Journal of Magnetic Resonance **88,** 552-570 (1990)

Bayesian Analysis. II. Signal Detection and Model Selection

G. LARRY BRETTHORST
Washington University,
Department of Chemistry
Campus Box 1134
1 Brookings Drive,
St. Louis, Missouri 63130-4899

ABSTRACT. In the preceding paper, Bayesian analysis was applied to the parameter estimation problem, given quadrature NMR data. Here Bayesian analysis is extended to the problem of selecting the model which is most probable in view of the data and all the prior information. In addition to the analytic calculation, two examples are given. The first example demonstrates how to use Bayesian probability theory to detect small signals in noise. The second example uses Bayesian probability theory to compute the probability of the number of decaying exponentials in simulated T_1 data. The Bayesian answer to this question is essentially a microcosm of the scientific method and a quantitative statement of Ockham's razor: theorize about possible models, compare these to experiment; and select the simplest model that "best" fits the data.

Introduction

The parameter estimation problem discussed in the preceding paper [1] tells how to estimate the parameters given a model; but it does not tell *how* to select that model, nor does it tell *when* the signal has been detected. For this, additional tools are required. For example, if the signal is known to be the sum of exponentially decaying sinusoids, it is not enough to answer the question "Given that there are n sinusoids, what is the best estimate of the frequencies and decay rates?" Answering this question is useful, but not sufficient. In most experimental situations one must also ask "What is the evidence for a signal?" and given that a signal has been detected "How many sinusoids are present?" This is easily done using Bayes' theorem and repeated applications of the procedures used in the previous paper.

The questions to be examined here are of the form: "Given a specified set S of s possible models $S \equiv \{f_1, \cdots, f_s\}$ and looking only within that set, which model is most probable in view of the data and all known 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 Refs. 2–5.

The probability of model f_j, conditional on the data D and the prior information I, is given by Bayes' theorem,

$$P(f_j|D,I) = \frac{P(f_j|I)P(D|f_j,I)}{P(D|I)} \qquad (1 \le j \le s), \tag{1}$$

where $P(f_j|D,I)$ is the posterior probability of model f_j given the data D and the prior information I. This is the term one wants to calculate for the set of models S. To calculate it, Bayes' theorem indicates that three terms must be computed: the first, $P(f_j|I)$, is the probability of model f_j given only the information I. This term represents one's state of knowledge about the models before obtaining the data D. The second term, $P(D|f_j,I)$, is the global likelihood of the data given model f_j and the prior information I. It represents how well the model fits the data. The third term,

$P(D|I)$, is the marginal probability of the data given only the prior information I,

$$P(D|I) = \sum_{k=1}^{s} P(f_k|I)P(D|f_k, I), \tag{2}$$

and is a normalization constant over all models.

Suppose the data $D \equiv \{d_1, \cdots, d_N\}$ have been sampled at discrete times $\{t_1, \cdots, t_N\}$; the times are not assumed to be uniform, nor unique, e.g., there could be repeated measurements (for example, as in quadrature). The only restriction is that the variance of the noise is assumed to be the same in the repeated data. If the true signal in the data is f_j then

$$d_i = f_j(t_i) + e(t_i), \tag{3}$$

where d_i is the data at time t_i, $e(t_i)$ represents the noise at time t_i, and $f_j(t_i)$ is the jth member of the set S at time t_i. In NMR, the signal f_j may typically be expressed as

$$f_j(t, \mathbf{\Theta}) = \sum_{k=1}^{m} B_k G_k(t, \mathbf{\Theta}), \tag{4}$$

where $G_k(t, \mathbf{\Theta})$ is one of m signal functions with amplitude B_k. There are a total of m amplitudes: $\mathbf{B} \equiv \{B_1, \cdots, B_m\}$, and $\mathbf{\Theta}$ is a set of r nonlinear parameters defined as $\mathbf{\Theta} \equiv \{\Theta_1, \cdots, \Theta_r\}$. If the signal functions are exponentially decaying sinusoids, then the linear $\mathbf{B}$ parameters are effectively the amplitudes and phases of the sinusoids, and the nonlinear $\mathbf{\Theta}$ parameters would be the frequencies and decay rate constants. The parameters $\mathbf{B}$ and $\mathbf{\Theta}$ are assumed completely different in every model f_j.

The global likelihood of the data $P(D|f_j, I)$ is obtained from the joint probability of the data, and the parameters $P(D, \mathbf{B}, \mathbf{\Theta}|f_j, I)$. To obtain the global likelihood of the data the dependence on amplitudes $\mathbf{B}$ and the nonlinear $\mathbf{\Theta}$ parameters must be removed. To remove the parameters from the problem, it is assumed that when the data were taken each of the parameters could take on only one value; i.e., each parameter is a constant through the run of data. However, it is not known which value was actually realized. The probability that a parameter had value a_j, where a_j is one member of a set of mutually exclusive values, $\{a_1, \cdots, a_n\}$, is the sum of the probabilities of the individual values. When the parameters are continuous the sums are replaced by integrals. Thus the global likelihood of the data is given by

$$P(D|f_j, I) = \int d\mathbf{B} d\mathbf{\Theta} P(D, \mathbf{B}, \mathbf{\Theta}|f_j, I). \tag{5}$$

In the preceding paper [1], only the amplitudes $\mathbf{B}$ were eliminated from consideration here both the amplitudes $\mathbf{B}$ and the nonlinear $\mathbf{\Theta}$ parameters must be eliminated. The product rule may be used to factor the joint density $P(D, \mathbf{B}, \mathbf{\Theta}|f_j, I)$ to obtain

$$P(D|f_j, I) = \int d\mathbf{\Theta} P(\mathbf{\Theta}|f_j, I) \left[\int d\mathbf{B} P(\mathbf{B}|\mathbf{\Theta}, f_j, I) P(D|\mathbf{B}, \mathbf{\Theta}, f_j, I) \right]. \tag{6}$$

The quantity in square brackets is the parameter estimation problem from the first paper [1] with four changes: (i) all numerical factors must be kept (they do not necessarily cancel when the distribution is normalized), (ii) all parameters are to be considered nuisance parameters, (iii) one must use fully normalized priors, and (iv) the normalization is over all members of the set S.

In the preceding paper [1], improper priors were used (prior probabilities that cannot be normalized) because little prior information about the parameters was assumed. Any uninformative prior will look like a constant over the range of values where the likelihood is sharply peaked. The uninformative prior then cancels when the distribution is normalized. The parameters estimated by

this procedure are the maximum-likelihood estimates, and, in the case of a Gaussian noise prior, the least-squares estimates. In model selection calculations improper priors cannot be used. This is easily seen; for example, if the prior is a bounded uniform prior, then as the bounds are allowed to go to infinity the bounds will not cancel from the posterior probability of the model, Eq. (1), unless every model contains exactly the same prior. Thus, the model with the largest number of parameters would automatically be excluded.

The Global Likelihood of the Data

The posterior probability of model f_j is computed assuming σ, the standard deviation of the noise to be known, then at the end of the calculation σ is eliminated to obtain $P(f_j|D,I)$. To compute $P(f_j|\sigma,D,I)$ three terms must be computed: the prior probability $P(f_j|I)$, the normalization constant $P(D|I)$, and the global likelihood $P(D|\sigma,f_j,I)$. In this calculation little prior information about the models and about the numerical values of the parameters is assumed. The prior probability of the models is assigned assuming no preference for any model in the set S. The appropriate uninformative prior given s possible outcomes is the uniform prior $1/s$, where s is the total number of models. Using $1/s$ as the prior, and substituting the marginal probability of the data, Eq. (2), the posterior probability of the model, Eq. (1), becomes

$$P(f_j|\sigma,D,I) = \frac{P(D|\sigma,f_j,I)}{\sum_{k=1}^{s} P(D|\sigma,f_k,I)}. \tag{7}$$

The prior $P(f_j|I) = 1/s$ has canceled, leaving only the global likelihood of the data.

To compute the global likelihood of the data, a Gaussian noise prior is assumed. As was explained earlier [1], a conservative calculation is thus performed. Again, any other maximum entropy noise prior has more compact support (lower entropy) than the Gaussian given the finite noise power assumption; thus any other maximum entropy noise prior will always make more optimistic estimates of the parameters and the models. Using a Gaussian noise prior, the global likelihood Eq. (6) may be written as

$$\begin{aligned} P(D|f_j,\sigma,I) &= (2\pi\sigma^2)^{-\frac{N}{2}} \int d\mathbf{\Theta} P(\mathbf{\Theta}|f_j,I) \int d\mathbf{B} P(\mathbf{B}|\mathbf{\Theta},f_j,I) \\ &\quad \times \exp\left\{ -\sum_{i=1}^{N} \frac{[d_i - f_j(t_i)]^2}{2\sigma^2} \right\}. \end{aligned} \tag{8}$$

Substituting the signal Eq. (4) for f_j in the global likelihood of the data Eq. (8) one obtains

$$\begin{aligned} P(D|f_j,\sigma,I) &= (2\pi\sigma^2)^{-\frac{N}{2}} \int d\mathbf{\Theta} P(\mathbf{\Theta}|f_j,I) \\ &\quad \times \int d\mathbf{B} P(\mathbf{B}|\mathbf{\Theta},f_j,I) \exp\left\{ -\sum_{i=1}^{N} \frac{[d_i - \sum_{k=1}^{m} B_k G_k(t_i,\mathbf{\Theta})]^2}{2\sigma^2} \right\}. \end{aligned} \tag{9}$$

To proceed, the prior probability of the parameters must be supplied. Little prior information is assumed about the numerical value of either set of parameters. Thus, the exact functional form of the prior is of little importance. The important factors are the prior range and the height of the prior at the maximum-likelihood point: it is only these that have a chance of surviving to the end of the calculation, and they may cancel.

The amplitudes $\mathbf{B}$ are location parameters, and the appropriate prior for a location parameter is a Gaussian – see Refs. 2–4, and Ref. 6 for a more extensive discussion of priors. Suppose all that

is known about the signal is that it carried a finite total power. Using the principle of maximum entropy to assign the prior, given finite total power assumption, results in the assignment of

$$P(\mathbf{B}|f_j,\mathbf{\Theta},\delta,I)=\sqrt{\lambda_1\cdots\lambda_m}(2\pi\delta^2)^{-\frac{m}{2}}\exp\left\{-\sum_{i=1}^{N}\sum_{k=1}^{m}\frac{\sum_{l=1}^{m}B_kB_lG_k(t_i)G_l(t_i)}{2\delta^2}\right\}, \tag{10}$$

a Gaussian. The variance δ^2 expresses how uncertain one is of the total power carried by the signal, and λ_j is the jth eigenvalue of the g_{jk} matrix Eq. (13) below. It will be assumed that the data determine the total power carried by the signal much better than the prior information: so $\delta^2 \gg \sigma^2$ will be assumed. With the addition of the new parameter δ, assumed known, the global likelihood of the data Eq. (9) becomes

$$\begin{aligned} P(D|f_j,\delta,\sigma,I) &= (2\pi)^{-\frac{N+m}{2}}\sigma^{-N}\delta^{-m}\int d\mathbf{B}d\mathbf{\Theta}\sqrt{\lambda_1\cdots\lambda_m}P(\mathbf{\Theta}|f_j,I) \\ &\times\exp\left\{-\sum_{i=1}^{N}\frac{[d_i-\sum_{k=1}^{m}B_kG_k(t_i)]^2}{2\sigma^2}\right\} \\ &\times\exp\left\{-\sum_{i=1}^{N}\frac{\sum_{k=1}^{m}\sum_{l=1}^{m}B_kB_lG_k(t_i)G_l(t_i)}{2\delta^2}\right\}. \end{aligned} \tag{11}$$

To proceed, this equation must be integrated with respect to the amplitudes $\mathbf{B}$ and with respect to the nonlinear $\mathbf{\Theta}$ parameters.

The integrals over the amplitudes $\mathbf{B}$ are done first. The details of the amplitude integration are very similar to what was done in the preceding paper [1] and in [2]. First a change of variables

$$B_k=\sum_{l=1}^{m}\frac{A_le_{lk}}{\sqrt{\lambda_l}} \quad \text{and} \quad A_k=\sqrt{\lambda_k}\sum_{l=1}^{m}B_le_{kl} \tag{12}$$

is introduced, where e_{lk} is the lth component of the kth normalized eigenvector of the matrix g_{kl} defined as

$$g_{kl}\equiv\sum_{i=1}^{N}G_k(t_i)G_l(t_i), \tag{13}$$

and λ_k is the kth eigenvalue. Next a change of function

$$H_l(t)=\frac{1}{\sqrt{\lambda_l}}\sum_{k=1}^{m}e_{lk}G_k(t) \tag{14}$$

is introduced, where the orthonormal model functions H_j have the property

$$\sum_{i=1}^{N}H_j(t_i)H_k(t_i)=\delta_{j\,k}. \tag{15}$$

With these substitutions, the global likelihood of the data Eq. (11) becomes

$$\begin{aligned} P(D|f_j,\delta,\sigma,I) &= (2\pi)^{-\frac{N+m}{2}}\sigma^{-N}\delta^{-m}\int d\mathbf{A}d\mathbf{\Theta}P(\mathbf{\Theta}|f_j,I) \\ &\times\exp\left\{-\sum_{k=1}^{m}\frac{A_k^2}{2\delta^2}-\frac{N}{2\sigma^2}\left[\overline{d^2}-\frac{2}{N}\sum_{k=1}^{m}A_kh_k+\frac{1}{N}\sum_{k=1}^{m}A_k^2\right]\right\}, \end{aligned} \tag{16}$$

where $\overline{d^2}$ is the mean-square data value,

$$\overline{d^2} \equiv \frac{1}{N}\sum_{i=1}^{N} d_i^2, \tag{17}$$

h_k is the projection of the data onto the kth orthonormal model function

$$h_k \equiv \sum_{i=1}^{N} H_k(t_i) d_i, \tag{18}$$

and the volume elements are given by

$$d\mathbf{B} = d\mathbf{A}\lambda_1^{-\frac{1}{2}} \cdots \lambda_m^{-\frac{1}{2}}. \tag{19}$$

Performing the integrals over the amplitudes one obtains

$$\begin{aligned} P(D|f_j,\delta,\sigma,I) &= (2\pi\sigma^2)^{-\frac{N}{2}} \left[\frac{\sigma^2+\delta^2}{\sigma^2}\right]^{-\frac{m}{2}} \\ &\times \int d\mathbf{\Theta} P(\mathbf{\Theta}|I) \exp\left\{-\frac{N\overline{d^2}}{2\sigma^2} + \frac{m\overline{h^2}\delta^2}{2\sigma^2(\delta^2+\sigma^2)}\right\}, \end{aligned} \tag{20}$$

where

$$\overline{h^2} \equiv \frac{1}{m}\sum_{k=1}^{m} h_k^2 \tag{21}$$

is the mean-square projection of the orthonormal model functions onto the data.

By assumption, the prior uncertainty in the amplitudes, represented by δ^2, is much greater than σ^2, so the global likelihood of the data Eq. (20) may be simplified somewhat to obtain

$$\begin{aligned} P(D|f_j,\delta,\sigma,I) &\approx (2\pi)^{-\frac{N}{2}} \sigma^{m-N} \delta^{-m} \\ &\times \int d\mathbf{\Theta} P(\mathbf{\Theta}|I) \exp\left\{-\frac{[N\overline{d^2} - m\overline{h^2}]}{2\sigma^2} - \frac{m\overline{h^2}}{2\delta^2}\right\}. \end{aligned} \tag{22}$$

This equation represents the first step in the model selection process and is effectively the parameter estimation problem from the preceding paper [1]. It is almost what is needed to investigate signal detection problems.

Signal Detection

Suppose the nonlinear $\mathbf{\Theta}$ parameters are known (or given in the sense of investigating a range of possible values), then Bayes' theorem indicates the global likelihood of the data given the model f_j is

$$P(D|f_j,\mathbf{\Theta},\delta,\sigma,I) = (2\pi)^{-\frac{N}{2}} \sigma^{m-N} \delta^{-m} \exp\left\{-\frac{[N\overline{d^2} - m\overline{h^2}]}{2\sigma^2} - \frac{m\overline{h^2}}{2\delta^2}\right\}. \tag{23}$$

Because the $\mathbf{\Theta}$ parameters are assumed known, the prior $P(\mathbf{\Theta}|f_j,I)$ appearing in the global likelihood of the data, Eq. (22), does not appear, nor does the integral over these parameters, and $\mathbf{\Theta}$ was added to $P(D|f_j,\mathbf{\Theta},\delta,\sigma,I)$ to indicate that these parameters are given. This formulation is nearly what is needed to investigate signal detection problems. To see this, suppose one is trying to detect

a sinusoid in noise. The detection system is never perfect and the data will have a small constant component. There are two models: (i) a constant and (ii) a sinusoid plus a constant. Now assume that the frequency ω is given. The global likelihood of the data given the nonlinear parameters Eq. (23), may be used to ask "Given the choice between a constant and a sinusoid plus a constant, which model best accounts for the data?" By varying the given frequency ω through a range of values one could "see" which frequencies were being supported by the data and how strongly they are being supported relative to the alternatives. In a following paper [7] this example is worked in considerable detail.

Unfortunately, the global likelihood of the data Eq. (23) requires one to know δ^2 and the noise variance σ^2. Suppose neither is known. To eliminate these parameters apply the product and the sum rules from conditional probability theory: multiply by the appropriate normalized prior probability and integrate over the parameters. Both σ and δ are scale parameters, and Jeffreys [3] demonstrated that the appropriate uninformative prior for a scale parameter v is the Jeffreys' prior $1/v$. This prior is an improper prior because it is not normalizable on the interval zero to infinity. It was noted earlier that the use of improper priors was to be avoided in model selection problems. The Jeffreys' prior can be normalized if lower bound a_v and upper bound b_v are introduced. Although this would replace one unknown parameter by two, it is the preferred course. Taking

$$P(\sigma,\delta|a_\sigma,b_\sigma,a_\delta,b_\delta,I)=P(\sigma|a_\sigma,b_\sigma,I)P(\delta|a_\delta,b_\delta,I)=\frac{1}{\log(b_\sigma/a_\sigma)\sigma\log(b_\delta/a_\delta)\delta} \tag{24}$$

as the joint normalized prior probability for σ and δ, and multiplying the global likelihood of the data Eq. (23) by this prior, and integrating with respect to δ and σ, one obtains

$$\begin{aligned} P(D|f_j,\mathbf{\Theta},a_\sigma,b_\sigma,a_\delta,b_\delta,I) &= \frac{(2\pi)^{-\frac{N}{2}}}{\log(b_\sigma/a_\sigma)\log(b_\delta/a_\delta)}\int_{a_\sigma}^{b_\sigma}\int_{a_\delta}^{b_\delta}d\sigma d\delta\sigma^{m-N-1}\delta^{-m-1} \\ &\quad\times\exp\left\{-\frac{[N\overline{d^2}-m\overline{h^2}]}{2\sigma^2}-\frac{m\overline{h^2}}{2\delta^2}\right\}. \end{aligned} \tag{25}$$

If the lower and upper bounds are wide (the variances are not known well), then these integrals may be done approximately to obtain

$$P(D|f_j,\mathbf{\Theta},a_\sigma,b_\sigma,a_\delta,b_\delta,I)\approx\frac{\Gamma(\frac{m}{2})\Gamma(\frac{N-m}{2})}{4\log(b_\sigma/a_\sigma)\log(b_\delta/a_\delta)}\left[\frac{m\overline{h^2}}{2}\right]^{-\frac{m}{2}}\left[\frac{N\overline{d^2}-m\overline{h^2}}{2}\right]^{\frac{m-N}{2}}, \tag{26}$$

where $\Gamma(x)$ is a Gamma function of argument x. If the global likelihood of the data given the nonlinear parameters Eq. (26) is substituted into the posterior probability of the model, Eq. (7), $\log(b_\sigma/a_\sigma)$ will always cancel provided the variance σ^2 is assumed unknown in all models, and $\log(b_\delta/a_\delta)$ will cancel provided *every model* contains at least one amplitude. Making these assumptions the irrelevant terms may be dropped to obtain

$$P(D|f_j,\mathbf{\Theta},I)\propto\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{N-m}{2}\right)\left[\frac{m\overline{h^2}}{2}\right]^{-\frac{m}{2}}\left[\frac{N\overline{d^2}-m\overline{h^2}}{2}\right]^{\frac{m-N}{2}}. \tag{27}$$

This is the form of the global likelihood that is needed for signal detection problems.

If the models are regression or polynomial models, one can use these equations to calculate the probability of the expansion order. For examples of this calculation see Refs. 2,4.

Signal Detection – Exponential Decay

In many experiments it is often very difficult to tell if the "signal" has been detected. The equations as developed up to this point can be used for this purpose, but to do so one must state exactly what

one means by a "signal" and by "no signal." In this example "no signal" will be represented by a constant model

$$f_1(t) = B_1, \tag{28}$$

and the "signal" will be represented by a constant plus an exponential

$$f_2(t) = B_1 + B_2 \exp\{-\alpha t\}. \tag{29}$$

The decay rate constant α is supposed given in the sense that various values of α will be tested to see which, if any, are supported by the data.

In this demonstration, computer generated data are used. The data are a simulated on-resonance FID generated from

$$d(t_i) = 1 + 1.5 \exp\{-2t_i\} + e(t_i) \qquad (1 \leq i \leq 500). \tag{30}$$

The sampling time is taken to be a millisecond, data are taken every millisecond for 0.5 s; therefore obtaining 500 data values. The noise was generated from a unit normal random number generator; e.g., the true variance of the noise is one. Two different data sets were generated; the first, Fig. 1A, does not contain the exponential component and the second, Fig. 1C, does. The peak signal-to-RMS-noise ratio in Fig. 1C is 1.5, or 2.5 if the constant is included in the signal. *The noise is exactly the same in the two data sets.* Any differences obtained must be due to the presence of the signal. Visual examination of the two plots does disclose that Fig. 1C has a gentle slope to it. However, one could not be very sure that one has detected a signal nor present very convincing arguments to a skeptic that the signal is present.

Given these two models, Eqs. (28) and (29), one would like to ask what is the evidence in favor of model f_2? To answer this question, one computes the posterior probability of the two models using the global likelihood of the data, Eq. (27). To give an adequate graphical representation, a plot of the ratio

$$K \equiv \frac{P(f_2|\mathbf{\Theta}, D, I)}{P(f_1|\mathbf{\Theta}, D, I)} = \frac{P(D|f_2, \mathbf{\Theta}, I)}{P(D|f_1, \mathbf{\Theta}, I)} \tag{31}$$

will be computed; this ratio is called "the odds." Because the odds will tend to be rapidly varying, even for the data shown in Fig. 1A and Fig. 1C, $10\log_{10}(K)$ is plotted. This function is called "the evidence" and has units of decibels. If the evidence is 0 dB then the odds are 1 to 1: neither model is to be preferred. If the evidence is 20 dB then the odds are 100 to 1 in favor of the exponential model, and if the evidence is -20 dB then the odds are 100 to 1 in favor of the constant model.

Figure 1B is a plot of the evidence for the data shown in Fig. 1A. Notice that for all values of the decay rate constant α, the evidence is in favor of the constant model. From Fig. 1B one concludes that the data shown in Fig. 1A do not contain any positive evidence in favor of the exponential model: no signal has been detected. Figure 1D contains the evidence for a signal in the data shown in Fig. 1C. Here for all values of the α smaller than about 50 s^{-1}, there is positive evidence in favor of the signal of interest. For small values of α it is a bet of better than 10,000 to 1 in favor of the signal of interest.

One would expect intuitively that for large values α, the odds ratio should favor the constant model and indeed Fig. 1D clearly shows this. This occurs because for large values of the decay rate constant, the exponential signal function decays to zero in the first few time samples. The exponential model cannot fit the data any better than the constant model. However, it has two amplitudes to estimate while the constant model only has one. The larger model is being penalized because of its extra degree of freedom.

Given that the noise is identical in both data sets, Fig. 1A and Fig. 1C, the positive odds seen in Fig. 1D must be due to the presence of the signal. Thus, Bayesian probability theory finds good evidence for the exponential in data with low peak signal-to-RMS-noise ratio.

Figure 1: Signal Detection

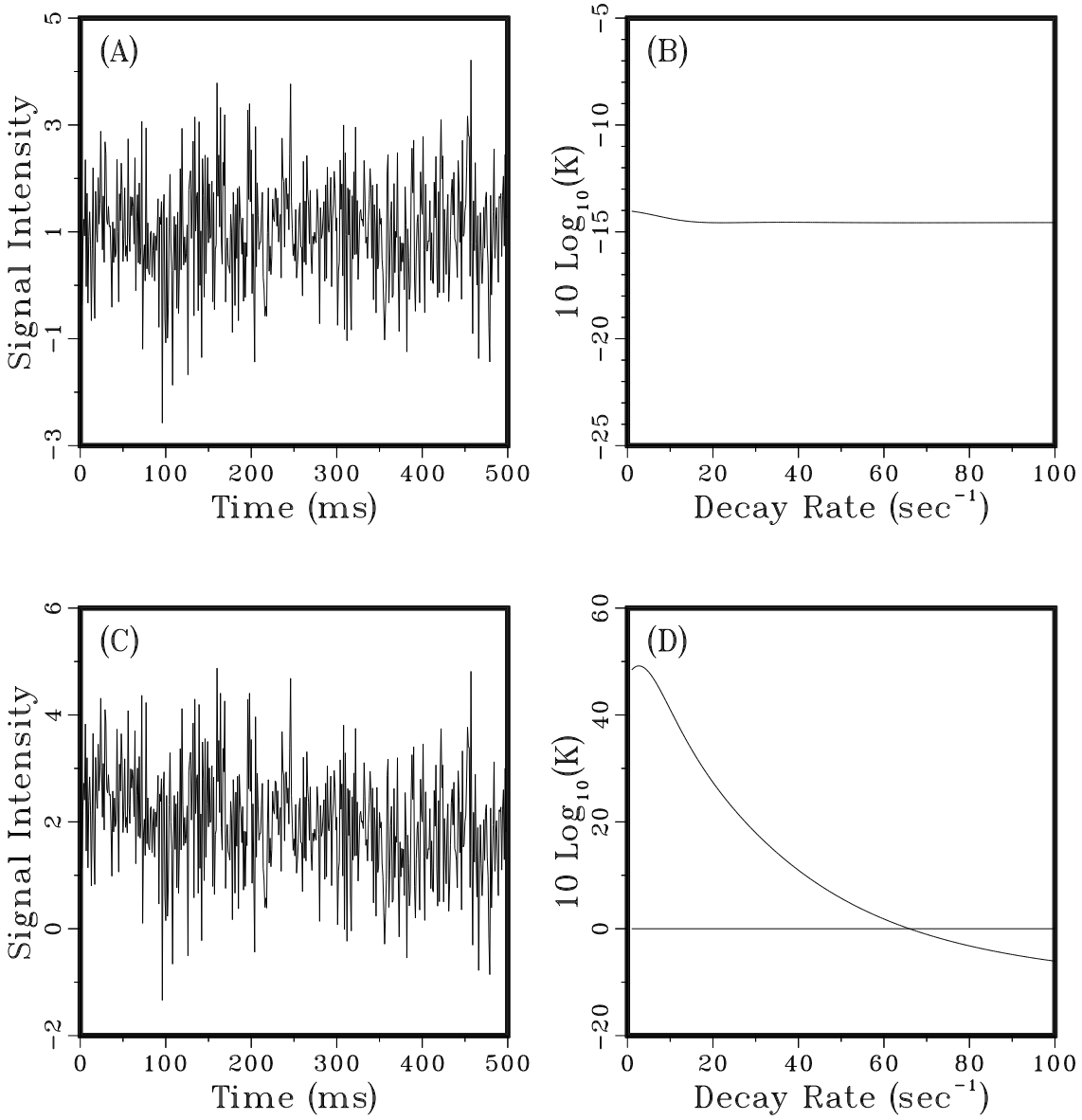

Fig. 1. (A) The data contain a constant component of amplitude one plus Gaussian white noise with standard deviation one (see text for the details of data preparation). Probability theory was then used to ask if there was any evidence for an exponential component in these data; this is shown in (B). For all values of the decay rate constant, it is a bet of approximately 100 to 1 against the exponential model. Thus, probability theory correctly indicates there is no signal. (C) The data from (A), but with a decaying exponential of unit amplitude and decay rate constant of 2 s^{-1} added to it. The evidence in favor of the exponential model was then computed; this is displayed in (D). Because the only difference in these two data sets is the signal, the change in the evidence must be due to the presence of the signal. In (D), for small values of the decay rate constant, it is a bet of better than 10,000 to 1 in favor of the exponential model.

Model Selection

Suppose the signal has been detected and one wants to determine the best model for the data. The global likelihood of the data given the nonlinear parameters, Eq. (27), cannot be used for this, because the nonlinear $\mathbf{\Theta}$ parameters are not known in general. One must go back to the global likelihood, Eq. (22), and carry out the remaining integrals over the nonlinear $\mathbf{\Theta}$ parameters. For an arbitrary model, this is obviously impossible. To obtain a useful analytic result, a number of approximations must be made.

Primarily, one must assume that the data determine the nonlinear $\mathbf{\Theta}$ parameters well. This is not equivalent to saying that the parameters have been well resolved, or even resolved, only that outside of some region, the posterior probability falls off rapidly: there exists a region in parameter space around which the sufficient statistic $\overline{h^2}$ may be approximated by a Taylor expansion. Fortunately, this will be the case in many problems concerning NMR. However, it will not be true in every conceivable case. In the event there is no well-defined maxima, one has no choice but to do the integrals numerically. Additionally, it will be assumed that the maximum in the posterior probability is unique: that is, there is only a single maximum. This is not the case for several important problems, and the modifications needed to account for multiple identical maxima will be noted latter.

To approximate the integrals over the nonlinear $\mathbf{\Theta}$ parameters, a Taylor expansion of $\overline{h^2}$ around the maximum of the posterior probability will be used. The values that maximize the posterior probability for model f_j are designated as $\hat{\mathbf{\Theta}} \equiv \{\hat{\Theta}_1, \cdots, \hat{\Theta}_r\}$. Taylor expanding $\overline{h^2}$ about $\hat{\mathbf{\Theta}}$ one obtains

$$\overline{h^2} \approx \overline{h^2}\big|_{\hat{\mathbf{\Theta}}} - \sum_{k=1}^{r}\sum_{l=1}^{r} \frac{b_{kl}}{m}(\hat{\Theta}_k - \Theta_k)(\hat{\Theta}_l - \Theta_l), \tag{32}$$

where

$$b_{kl} \equiv -\frac{\partial^2 m\overline{h^2}}{2\partial\Theta_k\,\partial\Theta_l}\bigg|_{\hat{\mathbf{\Theta}}}. \tag{33}$$

With this approximation the global likelihood of the data, Eq. (22), may be written as

$$\begin{aligned} P(D|f_j,\delta,\sigma,I) \;&\approx (2\pi)^{-\frac{N}{2}}\sigma^{m-N}\delta^{-m}\exp\left\{-\frac{N\overline{d^2}}{2\sigma^2}+\frac{mu\overline{h^2}}{2}\right\}\bigg|_{\hat{\mathbf{\Theta}}} \\ &\times\int d\mathbf{\Theta}P(\mathbf{\Theta}|f_j,I)\exp\left\{-u\sum_{k,l=1}^{r}\frac{b_{kl}(\hat{\Theta}_k-\Theta_k)(\hat{\Theta}_l-\Theta_l)}{2}\right\}, \end{aligned} \tag{34}$$

where

$$u \equiv \frac{1}{\sigma^2} - \frac{1}{\delta^2}.$$

To perform the integrals, the prior probability $P(\mathbf{\Theta}|f_j, I)$ must be assigned. In this approximation the nonlinear $\mathbf{\Theta}$ parameters are location parameters. Little prior information about the numerical values of the nonlinear $\mathbf{\Theta}$ parameters is assumed: it is assumed that the parameters could be either positive or negative and that they have a finite mean-square value.

Clearly, in any given problem, this may not be the case. For example, if the model is an exponentially decaying sinusoid, then the frequency could be positive or negative; but the decay rate constant is chosen so that the signal decays in time. Conceivably one could work this problem using this information and one would obtain slightly better results. But they would not be much better. It is only when the information in the prior is comparable to the information in the data that the prior probability can make any real difference in parameter estimation problems or in model selection problems.

Using the finite mean-square assumption in a maximum entropy calculation results in the assignment of a Gaussian prior probability:

$$P(\mathbf{\Theta}|\gamma, f_j, I) = (2\pi\gamma^2)^{-\frac{r}{2}} \exp\left\{-\sum_{k=1}^{r} \frac{\Theta_k^2}{2\gamma^2}\right\}. \tag{35}$$

The expected values of the nonlinear parameters over the prior information are given by

$$(\Theta_k)_{\rm est} = 0 \pm \gamma \tag{36}$$

at one standard deviation. The parameter γ expresses how strongly the nonlinear $\mathbf{\Theta}$ parameters are believed to be near zero. Equation (35) will not express a strong opinion about the parameters provided γ is large compared to the determination of the nonlinear $\mathbf{\Theta}$ parameters by the data. As with δ, γ is assumed to be known for now. At the end of the calculation one can eliminate γ if it is unknown. Using this prior probability of the nonlinear $\mathbf{\Theta}$ parameters, Eq. (35), in the global likelihood of the data, Eq. (34) becomes approximately

$$\begin{aligned} P(D|f_j, \gamma, \delta, \sigma, I) \quad &\approx (2\pi)^{-\frac{(N+r)}{2}} \sigma^{m-N} \delta^{-m} \gamma^{-r} \exp\left\{-\frac{N\overline{d^2}}{2\sigma^2} + \frac{mu\overline{h^2}}{2}\right\}\Bigg|_{\hat{\mathbf{\Theta}}} \\ &\times \int d\mathbf{\Theta} \exp\left\{-\sum_{k=1}^{r} \frac{\Theta_k^2}{2\gamma^2}\right\} \\ &\times \exp\left\{-u\sum_{k=1}^{r}\sum_{l=1}^{r} b_{kl} \frac{(\hat{\Theta}_k - \Theta_k)(\hat{\Theta}_l - \Theta_l)}{2}\right\}. \end{aligned} \tag{37}$$

To proceed the integrals over the nonlinear $\mathbf{\Theta}$ parameters must be done.

Little prior information about the nonlinear $\mathbf{\Theta}$ parameters has been assumed. By little prior information it is meant that the prior probability looks like a constant over the range of values where the likelihood of the data is strongly peaked. The integral may be approximated by evaluating the prior probability at $\hat{\mathbf{\Theta}}$ and, because it is essentially constant, it may be removed from the integral. With the indicated approximations the integral may be written as

$$\begin{aligned} P(D|f_j, \gamma, \delta, \sigma, I) \quad &\approx (2\pi)^{-\frac{(N+r)}{2}} \sigma^{m-N} \delta^{-m} \gamma^{-r} \exp\left\{-\frac{N\overline{d^2}}{2\sigma^2} + \frac{mu\overline{h^2}}{2} - \frac{r\overline{\Theta^2}}{2\gamma^2}\right\}\Bigg|_{\hat{\mathbf{\Theta}}} \\ &\times \int d\mathbf{\Theta} \exp\left\{-u\sum_{k=1}^{r}\sum_{l=1}^{r} b_{kl} \frac{(\hat{\Theta}_k - \Theta_k)(\hat{\Theta}_l - \Theta_l)}{2}\right\}, \end{aligned} \tag{38}$$

where

$$\overline{\Theta^2} \equiv \frac{1}{r}\sum_{k=1}^{r} \Theta_k^2 \tag{39}$$

is the mean-square nonlinear $\mathbf{\Theta}$ parameter. The integral over the nonlinear $\mathbf{\Theta}$ parameters may now be done to obtain

$$\begin{aligned} P(D|f_j, \gamma, \delta, \sigma, I) \quad &\approx (2\pi)^{\frac{-N}{2}} \sigma^{r+m-N} \delta^{-m} \gamma^{-r} \; v_1^{-\frac{1}{2}} \cdots v_r^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{r\overline{\Theta^2}}{2\gamma^2} - \frac{m\overline{h^2}}{2\delta^2} - [\frac{N\overline{d^2} - m\overline{h^2}}{2\sigma^2}]\right\}\Bigg|_{\hat{\mathbf{\Theta}}}, \end{aligned} \tag{40}$$

where v_j is the jth eigenvalue of the matrix b_{jk}. If the three variances δ^2, γ^2, and σ^2 are known, then the problem is completed and the global likelihood is given by Eq. (40).

Suppose that the three variances are unknown; then the product and sum rules may be used to eliminate them from the problem. The three variances, σ^2, δ^2, and γ^2, are scale parameters and a normalized Jeffreys' prior is used in the calculation. The integrals over these three variances are identical to what was done earlier and the details of the calculation are omitted. The global likelihood of the data, when the three variances are unknown, is given approximately by

$$\begin{aligned} P(D|f_j,I) \quad &\approx \frac{\Gamma(\frac{m}{2})}{2\log(b_\delta/a_\delta)}\left[\frac{m\overline{h^2}}{2}\right]^{-\frac{m}{2}} \frac{\Gamma(\frac{r}{2})}{2\log(b_\gamma/a_\gamma)}\left[\frac{r\overline{\Theta^2}}{2}\right]^{-\frac{r}{2}} v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}} \\ &\times\frac{\Gamma(\frac{N-m-r}{2})}{2\log(b_\sigma/a_\sigma)}\left[\frac{N\overline{d^2}-m\overline{h^2}}{2}\right]^{-\frac{N-m-r}{2}}\Bigg|_{\hat{\mathbf{\Theta}}}. \end{aligned} \tag{41}$$

So long as one compares models with the same types of parameters, the prior ranges on δ, γ, and σ cancel and the global likelihood of the data, Eq. (41), simplifies to

$$\begin{aligned} P(D|f_j,I) \quad &\approx \Gamma(\frac{m}{2})\Gamma(\frac{r}{2})\Gamma(\frac{N-m-r}{2})\; v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}} \\ &\times\left[\frac{m\overline{h^2}}{2}\right]^{-\frac{m}{2}}\left[\frac{r\overline{\Theta^2}}{2}\right]^{-\frac{r}{2}}\left[\frac{N\overline{d^2}-m\overline{h^2}}{2}\right]^{-\frac{N-m-r}{2}}\Bigg|_{\hat{\mathbf{\Theta}}}. \end{aligned} \tag{42}$$

This formulation of the global likelihood will be used to address model selection questions in several examples. In the following section the example began earlier is continued and the global likelihood of the data, Eq. (42), is used to compute the posterior probability of the number of exponentials in the data. Additional examples of its use are given in the following paper [7].

Before this example can be worked out, one missing factor must be discussed. When this calculation was performed, it was assumed that there existed a single maximum in the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters. Clearly, for nested models this is not the case. For example, suppose the model contains two decaying exponentials, $\mathbf{\Theta} = \{\theta_1, \theta_2\}$, with a well-defined maximum at $\theta_1 = \hat{\theta}_1$ and $\theta_2 = \hat{\theta}_2$, then by symmetry there must be another maximum at $\theta_2 = \hat{\theta}_1$ and $\theta_1 = \hat{\theta}_2$. The integral over the nonlinear $\mathbf{\Theta}$ parameters did not take this symmetry into account. This is easily corrected by multiplying the global likelihood of the data by a numerical factor equal to the number of maxima. This multiplicity factor is just the factorial of the number of identical signals present in the model. But this approximation assumes that these maxima do not overlap to any significant degree. If the maxima overlap, then the approximations made in this calculation are not valid and one must do the integrals numerically.

Model Selection – Exponential Decays

The model selection calculation just completed makes a number of approximations, and it is important to understand when these approximations are valid. To illustrate how the calculation will break down, and how to use the calculation correctly, an example will be given using decaying exponentials as the signal functions. These signal functions are about the most pathological signal functions one can use. They are pathological in the sense that there are no conditions under which the presence of a second exponential does not interfere with the estimation of the first exponential: that is the g_{jk} matrix Eq. (13) is never diagonal. This means that if one does not know the number of exponentials in the data and attempts to estimate the values of the decay rate constants with a model that uses the wrong number of exponentials, the estimates will be incorrect. The results are incorrect in the sense that, in the limit as the noise goes to zero the estimates from probability theory, maximum likelihood, and least squares go to complicated weighted averages, not to the "true" values.

The posterior probability of the number of exponentials, r, may be computed from the global likelihood of the data Eq. (42). To see this, one starts from Bayes' theorem and computes the probability of the number of exponentials, r, given the data and the prior information. This is given by

$$P(r|D,I) = \frac{P(r|I)P(D|r,I)}{P(D|I)}, \tag{43}$$

where $P(r|I)$ is the prior probability of the number of exponentials in the data, $P(D|r,I)$ is the probability of the data given the number of exponentials, and $P(D|I)$ is a normalization constant. For this problem, the set of models S may be written

$$f_r(t) = \sum_{k=1}^{r} B_k \exp\{-\alpha_k t\}, \tag{44}$$

where $r = \{1, 2, \cdots, s\}$ is the number of exponentials in the model. The posterior probability $P(f_r|D,I)$ is the probability of model f_r given the data and the prior information, but this is the probability of the number of exponentials, r. Changing notation one may write

$$P(r|D,I) = P(f_r|D,I) = \frac{P(f_r|I)P(D|f_r,I)}{P(D|I)}. \tag{45}$$

Using a uniform prior $P(f_r|I) = 1/s$, where s is an upper bound on the number of exponentials in the data, the posterior probability of the number of exponentials, r, becomes

$$P(r|D,I) = \frac{P(D|f_r,I)}{\sum_{j=1}^{s} P(D|f_j,I)}, \tag{46}$$

which may be computed from the global likelihood of the data, Eq. (42).

The upper bound s can be set in one of several ways: one can set s from prior information, or one can compute $P(D|f_r,I)$ for $r = \{1, 2, \text{etc.}\}$ until the unnormalized global likelihood has a maximum. When this occurs, one will be fairly sure of having found the correct number of exponentials in the data. In fact this is a slight misstatement. What one would have found is the *minimum number of decaying exponentials necessary to represent the signal down to the estimated noise level.* If the signal is composed of exponentials, then this is the answer one wants. If the signal is not exponential then one will still get the number of exponentials necessary to represent the data down to the estimated noise level. Bayesian probability theory will answer the question asked of it, and it will do so optimally. If it turns out to be a inappropriate question, one will still get an optimal answer.

In the following paper [7] Bayesian techniques are applied to sinusoidal models of FID data. There it is shown that, *from a Bayesian parameter estimation point of view,* the discrete Fourier transform answers a very specific question about frequency estimation. The discrete Fourier transform will give nearly optimal frequency estimates under a variety of conditions, but not all conditions encountered in FID data. If the only question one asks is the one answered by the discrete Fourier transform, then sometimes one will get an optimal answer to an inappropriate question.

To demonstrate how to compute the probability of the number of exponentials, r, computer generated data are used. The data were generated from

$$d(t_i) = 100\exp\{-30t_i\} + 50\exp\{-50t_i\} + e(t_i) \qquad (1 \le i \le 100). \tag{47}$$

The sampling time will again be a millisecond. The total duration of the data will be 0.1 s, so there are $N = 100$ data values, and the noise $e(t_i)$ was generated from a normal random number generator with unit variance. The peak signal-to-RMS-noise ratio is 150.

The traditional way to look for evidence of multiple exponential decay is to plot the data on a semilogarithmic plot, Fig. 2A. If the data contain a single exponential then this plot will be a

Figure 2: Model Selection

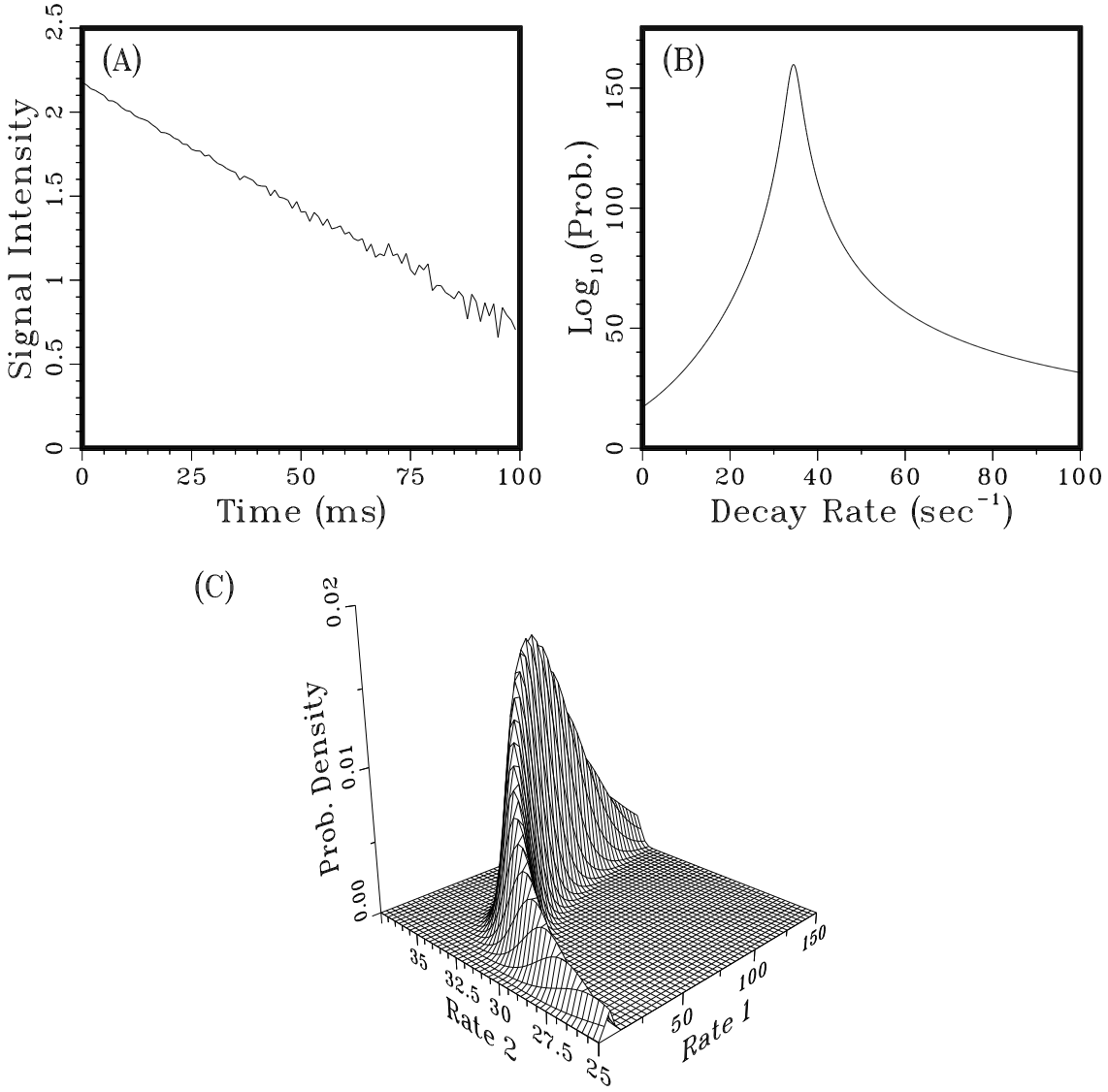

Fig. 2. (A) The data contain two decaying exponentials (see text for details). The base 10 logarithm of these data have been plotted in (A). Conventional wisdom would hold that if there is evidence of multiple exponentials in these data, this plot will not be a straight line. To the eye, the data in (A) do not significantly deviate from a straight line, although use of a ruler does suggest the presence of the second exponential. The base 10 logarithm of the posterior probability of a single decay rate constant is shown in (B). Because this probability density is so sharply peaked, the approximate calculation done in this paper is working well. Next the fully normalized posterior probability of two-exponential decay rate constants was computed and is shown as the surface plot (C). For the calculation of the posterior probability of the two-exponential model to be valid, the contours of constant probability density should be elliptical, and enclose most of the posterior probability. In (C) the Gaussian approximation is beginning to break down. The posterior probability of the two-exponential model was then computed, and it was found that the two-exponential model is favored by almost a million to one over the one-exponential model. When one tries to fit a three-exponential model, the Gaussian approximation is not valid and one must resort to numerical integration procedures.

straight line, and if there are multiple exponential components it will be a curve. From the plot one cannot see any marked deviations from a straight line, although the use of a ruler does suggest the presence of a second exponential.

The global likelihood of the data, Eq. (42), was computed for a model containing one exponential. This computation is done by first locating the maximum of the posterior probability of a single decay rate constant, and then computing the b_{jk} matrix, Eq. (33). A plot of the base 10 logarithm of the probability for a single exponential decay rate constant is shown in Fig. 2B. Notice that for one decaying exponential, the posterior probability rises approximately 150 orders of magnitude above the noise. When the posterior probability density function is this sharply peaked, the Gaussian approximation is working at its best.

Next, the posterior probability of two exponential decay rate constants was computed. There are two maxima in this probability density function; one of these is shown as a surface plot in Fig. 2C. In Fig. 2C the shorter of the two decay rates has been relatively well determined while the longer decay rate has been very poorly determined. This posterior probability density is not even approximately elliptical, so the Gaussian approximation is beginning to break down. However, it has not totally broken down, because most of the posterior probability is contained within the central maxima and it is approximately elliptical.

The Gaussian approximation could be improved in two ways: one could take more data or one could improve the signal-to-noise ratio. The height of the logarithm of the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters scales linearly with the number of data values. So sampling two times faster, thus doubling the number of data values over the region where there is signal, would double the height of the logarithm of the posterior probability. Because the posterior probability would be raised to an additional power of N, the Gaussian approximation would be improved. The height of the logarithm of the posterior probability also scales like the square of the signal-to-noise ratio. So if one doubles the signal-to-noise ratio, the height would be $4N$ after doubling. Thus doubling the signal-to-noise ratio will improve the Gaussian approximation much more than doubling the sampling rate. Of course the posterior probability would still have the wings, but they would be down many orders of magnitude and would be insignificant.

The Gaussian approximation was then used to compute the global likelihood for the model containing two decaying exponential signals. The global likelihood increased by roughly six orders of magnitude when going from the one-exponential model to the two-exponential model, indicating strong evidence in favor of the second exponential.

The posterior probability of three exponential decay rate constants was then computed and no well-defined maximum exists in the posterior probability density. Thus the Gaussian approximation made in this paper is not even approximately valid. Bayesian probability theory is still valid; but the integrals over the nonlinear $\mathbf{\Theta}$ parameters must be done numerically. One must go back to global likelihood of the data, Eq. (20), assign a prior probability to the three decay rate constants, and carry out the integrals over these parameters. For small numbers of parameters this is possible. Here a Monte Carlo integration was used and it indicates that the two-exponential model is preferred to the three-exponential model by about 20 to 1.

But detection, model selection, and parameter estimation are not the same problem: being able to say that two exponentials fit the data significantly better than one exponential is *not* the same as saying the parameters have been accurately determined. When two exponentials have nearly the same decay rate constants, the sum of the amplitudes will be well resolved, and unless the peak signal-to-RMS-noise ratio is very high, the difference in the amplitudes will be poorly resolved. Here only the sum of the amplitudes has been accurately determined. However, now that it is known that two exponentials are present, one can take steps to improve the resolution.

Discussion

At the onset it was assumed that the models would be chosen from the set S. To confine ourselves to the set S is not to assert dogmatically that there are no other possibilities; one may assign prior probabilities $P(f_j|I)$ $(1 \leq j \leq s)$ which do not add up to one:

$$\sum_{j=1}^{s} P(f_j|I) = a < 1. \tag{48}$$

Then one is assigning a prior probability $(1-a)$ to some unknown proposition:

$$\text{SE} \equiv \text{“Something Else not yet thought of.”}$$

But until SE is specified it cannot enter into a Bayesian analysis; probability theory can only compare the specified models $\{f_1, \cdots, f_s\}$ with each other.

This can be explicitly demonstrated. If one tries to include SE in the set of hypotheses, one can calculate the posterior probabilities of the models $\{f_1, \cdots, f_s\}$ and SE to obtain

$$P(f_j|D,I) = \frac{P(f_j|I)P(D|f_j,I)}{P(D|I)}, \tag{49}$$

and

$$P(\text{SE}|D,I) = \frac{P(\text{SE}|I)P(D|\text{SE},I)}{P(D|I)}. \tag{50}$$

But this is numerically indeterminate even if $P(\text{SE}|I) = 1-a$ is known, because $P(D|\text{SE},I)$ is undefined until that Something Else is specified. The denominator $P(D|I)$ is also indeterminate, because

$$\begin{aligned} P(D|I) &= \textstyle\sum_{j=1}^{s} P(f_j, D|I) + P(\text{SE}, D|I) \\ &= \textstyle\sum_{j=1}^{s} P(D|f_j, I)P(f_j|I) + P(D|\text{SE}, I)P(\text{SE}|I). \end{aligned} \tag{51}$$

But the relative probabilities of the specified models are still well defined, because the indeterminate cancels out:

$$\frac{P(f_i|D,I)}{P(f_j|D,I)} = \frac{P(f_i|I)P(D|f_i,I)}{P(f_j|I)P(D|f_j,I)}. \tag{52}$$

These relative probabilities are independent of what probability $(1-a)$ one assigns to Something Else, so one shall get the same results if one just ignores Something Else altogether and acts as if $a = 1$. In other words, while it is not wrong to introduce an unspecified Something Else into a probability calculation, no useful purpose is served by doing so.

Earlier it was indicated that normalized priors must be used for the model selection problem, and this can be seen in Eqs. (42) and (23). If one had used improper priors, neither the logarithm factors associated with normalizing the Jeffreys' prior nor the factors of the form $[X]^{-m}$ would have appeared. When models with amplitudes $\mathbf{B}$, nonlinear $\mathbf{\Theta}$ parameters, and unknown variance of the noise σ^2 are compared, the logarithmic factors cancel: the prior ranges on the parameters are unimportant. However, the factors of the form $[X]^{-m}$ do not cancel. It is these “Ockham” factors [5] that allow conditional probability theory to distinguish between models. Had improper priors been used, the likelihood factor on the right in the global likelihood, Eq. (42), would have been obtained. This factor is the likelihood of the data found in the preceding paper [1] and represents how well the model fits the data. It is closely related to the reciprocal of χ^2. It will always increase for hierarchical models as the number of signal functions m is increased. As was noted by Jaynes [8], larger models require one to smear one's prior information over a larger hypervolume in parameter space; for the larger model to be preferred, the likelihood of the data must increase more than what one would expect from fitting the noise.

Improper priors effectively smear the prior information over an infinite region. If a bounded uniform prior had been used in the calculation, and the bound allowed to go to infinity, the infinities would not cancel from the posterior probability of the model, Eq. (7). When models with different numbers of parameters are compared, the models containing the infinities are automatically excluded. Conversely, if maximum-likelihood model selection is used, by default a uniform prior has been used; but it is not bounded. Effectively the model with the larger number of model functions has been multiplied by infinity and, provided the models are hierarchical, the *larger model will always be accepted.* This does not occur in the Bayesian posterior probability because the global likelihood of the data is essentially the product of the likelihood times the prior probability of the model parameters. When the projection of a new model function onto the data is comparable to what is expected from expanding noise, the global likelihood of the data will remain essentially unchanged, while the prior probability of the model parameters will go down. Thus, the Bayesian answer to the model selection problem is essentially a microcosm of the scientific method: theorize about possible models, compare these to experiment, and when two models explain the data equally well select the simplest model.

Summary and Conclusions

The calculation presented in this paper makes use of full Bayesian probability theory to calculate, approximately, the posterior probability of a model. While Bayesian probability theory is always valid for this problem, the approximate calculation presented here is valid whenever there is a well-defined maximum in the parameter space. The calculation is general in the sense that it can be used to compare any models, not necessarily hierarchical models, as in the example.

This calculation is meant as a demonstration of how to use Bayesian probability theory to determine the "best" model, and while the calculation will be valid in a wide variety of problems, it will not always be valid and one must keep this firmly in mind. Additionally, there are many model selection problems where additional relevant prior information is available. To solve these problems one should go back to Bayes' theorem and solve the problem from first principles taking all of the available prior information into account. When this is done the results could be very different from what has been done here.

Last, the model selection example specifically demonstrated how this calculation breaks down, and could leave one with the feeling that the procedures are not worth the trouble. In a following paper [7], these procedures are applied to models which contain sinusoidal signals. These signals are almost the opposite of the exponential signals examined here: where exponentials are almost always ill behaved, sinusoids are almost always well behaved, and the calculations will perform well under a wide variety of conditions.

Acknowledgments

This work was supported by NIH grant GM-30331, J. J. H. Ackerman principal investigator. The encouragement of Professor J. J. H. Ackerman is greatly appreciated as are the editorial comments of Dr. C. Ray Smith and extensive conversations with Professor E. T. Jaynes. Last, I would like to thank Professor John Skilling for supplying us with the Monte Carlo integration routine used in the example.

References

[1] G. L. Bretthorst, *J. Magn. Reson.* **88,** pp. 533-551 (1990).

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

[3] H. Jeffreys, "Theory of Probability," Oxford Univ. Press, London, 1939; later editions, 1948, 1961.

[4] A. Zellner, "Bayesian Statistics" (J. M. Bernardo, Ed.), Valencia Univ. Press, Valencia, Spain, 1980.

[5] S. F. Gull, in "Maximum Entropy and Bayesian Methods in Science and Engineering" (G. J. Erickson and C. R. Smith, Eds.), Vol. 1, pp.53-75, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

[6] E. T. Jaynes, *IEEE Trans. Syst. Sci. Cybernets,* SSC-4, 227 (1968); reprinted in "Papers on Probability, Statistics and Statistical Physics," (R. D. Rosenkrantz, Ed.), D. Reidel, Boston, 1983.

[7] G. L. Bretthorst, *J. Magn. Reson.* **88,** pp. 571-595 (1990).

[8] E. T. Jaynes, JASA, September 1979, p. 740, review of "Inference, Methods, and Decision: Towards a Bayesian Philosophy of Science," by R. D. Rosenkrantz, D. Reidel, Boston.