Journal of Magnetic Resonance **88,** 571-595 (1990)

Bayesian Analysis. III. Applications to NMR Signal Detection, Model Selection and Parameter Estimation [1]

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

ABSTRACT. The two preceding articles developed the application of Bayesian probability theory to the problems of parameter estimation, signal detection, and model selection on quadrature NMR data in some generality. Here those procedures are used to analyze free induction decay data, when the models are sinusoidal. The exact relationship between Bayesian probability theory and the discrete Fourier-transform power spectrum is derived, and it is shown that the discrete Fourier-transform power spectrum is an optimal frequency estimator for a wide class of problems. Signal detection and model selection problems are then examined, and examples are given that demonstrate the ability of Bayesian probability theory to determine the best model of a process even when more complex models fit the data better.

Introduction

The traditional way to analyze a free induction decay is to treat the data as the real and imaginary parts of a complex data set, and then perform a discrete Fourier transform on the complex data. The frequencies and amplitudes are then estimated from the real part of the discrete Fourier transform. The real part of the discrete Fourier transform is called an *absorption spectrum.* Before taking the transform, zeros are often added to the end of the complex data to make a new, longer data set. Zero-"padding" (or "filling") the data and then performing a discrete Fourier transform essentially allows one to evaluate the Schuster periodogram [3] at smaller frequency intervals. Zero-padding does improve frequency resolution. A typical reason given for this is that, the absorption spectrum contains only half of the total information in the data and zero-padding to $2N$ data values allows one to recover the other half of the information [4]. Additionally, the complex time domain data are often multiplied by an apodization function, typically a decaying exponential (line broadening); this filter removes high-frequency oscillations from the absorption spectrum, suppresses side lobes, and increases the signal-to-noise ratio in the absorption spectrum.

Bayesian analysis of the *single* stationary sinusoidal frequency model and the *multiple well-separated* stationary sinusoidal frequency model will lead to a discrete Fourier-transform *power spectrum* as an optimal statistic for the estimation of multiple well-separated stationary frequencies. This calculation (given shortly) places the discrete Fourier transform in a new light and shows that under a variety of conditions no better estimate of the frequencies may be obtained than from the peak values of a zero-padded discrete Fourier-transform *power spectrum.* Additionally, Bayesian analysis of the single exponentially decaying sinusoidal model leads to a discrete Fourier-transform power spectrum of the complex data multiplied by an appropriate exponentially decaying apodizing function as the proper statistic for estimating the frequency and decay rate when the data contain a single exponentially decaying sinusoid. Thus, the discrete Fourier-transform *power spectrum* will

[1] All of the computer codes and the data used in these three papers are available from the author.

be shown to be the proper statistic for frequency estimation under a variety of conditions, and the common practices of zero-padding and exponential apodizing will be interpreted in an entirely new way.

Multiple Stationary Frequency Estimation

Suppose the data to be analyzed is an FID with a single resonance and the signal was sampled so fast and one is far enough off-resonance that, to first approximation, the signal may be considered a stationary (nondecaying) sinusoid. How would one use the procedures developed in the previous paper [1] to determine the resonance frequency? To answer this question, one first states exactly what prior information (in the form of a model) is to be incorporated into the calculation. As was discussed previously [1], for quadrature data there are two data sets, each 90° out of phase. The real data d_R has a model of the form

$$d_R(t_i) = f_R(t_i) + e(t_i) \quad (1 \leq i \leq N), \tag{1}$$

and

$$d_I(t_i) = f_I(t_i) + e(t_i) \quad (1 \leq i \leq N) \tag{2}$$

for the imaginary data, where f_R and f_I define what is meant by the signal and $e(t_i)$ is a random noise component at time t_i. The quadrature model for a single stationary sinusoid may be written as

$$f_R(t) = \sum_{j=1}^{2} B_j U_j(t) = B_1 \sin(\omega t) + B_2 \cos(\omega t), \tag{3}$$

for the real channel, and

$$f_I(t) = \sum_{j=1}^{2} B_j V_j(t) = B_1 \cos(\omega t) - B_2 \sin(\omega t) \tag{4}$$

for the imaginary channel, where the parameters B_1 and B_2 are effectively the amplitude and phase of the sinusoid, and ω is a parameter representing the frequency and is to be estimated from the data. For the purposes of analyzing the data for resonances, only the frequency ω is of interest. The other two parameters, B_1 and B_2, will be considered as nuisances and the problem will be formulated independent of them.

To compute the posterior probability of the frequency, one computes the g_{jk} matrix defined as

$$g_{jk} = \sum_{i=1}^{N} U_j(t_i)U_k(t_i) + V_j(t_i)V_k(t_i). \tag{5}$$

For this model, when the data are uniformly sampled, this matrix is particularly simple:

$$g_{jk} = \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix}, \tag{6}$$

where dimensionless units have been used. In these units the frequency ω takes on values ranging from $-\pi \leq \omega \leq \pi$, and the time increments take on values $t_i = \{0, 1, \cdots, N-1\}$, where N is the total number of data values in one channel. The statistic $\overline{h^2}$, Eq. (31) from Ref. 1, is then given by

$$\overline{h^2} = \frac{1}{2N}\left\{[C_R(\omega) + S_I(\omega)]^2 + [S_R(\omega) - C_I(\omega)]^2\right\}, \tag{7}$$

where

$$C_R(\omega) \equiv \sum_{i=1}^{N} d_R(t_i)\cos(\omega t_i), \tag{8}$$

and

$$S_R(\omega) \equiv \sum_{i=1}^{N} d_R(t_i)\sin(\omega t_i) \tag{9}$$

are the cosine and sine transforms of the real data, and $C_I(\omega)$ and $S_I(\omega)$ are the transforms for the imaginary data. The quantity $\overline{h^2}$ summarizes *all of the information* in the data relevant to the problem of estimating the value of a single stationary frequency, and is called a *sufficient statistic.* The sufficient statistic, Eq. (7), is then substituted into posterior probability of the nonlinear $\mathbf{\Theta}$ parameters, Eq. (32) from Ref. 1, if the variance of the noise is unknown, or into Eq. (33) from Ref. 1 if the variance of the noise is known. For this demonstration, the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters given the variance of the noise, Eq. (32) from Ref. 1, will be used. The posterior probability of a single stationary sinusoidal frequency is given by

$$P(\omega|\sigma, D, I) \propto \exp\left\{\frac{\overline{h^2}}{\sigma^2}\right\} = \exp\left\{\frac{[C_R(\omega)+S_I(\omega)]^2+[S_R(\omega)-C_I(\omega)]^2}{2N\sigma^2}\right\}. \tag{10}$$

The exact relation of the sufficient statistic $\overline{h^2}$ to a discrete Fourier transform will be discussed shortly. First these results are generalized to a model which contains multiple well-separated stationary sinusoids.

When the data are known to contain r stationary sinusoidal frequencies, the model signal in the real channel may be written

$$f_R(t) = \sum_{j=1}^{2r} B_j U_j(t) = \sum_{j=1}^{r} B_j \sin(\omega_j t) + B_{r+j}\cos(\omega_j t), \tag{11}$$

and

$$f_I(t) = \sum_{j=1}^{2r} B_j V_j(t) = \sum_{j=1}^{r} B_j \cos(\omega_j t) - B_{r+j}\sin(\omega_j t), \tag{12}$$

for the imaginary channel, where phase coherency has not been assumed.

The posterior probability of the frequencies is computed from the g_{jk} matrix. This matrix has rank $2r$ and is given by

$$g_{jk} = g_{r+j,r+k} = \frac{\sin\{N(\omega_j-\omega_k)/2\}}{\sin\{(\omega_j-\omega_k)/2\}} \qquad (1 \le j,k \le r); \tag{13}$$

all other elements are zero. If the frequencies are well separated, $|N(\omega_j-\omega_k)| \gg 2\pi$ (i.e., during the acquisition period the differences in frequencies between the jth and kth component evolve through many cycles), or even separated by a moderate amount, $|N(\omega_j-\omega_k)| > 5\pi$ (e.g., a 3 Hz frequency difference for 10,000 Hz total bandwidth and 16K complex data points), the off-diagonal terms are small compared to the diagonal and the g_{jk} matrix may be approximated as

$$g_{jk} \approx N\delta_{jk} \qquad (1 \le j,k \le 2r), \tag{14}$$

where δ_{jk} is a Kronecker delta function.

When the g_{jk} matrix is diagonal, the vectors represented by sines and cosines are orthogonal, and the problem of estimating the r frequencies separates into r one-frequency problems. The

posterior probability of r well-separated stationary sinusoidal frequencies is just the product of the probabilities of the individual frequencies,

$$\begin{aligned} P(\omega_1, \cdots, \omega_r | \sigma, D, I) &\approx \exp\left\{ \sum_{j=1}^{r} \frac{\overline{h^2}(\omega_j)}{\sigma^2} \right\} \\ &= \exp\left\{ \sum_{j=1}^{r} \frac{[C_R(\omega_j) + S_I(\omega_j)]^2 + [S_R(\omega_j) - C_I(\omega_j)]^2}{2N\sigma^2} \right\}, \end{aligned} \tag{15}$$

where $\overline{h^2}(\omega_j)$ is just the sufficient statistic for single frequency estimation, Eq. (7), evaluated for each of the r frequencies.

As noted earlier, in NMR the traditional way to analyze data for frequencies is to take the quadrature data as a complex data set

$$d(t_i) = d_R(t_i) + i d_I(t_i), \tag{16}$$

and then perform a discrete Fourier transform on the data. The squared magnitude of the discrete Fourier transform of the complex data, the power spectrum, is given by

$$\left| \sum_{k=1}^{N} d(t_i) \mathrm{e}^{-i\omega t_k} \right|^2 = [C_R(\omega) + S_I(\omega)]^2 + [S_R(\omega) - C_I(\omega)]^2. \tag{17}$$

Up to the constant factor $1/2N$, the sufficient statistic, Eq. (7), is the squared magnitude of a discrete Fourier transform of the complex data. Therefore, the discrete Fourier-transform power spectrum is essentially the *logarithm* of the posterior probability of a single stationary sinusoidal frequency, or it may also be interpreted as the *logarithm* of the posterior probability of r stationary well-separated sinusoids.

The reason zero-padding gives improved frequency resolution is that it allows one to calculate the sufficient statistic or Schuster periodogram at smaller frequency intervals. Locating the r largest maxima in a Schuster periodogram is equivalent to locating the maximum of the posterior probability of r stationary well-separated frequencies. If the frequencies are nearly stationary, or evolve through many cycles in the FID, then the maximum of the posterior probability, Eq. (15), is nearly the best estimate of the frequencies one may obtain.

However, there are limitations to the validity of the discrete Fourier-transform power spectrum as an optimal frequency estimator. Specifically, the assumptions that went into the calculation must be at least approximately met. The exact conditions of validity are (i) there can be only one frequency in the data, (ii) the frequency must be stationary and sinusoidal, (iii) the noise must be white. If these three conditions are not met, the discrete Fourier-transform power spectrum will still give a valid, but conservative, estimate of the frequency and there will be other statistics which will give better (i.e., more precise) frequency estimates.

To illustrate the conservative nature of the estimates, suppose the first condition is violated and there are two stationary sinusoids in the data set. Because the power spectrum represents the *logarithm* of the posterior probability, if the frequencies and amplitudes differ from each other by even a small amount, the highest peak in the discrete Fourier-transform power spectrum is where all of the posterior probability is concentrated – see Ref. 5 for examples of this. Thus, when the posterior probability, Eq. (10), is normalized, only the highest peak will be significant; the other resonance will be considered noise. Probability theory will then estimate the amplitude, phase, and variance of the noise from a small region around the highest peak. Because the other sinusoid is considered noise, a larger standard deviation of the noise will result. The precision estimates for the frequency and the amplitude are proportional to the estimated standard deviation of the noise.

Thus, the accuracy of the estimates will be conservative in the sense that, when one includes the second sinusoid in the signal, the estimated standard deviation of the noise will be smaller, and consequently the precision of the estimates will improve.

Now suppose the second condition is violated and the frequency is nonstationary (decaying with time); when the frequency is estimated, probability theory takes the scalar-product between the data and the model, and the estimated frequency is the one which makes these vectors as close to parallel as is possible. This occurs when the estimated frequency is equal to the frequency in the data to within the noise; but the data decay, while the model does not. Again, there will be significant parts of the signal placed into the noise. This will again result in a large estimated standard deviation of the noise, and consequently the accuracy estimates will be worse. Including any type of decay that is reasonable for the signal in question will improve the resolution of the frequency and amplitude, because it will reduce the estimated standard deviation of the noise [5].

Now suppose the signal is periodic and stationary, but not sinusoidal. When the signal is not sinusoidal, the best fit will still occur when the estimated frequency matches the fundamental frequency in the data. But, there will be a significant misfit between the data and the model. This misfit will result in an increased standard deviation estimate and the accuracy estimates for the frequency will again be worse.

It is more difficult to see how knowledge that the noise is not white can be exploited to improve the parameter estimates. Nonwhite noise contains correlations. Correlations can be described by a new parameter ρ, the correlation coefficient. One performs a new probability calculation which incorporates these correlations. In this new probability calculation the accuracy of parameter estimates are all proportional to $\sigma\sqrt{1-\rho^2}$. If the noise is uncorrelated, $\rho = 0$; the accuracy estimates reduce to those derived in Ref. 1, and the noise is white; any other value of the correlation coefficient acts to reduce the effective standard deviation of the noise and improve the accuracy estimates [6].

If one has specific knowledge about how the data depart from the model (here model means both the model for the signal and the model for the noise), that information can be used in another probability calculation to obtain more precise parameter estimates. The discrete Fourier-transform power spectrum thus represents a conservative estimate of the frequency. But the discrete Fourier-transform power spectrum is approximately an optimal frequency estimator under a wider set of conditions than those just given. Specifically, the discrete Fourier-transform power spectrum is approximately an optimal frequency estimator for multiple well-separated stationary frequencies when (i) there are r stationary sinusoids present, (ii) the frequencies are well separated, and (iii) the noise is white. All of the previous comments are applicable with three additions. First, from the viewpoint of probability theory, the discrete Fourier transform answers a question about frequency estimation, not a question about signal detection, and not a question about model selection. If the number of frequencies r is not known from prior information, *nothing* in the discrete Fourier transform can tell one its value. Second, when the frequencies are not well separated, the discrete Fourier transform is *not even approximately* an optimal frequency estimator and can give misleading or even incorrect results – incorrect in the sense that better models will give better results. Third, when the frequencies are not stationary, the discrete Fourier transform is *never* an optimal frequency estimator, and there are *always* other statistics which give better estimates of the frequencies and decay rate constants.

This should not be interpreted as recommending against the use of the discrete Fourier transform. When the frequency has exponential decay, use of a discrete Fourier-transform *power spectrum* will produce nearly optimal frequency estimates, when the data are multiplied by an appropriate exponentially decaying apodizing function [5]. Thus, under many conditions encountered in NMR, one simply cannot obtain better estimates of the frequencies than from a zero-padded discrete Fourier-transform power spectrum of the complex data. Indeed, it is the place to start on all frequency estimation problems, but it represents the answer to a specific question. When the data contain nonstationary frequencies, when spectral lines overlap in the discrete Fourier transform,

when nonsinusoidal periodicities are present, or when the number of resonances is not known from prior information, then the discrete Fourier transform is answering an inappropriate question and other statistics will give better results.

Note, however, that it is the discrete Fourier-transform *power spectrum* that is the proper statistic for estimating the values of stationary frequencies, not the *absorption spectrum.* The power spectrum is, up to a scale factor, essentially the logarithm of the posterior probability of a single stationary sinusoidal frequency, or it may be interpreted as the logarithm of the posterior probability of multiple well-separated stationary sinusoidal frequencies. The absorption spectrum starts by throwing away half of the information in the data, and then effectively proceeds to take the square root of the logarithm of the posterior probability. This has the effect of so compressing the scale that unless the evidence for a frequency is overwhelming, one simply cannot see it: this will be demonstrated shortly. Additionally, when the imaginary part of the discrete Fourier transform is discarded, the phase information is *essentially discarded.* The common problems associated with phase twists and anti-phased peaks are the result of discarding this phase information. These problems do not occur in the power spectrum, because the power spectrum essentially estimates the phase and amplitude at each value of the frequency and then eliminates them from consideration.

Note also, that it is a question about frequency estimation that is being addressed by the discrete Fourier-transform power spectrum, not a question about amplitude estimation. Thus when one uses the amplitudes given by Eq. (39) in the previous paper [1], it is the amplitude of a stationary sinusoid that is estimated; not the amplitude of an exponentially decaying sinusoid. If the problem is amplitude estimation, then Bayes' theorem tells one how to do this; simply compute the posterior probability of the amplitudes independent of the frequencies, decay rate constants, and variance of the noise. Such a statistic would look very different from the discrete Fourier-transform power spectrum, or the absorption spectrum.

If the goal of an NMR experiment is to estimate the frequencies very accurately, then probability theory indicates how this should be done: the data must be sampled very rapidly (to obtain data before the signal has decayed into the noise), the peak signal-to-RMS-noise ratio should be very high, and a complete model of the data must be used in the analysis. Only then will the frequencies be estimated as accurately as is possible – see Ref. 5. The frequencies are estimated to an accuracy that depends directly on the estimated standard deviation of the noise, inversely on the amplitude of the sinusoid, and inversely on the square root of the number of data values. But probability theory tempers the $\sqrt{N}$ effect and indicates that it is only the number of data values in the region where the signal has not decayed to zero that is important [5]. The accuracy of the estimated parameters is proportional to the estimated standard deviation of the noise; the smaller the estimated standard deviation of the noise, the better the estimates of the parameters. The best estimates will be obtained when all systematic effects in the data have been included in the model, even when the effect is a nuisance effect (such as an instrumental artifact). There is no magic in Bayesian analysis: if one wants good parameter estimates, then one must think carefully about the model and use all the information available in the analysis of the data.

Many of the problems typically encountered with the use of the discrete Fourier-transform absorption spectrum are now easily explained. For example, suppose one has an FID that contains a single exponentially decaying sinusoid. However, the data are sampled well past the point where the sinusoid has decayed into the noise. When using a discrete Fourier transform, one must be very careful to ensure that data at the end (where no signal exists) are discarded; otherwise, no peak exists. These data are typically discarded by multiplying by a decaying exponential apodizing function. From the point of view of probability theory, the discrete Fourier transform answers a question about stationary frequency estimation, and all data values are equally relevant. Thus when data values near the end of the FID are included in the discrete Fourier transform, the evidence for stationary frequencies goes down, as it must; there are no stationary frequencies in the data. Bayesian analysis using models which incorporate decay do not have this problem, because data values are weighted. For exponential decay, data values are weighted exponentially and values at

late times are essentially assigned zero weight. By multiplying the data by an exponential apodizing function and computing the power spectrum, one is essentially computing the logarithm of the posterior probability of a single exponentially decaying sinusoid. Probability theory interprets this procedure not as filtering; rather, one is estimating the value of a frequency and decay rate constant. Because, the numerical value of this statistic is much greater than the unapodized power spectrum, the frequency and decay rate constant have been determined much more precisely. Additionally, in the apodized power spectrum it is essentially the peak height that determines the amplitude; not the area.

Baseline artifacts are now easily understood also. Suppose the FID contains two resonances: one very large resonance that decays into the noise in the first 20 or 30 data values and one small resonance that decays into the noise in approximately 1000 data values. In order to see the rapidly decaying component in the data, the amplitude of the rapidly decaying component must be several orders of magnitude larger than the slowly decaying component. Since acquisition does not begin at time $t = 0$, this is especially true when the rapidly decaying component decays away in only the first 2 or 3 data values. The intuitive picture given in the previous paper [1], indicates that the frequency estimated by the discrete Fourier transform will be the one for which the mean-square difference between the data and the model is a minimum. This minimum occurs when the single stationary sinusoidal model most resembles the rapidly decaying component in early times and the slowly decaying component in late times. Note, however, that it is not the oscillations that are being fitted but the envelope of the decay. The mean-square difference between the data and the model is minimized, because in the residuals the resonances are now nearly stationary. The peak, or feature, in the discrete Fourier transform is broad, because a wide range of frequencies give nearly the same mean-square difference. There is often no peak at the location of either the rapidly decaying component or the slowly decaying component; they simply get lost because decay is the dominate feature in these data, not oscillations. Indeed, there could be excellent evidence for both components in the data, and use of models which include decay can readily extract them – see Ref. 7 and Ref. 8 for examples of how to use Bayesian techniques to analyze such data.

The intuitive picture given in the previous paper [1] should also alert one to possible problems concerning the procedures typically used to correct these "baseline" artifacts. These baseline features are often modeled by a polynomial in the frequency domain. The coefficients for this polynomial are determined from the absorption spectrum, and the estimated polynomial is then subtracted from the absorption spectrum. But probability theory clearly indicates that the data are projected onto the model by taking scalar products. Here there are two models: (i) the sines and cosines, and (ii) the time-domain baseline model. If the time domain baseline model is not orthogonal to the sines and cosines (by orthogonal it is meant that the scalar product of the sine or cosine onto the polynomial must be zero), then when the baseline is subtracted from the absorption spectrum, some of the signal of interest will also be subtracted. This will change the estimated values of the frequencies, amplitudes, and decay rate constants. The amount of change will be directly related to how far from orthogonal the two models are.

Signal Detection

Nothing in the preceding discussion can determine when a signal has been detected or what model signal best characterizes the data, because the discrete Fourier transform is answering a question about frequency estimation. It implicitly assumes the functional form of the model, and it assumes one knows the number of resonances. Procedures were developed in the preceding paper [2] for dealing with signal detection and model selection problems. In this section, the signal detection procedures are applied to sinusoidal FID NMR data.

Suppose one has an FID that looks like that in Fig. 1. The data shown are the real channel of a ^{1}H FID using ethyl ether (diluted in C_6D_6) as the sample. The data contain $N = 512$ time samples

Figure 1: Simulated FID Data

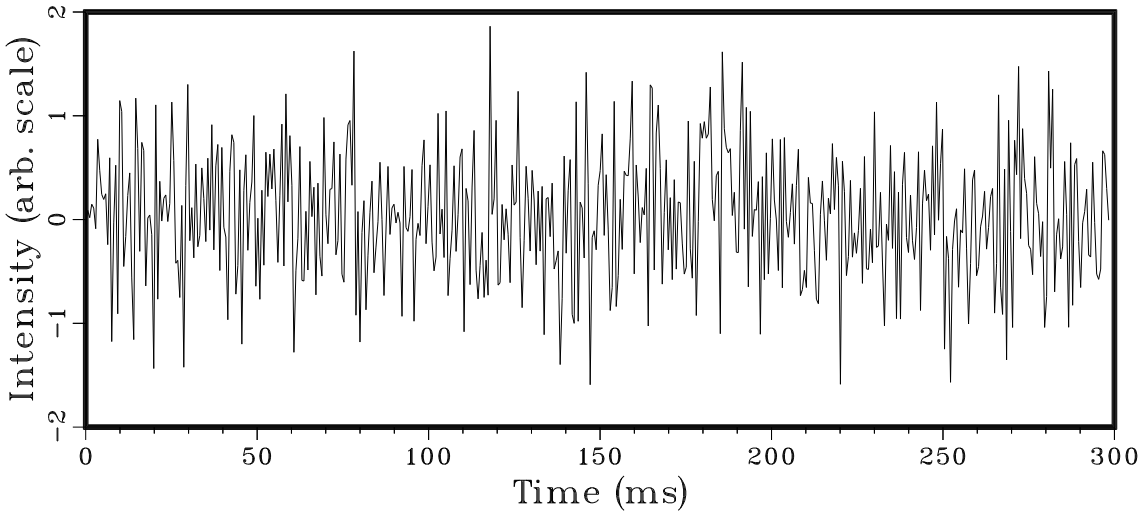

Fig. 1. The real channel of a ^{1}H NMR FID of ethyl ether, with $N = 512$ time samples per channel. The data span approximately 300 ms. Gaussian white noise was added to the data to make an extremely noisy data set. The peak signal-to-RMS-noise ratio in these time domain data is approximately 0.6.

per channel, span approximately 300 ms, and were taken using a Varian 500VXR spectrometer. Because the sample did not contain a reference compound, the ppm scale shown was set by defining the center frequency of the quartet [ethyl ether ^{1}H NMR spectrum consists of a triplet (-CH_3) and a quartet (-CH_2-)] to be 3.4 ppm relative to TMS. To illustrate signal detection, Gaussian white noise was added to the data, resulting in an extremely noisy data set. To the eye, it is very difficult to tell whether a signal is present. The discrete Fourier transform *absorption spectrum*, Fig. 2, is not much help. In the previous section, it was demonstrated that a power spectrum is the proper statistic to use when estimating the value of a single stationary frequency. Because of the way these data were sampled, the signal is very nearly stationary. Thus, the discrete Fourier-transform power spectrum is very nearly the proper statistic to use when estimating the value of a frequency. A *power spectrum* of these data, Fig. 3, does indeed give better evidence for the presence of a frequency. But the problem is signal detection, not parameter estimation, and, as has been emphasized before, nothing in the discrete Fourier transform can tell one whether the peak near 1.25 ppm is a real frequency or an artifact of the noise.

The procedures derived in Ref. 2 are applied here to determine if the peak near 1.25 ppm represents significant evidence in favor of a frequency, or if it is an artifact of the noise. To use those procedures one must state what is meant by the signal. Here two candidate models f_1 and f_2 are considered. Model f_1 will be a quadrature model of a constant,

$$f_1 \equiv \begin{cases} B_1 U_1 \\ B_1 V_1 \end{cases} = \begin{cases} f_R = B_1 & (1 \le i \le N) \\ f_I = B_1 & (N < i \le 2N) \end{cases}, \tag{18}$$

and represents *no signal.* Model f_2 will be a quadrature model of a stationary sinusoid plus a constant,

$$f_2 \equiv \begin{cases} \sum_{j=1}^{3} B_j U_j \\ \sum_{j=1}^{3} B_j V_j \end{cases} = \begin{cases} f_R & = B_1 + B_2 \sin(\omega t_i) + B_3 \cos(\omega t_i) & (1 \le i \le N) \\ f_I & = B_1 + B_2 \cos(\omega t_{i-N}) - B_3 \sin(\omega t_{i-N}) & (N < i \le 2N) \end{cases}, \tag{19}$$

Figure 2: Absorption Spectrum

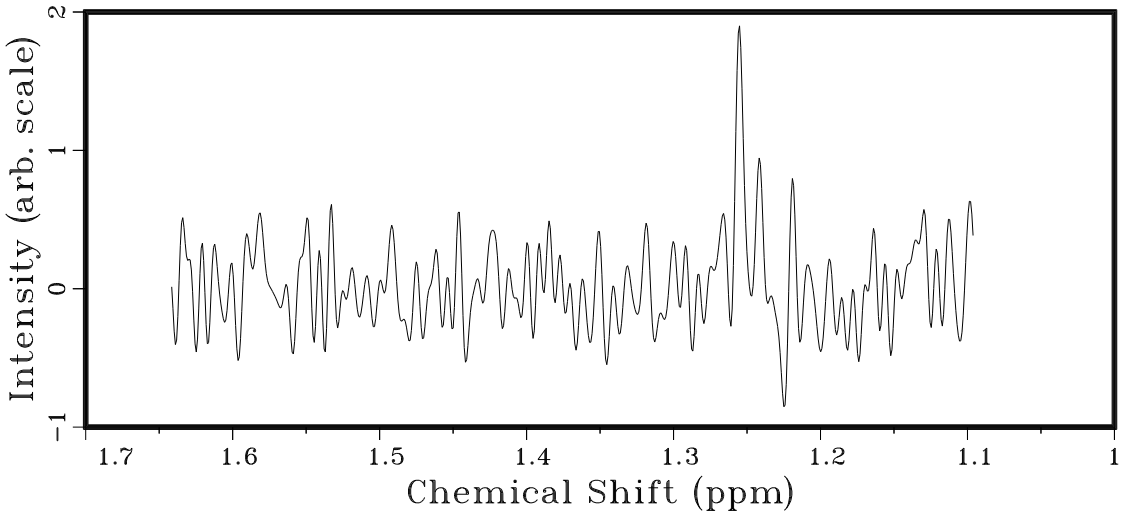

Fig. 2. The ^{1}H NMR absorption spectrum of the data shown in Fig. 1. In the region of the spectrum shown, the center frequency of the -CH_3 triplet is the dominant effect [ethyl ether ^{1}H NMR spectrum has a triplet (-CH_3) and a quartet (-CH_2-)]. One cannot tell from the absorption spectrum if any frequencies are present.

Figure 3: Power Spectrum

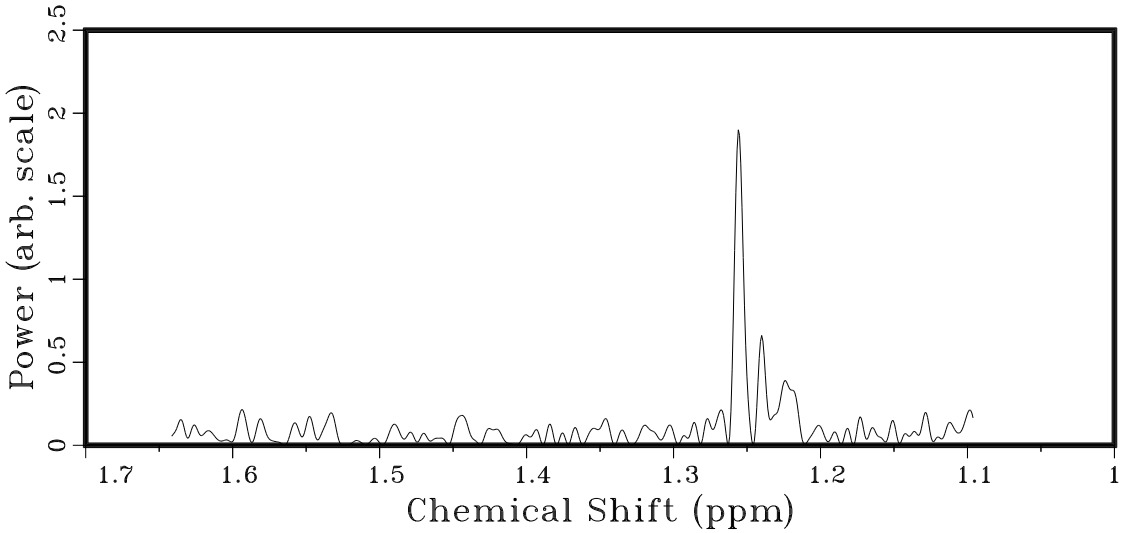

Fig. 3. Bayesian analysis indicates that a power spectrum is the appropriate statistic to use when estimating the value of a single stationary sinusoid. This is a power spectrum of the data shown in Fig. 1. This power spectrum does indeed give better evidence for the frequencies. But the question is signal detection, not parameter estimation, and nothing in the discrete Fourier transform can tell one if that peak near 1.25 ppm is noise or if it is an indication of a real frequency.

and represents *the signal.* This example will be worked in considerable analytic detail; then several numerical examples using the data in Fig 1 will be given. Decay will be included in model f_2 in the numerical examples. In the calculation, the global likelihood of data, Eq. (23) of Ref. 2, will be used. Thus, the prior uncertainty in the amplitudes δ^2, and the variance of the noise σ^2 will both be assumed known; then in the numerical examples, both of these parameters will be eliminated by using the global likelihood of the data, Eq. (27) from Ref 2. The global likelihood of the data independent of these parameters is not used in the analytic calculations simply because of its increased complexity.

The procedures derived in preceding paper [2] did not explicitly assume quadrature models. But in the earlier paper [1] it was demonstrated that quadrature models (when the variance of the noise is the same in both channels) are special cases of nonquadrature models. A quadrature model may be written in nonquadrature form by taking the data D, with data item d_i, to be

$$d_i = \begin{cases} d_R(t_i) & (1 \leq i \leq N) \\ d_I(t_{i-N}) & (N < i \leq 2N) \end{cases}. \tag{20}$$

The total number of data values is $2N$, where N is the number of data values per channel. The signal functions G_j are given by

$$G_j(t_i) = \begin{cases} U_j(t_i) & (1 \leq i \leq N) \\ V_j(t_{i-N}) & (N < i \leq 2N) \end{cases}. \tag{21}$$

The posterior probability of model f_1 is computed from Eq. (3) derived in Ref. 2 using the global likelihood of the data, Eq. (23) from Ref. 2. The global likelihood of the data, Eq. (23) from Ref. 2, is computed from the g_{jk} matrix. For model f_1 this matrix is just a single number

$$g_{jk} = g_{11} = \sum_{i=1}^{2N} G_1(t_i)G_1(t_i) = \sum_{i=1}^{N} U_1(t_i)U_1(t_i) + V_1(t_i)V_1(t_i) = 2N. \tag{22}$$

The sufficient statistic $\overline{h^2}$, Eq. (21) from [2], for model f_1 is given by

$$\overline{h^2} = 2N(\overline{d})^2, \tag{23}$$

where

$$\overline{d} = \frac{1}{2N}\sum_{i=1}^{N}[d_R(t_i) + d_I(t_i)] \tag{24}$$

is the average value of both the real and imaginary channels. If one has no prior knowledge about which model is in the data, and assigns each model a prior probability of 1/2, the posterior probability for model f_1 becomes

$$P(f_1|\sigma, \delta, D, I) = \frac{(2\pi)^{-N}\sigma^{1-2N}\delta^{-1}}{2}\exp\left\{-\frac{N\overline{d^2} - N(\overline{d})^2}{\sigma^2}\right\}, \tag{25}$$

where the variance σ^2 and δ^2 are assumed known, and a term in the exponential, $(\overline{d})^2/2\delta^2$, is assumed small compared to the retained terms. This assumption was the basis for all of the calculations done in the preceding paper [2]. This term was not dropped in Ref. 2, because it is not negligible when δ is removed as a nuisance parameter.

An exact solution for the posterior probability of model f_2 can be determined, because the eigenvalues and eigenvectors of the g_{jk} matrix may be computed analytically. Here, an approximate solution is given to simplify some of the analytic details. The solution derived is valid provided $|\omega N| \gg 2\pi$, i.e., there is no evidence of a low frequency. Basically, in avoiding the region in which

model f_2 reduces to f_1, one avoids taking careful limits when $\omega \to 0$. While this is an interesting limit, it is concerned more with probability theory than with spectrum analysis.

To compute the posterior probability of model f_2, one again computes the g_{jk} matrix. If the amount of data is large, $N \gg 1$, then $\sum_{i=1}^{N} \cos(\omega t_i) \approx \sum_{i=1}^{N} \sin(\omega t_i) \ll N$. The g_{jk} matrix is nearly diagonal and may be approximated as

$$g_{jk} \approx \begin{pmatrix} 2N & 0 & 0 \\ 0 & N & 0 \\ 0 & 0 & N \end{pmatrix}. \tag{26}$$

The sufficient statistic $\overline{h^2}$, Eq. (21) from [2], for model f_2 is then given by

$$\overline{h^2} \approx \frac{1}{3}\left\{2N(\overline{d})^2 + \frac{1}{N}(C_R + S_I)^2 + \frac{1}{N}(S_R - C_I)^2\right\}. \tag{27}$$

The posterior probability of model f_2 is then given by

$$P(f_2|\omega, \sigma, \delta, D, I) = \frac{\sigma^{3-2N}\delta^{-3}}{2(2\pi)^N} \exp\left\{-\frac{2N[\overline{d^2} - (\overline{d})^2] - (C_R + S_I)^2 - (S_R - C_I)^2}{2N\sigma^2}\right\}, \tag{28}$$

where a small term in the exponential was again ignored, and $P(f_2|I) = 1/2$ was used.

A posterior odds ratio was used to exhibit the results in the preceding paper [2] and that approach will be used here also. The odds ratio used is

$$K = \frac{P(f_2|\omega, \sigma, \delta, D, I)}{P(f_1|\sigma, \delta, D, I)} = \frac{\sigma^2}{\delta^2} \exp\left\{\frac{(C_R + S_I)^2 + (S_R - C_I)^2}{2N\sigma^2}\right\}, \tag{29}$$

and is the odds in favor of the sinusoidal model f_2. The result depends on two factors: the first factor, σ^2/δ^2, is like a ratio of student t distributions expressing the prior odds in favor of the simpler model. If this ratio is very small, one has strong evidence in favor of the simpler model. This is the regime were these equations are valid – this is a conservative signal detection calculation, where the prior odds are strongly in favor of the simpler model. The second factor is the posterior probability of a stationary harmonic frequency, Eq. (10), and represents how well the sinusoid fits the data. For this odds ratio to express evidence in favor of model f_2, the discrete Fourier-transform power spectrum must have a peak, which is large compared to the variance of the noise.

The posterior odds ratio, Eq. (29), expresses a bet. If the odds ratio is greater than 1, it is a bet in favor of the model containing the sinusoid. If the odds are less than 1, it is a bet in favor of the constant model – no signal of interest. If the odds are exactly 1, neither model is to be preferred. By experimenting with some numbers, a better understanding of the detection process, and how sensitive these formulae are to the prior information, can be obtained. Suppose one has data, which contain a sinusoidal signal of the form

$$d_i = \begin{cases} \hat{B}\cos(\hat{\omega}t_i) + e(t_i) & (1 \le i \le N) \\ \hat{B}\sin(\hat{\omega}t_{i-N}) + e(t_i) & (N < i \le 2N) \end{cases}, \tag{30}$$

where $\hat{B}$ is the true amplitude of the sinusoid, $\hat{\omega}$ is the true frequency, and $e(t_i)$ represents noise. If the peak signal-to-RMS-noise is large, the power spectrum will have a peak near $\omega \approx \hat{\omega}$ and at the peak the odds ratio will be given approximately by

$$K = \frac{\sigma^2}{\delta^2} \exp\left\{\frac{N\hat{B}^2}{4\sigma^2}\right\}. \tag{31}$$

If δ^2 were known, one could determine the peak signal-to-RMS-noise ratio needed to give even odds. When these equations were derived, the assumption was made that the prior probability

was essentially uniform over the region where the likelihood of the data is sharply peaked. So the calculation should not depend strongly on the value of δ^2. Almost any value of $\delta^2 \gg \sigma^2$, should work equally well. Suppose $\delta^2 = 100\sigma^2$; then terms involving $1/\delta^2$ are small, and the assumptions that went into the calculation are met. This assumption will be made for now, and later δ^2 will be made a million times larger to see what effect this has on the conclusions. Using the assumption $\delta^2 = 100\sigma^2$, and assuming there are $N = 512$ data values per channel, as in Fig. 1, and assuming the noise variance is one, then the even money bet is given by

$$K = 1 = \frac{1}{100} \exp\left\{\frac{512\hat{B}^2}{4}\right\}, \tag{32}$$

which gives

$$\hat{B} \approx \sqrt{\frac{\log(10^2)}{128}} \approx 0.19. \tag{33}$$

The even money bet occurs when the average noise fluctuations are five times larger than signal amplitude.

The posterior odds ratio depends on the exponential of the square of the amplitude: the posterior odds must rise very rapidly as a function of amplitude. To demonstrate this, suppose $\hat{B} = 0.3$ and the other parameters are unchanged, then the odds in favor of model f_2 becomes

$$K = (\frac{1}{100}) \exp\left\{\frac{512 \times 0.3^2}{4}\right\} \approx 1000. \tag{34}$$

The change in the amplitude from 0.19 to 0.3 has raised the odds to 1000 to 1 in favor of the sinusoidal model.

But these results depended on knowing the parameter δ^2, and an arbitrary value was assigned. Suppose $\delta^2 = 10^8\sigma^2$, how would this affect the calculation? Assuming all other parameters retained their values, the even money bet now occurs when

$$\hat{B} \approx \sqrt{\frac{\log(10^8)}{128}} \approx 0.38. \tag{35}$$

By increasing δ^2 from $100\sigma^2$ to $10^8\sigma^2$, the peak signal-to-RMS-noise ratio for the even money bet has increased form 0.19 to 0.38. Even though the results depend on knowing δ^2, they depend only weakly on this information, as one would expect. There are hardly any circumstances where an experimenter could not guess the values of these parameters to within an order of magnitude, let alone the six orders of magnitude variation examined here. Regardless, the second form of the global likelihood, Eq. (27) from Ref. 2, does not require knowledge of these parameters at all.

This discussion began by examining the FID shown in Fig. 1. In the traditional absorption spectrum, Fig. 2, one could not be very sure that a signal is present, nor could one readily convince a skeptic that detection has been accomplished. The power spectrum, Fig. 3, gives better evidence for the signal, but as was emphasized earlier the problem is detection, not parameter estimation. The discussion just completed demonstrates that small signals are detectable under idealized conditions, but what about in real data? Suppose the techniques are applied to the data shown in Fig. 1. For this demonstration the second form of the global likelihood, Eq. (27) from Ref. 2, will be used. When this equation is used the posterior odds ratio has *no* adjustable parameters. The posterior odds in favor of model f_2 are displayed in Fig. 4. Because K tends to be rapidly varying, even for the data shown in Fig. 1, $10\log_{10}(K)$ has been plotted. This quantity is called the evidence and has units of decibels. For these data, at the maximum, it is a bet of better than 10^7 to 1 in favor of the model containing the frequency. Thus, probability theory not only finds evidence for a frequency, it finds good evidence for the frequency.

Figure 4: Signal Detection – Without Decay

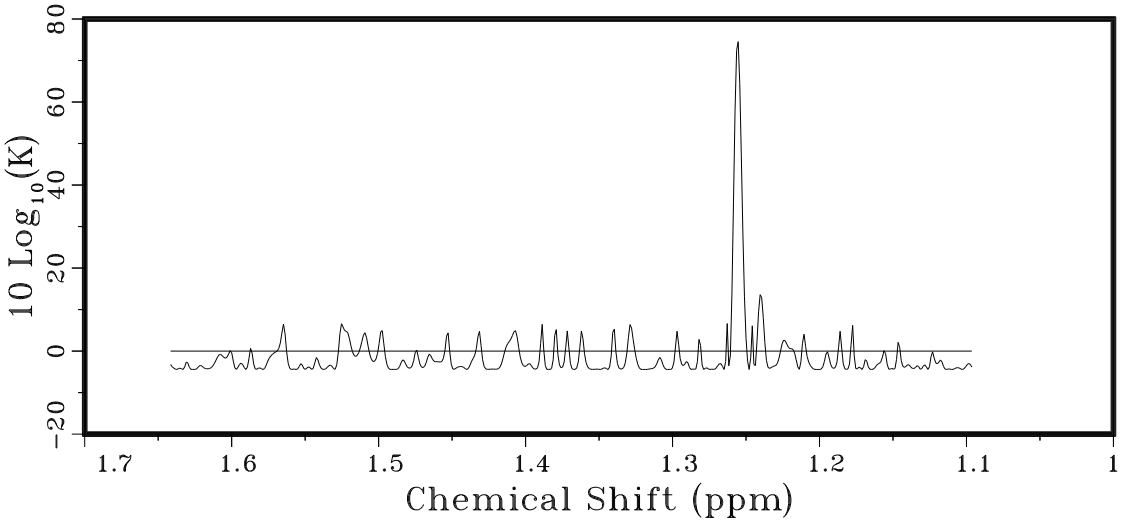

Fig. 4. The calculation in the preceding paper [2] can be used to test for the presence of a signal. Here 10 $\log_{10}(K)$ is plotted for the data shown in Fig. 1. This quantity is called the "evidence" and has units of decibels, see text for details. If the evidence is 30 dB, then it is a bet of 10^3 to 1 in favor of the sinusoidal model. If the evidence is -30 dB, it is a bet of 10^{-3} to 1 in favor of the constant model, and if the evidence is 0 dB, neither model is preferred. In these data it is a bet, at the maximum where the evidence is greater than 70, of better than 10^7 to 1 in favor of the frequency model.

But this is an FID, and all such NMR signals decay. The model just applied assumed that the sinusoid was stationary. For these data, this is a reasonable assumption; but not for FID data in general. There is nothing in the formalism that prevents one from including decay in the model. Suppose model f_2 is modified to include decay:

$$f_2 \equiv \begin{cases} f_R = B_1 + [B_2 \sin(\omega t_i) + B_3 \cos(\omega t_i)]\mathrm{e}^{-\alpha t_i} & (1 \le i \le N) \\ f_I = B_1 + [B_2 \cos(\omega t_{i-N}) - B_3 \sin(\omega t_{i-N})]\mathrm{e}^{-\alpha t_{i-N}} & (N < i \le 2N) \end{cases}, \tag{36}$$

and model f_1 remains unchanged. The posterior odds is now a function of the frequency ω and the decay rate constant α. This odds ratio is displayed as a contour plot in Fig. 5. The units are again decibels. Along the line $\alpha = 0$, this plot reduces to the previous odds ratio. If including decay has improved detection, then near the maximum odds as a function of frequency, the odds ratio should increase as one moves away from $\alpha = 0$. Indeed this does occur and the posterior odds increase from approximately 70 to 100 dB. At the maximum, it is a bet of approximately 10^{10} to 1 in favor of the model containing the frequency. Thus, *including decay* in this model was about half as important as including the frequency; but these data are nearly stationary. In typical FIDs including decay is about as important as including the frequency, and when a rapidly decaying component is present, including decay is *more* important than including the frequency.

Now that one knows, or at least is reasonably sure, that the peak near 1.25 ppm is an indication of a real frequency, the procedures derived in the earlier paper [1] may be used to estimate the values of the frequency and decay rate. Model f_2, Eq. (36), was used to estimate the frequency, decay rate, amplitude and phase. From this model one finds

$$\begin{aligned} (\omega)_{\mathrm{est}} &= 1.255 \pm 0.0015 \text{ ppm}, \\ (\alpha)_{\mathrm{est}} &= 4.1 \pm 1 \text{ s}^{-1}, \end{aligned} \tag{37}$$

Figure 5: Signal Detection – With Decay

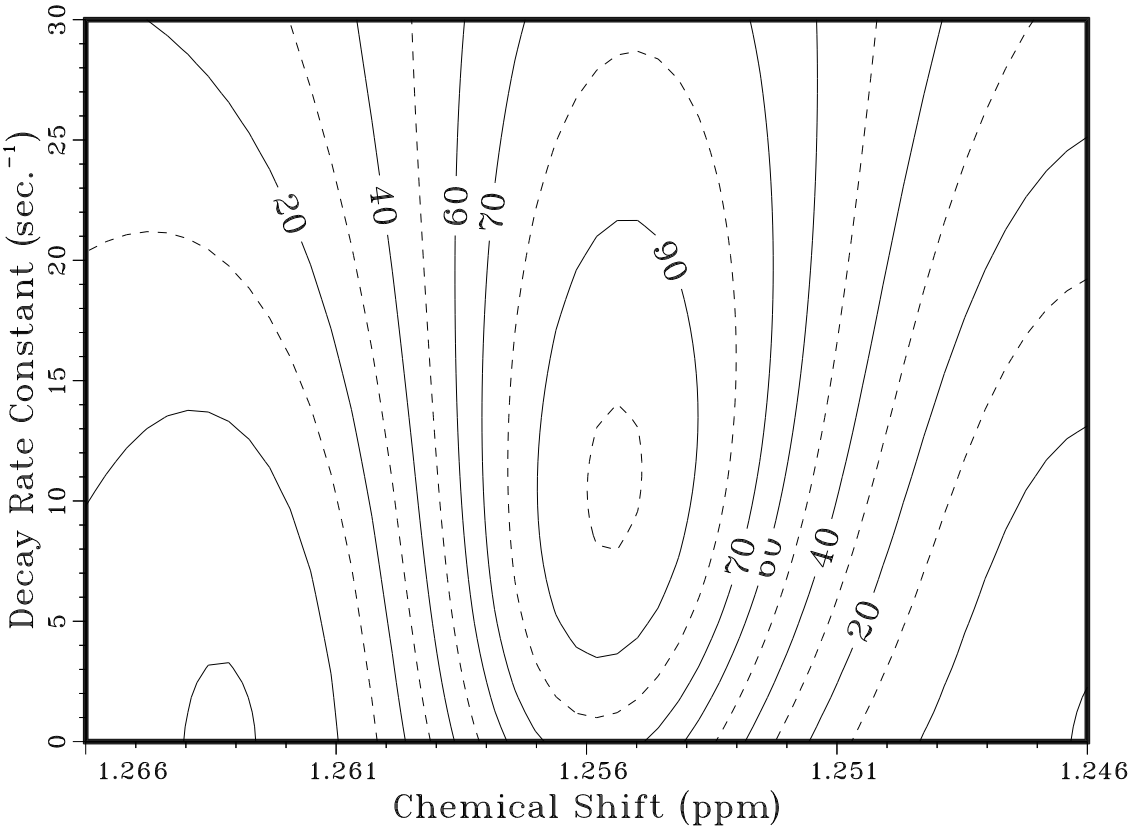

Fig. 5. Two-dimensional contour plot of the chemical shift (ppm) *vs* decay rate constant (s^{-1}). Contours are lines of constant "evidence," i.e., $10 \log_{10}(K)$ expressed in dB. The evidence in Fig. 4 does not include decay. When decay is included in the model, the evidence becomes a function of two parameters: the frequency and decay rate constant. A contour plot of the evidence, for the NMR FID data shown in Fig. 1, as a function of these parameters shows that there is a well-defined maxima. At the maximum, it is better than 10^{10} to 1 in favor of the frequency plus decay model over the constant model. Thus, including decay is about half as important as including the frequency.

and the amplitude B and phase Θ were estimated to be

$$\begin{aligned} (B)_{\text{est}} &= 8000 \pm 3000 \text{ (arb. units)}, \\ (\Theta)_{\text{est}} &= 5.2 \pm 0.3 \text{ rad.} \end{aligned} \tag{38}$$

The accuracy estimates are at two standard deviations. The peak signal-to-RMS-noise ratio of these data was estimated to be 0.68. Thus, in data where one cannot be sure of detection using the absorption spectrum, probability theory can tell one that the sinusoid is present, and then estimate the frequency to a precision which is 2.2 times better than the Rayleigh criterion [9]. This is not magic: probability theory simply makes use of all of the data and prior information. The precision estimates scale with the estimated standard deviation, and scale inversely with $\sqrt{N}$. Even though the estimated standard deviation of the noise is large, there are 1024 data values, and the signal is nearly stationary; thus all 1024 values are relevant, and a precise estimate of the frequency is possible. Note that even though the frequency is estimated to 1 part in 850, the decay rate is estimated only to 1 part in 4, and the amplitude is not even estimated to 1 part in 3. Frequencies are relatively easy to estimate; while decay rate constants, amplitudes, and phases are relatively difficult to estimate.

Model Selection

The preceding discussion allows one to determine how strongly a model containing given nonlinear parameters is supported relative to an alternative linear model. This is important, because it allows one to tell at a glance whether there is evidence for the signal of interest. But after the signal has been detected the problem changes from detection to either parameter estimation, or model selection. In this section the model selection procedures derived in the preceding paper [2] will be applied to FID data: first, to determine the number of resonances in the data and, second, to determine the "best" model for the data.

The data used in the analysis will again be a ^{1}H FID using an ethyl ether sample, but at a higher peak signal-to-RMS-noise level. The original FID contained 6848 real data values and 6848 imaginary data values, and spanned a total time of approximately 4.0 s. Using a discrete Fourier transform as the analysis tool, and the Rayleigh criteria for resolution ($1/2T$, where $T = 4.0$ s), the frequencies may be resolved to 0.125 Hz, if one uses all of the data. In the analysis using probability theory, only 512 real and 512 imaginary data values will be used. The resolution using probability theory will be contrasted to the resolution using the discrete Fourier transform. The real part of these data are shown in Fig. 6A. Because only a small part of the total FID is used, the absorption spectrum, Fig. 6B and Fig. 6C, show significant truncation artifacts. Just outside of the regions shown there are three other resonances due to residual water, the deuterated benzene solvent, and a spectrometer glitch. In the following calculations, these three frequencies will be accounted for with a model containing three independent exponentially decaying sinusoids, but the discussion will center on the seven lines shown. The peak signal-to-RMS-noise ratio of the time domain data is approximately 7.

To determine the number of spectral lines in the data, one computes the posterior probability of the number of spectral lines, r, given the data and the prior information. Applying Bayes' theorem, the posterior probability of the number of spectral lines, r, is given by

$$P(r|D,I) = \frac{P(r|I)P(D|r,I)}{P(D|I)}, \tag{39}$$

where $P(r|I)$ is the prior probability of the number of spectral lines, r. For this problem no prior information will be assumed about the number of lines in the data, and the prior will be taken to be a normalized uniform prior: $P(r|I) = 1/s$, where s is an upper bound on the number of frequencies. The global likelihood of the data, $P(D|r,I)$, represents how well the data fit the model and, in the

Figure 6: Ethyl Ether Data

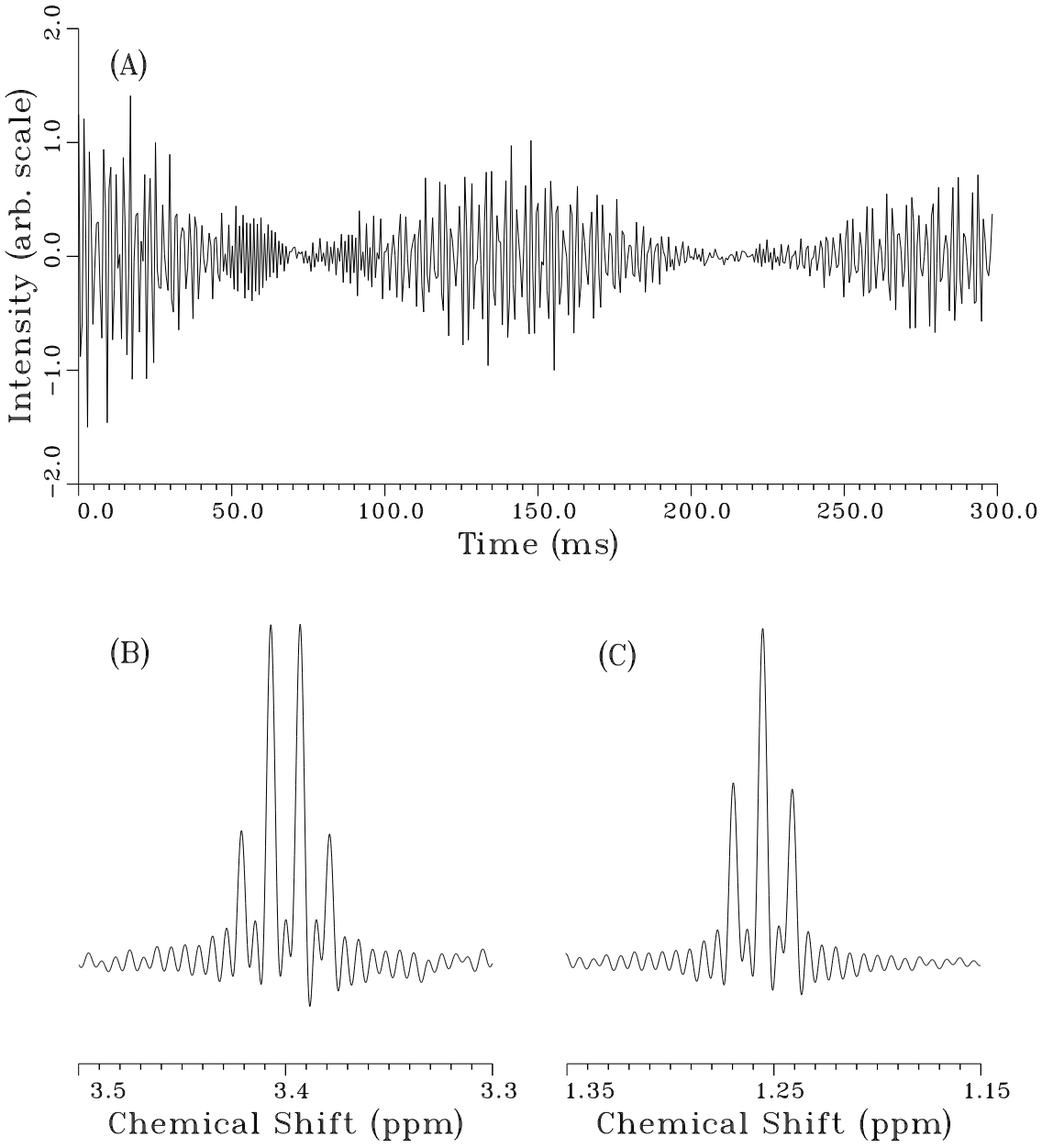

Fig. 6. The real channel of the time-domain ethyl ether data are shown in (A). There are $N = 512$ data values per channel. There are two regions of interest in the absorption spectrum: the region of the quartet (B) and the region of the triplet (C). Just outside of the two regions shown there are three other small resonances that are accounted for and ignored in the analysis – see text for details.

notation of the preceding paper [2], is $P(D|f_r, I)$. The probability of the data, given only the prior information $P(D|I)$, is a normalization constant. With these substitutions the posterior probability of the number of spectral lines, r, may be written as

$$P(r|D,I) = P(f_r|D,I) = \frac{P(D|f_r,I)}{\sum_{j=1}^{s} P(D|f_j,I)}. \tag{40}$$

This calculation is just an application of global likelihood of the data, Eq. (42) from Ref. 2.

In this calculation, the model signal will be taken to be the sum of exponentially decaying sinusoids of the form

$$f_2 \equiv \begin{cases} f_R & = \sum_{j=1}^{2r} [B_j \sin(\omega t_i) + B_{j+r} \cos(\omega t_i)] \mathrm{e}^{-\alpha_j t_i} & (1 \le i \le N) \\ f_I & = \sum_{j=1}^{2r} [B_j \cos(\omega t_{i-N}) - B_{j+r} \sin(\omega t_{i-N})] \mathrm{e}^{-\alpha_j t_{i-N}} & (N < i \le 2N) \end{cases}, \tag{41}$$

where r is the unknown number of spectral lines in the data, $\{B_1, \cdots, B_{2r}\}$ are effectively the unknown amplitudes and phases of the sinusoids, and $\{\omega_1, \cdots, \omega_r, \alpha_1, \cdots, \alpha_r\}$ are the unknown frequencies and decay rate constants. A great deal of information is known about the frequencies, decay rate constants, amplitudes, and phases of these sinusoids; all of this information will be incorporated into the models before this demonstration is finished. In the determination of the number of sinusoids, only the functional form of the resonances will be used. Phase coherences are not used, because some of the nuisance resonances are not in phase.

To compute the posterior probability of the number of spectral lines, r, one simply computes the global likelihood of the data, Eq. (19) from Ref. 2, using the model signal, Eq. (41), as the set of models S. Care must be taken when applying the global likelihood of the data, Eq. (19) from Ref. 2, because those equations were derived without the quadrature assumption. Here the number of model functions is $2r$, the number of nonlinear parameters is also $2r$, and the total number of data values is two times the number of data samples per channel, or 1024. Thus, for this problem the global likelihood of the data, Eq. (42) from Ref. 2, becomes

$$P(D|r,I) \approx \frac{r!\Gamma(r)\Gamma(512-2r)[r\overline{h^2}]^{-r}[r\overline{\Theta^2}]^{-r}}{\sqrt{v_1 \cdots v_{2r}}} \left[512\overline{d^2} - r\overline{h^2}\right]^{2r-512}\Bigg|_{\hat{\Theta}}, \tag{42}$$

where the definitions of these quantities may be found in Ref. 2. For r sinusoids with exponential decay there are $r!$ different peaks in the posterior probability distribution. The factor of $r!$ accounts for the contributions of these peaks to the integrals over the nonlinear $\mathbf{\Theta}$ parameters.

To compute the approximate posterior probability density for the model, one must locate the maximum of the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters. This optimization step requires good initial estimates of the frequencies. The orthogonality property of sines and cosines may be used to aid in determining the initial estimates, and in finding small resonances in an FID. The following procedure was used to determine the number of resonances in the data and has been found to be very effective in general:

1. Using the signal detection procedures described earlier, compute the evidence in favor of a model containing a constant plus a stationary sinusoid, Eq. (19), compared to a constant model, Eq. (18). This odds ratio may be computed from the posterior probability of the model, Eq. (23) in Ref. 2, if the variances are known, or from Eq. (27) from Ref. 2, if the variances are unknown. The posterior probability of the model given the variances Eq. (23) in Ref. 2 is the preferred form, because it is more sensitive to small resonances. The second form of the posterior probability, Eq. (27) in Ref. 2, is more difficult to use; because it must estimate the noise level given the single frequency model, and a single frequency model will not fit multiple decaying sinusoidal data well. Small resonances will initially be placed into the noise, although they will be detected later. This odds ratio is essentially a discrete Fourier-transform power spectrum, with the scale adjusted to an evidence scale.

2. If there are n_1 resonances with more than 100 dB evidence, construct a decaying sinusoidal model, with $r = n_1$ resonances. Resonances with more than 100 dB evidence have odds ratios of better than 10 billion to 1 in favor of the sinusoidal model, and are real effects in the data. If no resonance is above 100 dB, but some resonances have positive evidence, incorporate these "questionable" resonances into the model one at a time. If no resonances have positive evidence, then all resonances have probably been located and the procedure is terminated. The initial estimate of the decay rate constant is almost irrelevant. Pick an estimate that is reasonable: for example, for exponential decay use $3/T$, where T is the total sampling time.

3. Using a global-optimization routine, locate the maximum of the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters. The algorithm used in this simulation is a modification of the Levenberg-Marquardt method. The principle difference between it and the Levenberg-Marquardt method is that the posterior probability is maximized; χ^2 is not minimized. The values of the parameters that maximize the posterior probability are not equal to the values that minimize χ^2 in general, although for a single FID, using uninformative prior probabilities, they are. Thus, minimum χ^2 represents an approximation to maximum posterior probability that is not always valid. As with the discrete Fourier transform, unless one understands the limits under which the approximation is valid, one can misuse what is otherwise a valuable tool.

4. After locating the maximum of the posterior probability of the nonlinear $\mathbf{\Theta}$ parameters, compute the posterior probability of this model. If the posterior probability of the model increases, keep this model; otherwise reject the added component and test the remaining questionable resonances.

5. Take the difference between the data and the model and repeat this procedure using the residuals as the data in step (1). The first pass through this procedure will typically find all of the large resonances in the FID. Repeated passes will find small resonances, and resonances that are too close for a discrete Fourier transform to resolve. If on the second pass through the procedure one finds n_2 new resonances, one adds these resonances to the model, creating a model with $r = n_1 + n_2$ resonances. Steps (3) through (5) are repeated until no additional resonances can be found in the residuals.

This procedure is not foolproof. It relies on the orthogonal property of sines and cosines to leave well separated resonances in the residuals. When resonances are very close, or resonances with very small signal-to-noise ratio are present, sometimes no evidence for these resonances will show up in the residuals, and the procedure will fail. Implementing this procedure as an automated program is straightforward, and modifying the procedure to account for very rapidly decaying nuisance resonances is also straightforward. Additionally, the computation of the sine and cosine transforms may be done using procedures that run in $N \log(N)$ time, so the calculations may be done relatively quickly. But care must be taken in using any general routine such as this one. This routine represents the answer to a specific question, and sometimes that question may not be appropriate.

This procedure was used to determine the number of resonances in the ethyl ether data shown in Fig. 6A, with two changes. These changes were implemented in order to obtain the information presented in Fig. 7; there was nothing in the ethyl ether data that required these changes. The following changes were made. The automated program was forced to update the model one resonance at a time so the probability of each spectral line could be obtained and the three nuisance resonances were always included in the model. At each step in the procedure, the most probable frequency in the residuals was used as the initial estimate of the next frequency in the optimization step. The procedure was repeated until no additional effect could be found in the residuals. A plot of $\log_{10}[P(r|D,I)]$ is shown in Fig. 7A. The posterior probability rises 923 orders of magnitude and reaches a peak at the seven-frequency model.

To run the eight-frequency model, an initial estimate of the eighth frequency has to be obtained. As described above, these estimates are obtained by taking the difference between the data and the model (here, the model is $7 + 3 = 10$ exponentially decaying sinusoidal frequency model), and computing the odds in favor of a resonance in the residuals. This odds ratio for the residuals using a model with $7 + 3 = 10$ exponentially decaying sinusoidal frequencies is show in Fig. 7C. For all values of the frequency, the odds ratio is always in favor of the constant model; thus there was no evidence in the residuals for an additional resonance, and the procedure terminated. The probability of the number of spectral lines was then normalized over the seven models shown. This normalized posterior probability of the number of spectral lines is shown in Fig. 7B.

Note, that an eight frequency model was never actually tested by the automated procedure; so it is not known if the posterior probability of an eight resonance model will go up or not. All that is actually known is that there was no remaining evidence in the residuals for additional sinusoids. To do this problem correctly one must postulate the 8 plus 3 nuisance resonance model, and then compute the posterior probability by performing the integrals over the 11 frequencies and 11 decay rates. This is obviously very difficult, and because the odds ratio in Fig. 7C strongly indicates no additional effects are in the data, the assumptions that went into the Gaussian approximation are becoming questionable. In the preceding paper [2], when this happened, the integrals over the nonlinear decay rate constants were done numerically. Here, when the number of resonances in the model exceeds the number in the data, the same problem occurs; but it is not nearly as bad. The orthogonality property of sines and cosines is responsible for this. In determining the number of sinusoids in the data, the extra sinusoid simply expands the noise, and because most of the posterior probability is concentrated around the highest noise peak, the Gaussian approximation works to order of magnitude. The penalty against the more complex model is typically four orders of magnitude in these data, so if the approximate calculation is to order of magnitude, it hardly matters if the model is ruled out at 30,000 to 1 or 10,000 to 1 against. Either way the model will be rejected.

In addition to the concerns about the Gaussian approximation breaking down, one should also be concerned about the decaying exponential sinusoidal model. This model is at best an approximation. There could be other small effects in the data. There is some theoretical reason to expect that the decay will go into a power law for very late times and that radiative damping could change the local magnetic field in a time dependent way that could introduce a small chirp. Additionally small changes in the magnetic field could introduce other effects. If one looks too closely, one may find "frequencies" that correspond to expanding these effects on the decaying exponential sinusoidal model. That is not to say that one should not look, but if the only question one asks is, "How may decaying exponential sinusoids does it takes to expand the signal down to the noise?" probability theory will answer this question, and if the data do not contain exponentially decaying sinusoids, one will get an optimal answer to an inappropriate question.

With this in mind, note that a seven-independent-frequency model is not the optimal spectral model for this FID, because a triplet and quartet have very specific meanings in NMR. The triplet should have three equally spaced frequencies, with amplitudes in a ratio 1:2:1. There should be a single phase, and the resonances should decay at the same exponential rate. The quartet should also have equally spaced frequencies, with amplitudes in a ratio of 1:3:3:1; there should be a single phase, and they should decay with the same exponential rate. Last, the spacing of the frequencies in the triplet should be the same as the spacing of the frequencies in the quartet. Thus, there is a great deal of additional prior information that can be incorporated into this spectral calculation.

This additional information was used to demonstrate probability theory's ability to select the correct model even when the "true" model does not fit the data as well as others tested. The results of this calculation are summarized in Table 1. Each table entry represents a gradual simplification of the model starting with entry 1, which takes each resonance line as an independent frequency. This is the model used to determine the number of resonances in the FID. The last entry in the table is a triplet and a quartet model. Column three describes the model used for the three-frequency

Figure 7: Probability Of The Number Of Spectral Lines

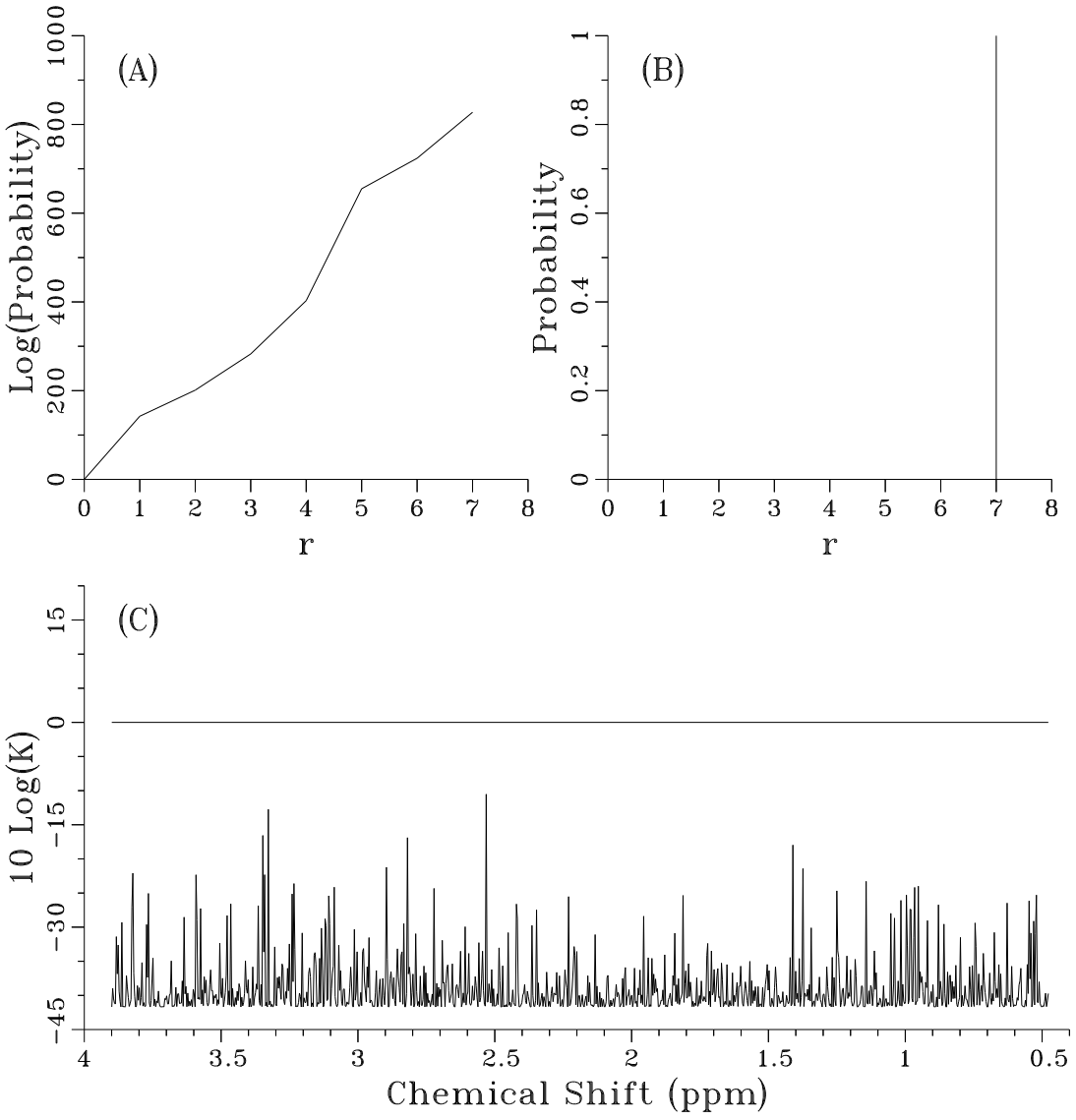

Fig. 7. The base 10 logarithm of the posterior probability of the number of spectral line components is plotted in (A). The normalization was set so that the base 10 logarithm of the three nuisance resonances was zero. The posterior probability increases over 923 orders of magnitude and reaches a maximum at seven frequencies. To determine an initial estimate for the eighth frequency, the evidence in favor of another resonance in the residual was computed (C). Because there is no positive evidence for another frequency, the search for spectral lines terminated with seven. The fully normalized posterior probability of the number of resonances has been plotted in (B). The probability is concentrated on the seven frequency model; this model has a probability of one to six decimal places.

Table 1: Model Selection

Model number	Base 10 Log $P(f_j \mid D, I)$	3 freq. region	4 freq. region	$< \sigma >$	Total parameters
1	0.0	3F	4F	464.2	41
2	2.5	3F	1D+2F	465.7	38
3	2.8	1D+1F	4F	466.2	38
4	4.2	3F	2D	467.7	35
5	4.8	1D+1F	1D+2F	467.6	35
phase coherent	5.8	3F	4F	468.2	35
6	7.7	T	4F	467.2	34
7	9.5	1D+1F	2D	469.6	32
8	9.9	T	1D+2F	468.6	31
9	10.3	3F	Q	468.6	30
10	11.8	T	2D	470.6	28
11	12.6	1D+1F	Q	470.5	27
12	18.0	T	Q	471.5	23

Note. A series of models beginning with model 1 (10 frequencies total = 7 in the region of interest + 3 nuisance frequencies) and ending with model 12 (a quartet + a triplet + 3 nuisance frequencies) were fit to the data. See text for a description of how to determine the models. The standard deviation of the noise, the number of parameters, and the posterior probability of each model were computed. Notice that model 1 fits the data the best (it has the smallest estimated standard deviation of the noise), and it also has the most parameters; while the model which least fits the data is model 12, the model with the fewest parameters. But in Bayesian probability theory additional parameters carry a penalty; unless they improve the fit more than what one would expect from fitting the noise the posterior probability does not increase. Thus, even though model 1 fits the data best, it is 18 orders of magnitude less probable than model 12.

region, and column four describes the model used for the four-frequency region. The abbreviation "**F**" stands for a single exponentially decaying sinusoidal frequency model; "**D,**" a doublet model; "**T,**" a triplet model; and "**Q,**" a quartet model. For example, the model for Table 1 entry 8 has the four-frequency region modeled as two independent frequency components plus a doublet, and it has the three-frequency region modeled as a triplet.

For each of the models, the base 10 logarithm of the posterior probability of the model, the estimated standard deviation of the noise, and the total number of parameters are shown. The normalization was set so that the base 10 logarithm of the posterior probability of Table 1 entry 1 was zero. Table 1 is ordered in increasing posterior probability order. The model that fits the data the best is entry 1 – it has the smallest estimated standard deviation of the noise. The model that fits the data the worst is entry 12, the quartet and triplet model. The model labeled "phase coherent" is a seven exponentially decaying sinusoidal model with a single phase. Phase coherence is an important piece of information and should be included in NMR models whenever possible. The triplet and quartet model, entry 12, is some 18 orders of magnitude more probable than the seven independent exponentially decaying sinusoidal model, entry 1. The second most probable model is entry 11, and it is some 5 orders of magnitude less probable than the triplet and quartet model, entry 12. Thus, probability theory strongly indicates that the model which "best" accounts for these data is the model that takes the three-frequency region to be a triplet and the four-frequency region to be a quartet, despite the fact that of all of the models tested, this model fits the data the least. As has been emphasized before, additional parameters carry a penalty; if those parameters do not expand the data better than what one would expect from fitting the noise, then the prior

will effectively eliminate the model from consideration. This effect is easily seen in Table 1; the posterior probability increases every time a parameter is removed.

Now that the correct model is known, the procedures derived in [1] may be applied to estimate the parameters. The separation Δ between the center of the triplet and the center of the quartet is

$$(\Delta)_{\text{est}} = 2.144750 \pm 0.000025 \text{ ppm} \tag{43}$$

at two standard deviations. The coupling constant J_T for the triplet is

$$(J_T)_{\text{est}} = 6.979 \pm 0.013 \text{ Hz} \tag{44}$$

at two standard deviations and the coupling constant J_Q for the quartet is

$$(J_Q)_{\text{est}} = 6.97 \pm 0.02 \text{ Hz} \tag{45}$$

at two standard deviations. This last estimate is the worst of the three, and it is some 6.25 times better than the Rayleigh criteria when all 6848 data values per channel are considered, and it is some 42 times better than the Rayleigh criteria when 512 data values per channel are considered.

Standard theory indicates that ratio of the total intensity of the triplet to the total intensity of the quartet should be 3/2. Thus the ratio of the amplitude of the small component of the triplet, B_T, to the amplitude of the small component of the quartet, B_Q, should be given by

$$\begin{aligned} 2(1+2+1)B_T &= 3(1+3+3+1)B_Q \\ \frac{B_T}{B_Q} &= 3.0. \end{aligned} \tag{46}$$

The estimated value of this ratio is

$$\frac{B_T}{B_Q} = 3.05 \pm 0.06 \tag{47}$$

at two standard deviations.

But these results were calculated from a model that allowed J_Q to be different from J_T, and it allowed B_Q to be in an arbitrary ratio with B_T. Physics clearly indicates that $J_Q = J_T$, and $B_T = 3B_Q$. This suggests that a 13th model, which incorporates this information, should be considered. The posterior probability of this 13th model was computed, and it was found to be five orders of magnitude higher than the probability of the 12th model. In this model there is only a single coupling constant J for both the triplet and quartet. This coupling constant is estimated to be

$$J = 6.977 \pm 0.011 \text{ Hz} \tag{48}$$

at two standard deviations, and the separation between the triplet and the quartet is given by

$$(\Delta)_{\text{est}} = 2.144799 \pm 0.000037 \text{ ppm}. \tag{49}$$

In this example, probability theory clearly indicates that all systematic effects in these data are in exact agreement with current theory. Indeed, this last model puts extremely strong constraints on any alternate theory. For probability theory to prefer an alternate theory, it must either fit the data much better using the same number of parameters or fit the data similarly using fewer parameters. But none of the models in Table 1 could significantly improve the fit. This means that the alternative theory must fit the data at least as well using fewer parameters. This will be difficult, because the 13th model contained only 5 parameters (one amplitude, one phase, the center frequency for the triplet, the center frequency of the quartet, and one exponential decay rate constant). The only reasonable possibility is that the alternative theory could specify the separation

frequency for the triplet and the quartet. But this separation frequency depends on the molecular dynamics, and this is very difficult to calculate.

In this calculation there were a total of 13 models tested, 14 if one includes the phase coherent model. The parameter estimation procedures developed in the earlier paper [1], could have been applied to any of these 14 models. Every one of these models would produce parameter estimates that may or may not agree with each other. If they disagree with each other, they do not disagree in any relevant sense, they simply answer different questions. Probability theory includes Fourier transforms, maximum likelihood, linear prediction, and least squares as special cases; if any of these procedures gives misleading or incorrect results, it is almost a given that the model assumed in the calculation was inappropriate for the data being analyzed. Before asking a parameter estimation question, one must be absolutely sure that the correct model has been incorporated into the calculation. Thus, parameter estimation questions are the last questions one should ask, not the first. It is only after one has carefully checked the alternatives that one can be reasonably sure of obtaining reasonable parameter estimates.

Summary and Conclusions

In these three papers, full Bayesian probability theory has been applied to the problems of parameter estimation, signal detection, model selection, and spectral estimation. In the previous paper [1], the procedures needed to estimate parameters using quadrature NMR models were derived. In the preceding paper [2], the procedures needed to detect signals and compare models were developed. In this paper, those procedures were used to demonstrate the relationship between Bayesian probability theory and the discrete Fourier transform, to detect small signals, to estimate parameters, and to test various alternative models against theory. These demonstrations illustrate that Bayesian probability theory contains a quantitative statement of Ockham's razor: postulate various theories, compare these to experiment, and when two theories explain the data equally well, prefer the simpler model.

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

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

[3] A. Schuster, *Proc. R. Soc. London* **77,** 136 (1905).

[4] E. D. Becker, J. A. Ferretti, and P. N. Gambhir, *Anal. Chem.* **51,** 1413 (1979).

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

[6] G. L. Bretthorst, "Maximum Entropy and Bayesian Methods" (J. Skilling, Ed.), p. 261, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

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

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

[9] Lord Rayleigh, *Philos. Maga.* **5,** 261 (1879).