in Maximum Entropy and Bayesian Methods, Wolfgang von der Linden, et. al. *eds.*, Kluwer, 1999.

"THE NEAR-IRRELEVANCE OF SAMPLING FREQUENCY DISTRIBUTIONS[1]"

G. LARRY BRETTHORST
Washington University
Dept. of Chemistry
St. Louis Mo 63130

Abstract. Jaynes, in his unfinished work on probability theory [2], discusses the Gaussian distribution extensively. In that discussion he demonstrates that when using a Gaussian distribution to estimate a location parameter, the only properties of the noise that enter the calculation are the first and second moments of the true noise; the underlying ensemble sampling distribution for the noise, the distribution from which the noise was actually sampled, completely cancels out of the inference. Exactly the same inferences will be made regardless of the underlying ensemble sampling distribution for the noise, provided the first and second moments of the true noise are the same. In this paper we review Jaynes' calculation, demonstrate it explicitly, generalize it to more complex models, and show that for regression style models the underlying ensemble sampling distribution for the noise is irrelevant to our inferences provided the mean and mean-square projection of the model onto the true noise are the same.

1. Introduction

The problem of assigning the joint probability for the noise, effectively the likelihood, has plagued probability theory since the time of Laplace. The outstanding question has always been, what probability density function should be assigned? It seems intuitively obvious that if one knows the underlying ensemble sampling distribution for the noise then one should assign this function as the likelihood. However, this is typically not done; rather one usually assigns a Gaussian distribution for the likelihood, and then proceeds to perform the calculation. Some authors have noted that this cavalier approach is unusually effective in practice [6]. The problem with the use of a Gaussian distribution for the likelihood is simply why are the estimates so good? Jaynes, in his unfinished work on probability theory [2], gives an answer to this question: namely that the Gaussian does not represent the

[1] In Chapter 7 of Jaynes' unfinished work on probability theory [2], there is a section by this title. In that section Jaynes gives convincing reasons why the Gaussian represents a state of knowledge about the true noise.

underlying ensemble sampling distribution of the noise at all, rather it represents a particularly uninformed state of knowledge, a state of knowledge that accurately reflects what the experimenter knows about the true noise in data.

To understand Jaynes' conclusion and how to resolve the problem of assigning the joint probability for the noise, one needs to go back to first principles and to understand what we are actually doing when we apply probability theory. This discussion, Section 2, will be done using an example taken from Nuclear Magnetic Resonance (NMR), but the example is unimportant. Virtually any scientific experiment could be substituted for the NMR example and it would not change the discussion. The rules of probability theory will be applied just enough to reach the point where the joint probability for the noise must be supplied, and we will show that it is the probability for the *noise actually realized in our data sample* that must be assigned. In Section 3, the principal of maximum entropy is used to derive the Gaussian probability density function, and there it is shown that the only information this probability density function incorporates is knowledge of the first and second moments of the *actual noise in the sample at hand.* In Section 4, Jaynes' example of estimating a location parameter [2] will be used to show that a Gaussian distribution uses only the first and second moment of the actual noise in the data, and that all other characteristics of the noise have been made irrelevant to our inferences. At the end of this section, we illustrate the near-irrelevance of ensemble sampling distributions for the noise with a numerical example. In Section 5, we will generalize the calculations to much more complex models, and show that for regression style models the ensemble sampling distribution for the noise is irrelevant provided the mean-square of the actual noise and the mean projection of the model onto the actual noise remains the same across different noise samples. When models contain nonlinear parameters we show that for any fixed value of the nonlinear parameters an additional condition is needed for invariance of the estimated location parameters: that the mean-square of the data must be the same across noise samples. If these conditions are met, then again, the underlying ensemble sampling distribution for the noise will cancel from the calculation, and we shall obtain the same results regardless of the underlying ensemble sampling distribution for the noise. At the end of this section, we again illustrate the calculation with a numerical example.

2. What Are We Doing When We Use Probability Theory?

When we are solving a problem using probability theory there is always some hypothesis about which we wish to make inferences. This hypothesis might be as simple as estimating a location parameter, or as complicated as determining the number of sinusoidal signals in NMR data. Hypotheses of this nature contain two implicit assumptions: we have some facts, usually data, and these facts can be used to measure the quantity of interest, *i.e.*, they are relevant. Measuring is the process of relating some aspect of reality to an appropriate standard. Mathematics is the science that quantifies the measuring process, *i.e.*, mathematics is the science of measurement. Parameter estimation using probability theory is the science of measurement when the facts do not uniquely determine the measurement.

As noted, measuring a quantity presupposes we have some data, and a mathematical relationship between these data and the quantity to be measured. *Without this mathematical relationship there can be no quantification or estimation.* This mathematical relationship is the process by which the standard of measurement is brought into the problem. The data are simply the facts on which the measurement is based. The relationship between the data and the measured quantity, the model, might include many different parameters about which we have no interest. These parameters are nonetheless part of the problem. For example, when estimating the number of resonances in an NMR experiment, the data, d_i, and number of resonances, $\hat{m}$, are related by

$$d_i = \sum_{j=1}^{\hat{m}} \hat{A}_j \cos(\hat{\omega}_j t_i + \hat{\theta}) \exp\{-\hat{\alpha}_j t_i\} + n_i, \tag{1}$$

where $\hat{A}_j$, $\hat{\omega}_j$, $\hat{\alpha}_j$ are the true amplitude, frequency, and decay rate constant of the jth resonance. The common phase is represented by $\hat{\theta}$, and n_i is the *the actual* noise in the ith data value. In this paper, the word "noise" will be used to refer to the true noise in a data sample, while the word "error" will be used to refer to a running index, a hypotheses about which we must infer, that appears in our probability theory calculations. This model, Eq. (1), contains many unknown quantities. In the estimation process these unknown quantities become hypotheses of the form "the amplitude of the jth sinusoid was A_j," where A_j indexes an entire series of hypotheses about which we must infer. In this example we designated the hypotheses as A_j and not $\hat{A}_j$ to distinguish the running index, A_j, from the true but unknown value of the amplitude $\hat{A}_j$. This distinction will become increasingly important in the calculations that follow. In estimating the number of resonances m all of the parameters, $A_j, \omega_j, \alpha_j, \theta$, are uninteresting and must eliminated from the probability for m using the rules for probability theory.

Note that in Eq. (1) we have assumed additive noise. This additivity assumption is not necessary. For example, the noise might be multiplicative, or it might contaminate the data in some other more complex way. The exact functional form is not important, what is important is that the functional relationship can be supposed known. If the exact way the noise enter the calculation is unknown, then many different assumptions about how the noise enter may be postulated and tested using probability theory. Each different assumption would constitute a different model or hypothesis that is to be tested.

Before we apply the rules of probability theory to determine what calculation is to be performed, we must state exactly what is known about the various parameters appearing in the model. We know the data, $D \equiv \{d_1 \cdots d_N\}$, and we know something about the amplitudes, the frequencies, decay rate constants and phase of the sinusoids. This is obviously true because we are analyzing a real physical experiment. The data did not come into existence in a vacuum. Indeed in NMR, the signals have certain general strengths that are known in advance – if this were not so then the spectrometer could not have been designed. Additionally, the sinusoids were sampled at some rate; this presupposes that the frequencies could not have been higher than a certain value, else we would have digitized them incorrectly. We also know something about the overall decay rate constants because the signal was sampled for only a finite amount of time, and this presupposes that we sampled the signal over the time interval that the signal was large. Indeed if the signal sampling rate or the signal acquisition time were inappropriate, the experimenter would simply reset these parameters and reacquire the data.

Last, the size of the noise is known at least approximately because a digitizer was used to acquire the data and the physical characteristics of the digitizer prohibit it from responding to signals smaller than a certain known value; nor can it sample signals larger than some known value because these signals would overflow the digitizer. If either of these conditions were to occur, the experimenter would detect this condition and correct it before any data were actually saved. So the general size of the noise is also known. We are going to designate the information about the noise as I_σ and any parameters associated with this information as $\{\sigma\}$, although, at this time, we are not going to specify either the information, I_σ, or the parameters, $\{\sigma\}$. The other general background information, *i.e.*, information about the frequencies, decay rate constants, amplitudes, etc., will be designated as I.

Next, we are going to apply the rules of probability and see where they lead us. We have supposed the parameter of interest to be the number of sinusoids m, all others being nuisance parameters. The probability for the number of parameters may be computed from the joint probability for all of the parameters:

$$P(m|DI_\sigma I) = \int d\{A\}d\{\omega\}d\{\alpha\}d\theta d\{\sigma\}P(\{A\}\{\omega\}\{\alpha\}\theta\{\sigma\}m|DI_\sigma I), \qquad (2)$$

where we have removed all of the uninteresting or nuisance parameters using the sum rule of probability theory. The notation $\{x\}$ means the collection of all of the parameters of the x type, so $\{A\} \equiv \{A_1 \cdots A_m\}$. Bayes' theorem [3] may be used to factor the joint probability for the parameters, to obtain

$$\begin{aligned} P(m|DI_\sigma I) \quad \propto \quad & \int d\{A\}d\{\omega\}d\{\alpha\}d\theta d\{\sigma\} \\ & \times\ P(\{A\}\{\omega\}\{\alpha\}\theta\{\sigma\}m|I_\sigma I) \\ & \times\ P(D|\{A\}\{\omega\}\{\alpha\}\theta\{\sigma\}mI_\sigma I), \end{aligned} \qquad (3)$$

where we have dropped the normalization constant. It is at this point that one must look carefully at this equation to see exactly what including the prior information about the noise has accomplished.

In probability theory as logic, probabilities are indicated 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. The notation $P(X)$ is meaningless because the information on which the truth of the hypothesis X is to be assessed has not been specified. With this in mind, look at $P(D|\{A\}\{\omega\}\{\alpha\}\theta\{\sigma\}mI_\sigma I)$. This term is the direct probability for the data given the truth of all of the parameters in the model. But if the true parameters are given, then

$$d_i - \sum_{j=1}^{m} A_j \cos(\omega_j t_i + \theta) \exp\{-\alpha_j t_i\} = e_i, \tag{4}$$

where, as noted above, we have introduced a notation without the hats to indicate that this equation is using the given values. Similarly, we have adopted the notation e_i for the given error value. This error value is a hypotheses about which we must infer. This hypotheses is of the form "the true noise value was e_i," where e_i is a running index and its numerical value would range over all valid noise amplitudes. Equation (4) 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:

$$\begin{aligned} P(D|\{A\}\{\omega\}\{\alpha\}\theta\{\sigma\}mI_\sigma I) &= \int de_1 \cdots de_N P(D\{e\}|\{A\}\{\omega\}\{\alpha\}\theta\{\sigma\}mI_\sigma I) \\ &= \int de_1 \cdots de_N P(D|\{e\}\{A\}\{\omega\}\{\alpha\}\theta mI) \\ &\quad \times P(e_1 \cdots e_N|\{\sigma\}I_\sigma), \end{aligned} \tag{5}$$

where $P(D|\{e\}\{A\}\{\omega\}\{\alpha\}\theta mI)$, the direct probability for the data given the errors and the parameters, is a delta function derived from Eq. (4). In this probability we dropped the dependence on the noise prior information and parameters because knowing the true error values, *i.e.*, knowing the noise values, determines things like the standard deviation for the noise and so renders knowledge of any statistics about the noise redundant. The last term, $P(e_1 \cdots e_N|\{\sigma\}I_\sigma)$, is the joint probability for the *actual noise in the data* given the noise parameters $\{\sigma\}$ and the information about the noise I_σ. In this probability we dropped the dependence on the model parameters because by assumption I_σ was all of the information about the noise. Last, the dependence on whether the noise was additive, multiplicative, or entered in some other more complex way was removed from this joint probability when we applied the rules of probability theory in Eq. (5). Indeed the joint probability for the noise, $P(e_1 \cdots e_N|\{\sigma\}I_\sigma)$, only depends on the existence of a model equation, the explicit inclusion of the prior information about the noise values, and the application of the rules of probability theory.

3. Assigning Distributions Using The Principle Of Maximum Entropy

The problem of assigning the direct probability for the data, the likelihood, has been reduced to the problem of assigning the joint probability for the actual noise in the data given whatever we know about the noise. The question that must be faced is how does one assign a joint probability density function that relates to the noise sample actually realized in the data? The easiest way to do this is state explicitly what is known about the noise and then use the principle of maximum entropy to assign this probability density function. Here, we will briefly review the principal of maximum entropy and then show in the following sections that probability density functions assigned using the principal of maximum entropy use only the information that is actually put into the assignment of these probabilities; and that, as a consequence, all noise samples having the same mean-square and the same projection onto the model will give identical location parameter estimates, irrespective of the underlying ensemble sampling distribution for which the noise sample was drawn.

Suppose one must assign a probability distribution for the ith value of a discrete parameter given the prior information I. This probability is denoted $P(i|I)$ $(1 \leq i \leq m)$. The Shannon entropy, defined as

$$H \equiv -\sum_{i=1}^{m} P(i|I) \log P(i|I), \tag{6}$$

is a measure of the amount of ignorance (uncertainty) in this probability distribution [4]. 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 a probability distribution, $P(i|I)$, that contains only the information I by maximizing H subject to the constraints represented by I. Because H measures the amount of ignorance in this 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: *i.e.*, $P(i|I)$ contains only the information I [4,2].

For the purposes of this discussion, we are going to pass to the limit of continuous parameters and ignore any subtleties that concern the use of probability density functions. As long as one maintains finite, normalizable probability density functions there are almost no circumstances where the use of continuous parameters will cause problems. When we do this we will assume that the appropriate prior probability for a location parameter is a uniform prior probability and we will use this uniform prior probability as the measure function in the entropy [5].

The probability that must be assigned is the joint probability for noise given that one knows the information I_σ and the parameters associated with this information $\{\sigma\}$. This probability, $P(e_1 \cdots e_N|\{\sigma\}I_\sigma)$, was derived in Eq. (5), where e_i is both a shorthand notation and an index: it stands for a hypothesis indexed by e_i of the form "the true value of the noise in the ith data value was e_i." As an index, e_i ranges over all valid values of the true noise; while the joint probability for the noise assigns a reasonable degree of belief to a particular set of hypotheses specified by the error indices.

To use the principle of maximum entropy to assign this probability there are two problems that must be solved. First, to use the principle of maximum entropy we must know what constraints we are going to use, and second we must know the numerical values of the parameters associated with these constraints. The second problem has been disposed of because when we included I_σ and $\{\sigma\}$ in the probability theory calculation we arrive at $P(e_1 \cdots e_N|\{\sigma\}I_\sigma)$ as the probability to be assigned. The parameters $\{\sigma\}$ and the prior information I_σ are given and so allow us to utilize the principle of maximum entropy. However, the first problem, has not yet been solved because we have not yet specified either I_σ or $\{\sigma\}$.

So the problem of assigning the joint probability for the noise has been reduced to specifying what is known about the noise, *i.e.,* specifying the constraints to be used in the calculation, and then applying the principal of maximum entropy. But what constraints should be used? As neither I_σ nor $\{\sigma\}$ have been specified, it would appear that we have an unlimited range of possible constraints. For example, should a constraint on correlations be included? If so, which of the many different types of correlations should be included? There are second order correlations of the form

$$\rho_s = \frac{1}{N-s}\sum_{i=1}^{N-s} n_i n_{i+s}, \tag{7}$$

where ρ_s is the known correlation coefficient and s is a measure of the correlation distance, as well as third, fourth, and higher order correlations. In addition to correlations, should a constraint on the moments of the noise be included? If so, on which moments should the joint probability for the noise depend? There are many different types of moments. There are power law moments of the form

$$\sigma^s = \frac{1}{N}\sum_{i=1}^{N} n_i^s, \tag{8}$$

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

Given that we have the information necessary to incorporate any or all of these constraints into the maximum entropy calculation, the question remains which constraints should be used? If there is information that suggests that the true noise is correlated, then a correlation constraint should be used. If there is information that suggests the higher moments can deviate significantly from what one would expect from constraints on the first few moments, then again a constraint on the higher moments should be included. But if all one knows is the general magnitude and scale of the noise, then *one is always better off to leave out constraints on higher moments and correlation coefficients*, 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.

This suggests that if we wish to be conservative, *i.e.,* we do not wish to make gratuitous assumptions about things we really do not know, and we do not wish to make excessive claims for the accuracy of our parameter estimates, then we should constrain the maximum entropy distribution as little as possible, while still reflecting what is actually known about the noise. If, for example, all one knows is something about the general magnitude of the noise, then a constraint on the absolute value of the errors would be appropriate; this would result in a Laplacian assignment for the joint probability for the noise. Additionally, if one also knows something about the general scale of the noise, for example they fluctuate around zero with some standard deviation, then, as we shall see, this would result in a Gaussian assignment. Both of these assignments would give very reasonable results.

The main difference between them is really one of functionality. Because of the many properties of the Gaussian [2], the Gaussian is easier and more convenient to use than the Laplacian; although with the advent of modern computers, the difficulties associated with the use of the Laplacian are disappearing.

In the next few paragraphs the principle of maximum entropy will be used to derive the joint probability for the noise. In this derivation we will use the first and second moments of the true noise as constraints in the maximum entropy calculation. Earlier, we gave a definition of the power law moments, Eq. (8), using the true noise. In the problem we now face we do not know the true noise and so the calculation shown in Eq. (8) cannot be used; all we know is that the joint probability for the noise is to be consistent with the given power law moments. So what does it mean for a probability density to be consistent with power law moments? The answer is simple; for the joint probability for the noise to be consistent with power law moments it must have the property that

$$\sigma^s = \langle e^s \rangle \equiv \frac{1}{N}\sum_{i=1}^{N}\int de_1 \cdots de_N \; e_i^s P(e_1 \cdots e_N|\{\sigma\}I_\sigma) \tag{9}$$

where σ^s is the known value of a particular moment, and the notation $\langle e^s \rangle$ means: compute the average of the expected value of the e_i^s. Consistent means that the expected moments of the joint probability density for the noise should be equal to the observed moments of the true noise, *i.e.,* the probabilities we assign should reflect what is actually known about the noise.

We are going to place constraints on the first two moments of the joint probability for the noise. If we designate the mean of the noise as μ and the standard deviation as σ, then the constraint on the first moment is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\int de_1 \cdots de_N \; e_i P(e_1 \cdots e_N|\{\sigma\}I_\sigma), \tag{10}$$

and the constraint on the second moment is given by

$$\sigma^2 + \mu^2 = \frac{1}{N}\sum_{i=1}^{N}\int de_1 \cdots de_N \; e_i^2 P(e_1 \cdots e_N|\{\sigma\}I_\sigma), \tag{11}$$

where the second moment has been written as $\sigma^2 + \mu^2$ to make the resulting Gaussian come out in standard notation.

We seek the joint probability density function for the noise that has highest entropy for the given sample mean and standard deviation. To find this distribution Eqs. (10) and (11) are rewritten so they sum to zero:

$$\mu - \frac{1}{N}\sum_{i=1}^{N}\int de_1 \cdots de_N \; e_i P(e_1 \cdots e_N|\{\sigma\}I_\sigma) = 0, \tag{12}$$

and

$$\sigma^2 + \mu^2 - \frac{1}{N}\sum_{i=1}^{N}\int de_1 \cdots de_N \; e_i^2 P(e_1 \cdots e_N|\{\sigma\}I_\sigma) = 0. \tag{13}$$

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

$$1 - \int de_1 \cdots de_N P(e_1 \cdots e_N | \{\sigma\} I_\sigma) = 0. \tag{14}$$

Because Eqs. (12)–(14) are each zero, each may be multiplied by a constant and added to the entropy of the joint probability density function without changing the entropy of that distribution. One obtains

$$\begin{aligned} H \; = \; & -\int de_1 \cdots de_N P(e_1 \cdots e_N | \{\sigma\} I_\sigma) \log P(e_1 \cdots e_N | \{\sigma\} I_\sigma) \\ & + \beta \left[1 - \int de_1 \cdots de_N P(e_1 \cdots e_N | \{\sigma\} I_\sigma) \right] \\ & + \delta \left[\mu - \frac{1}{N} \sum_{i=1}^{N} \int de_1 \cdots de_N \; e_i P(e_1 \cdots e_N | \{\sigma\} I_\sigma) \right] \\ & + \lambda \left[\sigma^2 + \mu^2 - \frac{1}{N} \sum_{i=1}^{N} \int de_1 \cdots de_N \; e_i^2 P(e_1 \cdots e_N | \{\sigma\} I_\sigma) \right], \end{aligned} \tag{15}$$

where β, δ, and λ are called Lagrange multipliers.

To find the distribution that maximum entropy will assign for the joint probability for the noise, this expression is maximized with respect to $P(e'_1 \cdots e'_N | \{\sigma\} I_\sigma)$. After some algebra, one obtains

$$P(e_1 \cdots e_N | \mu\sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\sum_{i=1}^{N} \frac{(e_i - \mu)^2}{2\sigma^2} \right\}, \tag{16}$$

where

$$\lambda = \frac{N}{2\sigma^2}, \qquad \delta = -\frac{N\mu}{\sigma^2}, \quad \text{and} \quad \beta = \frac{N}{2}\left[\log(2\pi\sigma^2) + \frac{\mu^2}{\sigma^2} \right] - 1, \tag{17}$$

and I_σ and $\{\sigma\}$ have been replaced by μ and σ in $P(e_1 \cdots e_N | \{\sigma\} I_\sigma)$.

There are several interesting points to note about this probability density function. First, this is a Gaussian distribution. However, the fact that the joint probability for the noise has been assigned to be a Gaussian makes no statement about the ensemble sampling distribution of the noise; rather, it says only that for the given mean and mean-square noise value, the joint probability density function for the noise should be maximally uninformative and that maximally uninformative distribution happens to be a Gaussian. Second, the joint probability for the noise 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 the joint probability for the noise 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 moments of this Gaussian, one obtains

$$\langle e^s \rangle = \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \sigma^{2s} \frac{\partial^s}{\partial \mu^s} \exp\left\{\frac{\mu^2}{2\sigma^2}\right\} \quad (s \geq 0) \tag{18}$$

which reduces to

$$\langle e^0 \rangle = 1, \qquad \langle e^1 \rangle = \mu, \quad \text{and} \quad \langle e^2 \rangle = \sigma^2 + \mu^2 \tag{19}$$

for $s = 0$, $s = 1$, and $s = 2$, just the given mean and mean-square noise values used to assign the probability density function. So as noted previously, the maximum entropy distributions makes no other assumptions about the noise above those already implied by the constraints. Fourth, for a given value of the sample mean and mean-square the joint probability for the noise has highest entropy. Consequently, when any operation is performed on an arbitrary distribution that the preserves mean and mean-square while discarding other information, *i.e.,* the operation increases the entropy of the distribution, then that distribution necessarily will move closer and closer to a Gaussian distribution regardless of the initial assignment. The Central Limit Theorem is one special case of this phenomenon – see Jaynes [2].

4. The Jaynes Example – Estimating A Location Parameter

Maximum entropy distributions are the only distributions that have sufficient statistics. Sufficient statistics are functions of the data, and therefore the noise, that summarize all of the information in the data relevant to estimating a location parameter. We would like to demonstrate how the sufficient statistics render the underlying ensemble sampling distribution for the noise irrelevant. We will illustrate this using a Gaussian distribution for the joint probability for the noise with the understanding that *any* maximum entropy distribution will have similar properties.

Suppose the true value of a location parameter is μ_0 and one has a data set D with data elements d_i such that

$$d_i = \mu_0 + n_i. \tag{20}$$

Now because μ_0 is the true value of the mean, the n_i are the true noise samples. The problem we face is to estimate μ_0. The hypothesis about which inferences are to be made is of the form "the true value of the location parameter is μ given the data D." The parameter μ is an index that specifies the various hypotheses; while the probability $P(\mu|D\sigma I_\sigma I)$ assigns a reasonable degree of belief to these hypotheses. Following what was done in Section 2, the posterior probability for μ is given by

$$P(\mu|\sigma D I_\sigma I) = P(\mu|I) P(D|\sigma\mu I_\sigma I) \tag{21}$$

where we have assumed that the value of the noise standard deviation, σ, is known for now. $P(\mu|I)$ is the prior probability for μ, and $P(D|\mu\sigma I_\sigma I)$ is the direct probability for the data given the model parameters. Because we are interested in the consequences of assigning a Gaussian distribution for the joint probability for the noise, we will take the prior for μ to be a constant.

In this example we are going to assume that the noise could take on both positive and negative values, and that when assigning the Gaussian we had no prior information that would lead us to favor either a positive or negative noise samples. Consequently, we will assign the joint probability for the noise to be a zero-mean Gaussian, namely

$$P(e_1 \cdots e_N|\sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{e_i^2}{2\sigma^2}\right\}. \tag{22}$$

From which one obtains

$$P(\mu|D\sigma I_\sigma I) \propto P(D|\mu\sigma I_\sigma I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right\} \tag{23}$$

as the posterior probability for μ. This equation may be rewritten as

$$P(\mu|D\sigma I_\sigma I) \propto (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{N}{2\sigma^2}\left[(\overline{d}-\mu)^2+s^2\right]\right\} \tag{24}$$

where the sufficient statistics are $\overline{d}$, the mean data value, and s^2 is the mean-square data value less the squared mean-data value, and these are just the constraints used in deriving the Gaussian distribution. Using Eq. (20), the mean data value is given by

$$\overline{d} \equiv \frac{1}{N}\sum_{i=1}^{N} d_i \tag{25}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\mu_0+n_i) \tag{26}$$

$$= \mu_0+\overline{n}, \tag{27}$$

with

$$\overline{n} \equiv \frac{1}{N}\sum_{i=1}^{N} n_i. \tag{28}$$

So the mean data value is equal to the true mean μ_0 plus the mean noise value $\overline{n}$. Similarly, s^2 is given by:

$$s^2 \equiv \overline{d^2}-(\overline{d})^2 \tag{29}$$

$$= \frac{1}{N}\sum_{i=1}^{N} d_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} d_i\right)^2 \tag{30}$$

$$= \frac{1}{N}\sum_{i=1}^{N} (\mu_0 + n_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} [\mu_0 + n_i]\right)^2 \tag{31}$$

$$= \overline{n^2} - (\overline{n})^2, \tag{32}$$

with

$$\overline{n^2} = \frac{1}{N}\sum_{i=1}^{N} n_i^2 \tag{33}$$

where $\overline{n^2}$ is the mean-square noise value and $(\overline{n})^2$ is the square of the mean-noise value. Therefore, s^2 is a function of only the true noise.

Using the mean ± standard deviation as an estimate of μ one obtains

$$(\mu)_{est} = \begin{cases} \overline{d} \pm \sigma/\sqrt{N} & \sigma \text{ known} \\ \overline{d} \pm s/\sqrt{N-3} & \sigma \text{ unknown} \end{cases} . \tag{34}$$

The actual error, Δ, is given by

$$\Delta = \overline{d} - \mu_0 = \overline{n} \tag{35}$$

which depends only on the *mean of the true 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 true noise when the standard deviation of the noise is not known. The underlying ensemble sampling distribution for the noise has canceled out of the calculation, and the only property of the noise that survives is the actual mean and mean-square of the true noise. *All other properties of the noise are irrelevant!* Exactly the same parameter estimates will result if the underlying ensemble sampling distribution of the noise changes, provided the mean and mean-square of the new noise sample is the same as in the previous sample – just the properties needed to represent what is actually known about the noise and to render what is *not* known about them irrelevant.

From this discussion it is clear that the Gaussian represents a particularly uninformed state of knowledge. And that the underlying ensemble sampling distribution for the noise has been made irrelevant in the sense that any set of noise drawn from any ensemble sampling distribution will result in the exact same posterior probability being computed, provided the mean and mean-square noise values for the different noise samples are same. However, we would like to explicitly demonstrate this using numerical examples. According to the results that have been derived so far, the only properties of the true noise that enter the calculations are the mean and mean-square. Consequently, if we analyze a new data set containing a new noise sample generated from an entirely different ensemble sampling distribution while preserving these two quantities, the resulting probability density functions will be identical.

To illustrate this point we have prepared three data sets. These are shown in Fig. 1. Each of these three data sets contain the same constant signal, with noise having the same mean and mean-square. However, the noise in each data set were generated from three very different ensemble sampling distributions. The first data set, panel (A), contains Gaussian white noise. The second, panel (B), was generated using a Gaussian random number generator, but the results from the random number generator were rounded to integer values. Last, the third data set, panel (C), does not contain noise at all; rather, this data set is a constant plus a deterministic sinusoidal signal that is not in our model. We could have ordered the sinusoid randomly to camouflage this unmodeled signal, but by making it obvious that we have an unaccounted for signal it will better make the point that everything that has been said in the paper is equally applicable to models that only account for part of the deterministic signal present in the data. Indeed the noise itself is a deterministic signal which we have not modeled in any regular fashion, so it should not be surprising that if our estimates are invariant when we change one deterministic signal, the noise, they would also be invariant when we change another deterministic signal, in this case an unmodeled sinusoid, provided the mean and mean-square of these residual components are the same. This last point will be illustrated with Panel (C).

To ensure that each of these data sets has exactly the same mean and mean-square, we first generated the three set of numbers that we are calling noise in Fig. 1. For each set we then computed the average "noise" value, and subtracted this average from each noise value. We then computed the mean-square noise value and used that mean-square value to scale each noise value within that set so that each noise set had the same second moment. Finally, we added the constant signal to each noise set to obtain the data shown in Fig. 1.

Next, the posterior probability for μ was computed for each of these three data sets and is displayed in Fig. 2. In the three panels, the solid line is $P(\mu|\sigma DI)$, and the dotted line is $P(\mu|DI)$. On the scale shown, these probability density functions are identical. If one examines the actual numbers used to generate these plots they are indeed identical. Please note that because we know that the noise in panel (B) were generated from a Gaussian random number generator that was rounded to the nearest integer it would always be possible to include this information in the calculation, and so make a better estimate of the parameter μ. Similarly, one look at the data in Fig. 1(C) and one would know that there is a signal present in the data that the model does not account for. If we included that sinusoid in the model we would be able to estimate the value of the location parameter down to the rounding errors of the procedure used to generate the data. However, the point is not that we can do better, the point is that these three noise samples, sampled from three very different ensemble sampling distributions, give identical estimates for the location parameter. The underlying ensemble sampling distribution for the noise is simply not relevant to the questions being answered by probability theory. And it does not matter if that ensemble sampling distribution is Gaussian white noise, unaccounted for signal, or any other type of noise one can imagine. The only properties of the noise utilized by probability theory were the mean and mean-square of the actual noise in our data.

Figure: 1 Three Data Sets Having The Same Mean and Mean-Square Noise

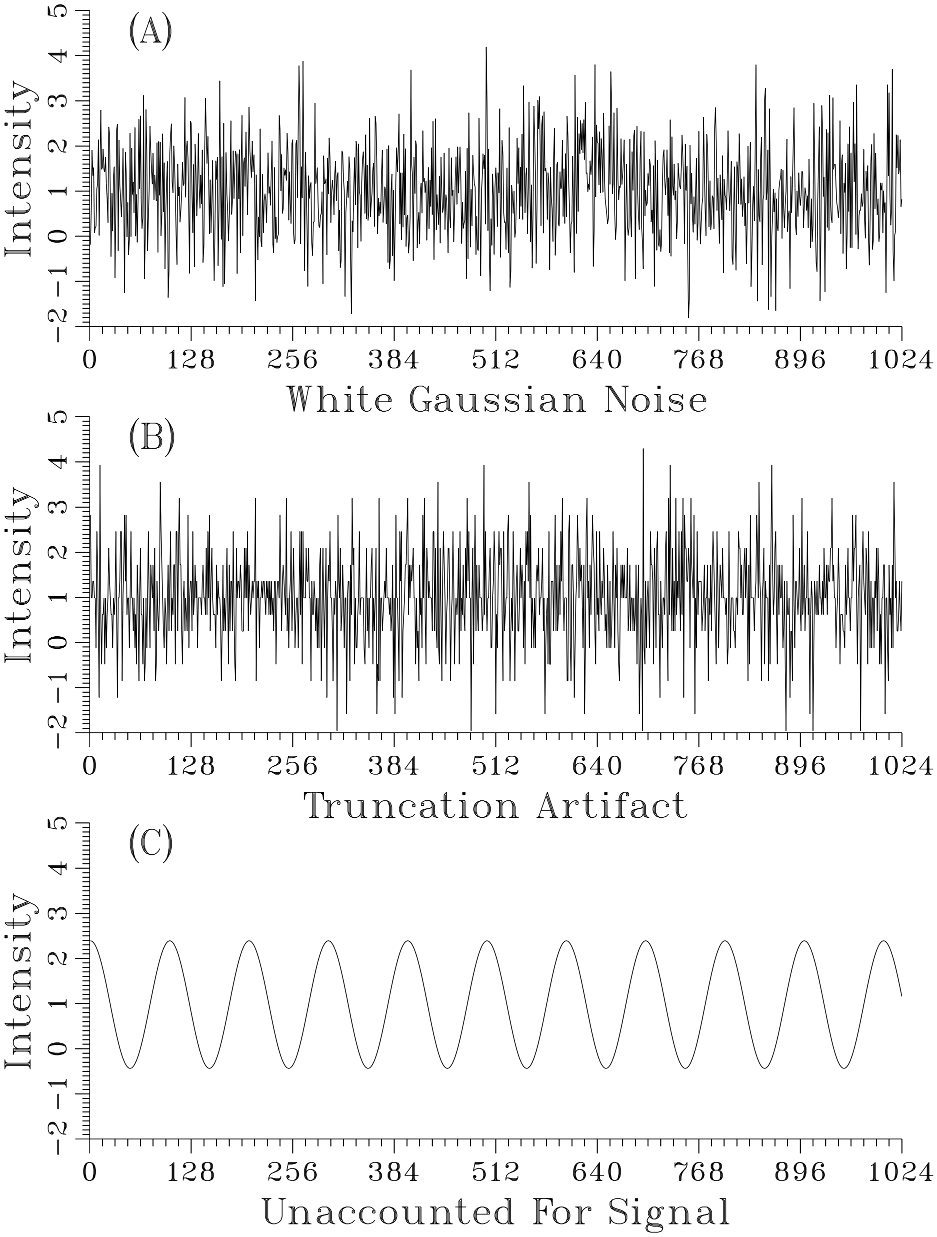

Fig. 1. Panel (A) is a constant plus Gaussian white noise, (B) a constant plus truncated Gaussian white noise, and (C) a constant plus a sinusoid. If you subtract the constant signal from each of these three data sets, the residuals have the same mean and mean-square. According to probability theory these three data sets should produce identical posterior probabilities for μ.

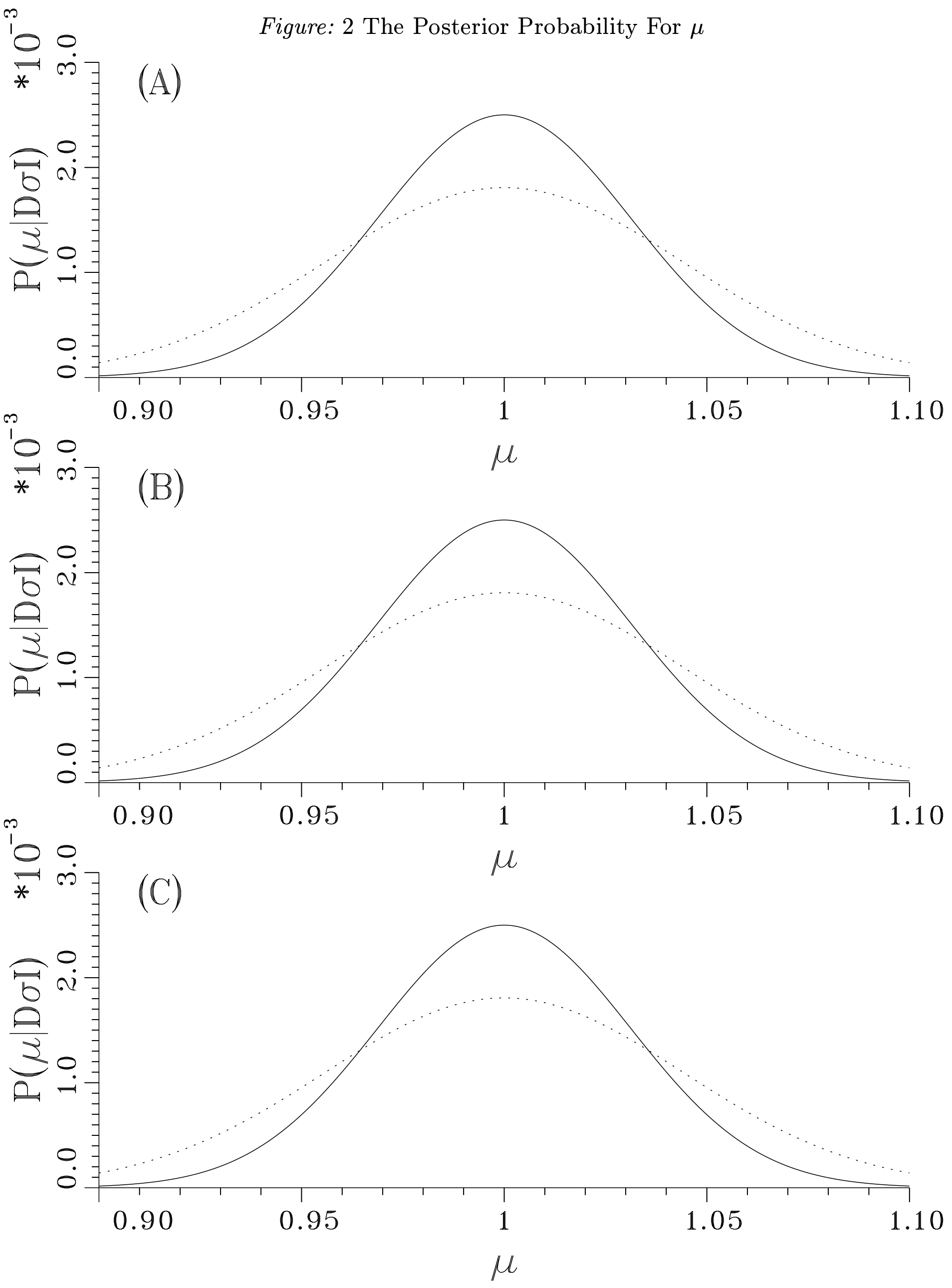

Fig. 2. The three panels are the posterior probabilities $P(\mu|\sigma DI)$, solid lines, and $P(\mu|DI)$, dotted lines, computed for the three data sets shown in Fig. 1. According to probability theory these three estimates for μ should be identical.

5. A General Nonlinear Model

The calculations given in the previous section are valid and correct as far as they go. They illustrate the near-irrelevance of ensemble sampling distributions for a particularly simple model. The question we would like to address in this section is: what happens when we have more complicated models? Do we still get the same invariance when the ensemble sampling distribution for the noise changes, and if not, how must we modify what was said in the previous section to account for these more complex models? In particular, what happens when the models contain both location and scale parameters?

The most general model that will be addressed is of the form [1]:

$$d_i = \sum_{j=1}^{m} \hat{B}_j G_j(\{\hat{\theta}\}, t_i) + n_i, \tag{36}$$

where $\hat{B}_j$ is the true amplitude of the jth model function $G_j(\{\hat{\theta}\}, t_i)$, and the $\{\hat{\theta}\}$ are the true values of a collection of parameters that appear in the model in a nonlinear fashion. Each model function is assumed to be evaluated at the same time that the data were taken. We have labeled this time as t_i, but whether this is time, space, or some other labeling is irrelevant. Additionally, we have indicated the data are one dimensional; however, whether or not the data are multidimensional is also irrelevant. Multidimensional data may always be labeled with a single index by an appropriate change of the indices. Last, n_i represents the true noise in the measurement.

What we have to say about the model is much easier to understand if we introduce an orthogonal model by the following change of parameters and functions:

$$\hat{B}_k = \sum_{j=1}^{m} \frac{\hat{A}_j \xi_{jk}}{\sqrt{\lambda_j}}, \tag{37}$$

where λ_j is the jth eigenvalue of the g_{jk} matrix, Eq. (38) below, and ξ_{jk} is the kth component of the jth eigenvector of the matrix

$$g_{jk} = \sum_{i=1}^{N} G_j(\{\hat{\theta}\}, t_i) G_k(\{\hat{\theta}\}, t_i). \tag{38}$$

In this new notation the model equation becomes

$$d_i = \sum_{j=1}^{m} \hat{A}_j H_j(\{\hat{\theta}\}, t_i) + n_i \tag{39}$$

with

$$H_j(\{\hat{\theta}\}, t_i) = \frac{1}{\sqrt{\lambda_j}} \sum_{k=1}^{m} \xi_{jk} G_k(\{\hat{\theta}\}, t_i), \tag{40}$$

and

$$\hat{A}_k = \sqrt{\lambda_k} \sum_{j=1}^{m} \hat{B}_j \xi_{kj}. \tag{41}$$

We call this model "orthogonal" because the model functions, $H_j(\{\hat{\theta}\}, t_i)$, have the property that

$$\sum_{i=1}^{N} H_j(\{\hat{\theta}\}, t_i) H_k(\{\hat{\theta}\}, t_i) = \delta_{jk} \tag{42}$$

where δ_{jk} is zero if $j \neq k$, and one otherwise. Note that introducing this change of variables and change of functions has not changed the form of the model function, *i.e.*, model Eq. (36) is formally identical to Eq. (39). Additionally, the transformation did not effect either the data, d_i, or the noise n_i. These are same in both equations.

We are going to repeat the calculations done in the previous section using this orthogonal model. The aim is to understand the effects of introducing a probability that has sufficient statistics and in particular we will concentrate on the Gaussian distribution. The object of this calculation is to understand the conditions under which the parameter estimates obtained using a Gaussian distribution are invariant when the underlying ensemble sampling distribution for the noise changes. To do this we must carefully differentiate between the true parameter values and the given parameter values in our calculations. For the amplitudes we will write A_j for the running index in our calculations. Similarly, $\{\theta\}$ will be the running index that specifies the collection of nonlinear parameters, and finally, e_i will be the given value of the error at time t_i.

In the first part of this calculation we are going to examine the estimates of the location parameters, the $\{A\}$. Consequently, the probability we will be interested in is $P(\{A\}|D\{\theta\}\sigma I)$ where, for the time being, we are going to assume that the nonlinear parameters are given. Last, because we are interested in the consequences of assigning a Gaussian for the joint probability for the noise we are going to assign a uniform prior probability for the amplitudes. With these assumptions and using Eq. (39), the posterior probability for the amplitudes is given by:

$$P(\{A\}|D\{\theta\}\sigma I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{Q}{2\sigma^2}\right\} \tag{43}$$

where Q is defined as

$$Q \equiv \sum_{j=1}^{m} \hat{A}_j^2 + 2N \sum_{j=1}^{m} \hat{A}_j \overline{H_j(\{\hat{\theta}\})n} + N\overline{n^2} - m\overline{h^2} + \sum_{j=1}^{m} (A_j - h_j)^2 \tag{44}$$

with

$$\begin{aligned} \overline{h^2} &= \frac{1}{m} \sum_{j=1}^{m} h_j^2, & (45) \\ h_j &= N\left(\overline{H_j(\{\theta\})n} + \sum_{k=1}^{m} \hat{A}_k \overline{H_k(\{\hat{\theta}\})H_j(\{\theta\})}\right), & (46) \\ \overline{H_k(\{\hat{\theta}\})H_j(\{\theta\})} &= \frac{1}{N} \sum_{i=1}^{N} H_k(\{\hat{\theta}\}, t_i) H_j(\{\theta\}, t_i), & (47) \end{aligned}$$

and

$$\overline{H_j(\cdot)n} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} H_j(\cdot, t_i)n_i \tag{48}$$

where we have left the argument of $\overline{H_j(\cdot)n}$ unspecified because of the two different uses in the above equations.

First, let us examine the estimates for the amplitudes. If we consider the case when the nonlinear parameters are given, then the probability for the amplitudes is given by

$$P(\{A\}|D\{\theta\}\sigma I) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}(A_j - h_j)^2\right\} \tag{49}$$

and the expected value for each amplitude is

$$\langle A_j\rangle = h_j = N\left(\overline{H_j(\{\theta\})n} + \sum_{k=1}^{m}\hat{A}_k\overline{H_k(\{\hat{\theta}\})H_j(\{\theta\})}\right). \tag{50}$$

The expected amplitude is just the projection of the orthogonal model function onto the data. However, the data consists of two terms: the signal and the noise. The total projection of model $H_j(\{\hat{\theta}\})$ onto the signal, the second term in Eq. (50), does not change if the ensemble sampling distribution for the noise changes. The only term that changes is the total projection of the model onto the noise, the first term in Eq. (50). Consequently, for the amplitude estimates to remain invariant when the ensemble sampling distribution for the noise changes, the projections $\overline{H_j(\{\theta\})n}$ must be the same from one noise sample to another.

Think of the data as an N dimensional vector space. The model constitutes an m dimensional subspace of the full vector space. The estimated amplitudes are invariant under change in ensemble sampling distribution for the noise if *the projection of the noise onto the m dimensional subspace is the same.* In the example given by Jaynes, there is only a single constant model function, $H_1 = 1/N$, and the projection of this model onto the noise is the mean noise value, so the estimated value of μ is invariant under change in noise sample if the mean value of the noise is the same between the two samples, just the condition derived earlier.

The posterior probability for the amplitudes was derived given the nonlinear parameters, including the variance of the noise. From this posterior distribution the expected values of the amplitudes were then computed. These expected amplitudes do not change if the standard deviation for the noise is not known. However, there is a complication that must be considered when σ is unknown. Probability theory will naturally lead one to use Student's t-distribution to estimate the amplitudes. If Student's t-distribution is used our uncertainty in the estimated amplitudes will be related to the total squared-data value. But the total squared-data value depends on the projection of the true signal onto the true noise, Eq. (44), and if the ensemble sampling distribution for the noise changes this total projection could change. Consequently, changing the total squared-data value would result in a change in our estimate of how uncertain we are of the amplitudes. To ensure this does not happen, there are two additional conditions that must be met: the mean-square of the true noise, and the projection of the signal onto the noise must remain the same when the ensemble sampling distribution for the noise changes. This statement is equivalent to saying that the total squared-data value must be the same, if the estimated amplitudes are to remain the same when the ensemble sampling distribution for the noise changes.

Before we discuss what is happening with the nonlinear parameters we will illustrate the near-irrelevance of ensemble sampling distributions for these more complex models with a numerical example. The exact functional form of the model used in this example is irrelevant provided the model is of the form shown in Eq. (36). In this example we will use

$$\text{Signal} = B_1 + B_2 t + B_3 \cos(\omega t)\exp(-\beta t) \tag{51}$$

with $B_1 = -123$, $B_2 = 1$, $B_3 = 100$, $\beta = 4/N$, $\omega = 1$, $N = 256$ and dimensionless time units are used, *i.e.*, $t_i = 0, 1, \cdots, N-1$. This signal is shown in Fig. 3 The signal is an exponentially decaying sinusoid that is decaying about a line.

To illustrate that the estimates for the amplitudes are invariant when the sampling distribution for the noise changes, provided we have the same projection of the model function onto the true noise, we have prepared three noise samples – see Fig. 4. In Fig. 4 panel (A) is Gaussian white noise, (B) rounded Gaussian noise, and (C) an unmodeled signal. To prepare these noise samples, we first fixed the nonlinear parameters – here these are ω and β. For this demonstration, we fixed these parameters to their true values. We did this simply to avoid the additional complication of ensuring that the projection of the true signal onto the noise was invariant across the noise samples. After fixing the nonlinear parameters, we computed the orthogonal models and then computed the projection of the orthogonal models onto the noise. These projections times the corresponding model vector were then subtracted from each noise set. This has the effect of making the noise orthogonal to the model. Last, we took the original projections from the first noise set, the Gaussian white noise, and added them to each of the three noise sets. This ensured that the projection of the orthogonal model onto each of the noise samples is exactly the same.

The posterior probability, $P(B_3|DI)$, was computed for the signal shown in Fig. 3 plus the noise shown in the three panels in Fig. 4. These three computed probabilities are shown in Fig. 5. We could have illustrated the invariance of the amplitude estimates using any of the three amplitudes in the model. However, we choose to show only the posterior probability for the amplitude of the sinusoid. Note that on the scale of the plots shown in Fig. 5, the three probabilities are identical and a Unix difference on the files used to generate these plots verifies that they are identical.

For regression models, the underlying ensemble sampling distribution for the noise is irrelevant, because the results one obtains are invariant when the sampling distribution for the noise changes, provided the projection of the model onto the noise and the mean-square noise value are the same. Thus the underlying ensemble sampling distribution for the noise completely cancels out of the calculation. For models containing nonlinear parameters, we again have a similar result. In these

Figure 3: The Signal $= B_1 + B_2 t + B_3 \cos(\omega t)\exp(-\beta t)$

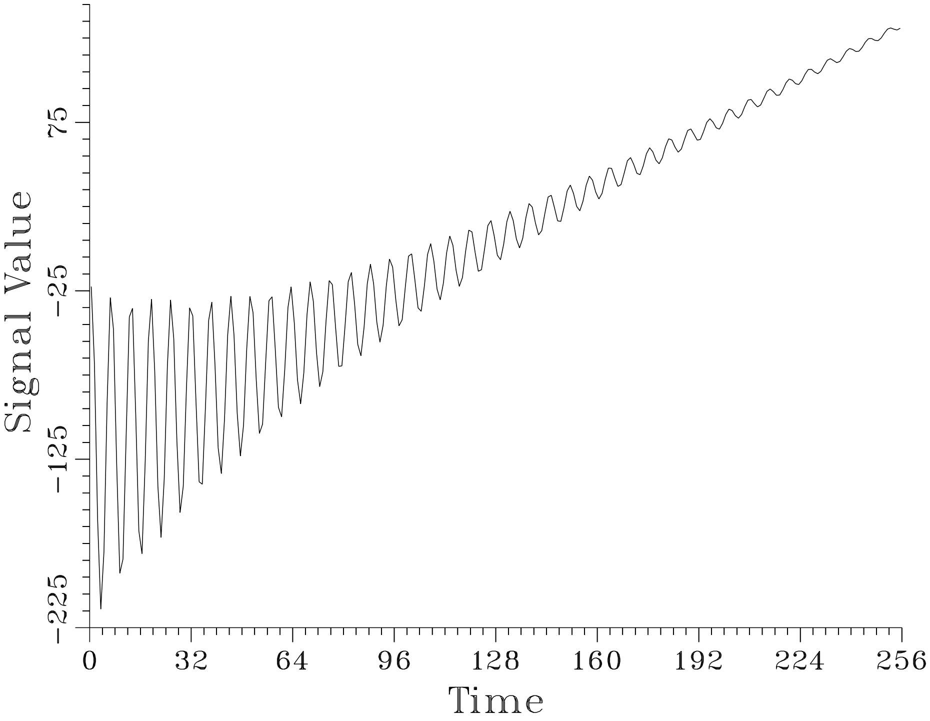

Fig. 3. This signal has parameters $B_1 = -123$, $B_2 = 1$, $B_3 = 100$, $\beta = 4/N$, $\omega = 1$, and $N = 256$. The data shown contain no noise. In order for the probability density functions for the location parameters to be invariant when the ensemble sampling distribution for the noise changes, the projection of the model onto the noise must be the same for each model vector.

models, the probability density functions for the amplitudes, given a fixed value of the nonlinear parameters, are invariant when the ensemble sampling distribution for the noise changes provided the projection of the orthogonal model onto the noise is the same in each noise sample. So again, the underlying noise sampling distribution is irrelevant for any given value of the nonlinear parameters. However, if our estimate of the uncertainty in these parameters is also to remain invariant when the ensemble sampling distribution for the noise changes, then the projection of the signal onto the noise and the mean-square noise value must also remain the same across noise samples. This last requirement is just the statement that the total squared-data value must be the same across noise samples.

The model we are using contains both location parameters, amplitudes, and parameters that appear in a nonlinear fashion. The question we would like to address is: "Are the estimates for the nonlinear parameters invariant when the

Figure 4: Three Different Noise Sets

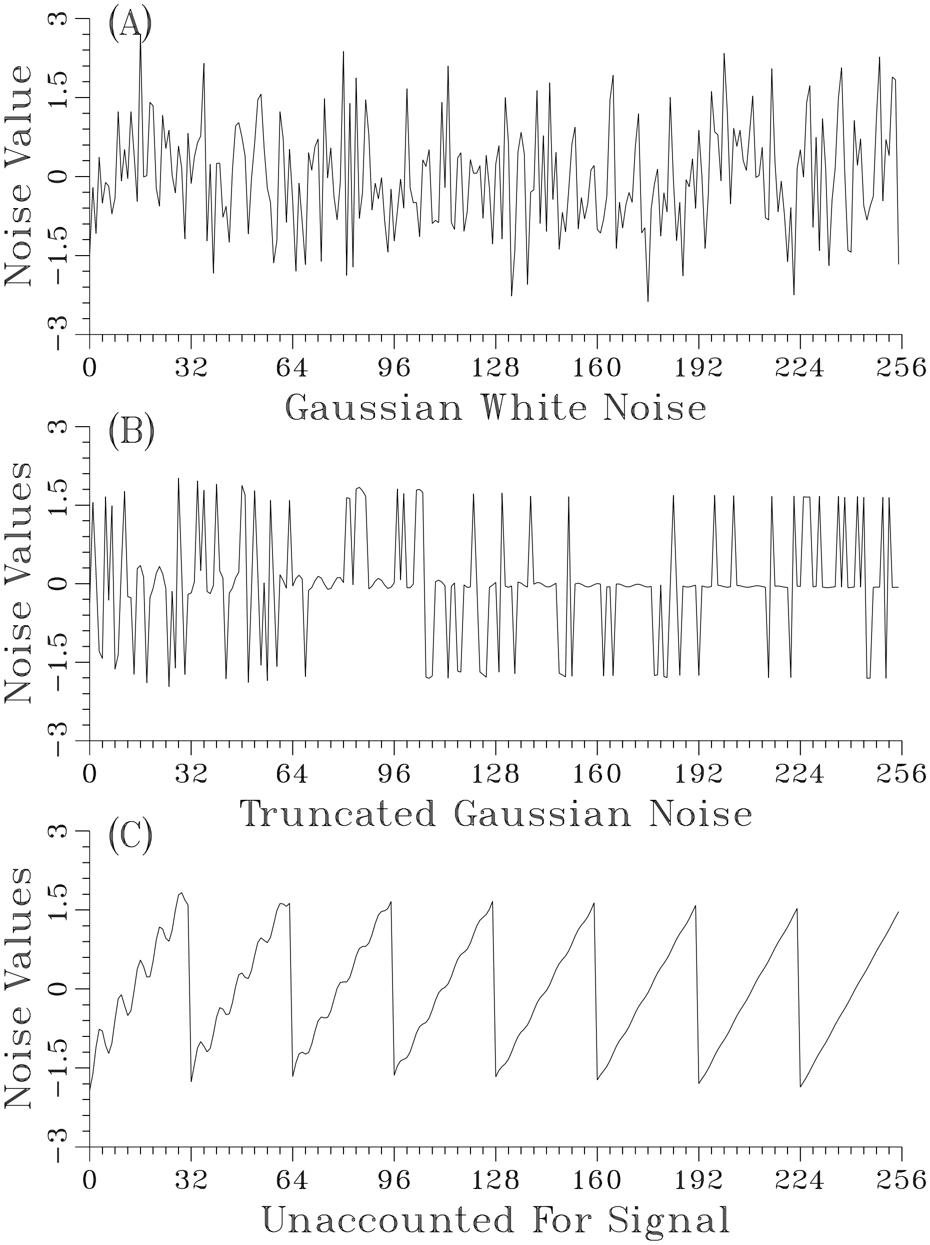

Fig. 4. Three noise sets were prepared: (A) Gaussian white noise, (B) rounded Gaussian white noise, and (C) deterministic signal. In preparing the noise the projection of the model onto the noise had to be removed for each noise set – see text for details.

Figure 5: The Posterior Probability For The Amplitude Of The Sinusoid

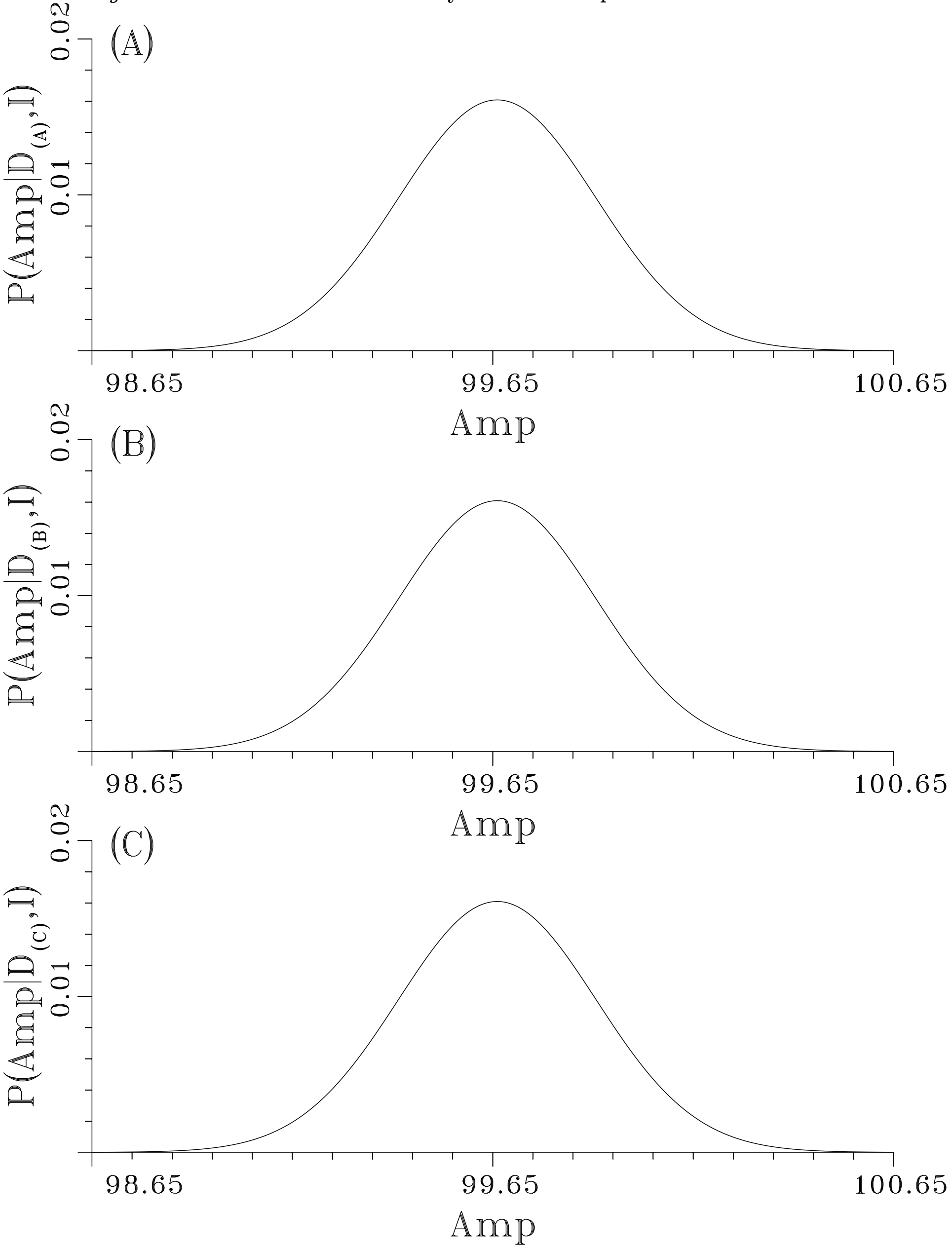

Fig. 5. The signal shown in Fig. 3 was added to the noise shown in Fig. 4 and the posterior probability for the amplitude of the sinusoid was computed, panels (A), (B), and (C) respectively. To the eye, these three probability density functions appear to be identical.

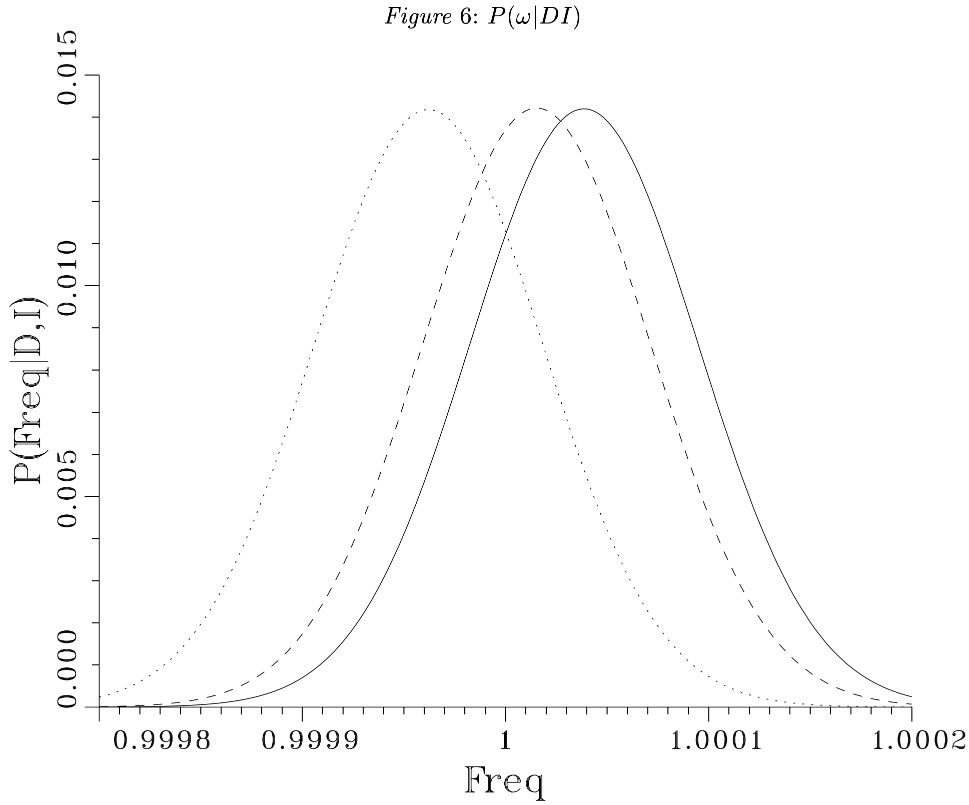

Fig. 6. The probability for the frequency was computed for the signal shown in Fig. 3 and the three noise sets shown in Fig. 4. These have been displayed here. The solid line used the Gaussian white noise, the dotted line used the rounded Gaussian noise, and the dashed line used the deterministic noise.

ensemble sampling distribution for the noise changes?" The short answer to this is "no," and we have illustrated this by plotting the posterior probability for the frequency, $P(\omega|D\alpha I)$ for the data shown in Fig. 3 using the noise shown in Fig. 4. These plots are shown in Fig. 6. The solid line is the posterior probability for the frequency using the signal shown in Fig. 3 plus the Gaussian white noise shown in Fig. 4(A). The dotted line used the rounded Gaussian noise, and the dashed line used the deterministic noise. From this plot it is obvious that the probability for the nonlinear parameters depends on the noise sample. The real questions that should be addressed are "Why do the estimates for the nonlinear parameters depend on the noise sample?" and "Why does assigning a Gaussian distribution for the joint probability for the noise usually give reasonable estimates for these nonlinear parameters?"

First, we will derive the dependence of the joint probability for the nonlinear parameters on the noise, and then we will discuss why assigning a Gaussian distributions for the joint probability for the noise usually gives reasonable results for these nonlinear parameters. The posterior probability for the $\{\theta\}$ parameters is given by

$$\begin{aligned} P(\{\theta\}|\sigma DI) &= \int dA_1 \cdots dA_m P(\{\theta\}\{A\}|DI) \\ &= \int dA_1 \cdots dA_m P(\{\theta\}\{A\}|I) P(D|\{\theta\}\{A\}I). \end{aligned} \tag{52}$$

Because we are interested in the consequences of assigning a Gaussian distribution for the joint probability for the noise, we will take the prior $P(\{\theta\}\{A\}|I)$ to be uniform. Equation (52) is an integral over the likelihood:

$$P(\{\theta\}|\sigma DI) = \int dA_1 \cdots dA_m (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{Q}{2\sigma^2}\right\}, \tag{53}$$

where Q was defined earlier, Eq. (44). Each of the integrals over the amplitudes is an uncoupled Gaussian quadrature integral and is trivial to evaluate. One obtains

$$P(\{\theta\}|\sigma DI) = (2\pi\sigma^2)^{-\frac{N-m}{2}} \exp\left\{-\frac{Q'}{2\sigma^2}\right\} \tag{54}$$

as the posterior probability for the nonlinear parameters, where Q' is

$$Q' = \sum_{j=1}^{m} \hat{A}_j^2 + 2N \sum_{j=1}^{m} \hat{A}_j \overline{H_j(\{\hat{\theta}\})n} + N\overline{n^2} - m\overline{h^2}. \tag{55}$$

The first three terms in the definition of Q' are the total squared-data value, and are the same as what was derived earlier in Eq. (44). The last term, $m\overline{h^2}$, is the total squared-projection of the data onto the model, Eq. (45), and is a function of the nonlinear parameters:

$$\begin{aligned} \overline{h^2} &= \frac{1}{m}\sum_{j=1}^{m} h_j^2 \\ &= \frac{N^2}{m}\sum_{j=1}^{m}\left(\overline{H_j(\{\theta\})n} + \sum_{k=1}^{m}\hat{A}_k \overline{H_k(\{\hat{\theta}\})H_j(\{\theta\})}\right)^2. \end{aligned} \tag{56}$$

From this equation we see why the estimates for the nonlinear parameters are not invariant when the ensemble sampling distribution for the noise changes. The reason is simply that the projection of the jth model function onto the true noise is a function of the $\{\theta\}$ parameters. The only way the posterior probability for the nonlinear parameters could be invariant when the ensemble sampling distribution for the noise changes is if the noise is orthogonal to the model for *all* values of the nonlinear parameters, a condition that could not possibly exist.

However, having said this, it is now obvious why the estimated value of the nonlinear parameters are typically the correct values. It is obvious from the form of Eq. (56) that if the noise goes to zero, the parameter values that maximize the posterior probability are the true parameter value. The confounding term is $\overline{H_j(\{\theta\})n}$. If we evaluate $\overline{h^2}$ at the true parameters, then

$$\overline{h^2} = \frac{1}{m}\sum_{j=1}^{m}\left(N\overline{H_j(\{\hat{\theta}\})n} + \hat{A}_j\right)^2 \tag{57}$$

and the condition that must be satisfied for the probability to have a peak near the true parameter values is:

$$|N\overline{H_j(\{\hat{\theta}\})n}| \ll |\hat{A}_j|; \tag{58}$$

that is, the total projection of the jth model function onto the noise must be much smaller than the true amplitude of the jth model function. This condition can be fulfilled in several different ways. For example, the signal-to-noise ratio might be high, or the noise might be made up of many positive and negative values which when projected onto the signal tend to cancel and thus sum to a small number.

Finally, substituting Eq. (57) into the definition of Q', Eq. (55), a remarkable simplification takes place:

$$Q' = N\overline{n^2} - \sum_{j=1}^{m}\left(N\overline{H_j(\{\hat{\theta}\})n}\,\right)^2. \tag{59}$$

Thus, Q' is just the total squared-true error less the total squared-signal functions projected onto the noise. If the signal functions and the noise are almost orthogonal, the term on the right is essentially zero and the parameter estimates will be good. However, if the signal functions have a significant projection onto the noise, then one may or may not obtain good parameter estimates, depending on the model. The more the signal functions look like the noise, the larger the confounding term, and the parameter estimates one obtains may not be very precise.

6. Summary And Conclusion

When one explicitly includes the parameters, $\{\sigma\}$, and information about the noise, I_σ, in the probability theory calculation, probability theory naturally leads one to assign the joint probability for the noise given both the parameters and the information, Eq. (5). In deriving this equation, it becomes clear that one must assign a joint probability for the noise given what is actually *known* about the true noise.

In this paper we concentrated on the problem of assigning the joint probability for the noise when we did not have very specific information about the noise. In particular we concentrated on the case where the general order and scale, the first two moments, of the noise were supposed known. The principle of maximum entropy was then used to assign the joint probability for the noise. Use of the principle of maximum entropy, with knowledge of the first two moments of the noise results in the assignment of a Gaussian for the joint probability for the noise.

Similarly, if we had used the total absolute value of the noise, a Laplacian would have been assigned. In assigning the joint probability for the noise, one must state explicitly what constraints are to be used in the calculation. Because of the way Eq. (5) was derived, any consistent set of constraints could have been incorporated into the assignment. However, in the case where little is known about the noise one is always better off to leave out constraints because the resulting maximum entropy probability density functions have higher entropy. Higher entropy distributions are by their very nature less informative then lower entropy distributions, and so are more conservative; they make allowances for every possible situation that could occur.

When the principle of maximum entropy is used to assign the joint probability for the noise, the resulting probability density functions have sufficient statistics. Sufficient statistics are functions of the data, and therefore the noise, that summarize all of the information in the data relevant to the problem being solved. As illustrated in the examples, the sufficient statistics are just the constraints used in assigning the joint probability for the noise. In the case of estimating a location parameter, this means that the parameter estimates are invariant when the ensemble sampling distribution for the noise changes, provided the mean and mean-square of the different noise samples are the same. For regression models, the estimated amplitudes remain the same when the ensemble sampling distribution for the noise changes, provided the total projection of the model onto the noise is the across noise sample. Similarly, for models containing nonlinear parameters one obtains the same result for a fixed or given value of the nonlinear parameters. Thus for any given value of the nonlinear parameters the estimated amplitudes one obtains are invariant when the ensemble sampling distribution for the noise changes.

References

1. G. Larry Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*, volume 48. Springer-Verlag, New York, New York, 1988.
2. E. T. Jaynes. *Probability Theory – The Logic of Science.* Copies of this manuscript are available by either anonymous FTP or WWW brouser from “bayes.wustl.edu”, 1993.
3. P. S. Laplace. *A Philosophical Essay on Probabilities.* Dover Publications, Inc., New York, 1951, original publication date 1814. Unabridged and unaltered reprint of Truscott and Emory translation.
4. C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 1948.
5. J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross–entropy. *IEEE Trans. Information Theory IT*, IT-26:26–37, 1980.
6. S. M. Stigler. *Annals of Statistics*, 5:1055–1098, 1977.