[19] Burg, John Parker, *Maximum Entropy Spectral Analysis, Ph.D. Dissertation,* (University Microfilms No. 75-25), Stanford University, 1975.

[20] Cohen, T. J. and P. R. Lintz, "Long Term Periodicties in the Sunspot Cycle," *Nature,* vol. 250, p. 398, 1974.

[21] Sonnet, C. P., "Sunspot Time Series: Spectrum From Square Law Modulation of the Half Cycle," *Geophysical Research Letters,* vol. 9 NO 12, pp. 1313-1316, 1982

[22] Bracewell, R. N., "Simulating the Sunspot Cycle," *Nature,* vol. 323, p. 516, Oct. 9, 1986.

References

[1] Bretthorst, G. Larry, *Bayesian Spectrum Analysis and Parameter Estimation,* Ph.D. thesis, University Microfilms Inc. Washington University, St. Louis, MO Aug. 1987

[2] Jaynes, E. T., "Bayesian Spectrum and Chirp Analysis," in *Proceedings of the Third Workshop on Maximum-Entropy and Bayesian Methods (1983),* ed. C. Ray Simth, D. Reidel, Boston 1987. (the third workshop was held in Laramie, Wyoming)

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

[4] Jaynes, E. T. *Papers on Probability, Statistics and Statistical Physics,* ed. R. D. Rosenkrantz, D. Reidel, Boston, 1983.

[5] Schuster, A., "The Periodogram an its Optical Analogy," *Proceedings of the Royal Society of London,* vol. 77, p. 136, 1905.

[6] Lord Rayleigh, *Philosophical Magazine,* vol. (5) 8, p. 261, 1879.

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

[8] Jeffreys, Sir Harold, *Theory of Probability,* Oxford University press, London, 1939. (Later editions, 1948, 1961)

[9] Jaynes, E. T., "Prior Probabilities," in *Papers on Probability, Statistics and Statistical Physics,* ed. R. D. Rosenkrantz, D. Reidel, Boston, 1983.

[10] Jaynes, E. T., "Marginalization and Prior Probabilities," in *Papers on Probability, Statistics and Statistical Physics,* ed. R. D. Rosenkrantz, D. Reidel, Boston, 1983.

[11] Waldmeier, M., in *The Sunspot Activity in the Years 1610-1960,* Schulthes, Zurich, 1961.

[12] Hooke, Robert and T. A. Jeeves, "'Direct Search' Solution of Numerical and Statistical Problems," *J. Assoc. Comp. Mach.,* p. 212, March 1962.

[13] Shaw, Derek, *Fourier Transform NMR Spectroscopy,* Elsevier Scientific Pub. Co., New York, 1976.

[14] Ganem, Joseph W. and R. E. Norberg, *Private Communication,* 1987.

[15] Abragam, A, *Principles of Nuclear Magnetism,* p. 187, Oxford Science Publications, London, 1961. reprint (1985)

[16] Beckett, Richard James, *The Temperature and Density Dependence of Nuclear Spin-Spin Interactions in Hydrogen-Deuteride Gas and Fluid,* Rutgers University Ph.D. Thesis, New Brunswick, New Jersey, 1979. (unpublished)

[17] Currie, Robert Guinn, *Private Communication,* 1985.

[18] Currie, Robert Guinn and Sultan Hameed, "Climatically Induced Cyclic Variations in United States Corn Yield and Possible Economic Implications," *presented at the Canadian Hydrology Symposium,* Regina, Saskatchewan, June 3, 1986.

```
      END

      SUBROUTINE SETGIJ(IN0,IVEC,GIJ,W,ALPHA)
      IMPLICIT REAL*08(A-H,0-Z)
      DIMENSION GIJ(IN0,IVEC)
C
C     THIS ROUTINE WILL EVALUATE THE MODEL FUNCTIONS AT
C     FREQUENCY W AND DECAY RATE ALPHA
C
      ADELTA=0.5D0*IN0 + 0.5D0
C
      DO 1000 I=1,IN0
C
      TIME=I - ADELTA
C
C     EVALUATE MODEL FUNCTION 1
      GIJ(I,1)=DCOS(W*TIME)*DEXP(ALPHA*TIME)
C
 1000 GIJ(I,2)=DSIN(W*TIME)*DEXP(ALPHA*TIME)
C
      RETURN
      END

      REAL FUNCTION ALIKE*8(W,ALPHA)
      IMPLICIT REAL*08(A-H,0-Z)
      DIMENSION DATA(512),HIJ(512,2),HI(2),GIJ(512,2),WA(2,2,20)
      COMMON    DATA,HIJ,HI,GIJ,WA,ZLE,IN0
C
C     THIS ROUTINE IS USED BY THE INTEGRATION ROUTINE
C     IT RETURNS THE VALUE OF THE INTEGRAND AT THE REQUESTED
C     VALUES
C
      CALL SETGIJ(IN0,2,GIJ,W,ALPHA)
C
      CALL PROB(IN0,2,DATA,GIJ,ZLE,HIJ,HI,H2,ST,STLE,SIG,PHAT,WA)
C
      ALIKE=ST
C
      RETURN
      END
```

```
      FHI =0.31D0
      DF  =(FHI - FLOW)/(ID-1)
      ALOW=-0.03D0
      AHI =-0.01D0
      DA  =(AHI - ALOW)/(ID-1)
C
C     THIS ROUTINE WILL SET THE NORMALIZATION CONSTANT
C
      ZLE=0D0
C
      CALL SETGIJ(IN0,2,GIJ,0.3D0,-.02D0)
C
      CALL PROB(IN0,2,DATA,GIJ,ZLE,HIJ,HI,H2,ST,STLE,SIG,PHAT,WA)
C
      ZLE=STLE
C
C     INTEGRATE THE STUDENT T-DISTRIBUTION AROUND THE MAXIMUM
C     THIS IS A 24POINT GAUSSIAN QUADRATURE ROUTINE
      AN2=XYINT(0.299D0,0.301D0,1,-.03D0,-.01D0,1,ALIKE)
C
      ZLE=ZLE + DLOG(AN2)
C
C     THIS LOOP EVALUATES THE NORMALIZED DISTRIBUTION
C     ON A 50 BY 50 GRID. THESE POINTS ARE USED IN THE
C     PLOT ROUTINES AND THEY WERE ALSO USED TO INTEGRATE
C     OUT THE DECAY OR THE FREQUENCY PARAMETERS
C
      TOTAL=0D0
      DO 2000 I=1,ID
      DO 2000 J=1,ID
C
      W=(I-1)*DF + FLOW
      ALPHA=(J-1)*DA + ALOW
C
      CALL SETGIJ(IN0,2,GIJ,W,ALPHA)
C
C     EVALUATE THE PROBABILITY DENSITY AT THESE POINTS
      CALL PROB(IN0,2,DATA,GIJ,ZLE,HIJ,HI,H2,ST,STLE,SIG,PHAT,WA)
C
      TOTAL=TOTAL + ESTP*DF*DA
C
      WRITE(7,3333)W,ALPHA,ST,SIG,PHAT
 2000 WRITE(6,3333)W,ALPHA,ST,SIG,PHAT
 3333 FORMAT(5D15.5)
C
      STOP
```

APPENDIX B.

An Example of how to Use Subroutine PROB

The following program was designed and used to prepare one example (the single harmonic frequency with Lorentzian decay) in the text. The steps needed to create this example may be generally described as follows: we read in the data; got an initial estimate of the log normalization constant, then integrate over the probability distribution, update the normalization constant, and evaluate the normalized probability distribution over the desired range of parameter values.

There are two basic steps involved in using subroutine PROB: first one must evaluate the nonorthogonal model functions at the desired values of the parameters; then, the subroutine PROB must be called to evaluate the "student t-distribution" for these parameter settings.

There are three routines in this example: the main line routine performs the steps just described; SETGIJ will evaluate the nonorthogonal model functions at the desired time points for the desired parameter values; ALIKE evaluates the probability density at the parameter values requested by the Gaussian quadrature routine. We have not included the integration routine since such routines are easily available, or easily written if need be.

```
      IMPLICIT REAL*08(A-H,O-Z)
      DIMENSION DATA(512),HIJ(512,2),HI(2),GIJ(512,2),WA(2,2,20)
      COMMON    DATA,HIJ,HI,GIJ,WA,ZLE,IN0
      EXTERNAL  ALIKE
C
C     CALL PROB(IN0,2,DATA,GIJ,ZLE,HIJ,HI,H2,ST,STLE,SIG,PHAT,WA)
C
C     IN0             THE NUMBER OF DATA POINTS
C     2               THE NUMBER OF MODEL FUNCTIONS
C     DATA            THE TIME SERIES
C     GIJ             THE NON-ORTHONORMAL MODEL FUNCTIONS
C     ZLE             LOG BASE E OF THE NORMALIZATION CONSTANT
C     HIJ             THE ORTHONORMAL MODEL FUNCTIONS
C     HI              THE PROJECTIONS OF THE DATA ONTO HIJ
C     H2              THE H**2 BAR STATISTIC
C     ST              STUDENT T-DISTRIBUTION
C     STLE            LOG BASE E OF THE STUDENT T-DISTRIBUTION
C     SIG             THE VARIANCE OF THE DATA
C     PHAT            POWER SPECTRAL DENSITY
C     WA              A WORK AREA USED BY THE SUBROUTINE
C
C
      IN0=512
      READ(8,1000)(DATA(I),I=1,IN0)
 1000 FORMAT(1X,19A4)
C
      ID=50
      FLOW=0.295D0
```

```
      CALL TRANS(INO,NMAX,AIJ,W(IM),W(IT),W(IE),W(IW),W(I2))
      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)
C
       DO 2000 I=1,NMAX
       DO 2000 J=1,NMAX
       TOTAL=0D0
       DO 1000 K=1,IO
 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*****************************************************************
C
C      SET UP THE ORTHOGONAL VECTORS
       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
```

```
 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 TOTA =TOTAL + HIJ(J,I)**2
      ANORM=DSQRT(TOTAL)
      DO 4000 J=1,IN0
 4000 HIJ(J,I)=HIJ(J,I)/ANORM
 5000 CONTINUE
      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= T + NMAX*NMAX
      IM=IE + NMAX*NMAX
      IW=IM + NMAX*NMAX
      I2=IW + NMAX*NMAX
```

$P(\{\omega\}\|DI)$	ST	output	The "student t-distribution" (28) 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. The calling routine in Appendix B has an example of how to do this.
STLE	STLE	output	This is the $\log_e$ of the "student t-distribution" (28). This field is always computed even when the normalization is zero.
$\langle\sigma\rangle$	SIG	output	This is the expected value of the noise variance σ as a function of the $\{\omega\}$ parameters (40) with $s = 1$.
$\hat{p}(\{\omega\})$	PHAT	output	This is the power spectral density (39) as a function of the $\{\omega\}$ parameters.
	WORK	scratch	This work area must be dimensioned $5m^2$. 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 continuous 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. 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 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)
    C   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
      Y2=0D0
      D0 1000 I=1,IN0
```

APPENDIX A.

A Computer Algorithm for Computing the Posterior Probability (28) for an Arbitrary Set of Model Equations.

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. 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, $\overline{h^2}$, $P(\{\omega\}|DI)$, $\langle\sigma\rangle$, 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.
G_{ij}	GIJ	input	This matrix contains the j nonorthogonal model functions [dimensioned as GIJ(INO,IFUN)] and evaluated at t_i.
ZLOGE	ZLOGE	i/o	This is the $\log_e$ of the normalization constant. On the initial call to this routine this field should be initialized to zero. The subroutine never computes the "student t-distribution" when ZLOGE is zero: instead the $\log_e$ 0 the "student t-distribution" is computed. It is up to the user to locate a value of $\log_e[P(\{\omega\}\|DI)]$ 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. For an example of this procedure see the driver routine in Appendix B.
$H_i(t_j)$	HIJ	output	These are orthonormal model functions (17) 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 (24) and (36).
$\overline{h^2}$	H2BAR	output	The sufficient statistic $\overline{h^2}$ (26) is always computed.

B. 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 doing Fourier transform analysis on a data set, then our analysis shows that one will always obtain answers of the form "What is the evidence of a single stationary harmonic frequency in these data?" This will be more than adequate if there are plenty of data and no evidence of complex phenomena. However, if the data show evidence for multiple frequencies or complex behaviors, the Fourier transform gives answers which can be misleading or incorrect in light of more realistic models.

A plot of the log of the "student t-distribution" using this as a model is the statistic to look for chirp. However, we now have two parameters to plot, not one. We have constructed a contour plot around the 11 year cycle, Fig. 16. 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, Fig. 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. 16.

There are 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 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. 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.

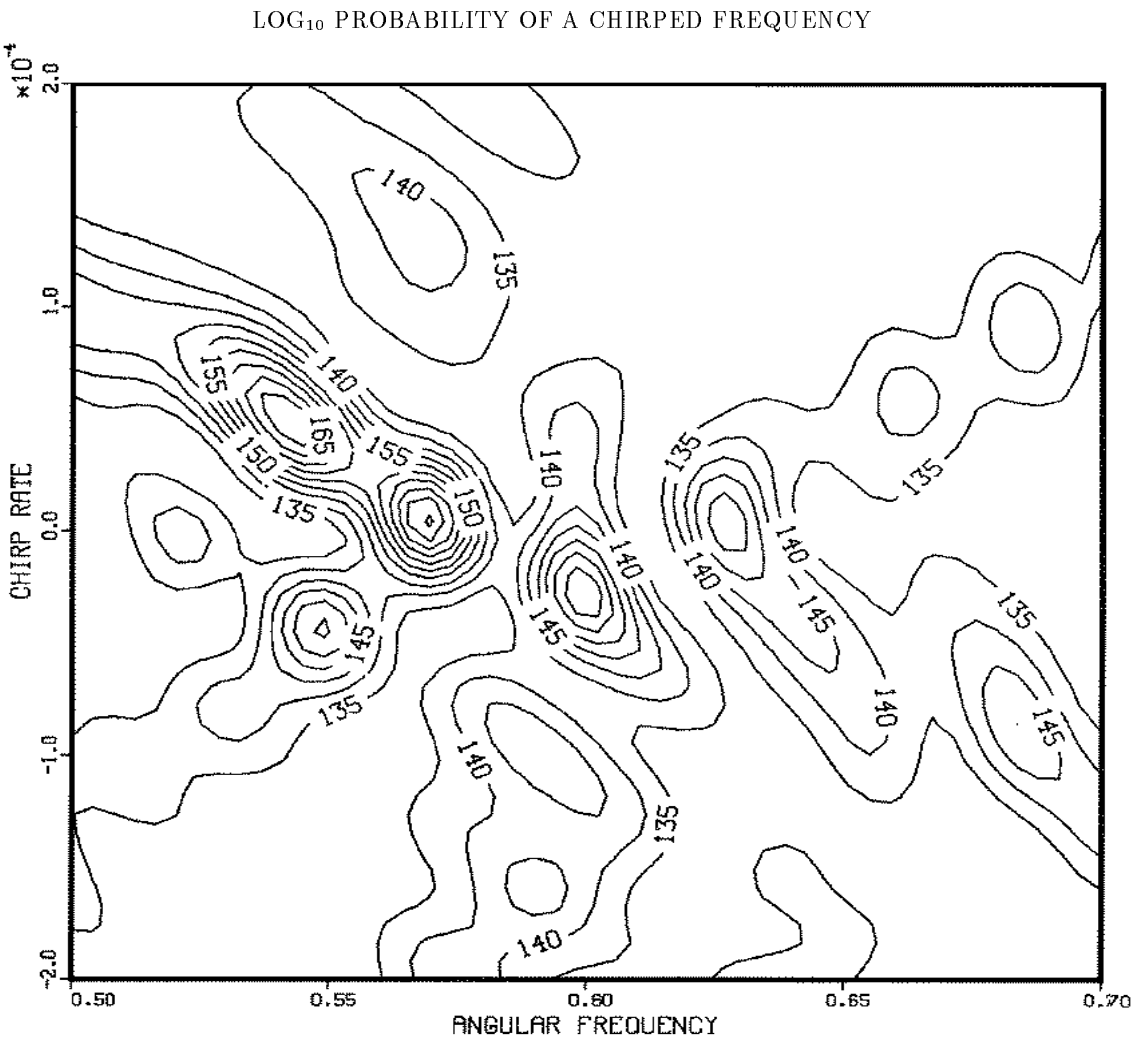
VII. SUMMARY AND CONCLUSIONS.

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

A. Summary

We began this analysis in Section II, by applying probability theory to estimate the spectrum of a data set that, we postulated, contained only a single sinusoid plus noise. In Section III, 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 Section IV, we estimated the nuisance parameters and calculated, among other things, the power spectral density, and the noise variance. In Section V, we specialized to spectrum analysis and explored some of the implications of the "student t-distribution" for this problem. At the end of Section V, we developed a procedure for estimating the accuracy of the $\{\omega\}$ parameters. In Section VI, we applied these analyses to a number of real time series with the aim of exploring 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. In the sunspot example we did not know which model was appropriate, so we applied a number of different models with the intention of discovering as much about them as possible.

Figure 16: 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.

We will take a more general approach and not constrain these amplitudes. We will simply allow probability theory to pick the amplitudes which fit the data best. Thus any result we find will have the Sonnet frequencies but the amplitudes and phases will be chosen to fit the data "better" than the Sonnet model. After integrating out the amplitudes we have only two parameters to determine, $omega_c$ and $omega_m$.

We located the maximum of the posterior probability density using he computer code in Appendix A, and using the pattern search routine. The "best" estimated value for $omega_c$ (in years) is approximately 21.0 years, and ω_m approximately 643 years. The values for these parameters given by Sonnet 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 value of $omega_c$ and $omega_m$ to be substantially different from those given by Sonnet. The most indicative value is the estimated variance for this model $\{\sigma^2\}_{\text{Sonnet}} = 605$. This is worse than that predicted for the simple 12 frequency model by almost a factor of 1.5 and is comparable to the fit achieved by a five frequency model.

We have so far investigated two variations of harmonic analysis on the relative sunspot numbers. Let us proceed to investigate a more complex case to see if there might be more going on in the relative sunspot numbers than just simple periodic behavior. These data have been looked at from this standpoint at least once before. Bracewell [22] has analyzed these numbers to determine if they could have a time-dependent "instantaneous phase." The model used by Bracewell can be written as

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

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 investigation.

An "instantaneous phase" in the notation we have been using is a chirp. Let $\phi(t)$ stand for the phase of the signal, and $omega$ 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 the first derivatives $\phi'(t)$ is zero. If this were not so then ω is not the frequency as presumed here. The Bracewell model can then be approximated as

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

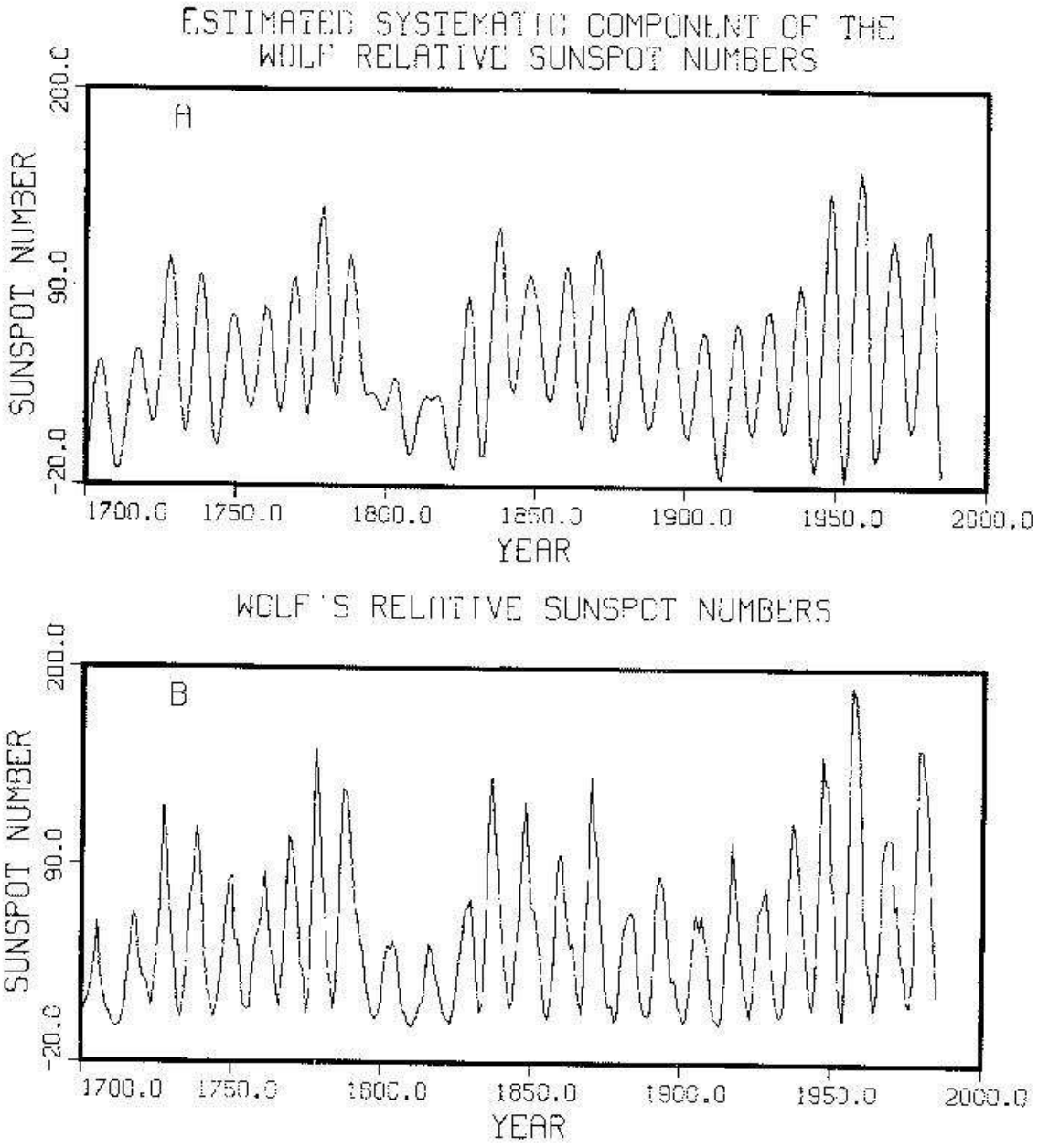
To second order, the Bracewell model is just a chirped frequency with a time varying envelope.

We can investigate the possibility of a chirped signal using

$$f(t) = B_1 + B_2 \cos(\omega t + \alpha t^2) + B_3 \sin(\omega t + \alpha t^2)$$

as the model, where α is the chirp rate, B_1 is a constant component, ω is the frequency of the oscillation, and B_2 and B_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.

Figure 15:



Not only can one obtain the estimated power carried by the signal, one can use the amplitudes to plot what probability theory has taken to be the signal with the noise removed. Of course a reconstruction of this nature is only as good as the model, Fig. 15(A). We have included the relative sunspot numbers, Fig. 15(B), for easy comparison. The predicted series can probably be made better by including some of the smaller cycles we ignored in this analysis.

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 the correct sufficient statistic for estimating a harmonic frequency. At least one of the predicted frequencies 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 addition to the power spectral density we can plot what this model thinks is the sunspot series – less the noise, We have repeated the plot of the sunspot numbers, Fig. 15(B) for comparison.

This simple 12 frequency model reproduces most, but not all of the features of the sunspot numbers. 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 12 frequency model. 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.

We chose to examine 12 frequencies because that was the number of frequencies used in a model proposed by Sonnet [21].

He has attempted to explain these numbers in terms of harmonic frequencies: 180, 90, and 45 are examples of harmonically related frequencies. In 1982, C. P. Sonnet [21] published a small paper in which the sunspot number spectrum was be explained using

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

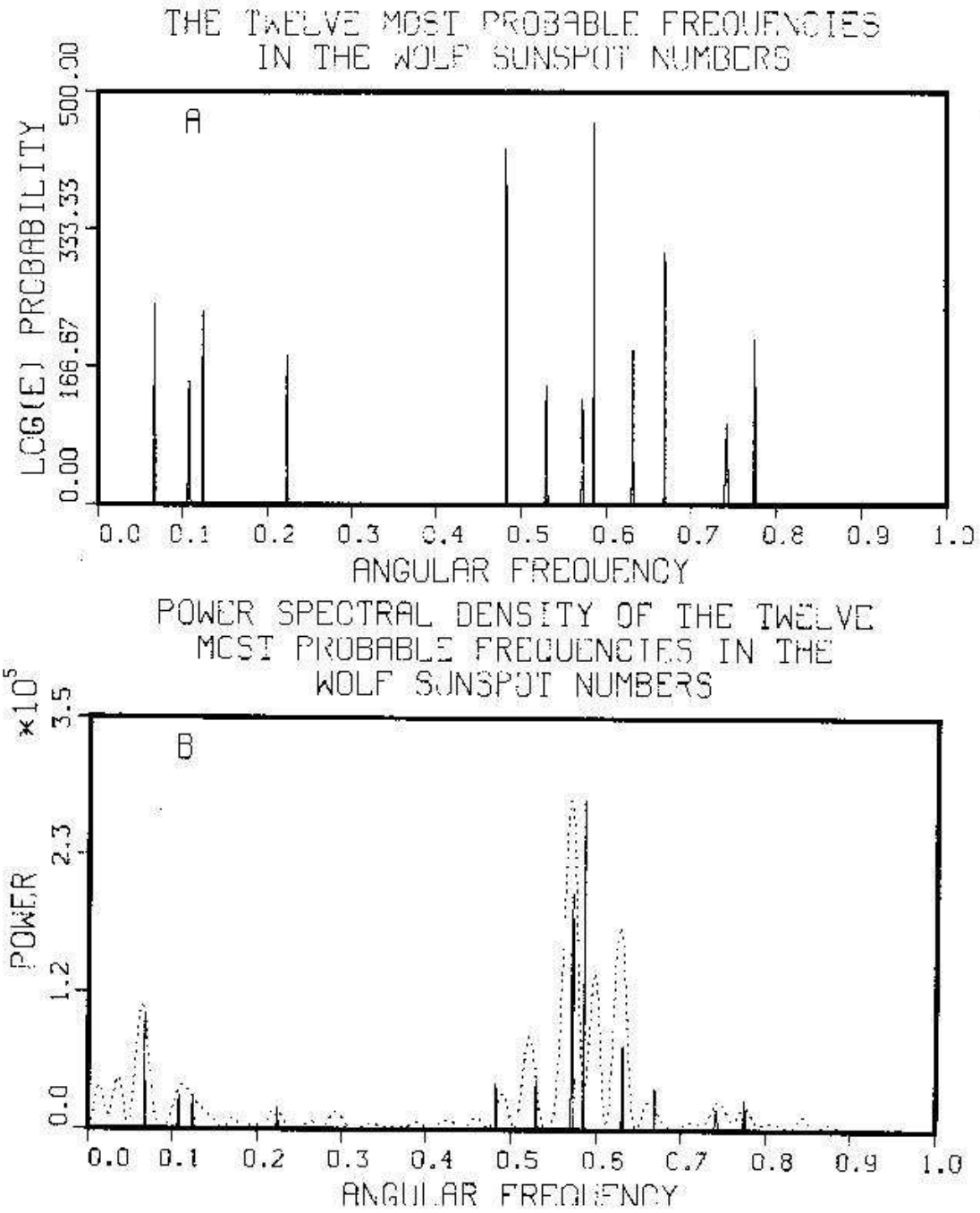
as a model, where Sonnet's estimate of magnetic cycle ω_m is approximately 90 years, and his estimate of the solar cycle ω_c is 22 years. The rectification effect is present here.

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 determine ω_c and ω_m. To do this we first square the term in brackets and then use trigonometric identities to reduce this mode to a form where probability theory can readily estimate the amplitudes and phases:

$$\begin{aligned}
f(t) &= A_1 + A_2\cos([\omega_m]t) && +A_3\sin([\omega_m]t)\\
&+ A_4\cos([2\omega_m]t) && +A_5\sin([2\omega_m]t)\\
&+ A_6\cos([\omega_c - 2\omega_m]t) && +A_7\sin([\omega_c - 2\omega_m]t)\\
&+ A_8\cos([\omega_c - \omega_m]t) && +A_9\sin([\omega_c - \omega_m]t)\\
&+ A_{10}\cos([\omega_c]t) && +A_{11}\sin([\omega_c]t)\\
&+ A_{12}\cos([\omega_c + \omega_m]t) && +A_{13}\sin([\omega_c + \omega_m]t)\\
&+ A_{14}\cos([\omega_c + 2\omega_m]t) && +A_{15}\sin([\omega_c + 2\omega_m]t)\\
&+ A_{16}\cos([2\omega_c - 2\omega_m]t) && +A_{17}\sin([2\omega_c - 2\omega_m]t)\\
&+ A_{18}\cos([2\omega_c - \omega_m]t) && +A_{19}\sin([2\omega_c - \omega_m]t)\\
&+ A_{20}\cos([2\omega_c]t) && +A_{21}\sin([2\omega_c]t)\\
&+ A_{22}\cos([2\omega_c + \omega_m]t) && +A_{23}\sin([2\omega_c + \omega_m]t)\\
&+ A_{24}\cos([2\omega_c + 2\omega_m]t) && +A_{25}\sin([2\omega_c + 2\omega_m]t).
\end{aligned}$$

Now Sonnet specifies the amplitudes, but not the phases [21].

Figure 14:



The posterior probability of twelve frequencies in the relative sunspot numbers Fig. 14(A). When plotted as normalized Gaussians the height represents the accuracy of the estimates not the power carried by them. Figure 14(B) has a power normalization. The peak value of the periodogram is an accurate estimate of the energy carried in a line so long as there are well separated resonances. However, around the 11 year period ($\omega \approx 0.58$) at least one of the estimated frequencies is located at a minimum of the periodogram.

Sonnet's model. We then applied a multiple frequency model using

$$f(t) = B_1 + \sum_{j=1}^{13} \{B_{j+1}\cos(\omega_j t) + B_{2j+1}\sin(\omega_j t)\}.$$

We computed the probability of the frequencies $\{\omega_1, \cdots, \omega_1 3\}$ using the computer code given in Appendix A. The pattern search routine discussed earlier was used to locate the maximum of this 13 dimensional space to five significant digits. Two of the 13 frequencies converged to the same numerical value, indicating that what we thought was two frequencies in Fig. 13(A) was in fact only one frequency. We removed one of these frequencies to obtain a 12 frequency model, and repeated the search using the previous values as our initial estimates. We computed the standard deviation using the procedure developed in Section V, equations (64-65). Last, we used the linear relations between the models (19) to compute the nonorthonormal amplitudes as well as their second moments. These are summarized as follows

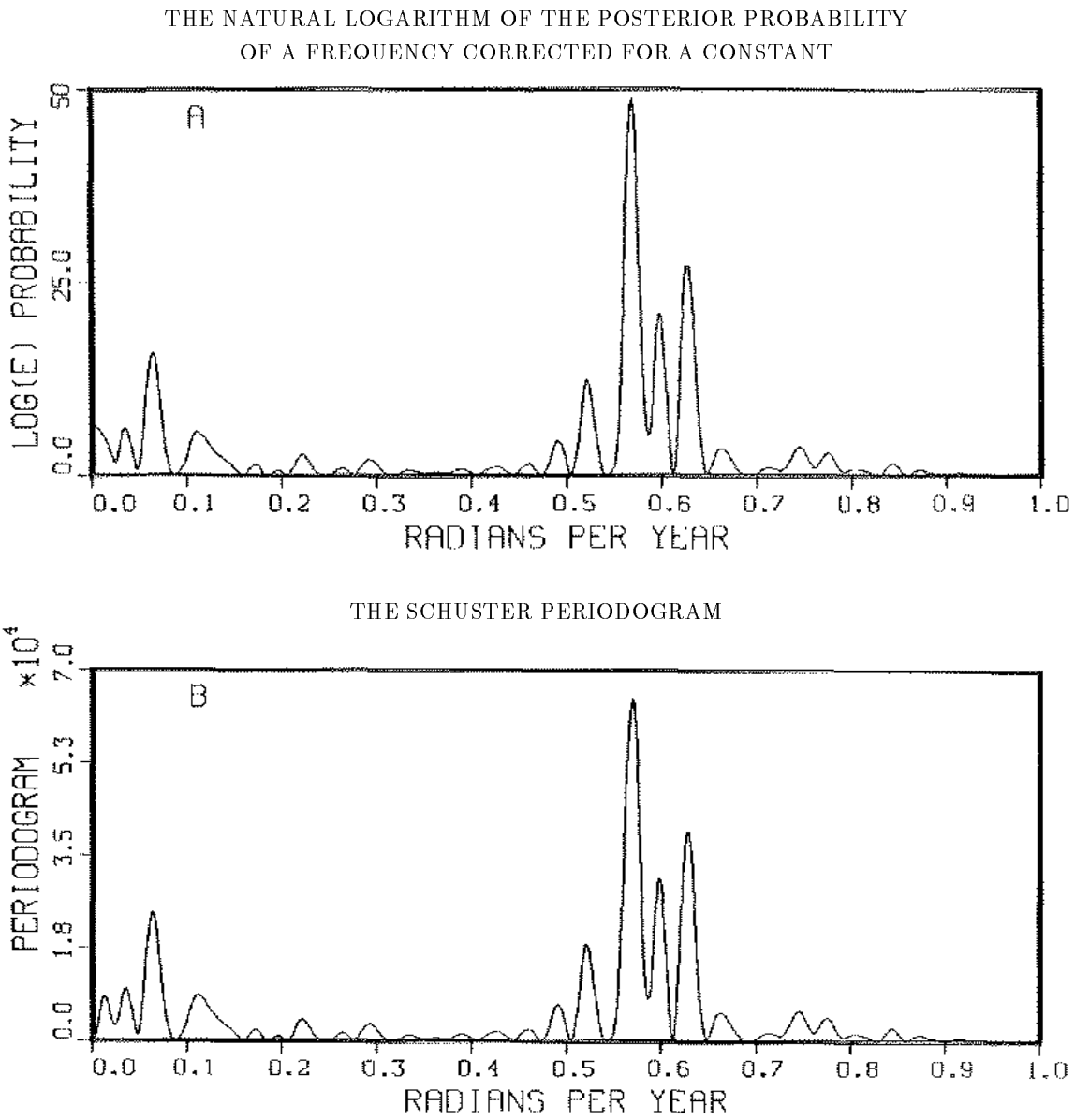
$\langle \hat{f} \rangle_{\text{est}}$	$\langle B_1 \rangle$	$\langle B_2 \rangle$	$\langle B_1^2 + B_2^2 \rangle$
95.29 ± 0.62 years	7.522	17.624	371.01
58.94 ± 0.38 years	11.165	5.247	157.91
51.07 ± 0.18 years	1.499	-8.789	84.73
28.16 ± 0.07 years	6.146	-3.522	54.07
13.03 ± 0.01 years	6.728	-1.637	51.08
11.86 ± 0.02 years	-15.613	-0.840	247.52
10.99 ± 0.02 years	-37.569	-10.329	1521.95
10.75 ± 0.01 years	23.071	-4.526	555.29
9.97 ± 0.01 years	13.509	-11.932	328.87
9.41 ± 0.01 years	1.971	7.038	58.63
8.48 ± 0.01 years	-9.222	4.655	109.81
8.12 ± 0.01 years	1.654	7.254	60.12

With these 12 frequencies and one constant the estimated noise variance is $\langle \sigma^2 \rangle_{\text{est}} = 398$, and the signal-to-noise ratio is 15.2. The constant term had a value of 48.22. We have plotted these 12 frequencies as normalized Gaussians 14(A) to get a better understanding of their determination. The best determined frequency is, of course, the 10.99 ± 0.016 year cycle. When we performed this calculation using the single frequency model our estimate was 11.04 ± 0.02 years; we have moved the estimated frequency over three standard deviations. This illustrates that the periodogram can give misleading estimates when there are multiple close frequencies. However, as long as they are reasonably well separated the estimates should not be off by very much.

We could not verify the 180 year period. We included this one in the original 13 frequencies. However, the pattern search consistently confounded it with the 95 year period. This could be due to poor searching procedures or it could indicate that the data do not show evidence for this frequency. Regardless, this frequency needs to be examined more closely. Additionally, there are a number of other small frequencies on the periodogram; we did not include these even though we suspect they are real frequencies.

We can plot an approximation to the power spectral density just by normalizing 14(A) to the appropriate power level, Fig. 14(B). The dotted line on this plot is the periodogram

Figure 13: Adding a Constant to the Model



The $\log_e$ of the marginal posterior probability of a single harmonic frequency plus a constant 13(A), and the periodogram 13(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 13(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 13(A) does not indicate the presence of a spurious low frequency peak, only a constant.

D. Wolf's relative sunspot numbers.

In 1848 Rudolf Wolf introduced the relative sunspot numbers as a measure of solar activity. These numbers, defined earlier, are available as yearly averages since 1700, Fig. 1(A). The importance of these numbers is primarily due to the fact that 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 discussed earlier. In addition to this cycle a number of others have been reported including 180, 90, 45, and a 22 year cycle as well as a number of others [20, 21].

We will apply probability theory to these numbers to see what can be learned. We must stress that in what follows we have no idea what the "true" model is, but can only examine a number of different possibilities. We begin by asking "What is the evidence for multiple harmonic frequencies in these data?"

These numbers have been analyzed before by many writers. We will contrast our results to those obtained recently by Sonnet [21] and Bracewell [22].

The analysis done by Sonnet concentrated on determining the spectrum of the relative sunspot numbers. He used the Burg [19] 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. 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. 13(A). We include the constant and allow probability theory to remove it the correct way, instead of subtracting the average from the data as was done in Section II. We do this to see if this theoretically correct way of eliminating a constant will make any difference in the predicted frequencies. We plot the log of the marginal posterior probability (28) using

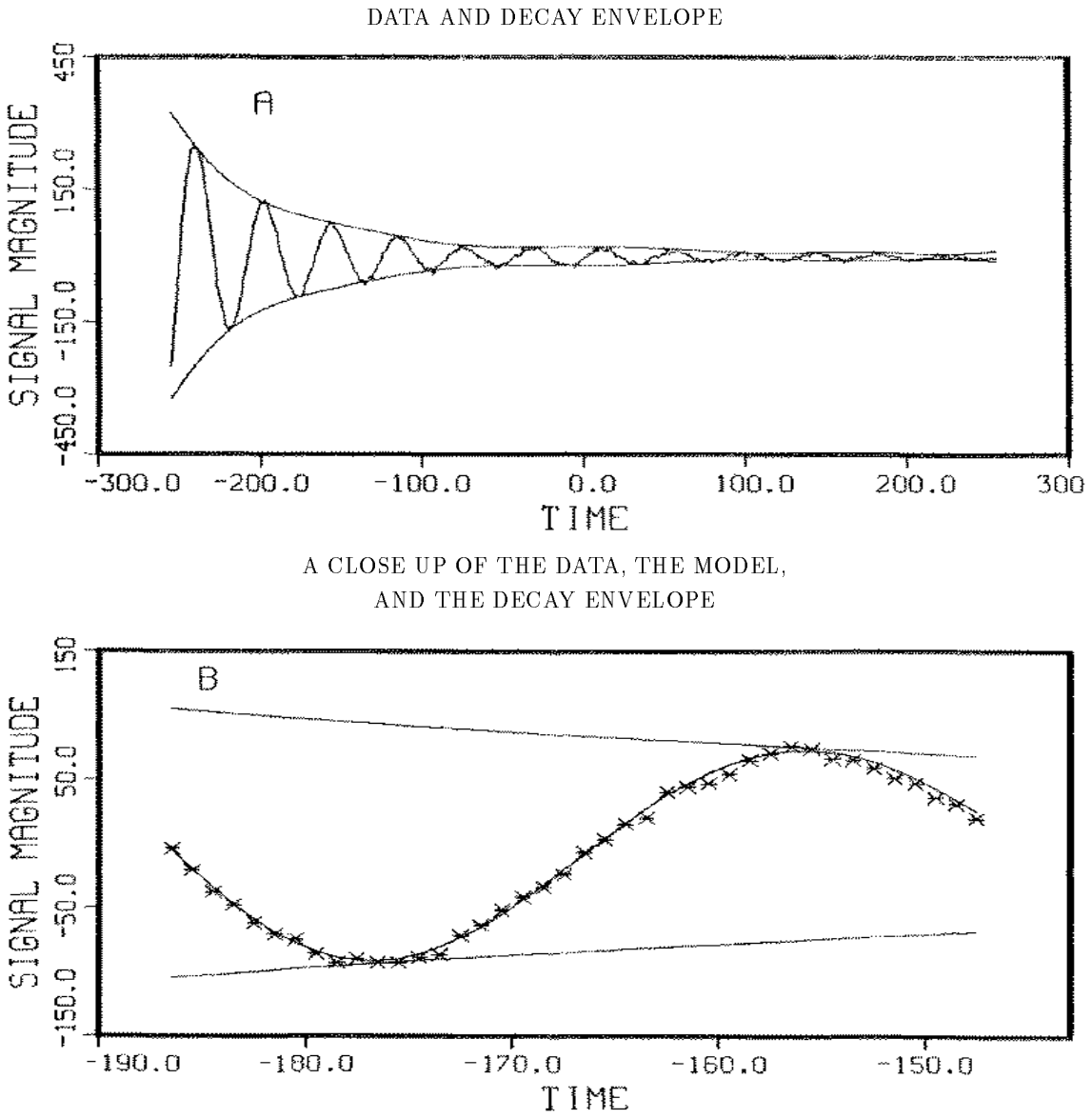
$$f(t) = B_1 + B_2 \cos \omega t + B_3 \sin \omega t$$

as the model. The periodogram, Fig. 13(B), is a sufficient statistic for harmonic frequencies if and only if the time series has zero mean. Under these conditions the periodogram must go to zero at $omega = 0$. In the periodogram, Fig. 13(B), that small 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. 13(A).

Now we examined each of the peaks in Fig. 13(A) with a two frequency model plus a constant to determine if there is any evidence for doublets. There are two, one located near 0.11 and another one near 0.72. Additionally, we examined the low frequency region very closely to see if there was evidence for a very low frequency and found none. The fact that we could not find it does not prove conclusively that the period is not there. We had to search for the peak in a high dimensional space. The peak is small, and the search routine could step over it.

We had to decide which peaks to include in the model; we simply took the 13 largest. We choose 13 because we wanted at least a 12 frequency model to be able to compare to

Figure 12: How Does an NMR Signal Decay?



The decay function in Fig. 12(A) comes down smoothly and then begins to oscillate. This is a real effect, and is not an artifact of the analysis. This type of behavior is characteristic of an inhomogeneous magnetic field. In Fig. 12(B) we have plotted a blow up of the data, the predicted signal, and the decay function.

parameters are to be removed by integration. We chose to eliminate $\{D_j B_1\}$ because there are 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 these $\{D_j B_1\}$ parameters from the linear relations between models (19). For this problem it is computationally simpler to search for the maximum of the probability distribution as a function of frequency *omega* and the ratio B_1/B_2, and then use equation (19) to compute the expansion coefficients. 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 10'th 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 10'th order expansion.

To solve this problem we again used the computer code in Appendix A, and the "pattern" search routine discussed earlier. We located the maximum of the two dimensional "student t-distribution" (28) and used the procedure given in Section V, equations (64-65), to estimate the standard deviation of the parameters. We find these to be

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

$$\left(\frac{B_2}{B_1}\right)_{est} = -.475 \pm 2.7 \times 10^{-3}.$$

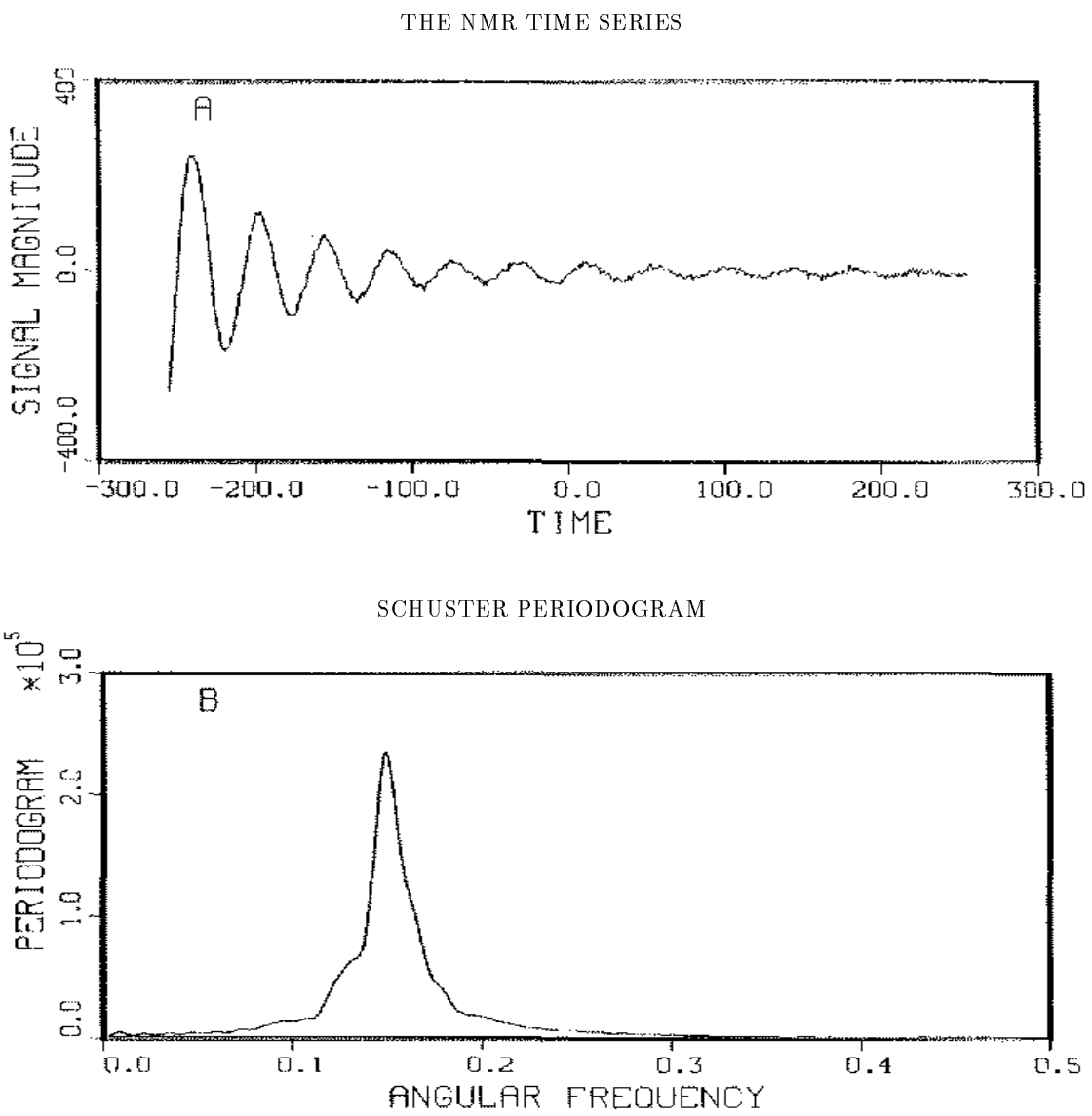
The variance of these data was $\overline{d^2} = 2902$, the estimated noise variance $\langle\sigma^2\rangle_{\rm est} \approx 27.1$, and the signal-to-noise ratio was 23.3.

After locating the maximum of the probability density we used the linear relations (19) between the orthonormal model and the nonorthonormal model to compute the expansion coefficients. As noted earlier there is an arbitrary choice in the scale (magnitude) of the decay function. 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. 12(A). In Fig. 12(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. 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 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. The most likely cause is the inhomogeneous magnetic field, because one can literally change a dial setting on the equipment and get the decay envelope to change shape [14].

In problems with multiple signals, or even with this D_2 signal, when the magnetic field is particularly inhomogeneous the decay function can show much stronger oscillations and even become negative.

Figure 11: A Second NMR Example - Decay Envelope Extraction



These NMR data, Fig. 11(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, Fig. 11(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.

Here the estimated errors represent one standard deviation of the posterior distribution. Generally, it is considered good policy to claim an accuracy corresponding to two standard deviations. Thus, given the spread in the estimates it appears there is indeed evidence for a frequency of a period 20.4 ± 0.2 years.

Now that the effects of removing a trend are better understood, we can proceed to a two frequency model plus a trend to see if we can verify Currie's two frequency results. Figure 10 is a plot of the log of this probability distribution after removing a fifth order trend. The behavior of this plot is the type one would expect when a two frequency model is applied to a data set that contains only one frequency. From this we cannot verify Currie's results. That is, for the three states taken as a whole these data show evidence for a oscillation near 20.4 years as he reports, but we do not find evidence for an 11 year cycle. This does not say that Currie's result is incorrect; he incorporated much more data into his calculation and to check it we would need to include data from at least a dozen more states. While this is a worthy project it is beyond the scope of this simple demonstration.

C. 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. 11(A), in this example are one channel of a pure D_2 spectrum. [14]. Figure 11(B), contains the periodogram for these data. For this demonstration we will use the first $N = 512$ data points because it contains most of the signal.

For D_2, theory indicates there is a single frequency with decay [15].

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

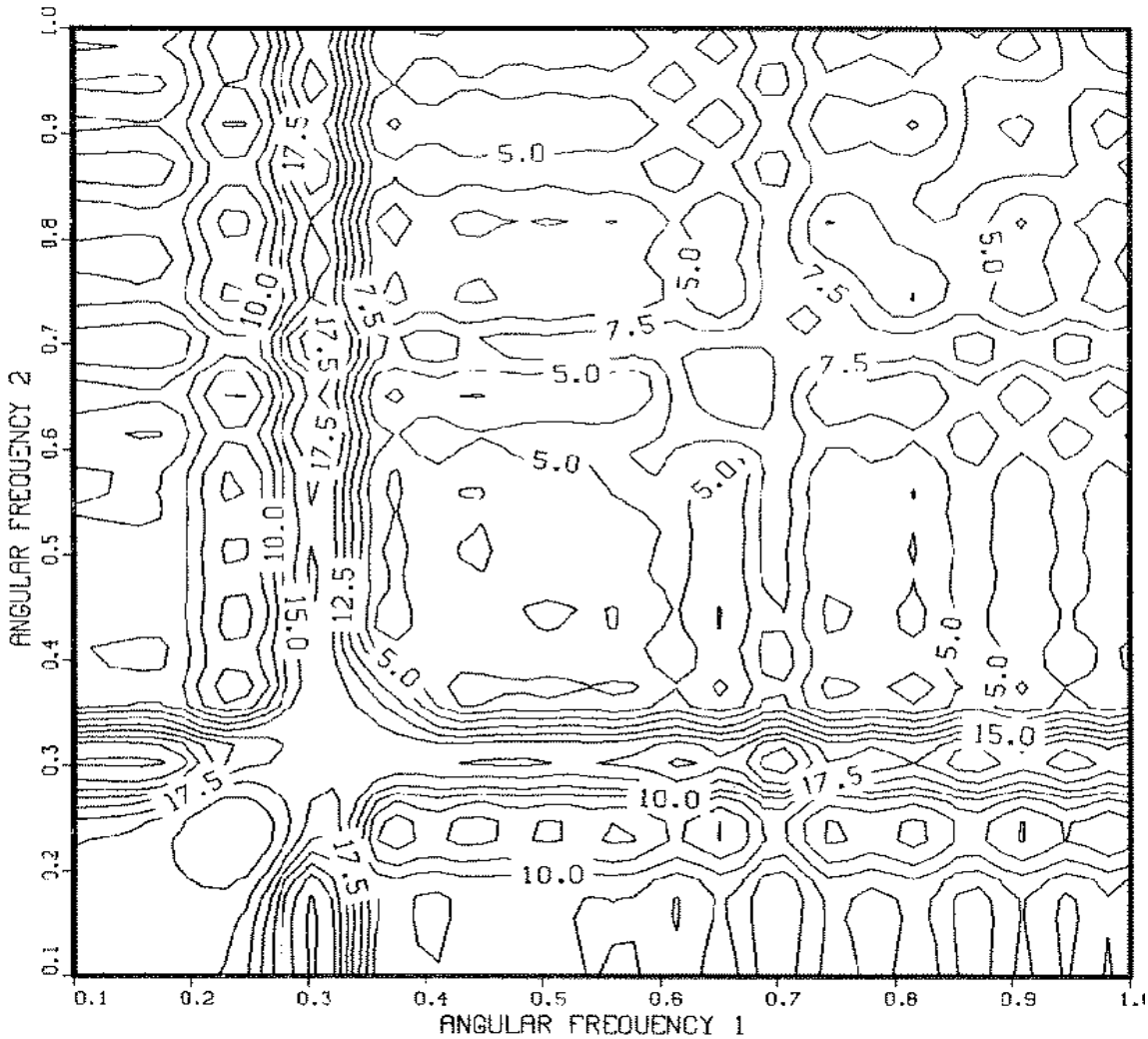
$$f(t) = \{B_1 \sin(\omega t) + B_2 \cos(\omega t)\} \sum_{j=0}^{r} D_j L_j(t)$$

where D_j are the expansion coefficients for the decay function, B_1 and B_2 are effectively the magnitude 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

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

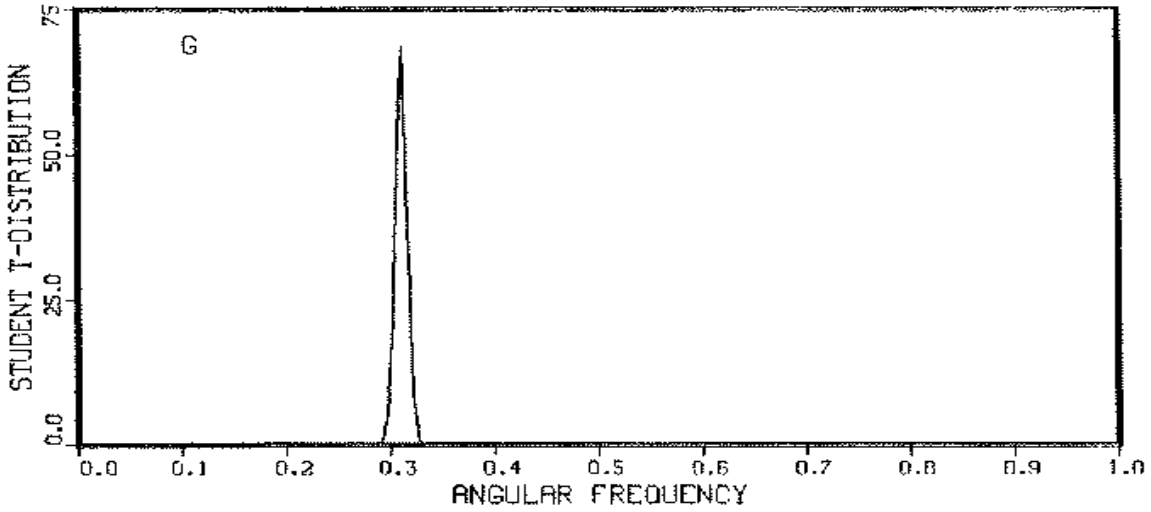
There is an indeterminacy in the overall scale. That is, the amplitude of the sinusoid and the amplitude of the decay $D(t)$ cannot both be determined. One of them is necessarily arbitrary. We chose the amplitude of the sine term to be one because it effectively eliminates one $\{\omega\}$ parameter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

Figure 10: Probability of Two Frequencies After Trend Correction

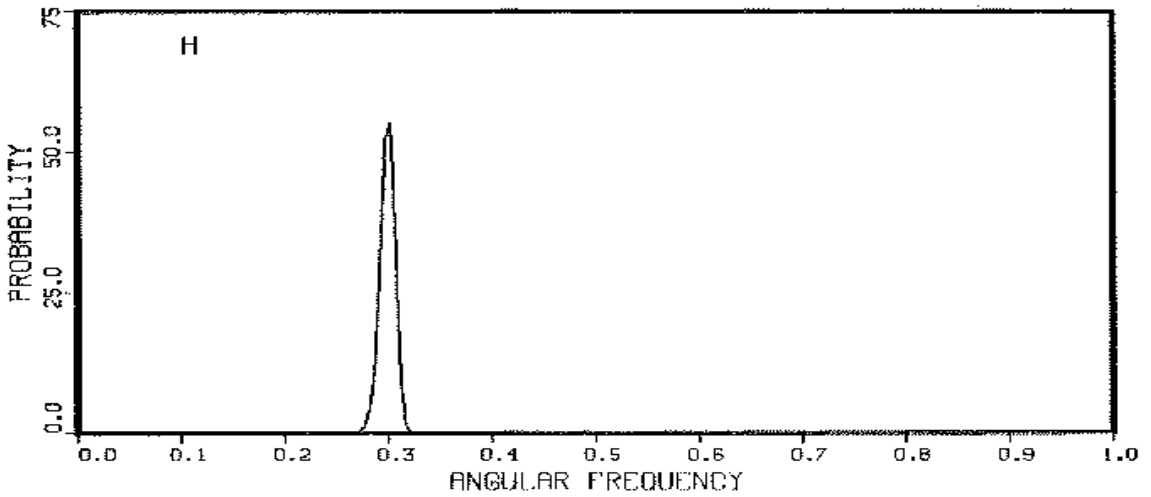


This is the natural logarithm of the probability of two common harmonic frequencies in the crop yield data with a fifth order trend. This type of structure is what one expects from the sufficient statistic when there is only one frequency present. Notice the maximum is located roughly along a vertical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

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

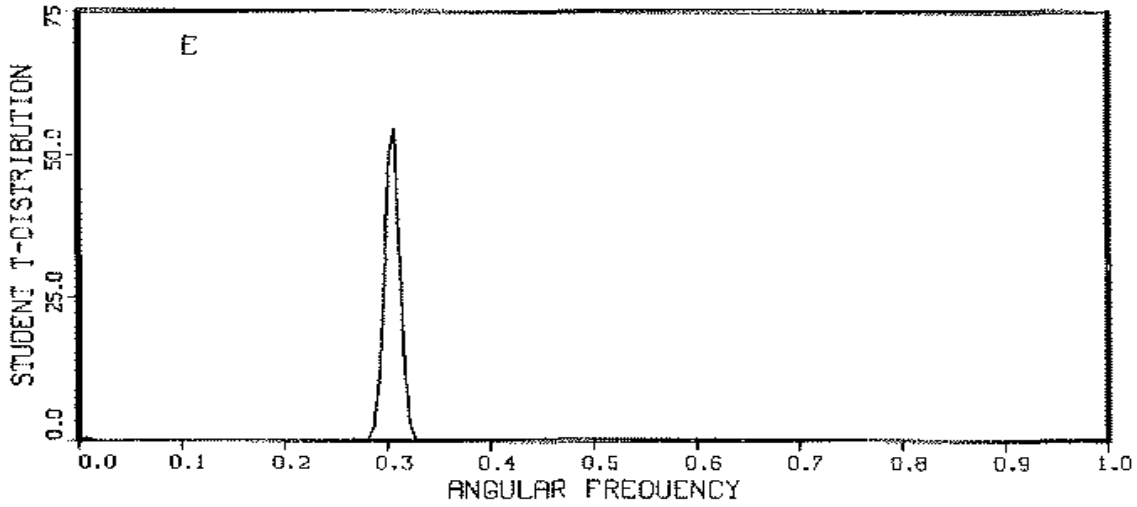


PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A SEVENTH ORDER TREND CORRECTION

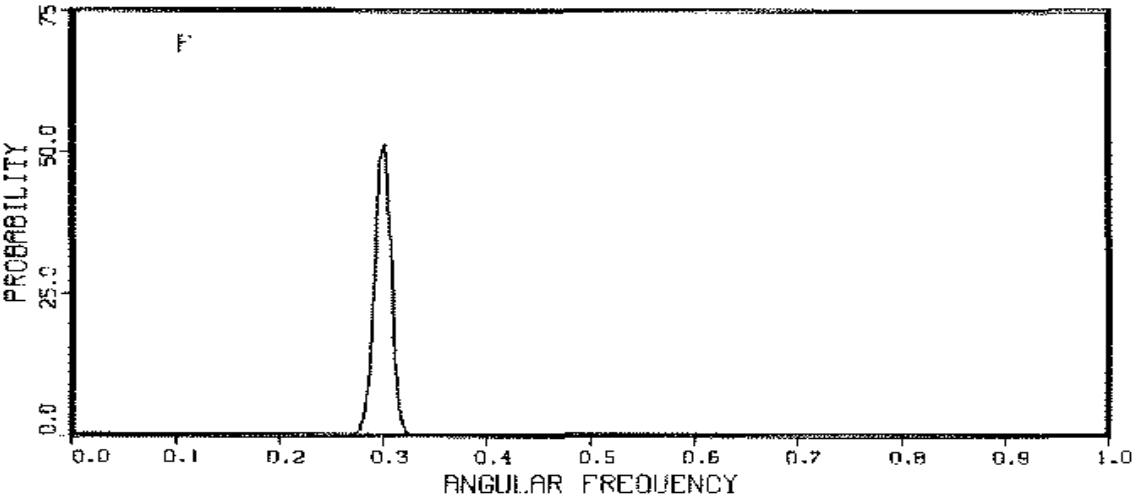


Increasing the expansion order further does not significantly affect the estimated frequency (G) and (H). If the expansion order is increased sufficiently, the expansion will begin to remove the harmonic oscillation; and the posterior probability density will gradually decrease in height.

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

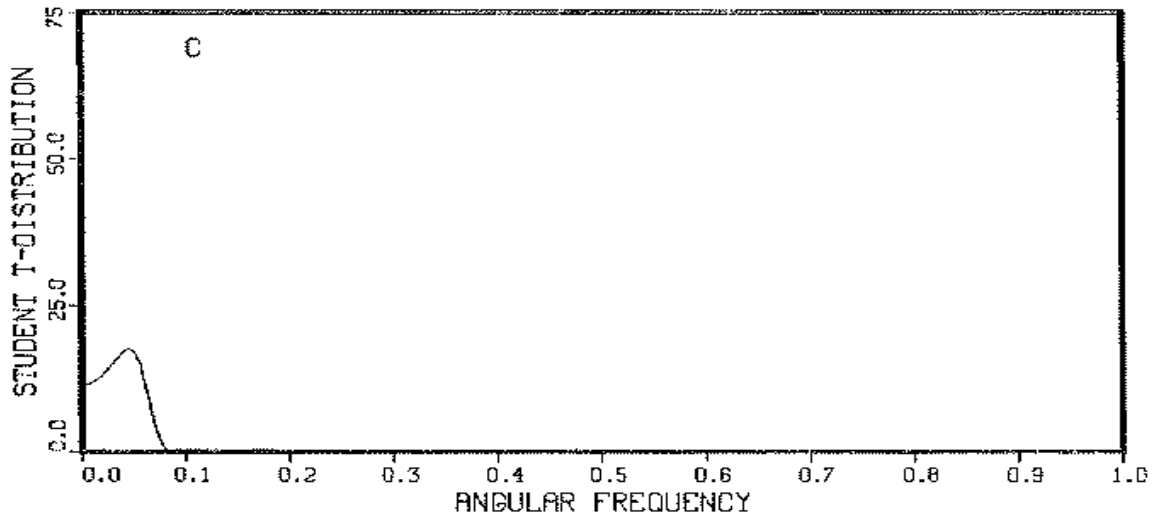


PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A FIFTH ORDER TREND CORRECTION

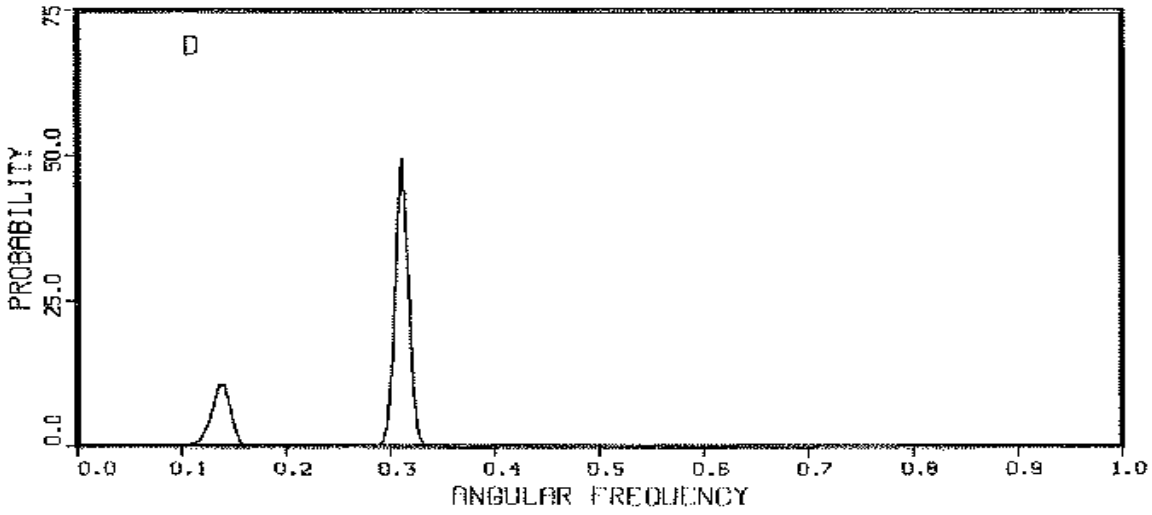


When the probability of a fourth-order trend plus a harmonic frequency is computed the trend is now completely gone and only the frequency at 20 years remains (E). When the expansion order is increased in (F) the frequency estimate is not essentially changed.

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

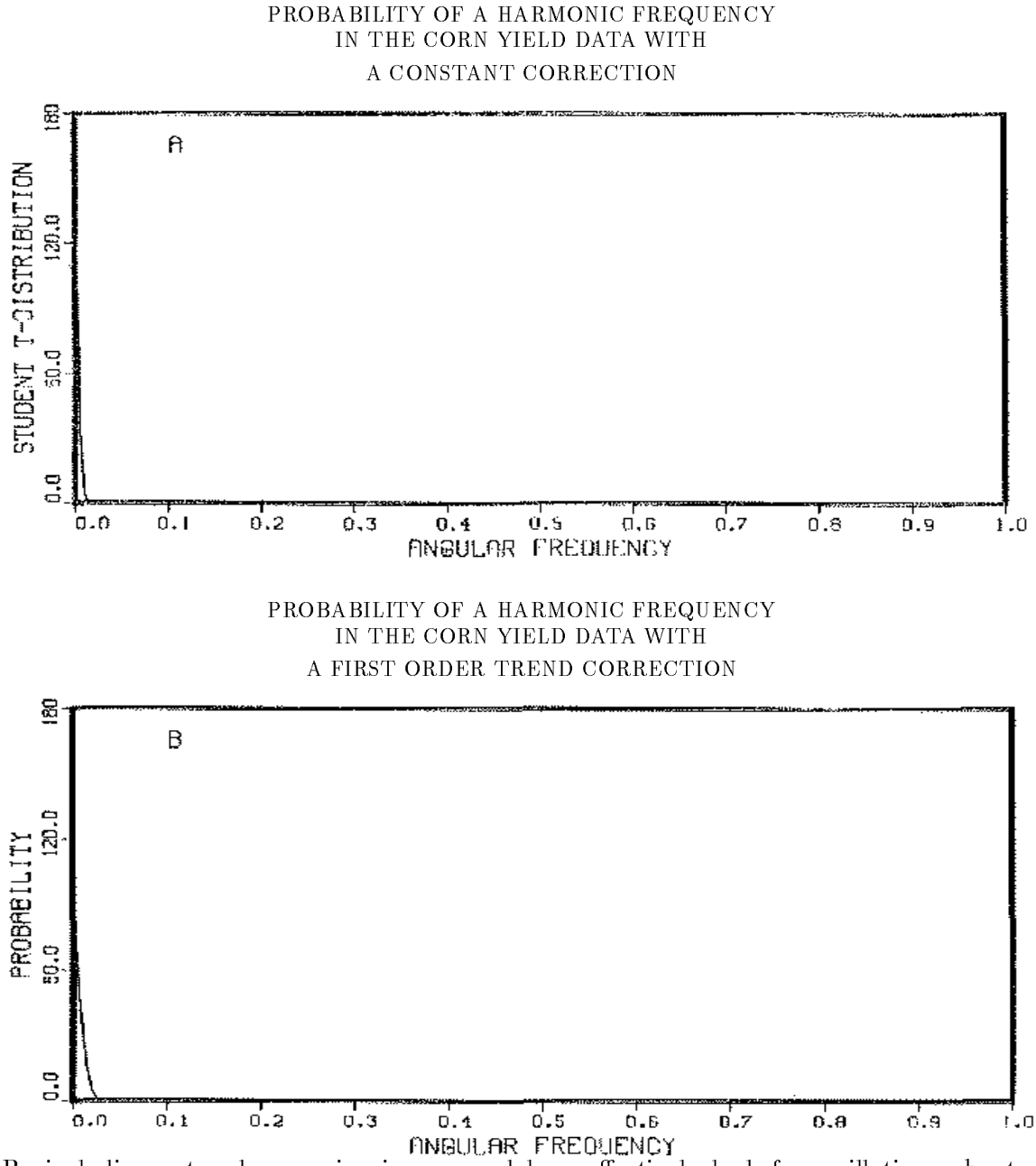


PROBABILITY OF A HARMONIC FREQUENCY
IN THE CORN YIELD DATA WITH
A THIRD ORDER TREND CORRECTION



The probability of a single harmonic frequency plus a second-order trend (C) continues to pick out the low frequency trend. However, the level and spread of the marginal posterior probability density is such that the trend has almost been removed. When the probability of a single harmonic frequency plus a third-order trend is computed, the probability density suddenly changes behavior. The frequency near 0.3 is now the dominant feature (D). The trend has not been completely removed; a small artifact persists at low frequencies.

Figure 9: The Joint Probability of a Frequency Plus a Trend



By including a trend expansion in our model we effectively look for oscillations about a trend. This is not the same as detrending, because the trend functions and the sine and cosine functions are never orthogonal. The zero order trend (or constant) plus a simple-harmonic-frequency model (A) is dominated by the trend. When we included a linear trend the height of the trend is decreased some, however the trend is still the dominant effect in the analysis.

of each "student t-distribution" (28) was computed, and added to obtain the log of the posterior probability of a single common harmonic frequency, Fig. 8(D).

What we would like to know is, "Are those small bumps in Fig. 8(D) indications of periodic behavior, or are they artifacts of the noise or trend?" To attempt to answer this, consider the following model function

$$f_j(t) = T_j(t) + B_{j,1}\cos(\omega t) + B_{j,2}\sin(\omega t)$$

where we have augmented the standard frequency model by a trend $T_j(t)$. The only parameter of interest is the frequency ω. The trend $T_j(t)$ is a nuisance function. To eliminate the nuisance function $T_j(t)$ we expand the trend in orthonormal polynomials $L_j(t)$. These orthonormal polynomials could be any complete set. We use the Legendre polynomials with an appropriate scaling of the independent variable to make them orthonormal on the region $(-49.5 \leq t \leq 49.5)$. This is the range of values used for the time index in the sine and cosine terms. After expanding the trend, the model function for the j'th measurement can be written

$$f_j(t) = \sum_{k=0}^{r} B_{j,k+1}L_k(t) + B_{j,r+2}\cos(\omega t) + B_{j,r+3}\sin(\omega t).$$

Notice, that if $r \leq\!=\!\leq 0$ the problem is reduced to the previous problem (68). The cosine and sine model functions have been renumbered to remain consistent with the notation used earlier.

The expansion order r must be set to some appropriate value. From looking at these data one sees that it will take at least a second order expansion to remove the trend. The actual expansion order for the trend is unknown. However, it will turn out that the estimated frequencies are insensitive to the expansion order, as long as the expansion is sufficient to represent the trend without representing the signal of interest. Of course, different orders would have very different implications about other questions than the ones we are asking; for example, predicting the future trend. That is an altogether more difficult problem than the one we are solving.

The effects of increasing the expansion order r can be demonstrated by plotting the posterior probability for several expansion orders; see Fig. 9(A) through Fig. 9(H). For expansion orders zero, Fig. 9(A) through expansion order 2, Fig. 9(C) the trend has not been removed: the posterior probability continues to pick out the low frequency trend. When a third order trend is used, Fig. 9(D) a sudden change in the behavior is seen. The frequency near $\omega \approx 0.31$ suddenly shows up, along with a spurious low frequency effect due to the trend. In expansion orders four through seven, Fig. 9(E) through Fig. 9(H) the trend has been effectively removed and the posterior probability indicates there is a frequency near 0.31 corresponding to a 20.4 year period.

The amount of variability in the frequency estimates as a function of the expansion order will show how strongly the trend expansion is affecting the estimated frequency. The frequency estimates for the fourth through seventh order expansions are

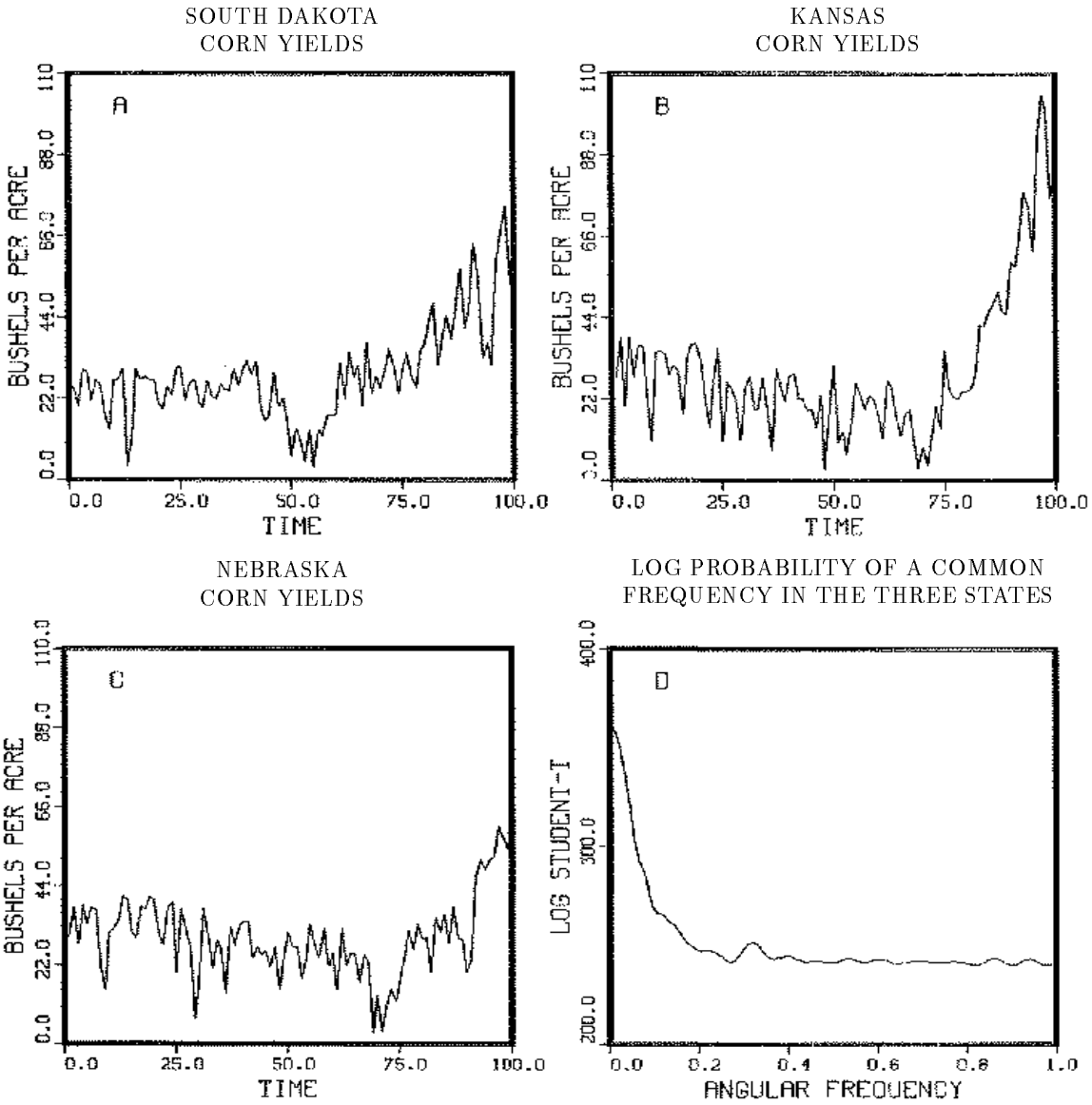
$$(f_4)_{est} = 20.60 \pm 0.08 \text{ years}$$

$$(f_5)_{est} = 20.47 \pm 0.09 \text{ years}$$

$$(f_6)_{est} = 20.20 \pm 0.07 \text{ years}$$

$$(f_7)_{est} = 20.47 \pm 0.09 \text{ years}.$$

Figure 8: Corn Crop Yields for Three Selected States



The three data sets analyzed were corn yields in bushels per acre for South Dakota, Fig. 8(A), Kansas, Fig. 8(B), and Nebraska, Fig. 8(C). The log probability of a single common frequency plus a constant is plotted in, Fig. 8(D). The question we would like to answer is "Is that small bump located at approximately 0.3, corresponding to a 20 year period, a real indication of a frequency or is it an artifact of the trend?" The sharp up turn in the yields occurs at about 1940 and is due to improved varieties, irrigation, etc.

B. Economic data: Corn crop yields.

Economic data are hard to analyze, in part because the data are frequently contaminated by large spurious effects, and the time series are often very short. Here we will examine one example of economic data to demonstrate how to remove some unknown and spurious effects. In particular, we will analyze one hundred year's worth of the corn crop data from three states (Kansas, South Dakota, and Nebraska), Fig. 8(A) through Fig. 8(C) [17].

We would like to know if there is any indication of periodic behavior in these data.

These data have been analyzed before. Currie [18] used a high pass filter and then applied the Burg algorithm [19] to the filtered data. Currie finds one frequency near 20 years which is attributed to the lunar 18.6 year cycle, and another at 11 years, which is attributed to the solar cycle.

There are three steps in Currie's analysis that are troublesome. First, the Burg algorithm is not optimal in the presence of noise (although it is for the problem it was formulated to solve). The fact that it continues to work means that the procedure is reasonably robust; that does not change the fact that it is fundamentally not appropriate to this problem [19].

Second, one has doubts about the filter: could it suppress the effect one is looking for or introduce other spurious effects? Third, to apply the Burg algorithm when the data consist of the actual values of a time series, the autoregression order (maximum lag to be used) must be chosen and there is no theoretical principle to determine this choice. We do not mean to imply that Currie's result is incorrect; only that it is provisional. We would like to apply probability theory as developed in this paper to check these results.

The first step in a harmonic analysis is simply to plot the data, Fig. 8(A) through Fig. 8(C) and the log of the posterior probability of a single harmonic frequency. In the previous example we generalized the analysis for two measurements. The generalization to an arbitrary number of measurements is just a repeat of the previous arguments and we give the result here for any number of measurements

$$P(\omega|DI) \propto \prod_{j=1}^{r} \left\{ \left[1 - \frac{m_j \overline{h^2}_j}{N_j \overline{d^2}_j} \right]^{\frac{m_j - N_j}{2}} \right\}. \tag{67}$$

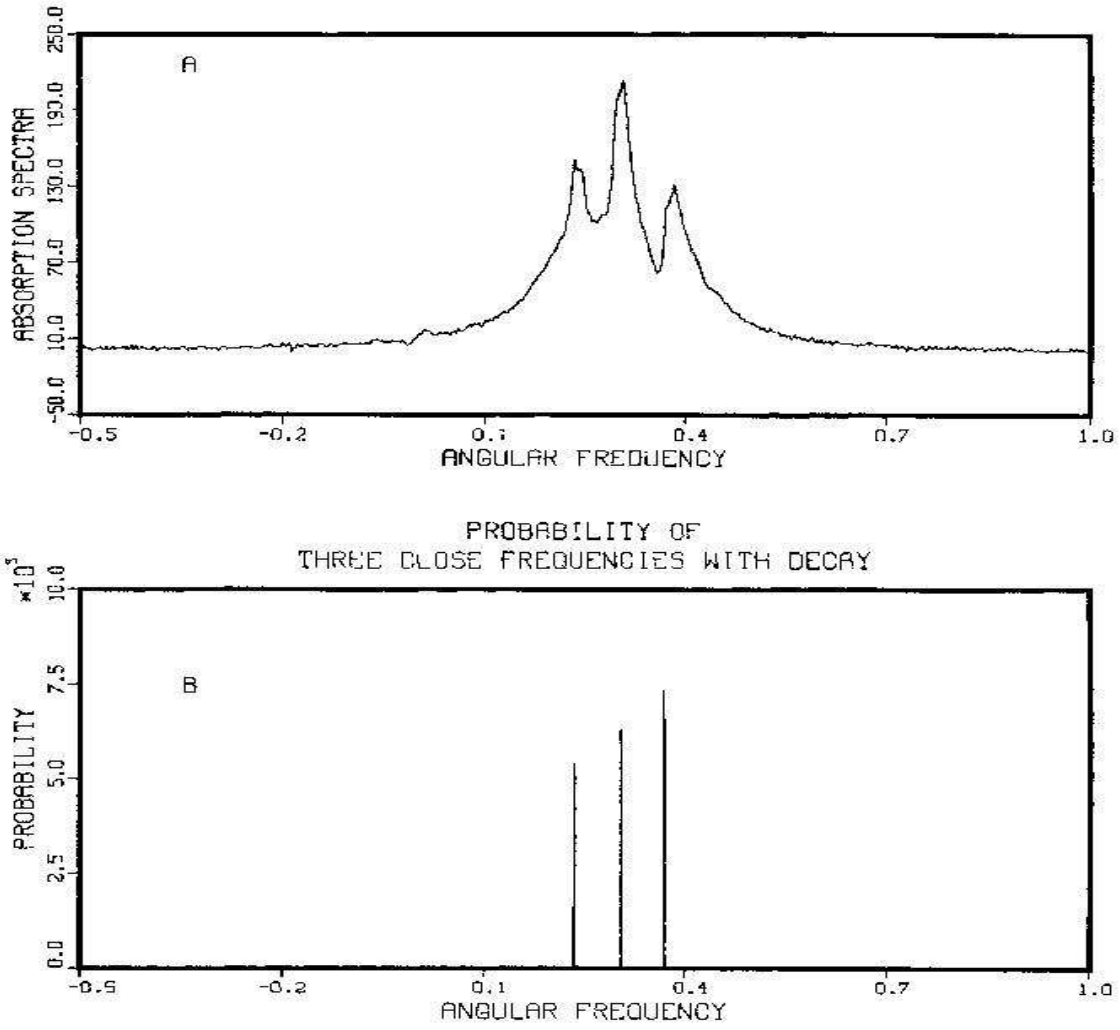
The subscripts refer to the j'th measurement, and each of the models have m_j amplitudes, and each data set contains N_j data values. Additionally it was assumed that the noise variance σ_j was unknown and possibly different for each measurement. The "student t-distributions" (28) for each measurement should be computed separately, thus estimating and eliminating the effects particular to one measurement, and then multiplied to obtain the posterior probability, for the common effects (67). As discussed earlier, this procedure leads to conservative estimates; if we incorporated more prior information (for example, if it were known that the σ_j are all equal) we would obtain slightly better results.

For this harmonic analysis we take the model to be a single sinusoid which oscillates about a constant. The model for one measurement may be written

$$f_j(t) = B_{j1} + B_{j2}\sin(\omega t) + B_{j3}\cos(\omega t). \tag{68}$$

We allow B_{j1}, B_{j2}, and B_{j3} to be different for each measurement; thus there are a total of nine amplitudes, one frequency, and three noise variances. To compute the posterior probability for each measurement, we used the computer code in Appendix A. The log

Figure 7: Comparison to an Absorption Spectrum



The absorption spectrum, Fig. 7(A) can be obtained from two channels by assuming the channels have the same amplitude and are 90° out of phase. A complex Fourier transform is calculated using one channel as the real and the other as the imaginary part of the signal. The global phase of the complex transform is adjusted until the real par has the largest possible area. The real part is the absorption spectrum. Of course the channels do not have the same amplitude and are not exactly 90° out of phase. It requires an extensive procedure to put these two channels into a usable form. Using the full-width at half maximum of the absorption spectrum to determine the accuracy estimate and converting to physical units one may determine the frequencies to about ±15 Hz. The probability analysis, Fig. 7(B) used a three frequency model with two decay rates. The estimated accuracy is ±0.02 Hz.

dimensional parameter space until we located the maximum of the distribution by the "pattern" searching procedure noted before. We then used the procedure given in Section V, equations (64-65), to estimate the standard deviation of the parameters. The derivatives which appear in this procedure were evaluated numerically. The results of this calculation are:

Parameter	estimate	$\pm$	standard deviation
ω_1	74.51	$\pm$	0.03 Hz
ω_2	96.68	$\pm$	0.02 Hz
ω_3	117.38	$\pm$	0.02 Hz
α_1	4.02	$\pm$	0.02 Hz
α_2	5.07	$\pm$	0.02 Hz

We also estimated the signal-to-noise ratio (42) for each channel:

$$\frac{\text{Signal}}{\text{Noise}} \text{ channel } 1 = 14.5$$

$$\frac{\text{Signal}}{\text{Noise}} \text{ channel } 2 = 14.2$$

and the estimated variance (40) with $s = 1$:

$$\langle\sigma\rangle \text{ channel } 1 \quad = \quad 10.9$$

$$\langle\sigma\rangle \text{ channel } 2 \quad = \quad 10.7.$$

The actual frequencies are of little importance in this experiment (the values are controlled by how close to 55Mhz a local oscillator is set). What is important is the relative separation of the three frequencies. The separation for the HD doublet $(\omega_3 - \omega_1)$ is 42.87 ± 0.04 Hz and the D_2 frequency is displaced from the center by $\Delta = \omega_2 - (\omega_3 + \omega_1)/2 = 0.74 \pm 0.04$ Hz. The separation frequency is in excellent agreement with previous measurements and with theory [16, 15].

The HD and D_2 components of the signal are known to have very different decay rates, [15] yet the values indicated by probability theory are nearly the same. We conclude that the inhomogeneous magnetic field has significantly affected the decay rates. The decay is substantially magnet limited.

We can compare these estimates directly to the absorption spectrum, Fig. 7(A). The absorption spectrum resolves the three frequencies. However, they are very close together. The reason the analysis of this experiment is so difficult with the absorption spectrum is that the full-width at half maximum for the D_2 peak, Fig. 7(A), is 16Hz. Figure 7(B) gives the estimates from (66). We have plotted these estimates as three normalized Gaussians each centered at the estimated frequency and having the same standard deviation as the estimated frequency. Clearly the resolution of these frequencies is much improved. With separately normalized distributions, the heights in Fig. 7(B) are indications of the accuracy of the three estimates, not of the power carried by the signal.

Fig. 6 we see there are a number of peaks near 0.3. There are many more peaks than theory indicates there should be. Is this evidence of more going on than theory predicts?

To answer this question we proceed to the next phase of the analysis and apply a two frequency model to each of these peaks. We know that these frequencies have some type of decay structure, we include a decay by adding an exponential factor. For this preliminary analysis we assume the same decay rate for all the frequencies. The model used in this investigation was

$$f_1(t) = [B_1 \cos(\omega_1 t) + B_2 \sin(\omega_1 t) + B_3 \cos(\omega_2 t) + B_4 \sin(\omega_2 t)]e^{-\alpha t}.$$

After searching each of these peaks we found there is one center frequency at $\omega \approx 0.3$ and two others at $\omega \approx 0.25$ and $\omega \approx 0.35$. On the periodogram, therefore, the two highest peaks are not indicative of two frequencies but of a single frequency located at the minimum between them. Here we have the opposite effect from what we saw in Section IV; there we had only one peak in the periodogram, but finer analysis showed that there were two close frequencies present. Here, we have two peaks in the periodogram, but finer analysis shows only one frequency to be present. These examples just illustrate one of the major results of this work: If one asks a question about a single harmonic frequency, when the data have evidence of multiple complex phenomena, one can get answers which are misleading or simply incorrect in the light of more realistic models. Peaks in the Fourier transform are not always an indication of the frequencies present.

We investigated this splitting in the periodogram further. It is caused by a phase reversal in the signal. That is, if the signal is a simple cosine times a decay function $D(t)$, then the splitting can appear when $D(t)$ becomes negative. This type of feature can be present in an NMR signal because due to magnetic field inhomogeneity, the "line" may be a complex superposition of many small signals which have slightly different frequencies.

The next step in the analysis is to construct a plausible model for the data and apply it. As was demonstrated earlier, the exact decay model is not needed to obtain good estimates of a frequency. What is needed is a decay model which is reasonable for the phenomenon being observed: one which decays down to the noise level on the same time scale as the "true" decay. We simply assume the decay is magnet limited and that the decay is Lorentzian. The model we use has three frequencies and two decay rates,

$$\begin{aligned} f_1(t) &= [B_1 \cos(\omega_1 t) + B_2 \sin(\omega_1 t)] \exp[-\alpha_1 t] \\ &+ [B_3 \cos(\omega_2 t) + B_4 \sin(\omega_2 t)] \exp[-\alpha_2 t] \\ &+ [B_5 \cos(\omega_3 t) + B_6 \sin(\omega_3 t)] \exp[-\alpha_1 t] \end{aligned}$$

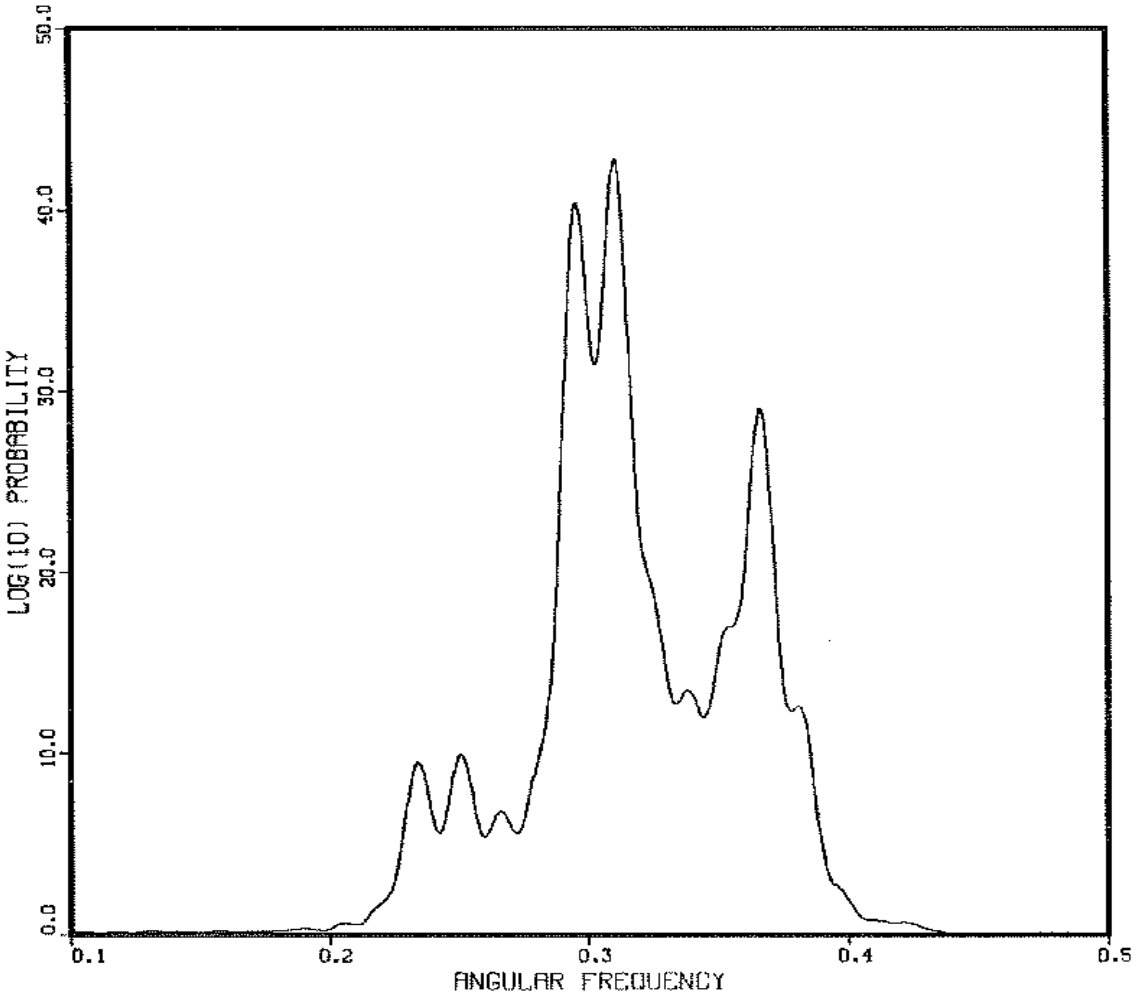
and similarly for channel 2

$$\begin{aligned} f_2(t) &= [B_7 \cos(\omega_1 t) + B_8 \sin(\omega_1 t)] \exp[-\alpha_1 t] \\ &+ [B_9 \cos(\omega_2 t) + B_{10} \sin(\omega_2 t)] \exp[-\alpha_2 t] \\ &+ [B_{11} \cos(\omega_3 t) + B_{12} \sin(\omega_3 t)] \exp[-\alpha_1 t]. \end{aligned}$$

There are five $\{\omega\}$ parameters, 12 amplitudes, and two noise variances. Probability theory has eliminated 14 of the nineteen parameters. We are primarily interested in the three frequencies; however, the decay rates will tell us just how strongly the inhomogeneous magnetic field is affecting the decay rates.

The computer code in Appendix A was used to evaluate the "student t-distribution" (28) for each channel, and these were multiplied to obtain (66). We searched in the five

Figure 6: The Log_{10} Probability of One Frequency in Both Channels



When more than one channel is present, the periodogram is not the proper statistic to be analyzed for indications of a simple harmonic frequency. The analysis for multiple channels indicates the proper statistic is essentially the sum of the periodograms for each channel weighted by the mean-square variance of the data. A phase reversal in this data produces the splitting in the central peak.

second. As was discussed earlier we are using dimensionless units. The relation to physical units are

$$f = \frac{\omega}{2\pi\Delta T} \text{ Hz} \qquad \text{Period} = \frac{2\pi\Delta T}{\omega} \text{ Seconds}$$

where f is the frequency in Hertz, ω is the unitless frequency in radians per unit step, and ΔT is the sampling time.

In these data there are a number of effects which we would like to investigate. First, the indirect J coupling [15] in the HD produces a doublet with a splitting of about 43Hz. The D_2 in the sample is also excited, its resonance is approximately in the middle of the HD doublet. One of the things we would like to determine is the shift of this frequency relative to the center of the HD doublet. In addition to the three frequencies there are two different characteristic decay times; the decay rate of the HD doublet is grossly different from that of D_2 [15].

However, an inhomogeneous magnetic field could mask the true decay: the decay could be magnet limited. We would like to know how strongly the inhomogeneous magnetic field has affected the decay.

The analysis we did in Section III, although general, did not use a notation appropriate to two channels. We will need to generalize the notation briefly; this is very straightforward and we will only outline it here. There are two different measurements of this signal, (assumed to be independent) and we designate these measurements as $d_1(t_i)$ and $d_2(t_i)$. The model functions will be abbreviated as $f_1(t)$ and $f_2(t)$ with the understanding that each measurement of the signal has different amplitudes, and noise variance, but the same $\{\omega\}$ parameters.

We can write the likelihood (15) immediately to obtain

$$L(f_1, f_2) \propto (\sigma_1\sigma_2)^{-N} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}[d_1(t_i) - f_1(t_i)]^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{N}[d_2(t_i) - f_2(t_i)]^2\right\}$$

Because the amplitudes and noise variance are assumed different in each channel we may remove these using the same procedure developed in Section III. After removing the nuisance parameters the marginal posterior probability of the $\{omega\}$ parameters is just the product of the "student t-distributions" (28) for each channel separately:

$$P(\omega|DI) \propto \left[1 - \frac{m\overline{h^2}_1}{N\overline{d^2}_1}\right]^{\frac{m-N}{2}} \left[1 - \frac{m\overline{h^2}_2}{N\overline{d^2}_2}\right]^{\frac{m-N}{2}} \tag{66}$$

where the subscripts refer to the channel number. As explained previously, (66) in effect estimates the noise level independently in the two channels.

A procedure for dealing with the multiple frequency problem was outlined in Section V, and we will apply that procedure here. The first step in any harmonic analysis is to plot the data and the log of the probability of a single harmonic frequency. If there is only one channel present, this is essentially the periodogram of the data, Fig. 5(B) and Fig. 5(D). When more than one channel is present the log probability of a single harmonic frequency is essentially the sum of the periodograms for each channel weighted by the appropriate variances. If the variances are unknown, then the appropriate statistic is the log of (66), Fig. 6.

Now as was noted in Section V, if the frequencies are well separated, a peak in the periodogram above the noise level is evidence of a frequency near that peak. From examining

Figure 5: Analyzing NMR Data

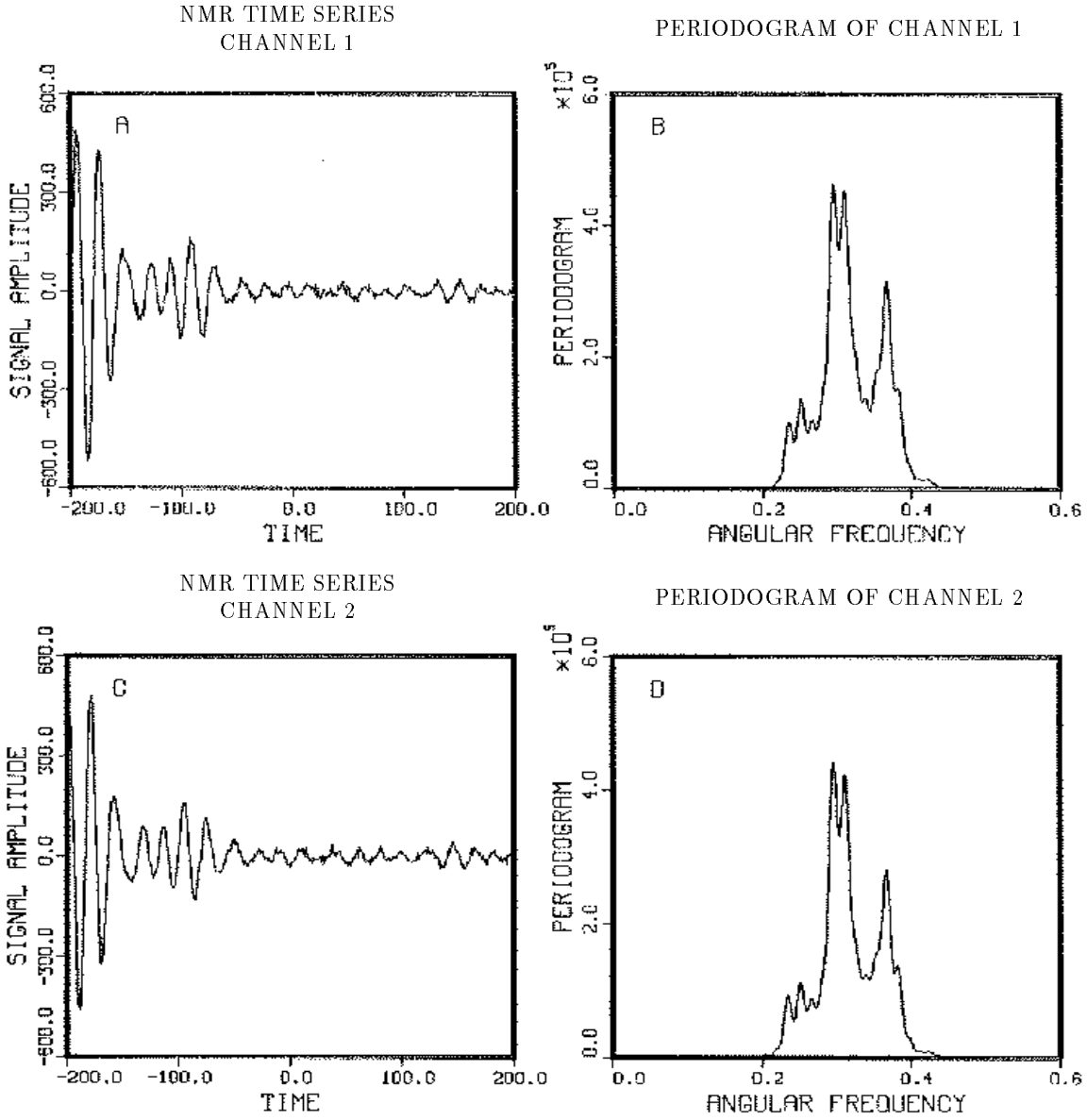


Fig. 5(A) and 5(C) are channel one and two of an NMR experiment. The data are a free-induction decay for a sample containing a mixture of D_2 and HD in a liquid phase. Theory indicates there should be three frequencies in these data: A D_2 singlet, and an HD doublet with a 43Hz separation. The singlet should be approximately in the center of the doublet. There were 2048 data points in each channel, the data were sampled at 0.0005 seconds. In the discrete Fourier transform, Fig. 5(B) and 5(D), the singlet appears to be split into a doublet. This is caused by a non-Lorentzian decay. The envelope for the decay actually goes negative producing the double peak.

We can now work backward to obtain the standard deviation for the ω_j as

$$\langle \omega_j^2 \rangle - \langle \omega_j \rangle^2 = \sum_{k=1}^{r} \frac{u_{kj}^2}{2v_k} \equiv \sigma_j^2.$$

Then the estimated ω_j parameters are

$$(\omega_j)_{est} = \hat{w}_j \pm \sigma_j \tag{65}$$

where $\hat{w}_j$ is the location of the maximum of the probability distribution as a function of the parameter ω_j.

For an arbitrary model the matrix H_{jk} cannot be calculated analytically; however, it can be evaluated numerically using the computer code given in Appendix A. We use a general searching routine to find the maximum of the probability distribution and then calculate this matrix numerically. The log of the "student t-distribution" is so sharply peaked that gradient searching routines do not work well. We use a "pattern" search routine described by Hooke and Jeeves [12].

VI. EXAMPLES: APPLICATIONS TO REAL DATA.

The this section is devoted to applications. We will apply these procedures to a number of time series including NMR signals, economic time series, and Wolf's relative sunspot numbers. We do this in a effort to show how these procedures can be used to obtain optimal parameter estimates and to show the power and versatility of these methods.

A. NMR time series.

NMR is one of the best examples of how the introduction of modern computing machines has revolutionized a branch of science. With the aid of computers more data can be taken and summarized into a useful form faster than has ever been possible before. The standard way to analyze an NMR experiment is to obtain a quadrature data set, with two separate measurements, 90° out of phase with each other, and to do a complex Fourier transform on the data [13].

The global phase of the discrete complex transform is adjusted until the real part (called an absorption spectrum) is as symmetric as possible. The frequencies and decay rates are then estimated from the absorption spectrum. There are, of course, good physical reasons why the absorption spectrum of the "true signal" is important to physicists. However, as we have emphasized repeatedly since Section II, the discrete Fourier transform is an optimal frequency estimator only when a single simple harmonic frequency is present.

We will apply the procedures developed in the previous sections to a time series from a real NMR experiment, and contrast our analysis to the one done using the absorption spectrum. The NMR data used are of a free-induction decay, [14], Fig. 5. The sample contained a mixture of 63% liquid Hydrogen-Deuterium (HD) and Deuterium (D_2) at 20.2°K. The sample was excited with a 55MHz pulse, and then detected using a standard mixer-modulation technique. The resulting signal is in the audio range where it has several oscillations at about 100Hz. The data were sampled at $\Delta T = 0.0005$ seconds, and $N = 2048$ samples were taken for each channel. The data therefore, span a time interval of about one

from the preceding steps, as initial guesses in the general problem. Locate the maximum of the multiple frequency posterior probability distribution. This will improve the estimates by removing the interference effects between them. Then determine the accuracy of the estimates.

To obtain the accuracy estimates we must compute both $\langle \omega_j \rangle$ and $\langle \omega_j^2 \rangle$ where ω_j is one member of the set of $\{\omega\}$ parameters. It might be a frequency, decay, chirp or any other parameter in the set. Then the estimates are given by

$$(\omega_j)_{est} = \hat{w}_j \pm \sigma_j$$

where

$$\sigma_j = \sqrt{\omega_j^2 - \omega_j^2}$$

and $\hat{w}_j$ is the location of the maximum of the "student t-distribution" for parameter ω_j. But if the number of parameters in this set is large there is virtually no hope of performing the integrals represented by $\langle \omega_j \rangle$ and $\langle \omega_j^2 \rangle$, either numerically or analytically. We will be forced to use an approximate result for σ_j.

We can approximate (27) by a Gaussian if we replace σ^2 in (27) by its expectation value (41). We can then Taylor expand to obtain a suitable Gaussian approximation. Define the matrix

$$H_{jk} \equiv -\frac{m}{4} \frac{\partial^2}{\partial \omega_j \partial \omega_k} \frac{\overline{h^2}}{\langle \sigma^2 \rangle}$$

then $P(\{\omega\}|D, \sigma, I)$ (27) is approximately given by

$$P(\{\omega\}|D, \sigma, I) = \frac{1}{z} \exp\left(\sum_{jk=1}^{r} H_{jk} \Delta_j \Delta_k \right)$$

where

$$\Delta_j \equiv \omega_j - \hat{w}_j$$

are just the Taylor expansion variables and z is a normalization constant.

We can transform the variables to an orthogonal set and then perform the r integrals just as we did with the amplitudes in Section III. These new variables are obtained from the eigenvalues and eigenvectors of H_{jk}. Let u_{jk} denote the k'th component of the j'th eigenvector of H_{jk} and let v_j be the eigenvalue. Then the orthogonal variables are given by

$$s_j = \sum_{k=1}^{r} \Delta_k u_{kj} \qquad \Delta_j = \sum_{k=1}^{r} s_k u_{jk}.$$

Making this change of variables we have

$$P(s|D, \sigma, I) \approx \prod_{k=1}^{r} \left(\frac{v_k}{\pi} \right)^{\frac{1}{2}} \exp\left(-\sum_{j=1}^{r} s_j^2 \right). \tag{64}$$

From (64) we can compute $\langle s_j \rangle$ and $\langle s_j^2 \rangle$. The Jacobian is just the determinant $|u_{jk}|$, but this is one since the transformation matrix u_{jk} is orthogonal. Of course $\langle s_j \rangle$ is zero and the expectation value $\rangle s_j s_k \rangle$ is given by

$$\langle s_j s_k \rangle = \frac{\delta_{jk}}{2 v_k}$$

where e_i has variance one and the index runs over the symmetric time interval ($-255.5 \leq i \leq 255.5$) by unit steps. This time series Fig. 4(A) has two simple harmonic frequencies. The two frequencies are separated by approximately one step in the discrete Fourier transform, $|\omega_1 - \omega_2| \approx 2\pi/512$.

From looking at the raw time series one might just barely guess that there is more going on than a simple harmonic frequency plus noise, because the oscillation amplitude seems to vary slightly. If we were to guess that there are two close frequencies, then by examining the data one can guess that the difference between these two frequencies is not more that one cycle over the entire time interval. If the frequencies were separated by more than this we would expect to see beats in the data. If there are two frequencies, the second frequency must be within 0.012 of the first in dimensionless units. This is in the region were the frequency estimates are almost but not quite confounded.

Now Fig. 4(B) the periodogram (continuous curve) and the discrete Fourier transform (open circles) show only a single peak. The single frequency model has estimated a frequency which is essentially the average of the two. The two frequency posterior probability density Fig. 4(C) show two well resolved, symmetrical maxima. Thus the inclusion of just this one simple additional fact has greatly enhanced our ability to detect the two signals.

From the contours in Fig. 4(C) it appears that the upper frequency is being determined more accurately than the lower one. However, to be sure of this we should integrate out one frequency to see the marginal posterior distribution of the other.

Now that we know the data contain two partially resolved frequencies, we could proceed to obtain data over a longer time span and resolve the frequencies. Regardless, it is now clear that what one can detect clearly depends on what question one asks.

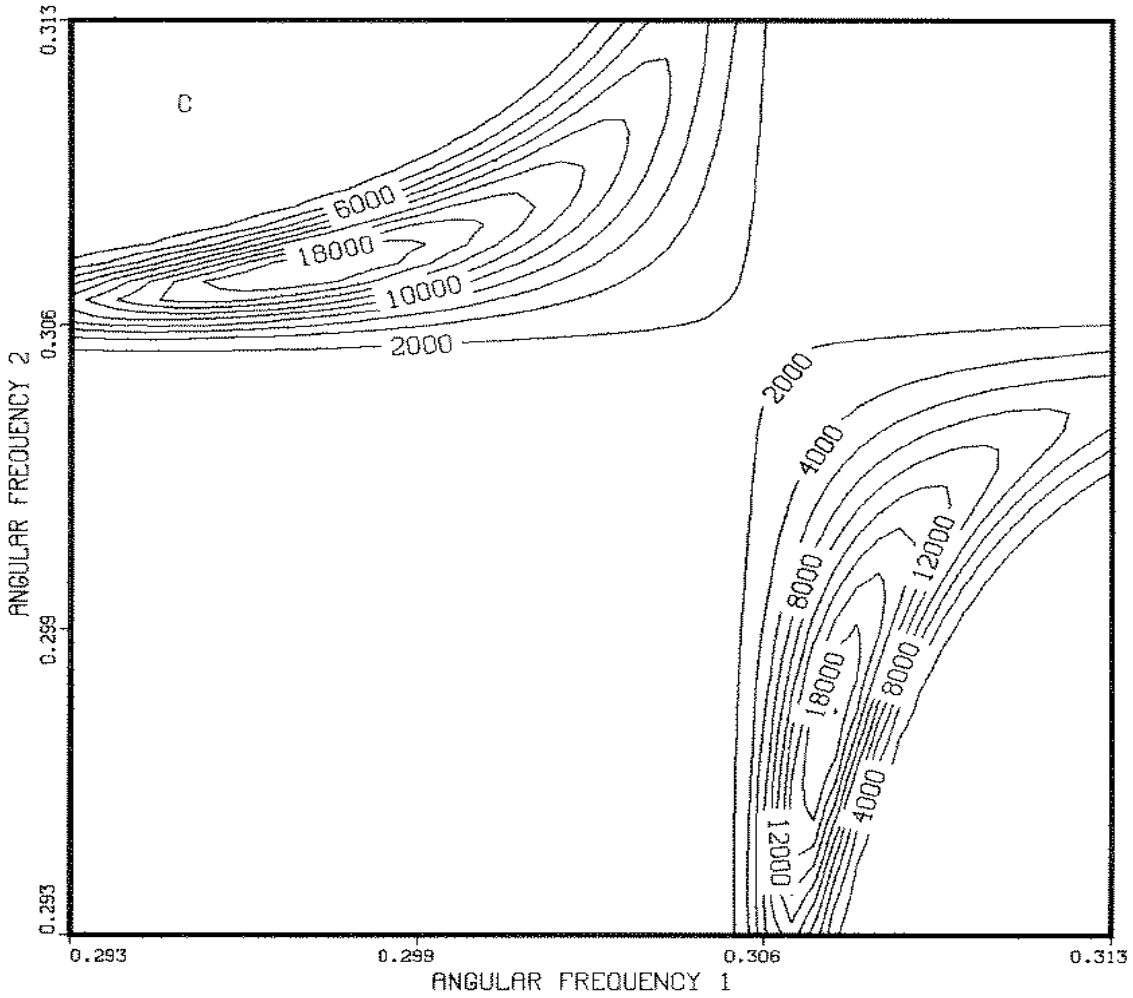
D. Multiple nonstationary frequencies estimation.

The general solution to this problem is given by the "student t-distribution." When the frequencies are harmonic (i.e. no decay or chirp) and well separated ($|\omega_j - \omega_k| > 2\pi/N$) the problem separates into multiple single frequency problems. When more than one frequency is close ($|\omega_j - \omega_k| \approx 2\pi/N$) we must use a more general model (around the close frequencies).

To understand this problem completely, one must look at the case when there are two nonstationary frequencies present. We already know the answer to this question: the Fourier transform would not work for estimating multiple nonstationary frequencies if the estimation problem did not separate. The details for this problem are just a straightforward generalization to the two frequency problem and we have not included them here.

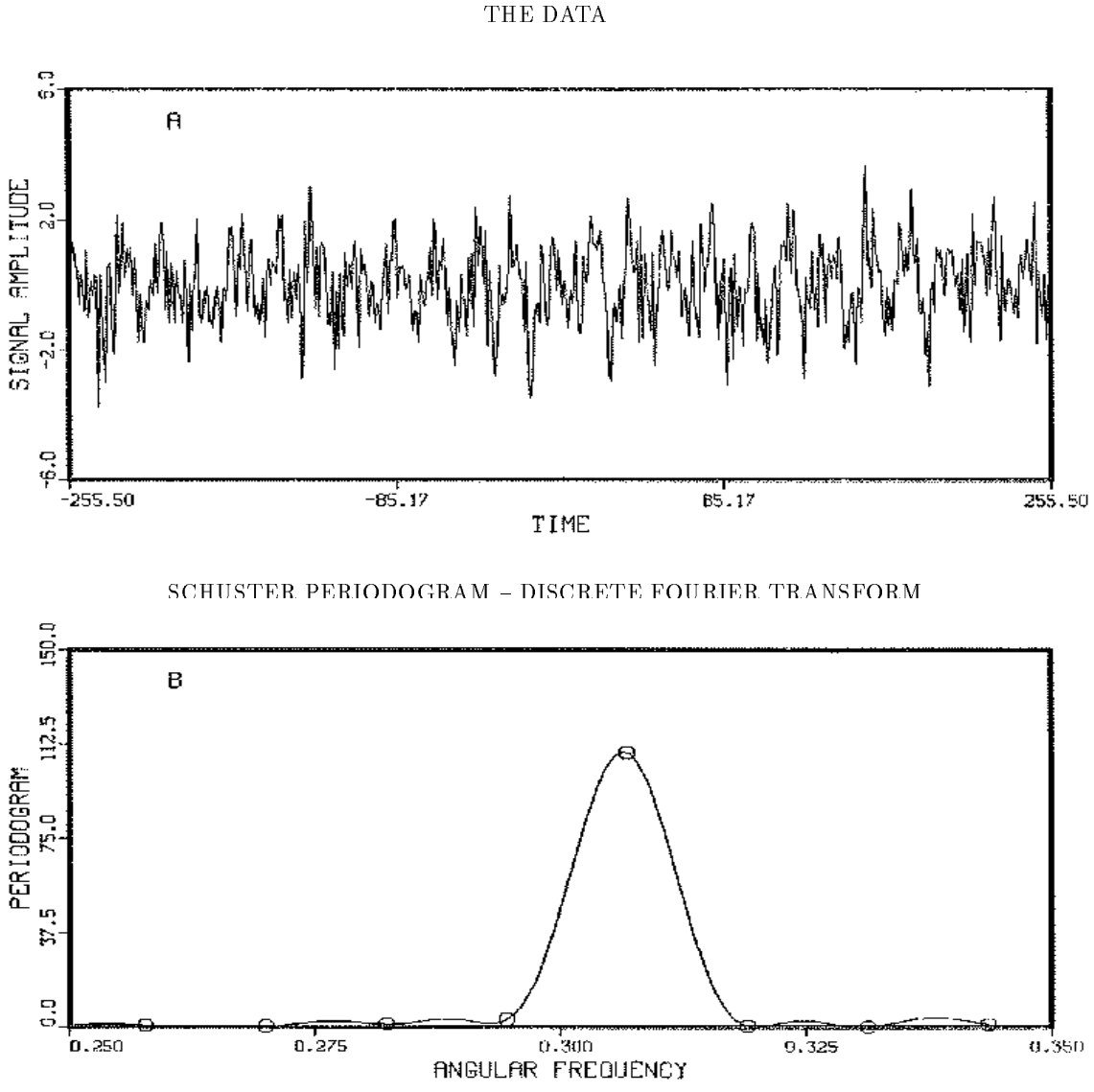
From what we have learned and from what was shown by Jaynes [2] a procedure for estimating multiple frequencies can now be given. First, compute the log of the posterior probability using a model with a single harmonic frequency plus a constant and look for peaks. If there are r well resolved peaks above the noise level, it is a good bet there are at least r frequencies present. Second, use a single frequency model (with decay if necessary) and locate the maximum of each peak in the Fourier transform. Third, use a two frequency model to examine each peak. The initial frequency estimates should be slightly above and below (about 1/4 step in the discrete Fourier transform) the location of the peak. If the peak is a single frequency, the two frequency model will confound it, because it cannot fit the data any better than does a one frequency model. If the peak is due to two resolvable frequencies it will find the second frequency. Remove any confounded parameters. Fourth, if one desires to know the "best" estimates of the parameters, then use the estimated values,

Figure 4: (Continued)



The two-frequency probability density, Fig. 4(C), clearly indicates the presence of two frequencies. The posterior odds ratio prefers the two-frequency model by 10^7 to 1.

Figure 4: Two Harmonic Frequencies – The Data



The data, Fig. 4(A), contain two frequencies. They are separated from each other by approximately a single step in the discrete Fourier transform. The periodogram, Fig. 4(B), shows only a single peak located between the two frequencies.

and when one signal is much larger than the other. When we prepared this table we used equation (27), not the "student t-distribution." In order to obtain the "best" theoretical estimates we did not include any noise in the data [just as noise was not included in (60)]. Had we used the "student t-distribution" the accuracy estimates would have been much much better. The estimates obtained are the "best" in the sense that in a real data set with $\sigma = 1$, containing $N = 1000$ data points the accuracy estimates one obtains will be, nearly always, slightly worse than those contained in Table 1.

Table 1			
Amplitudes/Description	$(\hat{f}_2 - \hat{f}_1) = 0.07$ Hz	$(\hat{f}_2 - \hat{f}_1) = 0.3$ Hz	$(\hat{f}_2 - \hat{f}_1) = 5.1$ Hz
The square magnitude is equal to signal 1 of signal 2	$(f_1)_{est} = \hat{f}_1 \pm 0.091$ $(f_2)_{est} = \hat{f}_2 \pm 0.091$	$(f_1)_{est} = \hat{f}_1 \pm 0.027$ $(f_2)_{est} = \hat{f}_2 \pm 0.027$	$(f_1)_{est} = \hat{f}_1 \pm 0.025$ $(f_2)_{est} = \hat{f}_2 \pm 0.025$
The square magnitude of signal 2 is four times larger than signal 1	$(f_1)_{est} = \hat{f}_1 \pm 0.091$ $(f_2)_{est} = \hat{f}_2 \pm 0.088$	$(f_1)_{est} = \hat{f}_1 \pm 0.027$ $(f_2)_{est} = \hat{f}_2 \pm 0.013$	$(f_1)_{est} = \hat{f}_1 \pm 0.025$ $(f_2)_{est} = \hat{f}_2 \pm 0.012$
The square magnitude of signal 2 is 128 times larger than signal 1	$(f_1)_{est} = \hat{f}_1 \pm 0.091$ $(f_2)_{est} = \hat{f}_2 \pm 0.034$	$(f_1)_{est} = \hat{f}_1 \pm 0.025$ $(f_2)_{est} = \hat{f}_2 \pm 0.002$	$(f_1)_{est} = \hat{f}_1 \pm 0.025$ $(f_2)_{est} = \hat{f}_2 \pm 0.002$

The three values of $(\omega_1 - \omega_2)$ examined correspond to $\delta = 1/4$, $\delta = 4$, and $\delta = 16$: roughly these correspond to frequency separations of 1/12, 1, and 5 steps in the discrete Fourier transform. We held the squared magnitude of signal one constant at one and the second is either 1, 4 or 128 times larger.

When the separation frequency is 0.07 Hz the frequencies are indistinguishable. The smaller component cannot be estimated accurately. As the magnitude of the second signal increases the estimated accuracy of the second signal becomes better as one's intuition would suppose it should (the signal looks more and more like one frequency).

When the separation frequency is 0.3 Hz or about one step in the discrete Fourier transform, the accuracy estimates indicate that the two frequencies are well resolved. By this we mean one of the frequencies would have to be moved by eleven standard deviations before it would be confounded with the other (two parameters are said to be confounded when probability theory estimates their values to be the same). This is true for all the amplitudes in the table; it does however, improve with increasing amplitude. According to probability theory, when two frequencies are as close together as one step in the discrete Fourier transform those frequencies are clearly resolvable.

When the separation frequency is 5.1Hz, the accuracy estimates clearly determine both frequencies. Additionally, the accuracy estimates for the smaller frequency are essentially 0.025 Hz which is the same as the estimate for a single harmonic frequency that we found previously (10). Examining Table 1 more carefully, we see that when the frequencies are separated by even a single step in the discrete Fourier transform, the accuracy estimates are essentially those for the single harmonic frequencies. The ability to estimate two close frequencies accurately is essentially independent of the separation frequency, as long as it is greater than or approximately equal to one step in the discrete Fourier transform.

To illustrate the two frequency probability density (59) we prepared a simple example, Fig. 4. This example was prepared from the following equation

$$d_i = \cos(0.3i + 1) + \cos(0.307i + 2) + e_i$$

To get these partial derivatives we Taylor expand (63) around the maximum located at $\hat{w}_1$ and $\hat{w}_2$ and then take the derivatives. The intermediate steps are of little concern and were carried out using an algebra manipulation package. Terms of order one compared to N were again ignored and, we have assumed the frequencies are close but distinct: we used the small angle approximations for the sine and cosine at the end of the calculation. The local variable δ [defined as $(\hat{w}_2 - \hat{w}_1)/2 \equiv \delta/N$] measures the distance between two adjacent frequencies. If δ is π then the frequencies are separated by one step apart in the discrete Fourier transform. The second partial derivatives of $\overline{h^2}$ are given by:

$$\left.\frac{\partial^2\overline{h^2}}{\partial\omega_1^2}\right|_{\substack{\omega_1=\hat{w}_1\\ \omega_2=\hat{w}_2}} \approx -(\hat{b}_1^2+\hat{b}_3^2)N^3\left(\frac{3\sin^2(\delta)-6\delta\cos(\delta)\sin(\delta)+\delta^2[\sin^2(\delta)+3\cos(\delta)]-\delta^4}{48\delta^3[\sin(\delta)-\delta][\sin(\delta)+\delta]}\right)$$

$$\left.\frac{\partial^2\overline{h^2}}{\partial\omega_2^2}\right|_{\substack{\omega_1=\hat{w}_1\\ \omega_2=\hat{w}_2}} \approx -(\hat{b}_2^2+\hat{b}_4^2)N^3\left(\frac{3\sin^2(\delta)-6\delta\cos(\delta)\sin(\delta)+\delta^2[\sin^2(\delta)+3\cos(\delta)]-\delta^4}{48\delta^3[\sin(\delta)-\delta][\sin(\delta)+\delta]}\right)$$

$$\left.\frac{\partial^2\overline{h^2}}{\partial\omega_1\partial 2_2}\right|_{\substack{\omega_1=\hat{w}_1\\ \omega_2=\hat{w}_2}} \approx (\hat{b}_1\hat{b}_2+\hat{b}_3\hat{b}_4)N^3\left(\frac{\delta^4\sin(\delta)+2\delta^3\cos(\delta)-3\delta^2\sin(\delta)+\sin^3(\delta)}{16\delta^3[\sin(\delta)-\delta][\sin(\delta)+\delta]}\right).$$

If the true frequencies $\hat{w}_1$ and $\hat{w}_2$ are separated by two steps in the discrete Fourier transform, $\delta = 2\pi$, we may reasonably ignore all but the δ^4 term to obtain

$$\left.\frac{\partial^2\overline{h^2}}{\partial\omega_1^2}\right|_{\substack{\omega_1=\hat{w}_1\\ \omega_2=\hat{w}_2}} \approx -\frac{(\hat{b}_1^2+\hat{b}_3^2)N^3}{48}$$

$$\left.\frac{\partial^2\overline{h^2}}{\partial\omega_2^2}\right|_{\substack{\omega_1=\hat{w}_1\\ \omega_2=\hat{w}_2}} \approx -\frac{(\hat{b}_2^2+\hat{b}_4^2)N^3}{48}$$

$$\left.\frac{\partial^2\overline{h^2}}{\partial\omega_1\partial\omega_2}\right|_{\substack{\omega_1=\hat{w}_1\\ \omega_2=what_2}} \approx -\frac{(\hat{b}_1\hat{b}_2+\hat{b}_3\hat{b}_4)N^3\sin(\delta)}{16\delta}.$$

The accuracy estimates reduce to equation (60) when the frequencies are will separated. When the frequencies have approximately the same amplitudes and δ is order of 2π (the frequencies are separated by two steps in the discrete Fourier transform) the interaction term is down by four and one expects the estimates to be nearly the same as those for a single frequency. Probability theory indicates that two frequencies which are as close together as two steps in a discrete Fourier transform do not interfere with each other in any significant way.

To better understand the maximum theoretical accuracy with which two frequencies can be estimated we have prepared Table 1. To make these estimates comparable to those obtained in Section II we have again assumed there $N = 1000$ data points and $\sigma = 1$. There are three regions of interest: when the frequency separation is small compared to a single step in the discrete Fourier transform; when the separation is of order one step; and when the separation is large. Additionally we would like to understand the behavior when the signals are of the same amplitude, when one signal is slightly larger than the other,

where

$$\begin{aligned}
\frac{1}{r^2} &= -\frac{\partial^2 \overline{h^2}}{\partial \omega_1^2}\bigg|_{\substack{\omega_1 = \hat{w}_1 \\ \omega_2 = \hat{w}_2}} \\
\frac{1}{s^2} &= -\frac{\partial^2 \overline{h^2}}{\partial \omega_2^2}\bigg|_{\substack{\omega_1 = \hat{w}_1 \\ \omega_2 = \hat{w}_2}} \\
\frac{1}{u^2} &= -\frac{\partial^2 \overline{h^2}}{\partial \omega_1 \partial \omega_2}\bigg|_{\substack{\omega_1 = \hat{w}_1 \\ \omega_2 = \hat{w}_2}}
\end{aligned}$$

where $\hat{w}_1$, $\hat{w}_2$ are the locations of the maxima of (59). If we have a uniformly sampled signal of the form

$$f_l = \hat{b}_1 \cos(\hat{w}_1 l) + \hat{b}_2 \cos(\hat{w}_2 l) + \hat{b}_3 \sin(\hat{w}_1 l) + \hat{b}_4 \sin(\hat{w}_2 l) + e_l \tag{61}$$

where $-T \leq l \leq T$, $2T+1 = N$, and $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$, $\hat{b}_4$ are the true amplitudes, $\hat{w}_1$, $\hat{w}_2$ are the true frequencies, and $e_t \ll$ the signal, then h_1 is given by the projection of H_1 (58) onto the data (61) to obtain

$$h_1 = \frac{1}{\sqrt{N + B(\omega_1, \omega_2)}} \sum_{l=-T}^{T} \cos(\omega_1 l) + \cos(\omega_2 l) f_l$$

where

$$\frac{B(\omega_1, \omega_2)}{2} \equiv \frac{1}{2} \sum_{l=-T}^{T} \cos(\omega_1 - \omega_2) l = \frac{1}{2} \frac{\sin \frac{1}{2} N(\omega_1 - \omega_2)}{\sin \frac{1}{2}(\omega_1 - \omega_2)} \tag{62}$$

For uniform time series these h_j may be summed explicitly using equation (62) to obtain

$$h_1 = \frac{1}{2\sqrt{N + B(\omega_1, \omega_2)}} \left\{ \hat{b}_1 [B(\hat{w}_1, \omega_1) + B(\hat{w}_1, \omega_2)] + \hat{b}_2 [B(\hat{w}_2, \omega_1) + B(\hat{w}_2, \omega_2)] \right\}$$

$$h_2 = \frac{1}{2\sqrt{N - B(\omega_1, \omega_2)}} \left\{ \hat{b}_1 [B(\hat{w}_1, \omega_1) - B(\hat{w}_1, \omega_2)] + \hat{b}_2 [B(\hat{w}_2, \omega_1) - B(\hat{w}_2, \omega_2)] \right\}$$

$$h_3 = \frac{1}{2\sqrt{N + B(\omega_1, \omega_2)}} \left\{ \hat{b}_3 [B(\hat{w}_1, \omega_1) + B(\hat{w}_1, \omega_2)] + \hat{b}_4 [B(\hat{w}_2, \omega_1) + B(\hat{w}_2, \omega_2)] \right\}$$

$$h_4 = \frac{1}{2\sqrt{N - B(\omega_1, \omega_2)}} \left\{ \hat{b}_3 [B(\hat{w}_1, \omega_1) - B(\hat{w}_1, \omega_2)] + \hat{b}_4 [B(\hat{w}_2, \omega_1) - B(\hat{w}_2, \omega_2)] \right\}.$$

We have kept terms corresponding to the differences in the frequencies. When the frequencies are close together it is only these terms which are important: the approximation is consistent with the others made.

The sufficient statistic $\overline{h^2}$ is then given by

$$\overline{h^2} = \frac{1}{4}(h_1^2 + h_2^2 + h_3^2 + h_4^2). \tag{63}$$

To obtain a Gaussian approximation for (59) one must calculate the second derivative of (63) with respect to ω_1 and ω_2. The problem is simple in principle but difficult in practice.

$$H_2(t) = \frac{1}{\sqrt{N-B}}\left\{\cos(\omega_1 t) - \cos(\omega_2 t)\right\},$$

$$H_3(t) = \frac{1}{\sqrt{N+B}}\left\{\sin(\omega_1 t) + \sin(\omega_2 t)\right\},$$

$$H_4(t) = \frac{1}{\sqrt{N-B}}\left\{\sin(\omega_1 t) - \sin(\omega_2 t)\right\}.$$

We can now write the sufficient statistic $\overline{h^2}$ in terms of these orthonormal model functions to obtain

$$\overline{h^2} = h_+^2 + h_-^2,$$

$$h_+^2 \equiv \frac{1}{4(N+B)}\left\{[P(\omega_1)+P(\omega_2)]^2 + [Q(\omega_1)+Q(\omega_2)]^2\right\},$$

$$h_-^2 \equiv \frac{1}{4(N-B)}\left\{[P(\omega_1)-P(\omega_2)]^2 + [Q(\omega_1)-Q(\omega_2)]^2\right\},$$

where P and Q are the sine and cosine transforms of the data as functions of the appropriate frequency. The factor of 4 comes about because for this problem there are $m = 4$ model functions. Using (26), the posterior probability that two distinct frequencies are present, given the noise variance σ^2, is

$$P(\omega_1, \omega_2 | D, I, \sigma) \propto \exp\left[\frac{2\overline{h^2}}{\sigma^2}\right]. \tag{59}$$

A quick check on the asymptotic forms of this will verify that when the frequencies are well separated one has $\overline{h^2} = \frac{1}{2}[C(\omega_1) + C(\omega_2)]$, and it has reduced to (55) and, when the frequencies are the same the second term goes smoothly to zero, and the first term goes into $\frac{1}{2}C(\omega)$, to reduce to (56) as expected.

When the frequencies are very close or far apart we can apply the results obtained by Jaynes

Jaynes chirp

concerning the accuracy of the frequency estimates:

$$(\omega_{est}) = (\omega_{max}) \pm \frac{\sigma}{A}\sqrt{48/N^3}. \tag{60}$$

In the region where the frequencies are close but distinct, (59) appears very different. We would like to understand what is happening in this region, in particular we would like to know just how well two close frequencies can be estimated. To understand this we will construct a Gaussian approximation similar to what was done for the case with Lorentzian decay. We Taylor expand the $\overline{h^2}$ in (59) to obtain

$$P(\omega_1, \omega_2 | D, I, \sigma) \approx \exp\left(-\frac{(\omega_1 - \hat{w}_1)^2}{2r^2\sigma^2} - \frac{(\omega_2 - \hat{w}_2)^2}{2s^2\sigma^2} - \frac{(\omega_1 - \hat{w}_1)(\omega_2 - \hat{w}_2)}{2u^2\sigma^2}\right)$$

The matrix g_{jk} has two redundant eigenvalues, and the probability distribution becomes

$$P(\omega|D,I,\sigma) \propto \exp\left[\frac{C(\omega)}{\sigma^2}\right]. \tag{56}$$

The probability density goes smoothly into the single frequency probability distribution along this axis of symmetry. Given that the two frequencies are equal, our estimate of them will be identical, in value and accuracy, to those of the one frequency case.

The problem of understanding the posterior probability density when there are two close but distinct frequencies must now be addressed. The matrix g_{jk} for this two frequency problem is readily diagonalized and the exact solution for the two frequency problem obtained. An approximate solution may be obtained that is valid in the same sense that the approximate solution to the single frequency problem obtained in Section II is valid. To obtain this approximate solution one needs only to examine the matrix g_{jk} and notice that the elements of this matrix consist of the diagonal elements given by:

$$C_{11} = \frac{N}{2} + \frac{\sin(N\omega_1)}{2\sin(\omega_1)} \approx \frac{N}{2},$$

$$C_{22} = \frac{N}{2} + \frac{\sin(N\omega_2)}{2\sin(\omega_2)} \approx \frac{N}{2},$$

$$S_{11} = \frac{N}{2} - \frac{\sin(N\omega_1)}{2\sin(\omega_1)} \approx \frac{N}{2},$$

$$S_{22} = \frac{N}{2} - \frac{\sin(N\omega_2)}{2\sin(\omega_2)} \approx \frac{N}{2},$$

and the off diagonal elements. The off diagonal terms are small compared to N unless the frequencies are specifically in the region of $\omega_1 \approx \omega_2$, then only the terms involving the difference $(\omega_1 - \omega_2)$ are large. We can approximate the off diagonal terms as:

$$C_{12} \approx S_{12} \approx \frac{1}{2}\sum_{l=-T}^{T} \cos\frac{1}{2}(\omega_1-\omega_2)l = \frac{1}{2}\frac{\sin\frac{1}{2}N(\omega_1-\omega_2)}{\sin\frac{1}{2}(\omega_1-\omega_2)} \equiv \frac{B}{2}. \tag{57}$$

When the two frequencies are well separated, (57) is of order one and is ignorable. When the two frequencies are nearly equal, then the off diagonal terms are large and are given accurately by (57). So the approximation is valid for all values of ω_1 and ω_2.

With this approximation for g_{jk} it is now possible to write a simplified solution for the two frequency problem. The matrix g_{jk} (16) is given approximately by

$$g_{jk} = \frac{1}{2}\begin{Bmatrix} N & B & 0 & 0 \\ B & N & 0 & 0 \\ 0 & 0 & N & B \\ 0 & 0 & B & N \end{Bmatrix}.$$

The orthonormal model functions (17) may now be constructed:

$$H_1(t) = \frac{1}{\sqrt{N+B}}\left\{\cos(\omega_1 t) + \cos(\omega_2 t)\right\}, \tag{58}$$

The model functions can then be used to construct the g_{jk} matrix. On a uniform grid this is given by

$$g_{jk} = \begin{bmatrix} C_{11} & C_{12} & 0 & 0 \\ C_{12} & C_{22} & 0 & 0 \\ 0 & 0 & S_{11} & S_{12} \\ 0 & 0 & S_{12} & S_{22} \end{bmatrix}$$

where

$$C_{jk} = \sum_{l=-T}^{T} \cos(\omega_j l)\cos(\omega_k l) = \frac{\sin(\frac{1}{2}N\omega_+)}{2\sin(\frac{1}{2}\omega_+)} + \frac{\sin(\frac{1}{2}N\omega_-)}{2\sin(\frac{1}{2}\omega_-)}$$

$$S_{jk} = \sum_{l=-T}^{T} \sin(\omega_j l)\sin(\omega_k l) = \frac{\sin(\frac{1}{2}N\omega_-)}{2\sin(\frac{1}{2}\omega_-)} - \frac{\sin(\frac{1}{2}N\omega_+)}{2\sin(\frac{1}{2}\omega_+)}$$

$$w_+ = \omega_j + \omega_k, \qquad (j,k = 1 \text{ or } 2)$$

$$w_- = \omega_j - \omega_k.$$

The eigenvalue and eigenvectors problem for g_{jk} splits into two separate problems each involving 2×2 matrices. The eigenvalues are

$$\lambda_{1\&2} = \frac{C_{11} + C_{22}}{2} \pm \sqrt{(C_{11} - C_{22})^2 + 4C_{12}^2}$$

$$\lambda_{3\&4} = \frac{S_{11} + S_{22}}{2} \pm \sqrt{(S_{11} - S_{22})^2 + 4S_{12}^2}.$$

We can go on and obtain the exact solution to this problem but that will not be necessary. When the frequencies are well separated $|\omega_1 - \omega_2| \gg 2\pi/N$, the eigenvalues reduce to $\lambda = N/2$. That is, g_{jk} goes into $N/2$ times the unit matrix. Then each of the model equations are effectively orthogonal and the sufficient statistic $\overline{h^2}$ reduces to

$$\overline{h^2} = \frac{2}{N}\left[C(\omega_1) + C(\omega_2)\right];$$

and the probability, when the variance is known, is given by

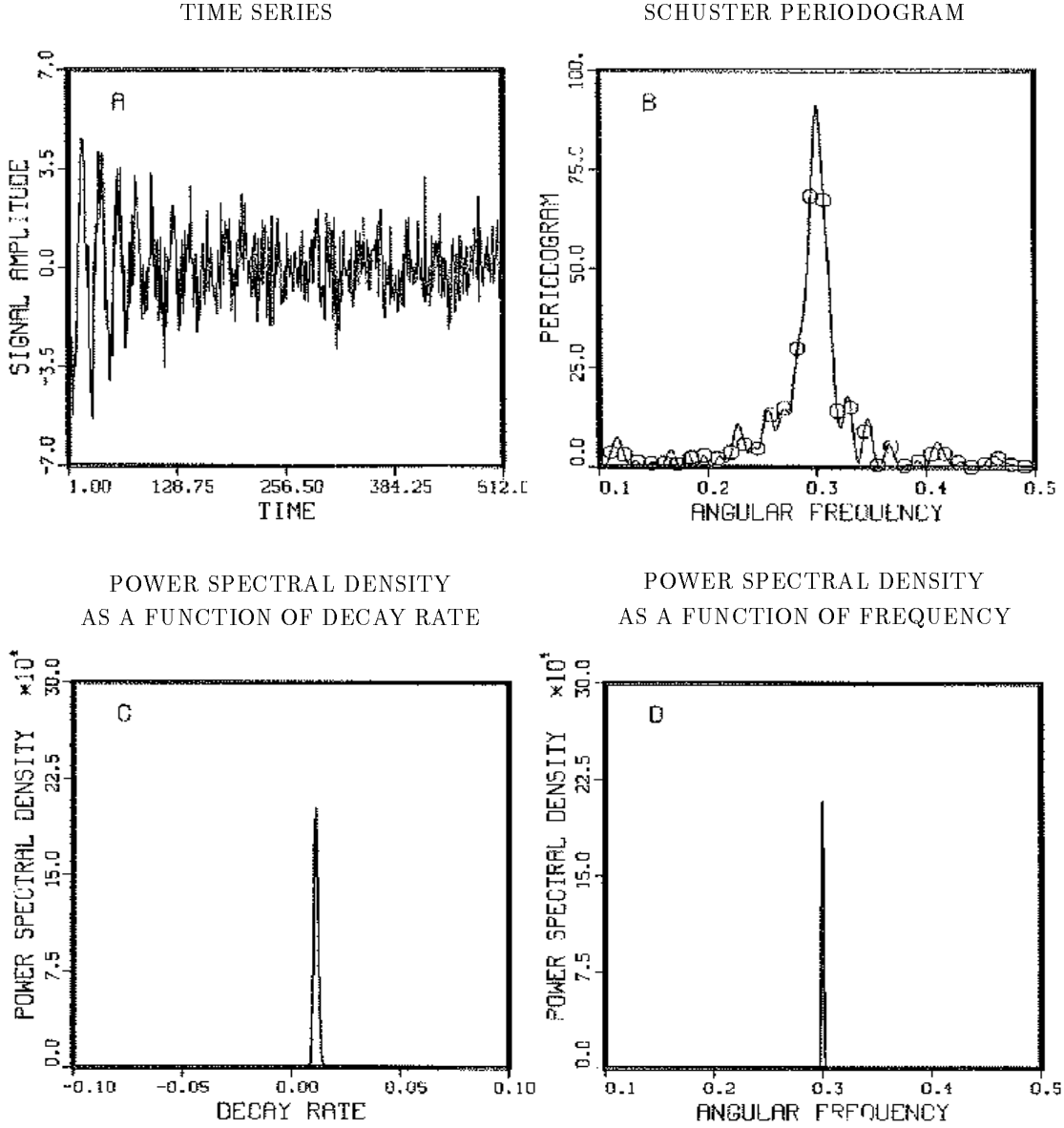
$$P(\omega_1\omega_2|D, I, \sigma) \propto \exp\left[\frac{C(\omega_1) + C(\omega_2)}{\sigma^2}\right]. \tag{55}$$

The problem has separated: one can estimate each of the frequencies separately. The maximum of the two frequency posterior probability density will be located at the two greatest peaks in the periodogram, in agreement with the common sense usage of the Fourier transform. A similar result holds for the general frequency estimation problem. Then the r frequencies, corresponding to the maximum of the joint posterior probability, are essentially the estimates obtained from the r biggest peaks in the periodogram [2].

The labels ω_1, ω_2, etc. for the frequencies in the model are arbitrary, and accordingly their joint probability density is invariant under permutations. That means, for the two frequency problem, there is an axis of symmetry running along the line $\omega_1 = \omega_2$. We do not know from (55) what is happening along that line. This is easily investigated when $\omega_1 = \omega_2 \equiv \omega$: the eigenvalues become

$$\lambda_1 = N, \quad \lambda_2 = 0, \quad \lambda_3 = N, \quad \lambda_4 = 0.$$

Figure 3: Single Frequency with Lorentzian Decay



The data Fig. 3(A) contain a simple frequency with a Lorentzian decay plus noise. In Fig. 3(B), the noise has significantly distorted the periodogram (continuous curve) and the fast Fourier transform (open circles). The power spectral density may be computed as a function of decay rate α by integrating over the frequency Fig. 3(C), or as a function of frequency ω by integrating over the decay Fig. 3(D).

(40) with $s = 1$. In addition there is an example of how to use this subroutine in Appendix B. In this power spectral estimate (53,54) (and in the computer code) we have assumed the estimated noise variance σ^2 is small compared to $\overline{h^2}$ and have ignored this term.

To illustrate some of these points we have prepared another example, Fig. 3. This time series was prepared from the following equation

$$d_j = 0.001 + \cos(0.3j + 1)e^{-0.01j} + e_j.$$

The $N = 512$ data samples were prepared in the following manner: first, we generated the data without the noise; we then computed the average of the data, and subtracted it from each data point, thus to ensure that the average value of the data was zero; we then repeated this process on the Gaussian white noise; next, we computed the average mean-square of the signal and the noise; and scaled the data by the appropriate ratio to make the signal-to-noise ratio of the data exactly one; last, we added the noise to the data. The time series clearly shows a small signal which rapidly decays away, Fig. 3(A). Figure 3(B), the periodogram (continuous curve) and the discrete Fourier transform (open circles) clearly show the Lorentzian line shape. The noise is now significantly affecting the periodogram: the periodogram is no longer an optimum frequency estimator.

Figures 3(C) and 3(D) contain plots of the power spectral density (53, 54). In Fig. 3(C) we have treated the frequency as a nuisance parameter and have integrated it out; as was emphasized earlier this is essentially the posterior probability distribution for α normalized to a power level rather than to unity. In Fig. 3(D) we have treated the decay as the nuisance parameter and have integrated it out. This gives the power spectral estimate as a function of frequency.

The width of these curves is a measure of the uncertainty in the determination of the these parameters. We have determined full-width at half maximum (numerically) for each of these and have compared these to the theoretical "best" estimates (51) and find

$$(\omega)_{est} = 0.2998 \pm 5.3 \times 10^{-4} \quad \text{and} \quad (\omega)_{\text{best}} t = 0.3000 \pm 3 \times 10^{-4},$$

$$(\alpha)_{est} = 0.0109 \pm 5.5 \times 10^{-4} \quad \text{and} \quad (\alpha)_{\text{best}} = 0.0100 \pm 8 \times 10^{-4}.$$

Converting to Hz, $5.3 \times 10^{-4}/2\pi N = 0.84$ Hz. The frequency estimate compares nicely with the "best" estimate, while our decay estimate is a little better. Given that the theoretical estimates were only approximations they are in good agreement with each other.

C. The Spectrum of Two Harmonic Frequencies.

We now turn our attention to the slightly more general problem of analyzing a data set which we postulate contains two distinct harmonic frequencies. The "student t-distribution" represented by equation (28) is, of course, the general solution to this problem. Unfortunately, that equation does not lend itself readily to understanding the probability distribution. In particular we would like to know what the behavior of these equations are in three different limits: first, when the frequencies are well separated; second, when they are close but distinct; and third, when they are so close as to be, for all practical purposes, identical. To investigate these we will solve, approximately, the two frequency problem.

The model equation for the two frequency problem is a simple generalization of the single harmonic problem:

$$f(t) = B_1 \cos(\omega_1 t) + B_2 \cos(\omega_2 t) + B_3 \sin(\omega_1 t) + B_4 \sin(\omega_2 t).$$

The first derivatives of $\overline{h^2}$ evaluated at $\omega = \hat{w}$ and $\alpha = \hat{\alpha}$ are zero, as they should be. The mixed second partial is also zero. This gives the second derivatives of $\overline{h^2}$ as

$$\left(\frac{\partial^2 \overline{h^2}}{\partial \omega^2}\right)_{\omega=\hat{w}} = -\frac{A_1^2}{4\hat{\alpha}^3} \quad \text{and} \quad \left(\frac{\partial^2 \overline{h^2}}{\partial \alpha^2}\right)_{\alpha=\hat{\alpha}} = -\frac{A_1^2}{32\hat{\alpha}^3}.$$

We can now expand $\overline{h^2}$ in a Taylor series about the maximum and normalizing the distribution gives

$$P(\omega\alpha|D,I,\sigma) \approx (4\pi\delta_\omega\delta_\alpha)^{-1} \exp\left[-\frac{(\omega-\hat{w})^2}{2\delta_\omega^2} - \frac{(\alpha-\hat{\alpha})^2}{2\delta_\alpha^2}\right]$$

$$(\alpha)_{est} = \hat{\alpha} \pm \delta_\alpha \quad \text{and} \quad (\omega)_{est} = \hat{w} \pm \delta_\omega$$

where

$$\delta_\alpha \approx \frac{5.6\sigma\hat{\alpha}^{\frac{3}{2}}}{A_1} \quad \text{and} \quad \delta_\omega \approx \frac{2\sigma\hat{\alpha}^{\frac{3}{2}}}{A_1} \tag{51}$$

The accuracy estimate δ_α for the decay parameter is almost a factor of 3 worse than the estimate δ_ω for the frequency. This result has been noted before but why it should be so was not understood. Our independent probability analysis clearly indicates that this must be the case.

How does this compare to the results obtained for the simple harmonic frequency? Converting to Hertz involves dividing these by $2\pi\Delta t$, for a signal with $N = 1000$, $\hat{\alpha} = 0.01$, $A_1/\sqrt{2}\sigma = 1$ and, including a factor of 2 to obtain the values at the full-width at half maximum we have the estimated accuracy for frequency and decay as

$$\omega = \hat{w} \pm 0.9\text{Hz} \quad \text{and} \quad \alpha = \hat{\alpha} \pm 2.5\text{Hz}.$$

This compares to 0.025Hz for a stationary signal with the same signal-to-noise ratio. This is a factor of 36 times larger and since the error varies like $N^{\frac{-3}{2}}$ we have effectively lost all but one tenth of the data. When we have reached the unitless time of $t = 100$ the signal is down by a factor of 2.7 and has all but disappeared into the noise.

We wish to plot the power spectral estimate as a function of frequency and decay. These are given by

$$\hat{p}(\omega) \approx m\frac{\int d\alpha \overline{h^2} P(\omega,\alpha|D,I)}{\int d\alpha d\omega P(\omega,\alpha|D,I)} \tag{53}$$

$$\hat{p}(\alpha) \approx m\frac{\int d\omega \overline{h^2} P(\omega,\alpha|D,I)}{\int d\alpha d\omega P(\omega,\alpha|D,I)} \tag{54}$$

where $P(\omega,\alpha|D,I)$ is taken from (28) using (45) as the model: then, $\hat{p}(\omega)$ is useful for estimating the frequency; and $\hat{p}(\alpha)$ is useful for estimating the decay rate. These integrals can be computed numerically. The computer code used to evaluate the "student t-distribution" in this paper (in fact in all of the examples in this work) is included in Appendix A. This appendix contains a general routine for evaluating the "student t-distribution" (28), the orthonormal amplitudes (36), the power spectral density (39), and the estimated variance

and the posterior probability of a frequency ω and a decay rate α is given by

$$P(\omega,\alpha|DI) \propto \left[1-\frac{P(\omega,\alpha)^2+Q(\omega,\alpha)^2}{Nc\overline{d^2}}\right]^{\frac{2-N}{2}}. \tag{49}$$

This approximation is valid provided there is plenty of data $N \gg 1$, and there is no evidence of a low frequency, there is no restriction on the range of α: if $\alpha > 0$ the signal is decaying with increasing time, if $\alpha < 0$ the signal is growing with increasing time, and if $\alpha = 0$ the signal is stationary. This equations is exactly analogous to (13) and reduces to (43) in the limit $\alpha \to 0$.

We would like to derive an estimate of the accuracy of the frequency and decay parameter estimates. To do this we can approximate the probability distribution $P(\omega,\alpha|D,I,\sigma)$ by a Gaussian. This may be done readily by assuming a form of the data, and then expanding $\overline{h^2}$ around the maximum of the probability distribution (49) as was done in Section II. From the second derivative we may obtain the desired (mean) $\pm$ (standard deviation) estimates. We take as the data

$$d(t) = A_1\cos(\hat{w}t)e^{-\hat{\alpha}t} \tag{50}$$

where $\hat{w}$ is the true frequency of oscillation and $\hat{\alpha}$ is the true decay rate. We have assumed only a cosine component to effect some simplifications in the discussion. It will be obvious at the end of the calculation that the result for an arbitrary signal phase can be obtained by replacing the amplitude A_1^2 by the squared magnitude $A^2 \equiv A_1^2 + A_2^2$.

The projection of the data (50) onto the model functions (47, 48) is:

$$h_1 = \frac{A_1}{2\sqrt{c}}\left\{\sum_{l=1}^{N}\cos(\omega-\hat{w})le^{-(\alpha+\hat{\alpha})l} + \sum_{l=1}^{N}\cos(\omega+\hat{w})le^{-(\alpha+\hat{\alpha})l}\right\}$$

and $h_2 \ll h_1$ and is ignored. The sums may be done explicitly using (46) to obtain

$$h_1 = \frac{A_1}{4\sqrt{c}}\left\{\frac{1-e^{-2Nv}}{e^{2v}-1}+\frac{1-e^{-2Nu}}{e^{2u}-1}\right\}$$

where

$$v = \frac{\alpha+\hat{\alpha}-i(\omega-\hat{w})}{2} \quad \text{and} \quad u = \frac{\alpha+\hat{\alpha}+i(\omega-\hat{w})}{2},$$

and $i = \sqrt{-1}$ in the above equations. Then the sufficient statistic $\overline{h^2}$ is given by:

$$\overline{h^2} = \frac{A_1^2}{16}\left[\frac{e^{2\alpha}-1}{1-e^{-2N\alpha}}\right]\left[\frac{1-e^{-2Nv}}{1-e^{2v}}+\frac{1-e^{-2Nu}}{1-e^{2u}}\right]^2$$

The region of the parameter space we are interested in is where the unitless decay rate is small compared to one, and $\exp(N\hat{\alpha})$ is large compared to one. In this region the true signal decays away in the observation time, but not before we obtain a good representative sample of it. We are not considering the case were the decay is so slow that the signal is nearly stationary, nor are we considering the case were the decay is so strong that the signal is gone within a small fraction of the observation time. Within these limits the sufficient statistic $\overline{h^2}$ is

$$\overline{h^2} \approx \frac{A_1^2\alpha}{4}\left[\frac{\alpha+\hat{\alpha}}{(\alpha+\hat{\alpha})^2+(\omega-\hat{w})^2}\right]^2.$$

Of course, the peak of the periodogram and the peak of the power spectral density occur at the same frequency. Indeed, for a simple harmonic signal the peak of the periodogram is the optimum frequency estimator. But in our problem (i.e. our model), the periodogram is not even approximately a valid estimator of the power spectrum, as Schuster supposed it to be. Consequently, even though these techniques give nearly the same frequency estimates, they give very different power spectral estimates.

B. The Simple Harmonic Signal with Lorentzian Decay.

The simple harmonic frequency problem just discussed may be generalized easily to include Lorentzian or Gaussian decay. We assume, for this discussion, that the decay is Lorentzian the generalization to other types of decay will become more obvious as we proceed. For a uniformly sampled interval the model we are considering is

$$f(l) = [\ B_1 \cos(\omega l) + B_2 \sin(\omega l)]e^{-\alpha l} \tag{45}$$

where l is restricted to values $(1 \leq l \leq N)$. We now have four parameters to estimate: the amplitudes B_1, B_2; the frequency ω; and the decay rate α. The solution to this problem is a straight forward application of the general procedures. The matrix g_{ij} (16) is given by

$$g_{ij} = \begin{bmatrix} \sum_{l=1}^{N} \cos^2(\omega l)e^{-2\alpha l} & \sum_{l=1}^{N} \cos(\omega l)\sin(\omega l)e^{-2\alpha l} \\ \sum_{l=1}^{N} \cos(\omega l)\sin(\omega l)e^{-2\alpha l} & \sum_{l=1}^{N} \sin^2(\omega l)e^{-2\alpha l} \end{bmatrix} .$$

These sums may be done explicitly or approximated in any number of ways. We will approximate them as follows:

$$c \equiv \sum_{l=1}^{N} \cos^2(\omega l)e^{-2l\alpha} \approx \sum_{l=1}^{N} \sin^2(\omega l)e^{-2l\alpha} \approx \frac{1}{2}\sum_{l=1}^{N} e^{-2l\alpha} = \frac{1}{2}\left[\frac{1 - e^{-2N\alpha}}{e^{2\alpha} - 1}\right] . \tag{46}$$

The off diagonal terms are at most the same order as the ignored terms; these terms are therefore ignored. The matrix g_{ij} can be written as

$$g_{ij} \approx \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} .$$

The orthonormal model functions may then be written as

$$H_1(l) = c^{-\frac{1}{2}} \cos(\omega l)e^{-\alpha l} \tag{47}$$

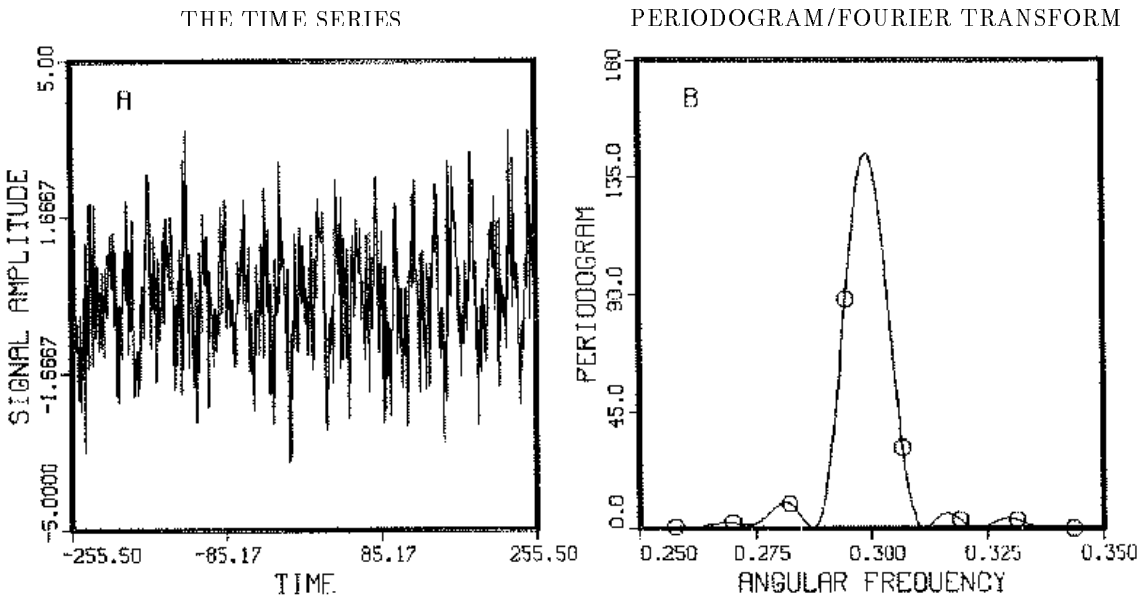
$$H_2(l) = c^{-\frac{1}{2}} \sin(\omega l)e^{-\alpha l} \tag{48}$$

The projections of the data onto the orthonormal model functions (24) are given by

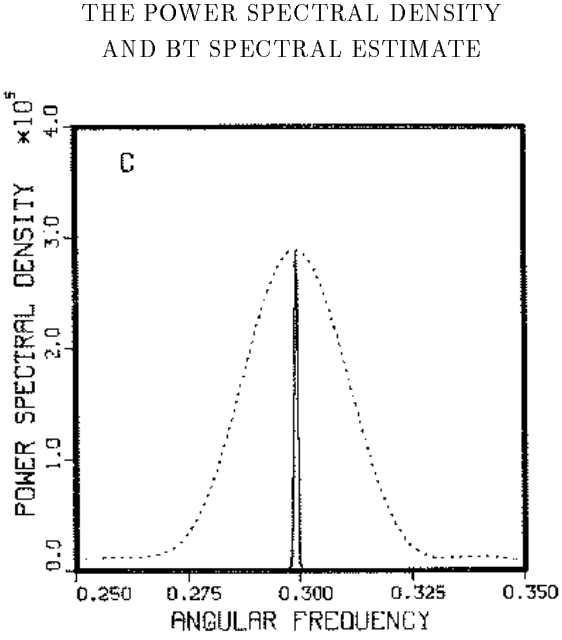
$$h_1 \equiv c^{-\frac{1}{2}} P(\omega, \alpha) = c^{-\frac{1}{2}} \sum_{l=1}^{N} d_l \cos(\omega l)e^{-\alpha l}$$

$$h_2 \equiv c^{-\frac{1}{2}} Q(\omega, \alpha) = c^{-\frac{1}{2}} \sum_{l=1}^{N} d_l \sin(\omega l)e^{-\alpha l}$$

Figure 2: Single Frequency Estimation



The data Fig. 2(A) contain a single harmonic frequency plus noise. There are 512 data points in the signal with $S/N \approx 1$. The Schuster periodogram, Fig. 2(B) solid curve, and the discrete Fourier transform, open circles, clearly show a sharp peak plus side lobes. These side lobes do not show up in the power spectral density, Fig. 2(C), because $\hat{p}(\omega) \approx 2C(\omega)P(\omega|DI)$; the normalized posterior probability is very sharply peaked around the maximum of the periodogram. The dotted line in Fig. 2(C) is a Blackman-Tukey spectrum with a Hanning window and 256 lag coefficients. If we had used a 1/10 lag as Tukey suggested the BT spectrum would have been nearly a flat line of this scale.



We would like to use the posterior probability to derive an estimate of the power spectral density $\hat{p}(\omega)$. We caution the reader again that the terms "power spectrum" or "spectral density" are used in the literature with several different meanings. Our meaning was defined previously as the expected power, over the joint posterior probability distribution of all the parameters, carried by the signal (not the noise), during the observation time. We made such an estimate in Section IV, but those estimates assumed the noise variance σ^2 was known. When the variance is unknown, the desired quantity is easily obtained from equation (39)

$$\hat{p}(\omega) \approx \left[\frac{P^2(\omega)}{c} + \frac{Q^2(\omega)}{s}\right] \frac{P(\omega|D,I)}{\int d\omega\ P(\omega|D,I)}$$

where we have dropped a term which is essentially the estimated variance of the noise. The estimated variance term can be neglected provided it is small compared to maximum of $\overline{h^2}$. This will occur whenever $\sum\langle A_j^2\rangle \gg \sigma^2$. In practice this approximation is good when one has a few hundred data points and a signal-to-noise ratio larger than about one. But if the number of data points is large, then this equation can be further simplified to obtain

$$\hat{p}(\omega) = C(\omega)\frac{d\omega P(\omega|D,I)}{\int d\omega\ P(\omega|D,I)} \tag{44}$$

$$P(\omega|D,I)\ \approx\ \left[1 - \frac{2C(\omega)}{N\overline{d^2}}\right]^{\frac{2-N}{2}}.$$

In N is large $P(\omega|D,I)$ is effectively a delta function; the peak value of $c(\omega)$ is approximately the total energy carried by the signal.

To obtain a better understanding of the use of this power spectral estimate, we have prepared an example: the data consist of a single harmonic frequency plus Gaussian white noise, Fig. 2. We generated these data from the following equation

$$d_j = 0.001 + \cos(0.3j + 1) + e_j$$

where j is a simple index running over the symmetric interval $-T$ to T in half integer steps $(2T+1 = 512)$, and e_i was a Gaussian distributed random number with unit variance. After generating the time series we computed its average value and subtracted it from each data point: this insures the data have zero mean value. Figure 2(A) is a plot of this computer simulated time series, Fig. 2(B) is a plot of the Schuster periodogram (continuous curve) with the discrete Fourier transform marked with open circles. The periodogram and the discrete Fourier transform have spurious side lobes, but these do not appear in the plot of the power spectral density Fig. 2(C) because, the processing in (39) will effectively suppress all but the very highest peak in the periodogram. This just illustrates numerically what we already knew analytically; it is only the very highest part of the periodogram that is important for estimation of a single frequency.

We have included a Blackman-Tukey spectrum estimate (dotted line) in Fig. 2(C) for comparison. The dotted line is a Blackman-Tukey spectrum using a Hanning window. The Blackman-Tukey spectrum has removed the side lobes at the cost of half the resolution in the discrete Fourier transform. The maximum lag was set at 256, i.e. over half the data. Had we used a lag of one-tenth as Tukey [3] advocates, the Blackman-Tukey spectrum would look nearly like a horizontal straight line on the scale of this plot.

with the simplest spectrum estimation problem. We do this because as was shown by Jaynes [2] when multiple, well-separated frequencies are present [$|\omega_j - omega_k| \gg 2\pi/N$], the spectrum estimation problem essentially separates into independent single-frequency problems. It is only when multiple frequencies are close together that we will need to use more general models.

A. The Simple Harmonic Spectrum.

The simplest frequency estimation problem one can discuss is the single frequency problem presented in Section II. For this problem, when the data are uniformly sampled in time the model can be written

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

where l is an index running over a symmetric time interval ($-T \leq l \leq T$) and ($2T+1 = N$). The matrix g_{ij} becomes

$$g_{ij} = \begin{bmatrix} \sum_{l=-T}^{l=T} \cos^2\omega l & \sum_{l=-T}^{l=T} \cos\omega l \sin\omega l \\ \sum_{l=-T}^{l=T} \cos\omega l \sin\omega l & \sum_{l=-T}^{l=T} \sin^2\omega l \end{bmatrix}.$$

For uniform time sampling the off diagonal terms are zero and the diagonal term may be summed explicitly to obtain

$$g_{ij} = \begin{bmatrix} c & 0 \\ 0 & s \end{bmatrix}$$

where c and s are given by

$$c = \frac{N}{2} + \frac{\sin(N\omega)}{2\sin(\omega)}$$

$$s = \frac{N}{2} - \frac{\sin(N\omega)}{2\sin(\omega)}.$$

Then the orthonormal model functions may be written as

$$H_1(t) = \frac{\cos(\omega t)}{\sqrt{c}}$$

$$H_2(t) = \frac{\sin(\omega t)}{\sqrt{s}}.$$

The posterior probability of a frequency ω in a uniformly sampled data set, independent of the signal amplitude, and phase, and the noise level, is given by equation (28). Substituting these model functions gives this as

$$P(\omega|D,I) \propto \left[1 - \frac{P(\omega)^2/c + Q(\omega)^2/s}{N\overline{d^2}}\right]^{\frac{2-N}{2}} \tag{43}$$

where $P(\omega)$ and $Q(\omega)$ are the squares of the real and imaginary parts of the discrete Fourier transform (7,8). Notice, when $N \gg 1$ the normalization constants c and s reduce to $N/2$ and (43) reduces to equation (13) found earlier.

D. The estimated variance σ.

One of the things that is of interest in an experiment is to estimate the noise power σ^2. We can obtain the expected value of σ as a function of the $\{\omega\}$ parameters; however, we can just as easily obtain $\langle sigma^s \rangle$ for any power s. Using equation (25), and the Jeffreys prior $1/\sigma$ we integrate:

$$\langle \sigma^s \rangle = \frac{\int_0^{+\infty} \frac{d\sigma}{\sigma} \sigma^s L(\{\omega\}, \sigma)}{\int_0^{+\infty} \frac{d\sigma}{\sigma} L(\{\omega\}, \sigma)}$$

to obtain

$$\langle \sigma^s \rangle = \Gamma\left(\frac{N-m-s}{2}\right) \Gamma\left(\frac{N-m}{2}\right)^{-1} \left\{ \frac{N}{2} \left[\overline{d^2} - \frac{m\overline{h^2}}{N} \right] \right\}^{\frac{s}{2}}. \tag{40}$$

For $s = 2$ this gives the estimated variance as

$$\langle \sigma^2 \rangle = \frac{N}{N-m-2} \left[\overline{d^2} - \frac{m\overline{h^2}}{N} \right]. \tag{41}$$

The estimate depends on the number m of expansion functions used in the model. The more model functions we use the smaller the last factor in (41), because by the Bessel inequality (29) the larger models fit the data better and $(\overline{d^2} - mN^{-1}\overline{h^2})$ decreases. But this should not decrease our estimate of σ^2 unless that factor decreases by more than we would expect from fitting the noise. The factor $N/(N-m-2)$ takes this into account; another example of sophisticated subtlety.

E. The estimated signal-to-noise ratio.

These results may be used to estimate the signal-to-noise ratio of the data. We define this as the square root of the (power carried by the signal) divided by the (mean power carried by the noise):

$$\frac{\text{Signal}}{\text{Noise}} = \left[\langle \sum_{j=1}^{m} A_j^2 \rangle N \sigma^2 \right]^{\frac{1}{2}}.$$

This may be obtained from equations (37)

$$\frac{\text{Signal}}{\text{Noise}} = \left\{ \frac{m}{N} \left[1 + \frac{\overline{h^2}}{\sigma^2} \right] \right\}^{\frac{1}{2}}. \tag{42}$$

A similar signal-to-noise ratio may be obtained when the noise variance σ is unknown by replacing σ in (44) by the estimated noise variance (42).

V. SPECTRAL ESTIMATION.

The previous sections surveyed the theory in generality. In this section we will specialize the analysis to frequency and spectrum estimates. Our ultimate aim is to derive explicit Bayesian estimates of the power spectrum and other parameters when multiple nonstationary frequencies are present. We will do this by proceeding through several stages beginning

Performing the integrals gives

$$\langle A_j A_k \rangle = h_j h_k + \sigma^2 \delta_{jk} \tag{37}$$

or, in view of (35), the posterior covariances are

$$\langle A_j A_k \rangle - \langle A_j \rangle \langle A_k \rangle = \sigma^2 \delta_{jk}.$$

The A_j parameters are uncorrelated [we defined the model functions $H_j(t)$ to ensure this], and each one is estimated to an accuracy $\pm\sigma$. Intuitively, we might anticipate this but we would not feel very sure of it.

The expectation value $\langle A_j A_k \rangle$ may be related back to the expectation value for the original model amplitudes by using equation (22):

$$\langle B_k B_l \rangle - \langle B_k \rangle \langle B_l \rangle = \sigma^2 \sum_{i=1}^{m} \frac{e_{ik} e_{il}}{\lambda_i}. \tag{38}$$

These are the explicit Bayesian estimates for the posterior covariances for the original model. These are the most conservative estimates (in the sense discussed before) one can make.

We can repeat these calculations for the second posterior moments in the case when σ is assumed unknown to see if obtaining explicit information about σ is of use. Of course, we expect the results to differ from the previous result since (38) depends explicitly on σ. Performing the required calculation gives

$$\langle A_j A_k \rangle = h_j h_k + \left(\frac{N}{N-2}\right)\left(\frac{2N-5}{2N-5-2m}\right)\left(\frac{2N-7}{2N-7-2m}\right)\left(\overline{d^2} - \frac{m\overline{h^2}}{N}\right)\delta_{jk}.$$

Comparing this with (37) shows that obtaining independent information about σ will affect the estimates of the second moments.

C. The power spectral density $\{\hat{\mathbf{p}}(\{\omega\})\}$.

Although not explicitly stated, we have calculated an estimate of the total energy of the signal. The estimated total energy of the signal is just $\sum \langle f^2(t_i) \rangle$, which in our orthonormal model is given by $\langle \sum A_j^2 \rangle$. Now we have computed this expectation value as a function of the $\{\omega\}$ parameters. We would like to express the total energy carried as a density. This is easily done, the power spectral density $\hat{p}(\{\omega\})$ is given by

$$\hat{p}(\{\omega\}) = \left[m\sigma^2 + m\overline{h^2}\right] \frac{P(\{\omega\}|DI\sigma)}{\int d\{\omega\} P(\{\omega\}|DI\sigma)}. \tag{39}$$

This function is the estimated energy carried by the signal (not the noise) per unit $\{\omega\}$.

That term of $m\sigma^2$ in (39) might be a little disconcerting to some; if (39) estimates the energy carried by the "signal" why does it include the noise power σ^2? If $\overline{h^2} \gg \sigma^2$ then the term is of no importance. But in the unlikely event $\overline{h^2} \gg \sigma^2$, then what is this term telling us? When these equations were formulated we essentially put in the fact that there is present noise of variance σ^2 and a signal in a subspace of m model functions. But then if $\overline{h^2} \gg \sigma^2$, there is only one explanation: the noise is such that its components on those m model functions just happened to cancel the signal. But if the noise just cancels the signal, the power carried by the signal must be equal to the power $m\sigma^2$ carried by the noise in those m functions; and that is exactly the answer one obtains. This is an excellent example of the sophisticated subtlety of Bayesian analysis.

A simple change of variables $u_j = (A_j - h_j)\sqrt{2\sigma^2}$ reduces the integrals to

$$\langle A_j(\{\omega\})\rangle = \frac{\int_{-\infty}^{+\infty} du_j \left\{\sqrt{2\sigma^2}u_j + h_j\right\} \exp[-u^2]}{\int_{-\infty}^{+\infty} du_j \exp[-u^2]}.$$

The first integral in the numerator is zero by symmetry and the second gives

$$\langle A_j(\{\omega\})\rangle = h_j(\{\omega\}). \tag{35}$$

This is the result one would expect. After all, we are expanding the data on an orthonormal set of vectors. The expansion coefficient is just the projection of the data onto the expansion vectors and that is what we find.

We can use these expected amplitudes $\langle A_j\rangle$ to calculate the expectation values of the amplitudes $\langle B_k\rangle$ in the nonorthonormal model. Using (21), these are given by

$$\langle B_k(\{\omega\})\rangle = \sum_{j=1}^{m} \frac{h_j e_{jk}}{\sqrt{\lambda_j}}.$$

Care must be taken in using this formula, because the dependence of the $\langle B_k\rangle$ on the $\{\omega\}$ is hidden. The functions h_j, the eigenvectors e_{kj} and the eigenvalues λ_j are all functions of the $\{\omega\}$ parameters. If one wishes to integrate over the $\{\omega\}$ parameters to obtain the best estimate of the B_k, then the integrals must be done over $\langle B_k(\{\omega\})\rangle$ times the probability density of the $\{\omega\}$ parameters, including the Jacobian (20).

We would like to compute $\langle A_j\rangle$ when the noise variance σ^2 is unknown to see if obtaining independent information about σ will affect these results. To do this we need the likelihood $L(\{A\},\{\omega\})$; as a function of $\{A\}$ and $\{\omega\}$ this is given by

$$L(\{\omega\},\{A\}) \propto \left[\overline{d^2} - \frac{m\overline{h^2}}{N} + \frac{1}{N}\sum_{i=1}^{m}(A_j - h_j)^2\right]^{-\frac{N}{2}}. \tag{36}$$

Using equation (36) and repeating the calculation for $\langle A_j\rangle$ one obtains the same result. Apparently it does not matter if we know the variance or not. We will make the same estimate of the amplitudes regardless. As with some of the other results discovered in this calculation, this is what one's intuition might have said; knowing σ affects the accuracy of the estimates but not their actual values. Indeed, the first moments were independent of the value of σ when the variance was known; it is hard to see how the first moments could suddenly become different when it is unknown.

B. The second posterior moments $\{\langle \mathbf{A_j A_k}\rangle\}$.

The second posterior moments $\langle A_j A_k\rangle$ cannot be independent of the noise variance σ^2, for that is what limits the accuracy of our estimates of the A_j. The second posterior moments when the variance is assumed known are given by

$$\langle A_j A_k\rangle = \frac{\int_{-\infty}^{+\infty} dA_1 \cdots dA_m A_j A_k L(\{\omega\},\{A\},\sigma)}{\int_{-\infty}^{+\infty} dA_1 \cdots dA_m L(\{\omega\},\{A\},\sigma)}.$$

accuracy considerably better than $\pm 10^{-5}$, if the amount of data N is large. It may, however, be hard to see at first glance how probability theory can justify this intuitive conclusion that we draw so easily.

But that is just what (28) and (34) tell us; Bayesian analysis leads us to it automatically and for any model functions. Even though you had no reason to expect it, if it turns out that the data can be fit almost exactly to a model function, then from the Bessel inequality (29) it follows that σ^2 must be extremely small and, if the other parameters are independent, they can all be estimated almost exactly.

IV. ESTIMATING THE NUISANCE PARAMETERS.

When the models had been rewritten in terms of these orthonormal model functions we were able to remove the nuisance parameters $\{A\}$ and σ. The integrals performed in removing the nuisance parameters were all Gaussian; therefore, one can always compute the moments of these parameters.

There are a number of reasons why these moments are of interest: the first moments of the amplitudes are needed if one intends to reconstruct the original model function $f(t)$; the second moments are related to the energy carried by the signal; the estimated noise variance σ^2 and the energy carried by the signal can be used to estimate the signal-to-noise ratio of the data. Thus the parameters $\{A\}$ and σ are not entirely "nuisance" parameters; it is of some interest to estimate them.

A.The expected amplitudes $\{\langle \mathbf{A_j} \rangle\}$.

To begin we will compute the expected amplitudes $\langle A_j \rangle$ in the case where the variance is assumed known. Now the likelihood (23) is a function of the $\{\omega\}$ parameters and to estimate the $\langle A_j \rangle$ independently of the $\{\omega\}$'s, we should integrate over these parameters. Because we have not specified the model functions we cannot do this once and for all. But we can obtain the estimated $\langle A_j \rangle$ as functions of the $\{\omega\}$ parameters. This gives us what would be the "best" estimate of the amplitudes if we knew the $\{\omega\}$ parameters. The expected amplitudes are given by

$$\langle A_j(\{\omega\}) \rangle = \frac{\int_{-\infty}^{+\infty} dA_1 \cdots dA_m A_j L(\{\omega\}, A, \sigma)}{\int_{-\infty}^{+\infty} dA_1 \cdots dA_m L(\{\omega\}, A, \sigma)}.$$

We will carry out the first integration in detail to illustrate the procedure, and later just give results. Using the likelihood (23) and having no prior information about A_j we assign a uniform prior and integrate over the $\{A_j\}$. Because the joint likelihood is a product of their independent likelihoods, all of the integrals except the one over A_j cancel:

$$\langle A_j(\{\omega\}) \rangle = \frac{\int_{-\infty}^{+\infty} dA_j A_j \exp\left\{-\frac{1}{2\sigma^2}\left[A_j^2 - 2A_j h_j\right]\right\}}{\int_{-\infty}^{+\infty} dA_j \exp\left\{-\frac{1}{2\sigma^2}\left[A_j^2 - 2A_j h_j\right]\right\}}.$$

estimates from the two experiments. Those parameters that are estimated to be about the same in the two experiments are probably real systematic effects. If some parameters are estimated to be quite different in the two experiments, they are almost surely spurious: i.e. not real effects but only artifacts of fitting the noise. The model should then be simplified, by removing the spurious parameters.

Unfortunately, a repetition is seldom possible with geophysical or economic time series, although one may split the data into two parts and see if they make about the same estimates. But repetition is usually easy and standard practice in the controlled environment of a physics experiment. Indeed, the physicist's common-sense criterion of a real effect is its reproducibility. Probability theory does not conflict with good common-sense judgment; it only sharpens it and makes it quantitative. A striking example of this is given in the scenario below.

Consider now the case that σ is completely unknown, where probability theory led us to (28). As we discussed in Section II, integrating σ out of the problem as a nuisance parameter is much like estimating σ from the data, and using that estimate in our equations; if σ is actually well determined by the data, the two procedures are essentially equivalent. We can see what estimate of σ is being made in (28) by comparing it to (27). Using the fact that if $x \ll 1$ and $N \gg 1$, $(1-x)^{-N} \approx \exp(Nx)$, (28) is crudely approximated by

$$P(\{\omega\}|D,I) \approx \exp\left\{\frac{N-m}{2}\,\frac{m\overline{h^2}}{N\overline{d^2}}\right\}$$

which corresponds to (27) with the variance σ^2 replaced with the estimate given by

$$(\sigma^2)_{est} = \frac{N}{N-m}\overline{d^2} = \frac{1}{N-m}\sum_{i=1}^{N} d_i^2. \tag{33}$$

In effect, probability theory tells us that we should suppose the first m degrees of freedom to be fit by the m model functions, and apportion the observed $\sum d_i^2$ to the remaining $(N-m)$ noise degrees of freedom. But this approximation is good only when $(N-m) \gg 1$ and $m\overline{h^2} \ll N\overline{d^2}$; i.e. there are many noise degrees of freedom and the fit to the data is poor. We shall presently find the exact mean value estimate of σ^2, which turns out to be [equations (40), (41)]

$$\langle\sigma^2\rangle = \frac{N}{N-m-2}\left(\overline{d^2} - \frac{m\overline{h^2}}{N}\right) \tag{34}$$

and agrees with (33) in this limit.

More interesting is the opposite extreme when (28) approaches a singular value. Consider the following scenario. You have obtained some data which are recorded automatically to six figures and look like this: $D = \{d_1 = 1.42316, d_2 = 1.50977, d_3 = 1.59638, \cdots\}$. But you have no prior knowledge of the accuracy of those data; for all you know, σ may be as large as 0.1 or even larger, making the last four digits garbage. But you plot the data, to determine a model function that best fits them. Suppose, for simplicity, that the model function is linear: $d_i = a + s_i + e_i$. On plotting d_i against i, you are astonished and delighted to see the data falling exactly on a straight line (i.e. to within the six figures given). What conclusions do you draw from this?

Intuitively, one would think that the data must be far "better" than had been thought; you feel sure that $\sigma < 10^{-5}$, and that you are therefore able to estimate the slope s to an

So that (30) becomes

$$\langle \overline{d^2} \rangle = \frac{m}{N} \overline{A^2} + \sigma^2.$$

Now, what value of $\overline{h^2}$ would he expect the data to generate? This is

$$\begin{aligned} \langle \overline{h^2} \rangle &= \frac{1}{m} \sum_{j=1}^{m} \langle h_j^2 \rangle \\ &= \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{i,k=1}^{N} \langle d_i d_k \rangle H_j(t_i) H_j(t_k) \right] \\ &= \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{i,k=1}^{N} (\langle d_i \rangle \langle d_k \rangle + \sigma^2 \delta_{ik}) H_j(t_i) H_j(t_k) \right]. \end{aligned} \tag{31}$$

But

$$\sum_{i=1}^{N} \langle d_i \rangle H_j(t_i) = \sum_{i=1}^{N} \sum_{l=1}^{m} A_l H_l(t_i) H_j(t_i) = \sum_{l=1}^{m} A_l \delta_{lj} = A_j$$

and (31) reduces to

$$\langle \overline{h^2} \rangle = \overline{A^2} + \sigma^2.$$

So he expects the left-hand side of the Bessel inequality (29) to be approximately

$$\langle \overline{d^2} \rangle - \frac{m \overline{h^2}}{N} \approx \frac{N-m}{N} \sigma^2. \tag{32}$$

This agrees very nicely with our intuitive judgment that as the number of model functions increases, we should be able to fit the data better and better. Indeed, when $m = N$, the $H_j(t_i)$ become a complete orthonormal set on S_N, and the data can always be fit exactly, as (32) suggests.

E. A simple diagnostic test.

If σ is known, these results give a simple diagnostic test for judging the adequacy of our model. Having taken the data, calculate $(N\overline{d^2} - m\overline{h^2})$. If the result is reasonably close to $(N-m)\sigma^2$, then the validity of the model is "confirmed" (in the sense that the data give no evidence against the model). On the other hand, if $(N\overline{d^2} - m\overline{h^2})$ turns out to be much larger than $(N-m)\sigma^2$, the model is not fitting the data as well as it should: it is "underfitting" the data. That is evidence either that the model is inadequate to represent the data (we need more model functions), or our supposed value of σ^2 is too low. The next order of business would be to investigate these possibilities.

It is also possible, although unusual, that $(N\overline{d^2} - m\overline{h^2})$ is far less than $(N-m)\sigma^2$; the model is "overfitting" the data. That is evidence either that our supposed value of σ is too large (the data are actually better than we expected), or that the model is more complex than it needs to be. By adding more model functions we can always improve the apparent fit, but if our model functions represent more detail than is really in the systematic effects at work, part of this fit is misleading: we are "fitting the noise."

A test to confirm this would be to repeat the whole experiment under conditions where we know the parameters should have the same values as before, and compare the parameter

D. An intuitive picture.

This gives us the following intuitive picture of the meaning of equations (25-28). The data $\{d_j, \cdots, d_N\}$ comprise a vector in an N-dimensional linear vector space S_N. The model equation

$$d_i = \sum_{j=1}^{m} A_j H_j(t_i) + e_i, \qquad (1 \leq i \leq N)$$

supposes that these data can be separated into a "systematic part" $f(t_i)$ and a white Gaussian "random part" e_i. Estimating the parameters of interest $\{\omega\}$ that are hidden in the model functions $H_j(t)$ amounts essentially to finding the values of the $\{\omega\}$ that permit $f(t)$ to make the closest possible (by the mean-square criterion) fit to the data. Put differently, probability theory tells us that the most likely values of the $\{\omega\}$ are those that allow a maximum amount of the mean-square data
dbar to be accounted for by the systematic term; from (29), those are the values that maximize $\overline{h^2}$.

However, we have N data points and only m model functions to fit to them. Therefore, to assign a particular model is equivalent to supposing that the systematic component of the data lies only in an m-dimensional subspace S_m of S_N. What kind of data should we then expect?

Let us look at the problem backwards for a moment. Suppose someone knows (never mind how he could know this) that the model is correct, and he also knows the true values of all the model parameters $(\{A\}, \{\omega\}, \sigma)$; call this the Utopian state of knowledge U; but he does not know what data will be found. Then the probability density that he would assign to any particular data set $D = \{d_1, \cdots, d_N\}$ is just our original sampling distribution (15):

$$P(D|U) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[d_i - f(t_i)]^2\right\}.$$

From this he would find the expectations and covariances of the data:

$$\langle d_i \rangle = f(t_i) \qquad (1 \leq i \leq N)$$

$$\langle d_i d_j \rangle - \langle d_i \rangle \langle d_j \rangle = (2\pi\sigma^2)^{-\frac{N}{2}} \int d^N x x_i x_j \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right] = \sigma^2 \delta_{ij}$$

therefore he would "expect" to see a value of $\overline{d^2}$ of about

$$\begin{aligned} \langle \overline{d^2} \rangle &= \frac{1}{N}\sum_{i=1}^{N}\langle d_i^2 \rangle \\ &= \frac{1}{N}\sum_{i=1}^{N}(\langle d_i^2 \rangle + \sigma^2) \\ &= \frac{1}{N}\sum_{i=1}^{N} f^2(t_i) + \sigma^2, \end{aligned} \tag{30}$$

but from the orthonormality (18) of the $H_j(t_i)$ we have

$$\sum_{i=1}^{N} f^2(t_i) = \sum_{l=1}^{N}\sum_{j,k=1}^{m} A_j A_k H_j(t_i) H_k(t_i) = \sum_{j=1}^{m} A_j^2.$$

(25) over σ gives

$$P(\{\omega\}|D,I) \propto \left[1 - \frac{m\overline{h^2}}{N\overline{d^2}}\right]^{\frac{m-N}{2}} . \tag{28}$$

This is again of the general form of the "student t-distribution" that we found before (13). But one may be troubled by the negative sign [in the big brackets (28)], which suggests that (28) might become singular. We pause to investigate this possibility by Bessel's famous argument.

C. The Bessel inequality.

Suppose we wish to approximate the data vector $\{d_1, \cdots, d_N\}$ by the orthogonal functions $H_j(t_i)$:

$$d_i = \sum_{j=1}^{m} a_j H_j(t_i) + \text{error}, \qquad (1 \leq i \leq N).$$

What choice of $\{a_1, \cdots, a_m\text{"}\}$" is "best?" If our criterion of "best" is the mean-square error, we have

$$\begin{aligned} 0 &\leq \sum_{i=1}^{N} \left(d_i - \sum_{j=1}^{m} a_j H_j(t_i) \right)^2 \\ &= N\overline{d^2} + \sum_{j=1}^{m} (a_j^2 - 2a_j h_j) \\ &= N\overline{d^2} - m\overline{h^2} + \sum_{j=1}^{m} (a_j - h_j)^2 \end{aligned}$$

where we have used (22) and the orthonormality (18). Evidently, the "best" choice of the coefficients is

$$a_j = h_j, \qquad (1 \leq j \leq m)$$

and with this best choice the minimum possible mean-square error is given by the Bessel inequality

$$\overline{d^2} - \frac{m}{N}\overline{h^2} \geq 0 \tag{29}$$

with equality if and only if the approximation is perfect. In other words, (28) becomes singular somewhere in the parameter space if and only if the model

$$f(t) = \sum_{j=1}^{m} A_j H_j(t)$$

can be fitted to the data exactly. But in that case we know the parameters by deductive reasoning, and probability theory becomes superfluous. Even so, probability theory is still working correctly, indicating an infinitely greater probability of the true parameter values than for any others.

B. Elimination of the nuisance parameters.

We are now in a position to proceed as before. Because the calculation is essentially identical to the single harmonic calculation we will proceed very rapidly. The likelihood can now be factored into a set of independent likelihoods for each of the A_j. It is now possible to remove the nuisance parameters easily. Using the joint likelihood (15) we make the change of function (17) and the change of variables (19) to obtain the joint likelihood of the new parameters

$$L(\{A\},\{\omega\},\sigma) \propto \sigma^{-N} \times \exp\left\{-\frac{N}{2\sigma^2}\left[\overline{d^2} - \frac{2}{N}\sum_{j=1}^{m} A_j h_j + \frac{1}{N}\sum_{j=1}^{m} A_j^2\right]\right\} \tag{23}$$

$$h_j \equiv \sum_{i=1}^{N} d_i H_j(t_i), \qquad (1 \leq j \leq m). \tag{24}$$

Here h_j is just the projection of the data onto the orthonormal model function H_j. In the simple harmonic analysis performed in Section II, the $P(\omega)$ and $Q(\omega)$ functions are the analogues of these h_j functions. However, the h_j functions are more general: we did not make any approximations in deriving them. The orthonormality of the H_j functions was used to simplify the quadratic term. This simplification makes it possible to complete the square in the likelihood and to integrate over the A_j's, or any selected subset of them.

As before, if one has prior information about these amplitudes, then here is where it should be incorporated. We will assume that no prior information is available, and thus obtain the most conservative estimates by assigning the amplitudes a uniform prior. Then performing the m integrations one obtains

$$L(\{\omega\},\sigma) \propto \sigma^{-N+m} \times \exp\left\{-\frac{N\overline{d^2} - m\overline{h^2}}{2\sigma^2}\right\} \tag{25}$$

where

$$\overline{h^2} \equiv \frac{1}{m}\sum_{j=1}^{m} h_j^2 \tag{26}$$

is the mean-square of the observed projections. This equation is the analogue of equation (6) in the simple harmonic calculation. Although it is exact and far more general, it is actually simpler in structure and gives us a better intuitive understanding of the problem, as we will see in the Bessel inequality below. In a sense $\overline{h^2}$ is a generalization of the periodogram to arbitrary model functions. In its dependence on the parameters $\{\omega\}$ it is a sufficient statistic for all of them.

Now if σ is known, then the problem is again completed provided we have no additional prior information. The joint posterior probability of the $\{\omega\}$ parameters, conditional on the data and our knowledge of σ, is

$$P(\{\omega\}|D,\sigma,I) \propto \exp\left\{\frac{m\overline{h^2}}{2\sigma^2}\right\}. \tag{27}$$

But if there is no prior information available about the noise, then σ is a nuisance parameter and can be eliminated as before. Using the Jeffreys prior $1/\sigma$ and integrating

then at least one of the model functions $G_j(t)$ is redundant and can be removed from the model without changing the problem.

We suppose that redundant model functions have been removed, so that g_{jk} is positive definite and of rank m in what follows. Let e_{kj} represent the j'th component of the k'th normalized eigenvector of g_{jk}; i.e. $\sum_{k=1}^{m} g_{jk}e_{lk} = \lambda_l e_{lj}$, where λ_l is the l'th eigenvalue of g_{jk}. Then the functions $H_j(t)$, defined as

$$H_j(t) = \frac{1}{\sqrt{\lambda_j}} \sum_{k=1}^{m} e_{jk} G_k(t), \tag{17}$$

have the desired orthonormality condition,

$$\sum_{i=1}^{N} H_j(t_i) H_k(t_i) = \delta_{jk}. \tag{18}$$

The model equation can now be rewritten in terms of these orthonormal functions as

$$f(t) = \sum_{k=1}^{m} A_k H_k(t).$$

The amplitudes B_k are linearly related to the A_k by

$$B_k = \sum_{j=1}^{m} \frac{A_j e_{jk}}{\sqrt{\lambda_j}} \quad \text{and} \quad A_k = \sqrt{\lambda_k} \sum_{j=1}^{m} B_j e_{kj}. \tag{19}$$

The volume elements are given by

$$dB_1 \cdots dB_m = \left| \frac{e_{lj}}{\sqrt{\lambda_j}} \right| dA_1 \cdots dA_m. \tag{20}$$

The Jacobian is a function of the $\{\omega\}$ parameters and is a constant so long as we are not integrating over these $\{\omega\}$ parameters. At the end of the calculation the linear relations between the A's and B's can be used to calculate the expected values of the B's from the expected value of the A's and the same is true of the second posterior moments

$$\langle B_k \rangle = \sum_{j=1}^{m} \frac{\langle A_j \rangle e_{jk}}{\sqrt{\lambda_j}} \tag{21}$$

$$\langle B_k B_l \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{e_{ik} e_{jl} \langle A_i A_j \rangle}{\sqrt{\lambda_i \lambda_j}}. \tag{22}$$

The two operations of making a transformation on the model functions and a change of variables will transform any nonorthonormal model of the form (14) into an orthonormal model (18). We still have a matrix to diagonalize, but this is done once at the beginning of the calculation. It is not necessary to carry out the inverse transformation if we are interested only in estimating the $\{\omega\}$ parameters, since these parameters are transferred into the $H_j(t)$ functions.

level, and find what probability theory has to say about the frequency alone. In addition, it has given us an indication about how to proceed to more general problems. If we had used a model where the quadratic term in the likelihood function did not simplify, we would have a more complicated analytical solution. Although any multivariate Gaussian integral can be done, the key to being able to remove the nuisance parameters easily, and above all, selectively was that the likelihood factored into independent parts. In the full spectrum analysis problem worked on by Jaynes, [2] the nuisance parameters were not independent, and the explicit solution required the diagonalization of a matrix that could be quite large. To understand an easier approach to complex models, suppose we have a model of the form

$$d_i = f(t_i) + e_i$$

$$f(t) = \sum_{j=1}^{m} B_j G_j(t). \tag{14}$$

The model functions, $G_i(t)$, are themselves functions of other parameters which we collectively label $\{\omega\}$ (these parameters might be frequencies, chirp rates, decay rates, or any other quantities one could encounter). Now if we substitute this model into the likelihood (3) the simplification that occurred in (4) does not take place:

$$L(\{B\}\{\omega\}\sigma) \propto \sigma^{-N} \exp\left\{ -\frac{N}{2\sigma^2}\left[\overline{d^2} - \frac{2}{N}\sum_{j=1}^{m}\sum_{i=1}^{N} B_j d_i G_j(t_i) + \frac{1}{N}\sum_{j=1}^{m}\sum_{k=1}^{m} g_{jk} B_j B_k\right]\right\} \tag{15}$$

$$g_{jk} = \sum_{i=1}^{N} G_j(t_i) G_k(t_i). \tag{16}$$

If the desired simplification is to take place the matrix g_{jk} must be diagonal.

A. The orthonormal model equations.

For the matrix g_{jk} to be diagonal the model functions G_j must be made orthogonal. This can be done by taking appropriate linear combinations of them. But care must be taken; we do not desire a set of orthogonal functions of a continuous variable t, but a set of vectors which are orthogonal when summed over the discrete sampling times t_i. It is the sum over t_i appearing in the quadratic term of the likelihood which must simplify.

To accomplish this, consider the real symmetric matrix g_{jk} defined above (16). Since for all $\sum x_j^2 > 0$,

$$\sum_{j,k=1}^{m} g_{jk} x_j x_k = \sum_{i=1}^{N}\left(\sum_{j=1}^{m} x_j G_j(t_i)\right)^2 \geq 0$$

g_{jk} is positive definite if it is of rank m. If it is of rank $r < m$, then the model functions $G_j(t)$ and/or the sampling times t_i were poorly chosen. That is, if a linear combination of the $G_j(t)$ is zero at every sampling point:

$$\sum_{j=1}^{m} x_j G_j(t_i) = 0, \qquad (1 \leq i \leq N)$$

just completed; even though we know this analysis is too simple to be realistic for these numbers. We have plotted the time series from 1700 to 1985, Fig. 1(A). A cursory examination of this time series does indeed show a cyclic variation with a period of about 11 years. Next we computed, Fig. 1(B) the Schuster periodogram (continuous curve) and the discrete Fourier transform (open circles); these clearly show a maximum with a period near 11 years. It is a theorem that the discrete Fourier transform contains all the information that is in the periodogram; but one sees that the information is much more apparent to the eye in the continuous periodogram. We then computed the "student t-distribution" (13), Fig. 1(C), to determine the accuracy of the frequency estimate without making any assumption about σ. Now because of the processing in equation (13) all details in the periodogram have been suppressed and only the peak at 11 years remains.

We determined the accuracy of the frequency estimate as follows: We located the maximum of the "student t-distribution," integrated about a symmetric interval, and recorded the enclosed probability at a number of points. This gives:

period in years		**accuracy in years**	**probability enclosed**
11.04	$\pm$	0.012	0.50
	$\pm$	0.015	0.62
	$\pm$	0.020	0.75
	$\pm$	0.026	0.90

as the error estimates. According to this, there is not one chance in ten that the true period differs from 11.04 years by more than ten days. At first glance, this appears too good to be true.

But what does raw eye-balling of the data give? In 285 years, there are about $285/11 = 26$ cycles. If we can count these to an accuracy of $\pm 1/4$ cycle, our period estimate would be about

$$(f)_{est} = 11 \text{ years} \pm 39 \text{ days}\,.$$

Probability averaging of the noise, as discussed above, would reduce this uncertainty by about a factor of $\sqrt{285/10} = 5.3$, giving

$$(f)_{est} = 11 \text{ years} \pm 7.3 \text{ days}, \quad \text{or} \quad (f)_{est} = 11 \ \pm 0.02 \text{ years}$$

which corresponds nicely with the result of the probability analysis.

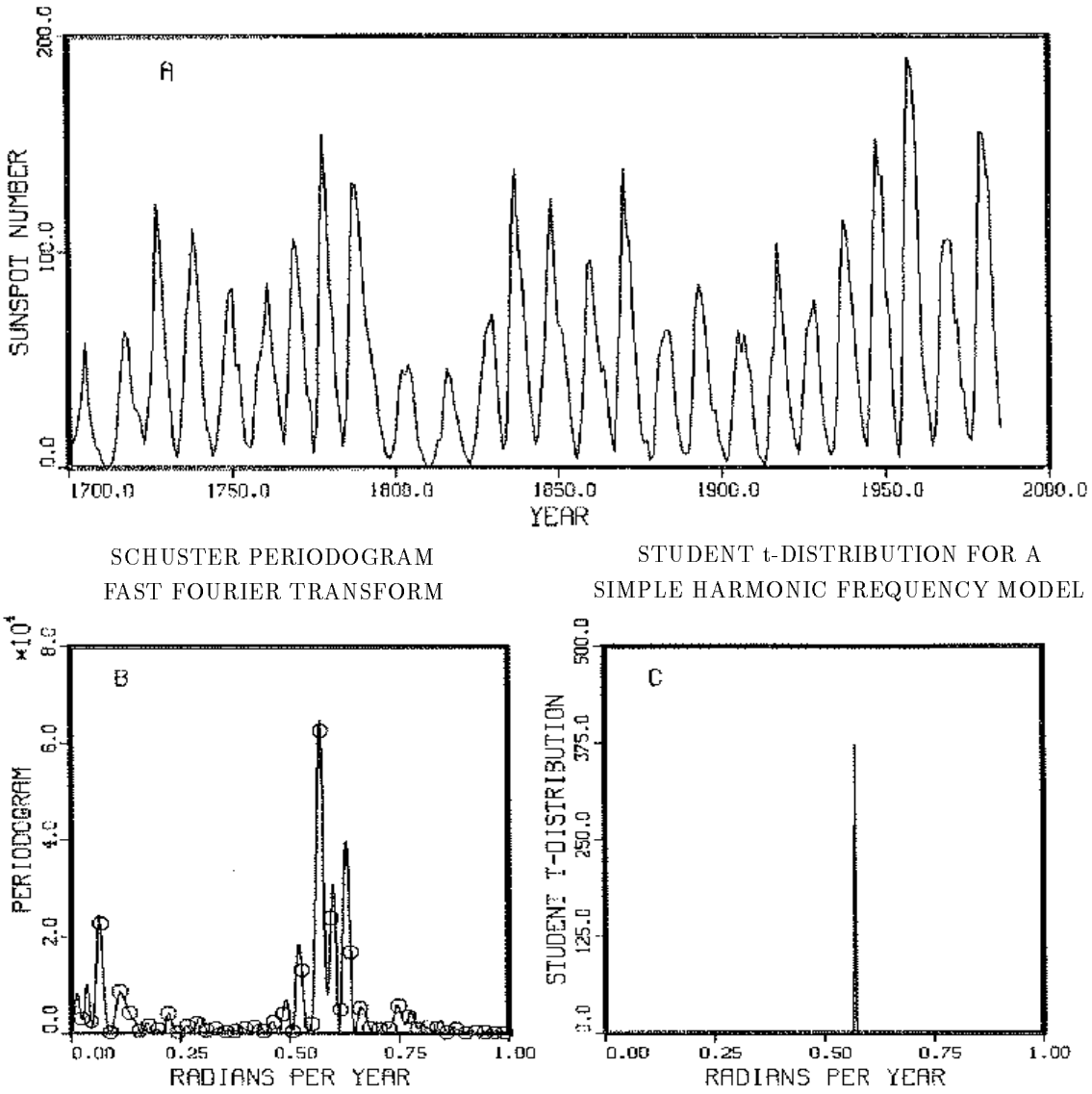
These results came from analyzing the data by a model which said there is nothing present but a single sinusoid plus noise. Probability theory, given this model, is obliged to consider everything in the data that cannot be fit to a single sinusoid to be noise. But a glance at the data shows clearly that there is more present than our model assumed: therefore, probability theory must estimate the noise to be quite large.

This suggests that we might do better by using a more realistic model which allows the "signal" to have more structure. Such a model can be fit to the data more accurately, therefore it will estimate the noise to be smaller. This should permit a still better period estimate!

III. THE GENERAL MODEL EQUATION PLUS NOISE.

These simple results already represent progress toward the more general spectral analysis problem because we were able to remove consideration of the amplitude, phase and noise

Figure 1: Wolf's Relative Sunspot Numbers



Wolf's relative sunspot numbers (A) have been collected on a yearly basis since 1700. The periodogram (B) contains evidence of several complex phenomena. In spite of this the single frequency model posterior probability density (C) picks out the 11.04 year cycle to an estimated accuracy of ± 10 days.

a nuisance parameter. We eliminate it in much the same way as the amplitudes were eliminated. Now σ is restricted to positive values and additionally it is a scale parameter. The prior which indicates "complete ignorance" of a scale parameter α is $d\alpha/\alpha = d\log\alpha$. This prior was first suggested by Sir Harold Jeffreys [8] some 50 years ago. It has since been derived by several different methods [9, 10] as being the only consistent prior which indicates "complete ignorance" of a scale parameter, by several different criteria of "consistent". Multiplying equation (6) by the Jeffreys prior and integrating over all positive values gives

$$P(\omega|D,I) \propto \left[1 - \frac{2C(\omega)}{N\overline{d^2}}\right]^{\frac{2-N}{2}} . \tag{13}$$

This is called a "student t-distribution" for historical reasons, although it is expressed here in very nonstandard notation. In our case it is the posterior probability density that a stationary harmonic frequency ω is present in the data when we have no prior information about σ.

This simple result shows explicitly why the discrete Fourier transform tends to peak at the location of a frequency when the data are noisy. Namely, the discrete Fourier transform is directly related to the probability that a simple harmonic frequency is present in the data, even when the noise level is unknown. Additionally, zero padding a time series (i.e. adding zeros at its end to make a longer series) and then taking the discrete Fourier transform of the padded series, is equivalent to calculating the Schuster periodogram at smaller frequency intervals. If the signal one is analyzing is a simple harmonic frequency plus noise, then the maximum of the periodogram will be the best estimate of the frequency in the absence of prior information about it.

If the signal is other than a single sinusoid, then the above analysis does not apply and the discrete Fourier transform may peak at the "incorrect" frequencies: i.e. frequencies different from those we wish to estimate. This occurs, not because the discrete Fourier transform is "wrong," but because it is answering what we should then regard as the "wrong" question. Put differently, the discrete Fourier transform is by definition the spectrum of the noisy data; but we are trying to use it to estimate a frequency in a particular model. If that model is other than a simple harmonic model (i.e. if there are several signals present, or the variation is periodic but not sinusoidal, or there is decay or chirp), there is no reason to expect the discrete Fourier transform to be a reasonable data analysis method for our different model. For each model, we must re-examine what probability theory has to say.

To apply these procedures to more complex signals we must generalize the formalism, this is done in Section III; for now we apply the simple result (13) to Wolf's relative sunspot numbers.

E. An example: Wolf's relative sunspot numbers.

Wolf's relative sunspot numbers are, perhaps, the most analyzed set of data in all of spectrum analysis. These numbers (defined as: $R = k[10g + f]$, where g is the number of sunspot groups, f is the number of individual sunspots, and k is used to scale different telescopes onto a common scale) have been collected on a yearly basis since 1700, and on a monthly basis since 1748 [11]. The exact physical mechanism which generates the sunspots is unknown and no complete theory exists. Different analyses of these numbers have been published more or less regularly since their tabulation began. Here we will analyze the sunspot numbers with a number of different models including the simple harmonic analysis

According to E. T. Jaynes, Tukey himself acknowledged [7] that his method fails to give optimum resolution, but held this to be of no importance because "real time series do not have sharp lines." Nevertheless, this misconception is so strongly held that there have been attacks on the claims of Bayesian/Maximum Entropy spectrum analysts to be able to achieve results like (10) when the assumed conditions are met. Some have tried to put such results in the same category with circle squaring and perpetual motion machines. Therefore we want to digress to explain the premise in very elementary physical terms why it is the Bayesian result (9) that does correspond to what a skilled experimentalist can achieve.

Suppose first that our only data analysis tool is our own eyes looking at a plot of the raw data of duration $T = 1$ sec., and that the unknown frequency f in (10) is 100Hz. Now anyone who has looked at a record of a sinusoid and equal amplitude wide-band noise, knows that the cycles are quite visible to the eye. One can count the total number of cycles in the record confidently (using interpolation to help us over the doubtful regions) and will feel quite sure that the count is not in error by even one cycle. Therefore by raw eyeballing of the data and counting the cycles, one can achieve an accuracy of

$$\delta f \approx \frac{1}{T} = 1 \text{ Hz}. \tag{12}$$

But in fact, if one draws the sine wave that seems to fit the data best, he can make a quite reliable estimate of how many quarter-cycles were in the data, and thus achieve

$$\delta f \approx \frac{1}{4T} = 0.25 \text{ Hz}$$

corresponding just to the periodogram width (11). Then the use of probability theory needs to surpass the naked eye by another factor of ten to achieve the Bayesian width (10).

What probability theory does is essentially to average out the noise in a way that the naked eye cannot do. If we repeat some measurement N times, any randomly varying component of the data will be suppressed relative to the systematic component by a factor of $N^{-\frac{1}{2}}$, the standard rule.

In the case considered, we assumed $N = 1000$ data points. If they were all independent measurements of the same quantity with the same accuracy, this would suppress the noise by about a factor of 30. But in our case not all measurements are equally cogent for estimating the frequency. Data points in the middle of the record contribute very little to the result; only data points near the ends are highly relevant for determining the frequency, so the effective number of observations is less than 1000. The probability analysis leading to (25) indicates that the "effective number of observations" is only about $N/10 = 100$; thus the Bayesian width (25) that results from the exponential peaking of the periodogram now appears to be, if anything, somewhat conservative. Indeed, that is what Bayesian analysis always does when we use smooth, uninformative priors for the parameters, because then probability theory makes allowance for all possible values that they might have. As noted before, if we had any cogent prior information about ω and expressed it in a narrower prior, we would be led to still better results; but they would not be much better unless the prior range became comparable to the width of the likelihood $L(\omega)$.

D. Elimination of the noise level σ.

The above analysis is valid whenever the noise variance (or power) is known. Frequently one has no independent prior knowledge of the noise. The noise variance σ^2 then becomes

where now, not distinguishing between N and $(N-1)$,

$$\delta f = \frac{\sigma}{2\pi A_1 T}\sqrt{48/N} = \ 1.1\frac{\sigma}{A_1 T\sqrt{N}} \text{ Hz.} \tag{9}$$

For example, if we have an RMS signal-to-noise ratio $= A_1/\sqrt{2}\sigma = 1$, and we take data every $\Delta t = 10^{-3}$ sec. for $T = 1$ second, thus getting $N = 1000$ data points, the theoretical accuracy for determining the frequency of a single steady sinusoid is

$$\delta f = \frac{1.1}{\sqrt{2000}} = 0.025 \text{ Hz} \tag{10}$$

while the Nyquist frequency for the onset of aliasing is $f_N = (2\Delta t)^{-1} = 500$ Hz, greater by a factor of 20,000.

To some, this result will be quite startling. Indeed, had we considered the periodogram itself to be a spectrum estimator, we would have calculated instead the width of its central peak. A noiseless sinusoid of frequency $\hat{w}$ would have a periodogram proportional to

$$C(\omega) \propto \frac{\sin^2\{N(\omega-\hat{w})/2\}}{\sin^2\{(\omega-\hat{w})/2\}}$$

thus the half-width at half amplitude is given by $|N(\hat{w}-\omega)| = \pi/4$ or $\delta\omega = \pi/2N$. Converting to physical units, the periodogram will have a width of about

$$\delta f = \frac{1}{4N\Delta t} = \frac{1}{4T} = 0.25 \text{ Hz} \tag{11}$$

just ten times greater than the value (10) indicated by probability theory. This factor of ten is the amount of narrowing produced by the exponential peaking of the periodogram in (7), even for unity signal-to-noise ratio.

But some would consider even the result (11) to be a little overoptimistic. The famous Rayleigh criterion [6] for resolving power of an optical instrument supposes that the minimum resolvable frequency difference corresponds to the peak of the periodogram of one sinusoid coming at the first zero of the periodogram of the second. This is twice (11):

$$\delta f_{\text{Rayleigh}} = \frac{1}{2T} = 0.5\text{Hz.}$$

There is widely believed "folk-theorem" among theoreticians without laboratory experience, which seems to confuse the Rayleigh limit with the Heisenberg uncertainty principle, and holds that (12) is a fundamental irreducible limit of resolution. Of course there is no such theorem, and workers in high resolution NMR have been routinely determining line positions to an accuracy that surpasses the Rayleigh limit by an order of magnitude, for thirty years.

The misconception is perhaps strengthened by the curious coincidence that (12) is also the minimum half-width that can be achieved by a Blackman-Tukey spectrum analysis [3] (even at infinite signal–noise ratio) because the "Hanning window" tapering function that is applied to the data to suppress side-lobes (the secondary maxima of $[\sin(x)/x]^2$ just doubles the width of the periodogram. Since the Blackman-Tukey method has been used widely by economists, oceanographers, geophysicists, and engineers for many years, it has taken on the appearance of an optimum procedure.

C. Resolving power.

To obtain the (mean) ± (standard deviation) approximation for the frequency ω we expand $C(\omega)$ about the peak

$$C(\omega) = C(\omega_{max}) - \frac{b^2}{2}(\omega - \omega_{max})^2 + \cdots$$

where

$$b^2 \equiv -C''(\omega_{max}) > 0$$

we have a Gaussian approximation

$$\langle \hat{p}(\omega) \rangle \approx 2C(\omega_{max}) \exp\left\{ -\frac{b^2(\omega - \omega_{max})^2}{2\sigma^2} \right\}$$

from which we would estimate of the frequency

$$\omega_{est} = \omega_{max} \pm \frac{\sigma}{b}.$$

The accuracy depends on the curvature of $C(\omega)$ at its peak. For example, if the data are composed of a single sine wave plus noise $e(t)$ of standard deviation σ

$$d_t = A_1 \cos(\hat{w}t) + e_t$$

and $\sigma \ll A_1$, then as found by Jaynes [2]:

$$\omega_{max} \approx \hat{w}$$

$$C(\omega_{max}) \approx \frac{NA_1^2}{4}$$

$$\omega_{est} \approx \hat{w} \pm \frac{\sigma}{A_1}\sqrt{48/N^3} \tag{8}$$

which indicates, as common sense would lead us to expect, that the accuracy depends on the signal-to-noise ratio, and quite strongly on how much data we have.

However, before comparing these results with experience we need to note that we are here using dimensionless units, since we took the data sampling interval to be 1. Converting to ordinary physical units, let the sampling interval be Δt seconds, and denote by f the frequency in Hz. Then the total number of cycles in our data record is

$$\frac{\hat{w}(N-1)}{2\pi} = (N-1)\hat{f}\Delta t = \hat{f}T$$

where $T = (N-1)\Delta t$ seconds is the duration of our data run. So the conversion of dimensionless ω to f in physical units is

$$f = \frac{\omega}{2\pi\Delta t} \text{ Hz}.$$

The frequency estimate (8) becomes

$$f_{est} = \hat{f} \pm \delta f \text{ Hz}$$

which expresses what the data and prior information have to tell us about ω, regardless of the value of θ.

Usually, the prior probabilities are independent:

$$P(\omega\theta|I) = P(\omega|I)P(\theta|I).$$

But even if they are not, the prior can be factored as

$$P(\omega\theta|I) = P(\omega|I)P(\theta|\omega, I)$$

so the calculation can always be organized as follows: calculate the "quasi-likelihood" of ω;

$$L(\omega) = \int d\theta P(D|\omega, \theta, I)P(\theta|\omega, I) \tag{5}$$

then, to within a normalization constant, the desired distribution for ω is

$$P(\omega|D, I) \propto P(\omega|I)L(\omega).$$

If we had prior information about the nuisance parameters (such as: they had to be positive, they could not exceed an upper limit, or we had independently measured values for them) then equation (5) would be the place to incorporate that information into the calculation. We assume no prior information about the amplitudes A_1 and A_2 and assign them a prior probability which indicates "complete ignorance of a location parameter." This prior is a uniform, flat, prior density; it is called an improper prior probability because it is not normalizable. In principle, we should approach an improper prior as the limit of a sequence of proper priors. However, in this problem there are no difficulties with the use of the uniform prior because the Gaussian cutoff in the likelihood function ensures convergence in (5), and the result is the same.

Upon multiplying and integrating the likelihood (4) with respect to A_1 and A_2 one obtains the joint quasi-likelihood of ω and σ:

$$L(\omega, \sigma) \propto \sigma^{-N+2} \times \exp\left\{-\frac{N}{2\sigma^2}\left[\overline{d^2} - \frac{2C(\omega)}{N}\right]\right\} \tag{6}$$

where

$$C(\omega) \equiv \frac{1}{N}\left[P^2(\omega) + Q^2(\omega)\right]$$

the Schuster periodogram $C(\omega)$, [5]. has appeared in a very natural way. If one knows the variance σ from some independent source and has no additional prior information about ω, then the problem is completed. The posterior probability density for ω is proportional to

$$P(\omega|D, \sigma, I) \propto \exp\left(\frac{C(\omega)}{\sigma^2}\right). \tag{7}$$

Because we have assumed no prior information about A_1, A_2, and ω this probability density will yield the most conservative estimate one can make from probability theory of ω and its probable accuracy.

where

$$P(\omega) = \sum_{i=1}^{N} d_i \cos(\omega t_i)$$

$$Q(\omega) = \sum_{i=1}^{N} d_i \sin(\omega t_i)$$

are the sine and cosine transforms of the data and

$$\overline{d^2} = \frac{1}{N}\sum_{i=1}^{N} d_i^2$$

is the observed mean-square data value. For a simplified preliminary discussion we have assumed the data have zero mean value (any nonzero average value has been subtracted from the data), and we simplified the quadratic term as follows:

$$\sum_{i=1}^{N} \cos^2(\omega t_i) = \frac{N}{2} + \frac{1}{2}\sum_{i=1}^{N} \cos(2\omega t_i) \approx \frac{N}{2}.$$

The neglected term is of order one, and is assumed small compared to N except for the isolated special case of $\omega \approx 0$. We have specifically eliminated this special case from consideration by subtracting off the constant term. A similar simplification occurs with the sine squared term. In addition, the cross term, $2A_1A_2\sum_{i=1}^{N}\cos(\omega t_i)\sin(\omega t_i)$, is at most of the same order as the terms we just ignored; therefore, this term is also ignored.

The assumption that this cross term is zero is equivalent to assuming the sine and cosine functions are orthogonal on the discrete time sampled region. Indeed, this is the actual case for uniformly spaced time intervals; however, even without uniform spacing this is a good assumption provided N is large. The assumption that the cross terms are zero by orthogonality will prove to be the key to generalizing this problem to more complex models, and eventually the assumptions that we are making now will become exact by a change of variables.

B. Elimination of nuisance parameters.

In a harmonic analysis one is usually interested only in the frequency ω. Then if the amplitude, phase, and the variance of the noise are unknown, they are referred to as nuisance parameters. The principles of probability theory uniquely determine how nuisance parameters should be eliminated. Suppose ω is a parameter of interest, and θ is a nuisance parameter. What we want is $P(\omega|D,I)$, the posterior probability (density) of ω. This may be calculated as follows: first calculate the joint posterior probability (or probability density) of ω and θ by Bayes' theorem:

$$P(\omega\theta|D,I) = P(\omega\theta|I)\frac{P(D|\omega,\theta,I)}{P(D|I)}$$

and then integrate out θ, obtaining the marginal posterior probability density for ω:

$$P(\omega|D,I) = \int d\theta P(\omega\theta|D,I)$$

different prior probabilities are small provided we have a reasonable amount of data. A good rule of thumb is that one more power of A^{-1} in the prior has about the same effect on our conclusions as having one more data point.

A. The likelihood function.

To construct the likelihood we take the difference between the model function, or "signal," and the data. If we knew the true signal, then this difference would be just the noise. We wish to assign a noise prior probability density which is consistent with the available prior information. The prior should be as uninformative as possible to prevent us from "seeing" things in the data which are not there. To derive this prior probability for the noise is a simple application of the principle of maximum entropy, or if the noise is known to be the result of many small independent effects, the central limit theorem of probability theory leads to the Gaussian form independently of the fine details. Regardless; the prior probability assignment will be the same:

$$P(e_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right).$$

Next we apply the product rule from probability theory to obtain the probability of a set of noise values $\{e_1, \cdots, e_N\}$ given by

$$P(e_1, \cdots, e_N) = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)\right]. \tag{2}$$

For a detailed discussion of why and when a Gaussian distribution should be used for the noise probability, see the original paper by Jaynes [2].

Additionally, the book of Jaynes' collected papers contains a discussion of the principle of maximum entropy and much more [4].

The probability that we should obtain the data $D = \{d_1 \cdots d_N\}$ given the parameters is

$$P(D|H,I) \propto L(A_1, A_2, \omega, \sigma) = \prod_{i=1}^{N} \sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}[d_i - f(t_i)]^2\right\}$$

$$L(A_1, A_2, \omega, \sigma) = \sigma^{-N} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[d_i - f(t_i)]^2\right\}. \tag{3}$$

The usual way to proceed is to fit the sum in the exponent. Finding the parameter values which minimize this sum is called least squares. The (in the Gaussian case) equivalent procedure of finding parameter values that maximize $L(A_1, A_2, \omega, \sigma)$ is called "maximum likelihood." The maximum likelihood procedure is more general than least squares: it has theoretical justification when the likelihood is not Gaussian. The departure of Jaynes was to use (3) instead in Bayes' theorem (1), and then to remove the phase and amplitude from further consideration by integration over these parameters. To do this we first expand (3)

$$L(A_1, A_2, \omega, \sigma) \propto \sigma^{-N} \exp\left\{-\frac{N}{2\sigma^2}\left[\overline{d^2} - \frac{2}{N}[A_1 P(\omega) + A_2 Q(\omega)] + \frac{1}{2}(A_1^2 + A_2^2)\right]\right\} \tag{4}$$

of spectrum analysis, for example, there is no mention of any model or any systematic signal at all. From the standpoint of probability theory, the class of problems for which the Blackman-Tukey method is appropriate has never been defined. In the problems we are considering, specification of a definite model (i.e. stating just what prior information we have about the phenomenon being observed) is essential; the information we can extract from the data depends crucially on which model we analyze.

In the following section we consider the simplest nontrivial model and analyze it in some depth to show some elementary but important points of principle in the technique of using probability theory with nuisance parameters and "uninformative" priors.

II. SINGLE STATIONARY SINUSOID PLUS NOISE.

We begin the analysis by constructing the direct probability. We think of this as the likelihood of the parameters, because it is the dependence of the likelihood function on the model parameters which concerns us here. The time series $y(t)$ we are considering is postulated to contain a single stationary harmonic signal $f(t)$ plus noise $e(t)$. The basic model is always: we have recorded a discrete data set $D = \{d_1, \cdots, d_N\}$; sampled from $y(t)$ at discrete times $\{t_1, \cdots, t_N\}$; with a model equation

$$d_i = y(t_i) = f(t_i) + e_i, \qquad (1 \leq i \leq N).$$

Different models correspond to different choices of the signal $f(t)$. We repeat the analysis originally done by Jaynes [2] using a different, but equivalent, set of model functions. We repeat this analysis for two reasons: first, by using a different formulation of the problem we can see how to generalize to multiple frequencies and more complex models; and second, to introduce a different prior probability for the amplitudes. This different prior simplifies the calculation but has almost no effect on the final result. The model we are considering in this section is

$$f(t) = A_1 \cos(\omega t) + A_2 \sin(\omega t)$$

which has three parameters (A_1, A_2, ω) that may be estimated from the data. The model used by Jaynes [2] was the same, but expressed in polar coordinates:

$$f(t) = A \cos(\omega t + \theta)$$

$$A = \sqrt{A_1^2 + A_2^2}$$

$$\tan\theta = -\frac{A_2}{A_1}$$

$$dA_1 dA_2 d\omega = A dA d\theta d\omega.$$

It is the factor A in the volume elements which is treated differently in the two calculations. Jaynes used a prior probability that initially considered equal intervals of A and θ to be equally likely, while we shall use a prior that initially considers equal intervals of A_1 and A_2 to be equally likely.

Of course, neither choice fully expresses all the prior knowledge we are likely to have in a real problem. This means that the results we find are conservative, and in a case where we have quite specific prior information about the parameters, we would be able to do somewhat better than in the following calculation. However, the differences arising from

completeness from the standpoint of the expert. Here we calculate a number of expectation values including the estimated amplitude of the signal, the variance of the data, and the power spectral density.

In Section V, we specialize the discussion to spectral estimates. In particular we discuss the estimation of multiple harmonic frequencies and their power spectra. We will then generalize the frequency and spectrum estimation problem to frequencies and spectra which are not stationary.

In Section VI, we apply the theory to a number of real time series including Wolf's relative sunspot numbers, some NMR data containing multiple close frequencies with decay, and to an economic time series which has a large trend. These analyses will give the reader a better feel for the types of applications and complex phenomena which can be investigated easily using Bayesian techniques.

The basic reasoning used in this work will be a straightforward application of Bayes' theorem: denoting by $P(A|B)$ the conditional probability that proposition A is true, given that proposition B is true, Bayes' theorem is

$$P(H|DI) = P(H|I)\frac{P(D|HI)}{P(D|I)}. \tag{1}$$

It is nothing but the probabilistic statement of an almost trivial fact: Aristotelian logic is commutative. That is, the propositions:

$$HD = \text{"Both } H \text{ and } D \text{ are true"}$$

$$DH = \text{"Both } D \text{ and } H \text{ are true"}$$

say the same thing, so they must have the same truth value in logic and the same probability, whatever our information about them. In the product rule of probability theory, we may then interchange H and D:

$$P(HD|I) = P(H|DI)P(D|I) = P(H|I)P(D|HI)$$

which is Bayes' theorem. In our problems, H is any hypothesis to be tested, D is the data, and I is the prior information. In the terminology of current statistical literature, $P(H|DI)$ is called the posterior probability of the hypothesis, given the data and the prior information. This is what we would like to compute for several different hypotheses concerning what systematic "signal" is present in our data. Bayes' theorem tells us that to compute it we must have three terms: $P(H|I)$ is the prior probability of the hypothesis (given only our prior information), $P(D|I)$ is the prior probability of the data (this term will always be absorbed into a normalization constant and will not change the distribution), and $P(D|HI)$ is called the direct probability of the of the data, given the hypothesis and the prior information. The direct probability is called the "sampling distribution" when the hypothesis is held constant and one considers different sets of data, and it is called the "likelihood function" when the data are held constant and one varies the hypothesis. Often, a prior probability distribution is called simply a "prior."

In a specific Bayesian probability calculation, we need to "define our model;" i.e. to enumerate the set $(H_1, H_2, \ldots)$ of hypotheses concerning the systematic signal that is to be tested by the calculation. A serious weakness of all Fourier transform methods is that they do not consider this aspect of the problem. In the widely used Blackman-Tukey [3] method

I. INTRODUCTION.

Experiments are performed in three general steps: first, the experiment must be designed; second, the data must be gathered; and third, the data must be analyzed. These three steps are highly idealized and no clear boundary exists between them. The problem of analyzing the data is one that should be faced early in the design phase. Gathering the data in such a way as to learn the most about a model is what doing an experiment is all about. It will do an experimenter little good to obtain a set of data that does not bear directly on the model to be tested.

In many experiments it is essential that one does the best possible job in analyzing the data. This could be true because no more data can be obtained, or one is trying to discover a very small effect. Furthermore, thanks to modern computers, sophisticated data analysis is far less costly than data acquisition, so there is no excuse for not doing the best job of analysis that we can. Unfortunately, the theory of optimum data analysis, which takes into account not only the raw data but also the prior knowledge that one has to supplement the data, is almost nonexistent. We hope to show the advantage of such a theory by developing a little of it, and applying the results to some real data.

In Section I we outline the calculation procedure used in this paper. The spectrum estimation problem is approached using probability theory and Bayes' theorem to remove the nuisance parameters.

In Section II, we analyze a time series which contains a single stationary harmonic signal plus noise, because it contains most of the points of principle that must be faced in the more general problem. In particular, we derive the probability that a signal of frequency ω is present, regardless of its amplitude, phase, and the variance of the noise. An example is given of numerical analysis of real data illustrating these principles.

In Section III, we discuss the types of model equations used, introduce the concept of an orthonormal model, and derive a transformation which will take any nonorthonormal model into an orthonormal model. Using these orthonormal models, we then generalize the simple harmonic analysis to arbitrary model equations and discuss a number of surprising features to illustrate the power and generality of the method.

In Section IV, we collect technical discussions of several side issues that are necessary for

in *Maximum-Entropy and Bayesian Methods in Science and Engineering,* **1,** *pp. 75-145,*
G. J. Erickson and C. R. Smith *Eds.,* Kluwer Academic Publishers,
Dordrecht the Netherlands, 1988.

Excerpts from
Bayesian Spectrum Analysis and Parameter Estimation

G. Larry Bretthorst

Department of Physics
Washington University
St. Louis, MO. 63130

ABSTRACT

Bayesian spectrum analysis is still in its infancy. It was born when E. T. Jaynes derived the periodogram [2] as a sufficient statistic for determining the spectrum of a time sampled data set containing a single stationary frequency. Here we extend that analysis and explicitly calculate the joint posterior probability that multiple frequencies are present, independent of their amplitude and phase, and the noise level. This is then generalized to include other parameters such as decay and chirp. Results are given for computer simulated data and for real data ranging from magnetic resonance to astronomy to economic cycles. We find substantial improvements in resolution over Fourier transform methods.

TABLE OF CONTENTS