in Maximum Entropy and Bayesian Methods, *Dartmouth, 1989*, pp. 53-79,
P. F. Fougère *(ed.) Kluwer Academic Publishers, The Netherlands.* (1990)

AN INTRODUCTION TO PARAMETER ESTIMATION USING BAYESIAN PROBABILITY THEORY

G. LARRY BRETTHORST
Washington University,
Department of Chemistry
1 Brookings Drive,
St. Louis, Missouri 63130

ABSTRACT. Bayesian probability theory does not define a probability as a frequency of occurrence; rather it defines it as a reasonable degree of belief. Because it does not define a probability as a frequency of occurrence, it is possible to assign probabilities to propositions such as "The probability that the frequency had value ω when the data were taken," or "The probability that hypothesis x is a better description of the data than hypothesis y." Problems of the first type are parameter estimation problems, they implicitly assume the correct model. Problems of the second type are more general, they are model selections problems and do not assume the model. Both types of problems are straight forward applications of the rules of Bayesian probability theory. This paper is a tutorial on parameter estimation. The basic rules for manipulating and assigning probabilities are given and an example, the estimation of a single stationary sinusoidal frequency, is worked in detail. This example is sufficiently complex as to illustrate all of the points of principle that must be faced in more realistic problems, yet sufficiently simple that anyone with a background in calculus can follow it. Additionally, the model selection problem is discussed and it is shown that parameter estimation calculation is essentially the first step in the more general model selection calculation.

1 Introduction

The problem of estimating the value of a parameter is basic to science and engineering, yet the procedures needed to solve this problem are rarely taught to scientists and engineers. This is mostly historical and has to do with the fact that probability theory, as traditionally interpreted, treats all probabilities as frequencies (here the word frequency is used in the sense of the number of times an event occurs). This interpretation of probability theory, called sampling theory, has no way of addressing the parameter estimation problem when only a single data set is available. To address the problem within sampling theory, one must imagine one is estimating the distribution of a random parameter within an ensemble of data sets. One then tries to determine the mean and standard deviation of this parameter within the ensemble. This is not the problem faced by scientist and engineers; there is typically only a single data set, and one is trying to determine the value the parameter had at the time the data was taken.

To solve this problem within probability theory, a wider interpretation is needed. This interpretation was first given by Laplace [1] and then essentially ignored for the next century. In Laplace's interpretation, probability theory is just common sense reduced to numbers, and a probability represents a reasonable degree of belief; not a frequency of occurrence. In the mid 1930, Jeffreys [2] rediscovered the works of Laplace and derived probability theory as an axiomatic theory of inference. Then in 1957, E. T. Jaynes [3] using the methodology of Shannon [4], the mathematics of Abel [5] and Cox [6], and the qualitative principles of Laplace [1], proved that if one represents a reasonable degree of belief as a real number, then the only consistent rules for manipulating probabilities are those given by Laplace. In this wider interpretation of probability theory, called Bayesian probability theory, problems of the form "What is the best estimate of a parameter one can make from the data and one's prior information?" make perfect sense. In this paper the basic principles of Bayesian probability theory are outlined and an example, the estimation of a single stationary sinusoidal frequency, is worked with each step in the calculation explained in detail. For an introduction to Bayesian probability theory see the works of Tribus [7] and Zellner [8] and for a derivation of the rules of probability theory see Jaynes [3].

2 The Rules of Probability Theory

There are only 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 arbitrary propositions then the product rule is

$$P(A, B|C) = P(A|C)P(B|A, C) \tag{1}$$

where $P(A, B|C)$ is the joint probability that "A and B is true given that C is true," $P(A|C)$ is the probability that "A is true given C is true," and $P(B|A, C)$ is the probability that "B is true given that both A and C are true." The notation "$|C$)" means conditional on the truth of proposition C. In Bayesian probability theory *all* probabilities are conditional. To use a notation such as $P(A)$ to stand for the probability of A does not make sense until the evidence on which it is based is given. Anyone using such notation either does not understand this or is being extremely careless in their notation. In either case one should be careful when interpreting such material.

In Aristotelian logic the proposition "A and B" is the same as "B and A" so the truth value of the propositions must be the same in the product rule. That is the probability of "A and B given C" must be equal to the probability of "B and A given C," so the order may be rearranged in Eq. (1) to obtain

$$P(B, A|C) = P(B|C)P(A|B, C). \tag{2}$$

This equation may be combined with Eq. (1) to obtain a seemingly trivial result

$$P(A|B, C) = \frac{P(A|C)P(B|A, C)}{P(B|C)}. \tag{3}$$

This is Bayes' theorem. It is named after Rev. Thomas Bayes, an 18th century mathematician who derived a special case of the theorem. Bayes' calculations [9] were published in

1763 after his death. Exactly what Bayes intended to do with the calculation, if anything, remains a mystery today. Nevertheless, today in Bayesian inference, this theorem is the starting point for all Bayesian calculations.

The sum rule of probability theory may be stated as

$$P(A+B|C) = P(A|C) + P(B|C) - P(A,B|C) \tag{4}$$

where $P(A+B|C)$ means the probability that "A or B is true given that C is true." If the propositions A and B are mutually exclusive, that is the probability $P(A,B|C)$ is zero, then the sum rule becomes

$$P(A+B|C) = P(A|C) + P(B|C). \tag{5}$$

The sum rule will prove to be useful in parameter estimation problems, because it allows one to investigate an interesting parameter while removing an uninteresting parameter from consideration. To see this, suppose one wished to compute the probability of A given C, but there is a third proposition B which must be taken into account. To make this more concrete, suppose C stands for the data, A the frequency of a sinusoidal oscillation, and B the amplitude of the sinusoid. Now suppose one wishes to compute the probability of the frequency given the data, $P(A|C)$. But there is this other parameter B; the way to proceed is to compute the joint probability of the frequency and the amplitude given the data, $P(A,B|C)$, and then use the sum rule to eliminate the amplitude B. Suppose, for arguments sake, the amplitude B could take on one of two mutually exclusive values $B = \{B_1, B_2\}$. If one computes the probability of the frequency and (amplitude 1 or amplitude 2) given the data then

$$P(A|C) \equiv P(A,[B_1+B_2]|C) = P(A,B_1|C) + P(A,B_2|C). \tag{6}$$

This probability distribution summarizes all the information in the data relevant to the proposition A, and $P(A|C)$ is called the marginal probability of A given C. The marginal probability does not depend on the amplitude at all. To see this, suppose that the proposition A had not been present, then $P(B_1+B_2|C) = P(B_1|C) + P(B_2|C) = 1$, thus when a parameter has been marginalized from a joint probability distribution the reference to that parameter is dropped because the probability distribution no longer depends on the parameter. Of course the amplitude could take on more than two values, for example $B_j = \{B_1, \cdots, B_m\}$, in which case the marginal probability distribution would become

$$P(A|C) = \sum_{j=1}^{m} P(A,B_j|C) \tag{7}$$

provided the amplitudes are mutually exclusive. In real problems, the parameter B would take on a continuum of values; but *as long as B is a constant through the run of the data* the sum rule becomes

$$P(A|C) = \int dB P(A,B|C), \tag{8}$$

where the integral is over all possible values of the parameter B. It is in this form that the sum rule will be used in the example to remove uninteresting or *nuisance* parameters. Note

that dB refers to a number, while B appearing inside $P(A, B|C)$ refers to a proposition. A notation could be invented to reflect this distinction, but it is unnecessary; provided that one realizes that when a capital letter appears outside of a probability symbol it refers to a number associated with a proposition, while inside the probability symbol it refers to the proposition.

3 Assigning Probabilities

The product rule Eq. (1), Bayes theorem Eq. (3), and the sum rule in the form of Eq. (8) are the only rules of probability theory needed to solve most problems of inference. But these rules tell one how to manipulate probabilities after they have been assigned; *there is nothing within the theory to tell one how to assign probabilities.* This must come from outside probability theory and is one of the major reasons why probability theory, as formulated by Laplace, was rejected. If probabilities are frequencies there is no problem in assigning their values; it is only when probabilities are interpreted as a reasonable degree of belief, that their assignment becomes a real question. There are many ways to address the assignment of probabilities – see Ref. [8, 10, 11, 12] for more on this. Here, the principle of maximum entropy will be used to assign all probabilities – see Ref. [13] for an extended discussion of maximum entropy.

Suppose one has a discrete probability distribution $P(i|I)$, where i stands for some proposition (for example, the probability of one of the faces of a die) and I represents the information on which the probability distribution is based, then Shannon's H theorem [4] states that

$$H \equiv -\sum_{i=1}^{N} P(i|I) \log P(i|I) \tag{9}$$

is a measure of the amount of ignorance in the probability distribution. Shannon's theorem is derived based on a qualitative requirement plus the requirement that the measure be consistent. The principle of maximum entropy then states that if one has some testable information I, one can assign a probability distribution to a proposition i such that $P(i|I)$ contains only the information I. This assignment is done by maximizing H subject to the constraints represented by the information I.

To demonstrate its use, suppose one has a die and wishes to assign a probability to each of the six faces. If nothing is known about the die except that the probabilities must total 1, then

$$1 - \sum_{j=1}^{6} P(j|I) = 0 \tag{10}$$

must be satisfied. If this constraint is satisfied, then one can multiply it by a constant β (called a Lagrange multiplier) and because the constraint is zero it can be add it to the entropy without changing the value of H:

$$H = -\sum_{j=1}^{6} P(j|I) \log P(j|I) + \beta \left[1 - \sum_{j=1}^{6} P(j|I) \right]. \tag{11}$$

But the probabilities and β are not known; they must be assigned. To assign the probabilities, H is constrained to be a maximum with respect to variations in the unknowns: the probabilities, and β. Because H measures the amount of ignorance in the probability distribution, constraining H to be a maximum is asking for the least informative (highest entropy) probability distribution subject to the known prior information. This maximum is located by differentiating H with respect to $P(k|I)$ and β, and setting derivatives equal to zero. Here there are six unknown probabilities and one unknown Lagrange multiplier. But when the derivatives are taken, there will be seven equations; thus all of the unknowns are completely determined. This system of equations is then solved for the values of $P(j|I)$ and for β. Taking the derivatives one obtains

$$\begin{aligned} -[\log P(k|I) + 1] + \beta &= 0, \\ 1 - \sum_{j=1}^{6} P(j|I) &= 0 \end{aligned}$$

from which one finds

$$P(j|I) = \frac{1}{6} \qquad \text{and} \qquad \beta = 1 - \log 6. \tag{12}$$

When nothing is known, except that the probability distribution should be normalized, the principle of maximum entropy reduces to the uniform prior. This is Laplace's principle of insufficient reason [1]. But the principle maximum entropy is much more general because it allows one to assign probabilities that are maximally uninformative, while still incorporating the known information. If information were available that indicated that the die was not honest, then this information could be used in a maximum entropy calculation to obtain a nonuniform probability distribution – see Ref. [14] for this calculation and much more on maximum entropy.

The principle of maximum entropy represents a way of assigning probabilities based only on the information that one actually possesses. In the previous example, that information was that the probabilities should total one. Almost any information can be incorporated into a maximum entropy calculation. In the example that follows, the joint probability $P(\mathbf{e}|I)$ of a set of noise values will be needed, where $\mathbf{e} \equiv \{e_1, \cdots, e_N\}$. One should read $P(\mathbf{e}|I)$ as the probability that the noise should have value e_1 at time t_1, and value e_2 at time t_2, etc. This probability density will be derived as a second example of the principle of maximum entropy.

In real experiments, not much is actually known about the noise. Typically when an experiment is performed, any systematic component in the data is defined to be the "signal" and the random part is defined to be noise. With this definition of "signal" the noise must have zero mean value

$$\frac{1}{N} \sum_{i=1}^{N} e_i = 0 \tag{13}$$

where N is the total number of data values. But even though the mean value of the noise is zero, it will have a mean-square value:

$$\frac{1}{N} \sum_{i=1}^{N} e_i^2 = \sigma^2. \tag{14}$$

Assuming the second moment of the noise exists, implies only that the noise is present; it does not assume that σ^2 is actually known. In addition to the second moment, there could be higher moments

$$\frac{1}{N}\sum_{i=1}^{N} e_i^x = \gamma_x \tag{15}$$

where x is an integer and γ_x represents the average value of the moment. Additionally, there could be correlations such that

$$\frac{1}{N-m}\sum_{i=1}^{N-m} e_i e_{i+m} = \rho_m \qquad (1 < m < N) \tag{16}$$

where ρ_m is a correlation coefficient. However without seeing several samples of the noise it is not known if correlations exist or if higher moments exist, except of course what is implied by Eq. (14). All that is actually known is that Eq. (14) must be satisfied.

Maximum entropy can be used to assign a probability density function that incorporates only what is actually known. If the noise has mean-square σ^2, then the expected value of e_j^2 is given by

$$\gamma_j\left[\sigma^2 - \int d\mathbf{e} e_j^2 P(\mathbf{e}|I)\right] = 0 \tag{17}$$

where γ_j is a Lagrange multiplier. This equation must be true for every e_j^2. All proper probability density functions must be normalized so

$$\beta\left[1 - \int d\mathbf{e} P(\mathbf{e}|I)\right] = 0. \tag{18}$$

If there are N data values, there are $N+1$ constraints and the entropy functional to be maximized is given by

$$\begin{aligned} H &= -\int d\mathbf{e} P(\mathbf{e}|\sigma,I)\log P(\mathbf{e}|\sigma,I) + \beta\left[1 - \int d\mathbf{e} P(\mathbf{e}|\sigma,I)\right] \\ &+ \sum_{j=1}^{N}\gamma_j\left[\sigma^2 - \int d\mathbf{e} e_j^2 P(\mathbf{e}|\sigma,I)\right] \end{aligned} \tag{19}$$

where the notation $P(\mathbf{e}|\sigma,I)$ has been adopted to indicate that it is the probability of the noise given σ and the information I.

This case is fundamentally different from the previous example, because $P(\mathbf{e}|\sigma,I)$ is a function of N continuous variables. Instead of having to determine a fixed number of unknown probabilities there are infinitely many probabilities corresponding to the continuous variables. These must be determined from the $N+1$ constraints. The principle of maximum entropy may still be used to assign the *probability density function* by taking the derivatives and solving the resulting system of equations. Taking the derivatives of H with respect to $P(\mathbf{e}|\sigma,I)$ gives

$$-\left[\log P(\mathbf{e}|\sigma,I) + 1\right] - \beta - \sum_{i=1}^{N}\gamma_i e_i^2 = 0 \tag{20}$$

as one of the equations, and taking the derivatives with respect to the Lagrange multipliers just returns the constraint equations:

$$1 - \int d\mathbf{e} P(\mathbf{e}|\sigma, I) = 0 \tag{21}$$

and

$$\sigma^2 - \int d\mathbf{e} e_j^2 P(\mathbf{e}|\sigma, I) = 0. \tag{22}$$

Solving this system of equations one finds the probability of the noise to be

$$P(\mathbf{e}|\sigma, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{e_i^2}{2\sigma^2}\right\} \tag{23}$$

where

$$\gamma_j = \frac{1}{2\sigma^2} \tag{24}$$

and

$$\beta = \log\left[\int d\mathbf{e} \exp\left\{-\sum_{i=1}^{N} \frac{e_i^2}{2\sigma^2}\right\}\right] - 1. \tag{25}$$

There are several interesting points to note about this probability density function. First, it was not assumed that the noise was correlated and the resulting maximum entropy probability density function does not contain correlations. Had Eq. (16) been used as a constraint, the resulting probability density function would have been very different. Second, no assumptions were made about the odd moments, and the resulting probability density function has no odd moments. Third, if one computes the expected value of the even moments one finds

$$(2\pi\sigma^2)^{-\frac{N}{2}} \int_{-\infty}^{+\infty} de_1 \cdots de_N e_j^{2n} \exp\left\{-\sum_{i=1}^{N} \frac{e_i^2}{2\sigma^2}\right\} = 1 \cdot 3 \cdot 5 (2n-1)\sigma^{2n}, \tag{26}$$

which for $n = 1$ reduces to

$$(2\pi\sigma^2)^{-\frac{N}{2}} \int_{-\infty}^{+\infty} de_1 \cdots de_N e_j^{2n} \exp\left\{-\sum_{i=1}^{N} \frac{e_i^2}{2\sigma^2}\right\} = \sigma^2, \tag{27}$$

just what one would expect to find. Thus maximum entropy has not introduced any spurious correlations or effects into the probability density function that were not already implicit in the constraints. Equation (23) is the least informative probability density function that is consistent with the given second moment. If one has information that correlations exist or that the odd moments are not zero, that information can always be used in a maximum entropy calculation to assign a probability density function to the noise, that probability density will always have more compact support (lower entropy) for a given value of σ than Eq. (23). Therefore, this new noise probability density function will always make more precise estimates of the parameters.

4 Example – Estimating a Frequency

In the second section the rules of probability theory were outlined, and in the third section the procedures for assigning probability distributions were given, in this section a nontrivial parameter estimation problem is worked. Each step in the calculation explained in detail. The example is sufficiently complex to illustrate all of the points of principle that must be faced in more complex problems, yet sufficiently simple that anyone with a background in calculus should be able to follow the mathematics, and additionally it gives an important and surprising result.

In this example the probability of a frequency of oscillation ω is computed conditional on the data D and the prior information I. This probability density will be computed from Bayes' theorem, but first, exactly what problem is being solved must be explained. In this example, there is a time series $y(t)$ that is being considered. The time series is postulated to contain a single stationary sinusoidal signal $f(t)$ plus noise. The basic model is: a discrete data set $D = \{d_1, \cdots, d_N\}$ has been recorded; it is sampled from $y(t)$ at discrete times $\{t_1, \cdots, t_N\}$; with model equation

$$d_i = y(t_i) = f(t_i) + e_i, \qquad (1 \leq i \leq N) \tag{28}$$

where e_i represents the numerical value of the noise at time t_i. It is possible in different problems to have other types of noise. For example, the noise could be multiplicative, or it could be simple digitizing noise. The signal $f(t)$ for a single sinusoidal frequency may be written

$$f(t) = B_1 \cos(\omega t) + B_2 \sin(\omega t) \tag{29}$$

which has three parameters (B_1, B_2, ω) that may be estimated from the data.

The problem to be solved is to compute the probability of the frequency ω conditional on the data and the prior information, this is abbreviated as $P(\omega|D, I)$. But Eq. (29) has two other parameters, effectively the amplitude and phase of the sinusoid. In this problem the two parameters B_1 and B_2 are referred to as *nuisance parameters,* because the probability distribution that is to be calculated does not depend on these parameters. To perform this calculation one applies Bayes' theorem to compute the joint probability of all of the parameters and then uses the sum rule, Eq. (8), to eliminate the nuisance parameters. Applying Bayes' theorem gives

$$P(\omega, B_1, B_2|D, I) = \frac{P(\omega, B_1, B_2|I)P(D|\omega, B_1, B_2, I)}{P(D|I)} \tag{30}$$

which indicates that to compute the joint probability density one must obtain three terms. The first term, $P(\omega, B_1, B_2|I)$, is the probability of the parameters given only the information I. This term is referred to as a prior probability density, or simply as a prior. The second term, $P(D|\omega, B_1, B_2, I)$, is the probability of the data given the parameters and the information I. This term is called the direct probability density of the data, and it is often referred to as a likelihood function when one considers a single data set. The third term, $P(D|I)$, is the probability of the data given only the information I, this term will be shown to be a normalization constant.

Equation (30) is the joint probability density of all the parameters including the amplitudes. The sum rule, Eq. (8), can be applied to remove the dependence on the amplitudes:

$$P(\omega|D,I) = \int dB_1 dB_2 \frac{P(\omega, B_1, B_2|I)P(D|\omega, B_1, B_2, I)}{P(D|I)}. \tag{31}$$

The calculation of the posterior probability density of a single stationary sinusoidal frequency is at least a four step problem: assign the three probability density functions, and then remove the dependence on the amplitudes by integration.

4.1 Assignment of the direct probability

The calculation of each of the three terms proceeds by applying the rules of probability theory and by supplying additional information I when necessary. Each of the terms will be taken separately starting with $P(D|\omega, B_1, B_2, I)$, and followed by $P(\omega, B_1, B_2|I)$. The probability density of the data given the value of the parameters is essentially asking how unlikely is the data, but the name of this term is a little misleading; it is not the probability of the data that is needed here but the probability of the noise. What is taken to be noise depends critically on what one takes to be signal. Equation (28) is a definition of what is meant by the noise. The probability of the noise was assigned in the previous section, Eq. (23). One can take the difference between the data and the model,

$$e_i = d_i - f(t_i) \tag{32}$$

and substitute this into Eq. (23) to obtain

$$P(D|f,\sigma,I) = (2\pi\sigma)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[d_i - f(t_i)]^2\right\} \tag{33}$$

as the probability of the data given the model f, where a new parameter σ^2 (the variance of the noise) has been added. Inserting the single stationary sinusoidal frequency model, Eq. (29), for f and changing notation $f \to \omega, B_1, B_2$ to indicate that it is the parameters that interests us, one obtains

$$P(D|\omega, B_1, B_2, \sigma, I) = (2\pi\sigma)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[d_i - B_1\cos(\omega t_i) - B_2\sin(\omega t_i)]^2\right\}. \tag{34}$$

The usual way to proceed is to fit the sum in the exponent. Finding the parameter values which minimize this sum is called "least-squares." The equivalent procedure, in this case, of finding parameter values that maximize $P(D|\omega, B_1, B_2, \sigma, I)$ is called "maximum-likelihood." The maximum-likelihood procedure is more general than least-squares: it has theoretical justification when the likelihood is not Gaussian.

In this calculation uniform sampling will be assumed, because it simplifies some of the analytic details. Expanding the exponent in Eq. (34) one obtains

$$P(D|B_1, B_2, \omega, \sigma, I) = (2\pi\sigma)^{-\frac{N}{2}} \exp\left\{-\frac{Q}{2\sigma^2}\right\} \tag{35}$$

where

$$Q \equiv N\overline{d^2} - 2[B_1 R(\omega) + B_2 I(\omega)] + B_1^2 c + B_2^2 s, \tag{36}$$

and

$$R(\omega) = \sum_{i=1}^{N} d_i \cos(\omega t_i) \tag{37}$$

$$I(\omega) = \sum_{i=1}^{N} d_i \sin(\omega t_i) \tag{38}$$

are the cosine and sine transforms of the data,

$$\overline{d^2} = \frac{1}{N}\sum_{i=1}^{N} d_i^2 \tag{39}$$

is the observed mean-square data value, and

$$c \equiv \sum_{i=1}^{N} \cos^2(\omega t_i) = \frac{N}{2} + \frac{\sin(N\omega)}{2\sin(\omega)} \tag{40}$$

$$s \equiv \sum_{i=1}^{N} \sin^2(\omega t_i) = \frac{N}{2} - \frac{\sin(N\omega)}{2\sin(\omega)} \tag{41}$$

where the convention $t_i = \{-T, -T+1, \cdots, T-1, T\}$, and $2T+1 = N$ has been adopted. Use of this convention means that frequencies are measured in radians and may take on values between zero and 2π. The cross term, $\sum_{i=1}^{N} \cos(\omega t_i)\sin(\omega t_i)$, is zero for uniform sampling.

The assignment of the direct probability of the data is now complete. However, the use of Eq. (23) has modified the problem to include a fourth parameter, the variance of the noise σ^2. Inserting Eq. (35) into Eq. (31) one obtains

$$P(\omega|\sigma, D, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \int dB_1 dB_2 \frac{P(\omega, B_1, B_2|I)\exp\{-Q/2\sigma^2\}}{P(D|I)} \tag{42}$$

as the posterior probability of the frequency, where the parameter σ is assumed known. If the variance is not known, at the end of the calculation the product and sum rules may be used to remove the variance from the problem, just as the amplitudes will be removed.

4.2 Assignment of the prior probability

Assigning the prior probability is one of the most controversial areas in Bayesian probability. Yet, to a Bayesian it is the most natural of things. No one would think of trying to solve any problem in everyday life without bringing to bear all of his prior experiences. In traditional frequency interpretation of probability theory assigning a prior probability makes no sense, because prior information has no frequency interpretation. The sampling theorist finds himself in the uncomfortable position of trying to solve a problem in which he is not allowed to use what he knows, while trying to justify his use of a particular model. Most of the controversy arises when one is trying to solve a problem in which one

has little prior information. If one has highly informative prior information, such as a prior measurement, there is little discussion on how to assign priors: one simply uses the posterior probability derived in analyzing the previous measurement as the prior probability for the current measurement. But this type of argument simply delays the problem of how to assign a probability to represent knowing little. In this tutorial the prior probabilities will be assigned to indicate that little, effectively no, prior information is available.

In assigning the prior probability $P(B_1, B_2, \omega|I)$ exactly what is known about the parameters will have to be stated. Before doing this note that the prior probability is the joint probability of the amplitudes and the frequency. But from the product rule, Eq. (1), this can be factored:

$$P(B_1, B_2, \omega|I) = P(B_1, B_2|\omega, I)P(\omega|I). \tag{43}$$

The prior probability of the frequency may be assigned completely independent of the values of the amplitudes.

Here the only thing known about the frequency is that the data has been sampled uniformly, thus frequency values greater than the Nyquist frequency are aliased. So if the experimenter was even reasonably competent, the frequency must be bounded between 0 and 2π radians. Using this bound and the normalization constraint in a maximum entropy calculation results in the assignment of

$$P(\omega|I) = \frac{1}{2\pi} \tag{44}$$

as the prior probability of the frequency. Of course this is not the only prior probability that could be assigned. Indeed in [11] very convincing arguments are given that demonstrate a frequency to be scale parameter for which a Jeffreys' prior [2] is appropriate. There is no contradiction in arriving at different prior probability assignments. The two different assignments correspond to being in different states of knowledge, and people with different prior information I will naturally make different assignments. But this probability assignment represent knowing little, effectively nothing, and regardless of what functional form one assigns to the prior; *if the prior is slowly varying compared to the direct probability, the prior will look like a constant over the range of values where the direct probability is sharply peaked* and its behavior outside of this region will make little effectively no difference in the results. It is only when the width of the prior is comparable to the width of the direct probability that it can have any significant effect.

Having assigned the prior probability of the frequency, Eq. (43) becomes

$$P(B_1, B_2, \omega|I) = \frac{P(B_1, B_2|\omega, I)}{2\pi}. \tag{45}$$

The probability of the amplitudes explicitly depends on a given value of the frequency. In this calculation it will be assumed that knowing the frequency tells one nothing about the amplitudes. This is not true in general, for example if the experiment is repeatable and a previous measurement is available, knowledge of the frequency would imply a great deal about the value of the amplitude. But if knowledge of the frequency does not tell one anything about the amplitudes then $P(B_1, B_2|\omega, I) = P(B_1, B_2|I)$ and the joint prior probability of all of the parameters may be written as

$$P(B_1, B_2, \omega|I) = \frac{P(B_1, B_2|I)}{2\pi}. \tag{46}$$

To proceed one must state exactly what is known about the amplitudes. Suppose for arguments sake one repeated this experiment a great number of times. The signal is a stationary sinusoid, when the experiment is repeated each of the amplitudes will take on both positive and negative values, (the phase will be different in each run of the data). Thus the average value of the amplitudes over many runs of the data will be zero, but the mean-square value of the amplitudes will be nonzero. *If one knows nothing else about the amplitudes, then applying the principle of maximum entropy will result in assigning a Gaussian prior probability density to the amplitudes:*

$$P(B_1, B_2|\delta, I) = (2\pi\delta^2)^{-1} \exp\left\{-\frac{B_1^2 + B_2^2}{2\delta^2}\right\}, \tag{47}$$

where δ^2 represents the uncertainty in the amplitudes. If this prior probability is to represent knowing little, then δ must be very large. But if δ is very large *this prior is effectively a uniform prior probability* over the range of values where the direct probability is peaked. Because this prior is to represent knowing little; δ will be taken to be very large, effectively infinite, and the Gaussian prior will be replaced by its limiting value: the uniform prior. This is the first example of what is called an improper prior probability density. An improper prior is a prior probability density that is not normalizable – and is not, strictly speaking, a probability density function at all. When performing a Bayesian probability calculation, one should always use a proper probability density, and then pass to the limit of an improper probability density at the end of the calculation – see Ref. [12] for more on this. But because the Gaussian is so strongly convergent the order of these operations may be interchanged. This is not always true and when in doubt the only safe course is to use a proper prior and then pass to the limit of an improper prior at the end of the calculation.

With the assignment of the uniform prior for the amplitudes the joint prior probability density of all of the parameters is given by

$$P(\omega, B_1, B_2|\delta, I) = \frac{1}{2\pi}. \tag{48}$$

This prior probability density may now be substituted into Eq. (42) to obtain

$$P(\omega|\sigma, D, I) = \frac{(2\pi\sigma^2)^{-\frac{N}{2}}}{2\pi} \int dB_1 dB_2 \frac{\exp\left\{-Q/2\sigma^2\right\}}{P(D|I)}. \tag{49}$$

4.3 Assignment of the prior probability of the data

The prior probability of the data, $P(D|I)$, is essentially a normalization constant in the parameter estimation problems. But to someone unfamiliar with Bayesian calculations this is not obvious, nor is it obvious how to calculate it. The way to proceed is to calculate the joint probability of the data and the parameters, $P(\omega, B_1, B_2, D|I)$. This can be factored using the product rule, Eq. (1), to obtain

$$P(\omega, B_1, B_2, D|I) = P(\omega, B_1, B_2|I)P(D|\omega, B_1, B_2, I). \tag{50}$$

The sum rule may now be applied to remove the dependence on the parameters,

$$P(D|I) = \int d\omega dB_1 dB_2 P(\omega, B_1, B_2|I)P(D|\omega, B_1, B_2, I). \tag{51}$$

Comparing this with Eq. (31) we see that this is just the constant needed to ensure that the total probability is one. So *in parameter estimation problems $P(D|I)$ is a normalization constant,* and in Eq. (49) the irrelevant constants can be dropped provided this probability density function is normalized at the end of the calculation. Dropping these constants, Eq. (49) becomes

$$P(\omega|\sigma, D, I) \propto \sigma^{-N} \int dB_1 dB_2 \exp\left\{-\frac{Q}{2\sigma^2}\right\}. \tag{52}$$

If σ is actually known there are some other constants which may be dropped. However, in real data, the variance of the noise is frequently not known and must be estimated, so the retained terms will be needed later.

Notice that this equation is essentially Eq. (35); the steps in assigning the uninformative priors were unnecessary, they simply cancel from the problem. People familiar with parameter estimation problems using uninformative priors simply skip the intermediate steps and go straight to Eq. (52). That could not be done here because until one has seen the intermediate steps, one simply does not know that the uninformative prior probabilities cancel. Indeed, in model selection problems even uninformative prior probabilities do not cancel, and improper priors simply cannot be used.

4.4 Elimination of the nuisance parameters

Now that all terms in the posterior probability density function have been assigned, Eq. (52) must be integrated with respect to B_1 and B_2. These are Gaussian integrals and any multivariant Gaussian integral can be done – see Ref. [11] for examples of this. In this problem the integrals are particularly simple because the term involving the product $B_1 B_2$ canceled. This cancellation occurred because uniform sampling was assumed. Had nonuniform sampling been assumed, the integrals would be tractable, but the results would be much more complex analytically – see Ref. [11] for this calculation. To do these integrals, the value of Q in Eq. (52) is replaced by its definition and the order of the terms is rearranged to obtain

$$\begin{aligned} P(\omega|\sigma, D, I) \quad &\propto \quad \sigma^{-N} \exp\left\{-\frac{N\overline{d^2}}{2\sigma^2}\right\} \\ &\times \quad \int_{-\infty}^{+\infty} dB_1 \exp\left\{-\frac{c}{2\sigma^2}\left[B_1^2 - 2B_1 R(\omega)/c\right]\right\} \\ &\times \quad \int_{-\infty}^{+\infty} dB_2 \exp\left\{-\frac{s}{2\sigma^2}\left[B_2^2 - 2B_2 I(\omega)/s\right]\right\}. \end{aligned}$$

The squares of each of these Gaussian integrals can be completed by adding and subtracting a constant from the exponent:

$$\begin{aligned} P(\omega|\sigma,D,I) &\propto \sigma^{-N}\exp\left\{-\frac{N\overline{d^2}-R(\omega)^2/c-I(\omega)^2/s}{2\sigma^2}\right\} \\ &\times \int_{-\infty}^{+\infty} dB_1 \exp\left\{-\frac{c}{2\sigma^2}\left[B_1-R(\omega)/c\right]^2\right\} \\ &\times \int_{-\infty}^{+\infty} dB_2 \exp\left\{-\frac{s}{2\sigma^2}\left[B_2-I(\omega)/s\right]^2\right\}. \end{aligned}$$

A change of variables

$$x=\sqrt{\frac{c}{2\sigma^2}}\left[B_1-\frac{R(\omega)}{c}\right] \qquad \text{and} \qquad dx=\sqrt{\frac{c}{2\sigma^2}}dB_1, \tag{53}$$

and

$$y=\sqrt{\frac{s}{2\sigma^2}}\left[B_2-\frac{I(\omega)}{s}\right] \qquad \text{and} \qquad dy=\sqrt{\frac{s}{2\sigma^2}}dB_2 \tag{54}$$

reduces the integrals to standard form

$$P(\omega|\sigma,D,I)\propto\frac{\sigma^{-N+2}}{\sqrt{cs}}\exp\left\{-\frac{N\overline{d^2}-R(\omega)^2/c-I(\omega)^2/s}{2\sigma^2}\right\}\int_{-\infty}^{+\infty}dx\mathrm{e}^{-x^2}\int_{-\infty}^{+\infty}dy\mathrm{e}^{-y^2} \tag{55}$$

where a factor of 2 was dropped. This numerical factor would be absorbed when the probability density is normalized. These integrals can now be done trivially; each contributing $\sqrt{\pi}$, which may also be dropped. This gives

$$P(\omega|\sigma,D,I)\propto\frac{\sigma^{-N+2}}{\sqrt{cs}}\exp\left\{-\frac{N\overline{d^2}-R(\omega)^2/c-I(\omega)^2/s}{2\sigma^2}\right\} \tag{56}$$

as the posterior probability of the frequency independent of the amplitudes.

If the variance of the noise is actually known, then there are several constants that will be absorbed when this probability density function is normalized. Assuming the variance of the noise known and dropping these constants, the posterior probability density is

$$P(\omega|\sigma,D,I)\propto\frac{1}{\sqrt{cs}}\exp\left\{\frac{R(\omega)^2/c+I(\omega)^2/s}{2\sigma^2}\right\}. \tag{57}$$

This is an exact result, and is the posterior probability of a single stationary sinusoidal frequency independent of the amplitude and phase of the sinusoid given the uniformly sampled data, the variance of the noise, and the prior information I.

In this form the sufficient statistic, $[R(\omega)^2/c+I(\omega)^2/s]$, is not very recognizable. There is an approximate result that is simpler and worth investigating. The functions c and s were defined earlier, Eqs. (40) and Eq.(41), and these equations explicitly depend on the frequency. Unless the frequency is near zero, the functions c and s are slowly varying and may be approximated by

$$c\approx\frac{N}{2} \qquad \text{and} \qquad s\approx\frac{N}{2}. \tag{58}$$

The posterior probability is approximately

$$P(\omega|\sigma, D, I) \propto \exp\left\{\frac{C(\omega)}{\sigma^2}\right\}, \tag{59}$$

where $C(\omega)$ is the Schuster periodogram [15] and may be defined as

$$C(\omega) = \frac{2R(\omega)^2 + 2I(\omega)^2}{N} = \frac{2}{N}\left|\sum_{k=1}^{N} d_k \exp\{-i\omega t_k\}\right|^2. \tag{60}$$

The Schuster periodogram is traditionally referred to as a discrete Fourier transform power spectrum. From the standpoint of Bayesian probability, the discrete Fourier transform power spectrum answers a very specific question about single stationary sinusoidal frequency estimation.

These simple results, Eqs. (57,59,60), show why the discrete Fourier transform tends to peak at the location of a frequency when the data are noisy. Namely, the discrete Fourier transform power spectrum is directly related to the probability that a single stationary sinusoidal frequency is present in the data. Additionally, zero padding a time series (i.e. adding zeros at its end to make a longer time series) and then performing the discrete Fourier transform of the padded series, is equivalent to calculating the Schuster periodogram at smaller frequency intervals. If the signal one is analyzing is a single stationary sinusoidal frequency plus noise, then the maximum of the periodogram will be the "best" estimate of the frequency one can make in the absence of additional prior information about it.

The discrete Fourier transform and the Schuster periodogram can now be seen in a entirely new light: the highest peak in the discrete Fourier transform is an optimal frequency estimator for a data set which contains a single stationary sinusoidal frequency in the presence of Gaussian white noise. Stated more carefully, the discrete Fourier transform will give optimal frequency estimates if six conditions are met:

1. The number of data values N is large,
2. There is no constant component in the data,
3. There is no evidence of a low frequency,
4. The data contain only one frequency,
5. The frequency is be stationary
   (i.e. the amplitude and phase are constant), and
6. The noise is white.

(61)

If any of these six conditions is not met, the discrete Fourier transform may give misleading or simply incorrect results in light of the more realistic models. Not because the discrete Fourier transform is wrong, but because it is answering what should be regarded as the wrong question. The discrete Fourier transform will always interpret the data in terms of a single stationary sinusoidal frequency model! The effects of violating one or more of these assumptions are shown in [11] and it is demonstrated that when the assumptions are violated, the range of parameter values that are consistent with the data is larger than when these conditions are met.

4.5 Eliminating the variance of the noise

Equation (57) is an exact result and is valid for any uniformly sampled data set while Eq. (59) is valid provided there is no evidence of a low frequency, and $N \gg 1$. Both results depend on knowing the variance of the noise. Frequently one has no independent knowledge of the noise. The noise variance σ^2 then becomes a nuisance parameter. It can be eliminated by first applying the product rule, to Eq. (56) followed by the sum rule Eq. (8). This gives

$$P(\omega|D,I) \propto \int_0^{+\infty} d\sigma P(\sigma|I) \frac{\sigma^{-N+2}}{\sqrt{cs}} \exp\left\{-\frac{N\overline{d^2} - R(\omega)^2/c - I(\omega)^2/s}{2\sigma^2}\right\} \tag{62}$$

as the posterior probability of the frequency independent of the amplitudes and the variance of the noise.

To perform this integral one must supply $P(\sigma|I)$, the prior probability of the variance of the noise σ^2. The derivation of the prior which indicates "complete ignorance" of a scale parameter was first given by Jeffreys [2] using invariance arguments. It has since been derived by Jaynes [12] using what is called the marginalization paradox, and by Zellner [8] using the principle of maximum entropy. This prior, $1/\sigma$, is called a Jeffreys prior and is the second example of an improper prior. The first example was the uniform prior when extended to an infinite region – see Ref. [11] for more on improper priors. As was mentioned earlier, when using improper priors one should begin by using a proper prior and then pass to the limit of an improper prior; only then can one be absolutely sure that the limits are well-behaved. But in parameter estimation problems using a Gaussian noise prior this limit is uneventful. Multiplying Eq. (56) by the Jeffreys prior gives

$$P(\omega|D,I) \propto \frac{1}{\sqrt{cs}} \int_0^{+\infty} d\sigma \sigma^{-N+1} \exp\left\{-\frac{Q}{\sigma^2}\right\} \tag{63}$$

as the posterior probability of the frequency, where Q is now taken to be

$$Q \equiv \frac{N\overline{d^2} - R(\omega)^2/c - I(\omega)^2/s}{2}. \tag{64}$$

The integral may be transformed into a gamma integral by the following change of variables:

$$\sigma = \sqrt{\frac{Q}{x}} \qquad \text{and} \qquad d\sigma = -\frac{\sqrt{Q}x^{-\frac{3}{2}}}{2}dx. \tag{65}$$

Using this change of variables Eq. (63) becomes

$$P(\omega|D,I) \propto \frac{Q^{-\frac{N-2}{2}}}{\sqrt{cs}} \int_0^{+\infty} dx x^{\frac{N-2}{2}-1} \mathrm{e}^{-x} \tag{66}$$

in standard form for a gamma integral and another factor of 2 was dropped. Performing the integral gives

$$P(\omega|D,I) \propto \frac{1}{\sqrt{cs}} \Gamma\left(\frac{N-2}{2}\right) \left[\frac{N\overline{d^2} - R(\omega)^2/c - I(\omega)^2/s}{2}\right]^{\frac{2-N}{2}} \tag{67}$$

where $\Gamma(x)$ is a gamma function of argument x. Numerous constants have already been dropped, these constants will cancel when this distribution is normalized. Here the gamma function and the factor of 2 may be dropped to obtain

$$P(\omega|D,I) \propto \frac{1}{\sqrt{cs}}\left[N\overline{d^2} - R(\omega)^2/c + I(\omega)^2/s\right]^{\frac{2-N}{2}} \approx \left[N\overline{d^2} - 2C(\omega)\right]^{\frac{2-N}{2}}. \qquad (68)$$

This is called a "Student t-distribution" for historical reasons, although it is expressed here in very nonstandard notation. In this case it is the posterior probability density that a single stationary sinusoidal frequency ω is present in the data when no prior information about the variance of the noise is available.

4.6 Resolving Power

At this point in the calculation the formal derivation of the posterior probability of the frequency is completed. But the problem of estimating the frequency of oscillation is essentially only half complete. In addition to estimating the value of the frequency, one needs to determine the accuracy of the estimate. Of course, this is given by the width of the probability density function, Eqs. (57-59), when the variance of the noise is known and Eq. (68) when the variance of the noise is unknown. But these equations do not aid understanding just how accurately the frequency has been determined; nor do they indicate what is to be gained by using Bayesian probability theory.

To understand what is to be gained from using Bayesian probability theory, an estimate of the width of the posterior probability is derived. The technique used has proven useful in a number of other examples [16, 17, 18]. To determine the precision of the frequency estimate, $C(\omega)$ is Taylor expanded about the maximum as a function of ω. Equation (59) is used in this derivation so that the results may be directly compared to the discrete Fourier transform power spectrum. This calculation results in a Gaussian approximation of the posterior probability from which the (mean) $\pm$ (standard deviation) estimates of the frequency ω is obtained. Expanding $C(\omega)$, about the maximum $\hat{\omega}$, one obtains

$$C(\omega) = C(\hat{\omega}) - \frac{b}{2}(\hat{\omega} - \omega)^2 + \cdots \qquad (69)$$

where

$$b \equiv -\left.\frac{\partial^2 C(\omega)}{\partial \omega^2}\right|_{\hat{\omega}} > 0. \qquad (70)$$

The Gaussian approximation is

$$P(\omega|D,\sigma,I) \simeq \left[\frac{2b}{\pi\sigma^2}\right]^{\frac{1}{2}} \exp\left\{-\frac{b(\hat{\omega}-\omega)^2}{2\sigma^2}\right\} \qquad (71)$$

from which the (mean) $\pm$ (standard deviation) estimate of the frequency is

$$\omega_{\rm est} = \hat{\omega} \pm \frac{\sigma}{\sqrt{b}}. \qquad (72)$$

The accuracy depends on the curvature of $C(\omega)$ at its peak, not on the height of $C(\omega)$. For example, if the data are composed of a sinusoid plus noise, e_i of standard deviation σ,

$$d_i = \hat{B}_1 \cos(\hat{\omega} t_i) + \hat{B}_2 \sin(\hat{\omega} t_i) + e_i \tag{73}$$

where $\hat{B}_1$ and $\hat{B}_2$ are the true amplitudes of the signal and $\hat{\omega}$ is the true frequency. A closed form for $C(\omega)$ may be obtained by noting

$$\begin{aligned}
R(\omega) &= \sum_{i=-T}^{+T} d_i \cos(\omega i) \\
&= \sum_{i=-T}^{+T} \hat{B}_1 \cos(\hat{\omega} i)\cos(\omega i) + \hat{B}_2 \sin(\hat{\omega} i)\cos(\omega i) + e_i \cos(\omega i) \\
&\approx \sum_{i=-T}^{+T} \hat{B}_1 \cos(\hat{\omega} i)\cos(\omega i) + \hat{B}_2 \sin(\hat{\omega} i)\cos(\omega i) \\
&= \frac{\hat{B}_1}{2}\left[\frac{\sin\frac{N}{2}(\hat{\omega}-\omega)}{\sin\frac{1}{2}(\hat{\omega}-\omega)} + \frac{\sin\frac{N}{2}(\hat{\omega}+\omega)}{\sin\frac{1}{2}(\hat{\omega}+\omega)}\right], \\
I(\omega) &= \sum_{i=-T}^{+T} d_i \sin(\omega i) \\
&= \sum_{i=-T}^{+T} \hat{B}_1 \cos(\hat{\omega} i)\sin(\omega i) + \hat{B}_2 \sin(\hat{\omega} i)\sin(\omega i) + e_i \sin(\omega i) \\
&\approx \sum_{i=-T}^{+T} \hat{B}_1 \cos(\hat{\omega} i)\sin(\omega i) + \hat{B}_2 \sin(\hat{\omega} i)\sin(\omega i) \\
&= \frac{\hat{B}_2}{2}\left[\frac{\sin\frac{N}{2}(\hat{\omega}-\omega)}{\sin\frac{1}{2}(\hat{\omega}-\omega)} - \frac{\sin\frac{N}{2}(\hat{\omega}+\omega)}{\sin\frac{1}{2}(\hat{\omega}+\omega)}\right].
\end{aligned}$$

The sums appearing in $R(\omega)$ and $I(\omega)$ were done by first changing the sines and cosines to exponential notation. In this notation the sums are of the form $\sum_{i=1}^{N} x^i$, which may be summed using the fact that $1/(1-x) = 1 + x + x^2 + \cdots$. In the above it is only the terms involving $\hat{\omega} - \omega$ that are large, so $C(\omega)$ may be approximated by

$$C(\omega) \approx \frac{(\hat{B}_1^2 + \hat{B}_2^2)}{4N}\left[\frac{\sin\frac{N}{2}(\hat{\omega}-\omega)}{\sin\frac{1}{2}(\hat{\omega}-\omega)}\right]^2. \tag{74}$$

So, as found by Jaynes [19], the estimate of the frequency is given by

$$C(\hat{\omega}) \simeq \frac{N}{4}(\hat{B}_1^2 + \hat{B}_2^2) \tag{75}$$

$$b \simeq \frac{(\hat{B}_1^2 + \hat{B}_2^2)N^3}{48} \tag{76}$$

$$(\omega)_{\rm est} = \hat{\omega} \pm \sigma\sqrt{\frac{48}{N^3(\hat{B}_1^2 + \hat{B}_2^2)}} \tag{77}$$

which indicates, as intuition would lead us to expect, that the accuracy depends on the signal-to-noise ratio, and quite strongly on how much data are available.

These results can be further compared with experience, but first note that dimensionless units have been used. To convert to ordinary physical units, let the sampling interval be Δt seconds, and denote by f the frequency in Hz. Then the total number of cycles in the data record is

$$\frac{\hat{\omega}(N-1)}{2\pi} = (N-1)\hat{f}\Delta t = \hat{f}T \tag{78}$$

where $T = (N-1)\Delta t$ seconds is the duration of our data run. So the conversion of dimensionless ω to f in physical units is

$$f = \frac{\omega}{2\pi\Delta t} \ \text{Hz}. \tag{79}$$

The frequency estimate Eq. (77) becomes

$$f_{\rm est} = \hat{f} \pm \delta f \ \ \text{Hz} \tag{80}$$

where now, not distinguishing between N and $(N-1)$,

$$\delta f = \frac{\sigma}{2\pi T}\sqrt{\frac{48}{N(\hat{B}_1^2 + \hat{B}_2^2)}} = \frac{1.1\sigma}{T\sqrt{N(\hat{B}_1^2 + \hat{B}_2^2)}} \ \text{Hz}. \tag{81}$$

From this one can see that the two most important factors for improving resolution are how long one samples (the T dependence) and the signal-to-noise ratio. The number of data values may be doubled in one of two ways, by doubling the total sampling time or by doubling the sampling rate. However, Eq. (81) clearly indicates that doubling the sampling time is to be preferred. This indicates that data values near the beginning and end of a record are most important for frequency estimation, which is in agreement with intuitive common sense.

Consider the following example: Suppose the RMS signal-to-noise ratio (i.e. ratio of RMS signal to RMS noise $\equiv$ S/N) of the data is S/N $= [(\hat{B}_1^2 + \hat{B}_2^2)/2\sigma^2]^{\frac{1}{2}} = 1$, and data is taken every $\Delta t = 10^{-3}$ sec. for $T = 1$ second, thus getting $N = 1000$ data points, then the theoretical accuracy for determining the frequency of a single, steady sinusoid is

$$\delta f = \frac{1.1}{\sqrt{2000}} = 0.025 \ \text{Hz} \tag{82}$$

while the Nyquist frequency for the onset of aliasing is $f_N = (2\Delta t)^{-1} = 500\text{Hz}$ – greater by a factor of 20,000.

To some, this result will be quite startling. Indeed, had the periodogram itself been considered to be a spectrum estimator, one would instead calculated the width of its central peak. A noiseless sinusoid of frequency $\hat{\omega}$ would have a periodogram proportional to

$$C(\omega) \propto \frac{\sin^2\{N(\omega - \hat{\omega})/2\}}{\sin^2\{(\omega - \hat{\omega})/2\}}. \tag{83}$$

The half-width at half amplitude is given by $|N(\hat{\omega}-\omega)/2| = \pi/4$ or $\delta\omega = \pi/2N$. Converting to physical units, the periodogram will have a width of about

$$\delta f = \frac{1}{4N\Delta t} = \frac{1}{4T} = 0.25 \text{ Hz} \tag{84}$$

just ten times greater than the value Eq. (82) indicated by probability theory. This factor of ten is the amount of narrowing produced by the exponential peaking of the periodogram in Eq. (59), even for peak signal-to-RMS noise ratio of one.

But some would consider even the result of Eq. (84) to be a little overoptimistic. The famous Rayleigh criterion [20] for resolving power of an optical instrument supposes that the minimum resolvable frequency difference corresponds to the peak of the periodogram of one sinusoid coming at the first zero of the periodogram of the second. This is twice Eq. (84):

$$\delta f_{\text{Rayleigh}} = \frac{1}{2T} = 0.5 \text{ Hz}. \tag{85}$$

There is a widely believed "folk-theorem" among theoreticians without laboratory experience which seems to confuse the Rayleigh limit with the Heisenberg uncertainty principle, and holds that Eq. (85) is a fundamental irreducible limit of resolution. Of course there is no such theorem, and workers in high resolution NMR have been routinely determining line positions to an accuracy that surpasses the Rayleigh limit by an order of magnitude for thirty years.

The misconception is perhaps strengthened by the curious coincidence that Eq. (85) is also the minimum half-width that can be achieved by a Blackman-Tukey spectrum analysis [21] (even at infinite signal-to-noise ratio) because the "Hanning window" tapering function that is applied to the data to suppress side-lobes (the secondary maxima of $[\sin(x)/x]^2$) just doubles the width of the periodogram. Since the Blackman-Tukey method has been used widely by economists, oceanographers, geophysicists, and engineers for many years, it has taken on the appearance of an optimum procedure.

According to E.T. Jaynes [22], Tukey himself acknowledged that his method fails to give optimum resolution, but held this to be of no importance because "real time series do not have sharp lines." Nevertheless, this misconception is so strongly held that there have been attacks on the claims of Bayesian/Maximum Entropy spectrum analysts to be able to achieve results like Eq. (82) when the assumed conditions are met. Some have tried to put such results in the same category with circle squaring and perpetual motion machines. Therefore, we want to digress to explain in very elementary physical terms why it is the Bayesian result, Eq. (81), that does correspond to what a skilled experimentalist can achieve.

Suppose first that our only data analysis tool is our own eyes looking at a plot of the raw data of duration $T = 1$ sec., and that the unknown frequency f in Eq. (82) is 100Hz. Now anyone who has looked at a record of a sinusoid and equal amplitude wide-band noise, knows that the cycles are quite visible to the eye. One can count the total number of cycles in the record confidently (using interpolation to help us over the doubtful regions) and will feel quite sure that the count is not in error by even one cycle. Therefore by raw "eyeballing" of the data and counting the cycles, one can achieve an accuracy of

$$\delta f \simeq \frac{1}{T} = 1 \text{ Hz}. \tag{86}$$

But in fact, if one draws the sine wave that seems to fit the data best, one can make a quite reliable estimate of how many quarter-cycles were in the data, and thus achieve

$$\delta f \simeq \frac{1}{4T} = 0.25\,\mathrm{Hz} \tag{87}$$

corresponding just to the periodogram width, Eq. (84).

Then the use of probability theory needs to surpass the naked eye by another factor of ten to achieve the Bayesian width, Eq. (82). What probability theory does is essentially average out the noise in a way that the naked eye cannot do. If some measurement is repeated N times, any randomly varying component of the data will be suppressed relative to the systematic component by a factor of $N^{-\frac{1}{2}}$, the standard rule.

In the case considered, there were $N = 1000$ data points. If they were all independent measurements of the same quantity with the same accuracy, this would suppress the noise by about a factor of 30. But in this case; not all measurements are equally cogent for estimating the frequency. Data points in the middle of the record contribute very little to the result; data points near the ends are highly relevant for determining the frequency, so the effective number of observations is less than 1000. The probability analysis leading to Eq. (82) indicates that the "effective number of observations" is only about $N/10 = 100$; thus the Bayesian width Eq. (82) that results from the exponential peaking of the periodogram now appears to be, if anything, somewhat conservative.

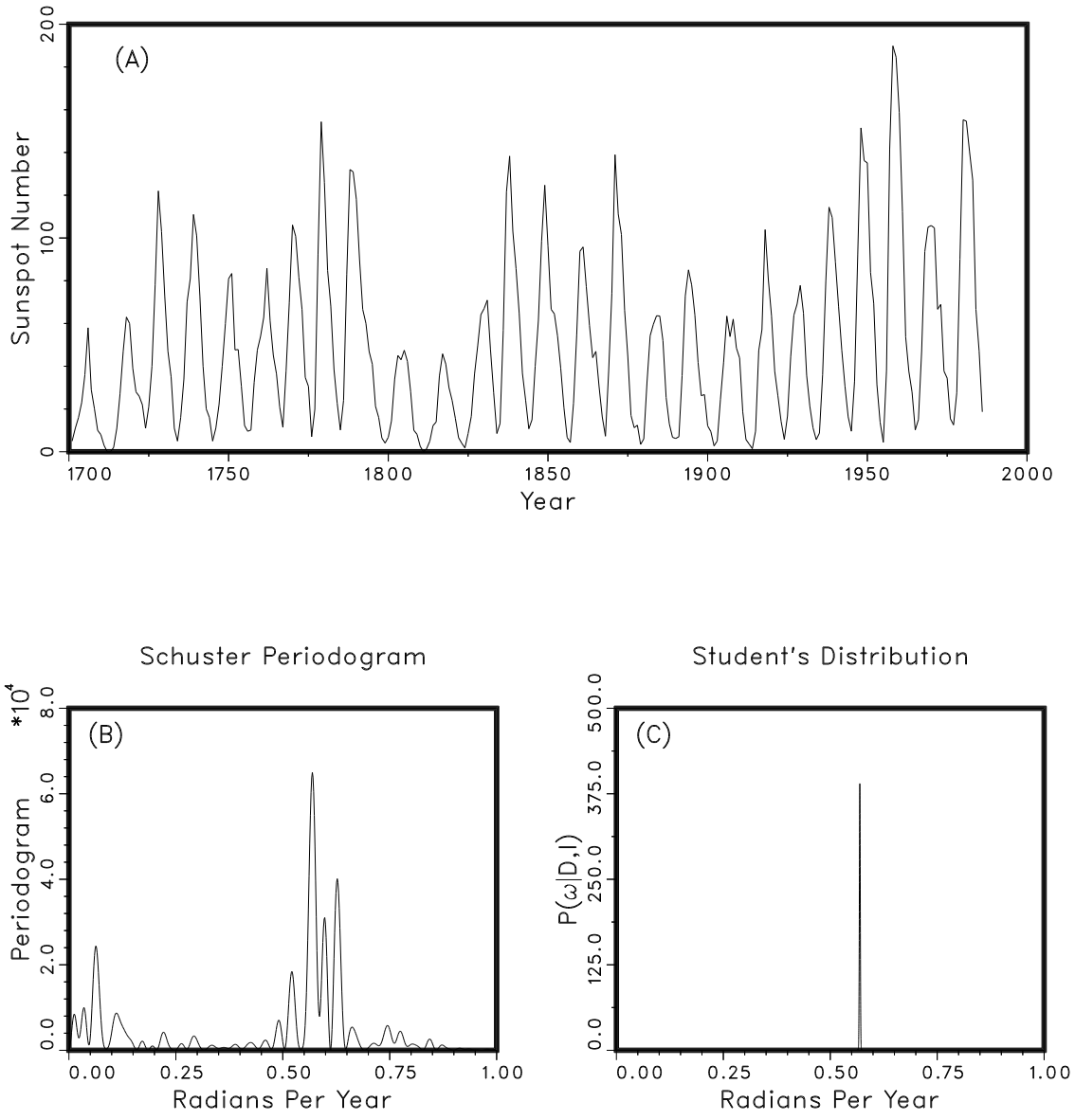
Indeed, that is what Bayesian analysis always does when smooth, uninformative priors for the parameters are used, because then probability theory makes allowance for all possible values that they might have. As noted before, if cogent prior information about ω was available and it was expressed in a narrower prior, still better results would be obtained; but they would not be much better unless the prior range became comparable to the width of the probability of the data.

4.7 Wolf's Relative Sunspot Numbers

Wolf's relative sunspot numbers [24] are, perhaps, the most analyzed set of data in all of spectrum analysis. As Marple [23] explains in more detail, these numbers (defined as: $W = k[10g + f]$, where g is the number of sunspot groups, f is the number of individual sunspots, and k is used to reduce different telescopes to a common scale) have been collected on a yearly basis since 1700 and on a monthly basis since 1748. The exact physical mechanism which generates the sunspots is unknown, and no complete theory exists. Different analyses of these numbers have been published more or less regularly since their tabulation began. Here the sunspot numbers will be analyzed with the stationary sinusoidal frequency model, even though this model is too simple to be realistic for these numbers.

The time series from 1700 to 1985 has been plotted in Fig. 1(A). A cursory examination of this time series does indeed show a cyclic variation with a period of about 11 years. The square of the discrete Fourier transform is a continuous function of frequency and is proportional to the Schuster periodogram of the data [Fig. 1(B), continuous curve]. The frequencies could be restricted to the Nyquist [25, 26] steps [Fig.1(B) open circles]; it is a theorem that the discrete Fourier transform on those points contains all the information that is in the periodogram, but one sees that the information is much more apparent to

Figure 1: Wolf's Relative Sunspot Numbers



Wolf's relative sunspot numbers (A) have been collected on a yearly basis since 1700. The periodogram (B) contains evidence of several complex phenomena. In spite of this the single frequency model posterior probability density (C) picks out the 11.04 year cycle to an estimated accuracy of ± 10 days.

the eye in the continuous periodogram. The Schuster periodogram or the discrete Fourier transform clearly shows a maximum with period near 11 years.

The "Student t-distribution," Eq. (68), was computed and displayed in Fig. 1(C). Now, because of the processing in Eq. (68), all details in the periodogram have been suppressed and only the peak at 11 years remains.

The accuracy of the frequency estimate was determined as follows: the maximum of the "Student t-distribution" was located, an integral was performed about this maximum in a symmetric interval, and the enclosed probability was recorded at a number of points. This gives a period of 11.04 years with the accuracy shown as follows:

period in years		accuracy in years	probability enclosed
11.04	$\pm$	0.015	0.62
	$\pm$	0.020	0.75
	$\pm$	0.026	0.90

According to this, there is not one chance in 10 that the true period differs from 11.04 years by more than 10 days. At first glance, this appears too good to be true. But what does raw "eyeballing" of the data give? In 285 years, there are about $285/11 \approx 26$ cycles. If cycles are counted to an accuracy of $\pm 1/4$ cycle, the period estimate would be about

$$(f)_{\rm est} = 11 \text{ years} \pm 39 \text{ days}. \tag{88}$$

Probability averaging of the noise, as discussed above Eq. (77), would reduce this uncertainty by about a factor of $\sqrt{285/10} = 5.3$, giving

$$(f)_{\rm est} = 11 \text{ years} \pm 7.3 \text{ days}, \quad \text{or} \quad (f)_{\rm est} = 11 \pm 0.02 \text{ years} \tag{89}$$

which corresponds nicely with the result of the probability analysis.

5 Model Selection

But these results come from analyzing the data by a model which assumes there is nothing present but a single sinusoid plus noise. Probability theory, given this model, is obliged to consider everything in the data that cannot be fit to a single sinusoid to be noise. But a glance at the data shows clearly that there is more present than our model assumed. Therefore, probability theory must estimate the noise to be quite large.

This suggests that a more realistic model, which allows the "signal" to have more structure, might do better. Such a model can be fit to the data more accurately; therefore it will estimate the noise to be smaller. This should permit a still better period estimate! But caution forces itself upon us; by adding more and more components to the model the data can always be fit more and more accurately. It is absurd to suppose that by mere proliferation of a model arbitrarily accurate estimates of a parameter can be extracted. There must be a point of diminishing returns – or indeed of negative returns – beyond which we are deceiving ourselves.

It is very important to understand the following point. In parameter estimation problems, probability theory always gives us the estimates that are justified by the information

that was actually used in the calculation. The parameter estimation procedures outlined in this tutorial assume that information to be absolutely true! If one puts false information into a parameter estimation calculation, then probability theory will give optimal estimates based on false information. These could be very misleading.

In real experimental data analysis, one is hardly ever sure of the true model; at best there may be a number of competing candidates. When one has a series of candidate models, probability theory can be used to rank various candidates. Therefore, it can answer questions of the form "Given a set $S \equiv \{f_1, \cdots, f_s\}$ of s possible models, which model is most probable in view of the data and all of one's prior information?" Therefore, probability theory can warn one that the hypothesis may not be true; but one must ask, probability theory will not volunteer the information. To understand how to ask, consider the data $D \equiv \{d_1, \cdots, d_N\}$ sampled at discrete times $\{t_1, \cdots, t_N\}$. If the true signal in the data is f_j, then the data may be modeled as

$$d_i = f_j(t_i) + e_i$$

where $f_j(t_i)$ is the jth member of the set S at time t_i. Suppose the model signal may be written

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

where $G_k(t, \mathbf{\Theta})$ is one of m signal functions with amplitude B_k, $\mathbf{B} \equiv \{B_1, \cdots, B_m\}$, and $\mathbf{\Theta}$ is a set of r nonlinear parameters defined as $\mathbf{\Theta} \equiv \{\Theta_1, \cdots, \Theta_r\}$. The parameters $\mathbf{B}$ and $\mathbf{\Theta}$ are assumed different in every model f_j.

From Bayes' theorem, Eq. (3), the probability of model f_j, conditional on the data D and the prior information I is given by

$$P(f_j|D, I) = \frac{P(f_j|I)P(D|f_j, I)}{P(D|I)} \qquad (1 \leq j \leq s) \tag{90}$$

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

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

The global-likelihood of the data, $P(D|f_j, I)$, is obtained from the joint posterior probability of the data and the parameters,

$$P(D|f_j, I) = \int d\mathbf{B} d\mathbf{\Theta} P(D, \mathbf{B}, \mathbf{\Theta}|f_j, I). \tag{91}$$

In the previous example, only the amplitudes B_1 and B_2 were eliminated from consideration. In model selection problems, both the amplitudes $\mathbf{B}$ and the nonlinear $\mathbf{\Theta}$ parameters must be eliminated from consideration.

The product rule of conditional probability theory may be used to factor $P(D, \mathbf{B}, \mathbf{\Theta}|f_j, I)$ to obtain

$$P(D|f_j, I) = \int d\mathbf{\Theta} P(\mathbf{\Theta}|f_j, I) \left[\int d\mathbf{B} P(\mathbf{B}|\mathbf{\Theta}, f_j, I) P(D|\mathbf{B}, \mathbf{\Theta}, f_j, I) \right]. \tag{92}$$

The term in brackets is just a generalized version of the parameter estimation problem with three changes: 1) all numerical factors must be kept, 2) all parameters are to be considered nuisance parameters, and 3) fully normalized prior probability density functions must be used. In the previous example improper priors were used (prior probabilities that cannot be normalized) because little prior information about the parameters was assumed. Any uninformative prior will look like a constant over the range of values where the likelihood of the parameters is sharply peaked. The uninformative prior would then cancel when the distributions were normalized. The parameters estimated by this procedure are the maximum-likelihood estimates and, in the case of a Gaussian noise prior, the least-squares estimates. However, improper priors cannot be used in Eq. (90) because they do not cancel. This is easily seen. For example, if the prior is a bounded uniform prior, then as the bounds are allowed to go to infinity, the bounds will not cancel in Eq. (90) – unless every model contains exactly the same prior. Thus, the model with the larger number of parameters would automatically be excluded. To perform the indicated calculation see Ref. [11] for details and Ref.[27] for several examples of its use.

6 Conclusions

In this tutorial the rules of Bayesian probability theory, the procedures for assigning probabilities, and a nontrivial example have been given. This example demonstrates the power of Bayesian probability theory and illustrates how to apply the procedures to real problems in parameter estimation. The relation between parameter and model selection has been discussed and it has been shown that the parameter estimation procedures are just one step in the more general model selection problem.

Acknowledgments

This work supported by NIH grant GM-30331, J. J. H. Ackerman principal investigator. The encouragement of Professor J. J. H. Ackerman is greatly appreciated as are the editorial comments of Dr. C. R. Smith and extensive conversations with Professor E. T. Jaynes.

References

[1] Laplace, P. S., *A Philosophical Essay on Probabilities,* unabridged and unaltered reprint of Truscott and Emory translation, Dover Publications, Inc., New York, 1951, original publication data 1814.

[2] Jeffreys, H., *Theory of Probability,* Oxford University Press, London, 1939; Later editions, 1948, 1961.

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

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

[5] Abel, N. H., *Crelle's Jour.,* Bd. 1 (1826).

[6] Cox, R. T., "Probability, Frequency, and Reasonable Expectations," *Amer. J. Phys.* **14,** pp. 1-13 (1946).

[7] Tribus, M., *Rational Descriptions, Decisions and Designs,* Pergamon Press, Oxford, 1969.

[8] Zellner, A., *An Introduction to Bayesian Inference in Econometrics,* John Wiley and Sons, New York, 1971; Second edition 1987.

[9] Bayes, Rev. T., "An Essay Toward Solving a Problem in the Doctrine of Chances," *Philos. Trans. R. Soc. London* **53,** pp. 370-418 (1763); 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.

[10] Jaynes, E. T., "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics,* SSC-4, pp. 227-241 (1968); reprinted in [13].

[11] Bretthorst, G. Larry, "Bayesian Spectrum Analysis and Parameter Estimation," in *Lecture Notes in Statistics* **48,** Springer-Verlag, New York, New York, 1988.

[12] Jaynes, E. T., "Marginalization and Prior Probabilities," in *Bayesian Analysis in Econometrics and Statistics,* A. Zellner, *ed.,* North-Holland Publishing Company, Amsterdam, 1980; reprinted in [13].

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

[14] Jaynes, E. T., "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, 1978; Reprinted in [13].

[15] Schuster, A., "The Periodogram and its Optical Analogy," *Proc. R. Soc. London* **77,** pp. 136 (1905).

[16] Bretthorst, G. Larry, *Bayesian Spectrum Analysis and Parameter Estimation,* Ph.D. thesis, Washington University, St. Louis, MO., available from University Microfilms Inc., Ann Arbor Mich. 1987; an "Excerpts from Bayesian Spectrum Analysis and Parameter Estimation," is printed in *Maximum-Entropy and Bayesian Methods in Science and Engineering* **1,** pp. 75-145, G. J. Erickson and C. R. Smith *Eds.,* Kluwer Academic Publishers, Dordrecht the Netherlands, 1988.

[17] Bretthorst, G. Larry, "Bayesian Spectrum Analysis on Quadrature NMR Data with Noise Correlations," *Maximum Entropy and Bayesian Methods,* pp. 261-274, J. Skilling *ed.,* Kluwer Academic Publishers, Dordrecht the Netherlands, 1989.

[18] Bretthorst, G. Larry and C. Ray Smith, "Bayesian Analysis of Signals from Closely-Spaced Objects," *Infrared Systems and Components III,* pp 93.104, Robert L. Caswell *ed., SPIE* Vol. **1050,** 1989.

[19] Jaynes, E. T., "Bayesian Spectrum and Chirp Analysis," in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems,* pp. 1-37, C. Ray Smith and G. J. Erickson, *Eds.,* Kluwer Academic Publishers, Dordrecht the Netherlands, 1987.

[20] Lord Rayleigh, *Philos. Mag.* **5,** p. 261 (1879).

[21] Blackman, R. B. and J. W. Tukey, *The Measurement of Power Spectra,* Dover Publications, Inc., New York, 1959.

[22] Tukey, J. W., several conversations with E. T. Jaynes, in the period 1980-1983.

[23] Marple, S. L., *Digital Spectral Analysis with Applications,* Prentice-Hall, New Jersey, 1987.

[24] Waldmeier, M., *The Sunspot Activity in the Years 1610-1960,* Schulthes, Zurich, 1961.

[25] Nyquist, H., "Certain Topics in Telegraph Transmission Theory," *Trans. AIEE,* pp. 617 (1928).

[26] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell Sys. Tech. J.* **3,** pp. 324 (1924).

[27] Bretthorst, G. Larry, "Bayesian Model Selection: Examples Relevant to NMR," *Maximum Entropy and Bayesian Methods,* J. Skilling *ed.,* pp. 377-388, Kluwer Academic Publishers, Dordrecht the Netherlands, 1989.