $p(E_{ab}, E_{cd}|D, I)$ of the events

$$E_{ab} \equiv (a < \omega < b)\,, \qquad E_{cd} \equiv (c < \omega < d)$$

is zero; one with orthodox indoctrination might be tempted to say that the "fluctuations" in non–overlapping frequency bands must be perfectly negatively correlated in our posterior distribution. Of course, in the present problem there are no such fluctuations; they are creations only of the Mind

[†] Bretthorst (1988) gives such a full chirpogram, computer generated as contour lines in the (ω, α) plane, for Wolf's famous sunspot data, long recognized as having a prominent spectrum peak at about 11 (years)$^{-1}$. But the chirpogram has its main peak off the ω–axis, indicating that a slightly chirped signal would fit Wolf's data considerably better than does any monochromatic signal.

Projection Fallacy which presupposes that probabilities are physically real things. The greater adaptability of Bayesian analysis arises in part from our recognition that probabilities are only indications of our own incomplete information; therefore we are free to change them whenever our information changes.

Now if two frequencies can be present, it turns out that the answer to question (A) will be essentially the same. But the answer to question (B) – or any other question that involves the joint posterior probabilities in two different frequency bands – will be different; for then it is possible for power to be in two different bands simultaneously and it is not obvious without calculation whether the correlations in energy transport in different bands is positive or negative.

If now we suppose that three signals may be present, the answers to questions (A) and (B) will not be affected; it makes a difference only when we ask a still more complicated question, involving joint probabilities for three different frequency bands; for example,

> *Question C*: What is your estimate of the power carried in the frequency band $(a < \omega < b)$, given the powers carried in $(c < \omega < d)$ and $(e < \omega < f)$?

and so on! In conventional spectrum estimation one asks only question (A); and for that our one–signal solution (27) requires no change, however many signals may be present. Our seemingly unrealistic assumption makes a difference only when we ask more complicated questions.

In this Section we prove these statements, or at least give the general solution from which they can be proved. For the moment we leave out the chirp; it is now clear how we can add it easily to the final results. The signal to be analyzed is a superposition of n signals like (1):

$$f(t) = \sum_{m=1}^{n} A_m \cos(\omega_m t + \theta_m) \tag{41}$$

sampled at instants $(t_1 \cdots t_N)$, which need not be uniformly spaced (our previous equally spaced version is recovered if we make the particular choice $t_m = -T + m - 1$). Using the notation

$$f_m \equiv f(t_m) \tag{42}$$

the joint likelihood for the parameters is now

$$L(A_i, \omega_i, \theta_i, \sigma) = \sigma^{-N} \exp\left[-\frac{1}{2\sigma^2} \sum_{m=1}^{N} (y_m - f_m)^2\right] = \sigma^{-N} \exp\left[-\frac{N}{2\sigma^2}(\overline{y^2} + Q)\right] \tag{43}$$

with the quadratic form

$$Q(A_i, \omega_i, \theta_i) = \overline{f^2} - 2\,\overline{yf} \tag{44}$$

in which, as always, we use overbars to denote averages over the data:

$$\overline{y^2} = N^{-1} \sum_{m=1}^{N} y_m^2 \tag{45}$$

$$\overline{yf} = N^{-1} \sum_{m} y_m f_m \tag{46}$$

$$\overline{f^2} = N^{-1} \sum_{m} f_m^2 \,. \tag{47}$$

To rewrite Q as an explicit function of the parameters, define the function

$$x(\omega) \equiv N^{-1} \sum_{m=1}^{N} y_m \exp(i\omega t_m) \tag{48}$$

which is the "complex square root of the periodogram", its projections

$$d_j \equiv Re\left[x(\omega_j)\, e^{i\theta_j}\right] = N^{-1}\sqrt{NC(\omega_j, 0)}\cos(\theta_j + \psi_j)\,, \quad 1 \le j \le n \tag{49}$$

where we have used Appendix C to write it in terms of the periodogram; and the matrix

$$M_{jk} \equiv N^{-1} \sum_{m=1}^{N} \cos(\omega_j t_m + \theta_j)\, \cos(\omega_k t_m + \theta_k)\,, \qquad 1 \le j, k \le n \tag{50}$$

Then

$$\overline{yf} = \sum_{j=1}^{n} d_j A_j \tag{51}$$

$$\overline{f^2} = \sum_{j,k=1}^{n} M_{jk} A_j\, A_k \tag{52}$$

and the quadratic form (44) is

$$Q = \sum_{jk} M_{jk} A_j A_k - 2 \sum_{j} d_j A_j \tag{53}$$

The likelihood (43) will factor in the desired way if we complete the square in Q. Define the quantities $(\hat{A}_1 \cdots \hat{A}_n)$ by

$$d_j = \sum_{k=1}^{n} M_{jk} \hat{A}_k\,, \qquad 1 \le j \le n \tag{54}$$

and note that, if the inverse matrix M^{-1} exists,

$$\sum_j d_j \hat{A}_j = \sum_{jk} M_{jk} \hat{A}_j \hat{A}_k = \sum_{kj} M^{-1}_{kj} d_k d_j\,. \tag{55}$$

Then

$$Q = \sum_{jk} M_{jk}(A_j - \hat{A}_j)(A_k - \hat{A}_k) - \sum_{jk} M_{jk} \hat{A}_j \hat{A}_k \tag{56}$$

and the joint likelihood function splits into three factors:

$$L = L_1\; L_2\; L_3 \tag{57}$$

with

$$L_1 = \sigma^{-N} \exp(-N\overline{y^2}/2\sigma^2) \tag{58}$$

$$L_2 = \exp\left\{-\frac{N}{2\sigma^2}\sum_{jk} M_{jk}\,(A_j - \hat{A}_j)\,(A_k - \hat{A}_k)\right\} \tag{59}$$

$$L_3 = \exp\left\{+\frac{N}{2\sigma^2}\sum_{jk} M_{jk}\hat{A}_j\,\hat{A}_k\right\}\ . \tag{60}$$

If σ is known, the factor L_1 may be dropped, since it will be absorbed into the normalization constant of the joint posterior distribution for the other parameters. If σ is unknown, then it becomes a "nuisance parameter" to be integrated out of the problem with respect to whatever prior probability $p(\sigma|I)$ describes our prior information about it. The most commonly used case is that where we are initially completely ignorant about it – or wish to proceed as if we were, in order to see what the consequences would be. As expounded in some detail elsewhere (Jaynes, 1980) the Jeffreys prior probability assignment $p(\sigma|I) = 1/\sigma$ is uniquely determined as the one which expresses "complete ignorance" of a scale parameter.

One of the fascinating things about Bayes' theorem is its efficiency in handling this situation. The noise level σ is highly relevant to our inferences; so if it is initially completely unknown, then it must be estimated from the data. But Bayes' theorem does this for us automatically in the process of integrating σ out of the problem. To see this, note that from (43) when we integrate out σ the quasi–likelihood for the remaining parameters becomes

$$\int_0^\infty L\,d\sigma/\sigma \propto \left[1 + (Q/\overline{y^2})\right]^{-N/2}\ . \tag{61}$$

But when N is reasonably large (*i.e.* we have enough data to permit a reasonably good estimate of σ), this is nearly the same as

$$\exp(-NQ/2\overline{y^2})\,. \tag{62}$$

In its dependence on $\{A_i, \omega_i, \theta_i\}$ this is just (43) with σ^2 replaced by $\overline{y^2}$. Thus, in effect, if we tell Bayes' theorem: "I'm sorry, I don't know what σ^2 is!" it replies to us, "That's all right, don't worry. We'll just replace σ^2 by the best estimate of σ^2 that we can make from the data, namely $\overline{y^2}$."

After doing this, we shall have the same quadratic form Q as before, and its minimization will locate the same "best" estimates of the other parameters as before. The only difference is that for small N the peak of (61) will not be as sharp as that of the Gaussian (43) so we are not quite so sure of the accuracy of our estimates; but that is the only price we paid for our ignorance of σ.

But if N is reasonably large it hardly matters whether σ is known or unknown. Supposing for simplicity, as we did before, that σ is known, the joint posterior density for the other parameters $\{A_i, \omega_i, \theta_i\}$ factors:

$$p(\{A_i.\omega_i, \theta_i\}|D, I) = p(\{A_i\}|\{\omega_i\theta_i\}, D, I)\cdot p(\{\omega_i, \theta_i\}|D, I) \tag{63}$$

in which, to explain this compact notation more explicitly, we have the joint conditional probability for the amplitudes A_i given the frequencies and phases:

$$p(\{A_j\}|\{\omega_j,\theta_j\},D,I) \propto \exp\left\{-\frac{N}{2\sigma^2}\sum_{jk} M_{jk}\,(A_j-\hat{A}_j)\,(A_k-\hat{A}_k)\right\} \qquad (64)$$

and the joint marginal posterior density for the frequencies and phases:

$$p(\{\omega_j,\theta_j\}|DI) \propto \exp\left\{+\frac{N}{2\sigma^2}\sum_{jk} M_{jk}\hat{A}_j\,\hat{A}_k\right\}\,. \qquad (65)$$

Eq. (64) says that $\hat{A}_j$ is the estimate of the amplitude A_j that we should make, given the frequencies and phases $\{\omega_j,\theta_j\}$, and (65) says that the most probable values of the frequencies and phases are those for which the estimated amplitudes are large. All this makes excellent common sense as soon as we see it; but nobody's intuition had seen it.

The above relations are, within the context of our model, exact and quite general. The number n of possible signals and the sampling times t_m may be chosen arbitrarily. Partly for that reason, to explore all the results that are in (63) – (65) would require far more space than we have here. We shall forego working out the interesting details of what happens to our conclusions when the sampling times are not equally spaced, what the answers to those more complicated questions like (B) and (C) are, and what happens to the matrix M in the limit when two frequencies coincide.

Somehow, as $\omega_j - \omega_k \to 0$, it must be that A_j and A_k become increasingly confounded (indistinguishable). In the limit there is only one amplitude where two used to be. Then will the rank of M drop from n to $(n-1)$? It is a recommended exercise to work this out for yourself in detail for the case $n = 2$, and see how as the two signals merge continuously into one there is a mixing of the eigenvectors of M much like the "level crossing" phenomenon of quantum theory.† At every stage the results make sense in a way that we do not think anybody's intuition can foresee, but which seems obvious after we contemplate what Bayes' theorem tells us.

Here we shall examine only the opposite limit, that of frequencies spaced far enough apart to be resolved, which of course requires that the number n of signals is not greater than the number N of observations. If our sampling points are equally spaced:

$$t_m = -(T+1)+m, \qquad 1 \le m \le N, \quad 2T+1 = N \qquad (66)$$

then M_{jk} reduces to

$$M_{jk} = N^{-1}\sum_{t=-T}^{T}\cos(\omega_j t+\theta_j)\,\cos(\omega_k t+\theta_k)\,. \qquad (67)$$

The diagonal elements are

$$M_{jj} = \frac{1}{2}+\frac{\sin N\omega_j}{2N\sin\omega_j}\,\cos 2\theta_j\,, \qquad 1 \le j \le n \qquad (68)$$

in which the second term becomes appreciable only when $\omega_j \to 0$ or $\omega_j \to \pi$. If we confine our frequencies to be positive, in $(0,\pi)$, then the terms in (67) with $(\omega_j+\omega_k)$ will never become large, and the off–diagonal elements are

$$M_{jk} \simeq \frac{\sin Nu}{2N\sin u}\,\cos(\theta_j-\theta_k)+O(N^{-1}) \qquad (69)$$

† If you are too lazy to work this out for himself, it is done for you in Bretthorst (1988).

where $u \equiv (\omega_j - \omega_k)/2$. This becomes of order unity only when the two frequencies are too close to resolve and that merging phenomenon begins. Thus as long as the frequencies ω_j are so well separated that

$$N|\omega_j - \omega_k| >> 1 \tag{70}$$

a good approximation will be simply

$$M_{jk} \simeq (1/2)\,\delta_{jk} \tag{71}$$

and the above relations simplify drastically. The amplitude estimates reduce to, from (49),

$$\hat{A}_j = 2d_j = 2\,N^{-1}\sqrt{NC(\omega_j, 0)}\,\cos(\theta_j + \psi_j) \tag{72}$$

and the joint posterior density for the frequencies and phases is

$$p(\{\omega_j, \theta_j|DI) \propto \exp\left[\frac{N}{\sigma^2}\sum_j d_j^2\right] = \exp\left[\sigma^{-2}\,\sum_j C(\omega_j, 0)\cos^2(\theta_j + \psi_j)\right] \tag{73}$$

while the joint posterior density for all the parameters is

$$p(\{A_j, \omega_j, \theta_j\}|DI) = \exp\left\{\sigma^{-2}\,\sum_{j=1}^{n}\left[A_j\sqrt{NC(\omega_j, 0)}\cos(\theta_j + \psi_j) - NA_j^2/4\right]\right\} \tag{74}$$

which is a product of independent distributions. If the prior probabilities for the phases $\{\theta_j\}$ were correlated, there would be an extra factor $p(\theta_1 \cdots \theta_n|I)$ in (74) that removes this independence, and might make an appreciable difference in our conclusions.

This could arise, for example, if we knew that the entire signal is a wavelet originating in a single event some time in the past; and at that initial time all frequency components were in phase. Then integrating the phases out of (74) will transfer that phase correlation to a correlation in the $\{A_j\}$. As a result the answers to questions such as (B) and (C) above would be changed in a possibly important way. This is another interesting detail that we cannot go into here; but merely note that all this is inherent in Bayes' theorem.

But usually our prior probability distribution for the phases is independent and uniform on $(0, 2\pi)$; *i.e.* we have no prior information about either the values of, or connections between, the different phases. Then the best inference we shall be able to make about the amplitudes and frequencies is obtained by integrating all the θ_j out of (73) and (74) independently. This will generate just the same $I_0(q)$ Bessel functions as before; and writing $q_j = C(\omega_j, 0)/2\sigma^2$, our final results are:

$$p(\{\omega_j\}|D, I) \propto \prod_{j=1}^{n} e^{q_j}\,I_0(q_j) \tag{75}$$

$$p(\{A_j, \omega_j\}|D, I) \propto \prod)j = 1^n \exp(-NA_j^2/4\sigma^2)\,I_0[A_j\sqrt{NC(\omega_j, 0)/\sigma^2}]\,. \tag{76}$$

But these are just products of independent distributions identical with our previous single–signal results (10), (26). We leave it as an exercise for the reader to show from this that our previous power spectrum estimate (27) will follow. Thus as long as we ask only question (A) above, our single–signal assumption was not a restriction after all.

At this point it is clear also that if our n signals are chirped, we need only replace $C(\omega_j, 0)$ by $C(\omega_j, \alpha_j)$ in these results, and we shall get the same answers as before to any question about frequency and chirp that involves individual signals, but not correlations between different signals.

11. CONCLUSION

Although the theory presented here is only the first step of the development that is visualized, we have thought it useful to give an extensive exposition of the Bayesian part of the theory.

No connection with AR models has yet appeared;[‡] but we expect this to happen when additional prior information is put in by entropy factors. In a full theory of spectrum estimation in the presence of noise, in the limit as the noise goes to zero the solution should reduce to something like the original Burg pure Maximum Entropy solution (it will not be exactly the same, because we are assuming a different kind of data).

For understanding and appreciating Bayesian inference, no theorems giving its theoretical foundations can be quite as effective as seeing it in operation on a real problem. Every new Bayesian solution like the present one gives us a new appreciation of the power and sophistication of Bayes' theorem as the true logic of science. It seeks out every factor in the model that has any relevance to the question being asked, tells us quantitatively how relevant it is – and relentlessly exposes how crude and primitive other methods were.

We could expand the example studied here to many volumes without exhausting all the interesting and useful detail contained in the general solution – almost none of which was anticipated by sampling theory or intuition.

APPENDIX A: OCEANOGRAPHIC CHIRP

We illustrate the phase cancellation phenomenon, for the spectral snapshot method described in the text, as follows. That method essentially calculates the periodogram of a data segment $\{y_t : -T \le t \le T\}$, or possibly a Blackman–Tukey smoothing of it (the difference, affecting only the quantitative details, is not crucial for the point to be made here). The periodogram is

$$X(\omega) = N^{-1} \,|\sum_t y_t\, e^{i\omega t}|^2 \,. \qquad \text{(A1)}$$

If y_t is a sinusoid of fixed frequency ν:

$$y_t = A \cos(\nu t + \theta) \qquad \text{(A2)}$$

then the periodogram reaches its peak value at or very near the true frequency:

$$X(\nu) = N A^2/4 \,. \qquad \text{(A3)}$$

But if the signal is chirped:

$$y_t = A \cos(\nu t + \alpha t^2 + \theta) \qquad \text{(A4)}$$

[‡] By 1994, ten years later, it still had not appeared; not because a search for it failed to find it, but because we were so overwhelmed by all the rich and useful detail in the 'plain vanilla' Bayesian solution that instead the writer and his colleagues have concentrated on producing about two dozen articles and a book dealing with it. The AR – entropy connections remain tasks for the future.

then the periodogram (A1) is reduced, broadened, and distorted. Its value at the center frequency is only about

$$X(\nu) = (NA^2/4) \left| N^{-1} \sum_t e^{i\alpha t^2} \right|^2 \tag{A5}$$

which is always less than (A3) if $\alpha \neq 0$. As a function of α, (A5) is essentially the Fresnel diffraction pattern of N equally spaced narrow slits. To estimate the phase cancellation effect, note that when $\alpha T^2 < 1$ the sum in (A5) may be approximated by a Fresnel integral, from whose analytic properties we may infer that when $\alpha T^2 \geq 1$, $X(\nu)$ will be reduced below (A3) by a factor of about

$$(\pi/4\alpha T^2)\left[1 - (2/\pi\alpha T^2)^{1/2}\cos(\alpha T^2 + \pi/4) + O(T^{-2})\right] \tag{A6}$$

It is clear from inspection of (A5) that the reduction is not severe if $\alpha T^2 \leq 1$, but (A6) shows that it can essentially wipe out the signal if $\alpha T^2 >> 1$.

Also, in the presence of chirp the periodogram (A1) exhibits "line broadening" and "line splitting" phenomena; for some values of αT^2 there appear to be two or more lines of different amplitudes. For graphs demonstrating this, see Barber & Ursell (1948).

Discovery of chirped ocean waves, originating from storms thousands of miles away, was attributed by Tukey *et al* to Munk and Snodgrass (1957). Their feat was termed "one of the virtuoso episodes in the annals of power spectrum analysis" showing the value of alertness to small unexpected things in one's data. In (Tukey, 1984) it is suggested that a Bayesian wears some kind of blinders that make him incapable of seeing such things, and that discovery of the phenomenon might have been delayed by decades – or even centuries – if Munk & Snodgrass had used Bayesian, AR, or Maximum Entropy methods.

The writer's examination of the Munk–Snodgrass article has led him to a different picture of these events. The chirped signals they found were not small; in the frequency band of interest they were the most prominent feature present. The signals consisted of pressure variations measured off the California coast at a depth of about 100 meters, attributed to a storm in the Indian ocean about 9,000 miles away. The periods were of the order of 20 seconds, decreasing at a rate of about 10% per day for a few days.

They took measurements every 4 seconds, accumulating continuous records of length $N = 2T = 6000$ observations, or 62/3 hours. Thus, from their grand average measured chirp rate of about $\alpha = 1.6 \times 10^{-7}$ sec^{-2} we get $\alpha T^2 = 24$, so the sum in (A5) wrapped several times around the Cornu spiral, and from (A6) phase cancellation must have reduced the apparent signal power at the center frequency to about 3% of its real value (this would have been about their best case, since nearer sources would give proportionally larger αT^2).

Such a phase cancellation would not be enough to prevent them from seeing the effect altogether, but it would greatly distort the line shape, as seems to have happened. Although they state that the greater length of their data records makes possible a much higher resolution (of the order of one part in 3000) than previously achieved, their actual lines are of complicated shape, and over 100 times wider than this.

As to date of discovery, this same phenomenon had been observed by Barber & Ursell in England, as early as 1945. A decade before Munk & Snodgrass, they had correlated chirped ocean waves arriving at the Cornwall coast with storms across the Atlantic, and as far away as the Falklands. In Barber & Ursell (1948) we find an analysis of the phase cancellation effect, which led them to limit their data records to 20 minutes, avoiding the difficulty. Indeed, the broad and complicated line shapes published by Munk & Snodgrass look very much like the spectra calculated by Barber & Ursell to illustrate the disastrous effects of phase cancellation when one uses too long a record.

The theory of this physical phenomenon, relating the chirp rate to the distance r to the source, was given in 1827 by Cauchy. In our notation, $\alpha = g/4r$, where g is the acceleration of gravity. For example, in the Summer of 1945 Barber and Ursell observed a signal whose period fell from 17.4 sec to 12.9 sec in 36 hours. These data give $\alpha = 5 \times 10^{-7}$ sec^{-2}, placing the source about 3000 miles away, which was verified by weather records. In May 1946 a signal appeared, whose period fell from 21.0 sec to 13.9 sec in 4 days, and came from a source 6000 miles away.

Up to 1947 Barber and Ursell had analyzed some 40 instances like this, without using anybody's recommended methods of spectrum analysis to measure those periods. Instead they needed only a home–made analog computer, putting a picture of their data on a rotating drum and noting how it excited a resonant galvanometer as the rotation speed varied!

Munk & Snodgrass surely knew about the phase cancellation effect, and were not claiming to have discovered the phenomenon, since they make reference to Barber & Ursell. It appears to us that they did not choose their record lengths with these chirped signals in mind, simply because they had intended to study other phenomena. But the signals were so strong that they saw them anyway – not as a result of using a data analysis method appropriate to find them, but in spite of an inappropriate method.

Today, it would be interesting to re–analyze their original data by the method suggested here. If there is a constant amplitude and chirp rate across the data record, the chirpogram should reach the full maximum (A3), without amplitude degradation or broadening, at the true center frequency and chirp rate; and should therefore provide a much more sensitive and accurate data analysis procedure.

APPENDIX B: WHY GAUSSIAN NOISE?

In what L. J. Savage (1954) called the "objectivist" school of statistical thought (nowadays more often called "orthodox" or "sampling theory"), assigning a noise distribution is interpreted as asserting or hypothesizing a statement of fact; *i.e.* a physically real property of the noise. It is, furthermore, a widely believed "folk-theorem" that if the actual *frequency* distribution of the noise differs from the *probability* distribution that we assigned, then all sorts of terrible things will happen to us; we shall be misled into drawing all sorts of erroneous conclusions. We wish to comment on both of these beliefs.

We are aware of no real problem in which we have the detailed information that could justify such a strong interpretation of our noise distribution at the beginning of a problem; nor do we ever acquire information that could verify such an interpretation at the end of the problem.

As noted in the text, an assigned noise distribution is a joint distribution for all the errors; *i.e.* a probability $p(e|I)$ assigned to the total noise vector $e = (e_1 \cdots e_n)$ in an n–dimensional space. Obviously, this cannot be a statement of verifiable fact, for the experiment will generate only one noise vector. Our prior distribution $p(e|I)$ defines rather the range of different *possible* noise vectors that we wish to allow for, only one of which will actually be realized.

In the problems considered here the information that would be useful in improving our spectrum estimates consists of correlations between the e_i (nonwhite noise). Such correlations cannot be described at all in terms of the frequencies of individual noise values. A correlation extending over a lag m is related to frequencies only of noise sequences of length $\geq m$. Even for $m = 3$ the number of possible sequences of length m is usually far greater than the length of our data record; so it is quite meaningless to speak of the "frequencies" with which the different sequences of length $m = 3$ appear in our data, and therefore equally meaningless to ask whether our noise probability distribution correctly describes those frequencies.

As these considerations indicate, the function of $p(e|I)$ cannot be to describe the noise; but rather to describe our *state of knowledge* about the noise. It is related to facts to this extent: we want to be fairly sure that we choose a set of possible vectors big enough to include the true one. This is a matter of being honest about just how much prior information we actually have; *i.e.* of avoiding unwarranted assumptions.

If in our ignorance we assign a noise distribution that is "wider" (for example, which supposes a greater mean–square error) than the actual noise vector, then we have only been more conservative – making allowance for a greater range of possibilities – than we needed to be. But the result is not that we shall see erroneous "effects" that are not there; rather, we shall have less discriminating power to detect small effects than we might have enjoyed, had we more accurate prior knowledge about the noise.

If we assign a noise distribution so "narrow" that the true noise vector lies far outside the set thought possible, then we have been dishonest and made unwarranted assumptions; for valid prior information could not have justified such a narrow distribution. Then as noted in the text we shall, indeed, pay the penalty of seeing things that are not there. The goal of stating, by our prior distribution, what we honestly do know – and nothing more – is the means by which we protect ourselves against this danger.

Tukey *et al* (1980) comment: "Trying to think of data analysis in terms of hypotheses is dangerous and misleading. Its most natural consequences are (a) hesitation to use tools that would be useful because 'we do not know that their hypotheses hold' or (b) unwarranted belief that the real world is as simple and neat as these hypotheses would suggest. Either consequence can be very costly $\cdots$. A procedure does not have hypotheses – rather there are circumstances where it does better and others where it does worse."

We are in complete agreement with this observation; and indeed would put it more strongly. While some hypotheses about the nature of the phenomenon may suggest a procedure – or even uniquely determine a procedure – the procedure itself has no hypotheses, and the same procedure may be suggested by many very different hypotheses. For example, as we noted in the text, the Blackman–Tukey window smoothing procedure was associated by them with the hypothesis that the data were a realization of a "stationary Gaussian random process". But of course nothing prevents one from applying the procedure itself to any set of data whatsoever, whether or not 'their hypotheses hold'. And indeed, there are "circumstances where it does better and others where it does worse".

But we believe also that probability theory incorporating Bayesian/Maximum Entropy principles is the proper tool – and a very powerful one – for (a) determining those circumstances for a given procedure; and (b) determining the optimal procedure, given what we know about the circumstances. This belief is supported by decades of theoretical and practical demonstrations of that power.

Clearly, while striving to avoid gratuitous assumption of information that we do not have, we ought at the same time to use all the relevant information that we actually do have; and so Tukey has also wisely advised us to think very hard about the real phenomenon being observed, so that we can recognize those special circumstances that matter and take them into account. As a general statement of policy, we could ask for nothing better; so our question is: how can we implement that policy in practice?

The original motivation for the Principle of Maximum Entropy (Jaynes, 1957) was precisely the avoidance of gratuitous hypotheses, while taking account of what is known. It appears to us that in many real problems the procedure of the Maximum Entropy Principle meets both of these requirements, and thus represents the explicit realization of Tukey's goal.

If we are so fortunate as to have additional information about the noise beyond the mean-square value supposed in the text, we can exploit this to make the signal more visible, because it reduces the measure W, or "volume", of the support set of possible noise variations that we have to allow for. For example, if we learn some respect in which the noise is not white, then it becomes in part predictable and some signals that were previously indistinguishable from the noise can now be separated.

The effectiveness of new information in thus increasing signal visibility, is determined by the reduction it achieves in the entropy of the joint distribution of noise values – essentially, the logarithm of the ratio W'/W by which that measure is reduced. The Maximum Entropy formalism is the mathematical tool that enables us to locate the new contracted support set on which the likely noise vectors lie.

It appears to us that the evidence for the superior power of Bayesian/Maximum Entropy methods over both intuition and "orthodox" methods is now so overwhelming that nobody concerned with data analysis – in any field – can afford to ignore it. In our opinion these methods, far from conflicting with the goals and principles expounded by Tukey, represent their explicit quantitative realization, which intuition could only approximate in a crude way.

APPENDIX C: DETAILS OF CALCULATIONS

Derivation of the Chirpogram. Expanding the cosine in (9) we have

$$\sum y_t \cos(\omega t + \alpha t^2 + \theta) = P \cos\theta - Q \sin\theta = (P^2 + Q^2)^{1/2} \cos[\theta + \tan^{-1}(Q/P)] \tag{C1}$$

where

$$P = \sum_t y_t \cos(\omega t + \alpha t^2) \tag{C2}$$

$$Q = \sum_t y_t \sin(\omega t + \alpha t^2) \tag{C3}$$

But $(P^2 + Q^2)$ can be written as the double sum

$$\begin{aligned} P^2 + q^2 &= \sum_{ts} y_t y_s [\cos(\omega t + \alpha t^2) \cos(\omega s + \alpha s^2) + \sin(\omega t + \alpha t^2) \sin(\omega s + \alpha s^2)] \\ &= \sum_{ts} y_t y_s [\cos \omega(t - s) + \alpha(t^2 - s^2)] \end{aligned} \tag{C4}$$

Therefore, defining $C(\omega, \alpha)$ as

$$C(\omega, \alpha) \equiv N^{-1}(P^2 + Q^2) \tag{C5}$$

and substituting (C1), (C4) into (9), the integral (7) over θ is the standard integral representation of the Bessel function:

$$I_0(x) = (2\pi)^{-1} \int_0^{2\pi} \exp(x \cos\theta)\, d\theta \tag{C6}$$

which yields the result (10) of the text.

Power Spectrum Derivations. From the integral formula

$$Z(a,b) \equiv \int_0^\infty e^{-ax^2}\, I_0(bx)\, dx = (\pi/4a)^{1/2}\, \exp(b^2/8a)\, I_0(b^2/8a) \tag{C7}$$

we obtain

$$\int_0^\infty e^{-ax^2}\, I_0(bx)\, dx = -\frac{\partial Z}{\partial a} = \frac{1}{4}\sqrt{\frac{\pi}{a^3}}\, [(1+2q)I_0(q) + 2qI_1(q)]\, e^q \tag{C8}$$

where $q \equiv b^2/8a$. In the notation of Eq. (12) we have $x = A$, $a = N/4\sigma^2$, $b^2 = NC(\omega,0)/\sigma^4$, therefore $q = C(\omega,0)/2\sigma^2$. Thus (C5) and (C6) become

$$\int_0^\infty dAL(A,\omega) = \sqrt{\frac{\pi\sigma^2}{N}}\, e^q\, I_0(q) \tag{C9}$$

$$\int_0^\infty dAA^2 L(A,\omega) = 2\sigma^3\sqrt{\frac{\pi}{N^3}}\, [(1+2q)I_0(q) + 2qI_1(q)]\, e^q \tag{C10}$$

from which (18), (20) of the text follow.

REFERENCES

N. F. Barber & F. Ursell (1948), "The Generation and Propagation of Ocean Waves and Swell", Phil. Trans. Roy. Soc. London, **A240**, pp 527–560.

J. M. Bernardo, *et al*, editors (1985), *Bayesian Statistics 2*, Elsevier Science Publishers, North–Holland. Proceedings of the Second Valencia International Meeting on Bayesian Statistics, Sept. 6–10, 1983.

R. B. Blackman & J. W. Tukey (1958), *The Measurement of Power Spectra*, Dover Publications, Inc., New York.

G. Larry Bretthorst (1988), *Bayesian Spectrum Analysis and Parameter Estimation*, Springer Lecture Notes in Statistics, #47. Numerous computer printouts with real and simulated data.

John Parker Burg (1967), "Maximum Entropy Spectral Analysis", in Proc. 37'th Meet. Soc. Exploration Geophysicists. Reprinted in *Modern Spectrum Analysis*, D. Childers, Editor, IEEE Press, New York. (1978).

John Parker Burg (1975), "Maximum Entropy Spectral Analysis", Ph.D. Thesis, Stanford University.

D. R. Griffin (1958), *Listening in the Dark*, Yale University Press, New Haven; see also *About Bats*, R. H. Slaughter & D. W. Walton, Editors, SMU Press, Dallas, Texas (1970).

B. Nikolaus & D. Grischkowsky (1983), "90 Fsec Tunable Optical Pulses Obtained by Two–Stage Pulse Compression", App. Phys. Lett.

S. F. Gull & G. J. Daniell (1978), "Image Reconstruction from Incomplete and Noisy Data", Nature, **272**, p. 686.

R. A. Helliwell (1965), *Whistlers and Related Ionospheric Phenomena*, Stanford University Press, Palo Alto, Calif.

E. T. Jaynes (1957), "Information Theory and Statistical Mechanics", Phys. Rev. **106**, 620; **108**, 171.

E. T. Jaynes (1973), "Survey of the Present Status of Neoclassical Radiation Theory", in Proceedings of the 1972 Rochester Conference on Optical Coherence, L. Mandel & E. Wolf, Editors, Pergamon Press, New York.

E. T. Jaynes (1980), "Marginalization and Prior Probabilities", in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner, Editor, North–Holland Publishing Co. Reprinted in E. T. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, a reprint collection, D. Reidel, Dordrecht–Holland (1983).

E. T. Jaynes (1981), "What is the Problem?", Proceedings of the Second ASSP Workshop on Spectrum Analysis, McMaster University, S. Haykin, Editor.

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

W. H. Munk & F. E. Snodgrass (1957), "Measurements of Southern Swell at Guadalupe Island", Deep–Sea Research, **4**, pp 272–286.

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

A. Schuster (1897), "On Lunar and Solar Periodicities of Earthquakes", Proc. Roy. Soc. **61**, pp. 455–465.

J. W. Tukey, P. Bloomfield, D. Brillinger, and W. S. Cleveland (1980), *The Practice of Spectrum Analysis*, notes on a course given in Princeton, N. J. in December 1980.

J. W. Tukey & D. Brillinger (1982), unpublished.

J. W. Tukey (1984), "Styles of Spectrum Analysis", Scripps Institution of Oceanography Reference Series 84–85, March 1984; pp. 100–103. An astonishing attack on all theoretical principles, including AR models, MAXENT, and Bayesian methods. John Tukey believed that one should not rely on any theory, but simply use his own intuition on every individual problem. But this made most Bayesian results inaccessible to him.