Bayesian Analysis Users Guide
Release 4.00, Manual Version 1

G. Larry Bretthorst
Biomedical MR Laboratory
Washington University School Of Medicine,
Campus Box 8227
Room 2313, East Bldg.,
4525 Scott Ave.
St. Louis MO 63110
http://bayes.wustl.edu
Email: larry@bayes.wustl.edu

October 21, 2016

Contents

List of Figures

List of Tables

Chapter 4

An Introduction to Bayesian Probability Theory

This Chapter is a tutorial on Bayesian probability theory. In it the procedures and principles needed to apply probability theory as extended logic will be discussed in detail. Primarily these procedures and principles will be illustrated using an example taken from NMR, the single frequency estimation problem. In this example we illustrate the assignment of probabilities and the use of uninformative prior probabilities. While we attempt to explain all of the steps in this calculation in detail, some familiarity with higher mathematics and Bayesian probability theory is assumed. For an introduction to probability theory see [32, 64, 66, 40]; for a derivation of the rules of probability theory see Jaynes [32, 31], and for an introduction to parameter estimation using probability theory see Bretthorst [3]. In this tutorial the sum and product rules of probability theory will be given and no attempt will be made to derive them. However, if one wishes to represent degrees of belief as real numbers, reason consistently, and have probability theory reduce to Aristotelian logic when the truth of the hypotheses are known, then the sum and product rules are the unique rules for conducting inference. For an extensive discussion of these points and much more, see Jaynes [32].

4.1 The Rules of Probability Theory

There are two basic rules for manipulating probabilities, the product rule and the sum rule; all other rules may be derived from these. If A, B, and C stand for three hypotheses, then the product rule states

$$P(AB|C) = P(A|C)P(B|AC), \tag{4.1}$$

where $P(AB|C)$ is the joint probability that "A and B are true given that C is true," $P(A|C)$ is the probability that "A is true given C is true," and $P(B|AC)$ is the probability that "B is true given that both A and C are true." The notation "$|C$)" means conditional on the truth of hypothesis C. In probability theory *all* probabilities are conditional. The notation $P(A)$ is not used to stand for the probability for a hypothesis, because it does not make sense until the evidence on which it is based is given. Anyone using such notation either does not understand that all knowledge is conditional, i.e., contextual, or is being extremely careless with notation. In either case, one should

be careful when interpreting such material. For more on this point see Jeffreys [33] and Jaynes [32].

In Aristotelian logic, the hypothesis "A and B" is the same as "B and A," so the numerical value assigned to the probabilities for these hypotheses must be the same. The order may be rearranged in the product rule, Eq. (4.1), to obtain:

$$P(BA|C) = P(B|C)P(A|BC), \tag{4.2}$$

which may be combined with Eq. (4.1) to obtain a seemingly trivial result

$$P(A|BC) = \frac{P(A|C)P(B|AC)}{P(B|C)}. \tag{4.3}$$

This is Bayes' theorem. It is named after Rev. Thomas Bayes, an 18th century mathematician who derived a special case of this theorem. Bayes' calculations [1] were published in 1763, two years after his death. Exactly what Bayes intended to do with the calculation, if anything, still remains a mystery today. However, this theorem, as generalized by Laplace [36], is the basic starting point for inference problems using probability theory as logic.

The second rule of probability theory, the sum rule, relates to the probability for "A or B." The operation "or" is indicated by a "+" inside a probability symbol. The sum rule states that given three hypotheses A, B, and C, the probability for "A or B given C" is

$$P(A + B|C) = P(A|C) + P(B|C) - P(AB|C). \tag{4.4}$$

If the hypotheses A and B are mutually exclusive, that is the probability $P(AB|C)$ is zero, the sum rule becomes:

$$P(A + B|C) = P(A|C) + P(B|C). \tag{4.5}$$

The sum rule is especially useful because it allows one to investigate an interesting hypothesis while removing an uninteresting or nuisance hypothesis from consideration.

To illustrate how to use the sum rule to eliminate nuisance hypotheses, suppose D stands for the data, ω the hypothesis "the frequency of a sinusoidal oscillation was ω," and B the hypothesis "the amplitude of the sinusoid was B." Now suppose one wishes to compute the probability for the frequency given the data, $P(\omega|D)$, but the amplitude B is present and must be dealt with. The way to proceed is to compute the joint probability for the frequency and the amplitude given the data, and then use the sum rule to eliminate the amplitude from consideration. Suppose, for argument's sake, the amplitude B could take on only one of two mutually exclusive values $B \in \{B_1, B_2\}$. If one computes the probability for the frequency and (B_1 or B_2) given the data one has

$$P(\omega|D) \equiv P(\omega[B_1 + B_2]|D) = P(\omega B_1|D) + P(\omega B_2|D). \tag{4.6}$$

This probability distribution summarizes all of the information in the data relevant to the estimation of the frequency ω. The probability $P(\omega|D)$ is called the marginal probability for the frequency ω given the data D.

The marginal probability $P(\omega|D)$ does not depend on the amplitudes at all. To see this, the product rule is applied to the right-hand side of Eq. (4.6) to obtain

$$P(\omega|D) = P(B_1|D)P(\omega|B_1D) \; + \; P(B_2|D)P(\omega|B_2D) \tag{4.7}$$

but

$$P(B_1|D) + P(B_2|D) = 1 \tag{4.8}$$

because the hypotheses are exhaustive. So the probability for the frequency ω is a weighted average of the probability for the frequency given that one knows the various amplitudes. The weights are just the probability that each of the amplitudes is the correct one. Of course, the amplitude could take on more than two values; for example if $B \in \{B_1, \cdots, B_m\}$, then the marginal probability distribution becomes

$$P(\omega|D) = \sum_{j=1}^{m} P(\omega B_j|D), \tag{4.9}$$

provided the amplitudes are mutually exclusive and exhaustive. In many problems, the hypotheses B could take on a continuum of values, but *as long as only one value of B is realized when the data were taken* the sum rule becomes

$$P(\omega|D) = \int dB P(\omega B|D) \tag{4.10}$$

where probabilities that have had one or more parameters removed by integration are frequently called marginal probabilities. Note that the B inside the probability symbols refers to the hypothesis; while the B appearing outside of the probability symbols is a number or index. A notation could be developed to stress this distinction, but in most cases the meaning is apparent from the context.

The sum and integral appearing in Eqs. (4.9,4.10) are over a set of mutually exclusive and exhaustive hypotheses. If the hypotheses are not mutually exclusive, one simply uses Eq. (4.4). However, if the hypotheses are *not* exhaustive, the sum rule *cannot* be used to eliminate nuisance hypotheses. To illustrate this, suppose the hypotheses, $B \in \{B, \cdots, B_m\}$, are mutually exclusive, but not exhaustive. The hypotheses B could represent various explanations of some experiment, but it is always possible that there is something else operating in the experiment that the hypotheses B do not account for. Let us designate this as

$$\text{SE} \equiv \text{“Something Else not yet thought of.”} \tag{4.11}$$

The set of hypotheses $\{B, \text{SE}\}$ is now complete, so the sum rule may be applied. Computing the probability for the hypothesis B_i conditional on some data D and the information I, where I stands for the knowledge that amplitudes B are not exhaustive, one obtains

$$P(B_i|DI) = \frac{P(B_i|I)P(D|B_iI)}{P(D|I)} \tag{4.12}$$

and for SE

$$P(\text{SE}|DI) = \frac{P(\text{SE}|I)P(D|\text{SE}I)}{P(D|I)}. \tag{4.13}$$

The denominator is the same in both these equation and is given by

$$\begin{aligned} P(D|I) &= \sum_{i=1}^{m} P(DB_i|I) + P(D\text{SE}|I) \\ &= \sum_{i=1}^{m} P(B_i|I)P(D|B_iI) + P(\text{SE}|I)P(D|\text{SE}I). \end{aligned} \tag{4.14}$$

But this is indeterminate because SE has not been specified, and therefore the likelihood, $P(D|\text{SE}I)$, is indeterminate even if the prior probability $P(\text{SE}|I)$, is known. However, the relative probabilities

$P(B_i|DI)/P(B_j|DI)$ are well defined because the indeterminacy cancels out. So there are two choices: either *ignore* SE and thereby assume the hypotheses B are complete or *specify* SE, thereby completing the set of hypotheses.

4.2 Assigning Probabilities

The product rule and the sum rule are used to indicate relationships between probabilities. These rules are not sufficient to conduct inference because, ultimately, the "numerical values" of the probabilities must be known. Thus the rules for manipulating probabilities must be supplemented by rules for assigning numerical values to probabilities. The historical lack of these supplementary rules is one of the major reasons why probability theory, as formulated by Laplace, was rejected in the late part of the 19th century. To assign any probability there is ultimately only one way, logical analysis, i.e., non-self-contradictory analysis of the available information. The difficulty is to incorporate only the information one actually possesses without making gratuitous assumptions about things one does not know. A number of procedures have been developed that accomplish this task: Logical analysis may be applied directly to the sum and product rules to yield probabilities [32]. Logical analysis may be used to exploit the group invariance of a problem [27]. Logical analysis may be used to ensure consistency when uninteresting or nuisance parameter are marginalized from probability distributions [29]. And last, logical analysis may be applied in the form of the principle of maximum entropy to yield probabilities [66, 27, 28, 60, 59]. Of these techniques the principle of maximum entropy is probably the most powerful, and in this tutorial it will be used to assign most probabilities.

In this tutorial there are three different types of information that must be incorporated into probability assignments: parameter ranges, knowledge of the mean and standard deviation estimates of several quantities, and some properties of the noise. Their assignment differs only in the types of information available. In the first case, the principle of maximum entropy leads to a bounded uniform prior probability. In the second and third cases, it leads to a Gaussian probability distribution. To understand the principle of maximum entropy and how these probability assignments come about, suppose one must assign a probability distribution for the ith value of a parameter given the information I. This probability is denoted $P(i|I)$ $(1 \leq i \leq m)$. The Shannon entropy, defined as

$$H \equiv -\sum_{i=1}^{m} P(i|I) \log P(i|I), \tag{4.15}$$

is a measure of the amount of ignorance (uncertainty) in this probability distribution [58]. Shannon's entropy is based on a qualitative requirement, the entropy should be monotonically increasing for increasing ignorance, plus the requirement that the measure be consistent. The principle of maximum entropy then states that if one has some information I, one can assign the probability distribution, $P(i|I)$, that contains only the information I by maximizing H subject to the information (constraints) represented by I. Because H measures the amount of ignorance in the probability distribution, assigning a probability distribution that has maximum entropy yields a distribution that is least informative (maximally ignorant) while remaining consistent with the information I: the probability distribution, $P(i|I)$, contains only the information I, and does not contain any additional information not already implicit in I [60, 59].

To demonstrate its use, suppose that one must assign $P(i|I)$ and nothing is known except that

the set of hypotheses is mutually exclusive and exhaustive. Applying the sum rule one obtains

$$\sum_{i=1}^{m} P(i|I) = 1. \tag{4.16}$$

This equation may be written

$$\sum_{i=1}^{m} P(i|I) - 1 = 0 \tag{4.17}$$

and because this equation sums to zero, any multiple of it may be added to the entropy of $P(i|I)$ without changing its value:

$$H = -\sum_{i=1}^{m} P(i|I) \log P(i|I) + \beta \left[1 - \sum_{i=1}^{m} P(i|I) \right]. \tag{4.18}$$

The constant β is called a Lagrange multiplier. But the probabilities $P(i|I)$ and the Lagrange multiplier β are not known; they must be assigned. To assign them, H is constrained to be a maximum with respect to variations in all the unknown quantities. This maximum is located by differentiating H with respect to both $P(k|I)$ and β, and then setting the derivatives equal to zero. Here there are m unknown probabilities and one unknown Lagrange multiplier. But when the derivatives are taken, there will be $m + 1$ equations; thus all of the unknowns may be determined. Taking the derivative with respect to $P(k|I)$, one obtains

$$\log P(k|I) + 1 + \beta = 0, \tag{4.19}$$

and taking the derivative with respect to β returns the constraint equation

$$1 - \sum_{i=1}^{m} P(i|I) = 0. \tag{4.20}$$

Solving this system of equations, one finds

$$P(i|I) = \frac{1}{m} \qquad \text{and} \qquad \beta = \log m - 1. \tag{4.21}$$

When nothing is known except the specification of the hypotheses, the principle of maximum entropy reduces to Laplace's principle of indifference [36]. But the principle of maximum entropy is much more general because it allows one to incorporate many different types of information.

As noted earlier, in the inference problem addressed in this chapter, there are three different types of information to be incorporated into probability assignments. The specification of parameter ranges occurs when the prior probabilities for various location parameters appearing in the calculation must be assigned. (A location parameter is a parameter that appears linearly in the model equation.) For these location parameters, the principle of maximum entropy leads to the assignment of a bounded uniform prior probability. However, care must be taken because most of these parameters are continuous and *the rules and procedures given in this tutorial are strictly valid only for finite, discrete probability distributions.* The concept of a probability for a hypothesis containing a continuous parameter, a probability density function, only makes sense when thought of as a limit. If the

preceding calculations are repeated and the number of hypotheses are allowed to grow infinitely, one will automatically arrive at a valid result as long as all probabilities remain finite and normalized. Additionally, the direct introduction of an infinity into any mathematical calculation is ill-advised under any conditions. Such an introduction presupposes the limit already accomplished and this procedure will cause problems whenever any question is asked that depends on how the limit was taken. For more on the types of problems this can cause see Jaynes [29], and for a much more extensive discussion of this point see Jaynes [32]. As it turns out, continuous parameters are not usually a problem, provided one always uses normalized probabilities. In this tutorial, continuous parameters will be used, but their prior probabilities will be normalized and the prior ranges will never be allowed to go to infinity without taking a limit.

The second type of information that must be incorporated into a probability assignment is knowledge of the mean and standard deviation of a parameter estimate. It is a straightforward exercise to show that, in this case, the principle of maximum entropy leads to a Gaussian distribution.

The third type of information that must be incorporated into a probability assignment is information about the errors or noise in the data. The probability that must be assigned is denoted $P(D|LI)$, the probability for the data given that the signal is L, where the data, D, is a joint hypothesis of the form, $D \equiv \{d_1 \dots d_N\}$; d_j are the individual data items, and N is the number of data values. If the signal L is given time t_j, then

$$d_j - L(t_j) = n_j \tag{4.22}$$

assuming that the noise is additive, and n_j is the misfit between the data and the model and is usually called the noise. Thus the probability for the data can be assigned if one can assign a probability for the noise.

To assign a probability for the noise, the question one must ask is, *what properties of the noise are to be used in the calculations*? For example, should the results of the calculations depend on correlations? If so, which of the many different types of correlations should the results depend on? There are second order correlations of the form

$$\rho'_s = \frac{1}{N-s} \sum_{j=1}^{N-s} n_j n_{j+s}, \tag{4.23}$$

where s is a measure of the correlation distance, as well as third, fourth, and higher order correlations. In addition to correlations, should the results depend on the moments of the noise? If so, on which moments should they depend? There are many different types of moments. There are power law moments of the form

$$\sigma'_s = \frac{1}{N} \sum_{j=1}^{N} n_j^s, \tag{4.24}$$

as well as moments of arbitrary functions, and a host of others.

The probability that must be assigned is the probability that one should obtain the data D, but from Eq. (4.22) this is just the probability for noise $P(e_1 \cdots e_N|I')$, where e_j stands for a hypothesis of the form "the value of the noise at time t_j was e_j, when the data were taken." The quantity e_j is an index that ranges over all valid values of the errors; while the probability for the noise, $P(e_1 \cdots e_N|I')$, assigns a reasonable degree of belief to a particular set of noise values. For the

probability for the noise to be consistent with correlations it must have the property that

$$\rho_s = \langle e_j e_{j+s} \rangle \equiv \frac{1}{N-s} \sum_{j=1}^{N-s} \int de_1 \cdots de_N \; e_j e_{j+s} P(e_1 \cdots e_N | I') \tag{4.25}$$

and for it to be consistent with the power law moments it must have the additional property that

$$\sigma_s = \langle e^s \rangle \equiv \frac{1}{N} \sum_{j=1}^{N} \int de_1 \cdots de_N \; e_j^s P(e_1 \cdots e_N | I') \tag{4.26}$$

where the notation $\langle\rangle$ denote mean averages over the probability density function.

In Eq. (4.23) and Eq. (4.24), the symbols ρ'_s and σ'_s were used to denote means or averages over the sample noise. These averages are the sample correlation coefficients and moments and they represent states of nature. In Eq. (4.25) and Eq. (4.26), the symbols ρ_s and σ_s are used to denote mean averages over the probability for the noise, and they represent states of knowledge. To use information in a maximum entropy calculation, that information must be known.

Assuming that none of these quantities are known, how can the principle of maximum entropy be used? Its use requires known information, and unless at least some of the ρ'_s and σ'_s are known, it would appear that maximum entropy cannot be used. However, this description of the problem is not what probability theory asks us to do. Probability theory asks us to assign $P(e_1 \cdots e_N | I')$, where I' represents the information on which this probability is based. Suppose for the sake of argument that that information is a mean, ν, and standard deviation, σ, then what probability theory asks us to assign is $P(e_1 \cdots e_N | \nu\sigma)$. This expression should be read as the joint probability for all the errors given that the mean and standard deviation of the errors. According to probability theory, in the process of assigning the probability for the errors, we are to assume that both ν and σ are known or given values. This is a very different state of knowledge from knowing that the mean and standard deviation of the sampling distribution are ν and σ. If we happen to actually know these values, then there is less work to do when applying the rules of probability theory. However, if their values are unknown, we still seek the least informative probability density function that is consistent with a fixed or given mean and standard deviation. The rules of probability theory are then used to eliminate these unknown nuisance hypotheses from the final probability density functions.

But which of these constraints should be used? The answer was implied earlier by the way the question was originally posed: what *properties* of the errors are to be used in the calculations? The class of maximum entropy probability distributions is the class of all probability density functions for which sufficient statistics exist. A sufficient statistic is a function of the data that summarizes all of the information in the data relevant to the problem being solved. These sufficient statistics are the sample moments that correspond to the constraints that were used in the maximum entropy calculation. For example, suppose we used the first three correlation coefficients, ρ_1, ρ_2, and ρ_3, as defined by Eq. (4.25) in a maximum entropy calculation, then the parameter estimates will depend only on the first three correlation coefficients of the data and our uncertainty in those estimates will depend on ρ_1, ρ_2, and ρ_3 if they are known, and on the first three correlation coefficients of the errors if ρ_1, ρ_2, and ρ_3 are not known. *All* other properties of the errors have been made irrelevant by the use of maximum entropy. So the real question becomes, what does one know about the errors before seeing the data? If there is information that suggests the errors may be correlated, then by all means a correlation constraint should be included. Additionally, if one has information that suggests the higher moments of the noise can deviate significantly from what one would expect from

a Gaussian distribution, then again a constraint on the higher moments should be included. But if one has no information about higher moments and correlations, then one is always better off to leave those constraints out of the maximum entropy calculation, because the resulting probability density function will have higher entropy. Higher entropy distributions are by definition less informative and therefore make more conservative estimates of the parameters. Consequently, these higher entropy probability density functions are applicable under a much wider variety of circumstances, and typically they are simpler and easier to use than distributions having lower entropy.

In assigning the probability density function for the errors, it will be assumed that our parameter estimates are to depend only on the mean and variance of the errors in the data. The appropriate constraints necessary are on the first and second moments of the probability density function. The constraint on the first moment is given by

$$\nu = \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j P(e_1 \cdots e_N | I') \tag{4.27}$$

and by

$$\sigma^2 + \nu^2 = \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j^2 P(e_1 \cdots e_N | I') \tag{4.28}$$

for the second moment, where ν and σ^2 are the fixed or given values of the mean and variance. Note the second moment of the probability distribution, Eq. (4.28), is written as $\sigma^2 + \nu^2$, to make the resulting probability density function come out in standard notation.

We seek the probability density function that has highest entropy for a fixed or given value of σ^2 and ν. To find this distribution Eq. (4.27) and Eq. (4.28) are rewritten so they sum to zero:

$$\nu - \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j P(e_1 \cdots e_N | I') = 0, \tag{4.29}$$

and

$$\sigma^2 + \nu^2 - \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j^2 P(e_1 \cdots e_N | I') = 0. \tag{4.30}$$

Additionally, the probability for finding the noise values somewhere in the valid range of values is one:

$$1 - \int de_1 \cdots de_N P(e_1 \cdots e_N | I') = 0. \tag{4.31}$$

Because Eq. (4.29) through Eq. (4.31), sum to zero, each may be multiplied by a constant and added

to the entropy of this probability density function without changing its value, one obtains

$$\begin{aligned} H &= -\int de_1 \cdots de_N P(e_1 \cdots e_N|I') \log P(e_1 \cdots e_N|I') \\ &+ \beta \left[1 - \int de_1 \cdots de_N P(e_1 \cdots e_N|I')\right] \\ &+ \delta \left[\nu - \tfrac{1}{N}\sum_{j=1}^{N} \int de_1 \cdots de_N \; e_j P(e_1 \cdots e_N|I')\right] \\ &+ \lambda \left[\sigma^2 + \nu^2 - \tfrac{1}{N}\sum_{j=1}^{N} \int de_1 \cdots de_N \; e_j^2 P(e_1 \cdots e_N|I')\right] \end{aligned} \tag{4.32}$$

where β, δ, and λ are Lagrange multipliers. To obtain the maximum entropy distribution, this expression is maximized with respect to variations in β, δ, λ, and $P(e'_1 \cdots e'_N|I')$. After a little algebra, one obtains

$$P(e_1 \cdots e_N|\nu\sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{j=1}^{N} \frac{(e_j - \nu)^2}{2\sigma^2}\right\}, \tag{4.33}$$

where we have replaced I' by the information actually used in assigning this probability density function,

$$\lambda = \frac{N}{2\sigma^2}, \qquad \delta = -\frac{N\nu}{\sigma^2}, \quad \text{and} \quad \beta = \frac{N}{2}\left[\log(2\pi\sigma^2) + \frac{\nu^2}{\sigma^2}\right] - 1. \tag{4.34}$$

There are several interesting points to note about this probability density function. First, this is a Gaussian distribution. However, the fact that the prior probability for the errors has been assigned a Gaussian makes no statement about the sampling distribution of the errors; rather it says only that for a fixed value of the mean and variance the probability density function for the errors should be maximally uninformative and that maximally uninformative distribution happens to be a Gaussian. Second, this probability assignment apparently does not contain correlations. The reason for this is that a constraint on correlations must lower the entropy. By definition a probability assignment with lower entropy is more informative, and so must make more precise estimates of the parameters. Instead of saying this probability density function does not contain correlations, it would be more correct to say that this probability density function makes allowances for *every possible correlation* that could be present and so is less informative than correlated distributions. Third, if one computes the expected mean value of the moments, one finds

$$\langle e^s \rangle = \exp\left\{-\frac{\nu^2}{2\sigma^2}\right\} \sigma^{2s} \frac{\partial^s}{\partial \nu^s} \exp\left\{\frac{\nu^2}{2\sigma^2}\right\} \quad (s \geq 0) \tag{4.35}$$

which reduces to

$$\langle e^0 \rangle = 1, \qquad \langle e^1 \rangle = \nu, \quad \text{and} \quad \langle e^2 \rangle = \sigma^2 + \nu^2 \tag{4.36}$$

for $s = 0$, $s = 1$, and $s = 2$, just the constraints used to assign the probability density function. Fourth, for a fixed value of the mean and variance this prior probability has highest entropy. Consequently, when parameters are marginalized from probability distributions or when any operation

is performed on them that preserves mean and variance while discarding other information, those probability densities necessarily will move closer and closer to this Gaussian distribution regardless of the initial probability assignment. The Central Limit Theorem is one special case of this phenomenon—see Jaynes [32].

Earlier it was asserted that maximum entropy distributions are the only distributions that have sufficient statistics and that these sufficient statistics are the only properties of the data, and therefore the errors, that are used in estimating parameters. We would like to demonstrate this property explicitly for the Gaussian distribution [32, 12]. Suppose the value of a location parameter is ν_0 and one has a measurement such that

$$d_j = \nu_0 + n_j. \tag{4.37}$$

The hypothesis about which inferences are to be made is of the form "the true value of the mean is ν given the data, D." Assigning a Gaussian as the prior probability for the errors, the likelihood function is then given by

$$P(D|\nu\sigma I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(d_j-\nu)^2\right\}. \tag{4.38}$$

The posterior probability for ν may be written as

$$P(\nu|D\sigma I) \propto (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{N}{2\sigma^2}([\overline{d}-\nu]^2+s^2)\right\} \tag{4.39}$$

where a uniform prior probability was assigned for ν. The mean data value, $\overline{d}$, is given by

$$\overline{d} = \frac{1}{N}\sum_{j=1}^{N} d_j = \nu_0 + \overline{n} \tag{4.40}$$

where $\overline{n}$ is the mean value of the errors. And s^2 is given by

$$s^2 = \overline{d^2} - (\overline{d})^2 = \frac{1}{N}\sum_{j=1}^{N} d_j^2 - \left(\frac{1}{N}\sum_{j=1}^{N} d_j\right)^2 = \overline{n^2} - (\overline{n})^2 \tag{4.41}$$

where $(\overline{n})^2$ is the mean square of the noise vales. From which one obtains

$$(\nu)_{est} = \begin{cases} \overline{d} \pm \sigma/\sqrt{N} & \sigma \text{ known} \\ \overline{d} \pm s/\sqrt{N-3} & \sigma \text{ unknown} \end{cases} \tag{4.42}$$

as the estimate for ν. The actual error, Δ, is given by

$$\Delta = \overline{d} - \nu_0 = \overline{n} \tag{4.43}$$

which depends only on the *mean of the noise values*; while our accuracy estimate depends only on σ if the standard deviation of the noise is known, and *only on the mean and mean-square* of the noise values when the standard deviation of the noise is not known. Thus the underlying

sampling distribution of the noise has completely canceled out and the only property of the errors that survives is the actual mean and mean-square of the noise values. *All* other properties of the errors have been made irrelevant. Exactly the same parameter estimates will result if the underlying sampling distribution of the noise is changed, provided the mean and mean-square of the new sampling distribution is the same, just the properties needed to represent what is actually known about the noise, and to render what is *not* known about it irrelevant. For more on the subject of how maximum entropy probability assignments render the underlying sampling distribution nearly irrelevant see [32] and [12].

In Section 4.1 the sum and product rules of probability theory were given. In Section 4.2 the principle of maximum entropy was used to demonstrate how to assign probabilities that are maximally uninformative while remaining consistent with the given prior information. In the following section a nontrivial parameter estimation problem is given. Each step in the calculation is explained in detail. The example is complex enough to illustrate all of the points of principle that must be faced in more complicated problems, yet sufficiently simple that anyone with a background in calculus should be able to follow the mathematics. In addition, the problem, single frequency estimation, is of interest to the general NMR community and in the process of solving the problem we will uncover a number of new and interesting features about the discrete Fourier transform and the bandwidth of nonuniformly sampled data.

4.3 Example: Parameter Estimation

Probability theory tells one what to believe about a hypothesis C given all of the available information or evidence $E_1 \cdots E_n$. This is done by computing the posterior probability for hypothesis C conditional on all of the evidence $E_1 \cdots E_n$. This posterior probability is represented symbolically by

$$P(C|E_1 \cdots E_n). \tag{43}$$

It is computed from the rules of probability theory by repeated application of the sum and product rules and by assigning the probabilities so indicated. This is a general rule and there are no exceptions to it: *ad hoc devices have no place in probability theory.* Given the statement of a problem, the rules of probability theory take over and will lead every person to the same unique solution, provided each person has exactly the same evidence.

To someone unfamiliar with probability theory, how this is done is not obvious; nor is it obvious what must be done to obtain a problem that is sufficiently well defined to permit the application of probability theory as logic. Consequently, in what follows all of the steps in computing $P(C|E_1 \cdots E_n)$ will be described in detail. To compute the probability for any hypothesis C given some evidence $E_1 \cdots E_n$, there are five basic steps, which are not necessarily independent:

1. *Define The Problem:* State in nonambiguous terms exactly what hypothesis you wish to make inferences about.

2. *State The Model:* Relate the hypothesis of interest to the available evidence $E_1 \cdots E_n$.

3. *Apply Probability Theory:* The probability for hypothesis C conditional on all the available evidence $E_1 \cdots E_n$ is computed from Bayes theorem. The sum rule is then applied to eliminate nuisance hypotheses. The product rule is then repeatedly applied to factor joint probabilities to obtain terms which cannot be further simplified.

4. *Assign The Probabilities:* Using the appropriate procedures, translate the available evidence into numerical values for the indicated probabilities.

5. *Evaluate The Integrals and Sums:* Evaluate the integrals and sums indicated by probability theory. If the indicated calculations cannot be done analytically, then implement the necessary computer codes to evaluate them numerically.

Each of these steps will be systematically illustrated in the following example, the single frequency estimation problem when the data are nonsimultaneously nonsimultaneously sampled.

4.3.1 Define The Problem

The problem to be solved is to estimate the frequency of a decaying sinusoid given some data and whatever prior information we have. The Bayesian calculations presented in this chapter will be for quadrature NMR data that has been sampled at differing times and with differing numbers of data values in each channel. We will derive the solution to the frequency estimation problem given an exponentially decaying sinusoidal model. Then, through a series of simplifications, we will reduce this calculation to the case of frequency estimation for a stationary sinusoid given real (nonquadrature) data. In the process of making these simplifications, we will encounter the Lomb-Scargle periodogram [39, 54, 55, 56], the Schuster periodogram [57] and a weighted power spectrum as sufficient statistics for frequency estimation. Because each sufficient statistic will have been derived from the rules of probability theory we will see the exact conditions under which each is an optimal frequency estimator. Thus, by making these simplifications, we will see how probability theory generalizes the discrete Fourier transform to handle nonuniformly nonsimultaneously sampled data and what is happening to the aliasing phenomenon in these data. Before doing this, we need to understand the discrete Fourier transform and the phenomena of aliasing.

4.3.1.1 The Discrete Fourier Transform

When the data consist of uniformly sampled time domain data containing some type of harmonic oscillations, the discrete Fourier transform is almost universally used as the frequency estimation technique. This is done for a number of reasons, but primarily because the technique is fast and experience has shown that the frequency estimates obtained from it are often very good. The discrete Fourier transform, $\mathcal{F}(f_k)$, is defined as:

$$\mathcal{F}(f_k) = \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\{2\pi f_k t_j \mathbf{i}\} \tag{4.44}$$

where $\mathbf{i} = \sqrt{-1}$, $\mathbf{d}(t_j)$ is the complex discretely sampled data,

$$\mathbf{d}(t_j) = d_R(t_j) + \mathbf{i} d_I(t_j), \tag{4.45}$$

and is composed of real, $d_R(t_j)$, and imaginary, $d_I(t_j)$, data samples; N is the total number of complex data samples and f_k is the frequency. For uniformly sampled data, the times are given by

$$t_j = j\Delta T, \qquad j \in \{0, 1 \ldots N-1\}, \tag{4.46}$$

where ΔT is the dwell time, the time interval between data samples, and the frequencies are given by

$$f_k = \frac{k}{N\Delta T} \qquad k \in \left\{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}\right\}. \tag{4.47}$$

These frequencies are the ones at which a discrete Fourier transform is exactly equal to the continuous Fourier transform of a bandlimited function [49, 48]. The largest frequency interval free of aliases for a bandlimited function is given by

$$-f_{Nc} \leq f \leq f_{Nc} \tag{4.48}$$

and is called the bandwidth. The frequency f_{Nc} is called the Nyquist critical frequency and is given by

$$f_{Nc} = \frac{1}{2\Delta T}. \tag{4.49}$$

Nothing would prohibit one from taking f_k as a continuous variable and evaluating Eq. (4.44) at different frequencies. Indeed, this is exactly what the common practice of zero-padding[1] does. After all, adding zero to a sum does not change that sum, so for any given frequency zero padding has no effect in Eq. (4.44). The only effect is in Eq. (4.47): changing N changes the frequencies at which Eq. (4.44) is evaluated.

If we expand the right-hand side of Eq. (4.44), we obtain

$$\mathcal{F}(f_k) = R(f_k) + \mathbf{i}I(f_k) \tag{4.50}$$

where

$$R(f_k) = \sum_{j=0}^{N-1} [d_R(t_j)\cos(2\pi f_k t_j) - d_I \sin(2\pi f_k t_j)] \tag{4.51}$$

and

$$I(f_k) = \sum_{j=0}^{N-1} [d_R(t_j)\sin(2\pi f_k t_j) + d_I \cos(2\pi f_k t_j)] \tag{4.52}$$

are the real and imaginary parts of the discrete Fourier transform.

Three different ways of viewing the results of the discrete Fourier transform are common: the absorption spectrum, the power spectrum and the absolute-value spectrum. The absorption spectrum, the real part of an appropriately phased discrete Fourier transform, is commonly used in NMR. In NMR the sinusoids usually have the same phase; consequently if one multiplies Eq. (4.50) by $\exp\{\mathbf{i}(\theta + T_0 f_k)\}$, the phase of the sinusoids can be made to cancel from the discrete Fourier transform. The two parameters, θ and T_0, are the zero- and first-order phase corrections, and must be estimated from the discrete Fourier transform. An absorption spectrum's usefulness is limited to problems in which the sinusoids have the same phase parameters. This is common in NMR, but not with other physical phenomena, consequently, we will not discuss the absorption spectrum further.

The power spectrum is defined as

$$\text{Power}(f_k) = \frac{R(f_k)^2 + I(f_k)^2}{N} \tag{4.53}$$

[1] To zero pad a data set, one adds zeros to the end of a data set, sets N to the length of this new zero padded data set and then runs a fast discrete Fourier transform on the zero padded data.

Figure 4.1: Frequency Estimation Using The DFT

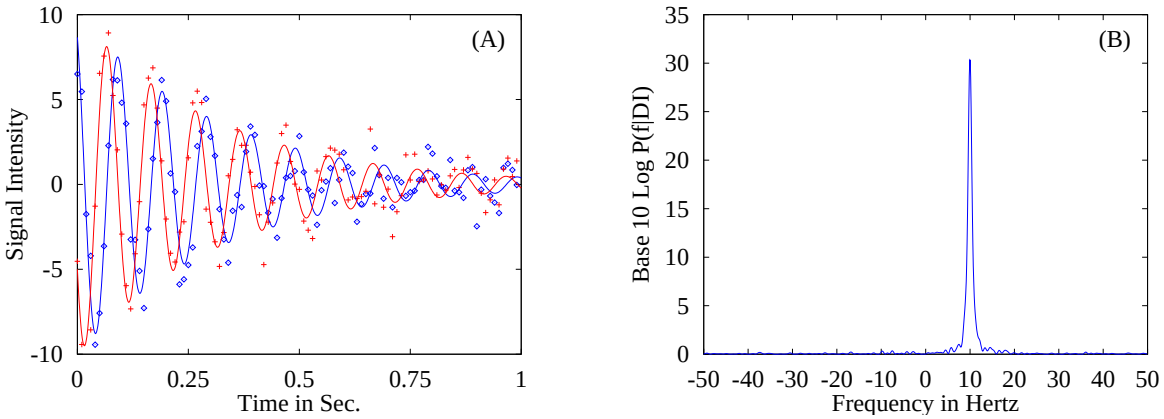

Figure 4.1: Panel (A) is computer simulated data. It contains a single exponentially decaying sinusoidal signal plus noise. The lines represent the real and imaginary parts of the sinusoid. The locations of the data values are denoted by the isolated characters. The Nyquist critical frequency for this data set is $f_{Nc} = 50$ Hz. Panel (B) is a plot of the base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid given these data.

and is the square of the absolute-value spectrum. It has been shown by Bretthorst [4, 5, 6] and Woodward [65], and as we will demonstrate shortly, the power spectrum is the sufficient statistic in a Bayesian calculation for the posterior probability for the frequency given a single stationary sinusoidal model with simultaneously sampled quadrature data.

In this chapter we will make several plots of the discrete Fourier transform and its generalizations to nonuniformly nonsimultaneously sampled data. To allow direct comparison of these plots we will always plot the same function of the data, the base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid independent of the phase, amplitude and variance of the noise, Eq. (4.101) below. For uniformly sampled quadrature data, this probability is a simple function of the power spectrum, see Bretthorst [2] for a more extended discussion of the relationship between the discrete Fourier transform and the posterior probability for a stationary frequency.

To illustrate the use of the discrete Fourier transform as a frequency estimation tool, suppose we had the data shown in Fig. 4.1(A). The signal in this simulated data is an exponentially decaying sinusoid of amplitude 10 plus Gaussian noise of zero mean and standard deviation one. These data were generated with a dwell time of $\Delta T = 0.01$ Sec. One hundred complex data values were generated at times ranging from 0 to 0.99 seconds. The frequency is 10 Hz, and the decay rate constant is 3 Sec.$^{-1}$. The real and imaginary data values are represented by the isolated characters in Fig. 4.1(A). The lines represent the real and imaginary parts of the true sinusoid. The Nyquist critical frequency for these data is one-half the inverse of the dwell time:

$$f_{Nc} = \frac{1}{2\Delta T} = \frac{1}{2(0.01 \text{ Sec.})} = 50 \text{ Hz.} \tag{4.54}$$

Figure 4.1(B) is a plot of the base 10 logarithm of the posterior probability over the bandwidth (-50 Hz $\leq f \leq$ 50 Hz). This base 10 logarithm starts at essentially zero and then increases some 30

Figure 4.2: Aliases

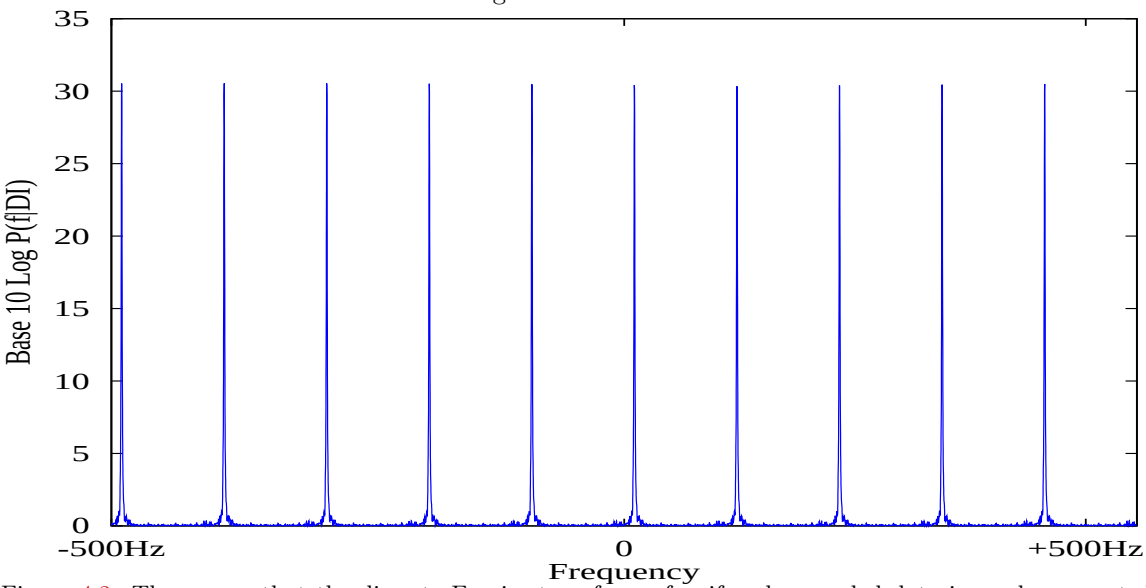

Figure 4.2: The reason that the discrete Fourier transform of uniformly sampled data is rarely computed at frequencies greater than the Nyquist critical frequency is simply that the discrete Fourier transform is a periodic function with a period equal to the bandwidth $2f_{Nc}$.

orders of magnitude, coming to a very sharp peak at 10 Hz.

4.3.1.2 Aliases

We would like to investigate the aliasing phenomenon. To do this we must evaluate the discrete Fourier transform outside the bandwidth and this cannot be done using the fast discrete Fourier transform. Because we are plotting the base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid, we can simply evaluate the posterior probability for any frequency and the sufficient statistic will be the power spectrum evaluated at that frequency; we are not restricted to the frequencies f_k specified by Eq. (4.47). The resulting plot is shown in Fig. 4.2. Outside of the interval $(-f_{Nc} \leq f \leq f_{Nc})$, the base 10 logarithm of the posterior probability, and thus the power spectrum and the discrete Fourier transform, are periodic functions of f with a period equal to the frequency interval spanned by the bandwidth. In Fig. 4.2, the frequency interval plotted is $(-10f_{Nc} \leq f \leq 10f_{Nc})$, so there should be 10 peaks in this range as Fig. 4.2 shows.

To understand why the discrete Fourier transform is a periodic function of frequency, suppose we wish to evaluate the discrete Fourier transform at the frequencies

$$f_k = \frac{k}{N\Delta T}, \qquad k = mN + k', \quad k' \in \left\{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}\right\}. \tag{4.55}$$

By itself the index k' would specify the normal frequency interval, Eq. (4.47), of a discrete Fourier transform. However, the integer m shifts this frequency interval up or down by an integer multiple

of the total bandwidth. If $m = 0$, we are in the interval $(-f_{Nc} \leq f_k \leq f_{Nc})$; if $m = 1$, we are in the interval $(f_{Nc} \leq f_k \leq 3f_{Nc})$, etc. If we now substitute Eqs. (4.55) and (4.46) into the discrete Fourier transform, Eq. (4.44), the reason the discrete Fourier transform is periodic becomes readily apparent

$$\mathcal{F}(f_{k'}) \equiv \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\left\{\frac{2\pi \mathbf{i}(mN + k')j}{N}\right\}, \tag{4.56}$$

$$= \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\{2\pi \mathbf{i} mj\} \exp\left\{\frac{2\pi \mathbf{i} k'j}{N}\right\}, \tag{4.57}$$

$$= \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\left\{\frac{2\pi \mathbf{i} k'j}{N}\right\}, \tag{4.58}$$

$$= \sum_{j=0}^{N-1} \mathbf{d}(t_j) \exp\{2\pi \mathbf{i} f_{k'} t_j\}. \tag{4.59}$$

In going from Eq. (4.57) to (4.58) a factor, $\exp\{\mathbf{i}(2\pi mj)\}$, was dropped because both m and j are integers, so $(2\pi mj)$ is an integer multiple of 2π, and the complex exponential is one. Aliases occur because the complex exponential canceled leaving behind a discrete Fourier transform on the interval $(-f_{Nc} \leq f_k \leq f_{Nc})$. The integer m specifies which integer multiple of the bandwidth is being evaluated and will always be an integer no matter how the data are collected. However, the integer j came about because the data were uniformly sampled. If the data had not been uniformly sampled the relationship, $t_j = j\Delta T$, would not hold, the complex exponential would not have canceled, and aliases would not have been present.

Because frequency estimation using the discrete Fourier transform was not derived using the rules of probability theory, there is no way to be certain that the estimates we obtain are the best we could do. What is needed is the solution to the single frequency estimation problem using Bayesian probability theory. Consequently, in the next section we analyze this problem, then in the following sections we will derive the conditions under which the discrete Fourier transform power spectrum is a sufficient statistic for single frequency estimation and we will see how probability theory generalizes the discrete Fourier transform to nonuniformly nonsimultaneously sampled data, and we will see the effect of these generalizations on the aliasing phenomenon.

4.3.2 State The Model—Single-Frequency Estimation

The problem is the estimation of the frequency of a single exponentially decaying sinusoid independent of the amplitude and phase of the sinusoid, given nonuniformly nonsimultaneously sampled quadrature data. First, what do we mean by nonuniformly nonsimultaneously sampled quadrature data? "Quadrature" simply means we have a measurement of the real and imaginary parts of a complex signal. So nonuniformly nonsimultaneously sampled quadrature data are measurements of the real and imaginary parts of a complex signal for which the measurements of the real and imaginary parts of the signal occur at different times. But if we have differing numbers of data samples with differing sample times, we really have two data sets: a real and an imaginary data set.[2] The

[2]Of course, when we say an "imaginary data set" we mean only that the data are a measurement of the imaginary part of the signal; not that the data are imaginary numbers.

real data set will be designated as $D_R \equiv \{d_R(t_1) \cdots d_R(t_{N_R})\}$, where d_R means a real data sample, t_i is the time the data sample was acquired, and N_R is the total number of data samples in the real data set. Similarly, the imaginary data set will be denoted by $D_I \equiv \{d_I(t_1) \cdots d_I(t_{N_I})\}$. We impose no restrictions on the number of data samples or the acquisition times in either channel. They could be the same or different, depending on the limit we investigate. Note that the times do not carry a channel indication so the context within the equations will have to establish which times we are referring to. If the equation context fails, we will clarify it in the text.

To perform any calculation using probability theory, the hypothesis of interest must be related to the information we actually possess. For the problem of estimating the frequency of a complex sinusoid, this means relating the frequency to the quadrature data through a model. If the complex data are given by Eq. (4.45), then the data and the sinusoid are related by

$$\mathbf{d}(t_j) = \mathbf{A} \exp\{-\mathbf{f} t_j\} + \mathbf{n}(t_j). \tag{4.60}$$

The complex amplitude is given by $\mathbf{A} = A_1 - \mathbf{i} A_2$, and is equivalent to the amplitude and phase of the sinusoid. The complex frequency, $\mathbf{f} = \alpha + 2\pi \mathbf{i} f$, contains two parameters: the decay rate constant, α, and the frequency, f. Note the minus signs in the definition of the complex amplitude and the one in Eq. (4.60). These signs correspond to a convention establishing what is meant by a positive frequency. The signs were chosen to model the data produced by a Varian NMR spectrometer. Other vendors use different conventions. Changing these conventions will change none of the conclusions that follow and very few of the actual details of the calculations. The decay rate constant, α, has units of inverse seconds, the frequency, f, Hertz, and the times, t_j, seconds. The quantity $\mathbf{n}(t_j)$ represents the complex noise at time t_j. Note that in this equation the times, t_j, simply designate the times at which we actually have data. If the datum happen to be a measurement of the real part of the signal, then the time would be associated with the real data set, and similarly for the imaginary part of the signal.

If we separate Eq. (4.60) into its real and imaginary parts, we have for the real part

$$d_R(t_i) = M_R(t_i) + n_R(t_i) \tag{4.61}$$

$$M_R(t_i) \equiv [A_1 \cos(2\pi f t_i) - A_2 \sin(2\pi f t_i)] \exp\{-\alpha t_i\} \tag{4.62}$$

and for the imaginary part we have

$$d_I(t_j) = M_I(t_j) + n_I(t_j) \tag{4.63}$$

$$M_I(t_j) \equiv -[A_1 \sin(2\pi f t_j) + A_2 \cos(2\pi f t_j)] \exp\{-\alpha t_j\}, \tag{4.64}$$

where $n_R(t_i)$ and $n_I(t_j)$ represent the noise in the real and imaginary data at times t_i and t_j. The quantity that we would like to estimate is the frequency, f, and we would like to estimate it independent of the amplitudes and variance of the noise. The decay rate constant, α, will sometimes be treated as a nuisance parameter, sometimes estimated, and sometimes taken as a given, depending of our purpose at the time. For now we will estimate it.

4.3.3 Apply Probability Theory

In probability theory as logic, all of the information about a hypothesis is summarized in a probability density function. For this problem, estimating the frequency of an exponentially decaying sinusoid,

the probability density function is designated as $P(f|D_R D_I I)$, which should be read as the posterior probability for the frequency given the real and imaginary data sets and the information I. The information I is all of the other information we have about the parameters appearing in the problem. Note that f appearing in this equation is not a parameter in any real sense; rather it is a hypotheses of the form "at the time the data were taken the frequency of the sinusoid was f." So by computing the posterior probability for various values of f we are ranking an entire series of hypotheses about the values of the frequency.

There are many different ways to compute the posterior probability for the frequency, $P(f|D_R D_I I)$, but they all lead to the same final result. Here we will simply apply Bayes' theorem:

$$P(f|D_R D_I I) = \frac{P(f|)P(D_R D_I|fI)}{P(D_R D_I|I)} \tag{4.65}$$

The three terms on the right-hand side of this equation are the prior probability for the frequency, $P(f|I)$ and it represents what was known about the frequency before seeing the data. The second term, $P(D_R D_I|fI)$, is the direct probability for the data given the frequency, and it represents what was learned about the frequency from the data. This term is often called the likelihood. Finally, the denominator, $P(D_R D_I|I)$, is a normalization constant and is computed using the product and sum rules of probability theory as

$$P(D_R D_I|I) = \int df P(D_R D_I f|I) = \int df P(f|I)P(D_R D_I|fI). \tag{4.66}$$

This term has many names, it is a direct prior probability, direct because it is a probability for the data; and prior because it depends only on our prior information. You will often hear the logarithm of this term called the evidence, and this probability is often called a Bayes factor. None of these terms really convey the importance of this term in model selection problems. However, our goal is parameter estimation, not model selection, so this direct probability is just a constant that we don't care about. If we agree to normalize the posterior probability for the frequency at the end of the calculation, then we can drop this normalization constant and one obtains:

$$P(f|D_R D_I I) \propto P(f|I)P(D_R D_I|fI). \tag{4.67}$$

The prior probability for the frequency is simplified enough that it could be assigned, however the direct probability for the data given the frequency, $P(D_R D_I|fI)$, is not nearly simplified enough to assign.

To see how to continue factoring this probability, note that $P(D_R D_I|fI)$ does not depend on the amplitude, phase or decay rate constant of the sinusoid, nor does it depend on the any properties of the errors. So this probability must be a marginal probability where the effects of the hypotheses indexed by these other parameters have been removed using the rules of probability theory. This marginal probability is computed from the joint probability for the data and these hypotheses given the frequency and prior information I, one can write

$$P(f|D_R D_I I) \propto P(f|I) \int dA_1 dA_2 d\alpha d\sigma P(A_1 A_2 \alpha \sigma D_R D_I|fI) \tag{4.68}$$

where σ is the standard deviation of the noise prior probability. Applying the product rule one obtains

$$P(f|D_R D_I I) \propto P(f|I) \int dA_1 dA_2 d\alpha d\sigma P(A_1 A_2 \alpha \sigma|I)P(D_R D_I|f A_1 A_2 \alpha \sigma I) \tag{4.69}$$

where the assumption was made that if the frequency is given to you, then it doesn't change the prior probabilities you would assign to the other parameters.

To compute the posterior probability for the frequency, we must perform a multidimensional integral over all of the parameters we are not interested in. As it will turn out, all of the integrals except the one over the decay rate may be done in close form. Consequently in much of what follows, the decay rate will be assumed known, in the sense of given and we will ignore it. The exception to this is the section on parameter estimation, and their we will marginalized over the decay rate numerically. However, before we can do this we must continue simplifying these probabilities, and before we can do the integrals we must assign numerical values to them.

First, we will simplify the joint posterior probability for the parameters. This probability, $P(\alpha A_1 A_2 \sigma|I)$, has four hypotheses as its arguments and one given. If we designate any one of these hypotheses as a, and all of the others as b, then this prior probability may be factored as $p(ab|I) = P(a|I)P(b|aI)$, where $p(a|I)$ is the prior probability for hypotheses a given I, and $P(b|aI)$ is the joint prior probability for all of the other hypotheses given knowledge of both a and I. Now suppose $P(b|aI) = P(b|I)$, that is to say, knowledge of a does nothing to change our prior probability assignment for b. This condition, known as logical independence, allows the prior probability to be factored into a set of independent prior for each parameter

$$P(\alpha A_1 A_2 \sigma|I) = P(\alpha|I)P(A_1|I)P(A_2|I)P(\sigma|I). \tag{4.70}$$

There are no additional simplifications that can be made to reduce these prior probabilities, and we will soon be forced to assign them numerical values. However, before doing that we must simplify the direct probability or likelihood.

We expect the two data sets to have the same noise standard deviation, because in NMR they are acquired by projecting the same signal onto orthogonal functions, a sine and a cosine. The fact that the two data sets are orthogonal projections, also means that each of the two data sets should contain unique and different information, i.e., they should be logically independent of each other. So the joint direct probability, $P(D_R D_I|f\alpha A_1 A_2 \sigma I)$, will factor as

$$P(D_R D_I|f\alpha A_1 A_2 \sigma I) = P(D_R|f\alpha A_1 A_2 \sigma I)P(D_I|f\alpha A_1 A_2 \sigma I) \tag{4.71}$$

a product of likelihoods, one for the real data and one for the imaginary data. Indeed, logical independence of these two data sets is almost forced upon us because we don't even know if the imaginary data exists, the number of imaginary data values could be zero.

If we now collect the prior probabilities from Eq. (4.70) and the likelihoods from Eq. (4.71) the posterior probability for the frequency becomes

$$\begin{aligned} P(f|D_R D_I I) &= \int dA_1\, dA_2\, d\alpha\, d\sigma \\ &\times \quad P(f|I)P(\alpha|I)P(A_1|I)P(A_2|I)P(\sigma|I) \\ &\times \quad P(D_R|f\alpha A_1 A_2 \sigma I)P(D_I|f\alpha A_1 A_2 \sigma I). \end{aligned} \tag{4.72}$$

None of the prior probabilities may be further simplified, however, the likelihoods could be further simplified, but if we are not taking correlations into account, need not be.

4.3.4 Assign The Probabilities

We have now reached the point where we need to assign numerical values to each of the probabilities appearing in Eq. (4.72). In all cases we are going to assign numerical values that represent what we actually know about the parameters. In most of these cases that may not be very much. For example, the prior probabilities for the amplitudes $P(A_1|I)$ or $P(A_2|I)$. We know they can be either positive or negative, and we know that there value is an upper, A_{max}, and lower bound, A_{min}. If that is all we know about these parameters then the principle of maximum entropy will lead us to assign a uniform prior probability:

$$P(A_x|I) = \begin{cases} \dfrac{1}{A_{max} - A_{min}} & A_{min} \leq A_x \leq A_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4.73}$$

where A_x represents either A_1 or A_2.

In a similar vain, we will assign a uniform prior probability for the frequency. For uniformly sampled data, the range of frequencies that are not aliases is simply related to the sampling rate. However, for nonuniformly sampled data, exactly what this range is, is an all together more interesting question, a question we will answer shortly, but not until we see how probability theory generalizes the discrete Fourier transform. Nonetheless we will assume that one can specify a range of values, f_{min} to f_{max} and that our prior expectation of the frequency is uniform over this region.

The standard deviation, σ, is a scale parameter. The completely uninformative prior probability for a scale parameter is the Jeffreys' prior [33],

$$P(\sigma|I) \propto \frac{1}{\sigma} \qquad 0 \leq \sigma < \infty. \tag{4.74}$$

However, this is not strictly speaking a probability at all, because it cannot be normalized. To make this a proper, i.e., normalized, probability one must introduce an upper, σ_{max}, and lower bound, σ_{min}, and compute the normalization constant:

$$P(\sigma|I) = \begin{cases} \dfrac{1}{\log(\sigma_{max}/\sigma_{min})\sigma} & \sigma_{min} \leq \sigma \leq \sigma_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4.75}$$

and then perform the integral over the valid range. If a Jeffreys' prior was desired, then at the end of the calculation one can pass to the limits $\sigma_{min} \to 0$, and $\sigma_{max} \to \infty$. It is only by following this safe, cautious procedures, that one can be sure of not inadvertently introducing singular mathematics into the calculations. However, in the process of doing the integrals over A_1, A_2 and σ we will assume that upper and lower bounds are very wide, much much wider than the maximum likelihood estimates of these parameters, and pass to the limit. We do this because it simplifies mathematics without introducing any complicating factors into this calculation.

The only remaining prior probability to be assigned is for the decay rate constant, $P(\alpha|I)$. Now the decay rate constant is a scale parameter and a bounded Jeffreys' prior would be appropriate for most conditions. However, one of the intentions of this chapter is to illustrate how probability theory generalizes the discrete Fourier transform to the case of nonuniformly nonsimultaneously sampled

data and for that demonstration a bounded Jeffreys' prior is inconvenient. Consequently, we will use one of two priors' for the decay rate constant, either a bounded Jeffreys' prior:

$$P(\alpha|I) = \begin{cases} \dfrac{1}{\log(\alpha_{max}/\alpha_{min})\alpha} & \alpha_{min} \leq \alpha \leq \alpha_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4.76}$$

when we are interested in estimating the decay rate constant, or we will assume that α is a given value $\hat{\alpha}$ and assign

$$P(\alpha|I) = \delta(\hat{\alpha} - \alpha) \tag{4.77}$$

when we are interested in illustrating the relationships between the discrete Fourier transform.

The only other prior probabilities to be assigned are the two direct probabilities $P(D_R|f\alpha A_1 A_2 \sigma I)$ and $P(D_I|f\alpha A_1 A_2 \sigma I)$ We will concentrate on assigning the direct probability for the real data, $P(D_R|f\alpha A_1 A_2 \sigma I)$, and after assigning it, how to assign $P(D_I|f\alpha A_1 A_2 \sigma I)$ will be obvious.

Probabilities are designated by the notation $P(X|I)$. The hypotheses appearing on the left-hand side of the vertical bar "|" are the hypotheses about which we are making inferences; while the hypotheses appearing on the right are given as true, *i.e.*, they specify the facts on which this probability is based. With this in mind, look at $P(D_R|f\alpha A_1 A_2 \sigma I)$. This term is the direct probability for the real data given the truth of all of the parameters in the model. But if the parameters are given, then from Eq. (4.61)

$$d_R(t_i) - M_R(t_i) = n_R(t_i). \tag{4.78}$$

The left-hand side of this equation contains the given parameter values, consequently the right-hand side contains the given error values. These given error values are thus hypotheses about which we must make inferences. These hypotheses are of the form "the true noise value was $n_R(t_i)$ at time t_i," where $n_R(t_i)$ is a running index and its numerical value would range over all valid noise amplitudes. Equation (4.78) is used in probability theory by introducing the joint probability for the data and the errors, and using the product and sum rules of probability theory to remove the dependence on the unknown error values:

$$\begin{aligned} P(D_R|f\alpha A_1 A_2 \sigma I) &= \int dn_R(t_1) \cdots dn_R(t_{N_R}) P(D_R\{n_R\}|f\alpha A_1 A_2 \sigma I) \\ &\propto \int dn_R(t_1) \cdots dn_R(t_{N_R}) P(\{n_R\}|\sigma) P(D_R|\{n_R\} f A_1 A_2 \alpha I) \end{aligned} \tag{4.79}$$

where $\{n_R\} \equiv \{n_R(t_1), n_R(t_2) \ldots, n_R(t_{t_{N_R}})\}$, $P(D_R|\{n_R\} f A_1 A_2 \alpha I)$, is the direct probability for the data given the errors and the parameters. The final term, $P(\{n_R\}|\sigma)$, is the prior probability for the errors given σ.

Most of the section on maximum entropy was devoted to deriving the Gaussian distribution as the prior probability appropriate to represent what is actually known about the errors. Using what was derived in Section 4.2 and assigning a Gaussian prior probability distribution for the errors one has

$$P(n_R(t_1) \cdots n_R(t_N)|\sigma) = (2\pi\sigma^2)^{-\frac{N_R}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N_R} n_R(t_i)^2\right\} \tag{4.80}$$

That leaves one final term that must be assigned, $P(D_R|\{n_R\} f A_1 A_2 \alpha I)$, the probability for the data given the errors and the parameters. But if we know the parameter values and we know the

error values, the by Eq. (4.78), $P(D_R|\{n_R\}fA_1A_2\alpha I)$, must be a delta function

$$P(D_R|\{n_R\}fA_1A_2\alpha I) = \delta\left[d_R(t_i) - M_R(t_i)\right]. \tag{4.81}$$

If we substitute this delta function, Eq. (4.81) and the prior probability for the errors, Eq.(4.80), into Eq. (4.79) one obtains

$$\begin{aligned} P(D_R|f\alpha A_1A_2\sigma I) &= \int dn_R(t_1)\cdots dn_R(t_{N_R}) \\ &\times (2\pi\sigma^2)^{-\frac{N_R}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N_R} n_R(t_i)^2\right\} \\ &\times \delta\left[d_R(t_i) - M_R(t_i)\right]. \end{aligned} \tag{4.82}$$

Finally, performing the integral over all of the n_R one obtains

$$P(D_R|f\alpha A_1A_2\sigma I) = (2\pi\sigma^2)^{-\frac{N_R}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N_R}\left[d_R(t_i) - M_R(t_i)\right]^2\right\} \tag{4.83}$$

as the direct probability for the data. In most expositions using probability one simply postulates a Gaussian probability density function and dispenses with all of the intermediate steps that tie everything together. Indeed in the original version of this paper, [13], that is exactly what the author did. However, here it was thought necessary to show how the rules of probability theory interact with maximum entropy to give both uninformative and informative prior probabilities.

Now having seen this derivation, we will simply skip all of the intermediate steps and make a Gaussian assignment for $P(D_I|f\alpha A_1A_2\sigma I)$. If we now substitute all of the probabilities into the posterior probability for the frequency one obtains

$$\begin{aligned} P(f|D_RD_II) &\propto \int dA_1\, dA_2\, d\alpha \frac{d\sigma}{\sigma} P(\alpha|I) \\ &\times \sigma^{-N_R}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=0}^{N_R-1}\left[d_R(t_i) - M_R(t_i)\right]^2\right\} \\ &\times \sigma^{-N_I}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=0}^{N_I-1}\left[d_I(t_j) - M_I(t_j)\right]^2\right\} \end{aligned} \tag{4.84}$$

where we have dropped some constants, as they will disappear when this probability is normalized and we have left the prior probability for the decay rate constant unspecified.

4.3.5 Evaluate The Sums and Integrals

Substituting Eqs. (4.62) and (4.64) for $M_R(t_i)$ and $M_I(t_j)$ into the joint posterior probability for the frequency and decay rate constant, Eq. (4.71), one obtains

$$P(f|D_RD_II) \propto \int dA_1\, dA_2\, d\alpha \frac{d\sigma}{\sigma} P(\alpha|I)\,\sigma^{-(N_R+N_I)}\exp\left\{-\frac{Q}{2\sigma^2}\right\} \tag{4.85}$$

where

$$Q \equiv (N_R + N_I)\overline{d^2} - 2\sum_{l=1}^{2} A_l T_l + \sum_{k,l=1}^{2} g_{kl} A_k A_l. \tag{4.86}$$

The mean-squared data value is defined as

$$\overline{d^2} = \frac{1}{N_R + N_I}\left[\sum_{i=0}^{N_R-1} d_R(t_i)^2 + \sum_{j=0}^{N_I-1} d_I(t_j)^2\right]. \tag{4.87}$$

The projection of the data onto the T_l model vector is given by

$$\begin{aligned} T_1 \quad \equiv \quad & \sum_{i=0}^{N_R-1} d_R(t_i)\cos(2\pi f t_i)\exp\{-\alpha t_i\} \\ - \quad & \sum_{j=0}^{N_I-1} d_I(t_j)\sin(2\pi f t_j)\exp\{-\alpha t_j\} \end{aligned} \tag{4.88}$$

for $l = 1$ and

$$\begin{aligned} T_2 \quad \equiv \quad & -\sum_{i=0}^{N_R-1} d_R(t_i)\sin(2\pi f t_i)\exp\{-\alpha t_i\} \\ & -\sum_{j=0}^{N_I-1} d_I(t_j)\cos(2\pi f t_j)\exp\{-\alpha t_j\} \end{aligned} \tag{4.89}$$

for $l = 2$. The matrix g_{kl} is defined as

$$g_{kl} \equiv \begin{pmatrix} a & c \\ c & b \end{pmatrix}, \tag{4.90}$$

with

$$a = \sum_{i=0}^{N_R-1} \cos^2(2\pi f t_i)\exp\{-2\alpha t_i\} + \sum_{j=0}^{N_I-1} \sin^2(2\pi f t_j)\exp\{-2\alpha t_j\} \tag{4.91}$$

where the sum over the cosine uses the t_i associated with the real data set, while the sum over the sine uses the t_j associated with the imaginary data set. Similarly, b is defined as

$$b = \sum_{i=0}^{N_R-1} \sin^2(2\pi f t_i)\exp\{-2\alpha t_i\} + \sum_{j=0}^{N_I-1} \cos^2(2\pi f t_j)\exp\{-2\alpha t_j\}, \tag{4.92}$$

where the sum over the sine uses the t_i associated with the real data set, while the sum over the cosine uses the t_j associated with the imaginary data set. Finally, c is defined as

$$\begin{aligned} c \quad = \quad & -\sum_{i=0}^{N_R-1} \cos(2\pi f t_i)\sin(2\pi f t_i)\exp\{-2\alpha t_i\} \\ & +\sum_{j=0}^{N_I-1} \sin(2\pi f t_j)\cos(2\pi f t_j)\exp\{-2\alpha t_j\}, \end{aligned} \tag{4.93}$$

where the sum over the cosine-sine product uses the t_i associated with the real data set, while the sum over the sine-cosine product uses the t_j associated with the imaginary data set.

The integrals over A_1 and A_2 are both Gaussian integrals and will be evaluated first. The way we will do this is to make a simple observation about Gaussian quadrature integrals and then use that observation to evaluate the integrals. The observation is simply that Gaussians are symmetric in the amplitudes. Because they are symmetric, all integrating with respect to the amplitudes does is to constrain the amplitudes to their maximum posterior probability estimates and it introduces a volume factor, the determinate. Consequently, the maximum of the posterior probability as a function of the amplitudes is given by the solution to

$$\sum_{l=1}^{2} g_{kl}\hat{A}_l = T_k. \tag{4.94}$$

For this simple 2×2 matrix, the solution is given by

$$\hat{A}_1 = \frac{bT_1 - cT_2}{ab - c^2} \tag{4.95}$$

and

$$\hat{A}_2 = \frac{aT_2 - cT_2}{ab - c^2}. \tag{4.96}$$

The sufficient statistic is thus given by

$$\overline{h^2} = \sum_{j=1}^{2} T_j \hat{A}_j. \tag{4.97}$$

This sufficient statistic can be rewritten as

$$\overline{h^2} \equiv \frac{bT_1^2 + aT_2^2 - 2cT_1T_2}{ab - c^2}. \tag{4.98}$$

While not obvious, it is this statistic that generalizes the discrete Fourier transform to nonuniformly nonsimultaneously sampled data. This statistic will reduce to the Lomb-Scargle periodogram, a weighted normalized power spectrum and the Schuster periodogram under appropriate conditions. finally the posterior probability for the frequency is given by

$$P(f|\sigma D_R D_I I) \propto \int d\alpha\, d\sigma\, \frac{P(\alpha|I)}{\sqrt{ab - c^2}} \sigma^{-(N_R+N_I-2)-1} \exp\left\{\frac{(N_R + N_I)\overline{d^2} - \overline{h^2}}{2\sigma^2}\right\} \tag{4.99}$$

where we have assumed that the upper and lower amplitude bounds on the amplitude integrals may be extended to plus and minus infinity without introducing an appreciable error into the integral.

The integral over the standard deviation of the noise prior probability, σ, can be transformed into a Gamma function:

$$\int_0^\infty \sigma^{-N-1} \exp\left\{-\frac{Q}{\sigma^2}\right\} = \frac{1}{2}\Gamma\left(\frac{N}{2}\right) Q^{-\frac{N}{2}} \tag{4.100}$$

from which one obtains

$$P(f|D_R D_I I) \propto \int d\alpha \, \frac{P(\alpha|I)}{\sqrt{ab-c^2}} \left[(N_R+N_I)\overline{d^2} - \overline{h^2}\right]^{-\frac{N_R+N_I-2}{2}} \tag{4.101}$$

which is of the form of Student's t-distribution, and we have dropped some constants that cancel when this probability density function is normalized. If the decay rate constant is a given, the prior, $P(\alpha|I)$, should be taken as a delta function and the integral in Eq. (4.101) evaluated analytically. Otherwise, the integral must be evaluated numerically.

4.3.6 How Probability Generalizes The Discrete Fourier Transform

We are now in a position to demonstrate how probability theory generalizes the discrete Fourier transform and what the effect of this generalization is on the aliasing phenomenon. We mentioned earlier that the sufficient statistic for frequency estimation is related to a power spectrum, and we demonstrate that next. Several simplifications must be made to reduce Eq. (4.98) to a power spectrum. First, we must be estimating the frequency of a stationary sinusoid. A stationary sinusoid has no decay, so $\alpha = 0$. Second, the data must be uniformly and simultaneously sampled. With these assumptions the matrix g_{kl}, Eq. (4.90), simplifies because $c = 0$ and $a = b = N_R = N_I = N$, where N is the total complex data values. The sufficient statistic, Eq. (4.98), reduces to

$$\overline{h^2} = \frac{R(f)^2 + I(f)^2}{N} \tag{4.102}$$

and is the power spectrum defined in Eq. (4.53). The functions $R(f)$ and $I(f)$ were defined earlier, Eqs. (4.51) and (4.52), and are the real and imaginary parts of the discrete Fourier transform.

The Schuster periodogram, or power spectrum, is the sufficient statistic for frequency estimation in uniformly simultaneously sampled quadrature data given a stationary sinusoidal model, but we already have two generalizations to the discrete Fourier transform that were not contained in the definition, Eq. (4.44). First, the frequency appearing in Eq. (4.102) is a continuous parameter; it is not in any way restricted to discrete values. Probability theory indicates that there is information in the data at frequencies between the f_k in the discrete Fourier transform. Second, the frequency, f, is bounded only by our prior information. Probability theory does not indicate that the frequency must be less than the Nyquist critical frequency. Consequently, probability theory is telling one that aliases are real indications of the presence of a frequency and absolutely nothing in the data can tell one which is the correct frequency, *only* prior information can do that.

The Schuster periodogram is an optimal frequency estimator for uniformly simultaneously sampled quadrature data. However, this is not true for real data where the statistic was originally proposed. To see this suppose we have a real data set, so that either $N_I = 0$ or $N_R = 0$. The g_{kl} matrix does simplify—many of the sums appearing in Eqs. (4.91)-(4.93) are zero—but the matrix remains nondiagional and so Eq. (4.98) is the sufficient statistic for this problem, and is numerically equal to the Lomb-Scargle periodogram. The Schuster periodogram, however, is never a sufficient statistic for either real uniformly or nonuniformly sampled data. It can only be derived as an approximation to the sufficient statistic for this problem. To derive it two approximations must be made in the g_{kl} matrix. First, the off-diagonal element must be much smaller than the diagonal and so can be approximated as zero. The second approximation assumes the diagonal elements may be approximated by $a = b = N/2$. Both approximations ignore terms on the order of $\sqrt{N}$ and are good

approximations when one has large amounts of data. In this discussion we have implicitly assumed that the times appearing in the cosine and sine transforms of the data making up the Schuster periodogram were evaluated at the times one actually has data. Trying to interpolate the data onto a uniform grid and then using a power spectrum is not justified under any circumstances.

Suppose now the signal is exponentially decaying with time, but otherwise the data remain uniformly and simultaneously sampled. Examining Eq. (4.90), we see that $c = 0$ and $a = b = N_{\text{eff}}$, where N_{eff} is an effective number of complex data values and is given by

$$N_{\text{eff}} = \sum_{i=0}^{N-1} \exp\{-2\alpha t_i\}. \tag{4.103}$$

The sufficient statistic becomes

$$\overline{h^2} = \frac{R(f,\alpha)^2 + I(f,\alpha)^2}{N_{\text{eff}}} \tag{4.104}$$

which is a weighted power spectrum. The effective number of complex data values is equal to the number of complex data values, N, when the decay rate constant, α, is zero, and is approximately, $1/2\alpha$, for densely sampled signals which decay into the noise. The reason for this behavior should be obvious: as the decay rate constant increases, for a fixed dwell time, fewer and fewer data values contribute to the estimation process. When the decay rate constant is large, the effective number of data values goes to zero and the data are uninformative about either the frequency or the decay rate constant.

The functions $R(f,\alpha)$ and $I(f,\alpha)$ are the real and imaginary parts of the discrete Fourier transform of the complex data weighted by an exponential of decay rate constant α. In a weighted discrete Fourier transform, one multiplies the complex data by a weighting function

$$\text{Complex Weighted Data} = \mathbf{d}(t_i)\exp\{-\alpha t_i\} \tag{4.105}$$

and then performs the discrete Fourier transform on the weighted data.

In spectroscopic applications many different weighting functions are used. Indeed Varian NMR software comes with exponential, Gaussian, sine-bell and a number of others. In all of these cases, use of the weighting function in conjunction with a discrete Fourier transform amounts to estimating the frequency of a single sinusoid having a decay envelope described by the weighting function. If the weighting function does not mimic the decay envelope of the signal, then these procedures are, from the standpoint of parameter estimation, less than optimal. Of course most of these weighting functions were developed with very different ideas in mind than parameter estimation. For example, the sine-bell is intended to increase resolution of multiple close lines, just as a Gaussian is used to transform the line shape of the resonances from Lorentzian to Gaussian in the hope that the Gaussian will be narrower in the frequency domain. Nonetheless, probability theory indicates that all of these procedures are estimating the frequency of a single sinusoid having a decay envelope described by the weighting function. The better this weighting function describes the true decay of the signal, the better the parameter estimates will be.

For exponential weighting, the spectroscopist must choose a value of α. This is typically done so that the decay envelope of the weighting function matches the decay of the signal. This is equivalent to trying to locate the maximum of the joint posterior probability for the frequency and decay rate constant, Eq.(4.85). If one makes a contour plot of this joint posterior probability, this plot will have nearly elliptical contours with the major axis of the ellipse nearly parallel to the decay rate

constant axis (decay rate constants are less precisely estimated than frequencies), while the minor axis will be nearly parallel to the frequency axis. Thus, estimates of the frequency and decay rate constant are nearly independent of each other. Consequently, if the spectroscopist can guess the decay rate constant, even approximately, the frequency estimate he obtains will be almost as good as that obtained by doing a thorough search for the location of the joint maximum.

It is commonly believed that matched weighting functions (matching the shape of the weighting function to the observed decay of the data) increases the signal-to-noise ratio in the resulting power spectrum at the expense of broadening the peak, thereby decreasing the frequency resolution of the discrete Fourier transform. This is correct if the discrete Fourier transform is used as a spectral estimation procedure and one then tries to estimate multiple frequencies from this spectrum. However, from the standpoint of probability theory, it is only the single largest peak in the weighted discrete Fourier transform power spectrum that is relevant to frequency estimation, and then it is only a very small region around the location of the maximum that is of interest. All of the details in the wings around that peak are irrelevant to the estimation problem. If there are multiple peaks in the power spectrum, probability theory will systematically ignore them: *the weighted power spectrum is the sufficient statistic for single frequency estimation*; it does not estimate multiple frequencies, although one can show that under many conditions the sufficient statistic for the multiple frequency estimation problem is related to the multiple peaks in a power spectrum [2]. Given a multiple-frequency model, probability theory will lead one to other statistics that take into account the nonorthogonal nature of the sinusoidal model functions. These multiple-frequency models always result in parameter estimates that are either better than or essentially identical to those obtained from a power spectrum. They will be essentially identical to the power spectrum results when multiple, very well separated sinusoids are present, and they will always be better when overlapping resonances are present—see Bretthorst [4, 5, 6, 2, 7, 8, 9] for details.

Suppose we have nonuniformly but simultaneously sampled data. What will happen to the resulting Bayesian calculations? When we repeat the Bayesian calculation we find that absolutely nothing has changed. The number of effective data values, N_{eff}, is given by Eq. (4.103), the matrix g_{kl} is given by Eq. (4.90), just as the sufficient statistic is given by Eq. (4.104), and the posterior probability for the frequency is given by Eq. (4.101). Here the generalization to the discrete Fourier transform accounts for the fact that the times must be evaluated explicitly, the formula, $t_j = j\Delta T$, does not hold and one must substitute the actual time, t_i and t_j, into the equations. Missing observations, make no difference whatever to probability theory. Probability theory analyzes the data you actually obtain, regardless of whether the data are uniformly sampled or not. Indeed, in Bayesian probability theory there is no such thing as a missing data problem.

Having allowed the data to be nonuniformly sampled, we are in a position to see what is happening to the aliasing phenomenon. However, before addressing this, there is one last generalization that we would like to make. This generalization allows the data to be nonsimultaneously sampled. Because the samples are nonsimultaneous, the sums appearing in the discrete Fourier transform, Eqs. (4.51) and (4.52), are no longer correct. The terms appearing in these equations are the sine and cosine transforms of the real and imaginary data sets. When the data were simultaneously sampled, the sine and cosine transforms could be combined into a single sum. For nonsimultaneous time samples, this cannot be done. Each sine and cosine transform must have an independent summation index. If you examine the projections of the model onto the nonuniformly nonsimultaneously sampled data, Eqs. (4.88) and (4.89), you will find this is exactly what probability theory has done. The function T_1 corresponds to the real part of the discrete Fourier transform and is the cosine transform of the real data minus the sine transform of the imaginary data; however, now it also accounts for the

nonuniform nonsimultaneous times. Similarly, up to a minus sign,[3] T_2 corresponds to the imaginary part of the discrete Fourier transform. So the discrete Fourier transform has been generalized in the sense that the sine and cosine transforms now have separate summation indices. However, there is more to this generalization than using separate summation indices.

In the previous examples the matrix g_{kl}, Eq. (4.90), was diagonal with diagonal elements equal to the effective number of data values in each channel. For simultaneously sampled data these diagonal elements are equal. For nonsimultaneously sampled data, the diagonal elements remain the effective number of data values in each channel, but these are no longer equal. For simultaneous data samples, the zero off-diagonal element means that the integrals over the amplitudes are completely independent of each other. In the model, the function multiplying each amplitude may be thought of as an N dimensional vector. With nonsimultaneous samples these vectors are linear combinations of each other. The magnitude of the off-diagonal element is a measure of how far from orthogonal these vectors are. Consequently, the sufficient statistic, Eq. (4.98), is now taking into account the fact that these vectors are not orthogonal, the real and imaginary data can have different numbers of data values and that the effective number of data values is a function of both frequency and decay rate constant.

4.3.7 Aliasing

Now that we have finished discussing the generalizations of the discrete Fourier transform to nonuniformly nonsimultaneously sampled data, we would like to investigate some of the properties of these calculations to show what has happened to the aliasing phenomenon. Earlier, when we investigated the discrete Fourier transform of uniformly sampled data, we showed that, for frequencies outside the bandwidth, the power spectrum was a periodic function of frequency. What will happen to these aliases when we use nonuniformly nonsimultaneously sampled data? Are the aliases still there? If not, where did they go? One thing that should be obvious is that in nonuniformly nonsimultaneously sampled data the Nyquist critical frequency does not apply, at least not as previously defined, because the times are nonuniformly sampled. Consequently, nonuniformly nonsimultaneously sampled data will not suffer from aliases in the same way that uniformly simultaneously sampled data does.

To demonstrate this, we will again use simulated data. The simulated data will have exactly the same signal as the data shown in Fig. 4.1(A). The only difference between these two data sets will be the times at which the data were acquired. The nonuniformly nonsimultaneously sampled simulated data are shown in Fig. 4.3(A). In generating the times at which we have data we had to choose a sampling scheme. On some spectrometers it is possible to sample data exponentially, so we choose exponential sampling. In an exponentially sampled data set, a histogram of the time samples would follow an exponential distribution. Thus, there are more data samples at short times, and exponentially fewer at longer times. However, the discussion here pertains to all nonuniformly nonsimultaneously sampled data, not just to data with times that are exponentially sampled. The base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid is shown in Fig. 4.3(B). This plot spans a frequency interval that is ten thousand times larger than the corresponding plot shown in Fig. 4.1(B). In Fig. 4.2(B), when we extended the region of interest to $\pm 10 f_{Nc} = \pm 500$ Hz we had 10 aliases. Yet here we have gone to ± 50 kHz and there are no aliases—where did the aliases go? Why do these data seem to have a bandwidth that is at least 10,000 times larger than in the first example?

[3] The minus sign comes about because of the use of Varian sign conventions in Eq. (4.60).

Figure 4.3: Nonuniformly Nonsimultaneously Sampled Sinusoid

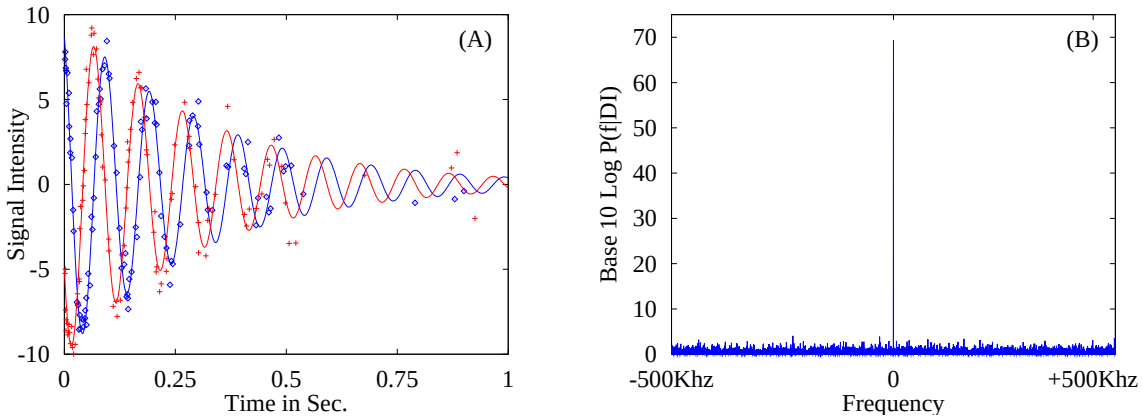

Figure 4.3: A Nonuniformly Nonsimultaneously Sampled Exponentially Decaying Sinusoid Panel (A) is computer simulated data. It contains the same exponentially decaying sinusoidal signal as shown in Fig. 4.1(A) plus noise of the same standard deviation. The lines represent the real and imaginary parts of the sinusoid. The location of the nonuniformly nonsimultaneously sampled data values are denoted by the isolated characters. Panel (B) is the base 10 logarithm of the posterior probability for the frequency given a stationary sinusoid model using these data. Note that this plot spans 10,000 times the Nyquist critical frequency for the uniformly sampled version of this data set shown in Fig. 4.1(A).

We showed earlier in Eqs. (4.56)-(4.59) that aliases come about because of the integer j used in specifying the uniformly sampled times, $t_j = j\Delta T$, in the discrete Fourier transform. In the present problem, nonuniformly nonsimultaneously sampled data, there is no ΔT such that all of the acquisition times are integer multiples of this time; not if the times are truly sampled randomly. However, all data and times must be recorded to finite accuracy. This is true even of the simulated data shown in Fig. 4.3(A). Consequently, there must be a largest effective dwell time, $\Delta T'$, such that all of the times (both the real and imaginary) must satisfy

$$t_l = k_l \Delta T' \qquad t_l \in \{\text{Real } t_i \text{ or Imaginary } t_j\} \tag{4.106}$$

where k_l is an integer. The subscript l was added to k to indicate that each of the times t_l requires a different integer k_l to make this relationship true. Of course, this was also true for uniformly sampled data: its just that for uniformly sampled data the integers were consecutive, $k_l = 0, 1, \cdots N-1$. The effective dwell time is always less than or equal to the smallest time interval between data items, and is the least common denominator for all of the times. Additionally, $\Delta T'$ is the dwell time one would have had to acquire data at in order to obtain a uniformly sampled data set with data items at each of the times t_i and t_j. The effective dwell time, $\Delta T'$, can be used to define a Nyquist critical frequency

$$f_{Nc} = \frac{1}{2\Delta T'}. \tag{4.107}$$

Aliases *must* appear for frequencies outside the bandwidth defined from $\Delta T'$.

The reason that aliases must appear for frequencies outside this bandwidth can be made apparent in the following way. Suppose we have a hypothetical data set that is sampled at $\Delta T'$. Suppose

Figure 4.4: Alias Spacing

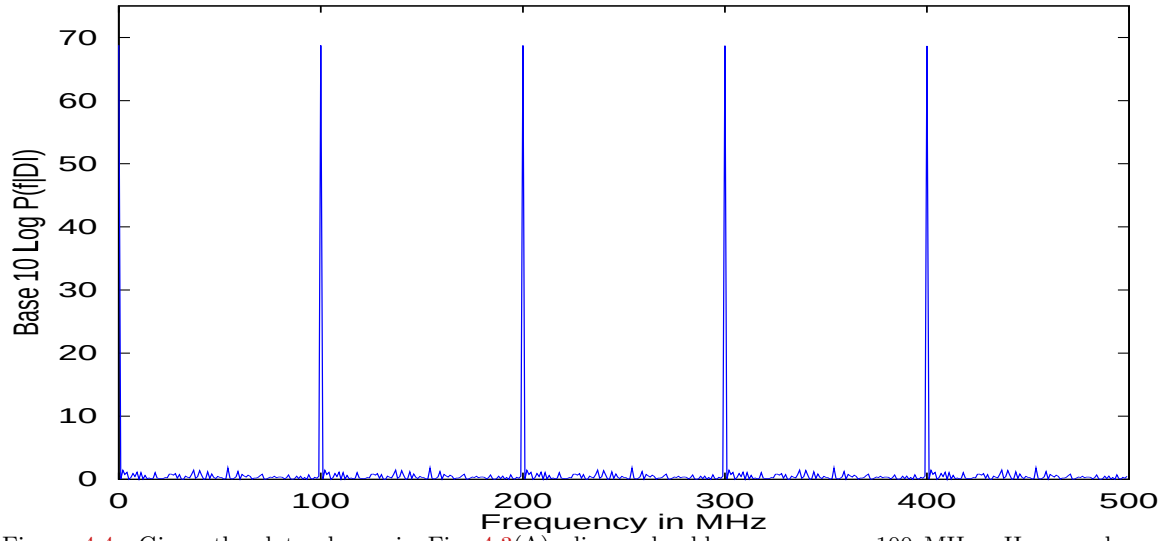

Figure 4.4: Given the data shown in Fig. 4.3(A) aliases should appear every 100 MHz. Here we have evaluated the logarithm of the posterior probability for the frequency of a stationary sinusoid at frequencies given by $10 + n \times 10^6$ Hz. Aliases should appear at $n = 100$, 200, 300, etc.

further, the hypothetical data are zero everywhere except at the times we actually have data, and there the data are equal to the appropriate $d_R(t_i)$ or $d_I(t_j)$. If we now compute the discrete Fourier transform of this hypothetical data set, then by the analysis done in Eqs. (4.56)-(4.59) the Nyquist critical frequency of this data set is $1/2\Delta T'$ and frequencies outside this bandwidth are aliased. Now look at the definitions of T_1 and T_2, Eqs. (4.88) and (4.89), for the data set we actually have. You will find that these quantities are just the real and imaginary parts of the discrete Fourier transform of our hypothetical data set. The zeros in the hypothetical data set cannot contribute to the sums in the discrete Fourier transform: they act only as place holders, and so the only part of the sums that survive are just where we have data. By construction that is just what Eqs. (4.88) and (4.89) are computing. So aliases *must* appear at frequencies greater than this Nyquist critical frequency.

For the data shown in Fig. 4.3, the times and the data were recorded to 8 decimal places in an ASCII file. This file was then used by another program to compute the posterior probability for the frequency, Eq. (4.98). Because the data were recorded to 8 decimal places, a good guess for the effective dwell time would be $\Delta T' = 10^{-8}$ Sec. This would correspond to a Nyquist critical frequency of $f_{Nc} = 5 \times 10^7$ Hz. The first alias of the 10 Hz frequency should appear at 100,000,010 Hz. If we evaluate the base 10 logarithm of the posterior probability at frequencies given $(10 + n \times 10^6)$ Hertz, we should see peaks at $n = 100$, 200, 300 etc. This plot is shown in Fig. 4.4. Note that the aliases are right at the expected frequencies. An extensive search from zero up to 100 MHz uncovered no aliases prior to the one just above 100 MHz. This suggests that the effective bandwidth of the 100 complex data values shown in Fig. 4.3(A) is 100 MHz! That is one million times larger than the bandwidth of the data shown in Fig. 4.1. Indeed, the effective dwell time, $\Delta T'$, was defined as the maximum time for which all of the t_i and t_j are integer multiples of $\Delta T'$. The Nyquist critical frequency computed from $\Delta T'$ is the *smallest* frequency for which the argument of the complex exponential in Eq. (4.57)

is an integer multiple of 2π. Consequently, $1/2\Delta T'$ is the Nyquist critical frequency for these data and there are no aliases within this implied bandwidth.

The fact that the data was recorded to 8 decimal places and that the bandwidth of this simulated data set is 10^8 Hz brings up another curious thing about aliases. Aliasing is primarily a phenomenon that concerns the times in the discretely sampled data, the data itself are almost irrelevant to aliasing. In the example we are discussing the times were recorded to 8 decimal places. If we had recorded the times to only 7 decimal places the aliases would have been in different places. If the last significant digit in the times is truncated, the bandwidth will be at least a factor of 10 lower. Of course we must qualify this somewhat, in the case we are discussing, a change in the 8'th decimal place changes the sines and cosines so little that the only noticeable effect is on the location of the aliases. However, if we were to continue truncating decimal places we will eventually reach the point where the times are so far from the correct values that we are essentially computing nonsense.

So far we have shown that the data have an effective bandwidth of $1/\Delta T'$ but that is not quite the same thing as showing that frequencies anywhere in this 100 MHz interval can be correctly estimated. There is another time, the minimum time between data values, ΔT_M, which might be of some importance. It might be thought that the data can show no evidence for frequencies outside a band of width $1/\Delta T_M$. In the data set shown in Fig. 4.3(A), $\Delta T_M = 0.00000488$ Sec. This would correspond to a bandwidth of roughly 200 kHz, a tremendous frequency range, but smaller than the effective bandwidth of 100 MHz by a factor of roughly 500. Which is correct?

The arguments given earlier prove that it is the effective dwell time, $\Delta T'$, and not the minimum interval between data values, ΔT_M, that is the important quantity. However, to illustrate that $\Delta T'$ is the critical time, it is a simple matter to generate data that contain a frequency in the range $([1/\Delta T_M] < f < [1/\Delta T'])$ and then compute the base 10 logarithm of posterior probability for the frequency of a stationary sinusoid over the entire range $(0 \leq f \leq 1/\Delta T')$. From a practical standpoint this is nearly impossible for data with a Nyquist critical frequency of 50 MHz; this frequency will have to be lowered. This can be done by truncating the exponentially sampled times to 5 decimal places, and then generating the simulated signal using these truncated times. Truncating the times lowers the Nyquist critical frequency to 50 kHz. Additionally, when these simulated data shown in Fig. 4.5(A) were generated, no times were generated closer together than 0.0001 Sec. This time corresponds to a Nyquist critical frequency of 5 kHz; so there is a factor of 10 difference between the Nyquist critical frequency and the frequency calculated from ΔT_M. The simulated acquisition parameters used in generating these data are similar to those used previously, *i.e.,* $N_R = N_I = 100$, $\alpha = 3$ Sec.$^{-1}$, Amplitude $= 10$, $\sigma = 1$. The only difference is that the simulated frequency is 50 kHz, a full factor of 5 higher than the minimum sampling time would give as the highest resolvable frequency, and equal to the Nyquist critical frequency for these data. However, the region plotted $(0 \leq f \leq 100$ kHz), has been shifted upward by 50 kHz, so the 50 kHz frequency is in the middle of the plotted region, and this region should be free of aliases. The base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid is shown in Fig. 4.5(B). It was evaluated at every 1 Hz from 0 to 100,000 Hz. This frequency resolution is more than enough to ensure that if aliases exist, multiple peaks would be present in this plot. Note that there is a single peak and it is located at 50 kHz, the frequency of the simulated resonance. So the critical time is indeed $\Delta T'$ and the bandwidth of these data is 100 kHz.

Having shown that the critical time is the effective dwell time $\Delta T'$ and that there are no aliases in the full bandwidth implied by $\Delta T'$, one might be tempted to think that that is the end of the story. However, that is not quite correct. While there are no aliases in the bandwidth implied by $\Delta T'$, it is possible for there to be multiple peaks which are artifacts related to the effective dwell

Figure 4.5: Which Is The Critical Time

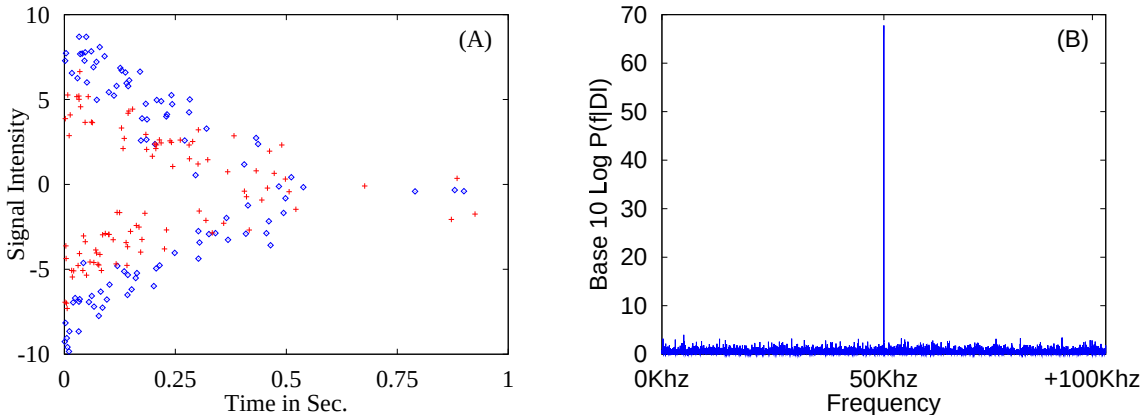

Figure 4.5: Which Is The Critical Time: $\Delta T'$ Or ΔT_M? Panel (A) is computer simulated data. Except for the frequency, it contains the same signal as that shown in Fig. 4.1(A). The frequency in these data is 50 kHz. The positions of the nonuniformly nonsimultaneously sampled data values are denoted by the isolated characters. Panel (B) is the base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid given these data.

time. These artifacts are not aliases in the sense that they are not exact replicas of the main peak; rather they are evidence for the resonance, and their height is directly related to how strongly the data indicate the presence of a resonance at that frequency. To see how multiple peaks might occur, suppose we have the data shown in Fig. 4.1(A); with 4 additional nonuniformly but simultaneously sampled complex data values at 0.001, 0.015, 0.025, and 0.037 seconds respectively.[4] These 4 complex data values were generated by sampling the same signal plus noise as shown in Fig 4.1(A), but at the four nonuniformly sampled times. Now, according to the analysis done in this chapter, the Nyquist critical frequency for the combined data is $1/(2 \times 0.001 \text{ Sec.}) = 500$ Hz; this bandwidth is exactly the same as the frequency interval shown in Fig. 4.2 where we had 10 aliases. In principle, these 4 complex data values should increased the bandwidth by a factor of 10 and should destroy the aliasing phenomenon in the ± 500 Hz frequency band. A plot of the base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid given the combined data is shown in Fig. 4.6(A). Note that there are 10 peaks in Fig. 4.6(A), just as there are 10 peaks in Fig. 4.2, but the peaks are not of the same height, so they are not true aliases. To determine which peak corresponds to the true frequency one must assign a bounded prior probability for the frequency (to eliminate the true aliases at frequencies above 500 Hz and below -500 Hz) and then normalize the posterior probability. The fully normalized posterior probability for the frequency of a stationary sinusoid is shown in Fig. 4.6(B). The fully normalized posterior probability density function has a single peak at 10 Hz, the true frequency. In this example, the 4 extra nonuniformly sampled complex data values were enough to rises the probability for the true frequency several orders of magnitude, and when the posterior probability for the frequency was normalized, the vast majority of the weight in the

[4]The fact that the 4 complex data values are simultaneously sampled is irrelevant. The results of this analysis would be the same regardless if these 8 total data items were simultaneously sampled or not.

Figure 4.6: Example, Frequency Estimation

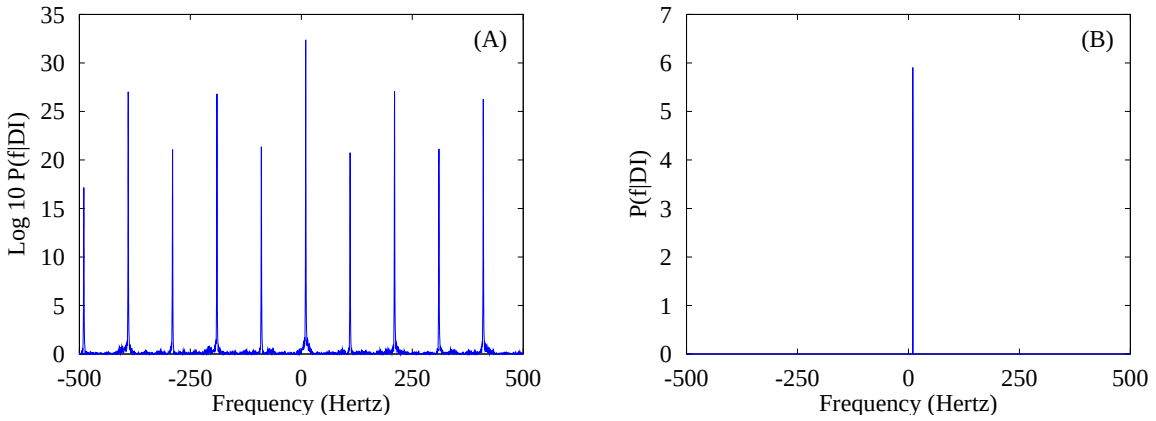

Figure 4.6: Panel (A) is the base 10 logarithm of the posterior probability for the frequency of a stationary sinusoid using the data shown in Fig. 4.1(A) with four additional data values with sample times given by 0.001, 0.015, 0.025, and 0.037 seconds respectively. These four data values were generated using the same signal and signal-to-noise ratio as those shown in Fig. 4.1(A). The only difference is the nonuniformly sampled times. Note that we now have multiple peaks, but they are not true aliases because they are not of the same height. Panel (B) is the fully normalized posterior probability for the frequency of a stationary sinusoid given these data, note that all of the peaks except the one at the true frequency, 10 Hz, have been exponentially suppressed.

posterior distribution was concentrated around the true frequency, consequently all of the spurious peaks seen in Fig. 4.6(A) were exponentially suppressed.

In preparing this example the number of nonuniformly sampled data values had to be chosen. Initially 6 complex nonuniformly sampled data values were generated. But this was abandon because, when the base 10 logarithm of the posterior probability for a stationary frequency was plotted, the spurious peaks were only about 5 percent of the main peak and so did not illustrate the points we wanted to make. However, it does indicate that it does not take many nonuniformly sampled data values to eliminate these spurious peaks. As few as 10 percent should completely eliminate them. Nonetheless, it is possible for the fully normalized posterior probability for the frequency to have multiple peaks in data that contain only a single frequency. If this happens, it indicates that the data simply cannot distinguish which of the possibilities are the true frequency. The only recourse will be to obtain more measurements, preferably at nonuniformly nonsimultaneously sampled times.

The preceding example reiterate what has been said several times: the discrete Fourier transform power spectrum, the Schuster periodogram, a weighted power spectrum, the Lomb-Scargle periodogram, and the generalizations to these presented in this chapter are sufficient statistics for frequency estimation given a *single* frequency model. Multiple peaks in the discrete Fourier transform and its generalizations are not necessarily evidence for multiple frequencies. The only way to be certain that multiple frequencies are present is to postulate models containing one, two, etc. frequencies and to then compute the posterior probability for these models. Depending on the outcome of that calculation, one can then estimate the frequencies from the appropriate model.

4.3.8 Parameter Estimates

So far the discussions have concentrated on the discrete Fourier transform, how probability theory generalizes it to nonuniformly nonsimultaneously sampled data, and how these generalizations affect aliases. Now we are going to discuss the effect of nonuniform nonsimultaneous sampling on the parameter estimates. In this discussion we are going to estimate the parameters using the data shown in Fig. 4.1(A) and in Fig. 4.3(A). These two data sets contain exactly the same signal and each data set contains Gaussian white noise drawn from a Gaussian random number generator of unit standard deviation. However, the noise realizations in each data set are different, and this will result in slightly different parameter estimates for each data set. Nonetheless these two data sets provide an excellent opportunity to demonstrate how nonuniform nonsimultaneous sampling affects the parameter estimates.

We will discuss estimation of the frequency, decay rate constant and the amplitude. We will not discuss estimation of the phase and standard deviation of the noise as these are of less importance. The posterior probability for the frequency, decay rate constant, and amplitude are shown in panels (A), (B) and (D) of Fig. 4.7 respectively. Each of these plots is the fully normalized marginal posterior probability for the parameter of interest independent of all of the other parameters appearing in the model. Panel (C) contains the absolute-value spectra computed from these two data sets and will be used to compare Fourier transform estimation procedures to the Bayesian calculations. The solid lines in these plots were computed from the nonuniformly nonsimultaneously sampled data shown in Fig. 4.3(A); while the curves drawn with open characters were computed using the uniformly sampled data shown in Fig. 4.1(A).

A Markov chain Monte Carlo simulation was used to compute the marginal posterior probability for each parameter. All of the parameters appearing in the model were simulated simultaneously, thus the Markov chain Monte Carlo simulation simulated the joint posterior probability for all

Figure 4.7: Estimating The Sinusoids Parameters

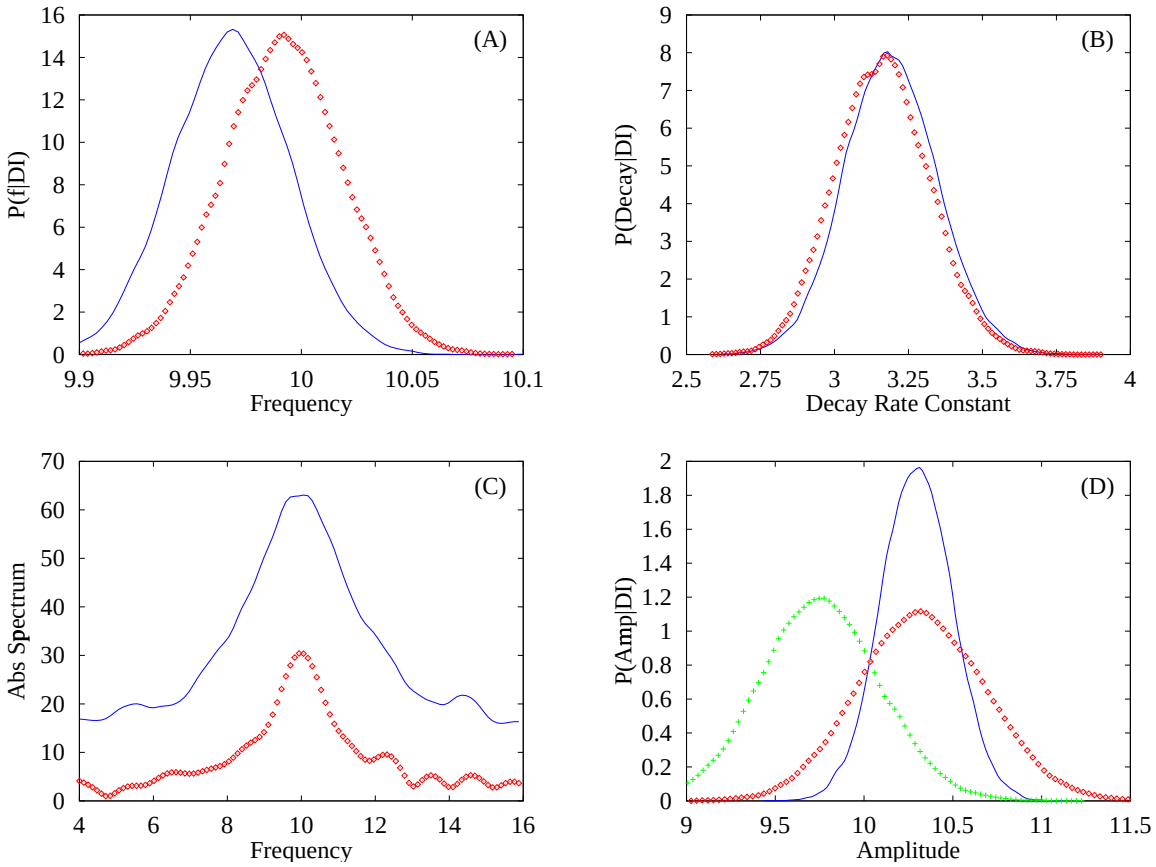

Figure 4.7: Estimating The Parameters The posterior probability of the parameter of interest was computed given the nonuniformly nonsimultaneously sampled data, solid lines, and was computed given the uniformly sampled data, open characters. Panel (A) is the posterior probability for the frequency, (B) the posterior probability of the decay rate constant, (D) the posterior probability for the amplitude. Panel (C) is the absolute-value spectrum computed for the uniformly sampled data and for the nonuniformly nonsimultaneously sampled data. The extra curve in panel (D), the plus signs, is the posterior probability for the amplitude computed from a nonuniformly nonsimultaneously sample data set having exactly the same signal with exactly the same signal-to-noise level but with times that were generated from a uniform random number generator.

the parameters. This was done for computational convenience; *i.e.,* it was easier to do a single Markov chain Monte Carlo simulation than to do five separate calculations, one for each parameter appearing in the model. Because the probability density functions shown in panels (A), (B) and (D) were formed by computing a histogram of the Markov chain Monte Carlo samples there are small irrelevant artifacts in these plots that are related to the number of samples drawn from the simulation. For more on Markov chain Monte Carlo methods and how these can be used to implement Bayesian calculations see Gilks [24] and Radford [46].

The marginal posterior probability for the frequency is shown in panel (A). This is the fully normalized marginal posterior probability for the frequency independent of all of the other parameters, Eq. (4.101). Note that the true frequency, 10 Hz, is well covered by the posterior probability computed from both the uniformly (open characters) and nonuniformly nonsimultaneously (sold line) sampled data. Also note that these distributions are almost identical in height and width. Consequently, both the uniform and nonuniformly nonsimultaneously sampled data have given the same parameter estimates to within the uncertainty in these estimates. Of course the details for each estimated differ, because the noise realizations in each data set differ. Consequently, the frequency estimate is not strongly dependent on the sampling scheme. Indeed this can be derived from the rules of probability theory with the following proviso: the two sampling schemes must cover the *same* total sampling time and must sample the signal in a reasonably dense fashion so that sums may be approximated by integrals, [2, 8]. Having said this, we must reemphasize that this is only true for frequency estimates using data having sampling schemes covering the *same* total sampling time; it is not true if the sampling times differ nor is it necessarily true of the other parameters appearing in the model. Indeed one can show that for a given number of data values, the precision of the frequency estimate for a stationary sinusoid is inversely proportional to the total sampling time, provided one samples the signal in a reasonably dense fashion. Thus, sampling 10 times longer will result in frequency estimates that are 10 times more precise. As noted in Bretthorst [2] this is equivalent to saying that for frequency estimation data values at the front and back of the data are most important in determining the frequency, because it is in these data that small phase differences are most highly magnified by the time variable. This could be very useful in some applications; but of course it will not help in NMR applications where the signal decays away.

We have also plotted the absolute-value spectra computed from these two data sets, Fig. 4.7(C). Note that the peaks of these two absolute-value spectra are at essentially the same frequency as the corresponding peaks in panel (A); although they are plotted on differing scales. If the absolute value spectrum is used to estimate the frequency, one would typically use the frequency of the peak as the estimate and then claim roughly the half-width-at-half-height as the uncertainty in this estimate. For these two data sets that is about 10 plus or minus 2 Hz. The two fully normalized posterior probabilities shown in panel (A) span a frequency interval of only 0.2 Hz. This frequency interval is roughly 6 standard deviations. Thus the frequency has been estimated to roughly 10 Hz with an uncertainty of $0.2/6 \approx 0.03$ Hz; a 60 fold reduction in the uncertainty in the frequency estimate.

One last note before we begin the discussion of estimating the decay rate constant, we reiterate that all of the details in the wings of the absolute-value spectrum shown in panel (C) are irrelevant to the frequency estimation process. The posterior probability for the frequency has peaked in a region that is very small compared to the scale of these wings, all of the information about the frequency estimate is contained in a very small region around the largest peak in the absolute-value spectrum.

The marginal posterior probability for the decay rate constant is shown in Fig. 4.7(B). Here we again find that the parameter estimates from both data sets are essentially identical in all of there

relevant details. Both probabilities peak at nearly the same value of the decay rate constant, both have nearly the same width, and therefore the same standard deviation; thus like frequency estimates, the estimates for the decay rate constants do not strongly dependent on the sampling scheme. In principle the accuracy of the estimates for the decay rate constants scale with time just like the frequency estimates, of course, with decaying signals this is of little practical importance. Note that the decay rate constant has been estimated to be about 3.2 ± 0.3 Sec.$^{-1}$ at one standard deviation. The true value is 3 Sec.$^{-1}$, so both sampling schemes give reasonable estimates of the decay rate. If one were to try and estimate the decay rate constant from the absolute-values spectrum, the half-width-at-half-height would normally be used, here that is about 2 Sec.$^{-1}$ and no claim about the accuracy of the estimate would be made.

The marginal posterior probability for the amplitude of the sinusoid is shown in Fig. 4.7(D). The amplitude, A, was defined in Eq. (4.60). In this chapter we did not directly talk about amplitude estimation (see Bretthorst [8] for a discussion of this subject), rather we treated the amplitudes of the sine and cosine model functions as nuisance parameters and removed them from the posterior probability for the other parameters. We did this because we wished to explore the relationships between frequency estimation using Bayesian probability theory and the discrete Fourier transform. However, the Markov chain Monte Carlo simulation used $A\cos(2\pi ft+\theta)\exp\{-\alpha t\}$ as the model for the real data, so it was a trivial matter to compute the posterior probability for the amplitude. If you examine Fig. 4.7(D) you will note that now we do have a real difference between the uniform (open characters) and the nonuniformly nonsimultaneously sampled data (solid lines). The amplitude estimates from the nonuniformly nonsimultaneously sampled data are a good factor of 2 more precise than the estimates from the uniformly sampled data. One might think that this is caused by the nonuniform nonsimultaneous sampling and this would be correct, but not for the obvious reasons. If you examine panel (D) you will note that we have plotted a third curve (plus signs). This curve is the posterior probability for the amplitude computed from data with the exact same signal and signal-to-noise ratio, but having times that are nonuniformly nonsimultaneously sampled where the times were generated from a uniform random number generator. We will call this data set the uniform-randomly sampled data. Note that the height and width of the posterior probabilities computed from both the uniformly and the uniform-randomly sampled data are essentially the same, so by itself the nonuniform nonsimultaneous sampling did not cause the amplitude estimates to improve. The amplitude estimate improved because exponential sampling gathered more data where the signal was large. The accuracy of the amplitude estimate is proportional to the standard deviation of the noise and inversely proportional to square root of the effective number of data values. Because exponential sampling gathered more data where the signal was large, its effective number of data values was larger and so the amplitude estimate improved. In this case, the improvement was about a factor of 2, so the exponential sampling had an effective number of data values that was about a factor of 4 larger than for the uniformly or uniform-randomly sampled data. This fact is also reflected in differing heights of the absolute value spectra plotted in Fig. 4.7(C). The peak height of an absolute value spectrum is proportional to the square root of the effective number of data values. In panel (C) the spectra computed from the uniformly sampled data set, open characters, is roughly a factor of 2 lower than the height of the spectrum computed from the exponentially sampled data set, solid line.

4.4 Summary and Conclusions

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 [31, 28, 30, 20]. 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 the following Chapters we will have need for both types of calculations. Most of the problems we will deal with are parameter estimation problems. However, whenever a model is selected in which an "unknown" options is invoked, a model selection calculation is be done. In addition, some of the calculations, such as the big peak/little peak calculation have a model selection calculation built into a parameter estimation problem.

Bibliography

[1] Rev. Thomas Bayes (1763), "An Essay Toward Solving a Problem in the Doctrine of Chances," *Philos. Trans. R. Soc. London*, **53,** pp. 370-418; reprinted in *Biometrika*, **45**, pp. 293-315 (1958), and *Facsimiles of Two Papers by Bayes,* with commentary by W. Edwards Deming, New York, Hafner, 1963.

[2] G. Larry Bretthorst (1988), "Bayesian Spectrum Analysis and Parameter Estimation," in *Lecture Notes in Statistics,* **48**, J. Berger, S. Fienberg, J. Gani, K. Krickenberg, and B. Singer (eds), Springer-Verlag, New York, New York.

[3] G. Larry Bretthorst (1990), "An Introduction to Parameter Estimation Using Bayesian Probability Theory," in *Maximum Entropy and Bayesian Methods,* Dartmouth College 1989, P. Fougère *ed.,* pp. 53-79, Kluwer Academic Publishers, Dordrecht the Netherlands.

[4] G. Larry Bretthorst (1990), "Bayesian Analysis I. Parameter Estimation Using Quadrature NMR Models" *J. Magn. Reson.*, **88**, pp. 533-551.

[5] G. Larry Bretthorst (1990), "Bayesian Analysis II. Signal Detection And Model Selection" *J. Magn. Reson.*, **88**, pp. 552-570.

[6] G. Larry Bretthorst (1990), "Bayesian Analysis III. Examples Relevant to NMR" *J. Magn. Reson.*, **88**, pp. 571-595.

[7] G. Larry Bretthorst (1991), "Bayesian Analysis. IV. Noise and Computing Time Considerations," *J. Magn. Reson.*, **93**, pp. 369-394.

[8] G. Larry Bretthorst (1992), "Bayesian Analysis. V. Amplitude Estimation for Multiple Well-Separated Sinusoids," *J. Magn. Reson.*, **98**, pp. 501-523.

[9] G. Larry Bretthorst (1992), "Estimating The Ratio Of Two Amplitudes In Nuclear Magnetic Resonance Data," in *Maximum Entropy and Bayesian Methods,* C. R. Smith *et al.* (eds.), pp. 67-77, Kluwer Academic Publishers, the Netherlands.

[10] G. Larry Bretthorst (1993), "On The Difference In Means," in Physics & Probability Essays in honor of Edwin T. Jaynes, W. T. Grandy and P. W. Milonni (eds.), pp. 177-194, Cambridge University Press, England.

[11] G. Larry Bretthorst (1996), "An Introduction To Model Selection Using Bayesian Probability Theory," in *Maximum Entropy and Bayesian Methods*, G. R. Heidbreder, ed., pp. 1-42, Kluwer Academic Publishers, Printed in the Netherlands.

[12] G. Larry Bretthorst (1999), "The Near-Irrelevance of Sampling Frequency Distributions," in *Maximum Entropy and Bayesian Methods,* W. von der Linden *et al.* (eds.), pp. 21-46, Kluwer Academic Publishers, the Netherlands.

[13] G. Larry Bretthorst (2001), "Nonuniform Sampling: Bandwidth and Aliasing," in *Maximum Entropy and Bayesian Methods in Science and Engineering,* Joshua Rychert, Gary Erickson and C. Ray Smith *eds.*, pp. 1-28, American Institute of Physics, USA.

[14] G. Larry Bretthorst, Christopher D. Kroenke, and Jeffrey J. Neil (2004), "Characterizing Water Diffusion In Fixed Baboon Brain," in Bayesian Inference And Maximum Entropy Methods In Science And Engineering, Rainer Fischer, Roland Preuss and Udo von Toussaint *eds.*, AIP conference Proceedings, **735**, pp. 3-15.

[15] G. Larry Bretthorst, William C. Hutton, Joel R. Garbow, and Joseph J.H. Ackerman (2005), "Exponential parameter estimation (in NMR) using Bayesian probability theory," *Concepts in Magnetic Resonance*, 27A, Issue 2, pp. 55-63.

[16] G. Larry Bretthorst, William C. Hutton, Joel R. Garbow, and Joseph J. H. Ackerman (2005), "Exponential model selection (in NMR) using Bayesian probability theory," *Concepts in Magnetic Resonance*, 27A, Issue 2, pp. 64-72.

[17] G. Larry Bretthorst, William C. Hutton, Joel R. Garbow, and Joseph J.H. Ackerman (2005), "How accurately can parameters from exponential models be estimated? A Bayesian view," *Concepts in Magnetic Resonance*, 27A, Issue 2, pp. 73-83.

[18] G. Larry Bretthorst, W. C. Hutton, J. R. Garbow, and Joseph J. H. Ackerman (2008), "High Dynamic Range MRS Time-Domain Signal Analysis," *Magn. Reson. in Med.*, **62**, pp. 1026-1035.

[19] V. Chandramouli, K. Ekberg, W. C. Schumann, S. C. Kalhan, J. Wahren, and B. R. Landau (1997), "Quantifying gluconeogenesis during fasting," *American Journal of Physiology*, **273**, pp. H1209-H1215.

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

[21] André d'Avignon, G. Larry Bretthorst, Marlyn Emerson Holtzer, and Alfred Holtzer (1998), "Site-Specific Thermodynamics and Kinetics of a Coiled-Coil Transition by Spin Inversion Transfer NMR," *Biophysical Journal*, **74**, pp. 3190-3197.

[22] André d'Avignon, G. Larry Bretthorst, Marlyn Emerson Holtzer, and Alfred Holtzer (1999), "Thermodynamics and Kinetics of a Folded-Folded Transition at Valine-9 of a GCN4-Like Leucine Zipper," *Biophysical Journal*, **76**, pp. 2752-2759.

[23] David Freedman, and Persi Diaconis (1981), "On the histogram as a density estimator: L_2 theory," *Zeitschrift f¨r Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**, 4, pp. 453-476.

[24] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (1996), "Markov Chain Monte Carlo in Practice," Chapman & Hall, London.

[25] Paul M. Goggans, and Ying Chi (2004), "Using Thermodynamic Integration to Calculate the Posterior Probability in Bayesian Model Selection Problems," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop*, **707**, pp. 59-66.

[26] Marlyn Emerson Holtzer, G. Larry Bretthorst, D. André d'Avignon, Ruth Hogue Angelette, Lisa Mints, and Alfred Holtzer (2001), "Temperature Dependence of the Folding and Unfolding Kinetics of the GCN4 Leucine Lipper via 13C alpha-NMR," *Biophysical Journal*, **80**, pp. 939-951.

[27] E. T. Jaynes (1968), "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics*, SSC-4, pp. 227-241; reprinted in [30].

[28] E. T. Jaynes (1978), "Where Do We Stand On Maximum Entropy?" in *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus *Eds.*, pp. 15-118, Cambridge: MIT Press, Reprinted in [30].

[29] E. T. Jaynes (1980), "Marginalization and Prior Probabilities," in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner *ed.,* North-Holland Publishing Company, Amsterdam; reprinted in [30].

[30] E. T. Jaynes (1983), "Papers on Probability, Statistics and Statistical Physics," a reprint collection, D. Reidel, Dordrecht the Netherlands; second edition Kluwer Academic Publishers, Dordrecht the Netherlands, 1989.

[31] E. T. Jaynes (1957), "How Does the Brain do Plausible Reasoning?" unpublished Stanford University Microwave Laboratory Report No. 421; reprinted in *Maximum-Entropy and Bayesian Methods in Science and Engineering* **1,** pp. 1-24, G. J. Erickson and C. R. Smith *Eds.,* 1988.

[32] E. T. Jaynes (2003), "Probability Theory—The Logic of Science," edited by G. Larry Bretthorst, Cambridge University Press, Cambridge UK.

[33] Sir Harold Jeffreys (1939), "Theory of Probability," Oxford Univ. Press, London; Later editions, 1948, 1961.

[34] John G. Jones, Michael A. Solomon, Suzanne M. Cole, A. Dean Sherry, and Craig R. Malloy (2001) "An integrated ^{2}H and ^{13}C NMR study of gluconeogenesis and TCA cycle flux in humans," *American Journal of Physiology, Endocrinology, and Metabolism,* **281,** pp. H848-H856.

[35] John Kotyk, N. G. Hoffman, W. C. Hutton, G. Larry Bretthorst, and J. J. H. Ackerman (1992), "Comparison of Fourier and Bayesian Analysis of NMR Signals. I. Well-Separated Resonances (The Single-Frequency Case)," *J. Magn. Reson.*, **98**, pp. 483–500.

[36] Pierre Simon Laplace (1814), "A Philosophical Essay on Probabilities," John Wiley & Sons, London, Chapman & Hall, Limited 1902. Translated from the 6th edition by F. W. Truscott and F. L. Emory.

[37] N. Lartillot, and H. Philippe (2006), "Computing Bayes Factors Using Thermodynamic Integration," *Systematic Biology*, **55** (2), pp. 195-207.

[38] D. Le Bihan, and E. Breton (1985), "Imagerie de diffusion in-vivo par rsonance," Comptes rendus de l'Acadmie des Sciences (Paris), **301** (15), pp. 1109-1112.

[39] N. R. Lomb (1976), "Least-Squares Frequency Analysis of Unevenly Spaced Data," *Astrophysical and Space Science*, **39**, pp. 447-462.

[40] T. J. Loredo (1990), "From Laplace To SN 1987A: Bayesian Inference In Astrophysics," in *Maximum Entropy and Bayesian Methods,* P. F. Fougere (ed), Kluwer Academic Publishers, Dordrecht, The Netherlands.

[41] Craig R. Malloy, A. Dean Sherry, and Mark Jeffrey (1988), "Evaluation of Carbon Flux and Substrate Selection through Alternate Pathways Involving the Citric Acid Cycle of the Heart by 13C NMR Spectroscopy," *Journal of Biological Chemistry*, **263** (15), pp. 6964-6971.

[42] Craig R. Malloy, Dean Sherry, and Mark Jeffrey (1990), "Analysis of tricarboxylic acid cycle of the heart using ^{13}C isotope isomers," *American Journal of Physiology*, **259**, pp. H987-H995.

[43] Lawrence R. Mead and Nikos Papanicolaou, "Maximum entropy in the problem of moments," *J. Math. Phys.* **25**, 2404–2417 (1984).

[44] K. Merboldt, Wolfgang Hanicke, and Jens Frahm (1969), "Self-diffusion NMR imaging using stimulated echoes," Journal of Magnetic Resonance, **64** (3), pp. 479-486.

[45] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953), "Equation of State Calculations by Fast Computing Machines," Journal of Chemical Physics. The previous link is to the Americain Institute of Physics and if you do not have access to Science Sitations you many not be able to retrieve this paper.

[46] Radford M. Neal (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," technical report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

[47] Jeffrey J. Neil, and G. Larry Bretthorst (1993), "On the Use of Bayesian Probability Theory for Analysis of Exponential Decay Data: An Example Taken from Intravoxel Incoherent Motion Experiments," *Magn. Reson. in Med.*, **29**, pp. 642–647.

[48] H. Nyquist (1924), "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, **3**, pp. 324-346.

[49] H. Nyquist (1928), "Certain Topics in Telegraph Transmission Theory," *Transactions AIEE*, **3**, pp. 617-644.

[50] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery (1992), "Numerical Recipes The Art of Scientific Computing Second Edition," Cambridge University Press, Cambridge UK.

[51] Emanuel Parzen (1962), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics* **33**, 1065–1076

[52] Karl Pearson (1895), "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Phil. Trans. R. Soc. A* **186**, 343–326.

[53] Murray Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics* **27**, 832–837 (1956).

[54] Jeffery D. Scargle (1981), "Studies in Astronomical Time Series Analysis I. Random Process In The Time Domain," *Astrophysical Journal Supplement Series*, **45**, pp. 1-71.

[55] Jeffery D. Scargle (1982), "Studies in Astronomical Time Series Analysis II. Statistical Aspects of Spectral Analysis of Unevenly Sampled Data," *Astrophysical Journal*, **263**, pp. 835-853.

[56] Jeffery D. Scargle (1989), "Studies in Astronomical Time Series Analysis. III. Fourier Transforms, Autocorrelation Functions, and Cross-correlation Functions of Unevenly Spaced Data," *Astrophysical Journal*, **343**, pp. 874-887.

[57] Arthur Schuster (1905), "The Periodogram and its Optical Analogy," *Proceedings of the Royal Society of London*, **77**, p. 136-140.

[58] Claude E. Shannon (1948), "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, **27**, pp. 379-423.

[59] John E. Shore, and Rodney W. Johnson (1981), "Properties of cross-entropy minimization," *IEEE Trans. on Information Theory*, **IT-27,** No. 4, pp. 472-482.

[60] John E. Shore and Rodney W. Johnson (1980), "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. on Information Theory*, **IT-26** (1), pp. 26-37.

[61] Devinderjit Sivia, and John Skilling (2006), "Data Analysis: A Bayesian Tutorial," Oxford University Press, USA.

[62] Edward O. Stejskal and Tanner, J. E. (1965), "Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient." Journal of Chemical Physics, **42** (1), pp. 288-292.

[63] D. G. Taylor and Bushell, M. C. (1985), "The spatial mapping of translational diffusion coefficients by the NMR imaging technique," Physics in Medicine and Biology, **30** (4), pp. 345-349.

[64] Myron Tribus (1969), "Rational Descriptions, Decisions and Designs," Pergamon Press, Oxford.

[65] P. M. Woodward (1953), "Probability and Information Theory, with Applications to Radar," McGraw-Hill, N. Y. Second edition (1987); R. E. Krieger Pub. Co., Malabar, Florida.

[66] Arnold Zellner (1971), "An Introduction to Bayesian Inference in Econometrics," John Wiley and Sons, New York.