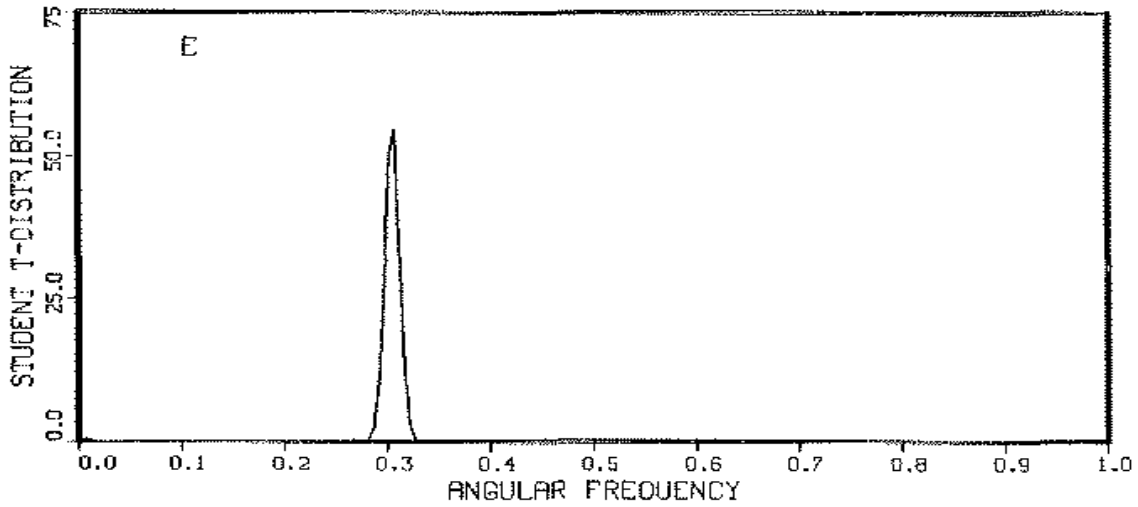
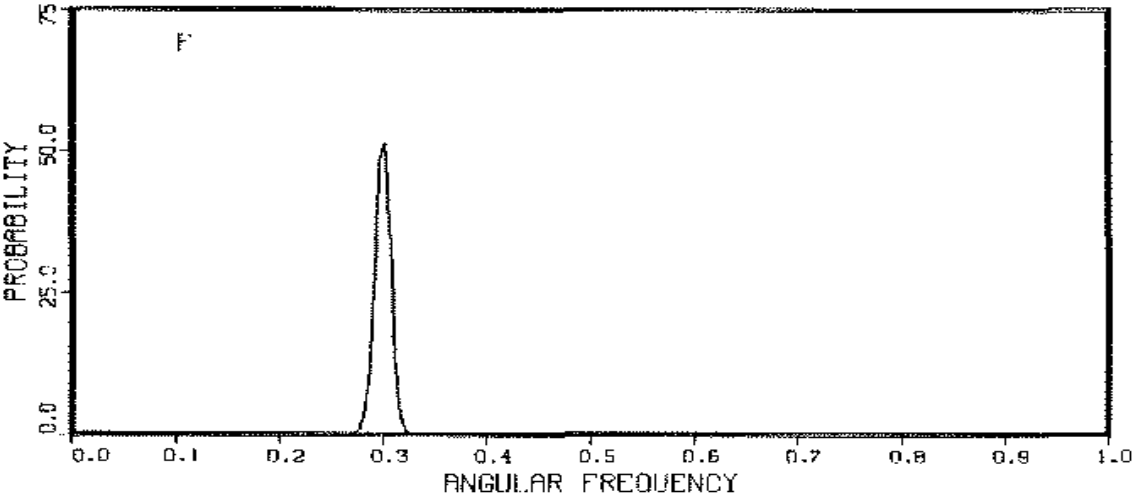
PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A FOURTH ORDER TREND CORRECTION

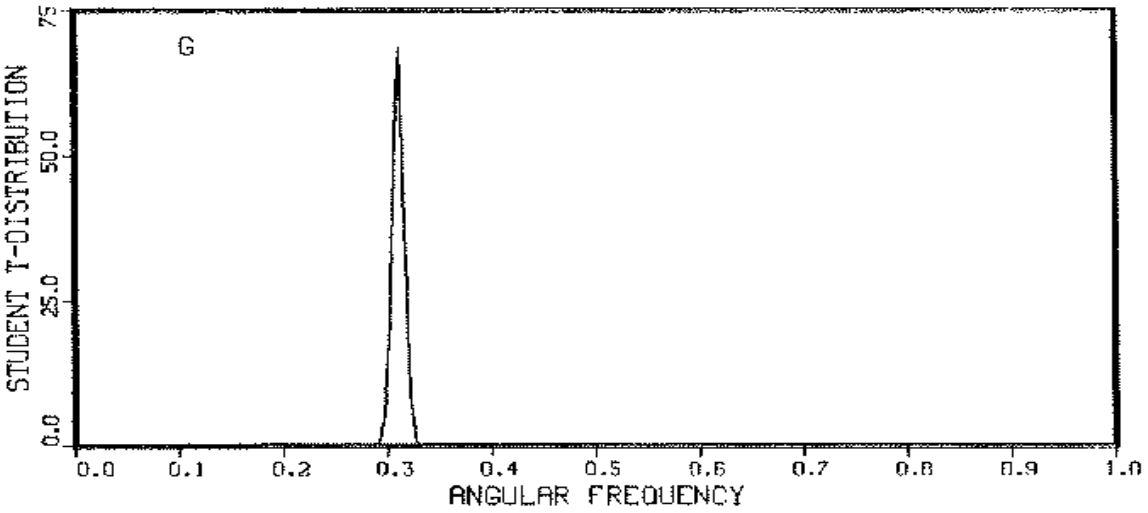


PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A FIFTH ORDER TREND CORRECTION

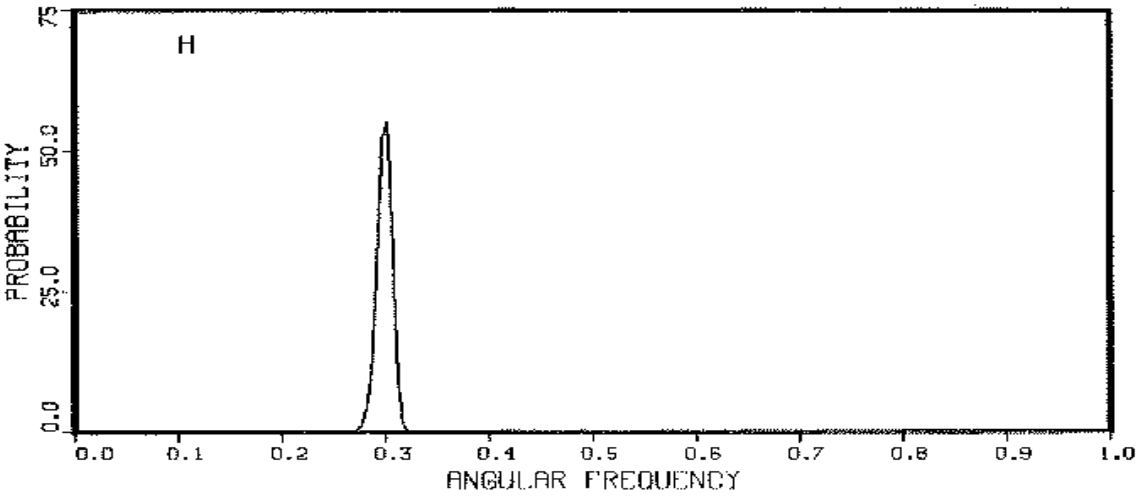


When the probability of a fourth-order trend plus a harmonic frequency is computed the trend is now completely gone and only the frequency at 20 years remains (E). When the expansion order is increased in (F) the frequency estimate is not essentially changed.

PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A SIXTH ORDER TREND CORRECTION

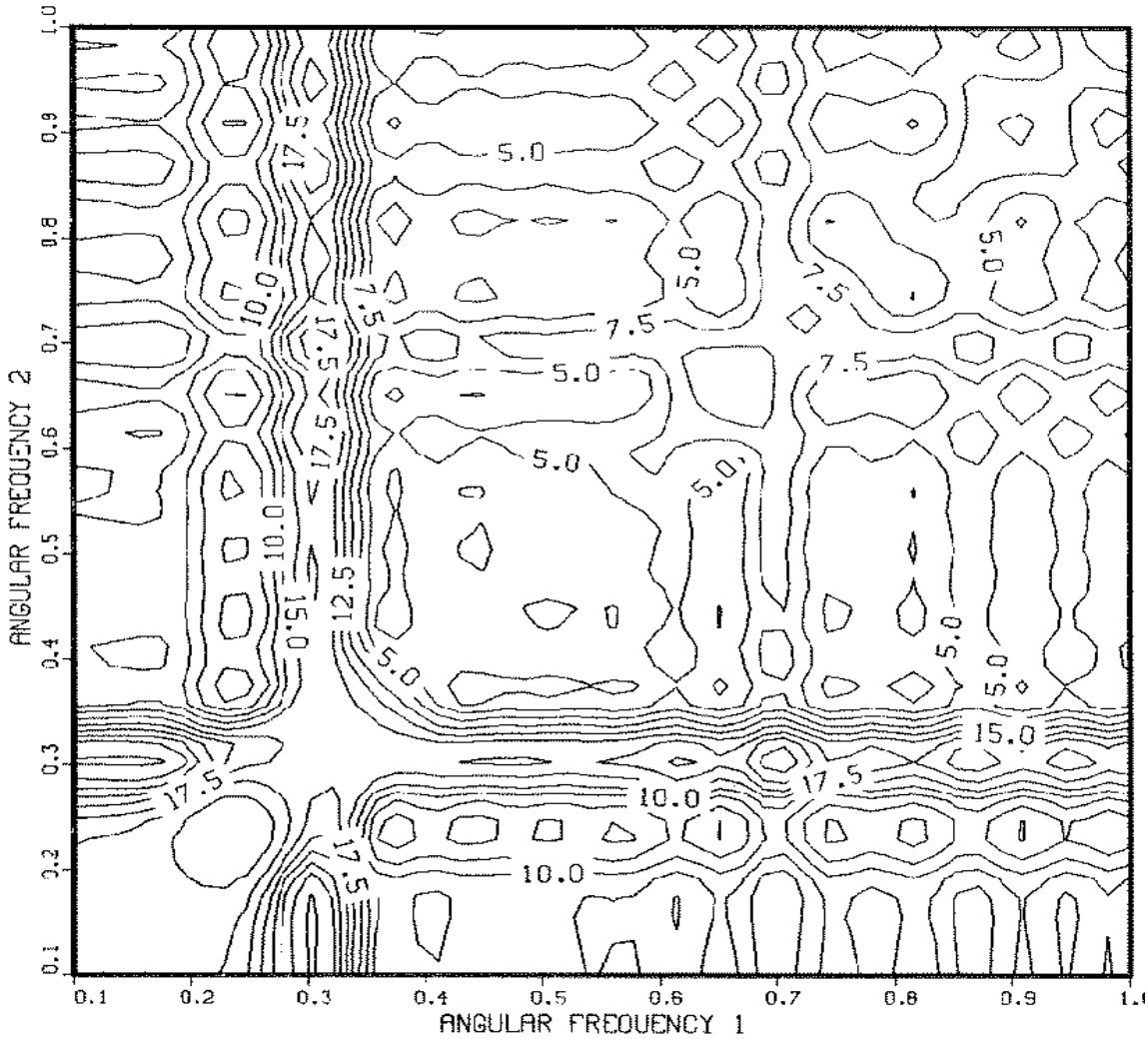


PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A SEVENTH ORDER TREND CORRECTION



Increasing the expansion order further does not significantly affect the estimated frequency (G) and (H). If the expansion order is increased sufficiently, the expansion will begin to remove the harmonic oscillation; and the posterior probability density will gradually decrease in height.

Figure 7.13: Probability of Two Frequencies After Trend Correction



This is the natural logarithm of the probability of two common harmonic frequencies in the crop yield data with a fifth order trend. This type of structure is what one expects from the sufficient statistic when there is only one frequency present. Notice the maximum is located roughly along a vertical and horizontal line at 0.3.

We did not seek to remove the trend from the data, but rather to eliminate its effect from the conclusions.

7.3 Another NMR Example

Now that the tools have been developed we can demonstrate how one can incorporate partial information about a model. In the corn crop example the trend was unknown, so it was expanded in orthonormal polynomials and integrated out of the problem, while we included what partial information we had in the form of the sine and cosine terms. In this NMR example let us assume that the decay function is of interest to us. We would like to determine this function as accurately as possible.

The data we used, Fig. 7.14(A), in this example are one channel of a pure D_2 spectrum [31]. Figure 7.14(B) contains the periodogram for these data. For this demonstration we will use the first $N = 512$ data points because they contain most of the signal.

For D_2, theoretical studies indicate there is a single frequency with decay [32]. Now we expect the signal should have the form

$$f(t) = [B_1 \sin(\omega t) + B_2 \cos(\omega t)]\, D(t),$$

where $D(t)$ is the decay function, and the sine and cosine effectively express what partial information we have about the signal. We will expand the decay function $D(t)$ to obtain

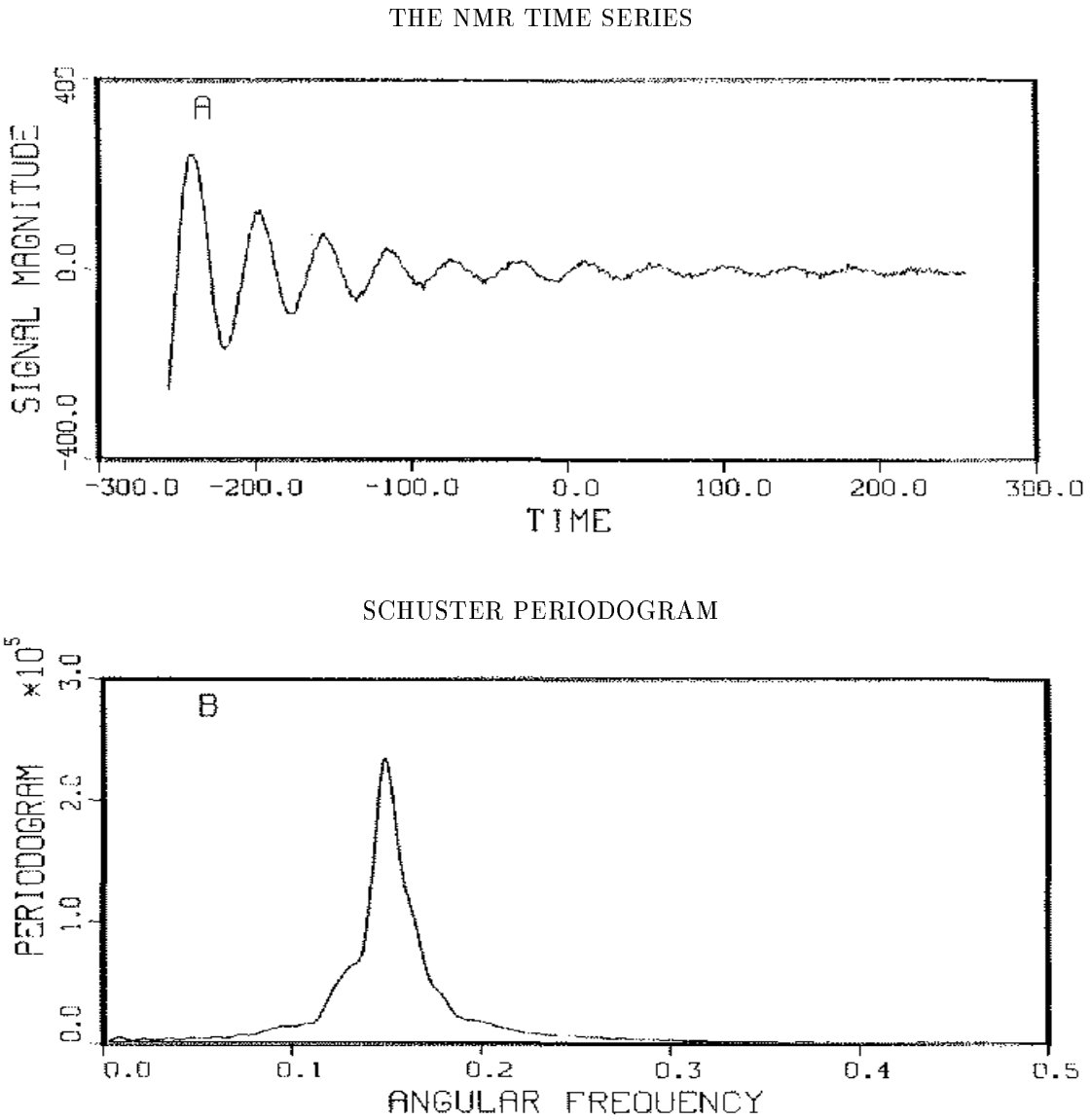
$$f(t) = [B_1 \sin(\omega t) + B_2 \cos(\omega t)] \sum_{j=0}^{r} D_j L_j(t)$$

where D_j are the expansion coefficients for the decay function, B_1 and B_2 are effectively the amplitude and phase of the sinusoidal oscillations, and L_j are the Legendre polynomials with the appropriate change of variables. This model can be rewritten as

$$f(t) = \sum_{j=0}^{r} D_j B_1 \left[L_j(t)[\sin(\omega t) + \frac{B_2}{B_1} \cos(\omega t)] \right].$$

There is an indeterminacy in the overall scale. That is, the amplitude of the sinusoid and the amplitude of the decay $D(t)$ cannot both be determined. One of them is necessarily arbitrary. We chose the amplitude of the sine term to be unity because it effectively eliminates one $\{\omega\}$ parameter from the problem. We have a choice, in this problem, on which parameters are to be removed by integration. We

Figure 7.14: A Second NMR Example - Decay Envelope Extraction



These NMR data (A) are a free-induction decay for a D_2 sample. The sample was excited using a 55MHz pulse and the signal detected using a mixer-demodulator. We used 512 data samples to compute the periodogram (B). We would like to use probability theory to obtain an estimate of the decay function while incorporating what little we know about the oscillations.

chose to eliminate $\{D_j B_1\}$ because there are many more of them, even though they are really the parameters of interest.

When we eliminate a parameter from the problem, it does not mean that it cannot be estimated. In fact, we can always calculate the parameters $\{D_j B_1\}$ from the linear relations between models, Eq. (4.2). For this problem it is simpler to search for the maximum of the probability distribution as a function of frequency ω and the ratio B_1/B_2, and then use Eq. (4.2) to compute the expansion coefficients D_j. If we choose to eliminate the amplitudes of the sine and cosine terms, then we must search for the maximum of the probability distribution as a function of the expansion parameters; there could be a large number of these.

We must again set the expansion order r; here we have plenty of data so in principle we could take r to be large. However, unless the decay is rapidly varying we would expect a moderate expansion of perhaps 5th to 10th order to be more than adequate. In the examples given here we set the expansion order to 10. We solved the problem also with the expansion order set to 5, and the results were effectively identical to the tenth order expansion.

To solve this problem we again used the computer code in Appendix E, and the "pattern" search routine discussed earlier. We located the maximum of the two dimensional "Student t-distribution," Eq. (3.17), and used the procedure given in Chapter 4, Eqs. (4.9) through (4.14), to estimate the standard deviation of the parameters. We find these to be

$$(\omega)_{\text{est}} = 0.14976 \pm 2 \times 10^{-5}$$

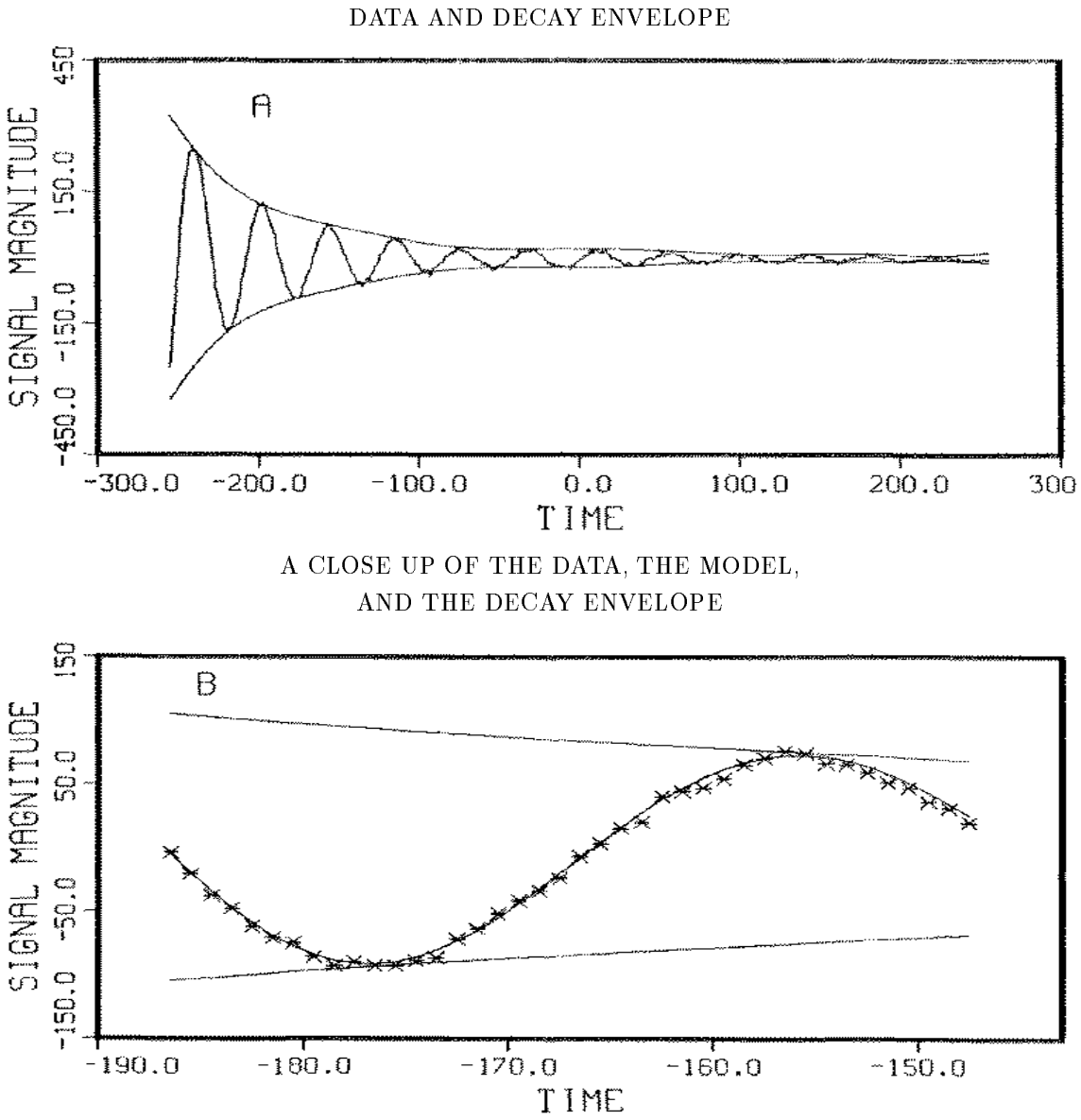
$$\left(\frac{B_2}{B_1}\right)_{\text{est}} = -.475 \pm 5 \times 10^{-3}$$

at two standard deviations. The variance of these data was $\overline{d^2} = 2902$, the estimated noise variance $(\sigma^2)_{\text{est}} \approx 27.1$, and the signal-to-noise ratio was 23.3.

After locating the maximum of the probability density we used the linear relations (4.2) between the orthonormal model and the nonorthonormal model to compute the expansion coefficients. We set the scale by requiring the decay function and the reconstructed model function to touch at one point near the global maximum. We have plotted the data and the estimated decay function, Fig. 7.15(A). In Fig. 7.15(B) we have a close up of the data, the decay function, and the reconstructed signal.

It is apparent from this plot that the decay is not Lorentzian or there is a second very small frequency present in the data. The decay function drops rapidly and then begins to oscillate. This is a real effect and is not an artifact of the procedure we are

Figure 7.15: How Does an NMR Signal Decay?



The decay function in (A) comes down smoothly and then begins to oscillate. This is a real effect, and is not an artifact of the analysis. In (B) we have plotted a blow up of the data, the predicted signal, and the decay function.

using. There are two possible interpretations: there could be a second small signal which is beating against the primary signal, or the inhomogeneous magnetic field could be causing it. When a sample is placed in a magnetic field each individual dipole in the field precesses at a well defined rate proportional to the local magnetic field. When the field is inhomogeneous (badly shimmed) a sample will resonate with an entire spectrum of frequencies around the principal frequency. Typically this distribution will manifest itself microscopically as a broadening or perhaps a splitting in lines: they become doubles and that is what we see here as this small oscillation. If we were to go back and look at this resonance very carefully we would find a second very small peak.

7.4 Wolf's Relative Sunspot Numbers

In 1848 Rudolph Wolf introduced the relative sunspot numbers as a measure of solar activity. These numbers, defined earlier, are available as yearly averages since 1700 – Fig. 2.1(A). The importance of these numbers is primarily because they are the longest available quantitative index of the sun's internal activity. The most prominent feature in these numbers is the 11.04 year cycle mentioned earlier. In addition to this cycle a number of others have been reported including cycles of 180, 90, 45, and a 22 years as well as a number of others [37], [38]. We will apply probability theory to these numbers to see what can be learned. We must stress that in what follows we do not know what the "true" model is, but can only examine a number of different possibilities. We begin by trying to determine the approximate number of degrees of freedom any reasonable model of these numbers should have.

7.4.1 Orthogonal Expansion of the Relative Sunspot Numbers

We can get a better understanding of the sunspot numbers if we simply expand these numbers in orthogonal vectors, and allow Eqs. (5.1) and (5.9) to indicate the number of expansion vectors needed to represent the data. This slight variant of the discrete Fourier transform will serve several useful purposes: it will give us an indication of the complexity of the data set, and it will indicate the noise level.

For this simple expansion we used sines and cosines. We generated the cosine

vectors using

$$H_j(t_i) = \frac{1}{\sqrt{c_j}} \cos\left(\frac{\pi j t_i}{N}\right)$$

$$c_j \equiv \sum_{i=1}^{N} \cos^2\left(\frac{\pi j t_i}{N}\right)$$

and the sine vectors using

$$H_k(t_i) = \frac{1}{\sqrt{s_j}} \sin\left(\frac{\pi k t_i}{N}\right)$$

$$s_j \equiv \sum_{i=1}^{N} \sin^2\left(\frac{\pi k t_i}{N}\right)$$

where $0 \leq k \leq N/2$ for the cosine components and $1 \leq k \leq N/2$ for the sine components. There are a total of 285 expansion vectors, and for this problem the time increments are one year. Next we computed h_k: the dot product between the data and the expansion vectors. Both the sine and cosine dot products were then squared and sorted into decreasing order.

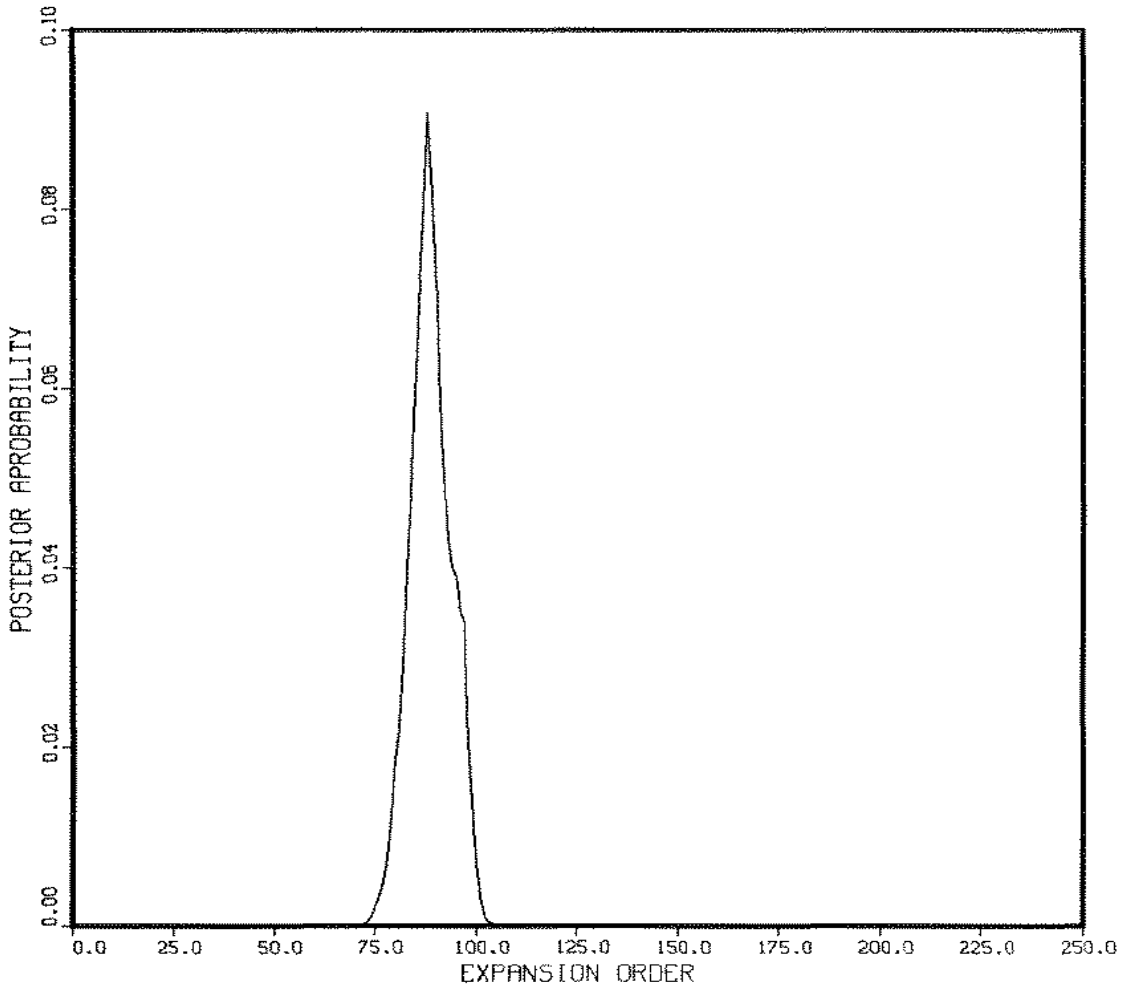
From these ordered projections we could then easily compute the probability of the expansion order E. For this problem this is essentially the posterior probability Eq. (5.9) with $r = 0$ and the terms associated with the $\{\omega\}$ set equal to 1. Because we are using an orthonormal expansion the Jacobian is unity. This simplifies Eq. (5.9) somewhat; we have

$$P(E|D,I) = \Gamma\left(\frac{E}{2}\right)\Gamma\left(\frac{N-E}{2}\right)\left[\frac{E(\overline{h^2})_E}{2}\right]^{-\frac{E}{2}}\left[\frac{N-E}{2}\langle\sigma^2\rangle\right]^{\frac{E-N}{2}},$$

where $(\overline{h^2})_E$ is the sufficient statistic computed with the E largest orthonormal projections. Figure 7.16 is a plot of the posterior probability of the model as a function of expansion order E. One can see from the plot that there is a peak in the probability around 90, and if one wants to be certain that all of the systematic component has been expanded, then the expansion order must be taken to be approximately 100. The estimated signal-to-noise ratio of these data is approximately 11.5, and the estimated standard deviation is about 5.

An orthogonal expansion of the data is about the worst model one could pick, in the sense of having the largest number of degrees of freedom. If we were to produce a model that reduced the total number of degrees of freedom by a factor 3 we would still have over thirty. For a simple harmonic frequency model, that would be 10 to 14 total frequencies. There are 286 data values, and the main period of roughly 11

Figure 7.16: The Probability of the Expansion Order



We expanded the Wolf sunspot numbers on orthonormal vectors and then used Eq. (5.9) to decide when to stop the expansion. This probability density indicates that the sunspot numbers are an extremely complex data set needing approximately 100 degrees of freedom to represent them.

years; that gives 26 cycles in the record. If each period has a unique amplitude, that still leaves approximately six to ten degrees of freedom to describe the shape of the oscillation. The implication of this is that Wolf's numbers are intrinsically extremely complicated, and no simple model for these numbers is going to prove possible. We will investigate them using a number of relatively simple models to see what can be learned.

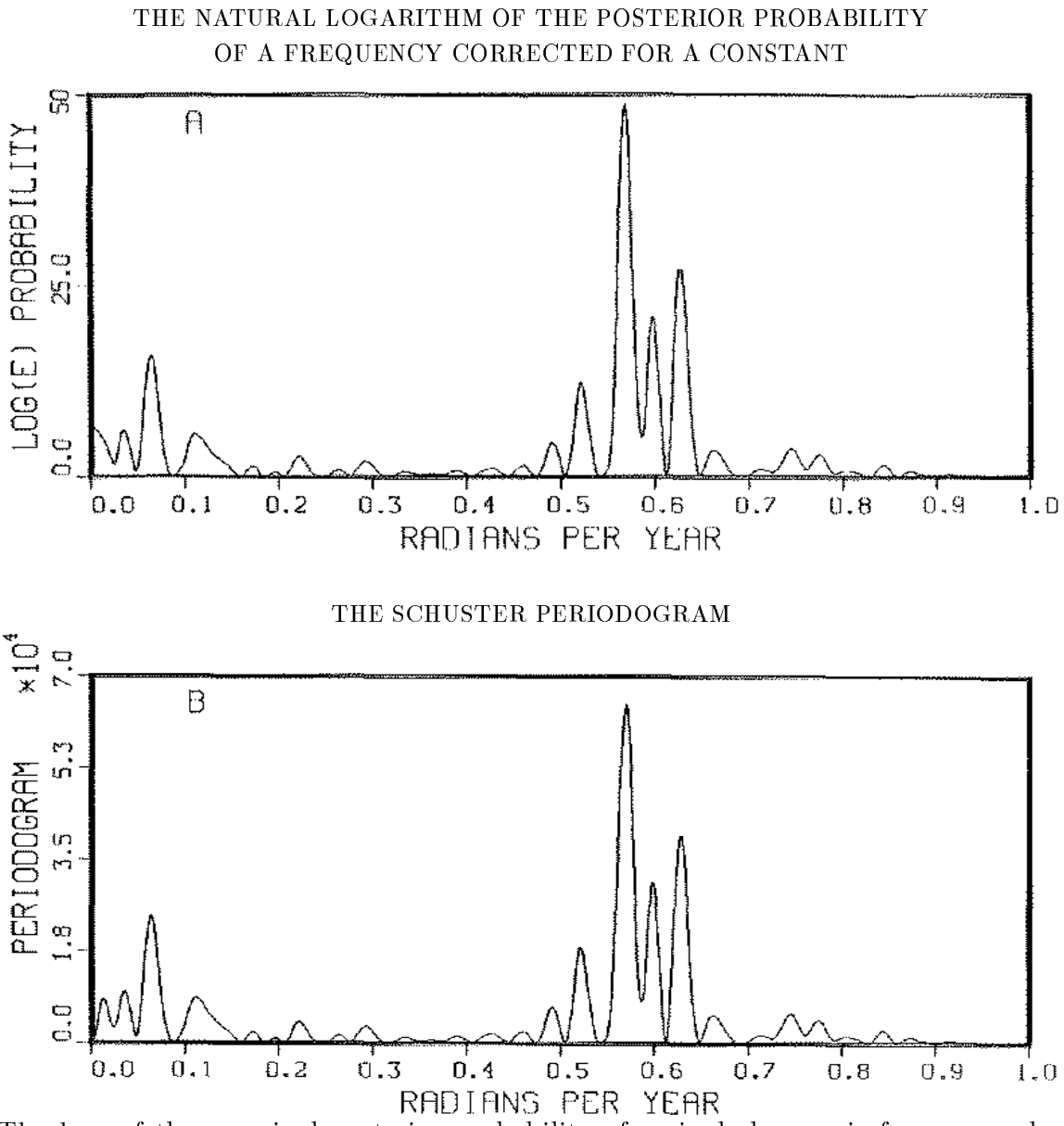
7.4.2 Harmonic Analysis of the Relative Sunspot Numbers

The second model we will investigate is the multiple harmonic frequency model. There are three degrees of freedom for each frequency, and with 100 degrees of freedom in the data, there is no chance of finding all of the structure in them. We will content ourselves with finding the first few frequencies and seeing how the results compare with the orthogonal expansion. Many writers have performed a harmonic analysis on these numbers. We will compare our results to those obtained recently by Sonett [38] and Bracewell [39]. The analysis done by Sonett concentrated on determining the spectrum of the relative sunspot numbers. He used the Burg [36] algorithm. This routine is extremely sensitive to the frequencies. In addition to finding the frequencies, this routine will sometimes shift the location of the predicted frequency, and it estimates a spectral density (a power normalized probability distribution), not the power carried in a line. Consequently, no accurate determination of the power carried by these lines has been done. As explained by Jaynes [40], the Burg algorithm yields the optimal solution to a certain well-defined problem. But in practice it is used in some very different problems for which it is not optimal (although still useful). We will use probability theory to estimate the frequencies, their accuracy, the amplitudes, the phases, as well as the power carried by each line.

Again, we plot the log of the probability of a single harmonic frequency plus a constant, Fig. 7.17(A). In this study, we include a constant and allow probability theory to remove it the correct way, instead of subtracting the average from the data as was done in Chapter 2. We do this to see if this theoretically correct way of eliminating a constant will make any difference in the evidence for frequencies. Thus we plot the log of the marginal posterior probability Eq. (3.17) using

$$f(t) = B_1 + B_2 \cos \omega t + B_3 \sin \omega t$$

Figure 7.17: Adding a Constant to the Model



The $\log_e$ of the marginal posterior probability of a single harmonic frequency plus a constant (A), and the periodogram (B) are almost identical. The periodogram is related to the posterior probability when σ^2 is known; for a data set with zero mean the periodogram must go to zero at zero frequency. The low frequency peak near zero in (B) is caused by subtracting the average from the data. The $\log_e$ of the marginal posterior probability of a single harmonic frequency plus a constant will go to zero at zero only if there is no evidence of a constant component in the data. Thus (A) does not indicate the presence of a spurious low frequency peak, only a constant.

as the model. The periodogram, Fig. 7.17(B), is a sufficient statistic for a single harmonic frequency if and only if the time series has zero mean. Under these conditions the periodogram must go to zero at $\omega = 0$. But this is the only difference visible; in the periodogram, Fig. 7.17(B), the low frequency peak near zero is a spurious effect due to subtracting the average value from the data. Probability analysis using a simple harmonic frequency plus a constant does not show any evidence for this period, Fig. 7.17(A).

Next we applied the general procedure for finding multiple frequencies. We started with the single frequency which best described the data, then computed the residuals and looked to see if there was evidence for additional frequencies in the residuals. The initial estimate from the residuals was then used in a two-frequency model. We continued this process until we had a nine-frequency model. Next we computed the standard deviation using the procedure developed in Chapter 4, Eqs. (4.9) through (4.14). Last, we used the linear relations between the models, Eq. (4.2), to compute the nonorthonormal amplitudes as well as their second moments. These are summarized as in Table . With these nine frequencies and one constant, the estimated standard deviation of the noise is $(\sigma)_{\text{est}} = 15$, and the signal-to-noise ratio is 14. The constant term had a value of 46.

We have plotted these nine frequencies as normalized Gaussians, Fig. 7.18(A), to get a better understanding of their determination. We plot in Fig. 7.18(B) an approximation to the line spectral density obtained by normalizing Fig. 7.18(A) to the appropriate power level. The dotted line on this plot is the periodogram normalized to the highest value in the power spectral density. This plot brings home the fact that when the frequencies are close, the periodogram is not even approximately a sufficient statistic for estimating multiple harmonic frequencies. At least one of the frequencies found by the nine-frequency model occurs right at a minimum of the periodogram. Also notice that the normalized power is more or less in fair agreement with the periodogram when the frequencies are well separated. That is because, for a simple harmonic frequency, the peak of the periodogram is indeed a good estimate of the energy carried in that line.

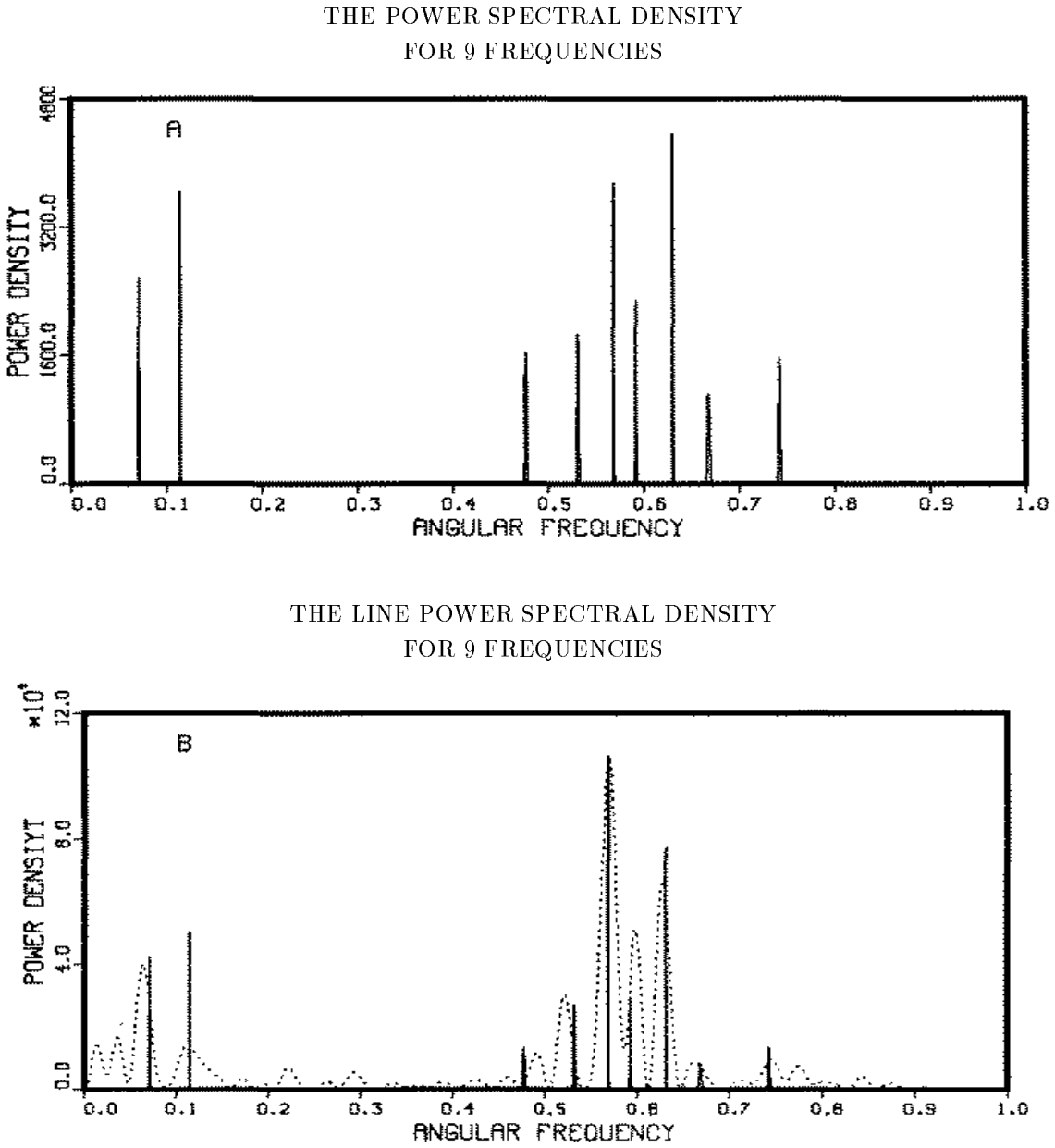
In Fig. 7.19(A), we can plot the simulated sunspot series. We have repeated the plot of the sunspot numbers, Fig. 7.19(B), for comparison. This simple nine-frequency model reproduces most of the features of the sunspot numbers, but there is still something missing from the model. In particular the data values drop uniformly to zero at the minima. This behavior is not repeated in the nine-frequency model.

Table 7.1: The Nine Largest Sinusoidal Components in the Sunspot Numbers

$\langle \hat{f} \rangle_{\text{est}}$	$\langle B_1 \rangle$	$\langle B_2 \rangle$	$\sqrt{B_1^2 + B_2^2}$
11.02 $\pm$ 0.01 years	-35	4.5	35
10.73 $\pm$ 0.03 years	1.0	19	19
9.98 $\pm$ 0.01 years	15	-10	18
88.08 $\pm$ 0.02 years	2.9	-17	17
53.96 $\pm$ 0.02 years	-10	-13	16
11.85 $\pm$ 0.01 years	-14	-2.2	14
48.44 $\pm$ 0.04 years	-9.8	-3.1	10
8.39 $\pm$ 0.03 years	-5.4	6.9	9
13.16 $\pm$ 0.03 years	4.7	-6.6	8

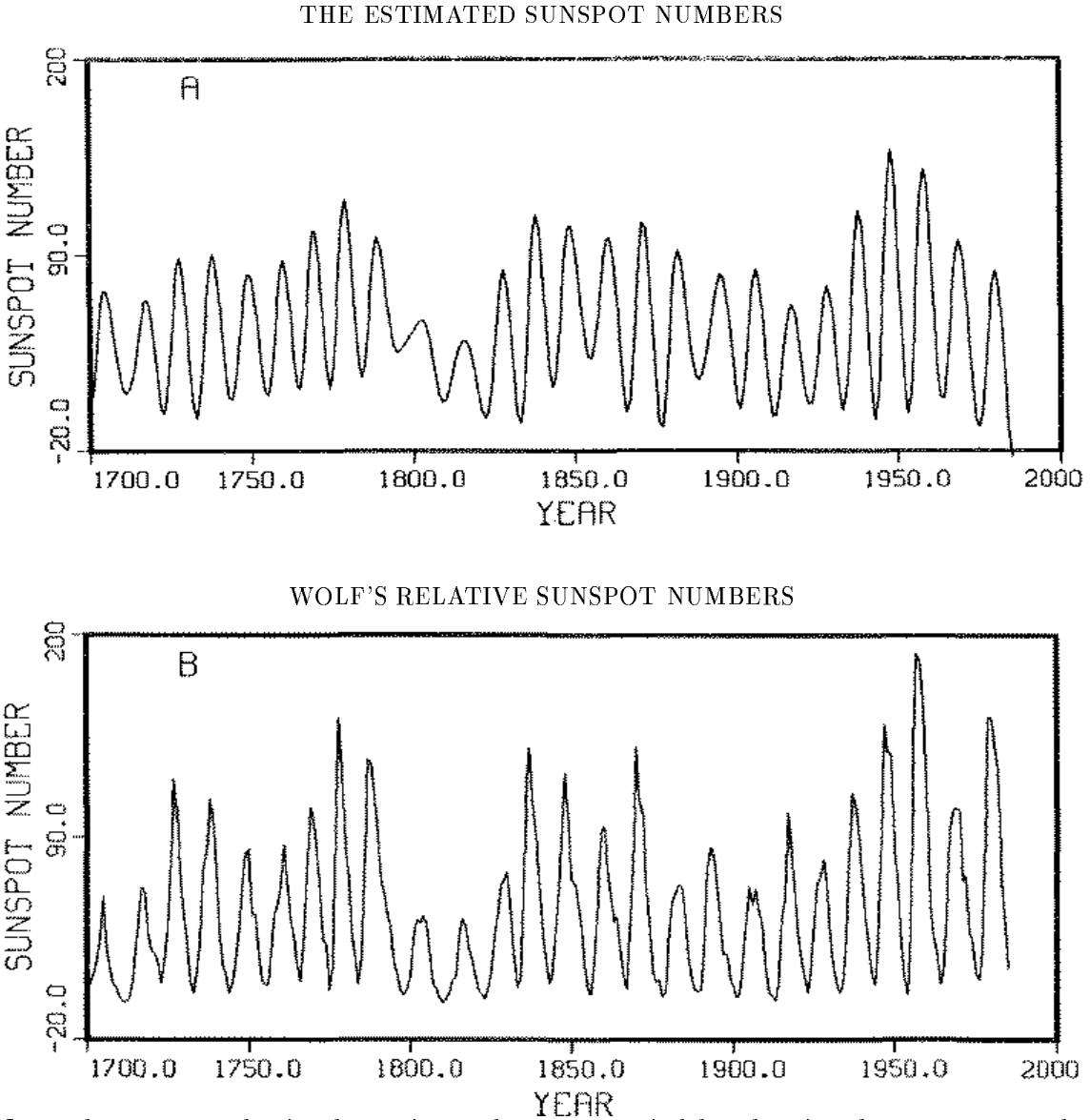
The first column is the frequency with an estimate of the variance of the posterior probability density; the second and third columns are amplitudes of the cosine and sine components and the last column is the magnitude of the signal. There are any number of effects in these data, but the largest is the 11 year cycle. We demonstrated in Section 6.1.4, page 76 that when the oscillations are nonharmonic the single frequency model can have spurious multiple peaks. It is only the largest peak in the marginal posterior probability density of a single harmonic frequency plus a constant that is indicative of the oscillation in the data. If we use a multiple harmonic frequency model, as we did here, probability theory will interpret these spurious peaks as frequencies. Probably all of the effects other than the 11 year cycle are artifacts of not knowing the correct model, which presumably involves nonsinusoidal and nonstationary oscillations.

Figure 7.18: The Posterior Probability of Nine Frequencies



The posterior probability of nine frequencies in the relative sunspot numbers (A), has nine well resolved peaks. In (B) we have a line spectral density. The peak value of the periodogram is an accurate estimate of the energy carried in a line as long as there is only one isolated frequency present.

Figure 7.19: The Predicted Sunspot Series



Not only can one obtain the estimated power carried by the signal, one can use the amplitudes to plot what probability theory has estimated to be the signal (A). We have included the relative sunspot numbers (B), for easy comparison.

Also, the data have sharper peaks than troughs, while our sinusoidal model, of course, does not. This is, as has been noted before, evidence of some kind of "rectification" process. A better model could easily reproduce these effects.

7.4.3 The Sunspot Numbers in Terms of Harmonically Related Frequencies

We used a harmonic model on the sunspot numbers so that a simple comparison to a model proposed by C. P. Sonett [38] could be done. He attempted to explain the sunspot numbers in terms of harmonic frequencies; 180, 90, and 45 are examples of harmonically related frequencies. In 1982, Sonett [38] published a paper in which the sunspot number spectrum was to be explained using

$$f(t) = [1 + \alpha \cos(\omega_m t)][\cos(\omega_c t) + \Delta]^2$$

as a model. Sonett's estimate of the magnetic cycle frequency ω_m is approximately 90 years, and his estimate of the solar cycle frequency ω_c is 22 years. The rectification effect is present here.

This model is written in a deceptively simple form and a number of constants (phases and amplitudes) have been suppressed. We propose to apply probability theory using this model to estimate ω_c and ω_m. To do this, we first square the term in brackets and then use trigonometric identities to reduce this model to a form in which probability theory can readily estimate the amplitudes and phases:

$$\begin{aligned}
f(t) = B_1 &+ B_2 \cos([\omega_m]t) &+ B_3 \sin([\omega_m]t) \\
&+ B_4 \cos([2\omega_m]t) &+ B_5 \sin([2\omega_m]t) \\
&+ B_6 \cos([\omega_c - 2\omega_m]t) &+ B_7 \sin([\omega_c - 2\omega_m]t) \\
&+ B_8 \cos([\omega_c - \omega_m]t) &+ B_9 \sin([\omega_c - \omega_m]t) \\
&+ B_{10} \cos([\omega_c]t) &+ B_{11} \sin([\omega_c]t) \\
&+ B_{12} \cos([\omega_c + \omega_m]t) &+ B_{13} \sin([\omega_c + \omega_m]t) \\
&+ B_{14} \cos([\omega_c + 2\omega_m]t) &+ B_{15} \sin([\omega_c + 2\omega_m]t) \\
&+ B_{16} \cos([2\omega_c - 2\omega_m]t) &+ B_{17} \sin([2\omega_c - 2\omega_m]t) \\
&+ B_{18} \cos([2\omega_c - \omega_m]t) &+ B_{19} \sin([2\omega_c - \omega_m]t) \\
&+ B_{20} \cos([2\omega_c]t) &+ B_{21} \sin([2\omega_c]t) \\
&+ B_{22} \cos([2\omega_c + \omega_m]t) &+ B_{23} \sin([2\omega_c + \omega_m]t) \\
&+ B_{24} \cos([2\omega_c + 2\omega_m]t) &+ B_{25} \sin([2\omega_c + 2\omega_m]t).
\end{aligned}$$

Now Sonett specifies the amplitudes of these, but not the phases [38]. We will take a more general approach and not constrain these amplitudes. We will simply allow probability theory to pick the amplitudes and phases which fit the data best. Thus any result we find will have the Sonett frequencies ω_m and ω_c, but the amplitudes and phases will be chosen in a way that will fit the data at least as well as does the Sonett model – possibly somewhat better. After integrating out the amplitudes we have only two parameters to determine, ω_c and ω_m.

We located the maximum of the posterior probability density using the computer code in Appendix E and the pattern search routine. The "best" estimated value for ω_c (in years) is approximately 21.0 years, and for ω_m approximately 643 years. The values for these parameters given by Sonett are $\omega_c = 22$ years and $76 < \omega_m < 108$ years with a mean value of $\omega_m \approx 89$ years. Our probability analysis estimates the values of ω_c to be about the same, and ω_m to be substantially different, from those given by Sonett. The most indicative value is the estimated standard deviation for this model: $\sigma_{\text{Sonett}} = 25.5$ years. By this criterion, this model is no better than a four-frequency model. Considering that a four-frequency model has 15 degrees of freedom compared to 29 for this model, we can all but exclude harmonically related frequencies as a possible explanation of the sunspot numbers. Of course, these conclusions refer only to an analysis of the entire run of data; if we considered the first century of the record to be unreliable and analyzed only the more recent data, a different conclusion might result.

7.4.4 Chirp in the Sunspot Numbers

We have so far investigated two variations of harmonic analysis of the relative sunspot numbers. Let us proceed to investigate a more complex case to see whether there is more structure in the relative sunspot numbers than just simple periodic behavior. These data have been looked at from this standpoint at least once before. Bracewell [39] has analyzed these numbers to determine whether they could have a time-dependent "instantaneous phase". The model used by Bracewell can be written as

$$f(t) = B_1 + \text{Re}\left[E(t)\exp(i\phi(t) + i\omega_{11}t)\right]$$

where B_1 is a constant term in the data, $E(t)$ is a time varying magnitude of the oscillation, $\phi(t)$ is the "instantaneous phase", and ω_{11} is the 11 year cycle.

This model does not incorporate any prior information into the problem. It is so general that any function can be written in this form. Nevertheless, the idea that the phase $\phi(t)$ could be varying slowly with time is interesting and worth investigating.

An "instantaneous phase" in the notation we have been using is a chirp. Let $\phi(t)$ stand for the phase of the signal, and ω its frequency. Then we may Taylor expand $\phi(t)$ around $t = 0$ to obtain

$$\omega t + \phi(t) \approx \phi_0 + \omega t + \frac{\phi''}{2} t^2 + \cdots,$$

where we have assumed $\phi'(t) = 0$. If this were not so then ω is not the frequency as presumed here. The Bracewell model can then be approximated as

$$f(t) = B_1 + E(t)[\cos(\omega t + \alpha t^2) + B_2 \sin(\omega t + \alpha t^2)].$$

Thus, to second order, the Bracewell model is just a chirped frequency with a time varying envelope.

We can investigate the possibility of a chirped signal using

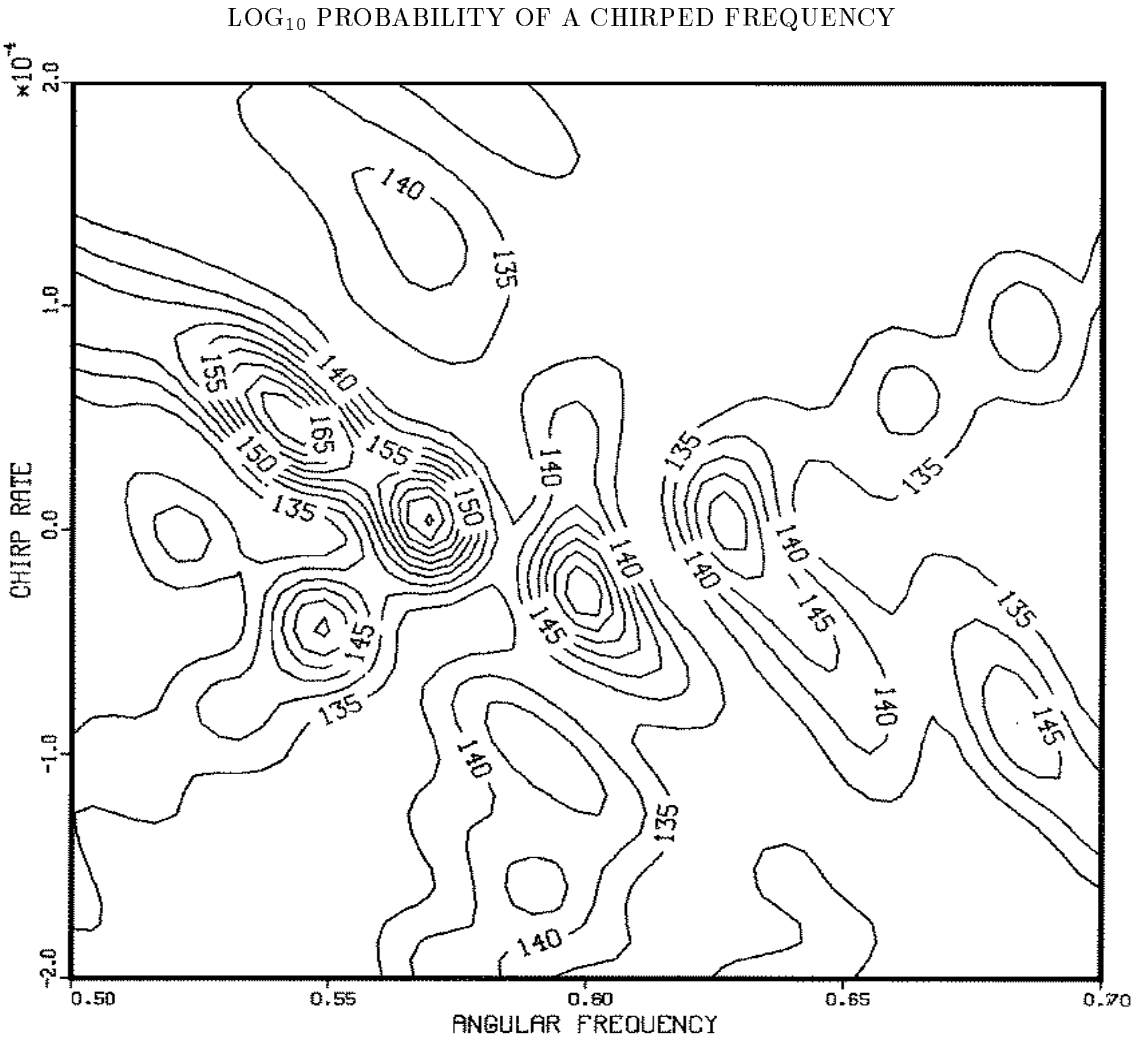
$$f(t) = C_1 + C_2 \cos(\omega t + \alpha t^2) + C_3 \sin(\omega t + \alpha t^2)$$

as the model, where α is the chirp rate, C_1 is a constant component, ω is the frequency of the oscillation, and C_2 and C_3 are effectively the amplitude and phase of the oscillation. This model is not a substitute for the Bracewell model. Instead this model is designed to allow us to investigate the possibility that the sunspot numbers contain evidence of a chirp, or "instantaneous phase" in the Bracewell terminology.

A plot of the log of the "Student t-distribution" using this model is the proper statistic to look for chirp. However, we now have two parameters to plot, not one. In Fig.7.20 we have constructed a contour plot around the 11 year cycle. We expect this plot to have a peak near the location of a frequency. It will be centered at zero chirp rate if there is no evidence for chirp, and at some nonzero value when there is evidence for chirp. Notice, that along the line $\alpha = 0$ this "Student t-distribution" is just the simple harmonic probability distribution studied earlier, see Fig. 2.1(A). As with the Fourier transform if there are multiple well separated chirped frequencies (with small chirp rates) then we expect there to be multiple peaks in Fig. 7.20.

There are indeed a number of peaks; the single largest point on the plot is located off the $\alpha = 0$ axis. The data contain evidence for chirp. The low frequencies also show evidence for chirp. To the extent that the Bracewell "instantaneous phase" may

Figure 7.20: Chirp in the Sunspot Numbers?



To check for chirp we take $f(t) = A_1 + A_2\cos(\omega t + \alpha t^2) + A_3\sin(\omega t + \alpha t^2)$ as the model. After integrating out the nuisance parameters, the posterior probability is a function of two variables, the frequency ω and the chirp rate α. We then plotted the $\log_e$ of the posterior probability. The single highest peak is located at a positive value of α: there is evidence of chirp.

be considered as a chirp, we must agree with him: there is evidence in these data for chirp.

In light of this discussion, exactly what these numbers represent and exactly what is going on inside the sun to produce them must be reconsidered. The orthogonal expansion on these numbers indicates that the complexity of these numbers is immense and no simple model will suffice to explain them. Given the total number of degrees of freedom it is likely that every cycle has a unique amplitude and a complex non-sinusoidal shape. In other words, different sunspot cycles are about as complicated in structure as are different business cycles in economic data. If that is true, the only frequency in these data of any relevance is probably the 11 year cycle; the other indications of frequency are just effects of the nonharmonic oscillation. Again, had we analyzed only the more recent data, the conclusions might have been different. Certainly we have not answered any real questions about what is going on; indeed that was not our intention. Instead we have shown how use of probability theory for data analysis can facilitate future research by testing various hypotheses more sensitively than could the traditional intuitive *ad hoc* procedures.

7.5 Multiple Measurements

The traditional way to analyze multiple (i.e. multi-channel) measurements is to average the data, and then analyze the averaged data. The hoped-for improvement in the parameter estimates is the standard $\sqrt{n}$ rule. To derive this rule one must assume that the signal and the noise variance, are the same in every data set, and that the noise samples were uncorrelated. Unfortunately, the conditions under which averaging works at its theoretical best are almost never realized in real experiments. Specifically, all experiments contain some effects which will not average out. These effects can become so significant, that the evidence for the signal can be greater in any one of the data sets that went into the average than it is in the averaged data (we will demonstrate this shortly). There are three main reasons why averaging may fail to give the expected $\sqrt{n}$ improvement in the parameter estimates: the experiment may not be reproducible, the model may incorrect, the noise within different data sets may be correlated.

In real physics experiments, reproducibility depends critically on the electronics repeating itself exactly every time. Of course this never happens; there are always

small differences. For example, to repeat an NMR experiment one must bring the sample to a stationary state (this may be far from equilibrium) and then further excite the sample using a high power radio transmitter. In a perfect world, every time one excited the sample it would be with a pulse of exactly the same amplitude and exactly the same shape as before. Of course this never happens; every repetition is a unique experiment having slightly different amplitudes, phases, and noise variance. These slight differences are enough to cause averaging to fail to give the $\sqrt{n}$ improvement when large numbers of data sets are averaged, even when the noise samples are independent.

The second source of systematic error is in our imprecise knowledge of the model. If the signal is exactly the same in each data set, of course the noise is reduced by averaging. Unfortunately, if we do not know the model exactly, then our model is only an approximation. When we fit the model to the data, some of the "true" signal will not be fit. This misfit of the model will be called noise by probability theory. But it is noise that is perfectly correlated in successive data sets, and does not average out. Thus the accuracy estimates will not improve, because the dominant contribution to the estimated noise variance will be the misfit between the model and the data.

If any systematic effect is present, averaging will fail to give the expected improvement; nevertheless, probability theory does not mislead us. We have stressed several times that the estimates one obtains from these procedures are conservative. That is, when the models misfit the data they still give the best estimates of the parameters possible under the circumstances, and yield conservative (wide) error estimates. This suggests that by analyzing each data set separately, and looking for common effects, we might be able to realize better estimates than by averaging. In this section we investigate the effects of multiple measurements and compare the results of a joint analysis (analyzing all of the data) to the analysis of the averaged data. We will do this analysis on a data set that most people would not hesitate to average. This is our first example where we apply Bayesian analysis to data which are not a time series.

The experiment we will consider is a simple diffraction experiment. A mercury vapor lamp was placed in front of a slit, and the light from the lamp passed through the slit and onto a screen. An electronic camera (a Charge Coupled Device – CCD) was placed behind the screen and used to image the intensity variations. The data for this analysis were kindly provided by W. H. Smith [41]. The image consists of a series of light and dark bars typical of such experiments. The pattern for the first row in the CCD is shown in Fig. 7.21(A). Figure 7.21(B) is a plot of the averaged

data.

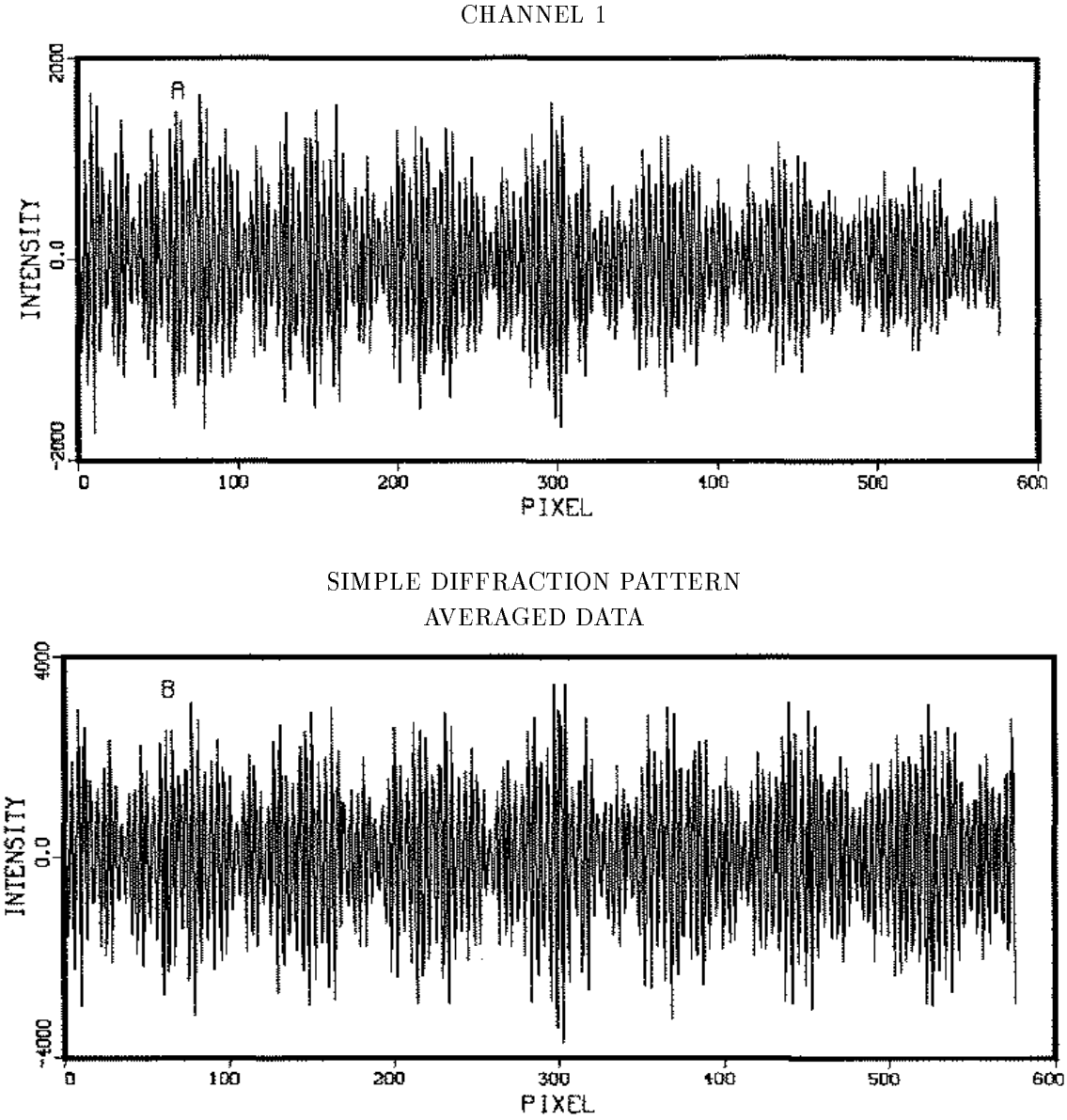
This particular device has 573 pixels in each row and there are 380 rows. Thus there are a total of 380 repeated measurements. The CCD was aligned parallel with the slit, so the image appearing in each row should be identical. In principle the data could be averaged to obtain a $\sqrt{380}$ improvement in the parameter estimates. However, the concerns mentioned earlier are all applicable here. The types of systematic effects that can enter this experiment are numerous, but a few of them are: the camera readout has small systematic variations from one row to the next; there can be intensity variations from the first to last row; and if the alignment of the camera is not perfect, there will be small phase drifts from the first to last row. Nonetheless, when one looks at these data, there is absolutely no reason to believe that averaging should not work.

We begin the analysis by plotting the $\log_{10}$ of the probability of a single harmonic frequency plus a constant. We plot this probability density for the first row of the CCD in Fig. 7.22(A), for the average of the 380 data sets in Fig. 7.22(B), and jointly for all data Fig. 7.22(C). One sees from the average data, Fig. 7.22(B), that there is indeed large evidence for a frequency near 1.6 in dimensionless units. That peak is some 133 orders of magnitude above the noise level. The second thing that one sees is that the peak from the first row of the CCD, Fig. 7.22(A), is some 136 orders of magnitude above the noise: the peak from one row has more evidence for frequencies than the average data! The third plot, Fig. 7.22(C), is from the joint analysis. That peak is some 55,000 orders of magnitude higher than the average data! The implications of this are indeed staggering. If one cannot average data in an experiment as simple as this one, then there are probably no conditions under which averaging is the way to proceed. Because the issues raised by this simple example are so important, we will pause to investigate some of the theoretical implications before proceeding with this example.

7.5.1 The Averaging Rule

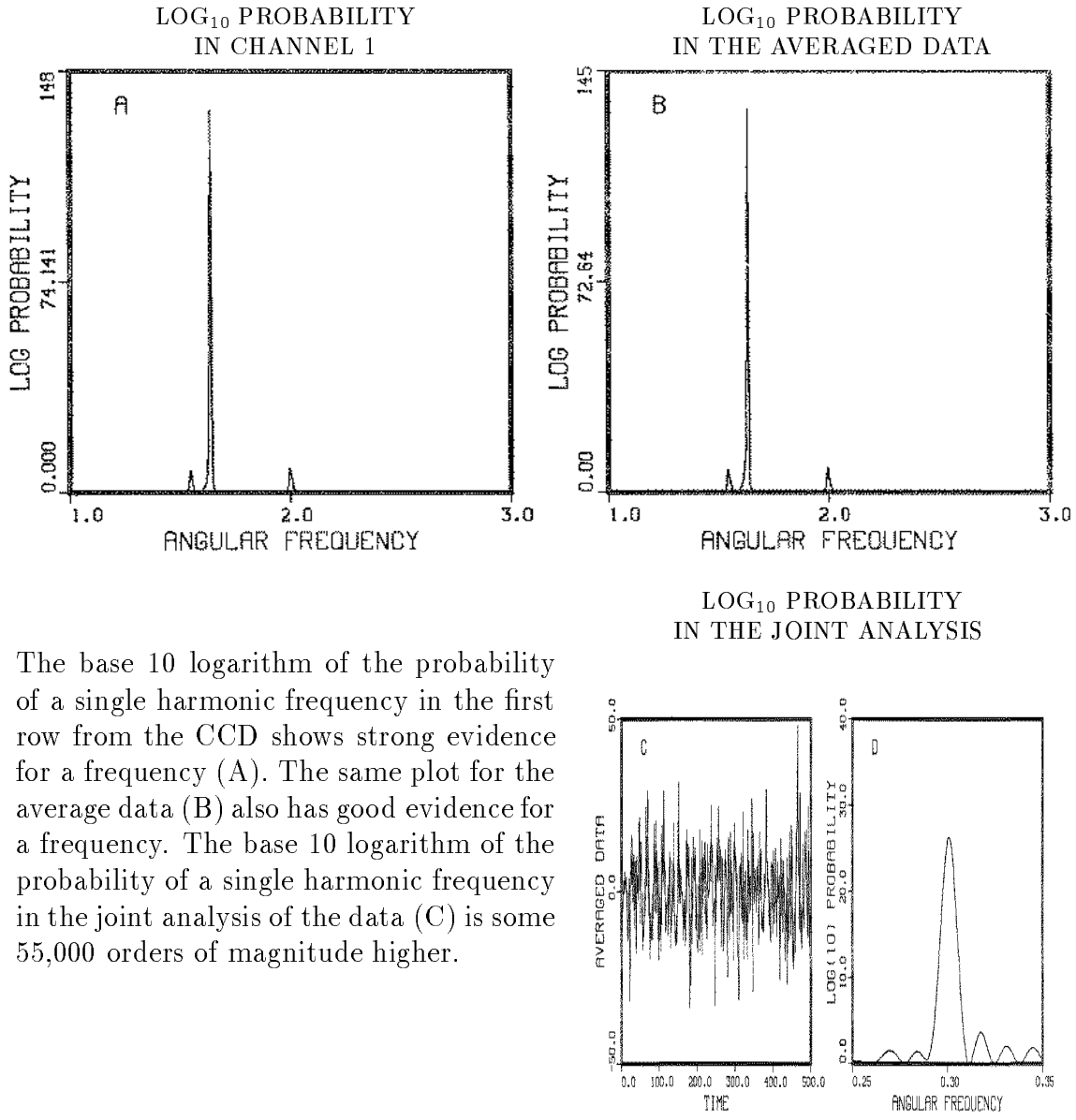
To derive the average rule one assumes a signal $f(t)$, and n sets of data $d(t_i)_j$ with noise variance σ^2. The signal $f(t)$ and the noise variance σ^2 are assumed to be the same in every data set. Then the probability that we should obtain a data set $d(t_i)_j$

Figure 7.21: A Simple Diffraction Pattern



An image was formed by placing a mercury vapor lamp behind a slit and allowing its light to shine through a slit and onto a screen, the CCD imaged the screen. The first row from the CCD is shown in (A). This particular CCD was 573 by 380 pixels, so there are 380 channels. The averaged data is shown in (B). The expected improvement in resolution is $\sqrt{380}$. However, if there are systematic errors in the data, the actual improvement realized will be less.

Figure 7.22: Log_{10} Probability of a Single Harmonic Frequency



The base 10 logarithm of the probability of a single harmonic frequency in the first row from the CCD shows strong evidence for a frequency (A). The same plot for the average data (B) also has good evidence for a frequency. The base 10 logarithm of the probability of a single harmonic frequency in the joint analysis of the data (C) is some 55,000 orders of magnitude higher.

is given by

$$P(D_j|f,\sigma,I) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d(t_i)_j - f(t_i))^2\right\}.$$

If the noise samples in different channels are independent, the probability that we should obtain all the data sets is just the product of the probabilities that we should obtain any one of the data sets:

$$P(D|f,\sigma,I) \propto \Pi_{j=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d(t_i)_j - f(t_i))^2\right\}.$$

This can be written as

$$P(D|f,\sigma,I) \propto \exp\left\{-\frac{n}{2\sigma^2}\sum_{i=1}^{N}[\overline{d(t_i)^2} - 2\overline{d(t_i)}f(t_i) + f(t_i)^2]\right\},$$

where $\overline{d(t_i)^2}$ is the mean square data value at time t_i, and $\overline{d(t_i)}$ is the mean data value, where "mean" signifies "average over the channels". This is almost a standard model-fitting problem with the data replaced by the average. The procedure is called "Brute Stacking" by geophysicists. The improvement comes from the factor of n multiplying the term in square brackets. We demonstrated that the accuracy estimates are all proportional to the square root of the variance, and here the variance is effectively σ^2/n – this gives the standard $\sqrt{n}$ improvement.

7.5.2 The Resolution Improvement

When multiple data sets are analyzed jointly, how much improvement in the parameter estimates can be expected? The resolution improvement depends on the curvature of the posterior probability density at the maximum. The general rule depends on the model; all we can say is that if all of the data sets have approximately the same evidence in them, then the logarithm of the posterior probability density is n times larger and something like the $\sqrt{n}$ improvement will be realized. We can demonstrate this for the single frequency estimation problem. Then the posterior probability of the frequency, given n repeated measurements and assuming the noise variance σ^2 is the same, is

$$P(\omega|\sigma,D,I) \approx \exp\left\{\sum_{j=1}^{n}\frac{C(\omega)_j}{\sigma^2}\right\} \tag{7.4}$$

where $C(\omega)_j$ is the Schuster periodogram evaluated for data set j. To obtain the accuracy estimates we expand the exponent about the "true" frequency $\hat{\omega}$ to obtain

$$P(\omega|D,I) \approx \exp\left\{-\sum_{j=1}^{n}\frac{b_j(\hat{\omega}-\omega)^2}{2\sigma^2}\right\},$$

where $b_j = -C_j''$ for the jth data set. If the data contain a single sinusoid such as $\hat{B}\cos(\hat{\omega}t)$, then b is given by Eq. (2.10). This gives the posterior probability density for a simple harmonic frequency when multiple measurement are present as

$$P(\omega|D,I) \approx \exp\left\{-\sum_{j=1}^{n}\frac{N^3\hat{B}_j^2}{96\sigma^2}(\hat{\omega}-\omega)^2\right\}.$$

The accuracy estimate is given by

$$(\omega)_{\text{est}} = \hat{\omega} \pm \sqrt{\frac{48\sigma^2}{N^3 n\overline{B^2}}}$$

where $\overline{B^2}$ is the mean square of the true amplitude. If all of the amplitudes are nearly the same height, this is just the standard $\sqrt{n}$ improvement.

The improvement realized is directly related to how well the assumptions in the calculation are met. In the case of averaging, the assumptions are that the amplitude, frequency, phase, and noise variance are the same in every data set. For the example just given we removed the assumption that the amplitudes had to be the same in every data set, consequently, $n\hat{B}^2$ was replaced by $n\overline{\hat{B}^2}$. If we further remove the assumption that the noise variance is the same in every data set, then $n\overline{\hat{B}^2}$ is replaced by $\sum_{j=1}^{n}\hat{B}_j^2/\sigma_j^2$.

When the assumed conditions are not met, the price one pays is in resolution. The procedure described by Eq. (7.2) is more general than averaging in that it allows the amplitudes, phases, and noise variance to be different for each data set and still allows one to look for common effects. When the true amplitudes, phases, and noise variance are the same in every data set this procedure reduces to averaging. Thus Eq. (7.2) represents a more conservative approach than averaging and will realize something approaching the $\sqrt{n}$ improvement under a wider variety of conditions, because it makes fewer assumptions.

7.5.3 Signal Detection

When multiple measurements are present we would like to understand what happens to the joint analysis as we increase the number of measurements. In other words

we typically average data when the signal-to-noise ratio is very bad. We do this because we think it allows one to reduce the noise; thus small signals can be detected. But what will happen with the joint analysis? The general answer for the joint analysis again depends on the model function. However, as noted earlier, if the evidence in each data set is roughly the same the sufficient statistic will be n times larger than when only a single measurement is present. Thus the evidence for a signal will build up in a manner similar to averaging. We will demonstrate how the evidence accumulates in the joint analysis using the single frequency model. The posterior probability density of a common harmonic frequency, when multiple measurements are available, is again given by Eq. (7.4). What we would like to know is how high is the peak in Eq. (7.4)? Again taking the data to be

$$d(t)_j = \hat{B}_j \cos(\hat{\omega} t)$$

the peak value of the periodogram is given approximately by

$$C(\hat{\omega})_j \approx \frac{N\hat{B}_j^2}{4}.$$

Assuming each data set has the same number of data values N, the maximum of the posterior probability density will be

$$P(\hat{\omega}|D,I) \propto \exp\left\{\sum_{j=1}^{n} \frac{N\hat{B}_j^2}{4\sigma^2}\right\}.$$

We can simplify this by using

$$\sum_{j=1}^{n} \hat{B}_j^2 = n\overline{B^2}$$

where $\overline{B^2}$ is the mean-square true amplitude. The evidence for a signal increases by the power of the number of data sets:

$$P(\hat{\omega}|D,I) \propto \exp\left\{\frac{nN\overline{B^2}}{4\sigma^2}\right\}.$$

If we allow the variance of the noise to be different in each data set $n\overline{B^2}$ will be replaced by a weighted average $\sum_{j=1}^{n} \hat{B}_j^2/\sigma_j^2$. Again we find the height of the posterior probability density depends directly on the assumptions made in the calculation. In the case of averaging, the log-height of the posterior probability density is n times larger than the height from one data set (provided the assumptions are met). If we relax the assumptions about the amplitude B, we replace B^2 in the average rule by

the mean-square true value. If we further relax the assumptions and allow the noise variances to be different, we obtain the weighted average of the true B^2 values. Thus we again have a more conservative procedure that will reduce to give the $\sqrt{n}$ rule when the appropriate conditions are met, under much wider conditions than averaging. In the case of the simple harmonic frequency, doubling the number of data sets is similar to doubling the number of time samples, while keeping the total sampling time fixed. This is not the best way to find a signal (doubling the signal-to-noise works better), but if no other course is available it will build up the probability density by essentially squaring the distribution for each doubling of the number of data sets.

7.5.4 The Distribution of the Sample Estimates

In the CCD example, we had 380 repeated measurements, and the maximum of the posterior probability was some 55,000 orders of magnitude above the noise. Each data set raised the posterior probability approximately 55,000/380 = 144 orders of magnitude. But when the data were averaged, small variations in the amplitude, phases, and variance of the noise caused systematic variations in the data which were nonsinusoidal. Probability theory automatically reduced both the height of the posterior density (i.e. it could not see the signal as well) and reduced the precision of the estimates. In this example the height was reduced from 55,000 to 133, and the accuracy was reduced from 6.8×10^{-8} to 0.00036; the error estimate of the averaged data is 5266 times larger than the estimate from the joint analysis. It thus appears that data averaging (Brute Stacking) is never better than a joint analysis of the data, and it is in general worse. Averaging does, of course, reduce the amount of computation; but with modern computers this is not an important consideration.

In this last example the improvement was very dramatic, but this was real experimental data; perhaps the reason averaging failed was some other effect in the data. We would like to show that the cause was the variation of the signal and the noise variance in the various data sets. To do this we will generate data from the following equation

$$d(t) = B\cos(0.3t + \theta) + e(t). \tag{7.5}$$

We will vary B, θ, and σ from one data set to another. We will then estimate the frequency in the averaged data and in a joint analysis of all the data, and show that these variations will cause averaging to fail to give the $\sqrt{n}$ improvement; while a joint analysis will continue to exhibit the expected improvement.

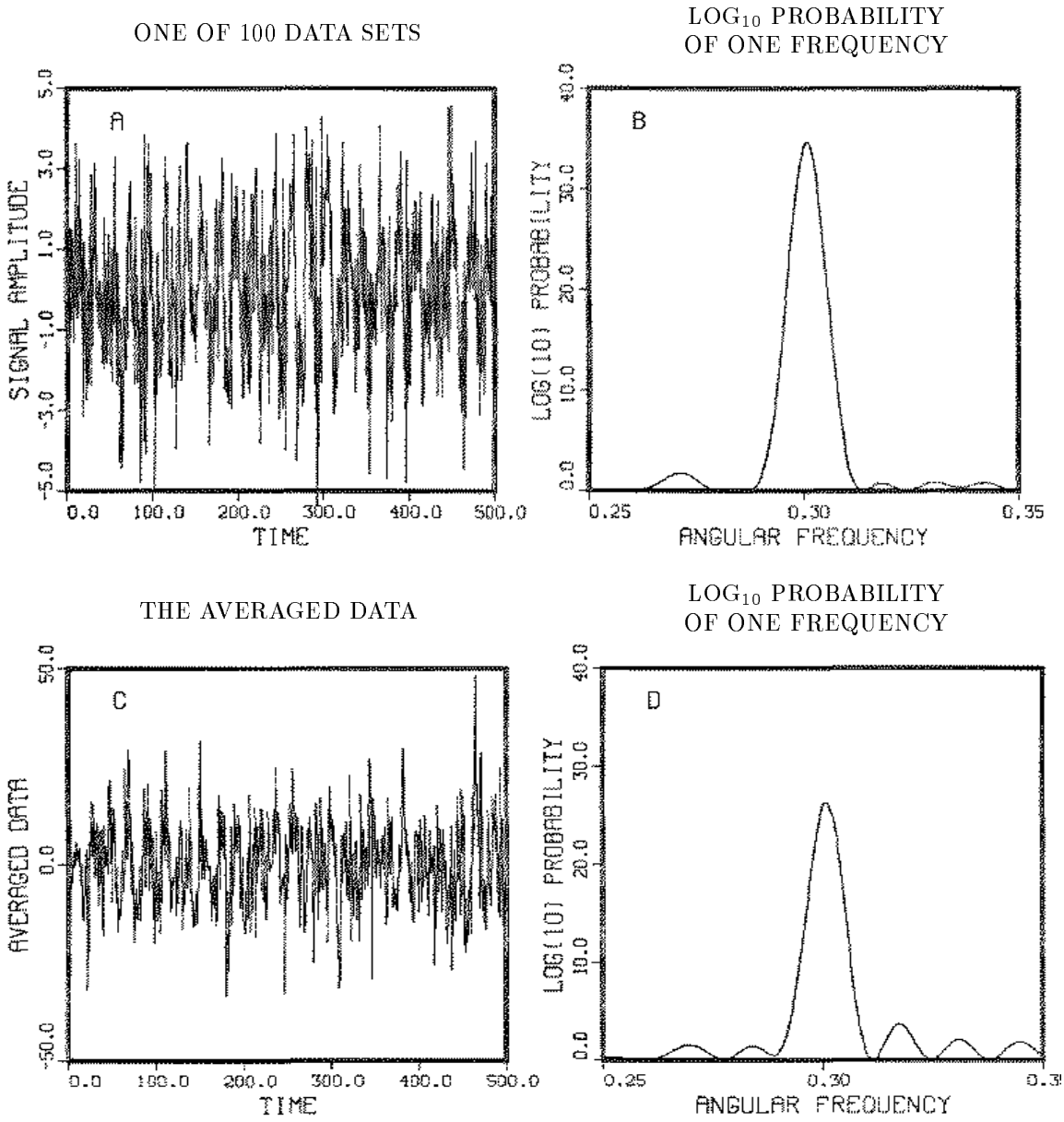
We generated multiple data sets from Eq. (7.5). Each data set we generated had a different amplitude B, phase θ, and noise variance σ^2. To generate the data we used three Gaussian random numbers with unit variance. We used one as the amplitude B, another as the phase θ, and the third to scale the noise. The noise was generated using a Gaussian random number generator with unit variance. We generated the noise and then multiplied it by the third random number. Using this procedure, the signal will average to zero.

We generated 100 data sets, each containing 512 data values. An example of one such data set is shown in Fig. 7.23(A). The $\log_{10}$ of the probability of a single harmonic frequency is shown in Fig. 7.23(B). We have also displayed the average data Fig. 7.23(C) and the $\log_{10}$ probability of a single frequency. In this particular case averaging has not completely cancelled the signal, however, one measurement, Fig. 7.23(A), has about a 10^9 times more evidence for a signal than the average data, Fig. 7.23(B). We estimated the frequency in the 100 data sets in a joint analysis and in the averaged data. We then selected three new random numbers and repeated this calculation 3000 times.

The results are summarized in Table 7.2. This table contains the actual estimates from a few of the 3000 sets of data analyzed. The second column is the estimated frequency from the average data. The third column is the squared difference between the true frequency and the estimate from the averaged data, (the variance of this estimate). The fourth column is the estimated frequency from the joint analysis, and the fifth column is the squared difference between the true frequency and the estimated frequency from the joint analysis. We averaged all 3000 entries (labeled average at the bottom of the table), and we computed the square root (the standard deviation) estimate for the variance (columns 3 and 5). The estimate from the averaged is a little better than it actually was in these data. When we estimated the frequency we had to give the search routine an initial estimate of the frequency. This locked the search routine onto the correct peak in the periodogram even though there was no clear peak in many of the data sets. This is analogous to estimating the averaged frequency with a strong prior.

For a single data set with unit signal-to-noise the "best" estimated frequency should be ± 0.0006 radians per step; the estimated standard deviation of the averaged data is about a factor of 2 larger than what one would obtain from one data set. Thus averaging has destroyed evidence in the data: any one data set contains more evidence for frequencies than the averaged data. If averaging were working

Figure 7.23: Example – Multiple Measurements



We generated 100 such data sets with different amplitude, phase, and noise variance, but the same frequency (see text for details). In (A) we have displayed one such data set. The $\log_{10}$ of the probability of a single harmonic frequency is displayed in (B). The average data (C) and the $\log_{10}$ probability of a single frequency in (D) show 10^{-9} times less evidence for a harmonic frequency than one data set.

Table 7.2: "Brute Stacking" vs. Joint Analysis

	Frequency Estimate Average	$(\langle\omega\rangle - 0.3)^2$	Frequency Estimate Joint Analysis	$(\langle\omega\rangle - 0.3)^2$
1	0.30018	3.52^{-8}	0.2999974	6.81^{-12}
2	0.29863	1.87^{-6}	0.2999896	1.08^{-10}
3	0.29987	1.47^{-8}	0.2999782	4.75^{-10}
4	0.30076	5.79^{-7}	0.2999995	1.64^{-13}
5	0.30044	2.00^{-7}	0.2999881	1.42^{-10}
6	0.29804	3.82^{-6}	0.2999998	2.78^{-14}
7	0.30024	5.77^{-8}	0.2999995	1.64^{-13}
8	0.30047	2.22^{-7}	0.2999985	2.18^{-12}
9	0.29966	1.09^{-7}	0.2999985	2.18^{-12}
10	0.29990	8.30^{-9}	0.3000088	7.90^{-11}
⋮	⋮	⋮	⋮	⋮
2999	0.30152	2.31^{-6}	0.2999969	9.16^{-12}
3000	0.29968	1.02^{-7}	0.3000008	7.16^{-13}
Average	0.29999	1.39^{-6}	0.2999995	3.02^{-11}
SD		1.17^{-3}		5.49^{-6}

We generated 3000 frequency estimates (see text for details). The frequency estimate in the second column was from the averaged data, and the third column is the variance for that estimate. The fourth and fifth columns are the estimates from the joint analysis. The row labeled "Average" is the average of the 3000 frequency and variance estimates. The last row is the square root of the average variance. Averaging actually appears a little better than it is in these data: when we estimated the frequency we had to supply an initial frequency estimate; this locked the search routine onto the correct peak in the periodogram, even though there was no clear peak above the noise in many of the averaged data sets. From a probability standpoint this is analogous to estimating the average frequency with a strong prior.

at its theoretical best we would expect that in 100 data sets the estimates should improve to $\pm 0.0006/\sqrt{100} = \pm 0.00006$ radians per step. Averaging is a factor of 20 times worse than it should be. But how has the joint analysis done? For 3000 such estimates (in unit signal-to-noise) the joint analysis can do no better than the $\sqrt{n}$ rule: $\pm 0.00006/\sqrt{3000} \approx \pm 0.000001$ radians per step, where we find ± 0.000005, about a factor of 5 larger; we conclude that the mean square weighted amplitude was about 1/25. The joint analysis has performed well; about $0.001/0.000005 = 185$ times better than averaging.

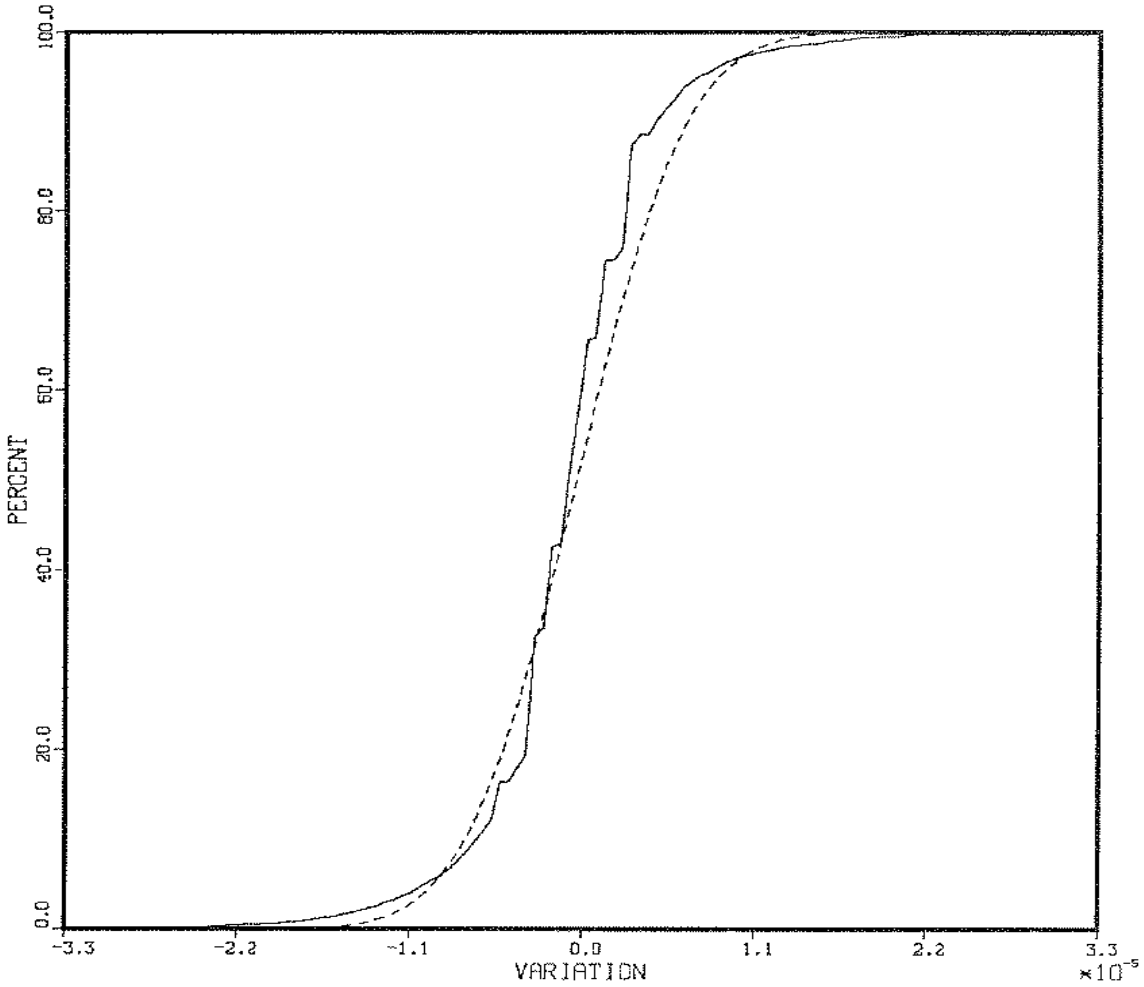
From the 3000 frequency estimates we computed a cumulative distribution of the number of estimates within one standard deviation, two deviations etc. of the true This distribution is displayed in Fig. 7.24. The solid line is the cumulative sample distribution, while the dashed line is the equivalent plot for a Gaussian having the same mean and standard deviation as the sample. The estimates resemble a Gaussian, but there are systematic differences. These differences are numerical in origin. We had to locate the maximum of the posterior probability, and this distribution is roughly 100 times more sharply peaked than a discrete Fourier transform. The pattern search routine we used moves the frequency by some predefined fixed amount. Typically it will move the frequency by only one or two steps, this tends to bunch the estimates up into discrete categories. We could fix this problem at the cost of much greater computing time.

7.5.5 Example – Multiple Measurements

We started this section by presenting a simple diffraction experiment and became sidetracked by some of the implications of the example. When we computed the sufficient statistic of the joint analysis we found the peak to be some 55,000 orders of magnitude higher than the peak for the averaged data. We have used the estimate of the frequency from that peak in several places; here, we plot the results of that analysis to give a better understanding of the determination of the frequency. We will estimate the frequency from the periodogram of the averaged data, from the "Student t-distribution" using the averaged data, and last using a joint analysis on all of the data.

The results of this analysis are displayed in Fig. 7.25. The normalization on all of these curves is arbitrary. If we took the periodogram of the averaged data as our frequency estimate we would have the broad peak in Fig. 7.25. However, probability

Figure 7.24: The Distribution of Sample Estimates

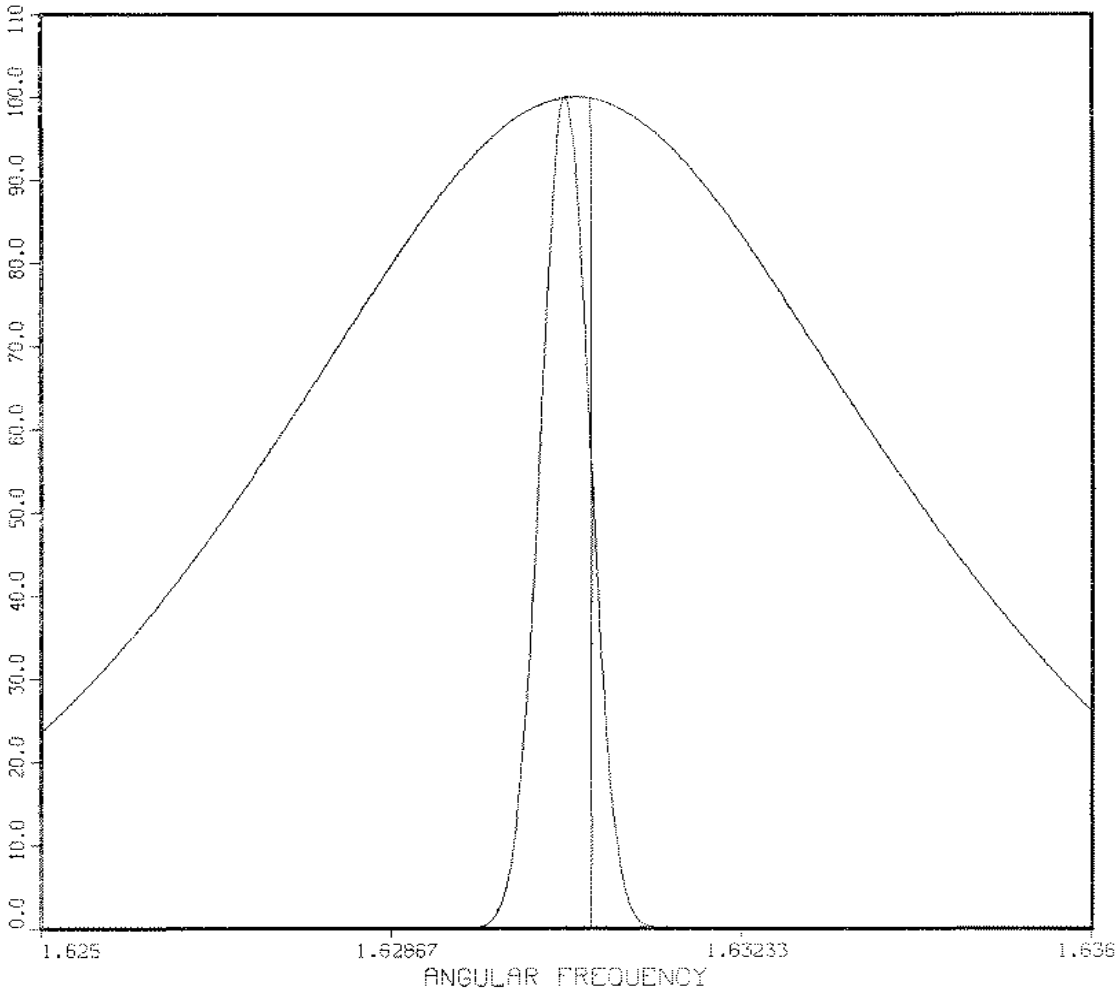


We generated 100 data sets with different amplitude, phase, and noise variance but the same frequency. From the 100 data sets we estimated the frequency. We then generated another 100 data sets and estimated the frequency. We repeated this process some 3000 times. Here we have plotted the cumulative percentage of estimates (solid line) falling within one, two, and three RMS standard deviations. The dashed line is the equivalent distribution for a Gaussian. The axis labels here correspond to two, four, and six standard deviations.

theory applied to the averaged data would narrow that peak by another factor of 10. The resulting posterior distribution is displayed as a sharp Gaussian inside the periodogram. We then estimated the frequency from all of the data using a joint analysis on all 380 data sets. The resulting posterior distribution is displayed as a Gaussian centered at the estimated frequency and having the same variance as our estimate. This is what appears as the vertical line just to the right of the Gaussian from the averaged data. From this we see that the joint analysis estimates the frequency much more precisely than does the analysis of the averaged data, and it estimates it to be rather different from that of the averaged data.

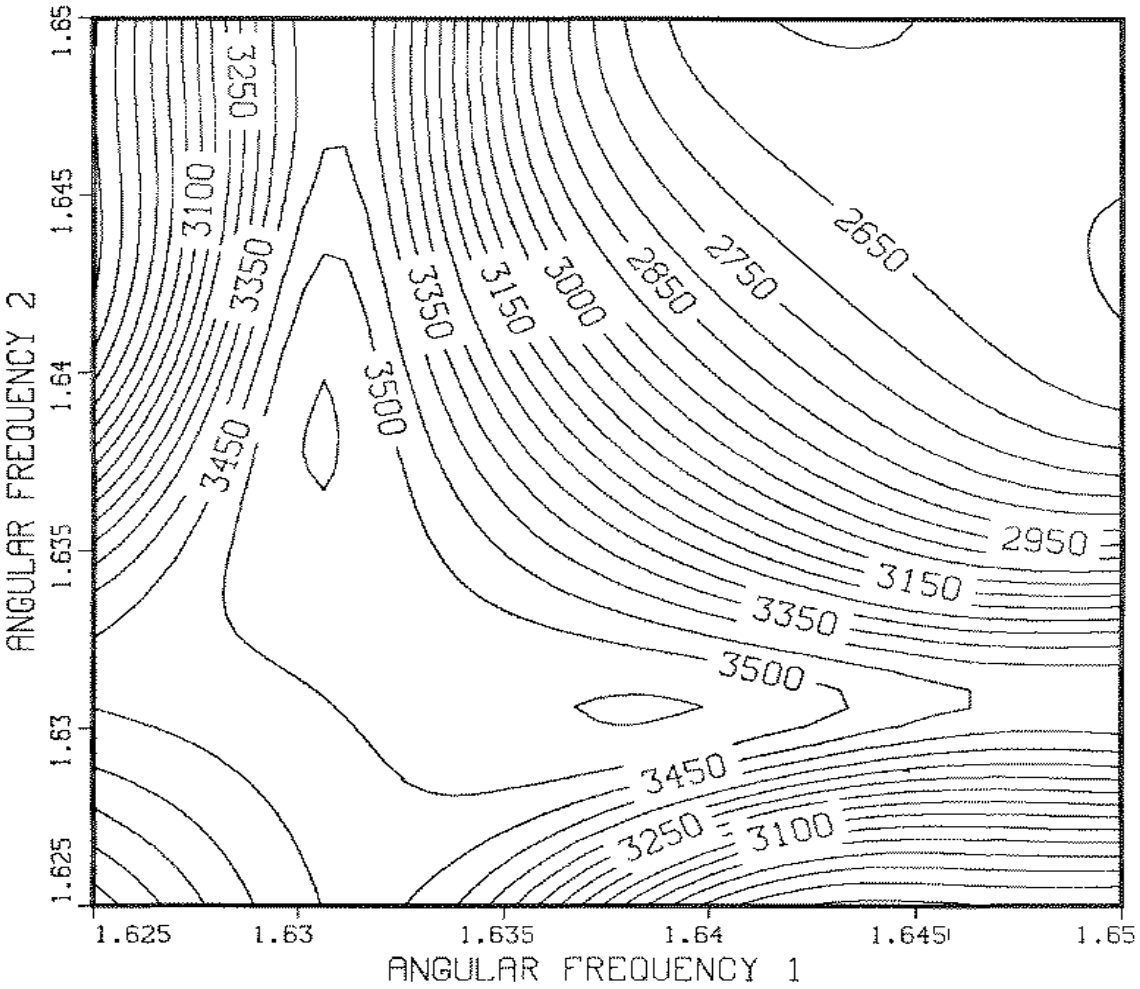
Before leaving this example we would like to apply one more simple analysis to these data. It is true that the small peaks to either side of the main peak in Fig. 7.22 are indications of frequencies. But could there be more that one frequency in the main peak? The spectrum of mercury has many lines in this main peak. Can we see them in these data? To determine whether the main peak has indications for more than one frequency, we compute the probability of two frequencies in this region using only the first five rows from the CCD (we used only the first five rows to reduce computation time). We plotted this as the contour plot in Fig. 7.26. If there is only one frequency present, we expect there to be two ridges in this plot, one extending horizontally and one vertically. On the other hand if there is more than one frequency in these data, there will be a small peak just to one side of $\omega_1 \approx \omega_2$. We see from Fig. 7.26 that there is indeed strong evidence of two frequencies in the main peak. This reinforces one of the things we noted earlier: What one can learn from a data set depends critically on what question one asks. R. A. Fisher once said "let the data speak for themselves". It appears that the data are more than capable of this, but they do not speak spontaneously; they need someone who is willing to ask the right questions, suggested by cogent prior information.

Figure 7.25: Example - Diffraction Experiment



The broad curve on this graph is the periodogram from the averaged data. The sharp Gaussian inside this line is the posterior distribution obtained from the averaged data. The sharp spike located just off-center is the Gaussian representing the posterior distribution from all 380 data sets.

Figure 7.26: Example - Two Frequencies



We examined the main peak in the joint analysis to see if there is any evidence for multiple frequencies. The results are presented as a contour plot of the $\log_{10}$ probability of two frequencies in these data. The plot shows clear evidence for two frequencies, even when only a few of the 380 data sets are analyzed, as was done here.

Chapter 8

SUMMARY AND CONCLUSIONS

In this study we have attempted to develop and apply some of the aspects of Bayesian parameter estimation to time series, even though the analysis as formulated is applicable to any data set, be it a time series or not.

8.1 Summary

We began this analysis in Chapter 2, by applying probability theory to estimate the spectrum of a data set that, we postulated, contained only a single sinusoid plus noise. In Chapter 3, we generalized these simple considerations to relatively complex models including the problem of estimating the spectrum of multiple nonstationary harmonic frequencies in the presence of noise. This led us to the "Student t-distribution": the posterior probability of the $\{\omega\}$ parameters, whatever their meaning. In Chapter 4, we estimated the parameters and calculated, among other things, the power spectral density, and the noise variance, and we derived a procedure for assessing the accuracy of the $\{\omega\}$ parameter estimates. In Chapter 6, we specialized to spectrum analysis and explored some of the implications of the "Student t-distribution" for this problem. In Chapter 7, we applied these analyses to a number of real time series with the aim of exploring and broadening some of the techniques needed to apply these procedures. In particular, we demonstrated how to use them to estimate multiple nonstationary frequencies and how to incorporate incomplete information into the estimation problem.

8.2 Conclusions

Perhaps the single biggest conclusion of this work is that what one can learn about a data set depends critically on what questions one asks. If one insists on taking the discrete Fourier transform of a data set, then our analysis shows that one will always obtain good answers to the question "What is the evidence of a single stationary harmonic frequency in these data?" This will be adequate if there are plenty of data and there is no evidence of complex phenomena. However, if the data show evidence for multiple frequencies or complex behavior, the discrete Fourier transform can give misleading or incorrect results in the light of more realistic models.

Although the use of integration to remove nuisance parameters is not new, and indeed the calculation in Chapter 3 has, to some degree, been done by every Bayesian who ever removed a nuisance parameter by integration, the realization of the degree of narrowing of the marginal joint posterior probability density that can be achieved by this is, to the best of our knowledge, new and almost startling. It indicates that, even though we might not be able to estimate an amplitude very precisely, the $\{\omega\}$ parameters often associated with an amplitude may be very precisely estimated. We can often improve the estimation of frequencies and decay rates by orders of magnitude over the estimates obtained from the discrete Fourier transform, least squares, or maximum likelihood. This is not to say that the actual estimates will be very different from those obtained from maximum likelihood or least squares – indeed, when little prior information is available the estimates of the parameters are the maximum likelihood estimates. The major difference is in the indicated accuracy of the estimates.

The principles of least squares or maximum likelihood provide no way to eliminate nuisance parameters, and thus oblige one to seek a global maximum in a space of much high dimensionality, which typically requires orders of magnitude more computation time. Having found this, they provide no way to assess the accuracy of the estimates other than the sampling distribution of the estimator – which is another even longer calculation. But it is a calculation that does not answer the real question of interest; it answers the "pre-data" question:

(Q1): "Before you have seen the data, how much do you expect the estimate to deviate from the true parameter value?"

The question of interest is the "post-data" one:

(Q2): "After getting the data, how accurately does the data set that you actually have determine the true values of the parameters?"

That these are very different questions with different answers in general, was recognized already by R. A. Fisher in the 1930's; he noted that in general two data sets that yield the same numerical value of the estimator, may nevertheless justify very different claims of accuracy. He sought to correct this by his device of conditioning on "ancillary statistics." But Jaynes [42] then showed that this conditioning is mathematically equivalent to using Bayes' theorem, as we have done here. Bayes' theorem, of course, always answers question (Q2), whether or not ancillary statistics exist.

The procedures for comparing models, Eq. (5.9), are perhaps new in the sense that we have extended the Bayesian calculation into the nonlinear $\{\omega\}$ parameters and by carefully keeping track of the normalization constants we were able eventually to integrate out all the parameters. This gives an objective way to compare models and to determine when additional effects are present in the data. Of course, as with any calculation, it will never replace the good sound judgment of the experimenter. The calculation can give a relative ranking of the various choices presented to it. It cannot decide which models to test.

Last, the improvement realized by these procedures when multiple measurements are present is quite striking. The analysis presented in Chapter 7 indicates that the traditional averaging rule will hold whenever the signal is exactly the same in every measurement. Yet in real experiments it is almost impossible to realize the true theoretical improvement. However, by computing the joint marginal posterior probability density of the common effects, the expected $\sqrt{n}$ can be obtained even in data sets where averaging clearly will not work. The implications of this for NMR and other fields are rather profound. Using these techniques we were able to improve resolution in NMR experiments by several orders of magnitude over the discrete Fourier transform; this is making it possible to examine extremely small effects that could not be examined before.

Appendix A

Choosing a Prior Probability

The question "How to choose the prior probability to express complete ignorance?" is interesting in itself, and it cannot be evaded in any problem of scientific inference that is to be solved by using probability theory and Bayes' theorem, but in which we do not wish to incorporate any particular prior information. In the case of the simple harmonic analysis performed in Chapter 2, there are four parameters to be estimated (B_1, B_2, ω, σ), and it is not obvious which choice of prior probabilities is to be preferred. Presumably, any prior probability distribution represents a conceivable state of prior information, but the problem of relating the distribution to the information is subtle and open-ended. You can always think more deeply and thus dredge up more prior information that you didn't think to use at first.

There are two questions one may consider to help in this. First, one should ask "Are the parameters logically connected?" That is, if we gain additional information about one of the parameters, does it change the estimates we would make about the others? If the answer is yes, then the parameters are not logically independent. It will be useful to find a representation where the parameters are independent.

Another useful question is "What are the invariances that the prior probability must obey?" That is, what transformations would convert the present problem into one where we have the same state of prior knowledge? Actually it is only this second question that is truly essential. However, using a representation in which the parameters are not logically independent will mean that the prior probabilities for all the parameters must be determined at once, by utilizing the properties of all the parameters.

In the two representations considered in Chapter 2, Cartesian versus polar, obtaining information about the frequency would rarely affect one's prior estimates of

the phase, amplitude, and noise level. Then the prior for the frequency will be independent of the other parameters, and the only invariance to be considered is some group of mappings S of ω onto itself. Later in this appendix we will derive the prior from the group of scale changes.

In the Cartesian representation, B_1 and B_2 are usually logically independent in the sense just noted, so we would assign them independent priors. In the polar notation the amplitude and phase are also logically independent, because obtaining information about either would not affect our prior estimate of the other. The volume elements transform as

$$dB_1 dB_2 = B dB d\theta$$

and so we want a probability density ρ with the two seemingly different forms:

$$\rho(B_1, B_2) dB_1 dB_2 = \rho(B, \theta) B dB d\theta$$

with

$$\rho(B_1, B_2) = f(B_1) f(B_2)$$

but also

$$\rho(B, \theta) = g(B) h(\theta).$$

But we rarely have prior information about θ, so we should take $h(\theta) = \text{const} = 1/2\pi$, $(0 \leq \theta \leq 2\pi)$. We are left with

$$f(B_1) f(B_2) = \frac{1}{2\pi} g(\sqrt{B_1^2 + B_2^2})$$

but setting $B_2 = 0$, this reduces to

$$f(B_1) f(0) = \frac{1}{2\pi} g(B_1)$$

so we have the functional equation

$$f(x) f(y) = f(\sqrt{x^2 + y^2}) f(0)$$

which a reasonable prior must satisfy. By writing this as

$$\log[f(x)] + \log[f(y)] = \log[f(\sqrt{x^2 + y^2})] + \log[f(0)]$$

the general solution is obvious; if a function $l(x)$ plus a function $l(y)$ is a constant plus a function only of $(x^2 + y^2)$ for all x, y the only possibility is

$$l(x) = ax^2 + b.$$

Thus, $f(x)$ must be a Gaussian; with $a = -1/2\sigma^2$ (the value of b is determined by normalization):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}.$$

To a modern physicist, this argument seems very familiar; it is just a two-dimensional version of Maxwell's original derivation of the Maxwellian velocity distribution [43]. However, historical research has shown that the argument was not original with Maxwell; ten years earlier the astronomer John Herschel [44] had given just our two-dimensional argument in finding the distribution of errors in measuring the position of a star. Thus the Gaussian prior that we use in Appendix B to illustrate the limit as $\sigma \to \infty$ to a uniform prior was not arbitrary; it is the only prior that could have represented our "uninformative" state of knowledge about these parameters. This is a good example of how one can relate prior probabilities to prior information by logical analysis.

In the calculation done by Jaynes [12] the prior used was $dAd\theta$, whereas ours amounts to taking instead $AdAd\theta$. The calculation performed in Chapter 2, and that done by Jaynes will differ from each other only in the fine details. We, effectively, assume slightly less information about the amplitude than Jaynes did, and so we make a slightly less conservative estimate of the frequency. This also simplifies the results by eliminating the Bessel functions found by Jaynes. However, as demonstrated in Appendix B, the differences introduced by the use of different priors to represent ignorance are negligibly small if we have any reasonable amount of data.

When we know that the parameters B_1, B_2, ω, σ are logically independent, how does one choose a prior to represent ignorance of ω and σ? Perhaps the easiest way is to exploit the invariances in the problem. The invariances we would like to exploit are the time invariances. There are two of these: first, the actual starting time of the experiment cannot make any difference; second, a small change in the sampling rate of the problem cannot make any difference provided the same amount of data is collected. To exploit these we apply a technique described by Jaynes [45].

Consider the following problem: we have two experimenters who are to take data on a stationary time series (the same problem described in Chapter 2). Each of these experimenters is free to set up and take the data in any way he sees fit. They do however measure the same time series, starting at slightly different times and using slightly different sampling rates. Now the first experimenter, called E, assigns to his

parameters a prior probability

$$P(B_1, B_2, \omega, \sigma|I) \propto G(B_1, B_2, \omega, \sigma) dB_1 dB_2 d\omega d\sigma$$

and the second experimenter called E' assigns to his a prior probability

$$P(B_1', B_2', \omega', \sigma'|I) \propto H(B_1', B_2', \omega', \sigma') dB_1' dB_2' d\omega' d\sigma'.$$

The model equation used by E is just the model used in Chapter 2,

$$f(t, B_1, B_2, \omega) = B_1 \cos(\omega t) + B_2 \sin(\omega t)$$

and E' uses the same equation but with the primed variables

$$f(t', B_1', B_2', \omega') = B_1' \cos(\omega' t') + B_2' \sin(\omega' t').$$

These two equations are related to each other by a simple transformation in the time variable $t' = \alpha t + t_0$ where α is related to the sampling rates and t_0 is the difference in their starting times. The relations between these two system are

$$\begin{aligned} \alpha\omega' = \omega, \qquad & \text{and} \qquad \alpha d\omega' = d\omega \\ B_1 = B_1' \cos(\omega' t_0) + B_2' \sin(\omega' t_0) & \\ B_2 = B_2' \cos(\omega' t_0) - B_1' \sin(\omega' t_0) & \\ dB_1 dB_2 = dB_1' dB_2' & \\ \sigma = \gamma\sigma' \qquad & \text{and} \qquad d\sigma = \gamma d\sigma'. \end{aligned} \tag{A.1}$$

The factor of α from the time transformation will be absorbed into the frequency as a scaling, because the number of cycles in a given interval ($\omega t/2\pi = \omega' t'/2\pi$) is an invariant. The squared magnitudes of their model functions are equal; the transformation introduces only an apparent phase change into the signal. In addition to the transformation for the frequency ω the variable σ will have an arbitrary scaling introduced into it.

Now we know that each of these experimenters has performed essentially the same experiment and we expect them to obtain nearly identical conclusions. Each of the experimenters is in the same state of knowledge about his experiment and we apply Jaynes' desideratum of consistency: "In two problems where we have the same prior information, we should assign the same prior probability" [45]. Because E and E' are in the same state of knowledge, H and G are the same functions. Thus we have

$$G(B_1, B_2, \omega, \sigma) dB_1 dB_2 d\omega d\sigma = G(B_1', B_2', \omega', \sigma') dB_1' dB_2' d\omega' d\sigma'.$$

We will solve for the dependence of the prior on the frequency and variance having already obtained the priors for B_1 and B_2. We substitute for ω and σ to obtain

$$G(B_1, B_2, \alpha\omega', \gamma\sigma') = \frac{G(B_1', B_2', \omega', \sigma')}{\gamma\alpha}.$$

This is a functional equation for the prior probability G. It is evident from (A.1) that G must be independent of B_1 and B_2, so the dependence of the prior on the parameters is now completely determined: the only prior which represents complete ignorance of ω, σ, B_1, and B_2 is

$$P(B_1, B_2, \omega, \sigma|I) \propto \frac{1}{\omega\sigma}.$$

This is the Jeffreys prior which we used for the standard deviation σ. Other more cogent derivations of the Jeffreys prior are known [46] but they involve additional technical tools beyond our present scope.

Of course, the realistic limits of the Jeffreys prior do not go all the way to zero and infinity; for example, we always know in advance that σ cannot be less than a value determined by the digitizing accuracy with which we record data; nor so great that the noise power would melt the apparatus. Likewise, as discussed earlier, we know that when the data have zero mean, our data do not contain a zero frequency component; nor can the data contain frequencies so high that they would not pass through our circuitry. Strictly speaking, then, a Jeffreys prior should always be taken between finite positive limits, and be normalized:

$$P(\sigma|I) = \left\{ \begin{array}{ll} A\sigma^{-1} & a < \sigma < b \\ \\ 0 & \text{otherwise} \end{array} \right\} \tag{A.2}$$

with $A^{-1} = \log(b/a)$. But then this prior gets multiplied by a likelihood of the form

$$L(\sigma) = \sigma^{-N} \exp\left\{-\frac{C}{\sigma^2}\right\}$$

which cuts off so strongly as $\sigma \to 0$ and $\sigma \to \infty$ that practically all the mass of the posterior distribution

$$P(\sigma|DI) \propto L(\sigma)P(\sigma|I) \tag{A.3}$$

is concentrated near the peak of (A.3), at $\sigma^2 = 2C/(N+1)$. In our examples, the exact conclusions from (A.2) differ from the limiting ones ($a \to 0, b \to \infty$) by amounts generally less than one part in 10^{20}, so in practice we need never introduce the limits a, b. Similarly, the prior limits on ω have negligible numerical effect, and need not be

introduced at all. In our calculation we used a uniform prior for the frequency instead of the Jeffreys prior simply to save writing, because we knew that the difference in the resulting frequency estimates would be negligibly small compared to the width of the posterior distributions (i.e. compared to the error $\delta\omega$ which was inevitable in any event).

Appendix B

Improper Priors as Limits

In the simple harmonic frequency problem Chapter 2 when we removed the amplitudes B_1 and B_2 by integration to get Eq. (2.6), we used a uniform prior probability density which we called an improper prior. In fact, such a function is not a probability density at all. When we use an improper prior, what we really mean is that our prior information is vague, that it carries negligible weight compared to the evidence of the data: the exact prior bounds are so wide that they are far outside the range indicated by the data. To perform the calculation (1.4) correctly, one could bound the parameter to be removed, integrate over the bounded region, and then take a limit as the bound is allowed to go to infinity; but for this problem the result is the same.

Alternatively, we could assume we have a previously measured value of the parameter and then take the limit as the uncertainty in that measurement becomes infinite. We will use a calculation very similar to this in a number of places in the text, and we give this calculation to demonstrate that the use of an improper prior to express "complete ignorance" cannot affect the results in any significant way. This will also show the effect of incorporating additional information into the calculation. Suppose we have some previously measured values for the amplitudes, designated as $\hat{B}_1$ and $\hat{B}_2$. We now proceed to calculate the expectation value of the amplitudes using a prior probability that takes this information into account.

Suppose the previous measured values $\hat{B}_1$ and $\hat{B}_2$ are known with an accuracy of $\pm\delta$ (interpreted as the standard deviation of a Gaussian error distribution for the previous measurements). The joint prior probability density of the true values B_1

and B_2 is the posterior distribution for the first measurement,

$$P(B_1, B_2|I) = \left[2\pi\delta^2\right]^{-1} \exp\left\{-\frac{1}{2\delta^2}[(\hat{B}_1 - B_1)^2 + (\hat{B}_2 - B_2)^2]\right\} \tag{B.1}$$

which becomes our informative prior for the second measurement. Then using Bayes' theorem, the posterior probability of the parameters is proportional to the product of the prior (B.1) and the likelihood (2.3):

$$P(B_1, B_2|D, I) = \left[2\pi\delta^2\right]^{-1} \left[2\pi\sigma^2\right]^{-\frac{N}{2}} \exp\left\{-\frac{X}{2\delta^2} - \frac{NY}{4\sigma^2}\right\}$$

where

$$X \equiv (\hat{B}_1 - B_1)^2 + (\hat{B}_2 - B_2)^2$$

$$Y \equiv B_1^2 + B_2^2 - 2\left(\frac{2R(\omega)}{N}B_1 + \frac{2I(\omega)}{N}B_2\right).$$

After a little algebra the posterior probability may be written as

$$P(B_1, B_2|D, I) = \left[2\pi\delta^2\right]^{-1} \left[2\pi\sigma^2\right]^{-\frac{N}{2}} \exp\left\{-\beta[(B_1 - E_1)^2 + (B_2 - E_2)^2]\right\}$$

$$\beta = \frac{\delta^2 + \sigma^2}{2\delta^2\sigma^2}$$

where

$$E_1 = \frac{\delta^2[2R(\omega)/N] + \sigma^2\hat{B}_1}{\delta^2 + \sigma^2} \tag{B.2}$$

$$E_2 = \frac{\delta^2[2I(\omega)/N] + \sigma^2\hat{B}_2}{\delta^2 + \sigma^2} \tag{B.3}$$

are the posterior expectations:

$$\langle B_1\rangle = E_1 \qquad \text{and} \qquad \langle B_2\rangle = E_2.$$

The posterior estimates are now weighted averages of the two measurements. This is a rather old result, first discovered by Laplace [47] but essentially forgotten for a century, until the modern development of Bayesian methods began to demonstrate that most of Laplace's results were correct and important.

To understand the full implications of this we will consider three special cases. First, when $\delta \ll \sigma$, the previous measurement is much better than the current one. Then

$$\langle B_1\rangle = \hat{B}_1 \qquad \text{and} \qquad \langle B_2\rangle = \hat{B}_2$$

which says to use the original measured value, a most pleasing result, since that is exactly what any physicist would have done anyway. Second, consider the case where $\sigma = \delta$. Then

$$\langle B_1 \rangle = \frac{1}{2}\left(\frac{2R(\omega)}{N} + \hat{B}_1\right) \qquad \text{and} \qquad \langle B_2 \rangle = \frac{1}{2}\left(\frac{2I(\omega)}{N} + \hat{B}_2\right)$$

which says the two measurements are of equal weight and one should average them. Again a most pleasing result, since that is exactly what one's intuition would have told one to do. Third, consider the case when $\delta \gg \sigma$ (one knows only that the two amplitudes must be bounded) then,

$$\langle B_1 \rangle = \frac{2R(\omega)}{N} \qquad \text{and} \qquad \langle B_2 \rangle = \frac{2I(\omega)}{N}. \tag{B.4}$$

This is the result obtained using the improper prior. In the limit as δ goes to infinity, the prior (B.1) goes smoothly into the uniform improper prior used in our calculation of Eq. (2.6), and the weighted averages go smoothly into (B.4).

The important point here is that if δ is appreciably greater than σ, the prior we use does not make any significant difference; as δ becomes larger, less information is conveyed by the prior measurement, and probability theory as indicated by (B.2) and (B.3) automatically assigns less weight to it. The result must depend mostly on the evidence in the data. In the limit as δ goes to infinity we have incorporated no prior information about the parameter, and the result must depend totally on the data.

Appendix C

Removing Nuisance Parameters

We illustrate in this appendix that integrating over a nuisance parameter is very much like estimating the parameter from the data and constraining it in the posterior probability to that value. We first estimate the amplitudes by calculating their posterior expectations, and then substitute them into the likelihood (2.3). If integrating over a nuisance parameter is nearly the same, we should obtain (2.6), or at the very least something very much like (2.6). We assume for this illustration that σ is known; then, using the likelihood Eq. (2.3), the expectation value of B_j, supposing ω known, is

$$\langle B_j \rangle = \frac{\int_{-\infty}^{+\infty} dB_1 dB_2 B_j L(B_1, B_2, \omega, \sigma)}{\int_{-\infty}^{+\infty} dB_1 dB_2 L(B_1, B_2, \omega, \sigma)}. \tag{C.1}$$

We take these as our estimates $\langle B_j \rangle$ (ω) in (2.3). Carrying out the required integrations gives the posterior expectation values of B_1 and B_2:

$$B_1^*(\omega) = \langle B_1(\omega) \rangle = \frac{2R(\omega)}{N}, \tag{C.2}$$

$$B_2^*(\omega) = \langle B_2(\omega) \rangle = \frac{2I(\omega)}{N},$$

where $R(\omega)$ and $I(\omega)$ are the cosine and sine transforms of the data, as defined in (2.4) and (2.5). Now these are substituted back into (2.3) to give

$$L(B_1^*, B_2^*, \omega, \sigma) \propto \sigma^{-N} \exp\left\{ -\frac{N}{2\sigma^2} [\overline{d^2} - \frac{2}{N} C(\omega)] \right\}. \tag{C.3}$$

But in its dependence on ω, this is just (2.6): integrating over the amplitudes with respect to the uniform prior has given us the same result as constraining them to their expectation values (C.1).

The two procedures are not always equivalent, as they happen to be here, but they can never be very different whenever we have enough information or data to make a good estimate of a nuisance parameter. In fact, these procedures would have been slightly different in this example if we had not assumed the noise variance σ^2 to be known. Then σ^2 would also become a nuisance parameter which we would remove by integration, and the "Student t-distribution" thus obtained would be raised to the $-N/2$ power instead of $(2-N)/2$ as was found in Chapter 2.

More generally, whenever a nuisance parameter is actually well determined by many data ($N \to \infty$), these two procedures become for all practical purposes equivalent. But when the data are too meager to determine the nuisance parameters very well, the *ad hoc* procedure (C.3) can be overoptimistic, leading us to think that we have determined ω more accurately than the data really justify; and if we have relevant prior information about the parameters the *ad hoc* method ignores it.

Appendix D

Uninformative Prior Probabilities

When we worked the single frequency problem in Chapter 2 we used a uniform prior for the amplitudes. In polar coordinates this prior is

$$P(B,\theta|I) \propto BdBd\theta$$

and leads to

$$P(\omega|\sigma,D,I) \propto \exp\left\{\frac{C(\omega)}{\sigma^2}\right\} \tag{D.1}$$

as the posterior probability of a single harmonic frequency, given the data and the noise variance σ^2. When Jaynes [12] worked this problem he performed the calculation in polar coordinates and supposed prior information I' for which

$$P(B\theta|I') \propto dBd\theta$$

as the prior for the amplitude and phase. He then arrived at

$$P(\omega|\sigma,D,I') \propto \exp\left\{\frac{C(\omega)}{2\sigma^2}\right\} I_0\left(\frac{C(\omega)}{2\sigma^2}\right) \tag{D.2}$$

where I_0 is a Bessel function of order zero. This is a very different looking result, given that the only difference in the two calculations was the prior used. How can such a simple change in the problem have such a dramatic effect on the answer, and just what effect did the use of these two different priors have on the results?

The main question we will pursue here is "What effect did this different prior have on the frequency estimate?" The answer to this question is surprising: since Eq. (D.1) and Eq. (D.2) are both functions of $C(\omega)$, they both reach their maximum at the same value $\omega = \hat{\omega}$; there is no difference at all in the actual frequency estimate! But there is

a difference in the curvatures of Eq. (D.1) and Eq. (D.2) at their common maximum $\hat{\omega}$, so there is a difference in the claimed accuracy of that estimate. Recalling that in the Gaussian approximation it is the second derivative of $\log(P(\omega|\sigma, D, I)$ that matters,

$$\frac{d^2}{d\omega^2} \log P(\omega|\sigma, D, I)\Big|_{\omega=\hat{\omega}} = \frac{1}{(\delta\omega)^2}$$

a short calculation gives for the standard deviations, from Eq. (D.1)

$$\delta\omega = \frac{\sigma}{\sqrt{C''(\hat{\omega})}}$$

and from Eq. (D.2)

$$\delta\omega' = \frac{\sigma}{\sqrt{C''(\hat{\omega})}} \left(\frac{2I_0}{I_0 + I_1}\right)^{\frac{1}{2}}$$

where the argument of the I_0 and I_1 Bessel functions is $C(\hat{\omega})/2\sigma^2$. The ratio of the error estimates is $q(C(\hat{\omega}/2\sigma^2))$, where

$$q(x) = \left(\frac{2I_0(x)}{I_0(x) + I_1(x)}\right)^{\frac{1}{2}}.$$

Substituting some numerical values for x we have

x	$q(x)$
0	1.414
1	1.176
2	1.086
4	1.036
8	1.016
> 18	$1 + (8x)^{-1}$.

Now if there is a single sinusoid present with amplitude B, the maximum of the periodogram will be about

$$C(\hat{\omega}) \approx \frac{NB^2}{4}.$$

With a signal-to-noise ratio of unity, the mean square signal $B^2/2 = \sigma^2$, so

$$\frac{C(\hat{\omega})}{2\sigma^2} \approx \frac{N}{4}.$$

If $N \geq 10$, there is less than a 6.5% difference in the error estimates, and when $N > 50$ the difference is less than 1%. Thus whenever we have enough signal-to-noise ratio or enough data to justify any frequency estimates at all, the differences are completely negligible.

Appendix E

Computing the "Student t-Distribution"

This subroutine was used to prepare all of the numerical analysis presented in this work. This is a general purpose implementation of the calculation that will work for any model functions and for any setting of the parameters, independent of the number of parameters and their values, and it does not care if the data are uniformly sampled or not. In order to do this, the subroutine requires five pieces of input data and one work area. On return one receives $H_i(t_j)$, h_i, $\overline{h^2}$, $P(\{\omega\}|D,I)$, $\langle\sigma\rangle$, and $\hat{p}(\{\omega\})$. The parameter list is as follows:

Parm	LABEL	i/o	Description/function
N	INO	input	The number of discrete time samples in the time series to be analyzed.
m	IFUN	input	This is the order of the matrix g_{jk} and is equal to the number of model functions.
d_j	DATA	input	The time series (length N): this is the data to be analyzed. Note: the routine does not care if the data are sampled uniformly or not.
g_{ij}	GIJ	input	This matrix contains the j nonorthogonal model functions [dimensioned as GIJ(INO,IFUN)] and evaluated at t_i.

Parm	LABEL	i/o	Description/function
ZLOGE	ZLOGE	i/o	This is the $\log_e$ of the normalization constant. The subroutine never computes the "Student t-distribution" when ZLOGE is zero: instead the $\log_e$ of the "Student t-distribution" is computed. It is up to the user to locate a value of $\log_e[P(\{\omega\}\|D,I)]$ close to the maximum of the probability density. This log value should then be placed in ZLOGE to act as an upper bound on the normalization constant. With this value in place the subroutine will return the value of the probability; then, an integral over the probability density can be done to find the correct value of the normalization constant.
$H_i(t_j)$	HIJ	output	These are orthonormal model functions Eq. (3.5) evaluated at the same time and parameter values as GIJ.
h_i	HI	output	These are projections of the data onto the orthonormal model functions Eq. (3.13) and Eq. (4.3).
$\overline{h^2}$	H2BAR	output	The sufficient statistic $\overline{h^2}$ Eq. (3.15) is always computed.
$P(\{\omega\}\|D,I)$	ST	output	The "Student t-distribution" Eq. (3.17) is not computed when the normalization constant is zero. To insure this field is computed the normalization constant must be set to an appropriate value.
STLE	STLE	output	This is the $\log_e$ of the "Student t-distribution" Eq. (3.17) and is always computed.
$\langle\sigma\rangle$	SIG	output	This is the expected value of the noise variance σ as a function of the $\{\omega\}$ parameters Eq. (4.6) with $s = 1$.
$\hat{p}(\{\omega\})$	PHAT	output	This is the power spectral density Eq. (4.15) as a function of the $\{\omega\}$ parameters.

Parm	LABEL	i/o	Description/function
	WORK	scratch	This work area must be dimensioned at least $5m^2$. The dimension in the subroutines was set high to avoid possible "call by value" problems in FORTRAN. On return, WORK contains the eigenvectors and eigenvalues of the g_{jk} matrix. The eigenvector matrix occupies m^2 contiguous storage locations. The m eigenvalues immediately follow the eigenvectors.

This subroutine makes use of a general purpose "canned" eigenvalue and eigenvector routine which has not been included. The original routine used was from the IMSL library and the code was later modified to use a public-domain implementation (an EISPACK routine). The actual routine one uses here is not important so long as the routine calculates both the eigenvalues and eigenvectors of a real symmetric matrix. If one chooses to implement this program one must replace the call (clearly marked in the code) with a call to an equivalent routine. Both the eigenvalues and eigenvectors are used by the subroutine and it assumes that the eigenvectors are normalized.

```
      SUBROUTINE PROB
     C (IN0,IFUN,DATA,GIJ,ZLOGE,HIJ,HI,H2BAR,ST,STLOGE,SIGMA,PHAT,WORK)
      IMPLICIT REAL*08(A-H,0-Z)
      DIMENSION DATA(IN0),HIJ(IN0,IFUN),HI(IFUN),GIJ(IN0,IFUN)
      DIMENSION WORK(IFUN,IFUN,20)
C
C
      CALL VECTOR(IN0,IFUN,GIJ,HIJ,WORK)
C
      H2=0D0
      D0 1600 J=1,IFUN
      H1=0D0
      D0 1500 L=1,IN0
 1500 H1=H1 + DATA(L)*HIJ(L,J)
      HI(J)=H1
      H2=H2 + H1*H1
 1600 CONTINUE
      H2BAR=H2/IFUN
C
      Y2=0D0
      D0 1000 I=1,IN0
 1000 Y2=Y2 + DATA(I)*DATA(I)
      Y2=Y2/IN0
C
```

```
      QQ=1D0 - IFUN*H2BAR / IN0 / Y2
      STLOGE=DLOG(QQ) * ((IFUN - IN0)/2D0)
C
      AHOLD=STLOGE - ZLOGE
      ST   =0D0
      IF(DABS(ZLOGE).NE.0D0)ST=DEXP(AHOLD)
C
      SIGMA=DSQRT( IN0/(IN0-IFUN-2) * (Y2 - IFUN*H2BAR/IN0) )
C
      PHAT = IFUN*H2BAR * ST
C
      RETURN
      END
      SUBROUTINE VECTOR(IN0,IFUN,GIJ,HIJ,WORK)
      IMPLICIT REAL*8(A-H,0-Z)
      DIMENSION HIJ(IN0,IFUN),GIJ(IN0,IFUN),WORK(IFUN,IFUN,20)
C
      DO 1000 I=1,IFUN
      DO 1000 J=1,IN0
 1000 HIJ(J,I)=GIJ(J,I)
C
      CALL ORTH0(IN0,IFUN,HIJ,WORK)
C
      DO 5000 I=1,IFUN
      TOTAL=0D0
      DO 4500 J=1,IN0
 4500 TOTAL=TOTAL + HIJ(J,I)**2
      ANORM=DSQRT(TOTAL)
      DO 4000 J=1,IN0
 4000 HIJ(J,I)=HIJ(J,I)/ANORM
 5000 CONTINUE
C
      RETURN
      END

      SUBROUTINE ORTH0(IN0,NMAX,AIJ,W)
      IMPLICIT REAL*8 (A-H,0-Z)
      REAL*8  AIJ(IN0,NMAX),W(NMAX)
C
      IT=1
      IE=IT + NMAX*NMAX
      IM=IE + NMAX*NMAX
      IW=IM + NMAX*NMAX
      I2=IW + NMAX*NMAX
C
      CALL TRANS(IN0,NMAX,AIJ,W(IM),W(IT),W(IE),W(IW),W(I2))
C
```

```
      RETURN
      END
      SUBROUTINE TRANS
     C(INO,NMAX,AIJ,METRIC,TRANSM,EIGV,WORK1,WORK2)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8  AIJ(INO,NMAX)
      REAL*8  METRIC(NMAX,NMAX),EIGV(NMAX)
      REAL*8  TRANSM(NMAX,NMAX),WORK1(NMAX),WORK2(NMAX)
      DO 2000 I=1,NMAX
      DO 2000 J=1,NMAX
      TOTAL=0D0
      DO 1000 K=1,INO
 1000 TOTAL=TOTAL + AIJ(K,I)*AIJ(K,J)
      METRIC(I,J)=TOTAL
 2000 CONTINUE
C*****************************************************************
C**** THIS CALL MUST BE REPLACED WITH THE CALL TO AN EIGENVALUE
C**** AND EIGENVECTOR ROUTINE
      CALL EIGERS(NMAX,NMAX,METRIC,EIGV,1,TRANSM,WORK1,WORK2,IERR)
C**** NMAX   IS THE ORDER OF THE MATRIX
C**** METRIC IS THE MATRIX FOR WHICH THE EIGENVALUES AND VECTORS
C****        ARE NEEDED
C**** EIGV   MUST CONTAIN THE EIGENVALUES ON RETURN
C**** TRANSM MUST CONTAIN THE EIGENVECTORS ON RETURN
C**** WORK1  IS A WORK AREA USED BY MY ROUTINE AND MAY BE USED
C****        BY YOUR ROUTINE.  ITS  DIMENSION IS NMAX
C****        IN THIS ROUTINE. HOWEVER IT MAY BE DIMENSIONED
C****        AS LARGE AS NMAX*NMAX WITHOUT AFFECTING ANYTHING.
C**** WORK2  IS A SECOND WORK AREA AND IS OF DIMENSION NMAX
C****        IN THIS ROUTINE, IT MAY ALSO BE DIMENSIONED AS
C****        LARGE AS NMAX*NMAX WITHOUT AFFECTING ANYTHING.
C*****************************************************************
      DO 5120 K=1,INO
      DO 3100 J=1,NMAX
 3100 WORK1(J)=AIJ(K,J)
      DO 5120 I=1,NMAX
      TOTAL=0D0
      DO 3512 J=1,NMAX
 3512 TOTAL=TOTAL + TRANSM(J,I)*WORK1(J)
 5120 AIJ(K,I)=TOTAL
      RETURN
      END
```

Bibliography

[1] Bretthorst G. L., (1987), Bayesian Spectrum Analysis and Parameter Estimation, Ph.D. thesis, Washington University, St. Louis, MO., available from University Microfilms Inc., Ann Arbor Mich.

[2] Robinson, E. A., (1982), "A Historical Perspective of Spectrum Estimation," *Proceedings of the IEEE,* 70, pp. 855-906.

[3] Marple, S. L., (1987), Digital Spectral Analysis with Applications, Prentice-Hall, New Jersey.

[4] Laplace, P. S., (1812), Théorie Analytique des Probabilités, Paris, (2nd edition, 1814; 3rd edition, 1820).

[5] Legendre, A. M., (1806), "Nouvelles Méthods pour la Détermination des Orbits des Cométes," Paris.

[6] Gauss, K. F., (1963 reprint) Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections, Dover Publications, Inc., New York.

[7] Cooley, J. W., P. A. Lewis, and P. D. Welch, (1967), "Historical Notes on the Fast Fourier Transform," *Proc. IEEE* 55, pp. 1675-1677.

[8] Brigham, E., and R. E. Morrow, (1967), "The Fast Fourier Transform," *Proc. IEEE Spectrum,* 4, pp. 63-70.

[9] Gentleman, W. M., (1968), "Matrix Multiplication and Fast Fourier Transformations," *Bell Syst. Tech. Journal,* 17, pp. 1099-1103.

[10] Cooley, J. W., and J. W. Tukey, (1965), "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation,* 19, pp. 297-301.

[11] Schuster, A., (1905), "The Periodogram and its Optical Analogy," *Proceedings of the Royal Society of London,* 77, pp. 136.

[12] Jaynes, E. T. (1987), "Bayesian Spectrum and Chirp Analysis," in Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems, C. Ray Smith, and G. J. Erickson, ed., D. Reidel, Dordrecht-Holland, pp. 1-37.

[13] Blackman, R. B., and J. W. Tukey, (1959), The Measurement of Power Spectra, Dover Publications, Inc., New York.

[14] Jaynes, E. T. (1983),Papers on Probability, Statistics and Statistical Physics, a reprint collection, D. Reidel, Dordrecht-Holland.

[15] Jeffreys, H., (1939), Theory of Probability, Oxford University Press, London, (Later editions, 1948, 1961).

[16] Lord Rayleigh, (1879), *Philosophical Magazine,* 5, pp. 261.

[17] Tukey, J. W., several conversations with E. T. Jaynes, in the period 1980-1983.

[18] Waldmeier, M., (1961), The Sunspot Activity in the Years 1610-1960, Schulthes, Zurich.

[19] Nyquist, H., (1928), "Certain Topics in Telegraph Transmission Theory," *Transactions AIEE*, pp. 617.

[20] Nyquist, H., (1924), "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal,* 3, pp. 324.

[21] Hooke, R., and T. A. Jeeves, (1962), "Direct Search Solution of Numerical and Statistical Problems," *J. Assoc. Comp. Mach.,* pp. 212-229.

[22] Wilde D. J., (1964), Optimum Seeking Methods, Prentice-Hall, Inc. Englewood Cliffs, N. J.

[23] Zellner, A., (1980), in Bayesian Statistics, J. M. Bernardo, ed., Valencia University Press, Valencia, Spain.

[24] Geisser, S., and J. Cornfield, (1963), "Posterior Distribution for Multivariate Normal Parameters," *Journal of the Royal Statistical Society*, B25, pp. 368-376.

[25] Zellner, A., (1971), An Introduction to Bayesian Inference in Econometrics, John Wiley and Sons, New York. Second edition, (1987).

[26] Cox, R. T., (1961), The Algebra of Probable Inference, Johns-Hopkins Press, Baltimore, Md.

[27] Tribus, M., (1969), Rational Descriptions, Decisions and Designs, Pergamon Press, Oxford.

[28] Schlaifer, R., (1959), Probability and Statistics for Business Decisions: an Introduction to Managerial Economics Under Uncertainty, McGraw-Hill Book Company, New York.

[29] Whittle, P., (1954), Appendix to H. Wold, Stationary Time Series, Almquist and Wiksell, Stockholm, pp. 200-227.

[30] Shaw, D., (1976), Fourier Transform NMR Spectroscopy, Elsevier Scientific Pub. Co., New York.

[31] Ganem, J. W., and R. E. Norberg, (1987), Private Communication.

[32] Abragam, A., (1961), Principles of Nuclear Magnetism, Oxford Science Publications, London.

[33] Beckett, R. J., (1979), The Temperature and Density Dependence of Nuclear Spin-Spin Interactions in Hydrogen-Deuteride Gas and Fluid,
Ph.D. thesis, Rutgers University, New Brunswick, New Jersey; available from University Microfilms Inc., Ann Arbor Mich.

[34] Currie, R. G., (1985), Private Communication.

[35] Currie, R. G., and S. Hameed, (1986), "Climatically Induced Cyclic Variations in United States Corn Yield and Possible Economic Implications," presented at the Canadian Hydrology Symposium, Regina, Saskatchewan.

[36] Burg, John Parker, (1975), Maximum Entropy Spectral Analysis, Ph.D. Thesis, Stanford University; available from University Microfilms Inc., Ann Arbor Mich.

[37] Cohen, T. J., and P. R. Lintz, (1974), "Long Term Periodicities in the Sunspot Cycle," *Nature,* 250, pp. 398.

[38] Sonett, C. P., (1982), "Sunspot Time Series: Spectrum From Square Law Modulation of the Half Cycle," *Geophysical Research Letters,* 9 pp. 1313-1316.

[39] Bracewell, R. N,. (1986), "Simulating the Sunspot Cycle," *Nature,* 323, pp. 516.

[40] Jaynes, E. T., (1982), "On the Rationale of Maximum-Entropy Methods", *Proceedings of the IEEE*, 70, pp. 939-952.

[41] Smith, W. H., and W. Schempp, (1987) private communication.

[42] Jaynes, E. T., (1976), "Confidence Intervals vs. Bayesian Intervals," in Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, W. L. Harper and C. A. Hooker, editors, D. Reidel Publishing Co., pp. 252-253; reprinted in [14].

[43] Maxwell, J. C., (1860), "Illustration of the Dynamical Theory of Gases. Part I. On the Motion and Collision of Perfectly Elastic Spheres," *Philosophical Magazine,* 56.

[44] Herschel, J., (1850), *Edinburgh Review,* 92, pp. 14.

[45] Jaynes, E. T., (1968), "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics,* SSC-4, pp. 227-241; reprinted in [14].

[46] Jaynes, E. T., (1980), "Marginalization and Prior Probabilities," in *Bayesian Analysis in Econometrics and Statistics,* A. Zellner, ed., North-Holland Publishing Company, Amsterdam; reprinted in [14].

[47] Laplace, P. S., (1814), A Philosophical Essay on Probabilities, Dover Publications, Inc., New York, (1951, unabridged and unaltered reprint of Truscott and Emory translation).

Index