Journal of Magnetic Resonance **88,** pp. 533-551 (1990)

Bayesian Analysis. I.
Parameter Estimation
Using Quadrature NMR Models

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

ABSTRACT. In the analysis of magnetic resonance data, a great deal of prior information is available which is ordinarily not used. For example, considering high resolution NMR spectroscopy, one knows in general terms what functional form the signal will take (e.g., sum of exponentially decaying sinusoids) and that, for quadrature measurements, it will be the same in both channels except for a 90° phase shift. When prior information is incorporated into the analysis of time domain data, the frequencies, decay rate constants, and amplitudes may be estimated much more precisely than by direct use of discrete Fourier transforms. Here, Bayesian probability theory is used to estimate parameters using quadrature models of NMR data. The calculation results in an interpretation of the quadrature model fitting that allows one to understand on an intuitive level what frequencies and decay rates will be estimated and why.

Introduction

Probability theory when interpreted as logic is a quantitative theory of inference, just as mathematics is a quantitative theory of deduction. Unlike the axioms of mathematics, the *desiderata* of probability theory do not assert that something is true; rather they assert that certain features of the theory are desirable. Stated broadly these *desiderata* are degrees of belief are represented by real numbers; probability theory when interpreted as logic must qualitatively correspond to common sense; and when the rules for manipulating probabilities allow the evidence to be combined in more than one way, one must reach the same conclusions, i.e., the theory must be self consistent. These qualitative requirements are enough to uniquely determine the content of the theory [1–4]. The rules for manipulating probabilities in this theory are just the standard rules of statistics plus Bayes' theorem. Thus Bayesian probability theory reinterprets the rules for manipulating probabilities. In this theory a probability represents a state of knowledge, not a state of nature.

Because a probability represents a reasonable degree of belief, a Bayesian can assign probabilities to hypotheses which have no frequency interpretation. Thus, problems such as: "What is the probability of a frequency ω, independent of the amplitude and phase of the sinusoid, given the data?" or "Given several possible models of the data, which model is most probable?" make perfect sense. The first question is a parameter estimation problem and assumes the model to be correct. The second question is more general; it is a model selection problem and does not assume the model to be correct.

In previous communications from this laboratory [5,6], select applications of probability theory to NMR spectroscopy were briefly described; but the theory was not given in detail. In this paper, the parameter estimation question is addressed using quadrature NMR models. These calculations lead to a qualitative interpretation of parameter estimation that allows one to understand on an intuitive level what values of the parameters (resonance frequencies and decay rate constants) will be estimated, and why.

In a second paper [7], the model selection question will be addressed and probability theory is used to answer questions of the form "Given a set of possible models $\{f_1, \cdots, f_s\}$, which model is most probable in view of the data and all of one's prior information?" The Bayesian answer to this question will be shown to include a *quantitative* statement of Ockham's razor. Examples of signal detection and model selection are given to illustrate the use of the calculation and the limits of its validity.

In a third paper [8], the calculations will be specialized to data containing sinusoidal signals. A number of specific and important examples are given. For those who wish to put off the mathematics in the first and second papers, the third paper is self-contained enough to allow independent reading. In that paper, the exact relation between the discrete Fourier transform and Bayesian probability theory will be derived, and it will be shown that under appropriate conditions, no better estimate of the resonance frequencies may be obtained than from the peak value of a zero-padded discrete Fourier transform *power spectrum.* A signal detection example is presented, and it is shown that Bayesian probability theory can detect signals and estimate frequencies in data where no peak unambiguously exists in traditional absorption spectra. Finally, the model selection calculation is applied to real NMR data.

The Posterior Probability of the Resonance Frequencies and Decay Rate Constants

Magnetic resonance theory typically describes the free induction decay time series as sinusoids with exponential or Gaussian decay. The initial work employing probability theory [9,10] did not incorporate all of the prior information about the NMR signals that one actually has. The functional form of the signal was utilized, but the data were analyzed as if two distinct measurements were available having the *same* frequencies and decay rate constants, but completely *different* amplitudes and phases. However, in quadrature NMR there is more information than this: the signal in the second channel is 90° out of phase with the signal in the first channel, but is otherwise identical. In this paper, the parameter estimation calculation is specialized to quadrature models. The posterior probability for the frequencies and decay rate constants is derived *independent* of the amplitudes, phases, and the variance of the noise. This allows an examination of the frequencies and decay rate constants without the (initially) "uninteresting" amplitudes, phases, and variance of the noise interfering with the estimation process.

The problem considered in this paper is: Given a quadrature detected data set, what is the "best" estimate of the frequencies and decay rate constants that one can make from the data and the prior information? By "best" one means *most probable.* Specifically one would like to compute the joint posterior probability density of the frequencies $\{\omega_1, \cdots, \omega_n\}$ and decay rate constants $\{\alpha_1, \cdots, \alpha_n\}$ *independent* of the amplitudes and phases, given a quadrature detected data set D and the prior information I. The total number of spectral lines n is assumed known in this paper; then in the second paper [7] the posterior probability of the number, n, of spectral lines, is computed and examples of its use are given in the third paper [8]. The posterior probability of the frequencies and decay rate constants is abbreviated $P(\boldsymbol{\Theta}|D, I)$, where the "$|D, I)$" means the probability density is conditional on the data D and the prior information I, and $\boldsymbol{\Theta}$ is defined as the set of nonlinear parameters $\boldsymbol{\Theta} \equiv \{\omega_1, \cdots, \omega_n, \alpha_1, \cdots, \alpha_n\}$.

To make progress on the parameter estimation problem, one must state exactly what prior information is to be incorporated into the calculation. In the general NMR calculation described herein, it is assumed that little prior information is available about the numerical values of the parameters; most of the prior information is rather in the functional form of the model i.e., the number and types of parameters that affect the data. The model employed assumes the data may be separated into a systematic part (signal) and a random part (noise). In a quadrature detected data set there are actually two models: one for the so-called real data (0° phase) and one for the imaginary data (90° phase). The model for the real data may be written

$$d_R(t_i) = f_R(t_i) + e(t_i), \tag{1}$$

where $f_R(t_i)$ is the model signal in the real channel and $e(t_i)$ represents the random part or noise. The separation of the data into a signal f_R and additive noise $e(t_i)$ is a hypothesis which is assumed to be true; it will be taken to be part of the general background information I. The assumption that the noise is additive will not be questioned in this calculation, although using probability theory different assumptions about how the noise enters the problem are easily tested. Similarly, the quadrature or imaginary data may be modeled as

$$d_I(t_i) = f_I(t_i) + e(t_i). \tag{2}$$

It is assumed the data are sampled at discrete times t_i, not necessarily uniform. The noise is represented symbolically as $e(t_i)$ in both channels. So far no assumptions are made about it (e.g., correlated *vs*not correlated), except that whatever the noise characteristics are in one channel, they are similar in the other.

The signals $f_R(t)$ and $f_I(t)$ are written as sums over functions U_j and V_j such that

$$f_R(t) = \sum_{j=1}^{m} B_j U_j(\mathbf{\Theta}, t), \tag{3}$$

and

$$f_I(t) = \sum_{j=1}^{m} B_j V_j(\mathbf{\Theta}, t), \tag{4}$$

where m is the total number of signal functions in one channel and B_j is the amplitude of the jth signal function. The signal functions $U_j(\mathbf{\Theta}, t)$ and $V_j(\mathbf{\Theta}, t)$ may be thought of as sinusoids with exponential or Gaussian decay, although nothing in the calculation will require this. Thus any quadrature data set that can be modeled by Eqs. (1)–(4) may be analyzed using this calculation. The functional form of the signal is another of those hypotheses that will be taken to be part of the general background information I. This assumption could also be tested using probability theory, but in this calculation that will not be done.

The signal functions $U_j(\mathbf{\Theta}, t)$ and $V_j(\mathbf{\Theta}, t)$ are functions of a *continuous* variable t which in this calculation is time; however, the data have been sampled only at discrete times $\{t_1, \cdots, t_N\}$. Additionally, $U_j(\mathbf{\Theta}, t)$ and $V_j(\mathbf{\Theta}, t)$ are assumed to be functions of other continuous parameters labeled $\mathbf{\Theta}$. These parameters are typically frequencies and decay rate constants. However, $\mathbf{\Theta}$ could also include the global phase of the sinusoids, or any other parameter needed to model the signal. The total number of nonlinear parameters is designated as r.

The quadrature (prior) information has been incorporated by assuming that the amplitudes B_j are the same in both channels. If phase coherence is included in the model, then the signal function will be of the form

$$f_R(t) = \sum_{j=1}^{m} B_j U_j(t) = \sum_{j=1}^{m} B_j \sin(\omega_j t + \theta) e^{-\alpha_j t}, \tag{5}$$

and

$$f_I(t) = \sum_{j=1}^{m} B_j V_j(t) = \sum_{j=1}^{m} B_j \cos(\omega_j t + \theta) e^{-\alpha_j t}, \tag{6}$$

where n is the number of frequencies, m is the number of signal functions, and r is the number of nonlinear parameters. For the model just given, $m = n$, $r = 2n+1$, and $\mathbf{\Theta} \equiv \{\omega_1, \cdots, \omega_n, \alpha_1, \cdots, \alpha_n, \theta\}$.

If the phase coherences are not explicitly included, then all of the phases may be written as amplitudes. In this case the signal functions will occur in pairs,

$$f_R(t) = \sum_{j=1}^{m} B_j U_j(t) = \sum_{j=1}^{n} \left[B_j \sin(\omega_j t) + B_{j+n} \cos(\omega_j t)\right] e^{-\alpha_j t}, \tag{7}$$

$$f_I(t) = \sum_{j=1}^{m} B_j V_j(t) = \sum_{j=1}^{n} \left[B_j \cos(\omega_j t) - B_{j+n} \sin(\omega_j t)\right] e^{-\alpha_j t}, \tag{8}$$

where $m = 2n$, $r = 2n$, and B_j and B_{j+n} are effectively the amplitude and phase of the jth resonance. Of course, it is possible that the amplitudes in one of the two channels could be scaled by a constant factor, and the two channels could be slightly out of phase. Each of these possibilities is handled easily by the general formalism.

The goal is to compute the joint posterior probability density for the nonlinear $\mathbf{\Theta}$ parameters (the frequencies and decay rates) independent of the amplitudes $\mathbf{B}$. According to Bayes' theorem [11], the posterior probability of all of the parameters (*including the amplitudes*) is given by

$$P(\mathbf{\Theta}, \mathbf{B}|D, I) = \frac{P(\mathbf{\Theta}, \mathbf{B}|I) P(D|\mathbf{\Theta}, \mathbf{B}, I)}{P(D|I)}. \tag{9}$$

To compute the joint posterior probability of the parameters $P(\mathbf{\Theta}, \mathbf{B}|D, I)$, one must evaluate three terms. The first term $P(\mathbf{\Theta}, \mathbf{B}|I)$ is the probability of the parameters given only the prior information I. This prior probability represents the state of knowledge about the parameters before this particular data set was taken; for example, a previous measurement would represent highly relevant prior information. However, in this calculation little prior information about the parameters will be assumed and broad uninformative prior probabilities will be used. The second term $P(D|\mathbf{\Theta}, \mathbf{B}, I)$ is called the direct probability of the data. It is often referred to as a "sampling distribution" when the model is held fixed and different data sets are examined, and it is often called a "likelihood function" when the data are held fixed and different parameter values are examined. The third term $P(D|I)$ is the marginal probability of the data given only the prior information. For this problem this term is a normalization constant, which may be ignored.

To remove the amplitudes from the problem, it is assumed that when the data were taken each of the amplitudes B_j could take on only one value; i.e., each B_j is a constant through the run of data. However, it is not known which value was actually realized in the experiment. 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 posterior probability of the frequencies and decay rates *independent of the amplitudes,* is given by

$$P(\mathbf{\Theta}|D, I) \propto \int d\mathbf{B} P(\mathbf{\Theta}, \mathbf{B}|I) P(D|\mathbf{\Theta}, \mathbf{B}, I). \tag{10}$$

It is this quantity that is computed for the general model, Eqs. (4) and (9).

To proceed one must assign either the prior probability $P(\mathbf{\Theta}, \mathbf{B}|I)$ or the direct probability of the data $P(D|\Theta, \mathbf{B}, I)$. The direct probability will be assigned first. The direct probability is the probability that this particular data set should have been obtained given that the signal is exactly known; but from the model Eqs. (1) and (2), this is just the probability of the noise. To assign the direct probability one must assign a prior probability to the noise. It is a prior probability because it depends only on what was known about the noise before this particular data set is analyzed.

In most experiments not much is actually known about the noise; but one thing is certain, when the data were taken, the noise carried a finite total power. This information is enough to assign a maximum entropy prior probability distribution. An extended discussion of the principle of maximum entropy is beyond the scope of this paper, but briefly, it is a theorem that maximum entropy probability distributions contain only the information I used in assigning them [ref1,4,12–14]. If an arbitrary procedure is used to assign a prior probability, regardless of the theoretical support that procedure may have, one can always find the information I that yields the same probability density from a maximum entropy calculation. If the information I is not the same as the stated assumptions used in the arbitrary procedure, then Shannon's theorem [14] guarantees that the arbitrary procedure assigned a probability distribution that either contains hidden assumptions or does not contain all of the information implicit in the stated prior information. In either case, one will reason inconsistently in the sense that one could reach conclusions that are not warranted on the basis of the stated assumptions, or one may not reach a conclusion that was warranted. Therefore, whenever possible, maximum entropy will be used in this calculation to assign prior probability distributions.

The known information about the noise is that the noise carried a finite but unknown total power. The maximum entropy calculation is straightforward and results in the assignment of a Gaussian. Thus, the probability that one should obtain a set of noise values $\mathbf{e} \equiv \{e(t_1), \cdots e(t_N)\}$ is

$$P(\mathbf{e}|\sigma, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{e(t_i)^2}{2\sigma^2}\right\}, \tag{11}$$

where the variance σ^2 is the average value of the noise power and has been assumed known for now.

Use of Gaussian noise prior probability is not the same thing as assuming Gaussian white noise. Indeed, as has been stressed before, probabilities in Bayesian probability theory represent states of knowledge, not facts of nature. The only fact of nature used in assigning the noise prior probability was that the noise has a finite total power. The Gaussian is simply the most conservative noise prior that is consistent with this fact. Indeed when using the Gaussian, probability theory will split the data into two categories: a noise category and a signal category. When the parameters are estimated, everything that cannot be placed into the signal will be placed into the noise. The accuracy of the parameter estimates depends on the estimated noise variance, and because everything not considered signal is noise, the accuracy estimates will be as wide as is consistent with the model and the data. If information were available that the noise were simple digitizing errors, Poisson in nature, or that correlations exist, that information could be used in a new maximum entropy calculation to obtain a noise prior probability that has lower entropy for a given total noise power. Because it has lower entropy, that new noise prior probability will be more informative than the Gaussian.

Having assigned a noise prior probability, one may proceed to calculate the direct probability of the data. Taking the difference between the data and the signal in the model equation Eq. (1), one obtains $e(t_i)$ as a function of $d_R(t_i)$. Substituting for $e(t_i)$ in the noise prior probability Eq. (11), the probability that one should obtain this particular noise sample in the real data is

$$P(d_R(t_1), \cdots, d_R(t_N)|f_R, \sigma, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{[d_R(t_i) - f_R(t_i)]^2}{2\sigma^2}\right\}. \tag{12}$$

For the imaginary channel a similar result holds:

$$P(d_I(t_1), \cdots, d_I(t_N)|f_I, \sigma, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{[d_I(t_i) - f_I(t_i)]^2}{2\sigma^2}\right\}, \tag{13}$$

where it is assumed that the average noise power in the two channels is the same, but the actual samples of noise in the two channels are independent. The probability that one should obtain the two data sets is just the product of the probability that one should obtain each of the data sets separately,

$$P(D|f_R, f_I, \sigma, I) = (2\pi\sigma^2)^{-N} \exp\left\{-\sum_{i=1}^{N} \frac{[d_R(t_i) - f_R(t_i)]^2 + [d_I(t_i) - f_I(t_i)]^2}{2\sigma^2}\right\}, \tag{14}$$

where D represents all of the data.

This is the direct probability of the data and is to be substituted back into Bayes' theorem, Eq. (9). However, first the exponent will be rearranged into a form more easily used. Substituting the model signal Eqs. (3) and (4) into the direct probability or likelihood, Eq. (14), the direct probability of the data given the parameters is written as

$$P(D|\mathbf{\Theta}, \mathbf{B}, \sigma, I) = (2\pi\sigma^2)^{-N} \exp\left\{-\frac{Q}{2\sigma^2}\right\}, \tag{15}$$

where

$$Q \equiv d_R \cdot d_R + d_I \cdot d_I - 2\sum_{j=1}^{m} B_j[d_R \cdot U_j + d_I \cdot V_j] + \sum_{j=1}^{m}\sum_{k=1}^{m} B_j B_k[U_j \cdot U_k + V_j \cdot V_k], \tag{16}$$

and $(\cdot)$ means the sum over the discrete times t_i, for example,

$$d_I \cdot V_j \equiv \sum_{i=1}^{N} d_I(t_i) V_j(t_i). \tag{17}$$

The models f_R and f_I in $P(D|f_R, f_I, \sigma, I)$ have been replaced by $\mathbf{\Theta}$ and $\mathbf{B}$ to indicate that one particular form of the model is now being considered and the model equations (the exponentially decaying sinusoids) are to be considered as part of the general background information I.

The direct probability of the data Eq. (15) may now be substituted back into Bayes' theorem to obtain the joint probability of the parameters. Before doing that, note that Bayes' theorem indicates that the joint posterior probability of the parameters given the data is essentially (i.e., , but for normalization) the product of the direct probability and the prior probability. Assignment of the direct probability of the data is complete, and the prior probability of the parameters will be formulated next.

To assign the prior probability of the parameters, exactly what is known about the parameters must be stated. In this problem, it is assumed that there is little, effectively no, prior information about the parameters. Because so little information about the parameters is assumed, knowledge of one parameter would not help in estimating the others. With this assumption, the prior will factor and may be written as

$$P(\mathbf{\Theta}, \mathbf{B}|I) = P(\mathbf{\Theta}|I)\, P(\mathbf{B}|I). \tag{18}$$

This is clearly not true in general. For example, if one has a previous measurement of a sinusoid, knowledge of the frequencies would imply a great deal about the amplitudes. So if cogent prior information is available it should be incorporated into the prior $P(\mathbf{\Theta}, \mathbf{B}|I)$. Having none, it will be assumed that the data determine the parameters much more precisely than the prior information. Therefore the prior information is vague and uninformative. If the prior does not vary appreciably over the range of parameter values indicated by the data, the conclusions are nearly independent of what the prior does outside that range. So, the exact functional form of the prior is relatively unimportant, as long as it does not express a strong opinion about the parameters.

To assign the prior probability for the amplitudes, one must state exactly what is known. In this problem what is known is that the signal, like the noise, carried a finite but unknown total power. But this information is enough to use maximum entropy to assign a prior probability to the amplitudes. This prior is again a Gaussian for exactly the same reason that the noise prior was a Gaussian. It has been assumed that the data determine the parameters much better than the prior, so the variance on the prior for the amplitudes is assumed to be very large. Therefore, over the region where the direct probability of the data is peaked, this prior will look like a uniform prior. In the limit, as the variance of the Gaussian goes to infinity, the Gaussian goes into a uniform prior and expresses no opinion about the parameters. This assumption will be made, and the Gaussian will be replaced by its limiting form.

This unbounded uniform prior is called an improper prior probability and strictly speaking is not a probability distribution at all; rather it is the limit of a sequence of proper priors, in the limit of infinite uncertainty in the parameters. Care must be exercised when using improper priors, because they effectively introduce an infinity into the calculation. In parameter estimation calculations this presents no problem, because the infinity always cancels when the distribution is normalized. This is not true in model selection calculations, and improper priors cannot be used [8].

The assignment of the prior for the frequencies and decay rate constants is simpler, because these parameters may be bounded. In dimensionless units the frequencies must take on values $-\pi \leq \omega_j \leq +\pi$, and surely in dimensionless units the decay rate constants may be bounded by $0 \leq \alpha_j < N$. If nothing else is known about them except the bounds, then maximum entropy will assign uniform priors probability density functions to these parameters.

With this choice of the prior, the joint posterior probability of the amplitudes, and nonlinear $\mathbf{\Theta}$ parameters given the variance of the noise, is just the original likelihood function

$$P(\mathbf{B},\mathbf{\Theta}|\sigma,D,I) \propto (2\pi\sigma^2)^{-N}\exp\{-\frac{Q}{2\sigma^2}\}, \tag{19}$$

where the standard deviation of the noise σ has been added to the posterior probability of the parameters to indicate that it is known. If σ is not known, then using the product and sum rules, it may also be removed from the problem.

This is the joint posterior probability for all of the parameters including the amplitudes. To remove the amplitudes, the sum rule from probability theory is applied: integrate the joint posterior probability density with respect to the unwanted parameters

$$P(\mathbf{\Theta}|\sigma,D,I) \propto \int_{-\infty}^{\infty}(2\pi\sigma^2)^{-N}\exp\left\{-\frac{Q}{2\sigma^2}\right\}d\mathbf{B}. \tag{20}$$

These integrals are multivariant Gaussian integrals and any multivariant integral of this form may be done analytically. To do the integral a change of variables is introduced,

$$B_k = \sum_{j=1}^{m}\frac{A_j e_{jk}}{\sqrt{\lambda_j}}, \quad \text{with} \quad A_k = \sqrt{\lambda_k}\sum_{j=1}^{m}B_j e_{kj}, \tag{21}$$

$$d\mathbf{B} \equiv dB_1\cdots dB_m = \lambda_1^{-\frac{1}{2}}\cdots\lambda_m^{-\frac{1}{2}}dA_1\cdots dA_m; \tag{22}$$

and a change of function is introduced:

$$R_k = \sum_{j=1}^{m}\frac{U_j e_{kj}}{\sqrt{\lambda_k}}, \quad \text{with} \quad f_R(t) = \sum_{j=1}^{m}A_j R_j(t), \tag{23}$$

$$I_k = \sum_{j=1}^{m}\frac{V_j e_{kj}}{\sqrt{\lambda_k}}, \quad \text{with} \quad f_I(t) = \sum_{j=1}^{m}A_j I_j(t), \tag{24}$$

where e_{jk} is the kth component of the jth eigenvector of the interaction matrix

$$g_{jk} \equiv \sum_{i=1}^{N}U_j(t_i)U_k(t_i) + V_j(t_i)V_k(t_i), \tag{25}$$

and λ_j is its jth eigenvalue. Then using the property

$$\sum_{i=1}^{N}R_j(t_i)R_k(t_i) + I_j(t_i)I_k(t_i) = \delta_{jk}, \tag{26}$$

called orthonormality, the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters Eq. (20) becomes

$$P(\mathbf{\Theta}|\sigma, D, I) \propto \sigma^{-2N}\lambda_1^{-\frac{1}{2}}\cdots\lambda_m^{-\frac{1}{2}}\int_{-\infty}^{\infty} dA_1\cdots dA_m \exp\left\{-\frac{Q'}{2\sigma^2}\right\}, \tag{27}$$

where

$$Q' = d_R\cdot d_R + d_I\cdot d_I - 2\sum_{j=1}^{m} A_j h_j + \sum_{j=1}^{m} A_j^2 \tag{28}$$

and

$$h_j \equiv d_R\cdot R_j + d_I\cdot I_j, \tag{29}$$

and some irrelevant numerical constants have been dropped. After completing the square and performing the m integrals one obtains the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters,

$$P(\mathbf{\Theta}|\sigma, D, I) \propto \sigma^{m-2N}\lambda_1^{-\frac{1}{2}}\cdots\lambda_m^{-\frac{1}{2}}\exp\left\{-\frac{d_R\cdot d_R + d_I\cdot d_I - m\overline{h^2}}{2\sigma^2}\right\}, \tag{30}$$

where

$$\overline{h^2} \equiv \frac{1}{m}\sum_{j=1}^{m} h_j^2 \tag{31}$$

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

If the variance of the noise σ^2 is known (as assumed so far), then the problem is completed. The posterior probability of the frequencies and decay rate constants, conditional on the data and the assumed knowledge of σ, is

$$P(\mathbf{\Theta}|\sigma, D, I) \propto \lambda_1^{-\frac{1}{2}}\cdots\lambda_m^{-\frac{1}{2}}\exp\left\{\frac{m\overline{h^2}}{2\sigma^2}\right\}. \tag{32}$$

But if σ is not known, then it too becomes a nuisance parameter to be removed by integration. Now σ is a scale parameter and is restricted to positive values. The completely uninformative prior probability of a scale parameter is the Jeffreys' prior $1/\sigma$ [15]. A maximum entropy derivation of this prior is beyond the scope of this paper, but one may be found in [16]. Like the unbounded uniform prior, the Jeffreys' prior is also an improper prior, and as was mentioned earlier, improper priors are not strictly speaking probability density functions at all; rather they are the limiting values of a sequence of proper priors and convey no information about the numerical value of a parameter. Their use in parameter estimation problems is helpful, because they make the mathematics easier; however, they cannot be used in model selection problems.

Multiplying the posterior probability of the nonlinear parameters Eq. (30) by a Jeffreys' prior for the standard deviation, one obtains the joint probability of the nonlinear parameters and the standard deviation. Integrating the joint posterior probability with respect to the standard deviation, σ, one obtains the posterior probability of the frequencies and decay rate constants independent of the amplitudes and variance of the noise:

$$P(\mathbf{\Theta}|D, I) \propto \lambda_1^{-\frac{1}{2}}\cdots\lambda_m^{-\frac{1}{2}}\left[1 - \frac{m\overline{h^2}}{d_R\cdot d_R + d_I\cdot d_I}\right]^{\frac{m-2N}{2}}. \tag{33}$$

The quantity $\overline{h^2}$ plays the role of a sufficient statistic and summarizes all of the information in the data for inferences about the nonlinear $\mathbf{\Theta}$ parameters. The posterior probability of the frequencies and decay rates is given by Eq. (32) if the variance of the noise is known, and by Eq. (33) if the variance of the noise is unknown. These equations are exact results and do not depend on uniform sampling, the data need not be a time series, and the models need not be sinusoidal. Any quadrature data set that can be modeled by Eqs. (1) and (2) having a signal of the form of Eqs. (3) and (4) can be used in these equations.

Expectations

In order to recover the model signal, Eqs. (3) and (4), an estimate of the amplitudes **B** is required. To estimate the accuracy of these parameters, the posterior covariances are needed, and an estimate of the variance σ^2 would be extremely useful in deciding the adequacy of the model. In general, the integrals over the nonlinear $\mathbf{\Theta}$ parameters cannot be performed exactly. Thus, a way to estimate the accuracy of these parameters must be devised. This presents only a minor problem, because the logarithm of the posterior probability is proportional to N, the number of data values. When the number of data values is large, the posterior probability will be so sharply peaked that good approximate results will be readily available.

The expected orthonormal amplitudes $\langle A_j \rangle$ are given by

$$\begin{aligned} E(A_j|\mathbf{\Theta},\sigma,D,I) = \langle A_j \rangle &= \frac{\displaystyle\int_{-\infty}^{\infty} A_j \exp\left\{-\frac{A_j^2 - 2h_j A_j}{2\sigma^2}\right\} dA_j}{\displaystyle\int_{-\infty}^{\infty} \exp\left\{-\frac{A_j^2 - 2h_j A_j}{2\sigma^2}\right\} dA_j} \\ &= h_j \end{aligned} \tag{34}$$

and the expected value of the amplitudes **B** are

$$E(B_k|\mathbf{\Theta},\sigma,D,I) = \langle B_k \rangle = \sum_{j=1}^{m} \frac{h_j e_{jk}}{\sqrt{\lambda_j}}. \tag{35}$$

These are the explicit Bayesian estimates for the amplitudes B_j. They are still functions of (i.e., , they are conditional on) the nonlinear $\mathbf{\Theta}$ parameters, and to remove this dependence one should in principle multiply these estimates by the posterior probability of the frequencies and decay rate constants and integrate over all remaining parameters. But if the data actually determine the frequencies and decay rates well, evaluating the amplitude estimates, Eq. (35), at the maximum of the posterior probability, Eq. (33), is nearly the same as computing these expectation values.

The posterior covariances of the orthonormal amplitudes **A** are easily computed. These are given by

$$\langle A_k A_l \rangle - \langle A_k \rangle \langle A_l \rangle = \sigma^2 \delta_{kl}, \tag{36}$$

where δ_{kl} is the Kronecker delta function. The posterior covariances for the amplitudes **B** are given by

$$\langle B_k B_l \rangle - \langle B_k \rangle \langle B_l \rangle = \sigma^2 \sum_{j=1}^{m} \frac{e_{jk} e_{jl}}{\lambda_j}. \tag{37}$$

From the posterior covariances the (mean) ± (standard deviation) estimate of the amplitudes **B** can be made. Defining δ_k^2 as

$$\delta_k^2 \equiv \langle B_k^2 \rangle - \langle B_k \rangle^2 = \sigma^2 \sum_{j=1}^{m} \frac{e_{jk}^2}{\lambda_j}, \tag{38}$$

one may estimate B_k to a precision of

$$(B_k)_{\text{est}} = \langle B_k \rangle \pm \delta_k \tag{39}$$

at one standard deviation. If one compares these amplitude estimates to those obtained from maximum likelihood, one will find they are the same; and the accuracy estimates are nearly the same as the standard error calculation typically given for maximum likelihood and least squares. This is not surprising; this calculation started with the same direct probability as least squares and maximum likelihood and did not include any additional prior information. Therefore, the direct probability must determine the parameters, and it has. However, there are significant differences for the nonlinear parameters and in the interpretation of these results. Additionally, this calculation can be extended to the model selection problem [ref7,8] which neither least squares nor maximum likelihood can approach.

An estimate of the variance of the noise is helpful in judging the adequacy of the model. This is given by

$$E(\sigma^2|\mathbf{\Theta},D,I)=\langle\sigma^2\rangle=\left[\frac{d_R\cdot d_R+d_I\cdot d_I-m\overline{h^2}}{2N-m-2}\right]. \tag{40}$$

As indicated earlier, everything probability theory cannot fit to the data is placed into the noise. The Gaussian noise prior is the least informative prior probability density that is consistent with the finite average noise power assumption.

Unlike the amplitudes $\mathbf{B}$ and the variance of the noise σ^2, one cannot calculate the expectation values of the nonlinear $\mathbf{\Theta}$ parameters analytically; that is, the integrals represented by

$$\langle\Theta_j\rangle=\int d\mathbf{\Theta}\;\Theta_j\;P(\mathbf{\Theta}|D,I) \tag{41}$$

cannot be done exactly. Nonetheless an estimate of these parameters and their probable accuracy must be obtained.

These estimates may be obtained by first noting that the joint posterior density is Eq. (30) when σ is known and Eq. (33) when it is not. But they are not very different provided *we have enough data for good estimates.* Writing the maximum attainable $\sum h_j^2$ as

$$\left(\sum_{j=1}^{m}h_j^2\right)_{max}=x \tag{42}$$

and writing the difference from the maximum as q^2, one has

$$\sum_{j=1}^{m}h_j^2=x-q^2 \tag{43}$$

and Eq. (33) becomes

$$\left[d_R\cdot d_R+d_I\cdot d_I-x+q^2\right]^{\frac{m-2N}{2}}\approx\exp\left\{-\frac{(2N-m)q^2}{2(d_R\cdot d_R+d_I\cdot d_I-x)}\right\}, \tag{44}$$

where the slowly varying Jacobian has been dropped. But this is nearly the same as

$$\left[d_R\cdot d_R+d_I\cdot d_I-x+q^2\right]^{\frac{m-2N}{2}}\approx\exp\left\{-\frac{q^2}{2\langle\sigma^2\rangle}\right\}, \tag{45}$$

where $\langle\sigma^2\rangle$ is Eq. (39) and is evaluated for the values $\hat{\mathbf{\Theta}}\equiv\{\hat{\Theta}_1,\cdots,\hat{\Theta}_r\}$ that maximize the posterior probability. Up to an irrelevant normalization constant, the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters around the location of the maximum posterior probability is given approximately by

$$P(\mathbf{\Theta}|\langle\sigma^2\rangle,D,I)\approx\exp\left\{\frac{m\overline{h^2}}{2\langle\sigma^2\rangle}\right\}, \tag{46}$$

where the slightly inconsistent notation $P(\mathbf{\Theta}|\langle\sigma^2\rangle,D,I)$ has been adopted to remind one that $\langle\sigma^2\rangle$ has been used, not σ^2. Of course if σ^2 is actually known then $\langle\sigma^2\rangle$ should be replaced by σ^2 in what follows.

Expanding $\overline{h^2}$ in a Taylor series around $\hat{\mathbf{\Theta}}$ one obtains

$$P(\mathbf{\Theta}|\langle\sigma^2\rangle, D, I) \propto \exp\left\{-\sum_{jk=1}^{r} \frac{b_{jk}}{2\langle\sigma^2\rangle}\Delta_j\Delta_k\right\}, \tag{47}$$

where

$$\begin{aligned} b_{jk} &\equiv -\frac{m}{2}\frac{\partial^2\overline{h^2}}{\partial\Theta_j\partial\Theta_k}\bigg|_{\hat{\mathbf{\Theta}}}, \\ \Delta_j &\equiv \hat{\Theta}_j - \Theta_j, \end{aligned} \tag{48}$$

from which the (mean) ± (standard deviation) approximations for the $\mathbf{\Theta}$ parameters can be made.

These estimates are obtained by computing the posterior covariances of the nonlinear $\mathbf{\Theta}$ parameters. These Gaussian integrals are evaluated by first changing to orthogonal variables and then performing the n integrals. The new variables are obtained from the eigenvalues and eigenvectors of b_{jk}. Let u_{jk} denote the kth component of the jth eigenvector of b_{jk} and let v_j be the eigenvalue. The orthogonal variables are given by

$$s_j = \sqrt{v_j}\sum_{k=1}^{r}\Delta_k u_{jk}, \qquad \Delta_j = \sum_{k=1}^{r}\frac{s_k u_{kj}}{\sqrt{v_k}}. \tag{49}$$

Making this change of variables, one has

$$P(\{s\}|\langle\sigma^2\rangle, D, I) \propto v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}}\exp\left\{-\sum_{j=1}^{r}\frac{s_j^2}{2\langle\sigma^2\rangle}\right\}. \tag{50}$$

From the local approximation of the posterior probability Eq. (50), the expected values of $\langle s_j\rangle$ and $\langle s_j^2\rangle$ can be computed. Of course $\langle s_j\rangle$ is zero and the expectation value $\langle s_j s_k\rangle$ is given by

$$\begin{aligned} \langle s_j s_k\rangle &= \frac{\displaystyle\int_{-\infty}^{\infty} ds_1\cdots ds_r v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}} s_j s_k \exp\left\{-\sum_{l=1}^{r}\frac{s_l^2}{2\langle\sigma^2\rangle}\right\}}{\displaystyle\int_{-\infty}^{\infty} ds_1\cdots\cdot ds_r v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}} \exp\left\{-\sum_{l=1}^{r}\frac{s_l^2}{2\langle\sigma^2\rangle}\right\}}, \\ &= \langle\sigma^2\rangle\delta_{kj} \end{aligned} \tag{51}$$

where δ_{kj} is a Kronecker delta function. The posterior covariances of the nonlinear $\mathbf{\Theta}$ parameters are given by

$$\langle\Theta_j\Theta_k\rangle - \langle\Theta_j\rangle\langle\Theta_k\rangle = \langle\sigma^2\rangle\sum_{l=1}^{r}\frac{u_{lj}u_{lk}}{v_l}. \tag{52}$$

Defining γ_k^2 as

$$\gamma_k^2 \equiv \langle\sigma^2\rangle\sum_{j=1}^{r}\frac{u_{jk}^2}{v_j}, \tag{53}$$

the (mean) ± (standard deviation) estimate of the nonlinear $\mathbf{\Theta}$ parameters is

$$(\Theta_j)_{\text{est}} = \hat{\Theta}_j \pm \gamma_j \tag{54}$$

at one standard deviation.

For an arbitrary model, the matrix b_{jk} cannot be calculated analytically; however, it can be evaluated numerically. The accuracy estimates of the nonlinear $\mathbf{\Theta}$ parameters and the amplitudes

B depend explicitly on the estimated noise variance. But the estimated variance is the mean-square difference between the model and the data. If the misfit is large, the variance is estimated to be large and the accuracy is estimated to be poor. Thus, when one states that the parameter estimates are conservative it is implied that, because everything probability theory cannot fit to the model is assigned to the noise, all of the parameter estimates are as wide as is consistent with the model and the data.

Discussion

Consider the data vector $d_i = \{d_R(t_1), \cdots, d_R(t_N), d_I(t_1), \cdots, d_I(t_N)\}$ which is to be approximated by the orthonormal functions $H_j(t_i)$

$$H_j \equiv \begin{cases} R_j(t_i), & 1 \le i \le N \\ I_j(t_{i-N}), & N < i \le 2N \end{cases} \tag{55}$$

then

$$\begin{aligned} d_i = &\ f(t_i) + \text{error} \\ = &\ \textstyle\sum_{j=1}^m A_j H_j(t_i) + \text{error} \qquad (1 \le i \le 2N). \end{aligned} \tag{56}$$

What choice of $\{A_1, \cdots, A_m\}$ is "best"? If the criterion of "best" is the mean-square error, then

$$\begin{aligned} 0 &\le \sum_{i=1}^{2N} \left[d_i - \sum_{j=1}^{m} A_j H_j(t_i) \right]^2 \\ &= d_R \cdot d_R + d_I \cdot d_I + \sum_{j=1}^{m} (A_j^2 - 2A_j h_j) \\ &= d_R \cdot d_R + d_I \cdot d_I - m\overline{h^2} + \sum_{j=1}^{m} (A_j - h_j)^2, \end{aligned} \tag{57}$$

where the sufficient statistic Eq. (29) and the orthonormality Eq. (26) were used. Evidently, the "best" choice of the coefficients is

$$A_j = h_j \qquad (1 \le j \le m) \tag{58}$$

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

$$d_R \cdot d_R + d_I \cdot d_I - m\overline{h^2} \ge 0 \tag{59}$$

with equality if, and only if, the approximation is perfect.

The Bessel inequality gives the following intuitive picture of the meaning of the modeling process Eqs. (20) thru (33). The quadrature data comprise a vector in a $2N$-dimensional linear vector space S_{2N}. The signal functions U_j and V_j define a $2N$ dimensional vector $G_j(t_i)$:

$$G_j(t_i) \equiv \begin{cases} U_j(t_i), & 1 \le i \le N \\ V_j(t_{i-N}), & N < i \le 2N \end{cases}, \tag{60}$$

where

$$d(t_i) = \sum_{j=1}^{m} B_j G_j(t_i) + e(t_i) \qquad (1 \le i \le 2N). \tag{61}$$

The orthonormal model functions H_j are related to G_j by the transformation

$$H_k \equiv \sum_{j=1}^{m} \frac{G_j e_{kj}}{\sqrt{\lambda_k}}, \tag{62}$$

and the amplitudes A_j are related to B_j by the change of variables

$$B_k = \sum_{j=1}^{m} \frac{A_j e_{jk}}{\sqrt{\lambda_j}}, \quad A_k = \sqrt{\lambda_k} \sum_{j=1}^{m} B_j e_{kj}, \tag{63}$$

where e_{jk} and λ_j are the eigenvectors and eigenvalues of the matrix g_{jk} which may be written as

$$g_{jk} = \sum_{i=1}^{2N} G_j(t_i) G_k(t_i). \tag{64}$$

This is just the parameter estimation problem for a nonquadrature data set with $2N$ data values [1,2]. Apparently parameter estimation using quadrature data, when the variance of the noise is the same in the two channels, is just a special case of nonquadrature parameter estimation.

The orthogonal model equation

$$d_i = \sum_{j=1}^{m} A_j H_j(t_i) + e(t_i) \qquad (1 \leq i \leq 2N) \tag{65}$$

expresses the assumption that these data can be separated into a "systematic part" $f(t_i)$ and a "random part" $e(t_i)$. Estimating the parameters of interest $\mathbf{\Theta}$ that are hidden in the model functions $H_j(t)$ amounts essentially to finding the values of $\mathbf{\Theta}$ that permit $f(t)$ to make the closest possible fit to the data by the mean-square criterion. Put differently, probability theory recognizes that the most likely values of $\mathbf{\Theta}$ are those that allow a maximum amount of the mean-square data to be accounted for by the systematic term; from the Bessel inequality Eq. (59) those are the values that *maximize* $\overline{h^2}$.

However, there are $2N$ 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_{2N}. What kind of data should one then expect?

Consider the problem backward for a moment. Suppose one knows (never mind how one could know this) that the model is correct, and one also knows the true values of all the model parameters $(\mathbf{A}, \mathbf{\Theta}, \sigma)$ – call this the Utopian state of knowledge U – but one does not know what data will be found. Then, the probability density that would be assigned to any particular data set $D = \{d_1, \cdots, d_{2N}\}$ is just the sampling distribution

$$P(D|U) = (2\pi\sigma^2)^{-N} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{2N} [d_i - f(t_i)]^2 \right\}. \tag{66}$$

From this the expectations and covariances of the data can be found:

$$\begin{aligned} E(d_i|U) = \langle d_i \rangle &= f(t_i) \qquad (1 \leq i \leq 2N) \\ \langle d_i d_j \rangle - \langle d_i \rangle \langle d_j \rangle &= (2\pi\sigma^2)^{-N} \int d^{2N}x \; x_i x_j \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{2N} x_i^2 \right\} \\ &= \sigma^2 \delta_{ij}; \end{aligned} \tag{67}$$

therefore one would "expect" to see a mean-square data value $\overline{d^2}$ of about

$$\begin{aligned} E(\overline{d^2}|U) = \langle \overline{d^2} \rangle &= \frac{1}{2N} \sum_{i=1}^{2N} \langle d_i^2 \rangle \\ &= \frac{1}{2N} \sum_{i=1}^{2N} (\langle d_i \rangle^2 + \sigma^2) \\ &= \sigma^2 + \frac{1}{2N} \sum_{i=1}^{2N} f^2(t_i). \end{aligned} \tag{68}$$

But from the orthonormality Eq. (26) of the $H_j(t_i)$, one can write

$$\begin{aligned}\sum_{i=1}^{2N} f^2(t_i) &= \sum_{l=1}^{2N} \sum_{j,k=1}^{m} A_j A_k H_j(t_i) H_k(t_i) \\ &= \sum_{j=1}^{m} A_j^2.\end{aligned} \tag{69}$$

So that the expected mean-square data value Eq. (68) becomes

$$\langle\overline{d^2}\rangle = \frac{m}{2N}\overline{A^2} + \sigma^2, \tag{70}$$

where $\overline{A^2}$ is the mean-square amplitude: $\overline{A^2} \equiv \sum_{j=1}^{m} A_j^2/m$. Now, what value of $\overline{h^2}$ does one expect the data to generate? This expectation is given by

$$\begin{aligned}E(\overline{h^2}|U) = \langle\overline{h^2}\rangle &= \frac{1}{m}\sum_{j=1}^{m}\langle h_j^2\rangle \\ &= \frac{1}{m}\sum_{j=1}^{m}\sum_{i,k=1}^{2N}\langle d_i d_k\rangle H_j(t_i)H_j(t_k) \\ &= \frac{1}{m}\sum_{j=1}^{m}\sum_{i,k=1}^{2N}(\langle d_i\rangle\langle d_k\rangle + \sigma^2\delta_{ik})H_j(t_i)H_j(t_k).\end{aligned} \tag{71}$$

But

$$\begin{aligned}\sum_{i=1}^{2N}\langle d_i\rangle H_j(t_i) &= \sum_{i=1}^{2N}\sum_{l=1}^{m} A_l H_l(t_i) H_j(t_i) \\ &= \sum_{l=1}^{m} A_l \delta_{lj} \\ &= A_j\end{aligned} \tag{72}$$

and the expected value of the sufficient statistic Eq. (71) reduces to

$$\langle\overline{h^2}\rangle = \overline{A^2} + \sigma^2. \tag{73}$$

So, given the Utopian state of knowledge, one expects the left-hand side of the Bessel inequality Eq. (59) to be approximately

$$2N\langle\overline{d^2}\rangle - m\overline{h^2} \approx (2N-m)\sigma^2. \tag{74}$$

This agrees very nicely with one's intuitive judgment that as the number of model functions increases, one should be able to fit the data better and better. Indeed, when $m = 2N$, the functions $H_j(t_i)$ become a complete orthonormal set on S_{2N}, and the data can always be fitted exactly, as Eq. (74) suggests.

If σ is known, these results give a simple diagnostic test for judging the adequacy of the model. Having taken the data, calculate $(d_R \cdot d_R + d_I \cdot d_I - m\overline{h^2})$. If the result is reasonably close to $(2N-m)\sigma^2$, 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 $(d_R \cdot d_R + d_I \cdot d_I - m\overline{h^2})$ turns out to be much larger than $(2N-m)\sigma^2$, the model is not fitting the data as well as it should: it is "underfitting" the data. This would be evidence that the model is inadequate to represent the data in the sense that more model functions or different model functions might be required or the 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 $(d_R \cdot d_R + d_I \cdot d_I - m\overline{h^2})$ is far less than $(2N - m)\sigma^2$; then the model is "overfitting" the data. This would be evidence either that the 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 the apparent fit can always be improved, but if the model functions represent more detail than is really in the systematic effects, part of this fit is misleading: one is *fitting the noise.*

A test to confirm this would be to systematically remove each parameter from the model and then, using Bayesian probability theory, calculate the probability of these models. The posterior probability of the models could then be compared to determine which parameters are spurious. The procedures for calculating the probability of the models have not yet been discussed and are the subject of the next paper [7]. There it is demonstrated that Bayesian probability theory contains a quantitative statement of Ockham's razor: When two models fit the data equally well, prefer the simpler model.

Consider now the case that σ is completely unknown, where probability theory led to the student t distribution, Eq. (33). Integrating over a nuisance parameter is very much like estimating the parameter from the data, and then using that estimate in our equations. If the parameter is actually well determined by the data, the two procedures are essentially equivalent. The expectation value of the variance $\langle\sigma^2\rangle$ is given by

$$\langle\sigma^2\rangle = \left[\frac{d_R \cdot d_R + d_I \cdot d_I - \sum_{j=1}^{m} h_j^2}{2N - m - 2}\right]. \qquad (75)$$

Constraining σ to this value, the posterior probability of the $\mathbf{\Theta}$ parameters is approximately

$$P(\mathbf{\Theta}|\langle\sigma^2\rangle, D, I) \approx \exp\left\{\frac{m\overline{h^2}}{2\langle\sigma^2\rangle}\right\}, \qquad (76)$$

just the approximation derived earlier. In effect, probability theory apportions the first m degrees of freedom to the signal, the next two to the variance, and the remaining $(2N - m - 2)$ should be noise degrees of freedom. Thus, as already stressed above, everything probability theory cannot fit to the signal will be placed in the noise, automatically.

More interesting is the opposite extreme when the student t distribution, Eq. (33), approaches a singular value. Consider the following scenario. One has obtained some data which are recorded automatically to six figures. But one has no prior knowledge of the accuracy of those data; in fact, σ may be such that all of the data are essentially noise. The data are plotted to determine a model function that best fits them. Suppose, for simplicity, that the model function is a stationary sinusoid. On plotting d_i against i, the data are seen to fall exactly on the sinusoid (i.e., to within the six figures given). What conclusions does one draw from this?

Intuitively, one would think that the data must be far "better" than had been thought; one feels sure that σ must be less than one part in a thousand, and that one is therefore able to estimate the frequency to an accuracy considerably better than a part in a thousand, if the number of data values $2N$ is large. It may, however, be hard to see at first glance how probability theory can justify this intuitive conclusion that one draws so easily.

But that is just what the parameter estimates Eqs. (39) and (54) tell one; Bayesian analysis leads one to it automatically and for any model functions. Even though one 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, Eq. (59), it follows almost certainly that σ^2 must be extremely small and, if the other parameters are independent, they can all be estimated almost exactly.

Summary and Conclusions

In high resolution NMR a great deal of prior information is available. This information can be used to great advantage when analyzing NMR data. Here the Bayesian data analysis calculation, originally done in [9], has been expanded to include the quadrature nature of NMR data. The calculation gives a simple intuitive picture of quadrature model fitting. The data may be thought of as a $2N$ dimensional vector, and the model may be thought of as an m dimensional vector. The frequencies and decay rate constants estimated are those which allow the model to make the closest approach to the data by the mean-square criterion. Although a simple picture of the model fitting process emerges from the calculation, what is to be gained by using probability theory has not yet been demonstrated. In later articles [7,8] the calculation is extended to the problem of determining the optimal model, and the use of these techniques on FID data using sinusoidal models is demonstrated. Examples of parameter estimation, signal detection, and model selection are given.

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.

References

[1] E. T. Jaynes, Stanford University Microwave Laboratory Report No. 421, 1957, reprinted in "Maximum-Entropy and Bayesian Methods in Science and Engineering" (G. J. Erickson and C. R. Smith, Eds.), Vol. 1, p. 1, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[2] R. T. Cox, "The Algebra of Probable Inference," Johns Hopkins Univ. Press, Baltimore, 1961.

[3] M. Tribus, "Rational Descriptions, Decisions and Designs," Pergamon Press, Elmsford, New York, 1969.

[4] E. T. Jaynes, " Papers on Probability, Statistics and Statistical Physics" (R. D. Rosenkrantz, Ed.), D. Reidel, Dordrecht, The Netherlands, 1983.

[5] G. L. Bretthorst, J. J. Kotyk, and J. J. H. Ackerman, *Magn. Reson. Med.* **9,** 282 (1989).

[6] G. L. Bretthorst, Chi-Cheng Hung, D. d'Avignon, and J. J. H. Ackerman, *J. Magn. Reson.* **79,** 369 (1988).

[7] G. L. Bretthorst, *J. Magn. Reson.* **88,** pp. 552-570 (1990).

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

[9] G. L. Bretthorst, "Bayesian Spectrum Analysis and Parameter Estimation," Ph.D. thesis, Washington University, St. Louis, Missouri, available from University Microfilms Inc., Ann Arbor, Michigan, 1987; an excerpt is printed in "Maximum-Entropy and Bayesian Methods in Science and Engineering" (G. J. Erickson and C. R. Smith, Eds.), Vol. 1, p. 75, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

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

[11] Rev. T. Bayes, *Philos. Trans. R. Soc. London* **53,** 370 (1763); reprinted in *Biometrika* **45,** 293 (1958), and "Facsimiles of Two Papers by Bayes," with commentary by W. Edwards Deming Hafner, New York, 1963.

[12] J. E. Shore, and R. W. Johnson, *IEEE Trans. on Inf. Theory,* **IT-26**(1), 26 1980.

[13] J. E. Shore, and R. W. Johnson, *IEEE Trans. on Inf. Theory* **IT-27**(1), 472 1981.

[14] C. E. Shannon, *Bell Syst. Tech. J.* **27,** 379 (1948).

[15] H. Jeffreys, "Theory of Probability," Oxford University Press, London 1939.

[16] A. Zellner, "An Introduction to Bayesian Inference in Econometrics," Wiley, New York, 1971.