BAYESIAN MODEL SELECTION: EXAMPLES RELEVANT TO NMR

G. LARRY BRETTHORST
Department of Chemistry
Campus Box 1134
Washington University
1 Brookings Drive
St. Louis, MO 63130

Abstract. The model selection problem is one of the most basic problems in data analysis. Given a data set one can always expand the model almost indefinitely. How does one pick a model which explains the data, but does not contain spurious features relating to the noise? Here we present the results of a Bayesian model selection calculation started in [1] and then extended in [2], and show that the Bayesian answer to this question is essentially a quantitative statement of Occams razor: When two models fit the evidence in the data equally well, choose the simpler model.

Introduction

When analyzing the results of an experiment it is not always known which model function applies. We need a way to choose between several possible models. This is easily done using Bayes' theorem. The first step in answering this question is to enumerate the possible models. Suppose we have a set S of s possible models with model functions $\{f_1, \cdots, f_s\}$. We are hardly ever sure that the "true" model is actually contained in this set. Indeed, the "set of all possible models" is not only infinite, but it is also quite undefined. It is not even clear what one could mean by a "true" model. Both questions may take us into an area more like theology than science.

The only questions we seek to examine are the ones that are answerable because they are mathematically well-posed. Such questions are of the form: "Given a specified set S_s of possible models $\{f_1, \cdots, f_s\}$ and looking only within that set, which model is most probable in view of all the data and prior information, and how strongly is it supported relative to the alternatives in that set?" Bayesian analysis can give a definite answer to such a question – see [2], [3], [4]. Here we give the results of the calculation done in [2] and present two numerical examples of its use.

The Relative Probability of Model f_j

Given a set S_s of models $\{f_1, \cdots, f_s\}$ and looking only within that set, which model best accounts for the data? We will take

$$f_j(t) = \sum_{k=1}^{m} A_k H_k(t, \{\omega\}) \tag{1}$$

as our model, where H_k are the orthonormal model functions defined in [1]. The subscript "j" refers to the jth member of the set S_s of models $\{f_1, \cdots, f_s\}$ with the understanding that the amplitudes $\{A\}$, the nonlinear $\{\omega\}$ parameters, the total number of model functions m, the total number of nonlinear parameters r, and the model functions $H_k(t, \{\omega\})$ are different for every f_j.

The use of the orthogonal models does not change the generality of the calculation because any arbitrary model may be transformed into an orthogonal model. If we have a nonorthogonal model

$$f_j(t) = \sum_{k=1}^{m} B_k G_k(t, \{\omega\}), \tag{2}$$

where G_k is the model function (for example a sine or Bessel function) and B_k is its amplitude, then we may transform this model into an orthogonal model, Eq. (1), as follows: compute the interaction matrix

$$g_{kl} = \sum_{i=1}^{N} G_k(t_i) G_l(t_i) \tag{3}$$

and from the kth eigenvalue λ_k of the interaction matrix, Eq. (3), and the lth component of the kth eigenvector e_{kl} compute the orthogonal model functions H_k given by

$$H_k(t) = \frac{1}{\sqrt{\lambda_k}} \sum_{l=1}^{m} e_{kl} G_l(t).$$

The orthogonal amplitudes A_k may be computed from a linear combination of the B_l:

$$A_k = \sqrt{\lambda_k} \sum_{l=1}^{m} B_l e_{kl}.$$

From Bayes' theorem we may compute the posterior probability of model f_j:

$$P(f_j|D, I) = \frac{P(f_j|I) P(D|f_j, I)}{P(D|I)} \quad \text{and} \quad P(D|I) = \sum_{j=1}^{s} P(f_j|I) P(D|f_j, I). \tag{4}$$

We will assume, for now, that the variance of the noise σ^2 is known and derive $P(f_j|\sigma, D, I)$; then at the end of the calculation, if σ is not known we will remove it. Thus symbolically the heart of the problem is to compute

$$P(D|\sigma, f_j, I) = \int d\{A\} d\{\omega\} P(\{A\}, \{\omega\}|I) P(D|\{A\}, \{\omega\}, \sigma, f_j, I). \tag{5}$$

When we solved this problem, there were several places where prior information had to be incorporated: first, when we assigned a noise prior; second, when we removed the amplitudes; third, when we removed the nonlinear $\{\omega\}$ parameters; and fourth, when we removed the variance of the noise. When we assigned a noise prior we assumed the second moment of the noise was given and using maximum entropy arrived at a Gaussian prior as the least informative prior probability for the noise for a given second moment. The amplitudes are location parameters, and when we removed the amplitudes, we used a Gaussian, centered at zero, whose variance δ^2 expressed how strongly we believed the amplitudes to be near zero. From the form of the model, we do not know if the nonlinear $\{\omega\}$ parameters were location parameters or scale parameters. However, when we did this calculation we made a local Gaussian approximation to the posterior probability (i.e. we assumed the data determine the parameters well and we assumed the data determine the parameters much more precisely than the prior information). In this approximation the $\{\omega\}$ parameters are location parameters; thus we used a Gaussian centered at zero with variance γ^2 to represent the prior information about the nonlinear parameters. Last, if the three variances σ^2, δ^2, and γ^2 were not known, we removed them using a normalized Jeffreys prior. The normalization constant for the Jeffreys prior expresses the prior information in the form of a permissible range of values for the variances. This range of values appears in the problem as a natural logarithm of the upper limit divided by the lower limit. We designated this ratio as R_σ for the variance of the noise and similarly for γ and δ.

If the three variances are actually known, then the direct probability of the data is approximately given by

$$\begin{aligned} P(D|\gamma,\delta,\sigma,f_j,I) &\approx (2\pi\delta^2)^{-\frac{m}{2}} \exp\left\{-\frac{m\overline{h^2}(\{\hat{\omega}\})}{2\delta^2}\right\} \\ &\times (2\pi\gamma^2)^{-\frac{r}{2}} \exp\left\{-\frac{r\overline{\omega^2}}{2\gamma^2}\right\} v_1^{-\frac{1}{2}} \cdots v_r^{-\frac{1}{2}} \\ &\times (2\pi\sigma^2)^{-\frac{N-m-r}{2}} \exp\left\{-\frac{N\overline{d^2} - m\overline{h^2}(\{\hat{\omega}\})}{2\sigma^2}\right\} \end{aligned} \tag{6}$$

where $\overline{\omega^2}$ is the mean-square estimated $\{\omega\}$ parameter

$$\overline{\omega^2} = (1/r)\sum_{k=1}^{r} \hat{\omega}_k^2,$$

$\overline{h^2}(\{\hat{\omega}\})$ is the mean-square value of the h_k functions

$$\overline{h^2}(\{\hat{\omega}\}) \equiv \frac{1}{m}\sum_{k=1}^{m} h_k^2\Big|_{\{\hat{\omega}\}},$$

h_k is the projection of the data onto the orthonormal model functions H_k

$$h_k \equiv \sum_{i=1}^{N} d(t_i) H_k(t_i, \{\hat{\omega}\}),$$

$\{\hat{\omega}\}$ is the location of the maximum posterior probability digitized as $\{\hat{\omega}_1, \cdots, \hat{\omega}_r\}$, and v_k is one of the eigenvalues of the matrix

$$b_{jk} \equiv -\frac{m}{2} \frac{\partial^2 \overline{h^2}}{\partial \omega_j \partial \omega_k}\Big|_{\{\hat{\omega}\}}.$$

If the three variances σ^2, δ^2, and γ^2 are not known, then they may be removed using a normalized Jeffreys prior. The global likelihood of the data is then approximately

$$\begin{aligned} P(D|f_j, I) \;\approx\; & \frac{\Gamma(m/2)}{2\log(R_\delta)} \left[\frac{m\overline{h^2}(\{\hat{\omega}\})}{2}\right]^{-\frac{m}{2}} \frac{\Gamma(r/2)}{2\log(R_\gamma)} \left[\frac{r\overline{\omega^2}}{2}\right]^{-\frac{r}{2}} v_1^{-\frac{1}{2}} \cdots v_r^{-\frac{1}{2}} \\ \times \; & \frac{\Gamma([N-m-r]/2)}{2\log(R_\sigma)} \left[\frac{N\overline{d^2} - m\overline{h^2}(\{\omega\})}{2}\right]^{\frac{m+r-N}{2}} \end{aligned} \tag{7}$$

where $\Gamma(x)$ is a gamma function of argument x. The three factors involved in normalizing the Jeffreys priors (R_σ, R_γ, R_δ) appear in every model; they always cancel as long as we are dealing with models having all three types of parameters. However, as soon as we try to compare a model involving two types of parameters to a model involving three types of parameters (e.g. a regression model to a nonlinear model) they no longer cancel: the prior ranges become important. One must think carefully about just what prior information one actually has about σ, γ, and δ and use that information to set the prior ranges.

Example – Multiple Exponential Decays

Two major problems in NMR are: determining the characteristic decay time of a signal (the so-called T_2 time), and determining how many resonances (frequencies) are in a free induction decay. We will give two examples of the use of Eq. (4), one on simulated multiple decaying exponential data, and one on simulated multiple nonstationary frequencies.

The data in T_2 experiments are a time series (typically nonuniformly sampled) which decays away exponentially. The problem is to determine how many exponentials are in the data, and to estimate the decay rates and their amplitudes. This is frequently done in one of two ways: least squares or curve stripping. In least squares, the experimenter will fit a model having one, two, three, etc. decaying exponentials to his data and then stop the process when (1) the model looks like the data, (2) the parameters are not physically meaningful, or (3) the parameters are not statistically

significant. In curve stripping, the data are plotted on a semi-logarithmic plot. On this plot a single exponential will appear as a straight line. The experimenter looks for how many straight lines are necessary to represent the data. The stopping criterion is typically set by the human eye. Neither least squares nor curve stripping has any theoretically justifiable stopping criterion; rather, they are intuitive procedures that give reasonable results.

Let us apply the procedures we have developed and see what probability theory can do on this problem. We take as our data, Fig. 1 (A), derived from

$$d(t_i) = 100\mathrm{e}^{-0.05t_i} + 50\mathrm{e}^{-0.02t_i} + n(0,1),$$

where $n(0,1)$ is a Gaussian random number with zero mean and standard deviation one. In this example we take $t_i = \{0, 0.5, 1, \cdots 100\}$, $N = 201$, with signal-to-noise ratio of approximately 60. We have displayed the data in Fig. 1 (A) and a semi-logarithmic plot of the data in Fig. 1 (B). We see little evidence of two decaying exponentials in these data.

To apply the procedures given here, we specify the set of models to be examined. We take

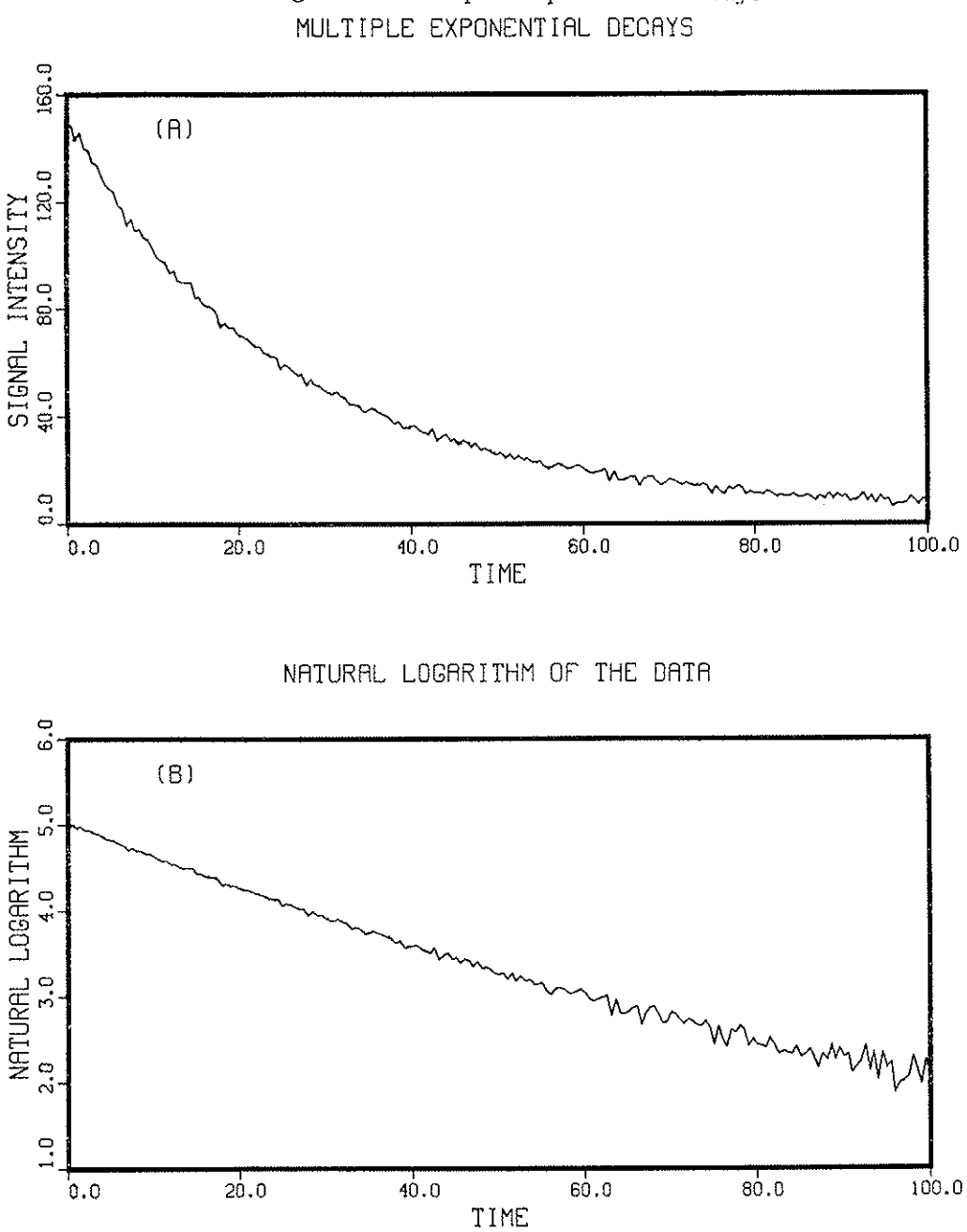
$$f_j(t) = \sum_{k=1}^{j} B_k \mathrm{e}^{-\alpha_k t} \quad (j = 1, 2, \cdots)$$

as the nonorthogonal models, Eq. (2). The question we would like to answer is: "What value of j is necessary to account for all systematic effects in the data?" To answer this question we compute the orthogonal model, Eq. (1), and from the orthogonal model we compute the global likelihood of the data, Eq. (7). From the global likelihood we compute the posterior probability, Eq. (4). To compute the posterior probability we must assign a prior probability, $P(j|D, I)$, for the number of exponential components in the data. Having little prior information about this, we use a uniform prior probability.

The results of this calculation are displayed in Fig. 2. There are six plots in this figure occurring in pairs. We have plotted the data (dotted line) in panels (A), (C), and (E). Included on these panels as the solid line is a one exponential model (A), a two exponential model (C), and a three exponential model (E). Panels (B), (D), and (F) contain a plot of the residuals (i.e. the difference between the data and the model) for the different models. The residuals for the one exponential model, Fig. 2 (B), show a systematic effect, and the posterior probability of this model is zero to eight decimal places. The residuals for the two exponential model, Fig. 2 (D), look like white noise, and the posterior probability of this model is 0.9984. The residuals for the three exponential model, Fig. 2 (F), show no noticeable improvement compared to the two exponential model. The posterior probability of this model is 0.0016. Given the three choices, Bayesian probability theory has correctly determined the number of exponentials in the data.

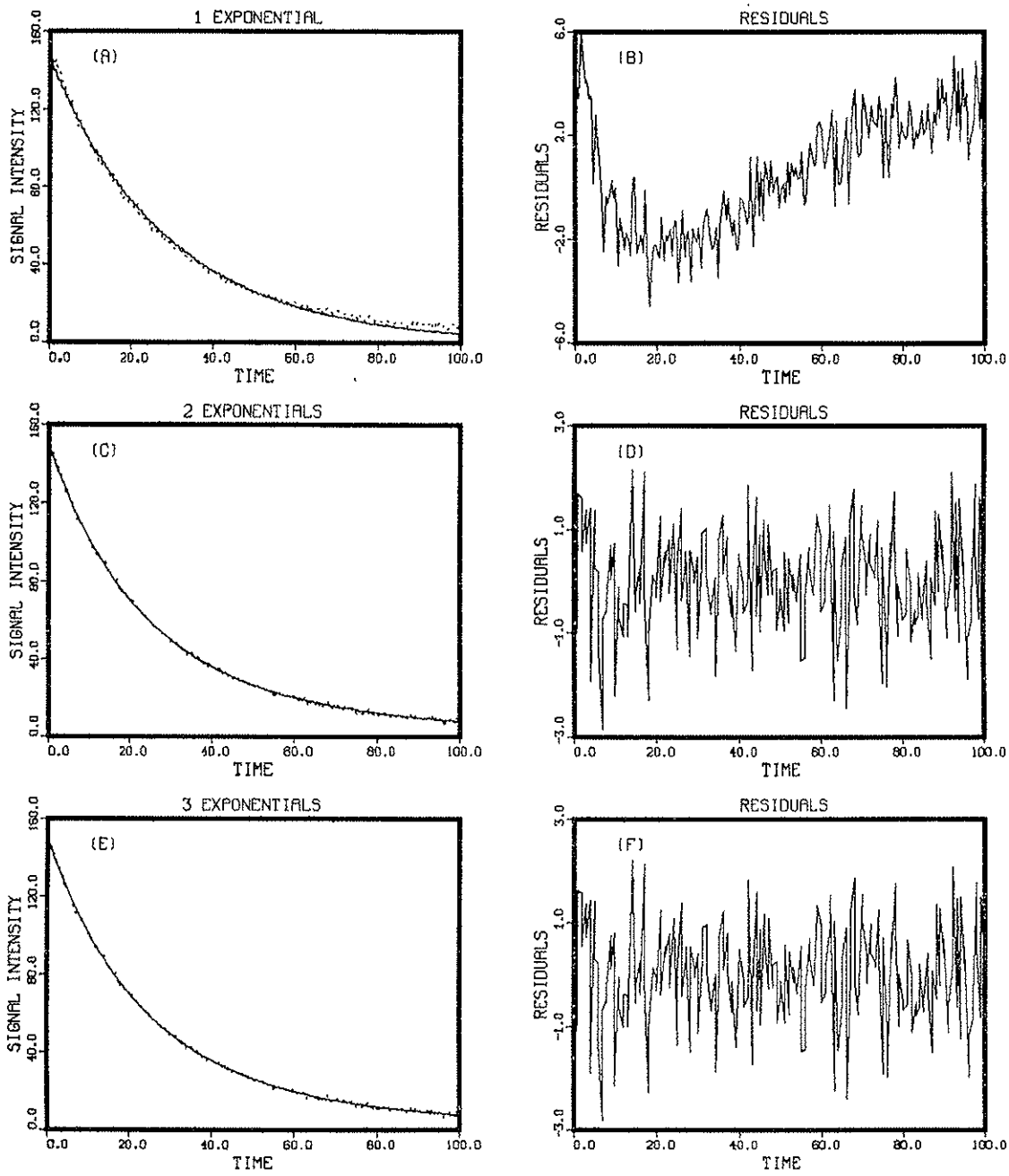
The choice between the two and three exponential model is very interesting. The residuals did not improve substantially for the three exponential model, and Bayes

Figure 1: Multiple Exponential Decays



The computer simulated data (A) contain two decaying exponentials – see text for the details. There are $N = 201$ data values with signal-to-noise ratio of approximately 60. Panel (B) is a plot of the natural logarithm of the data. From the plot there is little evidence for the second decaying exponential.

Figure 2: Multiple Exponential Decays



The data, (A) dotted line, were fit with a one exponential model, (A) solid line. The residuals are shown in (B). There is a systematic effect in the residuals. The probability of this mode is zero to 8 places. The data were then fit with a two exponential model, (C) dotted line. The residuals are shown in (D). The posterior probability of this model is 0.9984. A three exponential model is shown in (E) and the residuals in (F). There is no noticeable improvement. The posterior probability of this model is 0.0016.

theorem then tells us to choose the simpler model. The prior information is effectively doing this. When we did this calculation, we put in the fact that the amplitudes and decay rates should be taken to be zero unless the data clearly indicate otherwise. In the three exponential model, the data indicate the parameters should be nonzero. However, the prior probability of the model depends on the mean-square amplitude and the mean-square decay rate. When these factors are estimated to be nonzero, their prior probability is low. The fit for the three exponential model did not improve substantially: the probability of the data for the two and the three exponential model is about the same. So the posterior probability of the three exponential model (which is related to the product of these two) is low because the prior probability is low. Bayesian model selection is essentially a quantitative statement of Occams razor: when two models fit the evidence equally well, prefer the simpler model.

Example – Multiple Nonstationary Frequencies

Often an experimenter is faced with a data set that looks like Fig. 3 (A). The problem is to determine how many resonances are present. If the resonances were stationary, the experimenter could sample the signal longer. The discrete Fourier transform would then resolve the resonances. Unfortunately, the resonances are nonstationary, i.e. they decay away with time. Taking data for a longer period of time samples the noise, not the resonances, and no improvement is found. The only recourse is to get the information from the data available. In fact, probability theory indicates there is one thing the experimenter can do: sample the data faster [2], thus obtaining more data in the region where the resonances are big. This gives a $\sqrt{\Delta T}$ improvement in the estimates, where ΔT is the sampling time. But, if the experimenter uses the discrete Fourier transform as his analysis tool, instead of probability theory, this procedure will improve the line shape, but it will not help separate the resonances.

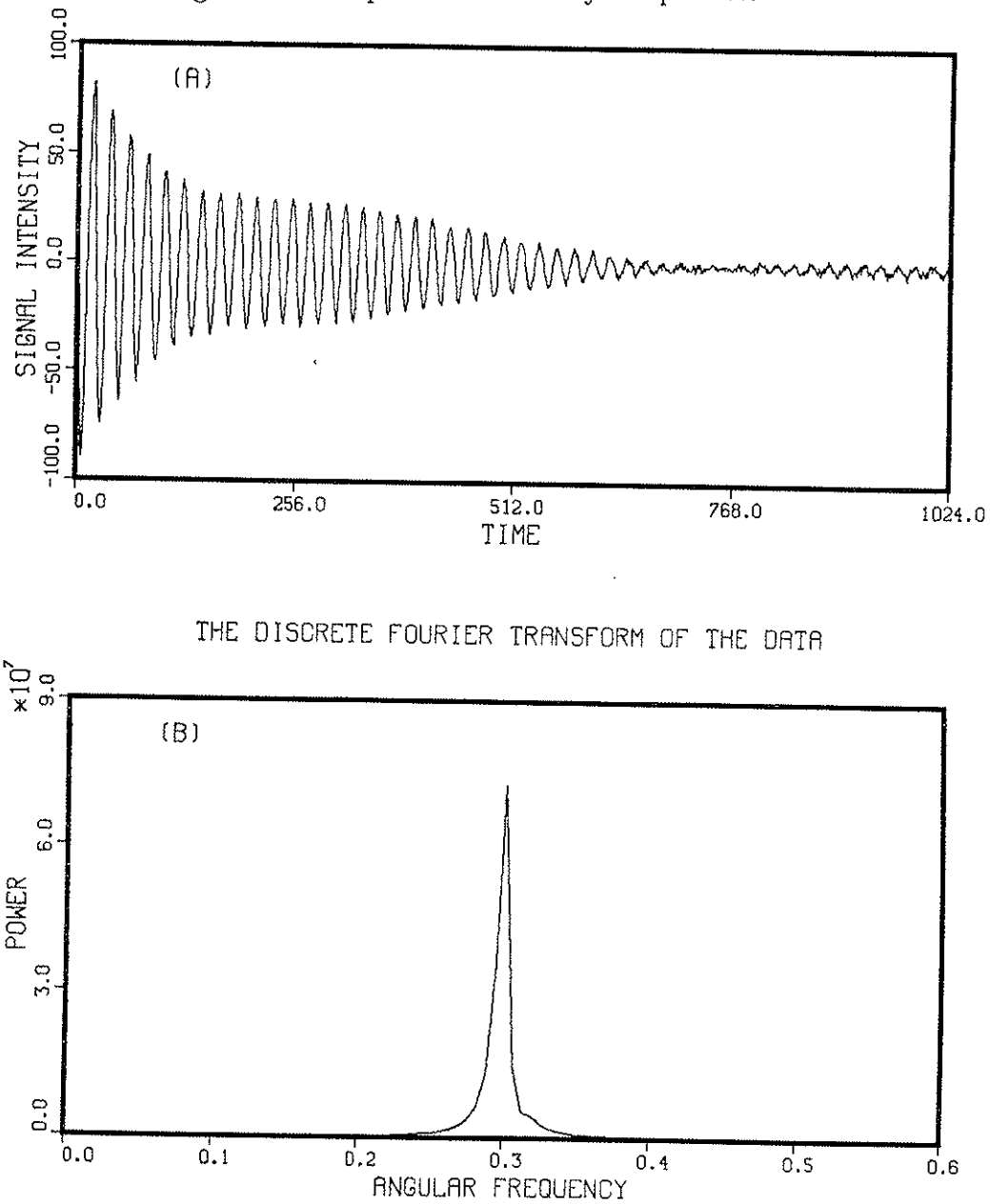
Let us see what probability theory can do on the multiple nonstationary frequency problem. We generated the data, Fig. 3 (A), in such a way that the discrete Fourier transform, Fig. 3 (B), has only a single peak. We then apply the results of the calculation to see if probability theory can determine the number of frequencies present. In this example, we generated the data from

$$d(t_i) = 100\cos(0.3t_i + 1)\mathrm{e}^{-.005t_i} + 25\cos(0.31t_i + 3)\mathrm{e}^{-.003t_i} + n(0,1),$$

with $N = 1024$, signal-to-noise ratio of approximately 20, and $t_i = \{0, 1, \cdots, 1023\}$. The discrete Fourier transform, Fig. 3 (B) has one peak near 0.3. There is no evidence in a discrete Fourier transform for the second frequency. However, we can look at the data and see the beats: the human eye is better at determining the presence of multiple close frequencies than a discrete Fourier transform.

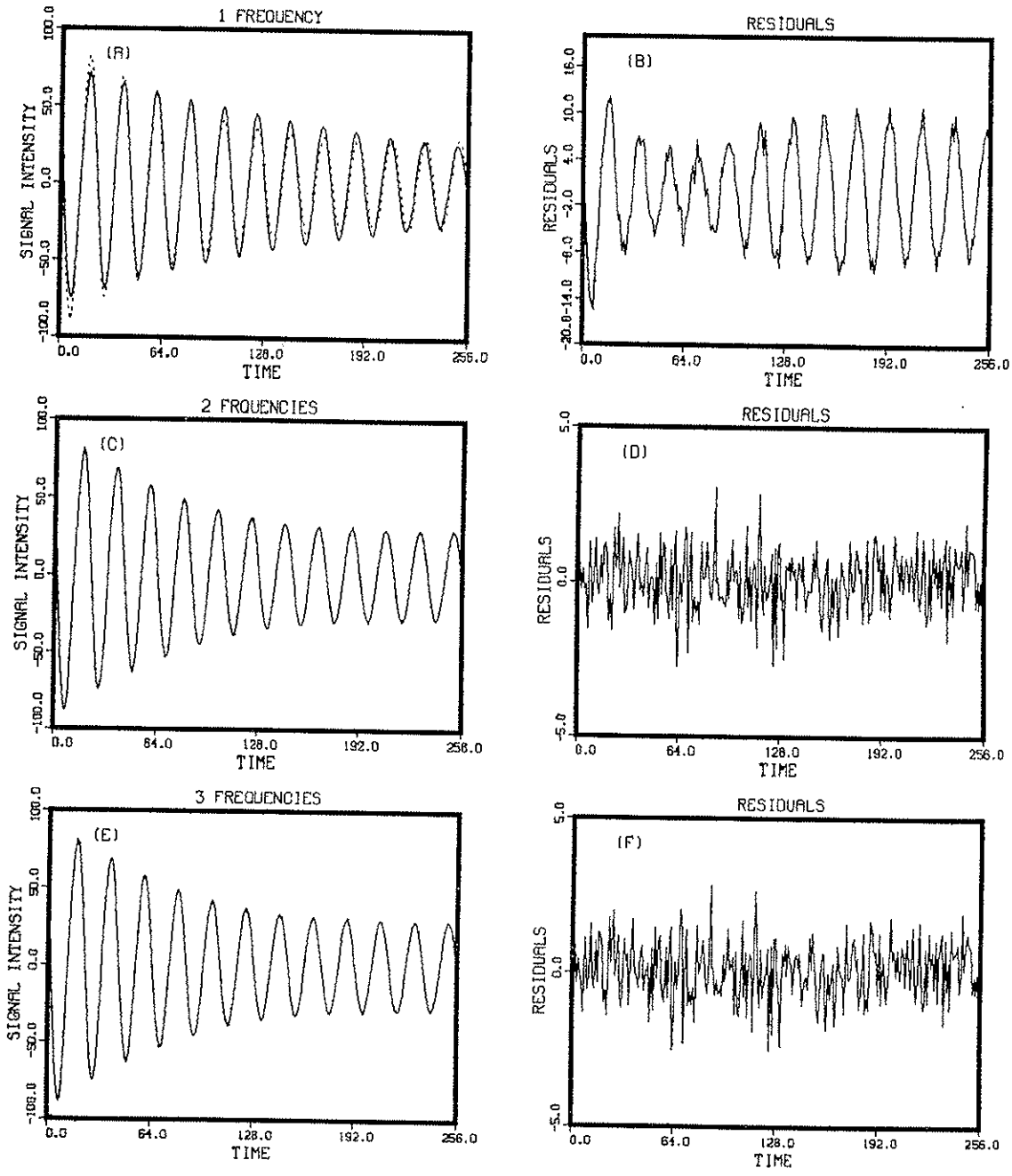
To compute the posterior probability of the model, we must state what set of

Figure 3: Multiple Nonstationary Frequencies



The computer simulated data in (A) contain two nonstationary frequencies. There are $N = 1000$ data values with signal-to-noise ratio of approximately 20. The discrete Fourier transform (B) shows only a single peak.

Figure 4: Multiple Nonstationary Frequencies



The data, (A) dotted line, were fit to a one frequency with decay model, (A) solid line. The residuals are shown in (B). The posterior probability of this model is zero to 690 places! We then fit a two frequency with decay model, (C) solid line. The residual are shown in (D). The posterior probability of this model is 1 to eight places. We then fit a three frequency model, (E) solid line. The residuals are shown in (F). The posterior probability of this model is zero to eight places.

models is to be examined. Here we used

$$f_j(t) = \sum_{k=1}^{j} (B_k \cos \omega t + B_{k+j} \sin \omega t)\, e^{-\alpha_k t}, \qquad (j = 1, 2, \cdots) \tag{8}$$

as the nonorthogonal models. The question we would like to ask is: "What value of j is needed to adequately describe the data?" To do this calculation we again fit a model containing one, two, three, etc. frequencies until the posterior probability of the data had a well defined peak.

We again displaye the results of this calculation as six plots, Fig. 4. The data (dotted line) are shown in (A), (C), and (E). We have included the model as a solid line; a one frequency model is shown in (A), a two frequency model in (C), and a three frequency model in (E). The residuals for the three models are shown in (B), (D), and (F) respectively.

We begin by fitting a one frequency with decay model. We compute the nonorthogonal model, Eq. (8), and then the orthogonal model, Eq. (1). From the orthogonal model we compute the global likelihood of the data, Eq. (7), and last we compute the posterior probability, Eq. (4), using a uniform prior on the models. The posterior probability of the one frequency with decay is zero to 690 decimal places, strong evidence indeed! Note the logarithm of the posterior probability increases like the number of data values. Here we have $N = 1024$ data values. Additionally, for this model, each sampled value significantly misfits the data: the data are very improbable in view of this model; the posterior probability of the model is extremely low.

We then fit a two frequency model to the data, Fig. 4 (C) – the residuals are shown in (D). Now the model and the data are essentially identical: the residuals (D) look like white noise. We then compute the posterior probability of this model and find it to be 1 to thirteen decimal places. Of course, all we knew at this point was that the two frequency model was strongly preferred to the one frequency model, so we proceeded to the three frequency model.

We then computed the three frequency model. The data and the model are displayed in Fig. 4 (E), and the residuals are shown in Fig. 4 (F). The model does not fit the data any better than one would expect from fitting the noise. The posterior probability of this model is zero to thirteen places. In this example not only does probability theory find the correct number of resonances, but also the evidence in these data is overwhelmingly in favor of the two frequency with decay model.

Conclusions

We have demonstrated in these two examples that Bayesian probability theory is capable of giving a quantitative interpretation to Occam's razor. These procedures are readily implemented and work well under conditions where more standard procedures fail. The multiple nonstationary frequency example illustrated that the human eye

is a better tool for determining the presence of multiple resonances than the discrete fourier transform. In data where the human eye outperforms the discrete Fourier transform, the Bayesian calculation gives overwhelming evidence for the frequencies. In the multiple decaying exponential example the human eye is no better than the more traditional techniques. However, the Bayesian analysis works under conditions where the more traditional tests fail and gives strong evidence in support of the correct hypothesis.

Acknowledgments

This work was partially supported by NIH grant GM-30331, J. J. H. Ackerman principal investigator, and by a gift from Nabisco brands. The encouragement of Dr. J. J. H. Ackerman and Professor E. T. Jaynes is greatly appreciated.

References

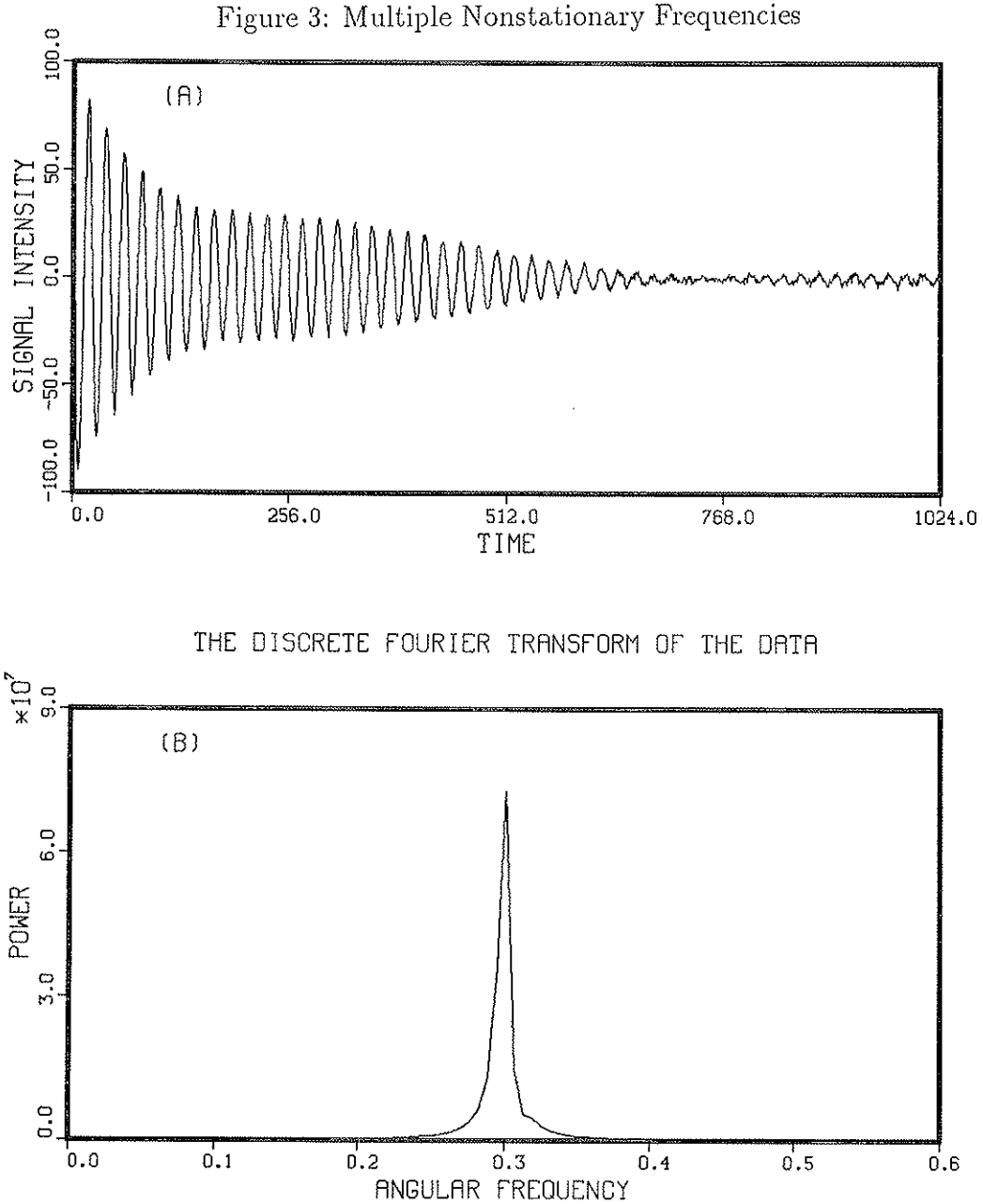
[1] Bretthorst G. L., (1987), Bayesian Spectrum Analysis and Parameter Estimation, Ph.D. thesis, Washington University, St. Louis, MO.; available from University Microfilms Inc., Ann Arbor, Mich.

[2] Bretthorst, G. L., (1988), Bayesian Spectrum Analysis and Parameter Estimation, in *Lecture Notes in Statistics,* Vol. 48, Springer-Verlag, New York, New York

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

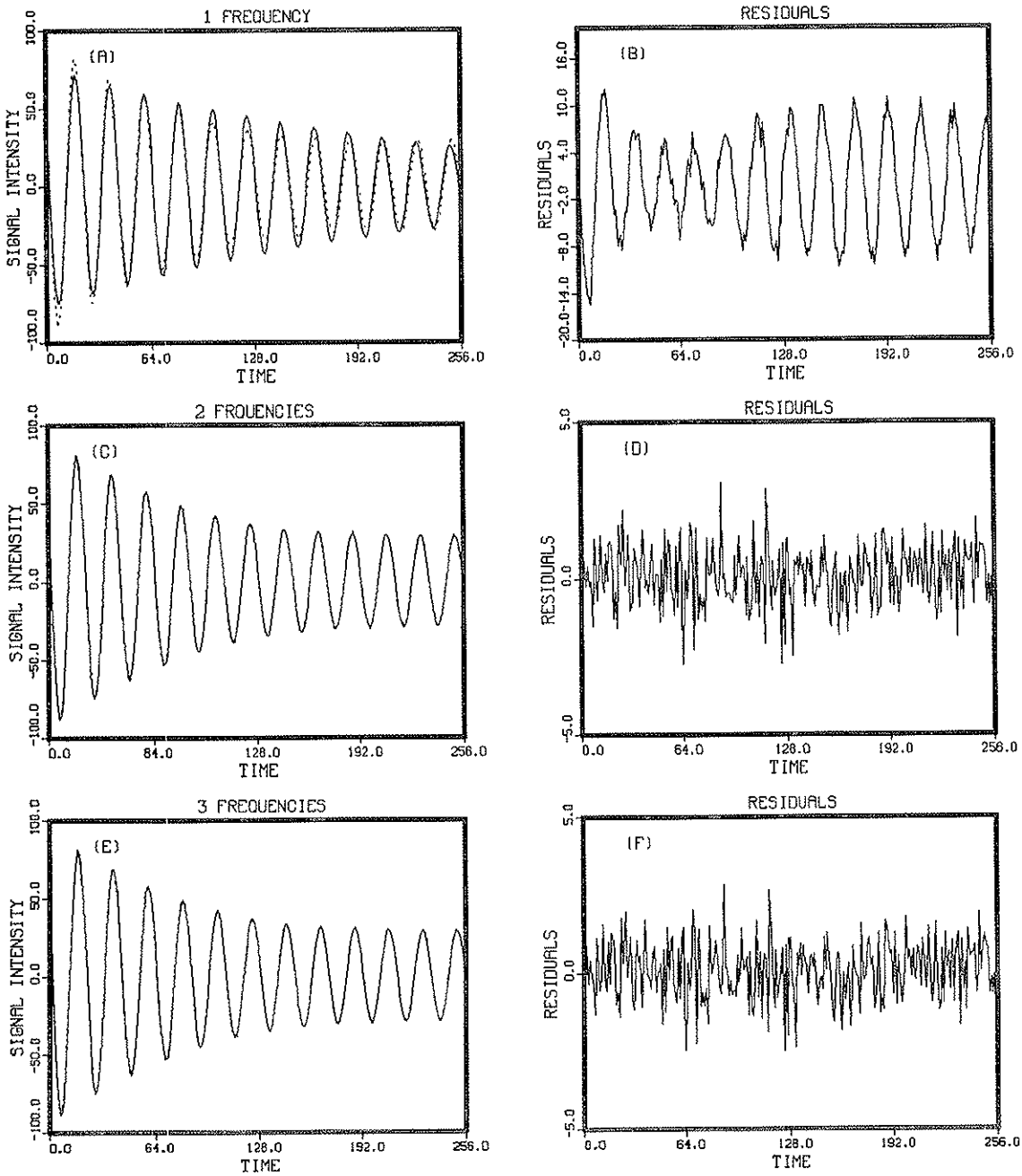
[4] Zellner, A., (1980), in Bayesian Statistics, J. M. Bernardo, ed., Valencia University Press, Valencia, Spain.

Figure 3: Multiple Nonstationary Frequencies



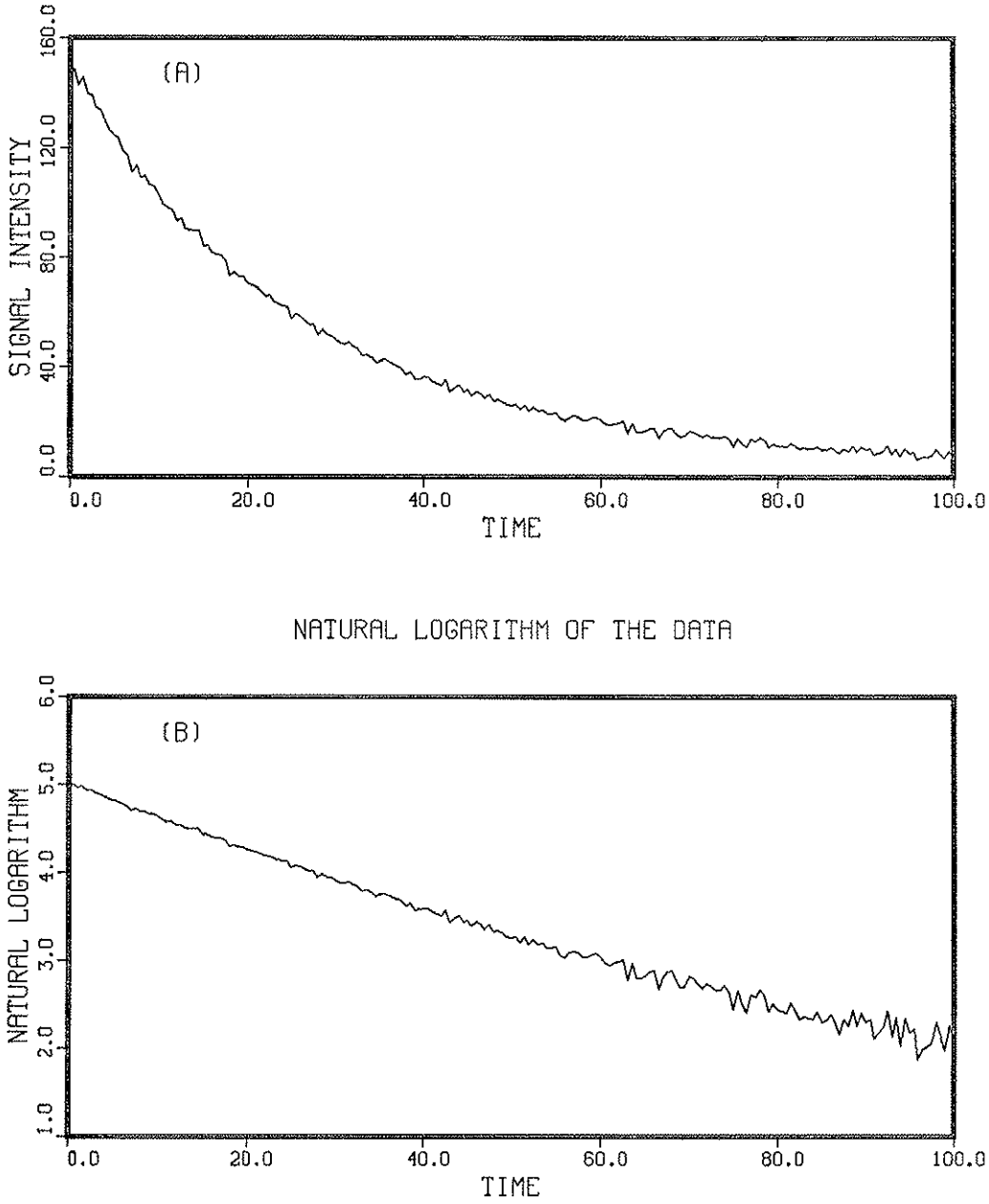
This computer simulated data (A) contain two nonstationary frequencies. There are $N = 1000$ data values with signal-to-noise ratio of approximately 20. The discrete Fourier transform (B) shows only a single peak.

Figure 4: Multiple Nonstationary Frequencies



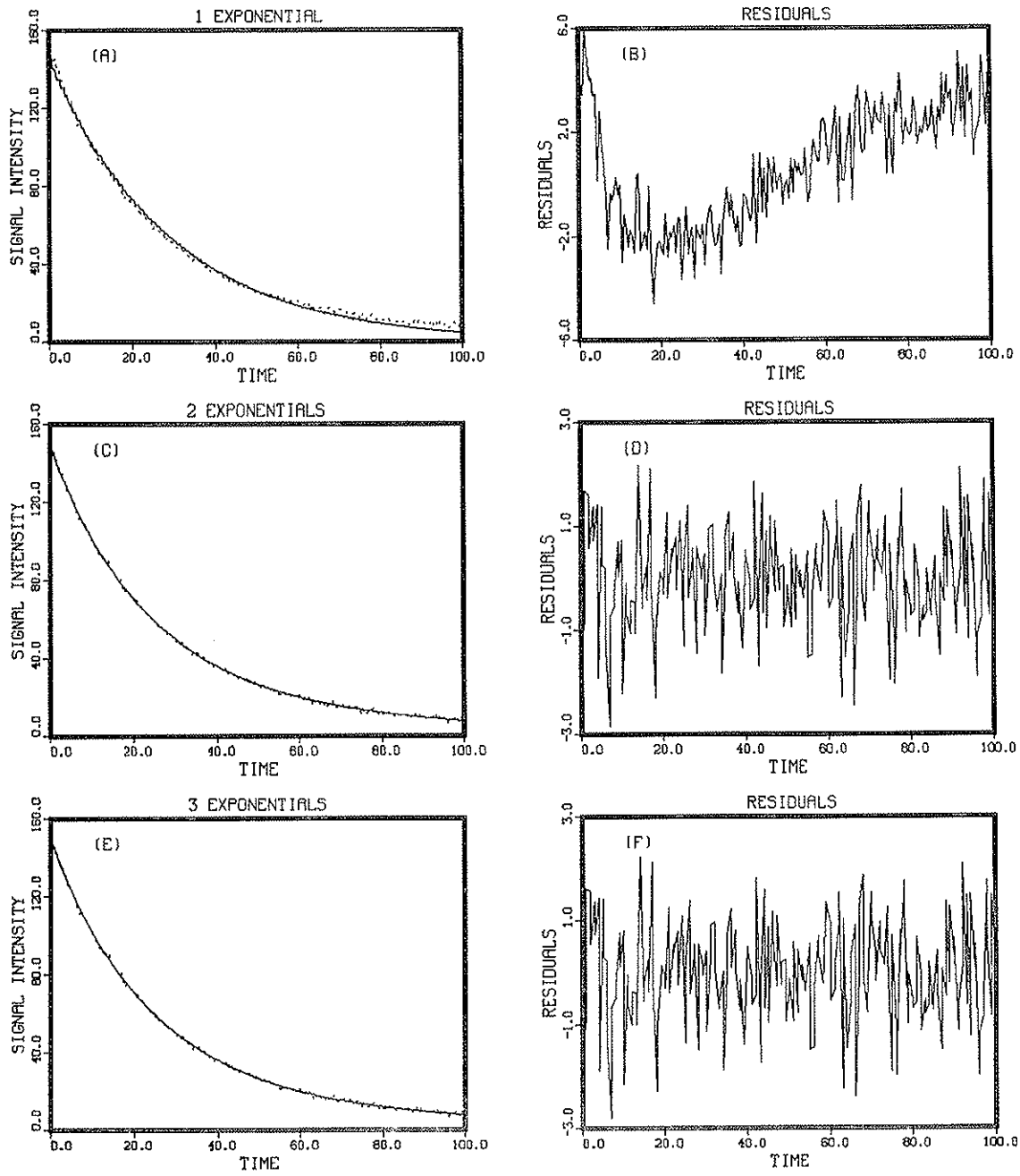
The data, (A) dotted line, were fit to a one frequency with decay model, (A) solid line. The residuals are shown in (B). The posterior probability of this model is zero to 690 places! We then fit a two frequency with decay model, (C) solid line. The residual are shown in (D). The posterior probability of this model is 1 to eight places. We then fit a three frequency model, (E) solid line. The residuals are shown in (F). The posterior probability of this model is zero to eight places.

Figure 1: Multiple Exponential Decays



This computer simulated data (A) contain two decaying exponentials, see text for the details. There are $N = 201$ data values with signal-to-noise ratio of approximately 60. Panel (B) is a plot of the natural logarithm of the data. From the plot there is little evidence for the second decaying exponential.

Figure 2: Multiple Exponential Decays



The data, (A) dotted line, were fit with a one exponential model, (A) solid line. The residuals are shown in (B). There is a systematic effect in the residuals. The probability of this modes is zero to 8 places. The data were then fit with a two exponential model, (C) dotted line. The residuals are shown in (D). The posterior probability of this model is 0.9984. A three exponential model is shown in (E) and the residuals in (F). There is no noticeable improvement. The posterior probability of this model is 0.0016.