AN INTRODUCTION TO MODEL SELECTION USING PROBABILITY THEORY AS LOGIC

G. Larry Bretthorst
Washington University
Department of Chemistry
1 Brookings Drive
St. Louis, Missouri 63130

ABSTRACT. Probability theory as logic is founded on three simple *desiderata:* that degrees of belief should be represented by real numbers, that one should reason consistently, and that the theory should reduce to Aristotelian logic when the truth values of the hypotheses are known. Because this theory represents a probability as a state of knowledge, not a state of nature, hypotheses such as "The frequency of oscillation of a sinusoidal signal had value ω when the data were taken," or "Model x is a better description of the data than model y" make perfect sense. Problems of the first type are generally thought of as parameter estimation problems, while problems of the second type are thought of as model selection problems. However, in probability theory there is no essential distinction between these two types of problems. They are both solved by application of the sum and product rules of probability theory. Model selection problems are conceptually more difficult, because the models may have different functional forms. Consequently, conceptual difficulties enter the problem that are not present in parameter estimation. This paper is a tutorial on model selection. The conceptual problems that arise in model selection will be illustrated in such a way as to automatically avoid any difficulties. A simple example is worked in detail. This example, (radar target identification) illustrates all of the points of principle that must be faced in more complex model selection problems, including how to handle nuisance parameters, uninformative prior probabilities, and incomplete sets of models.

Introduction

A basic problem in science and engineering is to determine when a model is adequate to explain a set of observations. Is the model complete? Is a new parameter needed? If the model is changed, how? Given several alternatives, which is best? All are examples of the types of questions that scientists and engineers face daily. A principle or theory is needed that allows one to choose rationally. Ockham's razor [1] is the principle typically used. Essentially, Ockham's razor says that objects should not be multiplied needlessly. This is typically paraphrased: "When two models fit the observations equally well, prefer the simpler model." This principle has proven itself time and time again as a valuable tool of science. From the standpoint of probability theory, the reason that Ockham's razor works is that simpler models are usually more probable. That simpler models are usually more probable was first argued by Jeffreys [2] and later explicitly demonstrated by Jaynes [3], Gull [4], and Bretthorst [5–8]. However, probability theory tempers Ockham's razor and will allow more complex models to be accepted when they fit the data significantly better or when they contain parameters that have higher initial probability.

This paper is a tutorial on model selection. In it the procedures and principles needed to

apply probability theory as extended logic to problems of model selection will be discussed in detail. Primarily these procedures and principles will be illustrated using an example taken from radar target identification. In this example we will illustrate the assignment of probabilities, the use of uninformative prior probabilities, and how to handle hypotheses that are mutually exclusive, but not exhaustive. While we attempt to explain all of the steps in this calculation in detail, some familiarity with higher mathematics and Bayesian probability theory is assumed. For an introduction to probability theory see the works of Tribus [9], Zellner [10], and Jaynes [11]; for a derivation of the rules of probability theory see Jaynes [11,12], and for an introduction to parameter estimation using probability theory see Bretthorst [13]. In this tutorial the sum and product rules of probability theory will be given and no attempt will be made to derive them. However, as indicated in the abstract, if one wishes to represent degrees of belief as real numbers, reason consistently, and have probability theory reduce to Aristotelian logic when the truth of the hypotheses are known, then the sum and product rules are the unique rules for conducting inference. For an extensive discussion of these points and much more, see Jaynes [11].

1 The Rules of Probability Theory

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

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

where $P(AB|C)$ is the joint probability that "A and B are true given that C is true," $P(A|C)$ is the probability that "A is true given C is true," and $P(B|AC)$ is the probability that "B is true given that both A and C are true." The notation "$|C)$" means conditional on the truth of hypothesis C. In probability theory *all* probabilities are conditional. The notation $P(A)$ is not used to stand for the probability for a hypothesis, because it does not make sense until the evidence on which it is based is given. Anyone using such notation either does not understand that all knowledge is conditional, i.e., contextual, or is being extremely careless with notation. In either case, one should be careful when interpreting such material. For more on this point see Jeffreys [2] and Jaynes [11].

In Aristotelian logic, the hypothesis "A and B" is the same as "B and A," so the numerical value assigned to the probabilities for these hypotheses must be the same. The order may be rearranged in the product rule, Eq. (1), to obtain:

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

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

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

This is Bayes' theorem. It is named after Rev. Thomas Bayes, an 18th century mathematician who derived a special case of this theorem. Bayes' calculations [14] were published in 1763, two years after his death. Exactly what Bayes intended to do with the calculation, if anything, still remains a mystery today. However, this theorem, as generalized by

Laplace [15], is the basic starting point for inference problems using probability theory as logic.

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

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

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

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

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

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

$$P(\omega|D) \equiv P(\omega[B_1 + B_2]|D) = P(\omega B_1|D) + P(\omega B_2|D). \tag{6}$$

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

The marginal probability $P(\omega|D)$ does not depend on the amplitudes at all. To see this, the product rule is applied to the right-hand side of Eq. (6) to obtain

$$P(\omega|D) = P(B_1|D)P(\omega|B_1D) \; + \; P(B_2|D)P(\omega|B_2D) \tag{7}$$

but

$$P(B_1|D) + P(B_2|D) = 1 \tag{8}$$

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

$$P(\omega|D) = \sum_{j=1}^{m} P(\omega B_j|D), \tag{9}$$

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

$$P(\omega|D) = \int dB P(\omega B|D). \tag{10}$$

Note that the B inside the probability symbols refers to the hypothesis; while the B appearing outside of the probability symbols is a number or index. A notation could be developed to stress this distinction, but in most cases the meaning is apparent from the context.

The sum and integral appearing in Eqs. (9,10) are over a set of mutually exclusive and exhaustive hypotheses. If the hypotheses are not mutually exclusive, one simply uses Eq. (4). However, if the hypotheses are *not* exhaustive, the sum rule *cannot* be used to eliminate nuisance hypotheses. To illustrate this, suppose the hypotheses, $B \in \{B, \cdots, B_m\}$, are mutually exclusive, but not exhaustive. The hypotheses B could represent various explanations of some experiment, but it is always possible that there is something else operating in the experiment that the hypotheses B do not account for. Let us designate this as

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

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

$$P(B_i|DI) = \frac{P(B_i|I)P(D|B_iI)}{P(D|I)} \tag{11}$$

and for SE

$$P(\text{SE}|DI) = \frac{P(\text{SE}|I)P(D|\text{SE}I)}{P(D|I)}. \tag{12}$$

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

$$\begin{aligned} P(D|I) &= \sum_{i=1}^{m} P(DB_i|I) + P(D\text{SE}|I) \\ &= \sum_{i=1}^{m} P(B_i|I)P(D|B_iI) + P(\text{SE}|I)P(D|\text{SE}I). \end{aligned} \tag{13}$$

But this is indeterminate because SE has not been specified, and therefore the likelihood, $P(D|\text{SE}I)$, is indeterminate even if the prior probability $P(\text{SE}|I)$, is known. However, the relative probabilities $P(B_i|DI)/P(B_j|DI)$ are well defined because the indeterminacy cancels out. So there are two choices: either *ignore* SE and thereby assume the hypotheses B are complete or *specify* SE, thereby completing the set of hypotheses. One of the main purposes of this tutorial is to illustrate this last alternative and to show how to apply it in real problems.

2 Assigning Probabilities

The product rule and the sum rule are used to indicate relationships between probabilities. These rules are not sufficient to conduct inference because, ultimately, the “numerical values” of the probabilities must be known. Thus the rules for manipulating probabilities must be supplemented by rules for assigning numerical values to probabilities. The historical lack of these supplementary rules is one of the major reasons why probability theory, as formulated by Laplace, was rejected in the late part of the 19th century. To assign any probability there is ultimately only one way, logical analysis, i.e., non-self-contradictory analysis of the

available information. The difficulty is to incorporate only the information one actually possesses without making gratuitous assumptions about things one does not know. A number of procedures have been developed that accomplish this task: Logical analysis may be applied directly to the sum and product rules to yield probabilities (Jaynes [11]). Logical analysis may be used to exploit the group invariances of a problem (Jaynes [16]). Logical analysis may be used to ensure consistency when uninteresting or nuisance parameter are marginalized from probability distributions (Jaynes [21]). And last, logical analysis may be applied in the form of the principle of maximum entropy to yield probabilities (Zellner [10], Jaynes [16,19], and Shore and Johnson [17,18]). Of these techniques the principle of maximum entropy is probably the most powerful, and in this tutorial it will be used to assign all probabilities.

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

$$H \equiv -\sum_{i=1}^{m} P(i|I) \log P(i|I), \tag{14}$$

is a measure of the amount of ignorance (uncertainty) in this probability distribution [22]. 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 testable information I, one can assign the probability distribution, $P(i|I)$, that contains only the information I by maximizing H subject to the information (constraints) represented by I. Because H measures the amount of ignorance in the probability distribution, assigning a probability distribution that has maximum entropy yields a distribution that is least informative (maximally ignorant) while remaining consistent with the information I: the probability distribution, $P(i|I)$, contains only the information I, and does not contain any additional information not already implicit in I [17,18].

To demonstrate its use, suppose that one must assign $P(i|I)$ and nothing is known except that the set of hypotheses is mutually exclusive and exhaustive. Applying the sum rule one obtains

$$\sum_{i=1}^{m} P(i|I) = 1. \tag{15}$$

This equation may be written

$$\sum_{i=1}^{m} P(i|I) - 1 = 0 \tag{16}$$

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

$$H = -\sum_{i=1}^{m} P(i|I)\log P(i|I) + \beta\left[1 - \sum_{i=1}^{m} P(i|I)\right]. \tag{17}$$

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

$$\log P(k|I) + 1 + \beta = 0, \tag{18}$$

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

$$1 - \sum_{i=1}^{m} P(i|I) = 0. \tag{19}$$

Solving this system of equations, one finds

$$P(i|I) = \frac{1}{m} \qquad \text{and} \qquad \beta = \log m - 1. \tag{20}$$

When nothing is known except the specification of the hypotheses, the principle of maximum entropy reduces to Laplace's principle of indifference [15]. But the principle of maximum entropy is much more general because it allows one to incorporate any type of testable information.

As noted earlier, in the inference problem addressed in this paper, there are three different types of information to be incorporated into probability assignments. The specification of parameter ranges occurs when the prior probabilities for various location parameters appearing in the calculation must be assigned. (A location parameter is a parameter that appears linearly in the model equation.) For these location parameters, the principle of maximum entropy leads to the assignment of a bounded uniform prior probability. However, care must be taken because most of these parameters are continuous and *the rules and procedures given in this tutorial are strictly valid only for finite, discrete probability distributions.* The concept of a probability for a hypothesis containing a continuous parameter, a probability density function, only makes sense when thought of as a limit. If the preceding calculations are repeated and the number of hypotheses are allowed to grow infinitely, one will automatically arrive at a valid result as long as all probabilities remain finite and normalized. Additionally, the direct introduction of an infinity into any mathematical calculation is ill-advised under any conditions. Such an introduction presupposes the limit already accomplished and this procedure will cause problems whenever any question is asked that depends on how the limit was taken. For more on the types of problems this can cause see Jaynes [21], and for a much more extensive discussion of this point see

Jaynes [11]. As it turns out, continuous parameters are not usually a problem, provided one always uses normalized probabilities. In this tutorial, continuous parameters will be used, but their prior probabilities will be normalized and the prior ranges will never be allowed to go to infinity without taking a limit.

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

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

$$d_j - L(r_j) = n_j \tag{21}$$

assuming that the noise is additive, and n_j is the true noise value. Thus the probability for the data can be assigned if one can assign a probability for the noise.

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

$$\rho'_s = \frac{1}{N-s}\sum_{j=1}^{N-s} n_j n_{j+s}, \tag{22}$$

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

$$\sigma'_s = \frac{1}{N}\sum_{j=1}^{N} n_j^s, \tag{23}$$

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

The probability that must be assigned is the probability that one should obtain the data D, but from Eq. (21) this is just the probability for noise $P(e_1 \cdots e_N|I')$, where e_j stands for a hypothesis of the form "the true value of the noise at position r_j was e_j, when the data were taken." The quantity e_j is an index that ranges over all valid values of the noise; while the probability for the noise, $P(e_1 \cdots e_N|I')$, assigns a reasonable degree of belief to a particular set of noise values. For the probability for the noise to be consistent with correlations it must have the property that

$$\rho_s = \langle e_j e_{j+s} \rangle \equiv \frac{1}{N-s}\sum_{j=1}^{N-s} \int de_1 \cdots de_N \; e_j e_{j+s} P(e_1 \cdots e_N|I') \tag{24}$$

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

$$\sigma_s = \langle e^s \rangle \equiv \frac{1}{N} \sum_{j=1}^{N} \int de_1 \cdots de_N \; e_j^s P(e_1 \cdots e_N | I') \tag{25}$$

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

In Eq. (22) and Eq. (23), the symbols ρ_s' and σ_s' were used to denote means or averages over the sample noise. These averages are the sample correlation coefficients and moments and they represent states of nature. In Eq. (24) and Eq. (25), the symbols ρ_s and σ_s are used to denote mean averages over the probability for the noise, and they represent states of knowledge. To use information in a maximum entropy calculation, that information must be testable, i.e., the moments and correlation coefficients must be known.

Assuming that none of these quantities are known, how can the principle of maximum entropy be used? Its use requires testable information, and unless at least some of the ρ_s' and σ_s' are known, it would appear that we have no testable information. However, this description of the problem is not what probability theory asks us to do. Probability theory asks us to assign $P(e_1 \cdots e_N | I')$, where I' represents the information on which this probability is based. Suppose for the sake of argument that that information is a mean, ν, and standard deviation, σ, then what probability theory asks us to assign is $P(e_1 \cdots e_N | \nu\sigma)$. This expression should be read as the joint probability for all the errors given that the mean of the errors is ν and the standard deviation of the errors is σ. According to probability theory, in the process of assigning the probability for the errors, we are to assume that both ν and σ are known or given values. This is a very different state of knowledge from knowing that the mean and standard deviation of the sampling distribution are ν and σ. If we happen to actually know these values, then there is less work to do when applying the rules of probability theory. However, if their values are unknown, we still seek the least informative probability density function that is consistent with a fixed or given mean and standard deviation. The rules of probability theory are then used to eliminate these unknown nuisance hypotheses from the final probability density functions.

But which of these constraints should be used? The answer was implied earlier by the way the question was originally posed: what *properties* of the errors are to be used in the calculations? The class of maximum entropy probability distributions is the class of all probability density functions for which sufficient statistics exist. A sufficient statistic is a function of the data that summarizes all of the information in the data relevant to the problem being solved. These sufficient statistics are the sample moments that correspond to the constraints that were used in the maximum entropy calculation. For example, suppose we used the first three correlation coefficients, ρ_1, ρ_2, and ρ_3, as defined by Eq. (24) in a maximum entropy calculation, then the parameter estimates will depend only on the first three correlation coefficients of the data and our uncertainty in those estimates will depend on ρ_1, ρ_2, and ρ_3 if they are known, and on the first three correlation coefficients of the true noise values if ρ_1, ρ_2, and ρ_3 are not known. *All* other properties of the errors have been made irrelevant by the use of maximum entropy. So the real question becomes, what does one know about the errors before seeing the data? If there is information that suggests the errors may be correlated, then by all means a correlation constraint should be included. Additionally, if one has information that suggests the higher moments of the noise can

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

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

$$\nu = \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j P(e_1 \cdots e_N | I') \tag{26}$$

and by

$$\sigma^2 + \nu^2 = \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j^2 P(e_1 \cdots e_N | I') \tag{27}$$

for the second moment, where ν and σ^2 are the fixed or given values of the mean and variance. Note the second moment of the probability distribution, Eq. (27), is written as $\sigma^2 + \nu^2$, to make the resulting probability density function come out in standard notation.

We seek the probability density function that has highest entropy for a fixed or given value of σ^2 and ν. To find this distribution Eq. (26) and Eq. (27) are rewritten so they sum to zero:

$$\nu - \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j P(e_1 \cdots e_N | I') = 0, \tag{28}$$

and

$$\sigma^2 + \nu^2 - \frac{1}{N}\sum_{j=1}^{N}\int de_1 \cdots de_N \; e_j^2 P(e_1 \cdots e_N | I') = 0. \tag{29}$$

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

$$1 - \int de_1 \cdots de_N P(e_1 \cdots e_N | I') = 0. \tag{30}$$

Because Eq. (28) through Eq. (30), sum to zero, they may each be multiplied by a constant and added to the entropy of this probability density function without changing its value,

one obtains

$$\begin{aligned} H &= -\int de_1 \cdots de_N P(e_1 \cdots e_N|I') \log P(e_1 \cdots e_N|I') \\ &+ \beta \left[1 - \int de_1 \cdots de_N P(e_1 \cdots e_N|I')\right] \\ &+ \delta \left[\nu - \tfrac{1}{N}\sum_{j=1}^{N} \int de_1 \cdots de_N \; e_j P(e_1 \cdots e_N|I')\right] \\ &+ \lambda \left[\sigma^2 + \nu^2 - \tfrac{1}{N}\sum_{j=1}^{N} \int de_1 \cdots de_N \; e_j^2 P(e_1 \cdots e_N|I')\right] \end{aligned} \tag{31}$$

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

$$P(e_1 \cdots e_N|\nu\sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{j=1}^{N} \frac{(e_j - \nu)^2}{2\sigma^2}\right\}, \tag{32}$$

where

$$\lambda = \frac{N}{2\sigma^2}, \qquad \delta = -\frac{N\nu}{\sigma^2}, \quad \text{and} \quad \beta = \frac{N}{2}\left[\log(2\pi\sigma^2) + \frac{\nu^2}{\sigma^2}\right] - 1 \tag{33}$$

and I' has been replaced by the fixed or given values of the moments.

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

$$\langle e^s \rangle = \exp\left\{-\frac{\nu^2}{2\sigma^2}\right\} \sigma^{2s} \frac{\partial^s}{\partial \nu^s} \exp\left\{\frac{\nu^2}{2\sigma^2}\right\} \quad (s \geq 0) \tag{34}$$

which reduces to

$$\langle e^0 \rangle = 1, \qquad \langle e^1 \rangle = \nu, \quad \text{and} \quad \langle e^2 \rangle = \sigma^2 + \nu^2 \tag{35}$$

for $s = 0$, $s = 1$, and $s = 2$, just the constraints used to assign the probability density function. Fourth, for a fixed value of the mean and variance this prior probability has highest

entropy. Consequently, when parameters are marginalized from probability distributions or when any operation is performed on them that preserves mean and variance while discarding other information, those probability densities necessarily will move closer and closer to this Gaussian distribution regardless of the initial probability assignment. The Central Limit Theorem is one special case of this phenomenon – see Jaynes [11].

Earlier it was asserted that maximum entropy distributions are the only distributions that have sufficient statistics and that these sufficient statistics are the only properties of the data, and therefore the errors, that are used in estimating parameters. We would like to demonstrate this property explicitly for the Gaussian distribution [11]. Suppose the true value of a location parameter is ν_0 and one has a measurement such that

$$d_j = \nu_0 + n_j. \tag{36}$$

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

$$P(D|\nu\sigma I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(d_j-\nu)^2\right\}. \tag{37}$$

The posterior probability for ν may be written as

$$P(\nu|D\sigma I) \propto (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{N}{2\sigma^2}([\overline{d}-\nu]^2+s^2)\right\} \tag{38}$$

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

$$\overline{d} = \frac{1}{N}\sum_{j=1}^{N} d_j = \nu_0 + \overline{n} \tag{39}$$

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

$$s^2 = \overline{d^2} - (\overline{d})^2 = \frac{1}{N}\sum_{j=1}^{N} d_j^2 - \left(\frac{1}{N}\sum_{j=1}^{N} d_j\right)^2 = \overline{n^2} - (\overline{n})^2 \tag{40}$$

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

$$(\nu)_{est} = \begin{cases} \overline{d} \pm \sigma/\sqrt{N} & \sigma \text{ known} \\ \overline{d} \pm s/\sqrt{N-3} & \sigma \text{ unknown} \end{cases} \tag{41}$$

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

$$\Delta = \overline{d} - \nu_0 = \overline{n} \tag{42}$$

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 values when the standard deviation of the noise is not

known. Thus the underlying sampling distribution of the noise has completely canceled out and the only property of the errors that survives is the actual mean and mean-square of the true noise values. *All* other properties of the errors have been made irrelevant. Exactly the same parameter estimates will result if the underlying sampling distribution of the noise is changed, provided the mean and mean-square of the new sampling distribution is the same, just the properties needed to represent what is actually known about the noise, the mean and mean-square, and to render what is *not* known about it irrelevant.

3 Example – Radar Target Identification

In Section 1 the sum and product rules of probability theory were given. In Section 2 the principle of maximum entropy was used to demonstrate how to assign probabilities that are maximally uninformative while remaining consistent with the given prior information. In this section a nontrivial model selection problem is given. Each step in the calculation is explained in detail. The example is complex enough to illustrate all of the points of principle that must be faced in more complicated model selection problems, yet sufficiently simple that anyone with a background in calculus should be able to follow the mathematics.

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

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

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

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

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

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

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

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

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

Each of these steps will be systematically illustrated in solving a simplified radar target identification problem. In the last section a numerical simulation is discussed.

3.1 Define The Problem

Probability theory solves specific problems in inference. It does this by summarizing ones state of knowledge about a hypothesis as a probability distribution. Thus, to solve an inference problem, one must first state the hypothesis of interest. Here the identification of radar targets will be used to illustrate how to solve model selection problems using probability. However, the subject of this paper is model selection, not radar target identification. For those interested in a more detailed discussion of the fundamentals of radar target identification using probability theory see Jaynes [23]. The hypothesis about which inferences are to be made is of the form "Target number k is being observed by the radar." The index k will represent a particular type of aircraft, or as the radar target identification community refers to them, a particular type of target. The first $\ell - 2$ of these hypotheses represent real aircraft (the known aircraft) and the last two are "The aircraft is NOT a known target," and "No target is in the data, this is a false alarm." The index k really specifies a series of different hypotheses of the form "Hypothesis k is the best description of this state of knowledge." The probability for the kth hypotheses is written $P(k|DI)$, where D is the data and I stands for all of the assumptions and prior information that go into making this a well defined problem. In this problem, as in all realistic problems, this list will be fairly long.

The kth hypothesis is the quantity about which inferences are to be made. The collection of all of these hypotheses is called a library, $L \equiv \{L_1, \ldots, L_\ell\}$, where ℓ is the total number of the hypothesis to be tested. The library is separated into three types of hypotheses: the "known," the "unknown," and the "no-target" hypotheses. Hypotheses one through $(\ell - 2)$ are the known aircraft. These might include the F15, and 747 and a host of others. When making inferences about the known hypotheses, the hypotheses are all of the form "The aircraft being observed is an F15" or "747," etc. In radar target identification, there are so many different types of aircraft, and the number of them changes so rapidly, that one can never be sure of having a hypothesis for all existing aircraft. That is to say, the set of known targets is *not exhaustive.* As was demonstrated earlier, the sum rule may be used to eliminate uninteresting or nuisance hypotheses, but only if the set of hypotheses is exhaustive. Here the hypotheses are mutually exclusive, but not exhaustive. Thus the sum rule cannot not be used unless the set of hypotheses is completed. The set of hypotheses may be made complete either by assuming the set of hypotheses is complete and there by forcing probability to choose from the given set of targets or by defining a model that completes the set. In the radar target identification problem, there is a requirement to be able to identify a hypothesis of the form "the target is NOT one of the known targets." This hypothesis will be number $(\ell - 1)$ in the library. The third class of hypotheses is the "no-target" hypothesis, i.e., no target is present in the data. This hypothesis will be designated as number ℓ.

The hypotheses about which inferences are to be made have now been defined. The needed probability distribution is given symbolically as $P(k|DI)$. However, the definitions of these hypotheses (k, D, and I) are still vague and could describe a host of different problems. To continue with the analysis of this problem, these hypotheses must be made more specific. The process of identifying the relationships between these hypotheses is a process of model building and it is to this task we now turn.

3.2 State The Model

Probabilities are conditional on evidence. Stating the model is the process of relating the hypotheses to that evidence. All types of evidence could be available. In this problem the evidence will consist of data, information about the orientation angle and range to the target, and information about parameter ranges. All of this evidence enters the calculations in exactly the same way, and it doesn't make any difference whether the evidence is data, parameter ranges, or strong prior information. It is all used to assign probabilities conditional on that evidence. To understand the evidence, one must first understand a little about the radar.

The radar is a fictional two-dimensional radar. Schematically, the radar is located at the origin of a polar coordinate system. These coordinates will be referred to as the radar coordinates; they are shown in Fig. 1. The radar captures three different types of data: range, Doppler velocity, and signature data. Only the signature data will be available to the target identification routines. Information from the range and Doppler velocity data will be available in the form of parameter estimates. Additionally, in the real radar target identification problem, information about the velocity, altitude, and acceleration could be used to help identify targets, because this information would effectively eliminate many different types of aircraft. However, in this tutorial, our attention will be restricted to the signature data and the range and Doppler velocity data will be used only to the degree necessary to locate the target in the signature data.

The range data is the vector position of the target as measured in the radar coordinates. Each measurement consists of three numbers: the vector range to target, R_0, Θ, and the time of the measurement. The radar gathers these range measurements periodically, about one measurement every second or so.

The Doppler velocity data is a scalar and represents the speed of the target as projected along the range vector. That is to say, it represents how fast the target is approaching the radar; it is not the target's velocity vector. These measurements are acquired at the same time as the range measurement.

The information needed by the identification calculation is the true range, R_c and orientation angle, ω, of the target. These are shown schematically in Fig. 2. The radar estimates these quantities from the measured range and Doppler velocity data. These inferred or measured values will be denoted as R_0 and Ω respectively. Inferring these quantities is an extensive calculation using probability theory. The details of these calculations are presented in Bretthorst [24]. The results of these inferences are available to the identification routines in the form of a (mean $\pm$ standard deviation) estimate of these quantities. These estimates are interpreted as probabilities in the form of

$$P(\omega|I_\Omega) = (2\pi\sigma_\Omega^2)^{-\frac{1}{2}} \exp\left\{-\frac{[\Omega-\omega]^2}{2\sigma_\Omega^2}\right\} \tag{44}$$

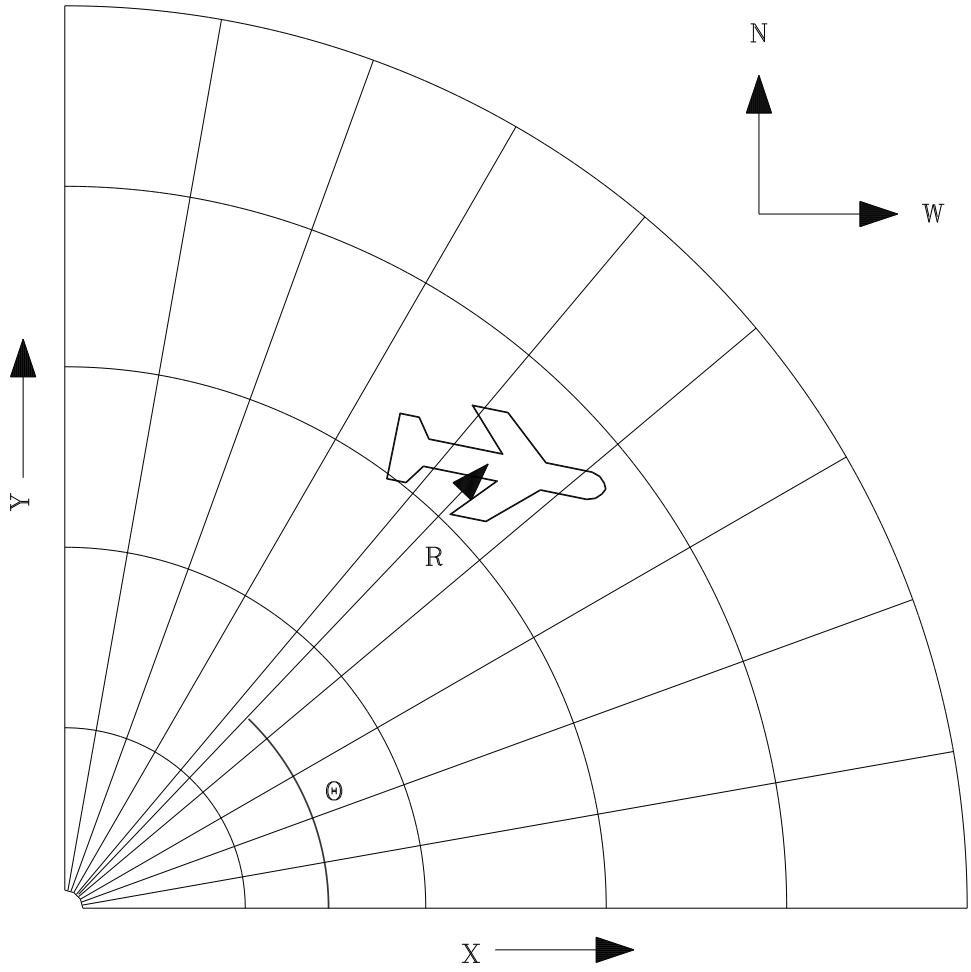

Fig. 1. The radar takes three different types of data: range, Doppler velocity, and signature data. The range data is the vector distance to the center of the target. The Doppler velocity data is the projection of the vector velocity onto the range vector, i.e., it is how fast the target is approaching the radar. Last, the signature data is the envelope of the reflected radar signal, as the signal crosses the target.

where σ_Ω^2 is the uncertainty in this estimate, and I_Ω is the information on which this probability is based. Equation (44) is the probability for a set of hypotheses. The hypotheses are of the form: "The true orientation angle of the target is ω." Similarly for the range to the target one has

$$P(R_c|I_R) = (2\pi\sigma_R^2)^{-\frac{1}{2}} \exp\left\{-\frac{[R_0 - R_c]^2}{2\sigma_R^2}\right\} \tag{45}$$

where σ_R^2 is the uncertainty in the estimated range, and I_R stands for the evidence on which the range estimate is based.

The radar gathers a third type of data, the signature data $D \equiv \{d_1, \cdots, d_N\}$, where N is the number of data values in a signature data set. If the radar were operating in the optical limit, the signature data would be the intensity of the reflected radar signal as the transmitted wave crosses the target. Data typical of this type of radar are shown in Fig. 3. The amplitudes of the peaks shown in Fig. 3 are a very sensitive function of the target orientation, while the locations of the peaks in the data represent the line of site distance to a scatterer (a surface orthogonal to the radar). Note that the radar is an envelope detector, so the signature data, as implied by Fig. 3, are positive. However, the radar does not operate in the optical limit, so the scattering center model is only an approximation. For high range resolution radars, this approximation appears adequate to represent isolated scatterers. It is not yet known if it is adequate to represent more complex interactions, like those between the radar and the engine cavities or propellers.

The signature data may be modeled as

$$d_j = L_k(r_j) + n_j \tag{46}$$

where d_j represents the data sampled at range r_j, $L_k(r_j)$ is the target signature evaluated at position r_j, and n_j is the noise in this measurement. The distances, r_j, correspond to distances across a target and these will be referenced to the center of the target.

The functional form of the signal is different for each of the three types of models. If the target is one of the known aircraft ($1 \le k \le \ell - 2$), then a scattering center model allows one to relate the target to the data:

$$d_j = B_0 + \sum_{l=1}^{N_k} B_l G(S_{kl}\cos(\phi_{kl} - \omega) - r_j + R_c) + n_j \qquad (1 \le k \le \ell - 2) \tag{47}$$

where k is the true target index, B_0 represents a dc offset in the data, B_l is the unknown amplitude of the lth scatterer, N_k is the number of scatterers, G is the peak shape function and is a fundamental characteristic of the radar, (S_{kl}, ϕ_{kl}) is the polar location of the scatterer in the target coordinates (polar coordinates on the target with the x axis orientated along the main axis of the aircraft), and (R_c, ω) are the true range and orientation angle of the target. The location of the scatterers (S_{kl}, ϕ_{kl}) and the number of scatterers, N_k, are known quantities and define what is meant by a known target. The constant term may be incorporated into the sum by rewriting the model as

$$d_j = \sum_{l=0}^{N_k} B_l G(S_{kl}\cos(\phi_{kl} - \omega) - r_j + R_c) + n_j \tag{48}$$

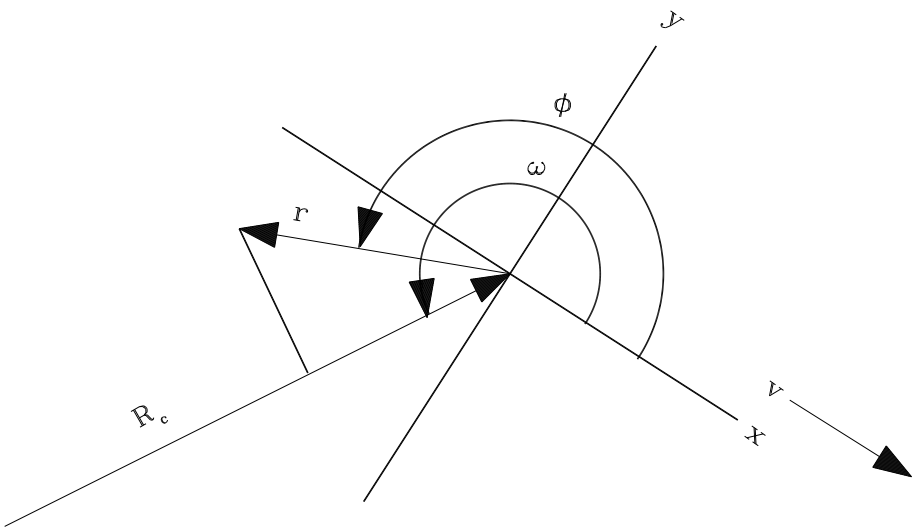

Fig. 2. The observation angle is the difference between the angular location of a scatterer, ϕ, and the orientation angle of the target, ω. These angles are measured in the local target coordinates. The target is orientated along its velocity vector so the observation angle is calculated from the range and velocity vectors of the target.

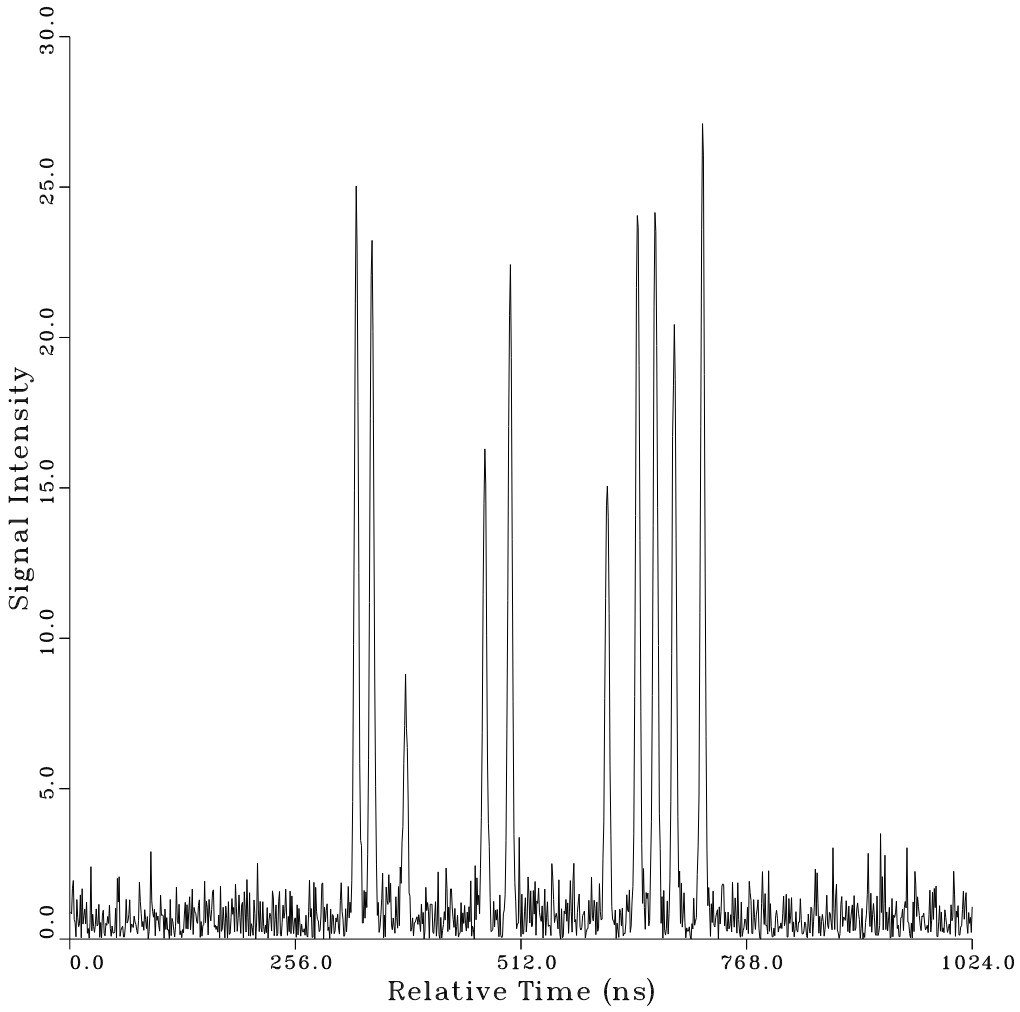

Fig. 3. The signature data represents the intensity of the received signal as the radar signal crossed the target. Locations on the target orthogonal to the radar reflect a large signal, while other locations scatter the radar signal off into space. The peak shape is a characteristic of the radar, while the intensity of the return is a complicated function of range, orientation, and the electromagnetic properties of the target surface.

with the understanding that the signal function G is a constant, for $l = 0$.

The simplest of the three types of models is the no-target model ($k = \ell$). In this model there are no scatterers, only noise. But the noise is positive, the radar is an envelope detector, so the signature data will contain a constant, corresponding to the dc component of the rectified noise. This model may be written as

$$d_j = B_0 + n_j. \tag{49}$$

In addition to the known and no-target hypotheses, the radar must also identify the unknown target. Any type of model that has the ability to expand the data on a complete set is a suitable model for the unknown. However, when expanding the data on a complete set, it is always advisable to choose a basis which captures the essence of the signal. The signal from an unknown aircraft contains an unknown number of scatterers of known peak shape, so an appropriate model would be

$$d_j = \sum_{l=0}^{N_\nu} B_l G(S_{[\ell-1]l} - r_j) + n_j \tag{50}$$

where N_ν is the unknown number of scatterers and the other symbols retain their meaning. From a single data set, there is no information about the angular location of the scatterers, and so no need to include the rotation (the cosine) or to reference the scatterers to the center of the target. Consequently, R_c, ϕ, and ω do not appear in Eq. (50).

The problem now has enough structure to begin the process of applying the rules of probability theory. However, the models are still incomplete in the sense that all of the information available has not yet been supplied. To give just one example, there are a number of amplitude parameters in these models. These amplitudes represent the intensity of the reflected signal. A great deal is known about the possible range of values for these amplitudes. Eventually, probability theory will ask us to supply this information. But supplying it will be delayed until the form in which this information is needed is known.

In one sense delaying this hides some of the beauty of probability theory as logic, because it will appear as if the prior information is being handled differently from the data. In fact this is not the case. For notational convenience, what will be done is that information other than data will be represented as I in all probability symbols. When manipulating probabilities, I must be thought of exactly as if it were any other hypothesis. When prior probabilities are assigned these will typically depend only on I. At that time it must be asked exactly what information is available about the hypothesis and then that information must be used in assigning the probability. If all of the information had been made explicit at the beginning of the calculations this last step would not be necessary because each probability would automatically indicate the evidence on which it is to depend. So by delaying the process of identifying the prior information the notation has been simplified, but at the expense of making prior information seem somehow different from data; which it is not.

3.3 Apply Probability Theory

The problem is to determine which of the hypotheses k is most probable in view of the data and all of one's prior information. This posterior probability is denoted by $P(k|DI)$. To

calculate this posterior probability, one applies Bayes' theorem, Eq. (3), to obtain:

$$P(k|DI) = \frac{P(k|I)P(D|kI)}{P(D|I)}. \tag{51}$$

To compute the posterior probability for the target k one must assign three terms. The first term, $P(k|I)$, is the probability for the target given only the information I. This term is referred to as a prior probability, or simply as a "prior" and represents what was known about the presence of this target before obtaining the data D. The second term, $P(D|kI)$, is the probability for the data given that the true hypothesis is k. This term is referred to as the marginal likelihood of the data for reasons that will become apparent shortly. The third term, $P(D|I)$, is the global likelihood for the data, and is a normalization constant.

The prior probability, $P(k|I)$, is sufficiently simplified that it could be assigned. Depending on the location of the radar, there could be either a great deal or very little prior information available. For example, if the radar were located at a civilian airport the types of aircraft one would expect to observe would be very different from what one would expect to observe on an aircraft carrier. Additionally, it is always possible that the radar has just been installed and there is no historical information on which to base a prior. This latter assumption will be used in this tutorial and the principle of maximum entropy will lead us to assign a uniform prior probability to this term.

The global likelihood for the data, $P(D|I)$, is a normalization constant. The way to calculate it is to calculate the joint probability for the data and the model, $P(Dk|I)$, and then apply the sum rule to eliminate k from consideration:

$$P(D|I) = \sum_{k=1}^{\ell} P(Dk|I). \tag{52}$$

This can be factored using the product rule, Eq. (1), to obtain:

$$P(D|I) = \sum_{k=1}^{\ell} P(k|I)P(D|kI). \tag{53}$$

Note, that, as asserted earlier, this term is a sum over all values appearing in the numerator, so it is just the constant needed to ensure the total probability is one. The global likelihood may now be substituted back into the posterior probability for the kth hypothesis, Eq. (51), to obtain

$$P(k|DI) = \frac{P(k|I)P(D|kI)}{\sum_{\eta=1}^{\ell} P(\eta|I)P(D|\eta I)} \tag{54}$$

where the summation index was changed to avoid confusion.

To simplify some of the notation in what follows, the normalization constant will be dropped, and the equal sign will be replaced a proportionality sign. At the end of the calculations the normalization constant will be computed. With this change, the posterior probability for the models becomes

$$P(k|DI) \propto P(k|I)P(D|kI). \tag{55}$$

The only remaining term that must be addressed is the marginal likelihood for the data $P(D|kI)$. The model hypothesis explicitly appears in this term. There are three different types of models each having different parameterizations; consequently there are three distinct applications of the rules of probability theory needed to simplify this term. The no-target model is by far the simplest of the three and it will be dealt with first.

Apply Probability Theory Given The No-Target Model

The marginal likelihood is computed from the joint likelihood of the data and the nuisance hypotheses or parameters. The sum rule is then used to remove the dependence on the nuisance parameters. For the no-target hypothesis there is only a single nuisance parameter, B_0, so the marginal likelihood is given by

$$P(D|\ell I) = \int dB_0 P(DB_0|\ell I) \tag{56}$$

where the integral is over all possible values of the constant B_0, and k has been replaced by ℓ to indicate that it is the marginal likelihood of the no-target model that is being computed. It should now be apparent why $P(D|kI)$ is called a marginal likelihood. It is a likelihood because it is the probability for the data given the model. It is a marginal probability because, to compute it, one must marginalize over all nuisance parameters appearing in the model.

To continue with the calculation, the product rule, Eq. (1), is applied to the right-hand side of the marginal likelihood, Eq. (56), to obtain:

$$P(D|\ell I) = \int dB_0 P(B_0|I) P(D|B_0 \ell I) \tag{57}$$

where it has been assumed that the constant dc offset (which is a characteristic of the noise) does not depend on which target is present, and $P(D|B_0\ell I)$ is the direct probability for the data given the hypothesis, or the likelihood function. Substituting the marginal likelihood into the posterior probability, Eq. (55), one obtains

$$P(\ell|DI) \propto P(\ell|I) \int dB_0 P(B_0|I) P(D|B_0\ell I). \tag{58}$$

Given the assumptions made, these probabilities may not be further simplified; the only recourse is to assign them numerical values and perform the indicated integral. These probabilities will be assigned in Section 3.4 and the integrals evaluated in 3.5.

Apply Probability Theory Given The Known Target Model

There are three types of models, so three applications of the rules of probability theory are needed to simplify the marginal likelihoods. The previous subsection dealt with the marginal likelihood for the no-target model; here the marginal likelihood for the known target hypothesis will be simplified. As was indicated previously, the marginal likelihood of the data is computed from the joint likelihood of the data and the nuisance parameters. For the known targets these parameters are the amplitudes, B, the true position R_c, and orientation angle of the target ω. The position of the scatterer (S_{kl}, ϕ_{kl}) and the number

of scatterers, N_k, are known. The marginal likelihood for the data given the known target hypothesis is given by

$$P(D|kI) = \int dB d\omega dR_c P(DB\omega R_c|kI) \qquad (1 \le k \le \ell - 2) \tag{59}$$

where the range on the integrals will be discussed later. Applying the product rule, to the right-hand side of the marginal likelihood one obtains

$$P(D|kI) = \int dB d\omega dR_c P(B\omega R_c|kI) P(D|B\omega R_c kI) \qquad (1 \le k \le \ell - 2) \tag{60}$$

where $P(B\omega R_c|kI)$ is the joint prior probability for the nuisance parameters given the known target hypothesis and the prior information I, and $P(D|B\omega R_c kI)$ is the likelihood of the data given the model parameters.

In the previous example there was only a single nuisance hypothesis or parameter, the dc offset, so after factoring the joint-likelihood the calculation was essentially finished. In this example there are many additional hypotheses which requires many additional applications of the product rule. The process is begun by applying the product rule to the joint prior probability for the parameters:

$$P(B\omega R_c|kI) = P(R_c|kI) P(B\omega|R_c kI) \tag{61}$$

where $P(R_c|kI)$ is the prior probability for the range to the target, and $P(B\omega|R_c kI)$ is the joint prior probability for the amplitudes and the orientation angle given the true target k, and the range R_c. In both these probabilities, the identity of the target is given. However, knowing the target identity may or may not help one in assigning either of these terms. When assigning the prior probability for the range to target, $P(R_c|kI)$, knowing the target index, k, would enable one to limit the range of valid values, because the length of the target k would be known. But compared to the six inch range resolution of the radar, knowing the total length of the target is essentially irrelevant. Consequently, it will be assumed that knowing the target identity does not increase our state of knowledge about the range to target and the reference to hypothesis k will be dropped from $P(R_c|kI)$ giving $P(R_c|I)$.

In the case of the joint prior probability for the amplitudes and the orientation angle, $P(B\omega|R_c kI)$, knowing which target is present does not increase our state of knowledge about either the amplitudes or the orientation angle, because the intensity of a scatterer is determined by constructive and destructive interference of the radar waves in the reflected signal. Because the size of the target is large relative to the wavelength of the transmitted signal, large changes in the amplitudes occur for small changes in the orientation angle. But the orientation angles are known only to about one or two degrees. Consequently, knowing the true hypothesis k does not improve our state of knowledge about the amplitudes. And because our state of knowledge about the amplitudes does not improve, there is no additional information about the orientation angle of the target. So whether or not the true target is known does not change our state of knowledge about the amplitudes or the orientation angle. As a result the reference to true hypothesis k may be dropped from the right-hand side of the prior, giving

$$P(B\omega R_c|kI) = P(R_c|I) P(B\omega|R_c I). \tag{62}$$

The previous discussion is one of deciding the logical independence of two or more hypotheses. It occurs in every problem in probability theory. Sometimes probabilities are logically independent and sometimes they are not; each case must be decided based on what one knows. When hypotheses are logically independent, the independent hypotheses may be dropped from the right-hand side of the appropriate probability. However, if the hypotheses are logically dependent, then one must follow the rules of probability theory to obtain valid results.

To illustrate that nonsense may be obtained if logical dependence is ignored, we give one of E. T. Jaynes' favorite examples: suppose someone polled every person in England about the height of the queen-mother. Then the probability for her height, H, given the responses $d_1, \ldots, d_n$ and the prior information I would be written:

$$P(H|d_1 \ldots d_n I) = P(H|I)P(d_1 \ldots d_n|HI). \tag{63}$$

Assuming logical independence, one obtains

$$P(H|d_1 \ldots d_n I) = P(H|I)P(d_1|HI)P(d_2|HI)\ldots P(d_n|HI) \tag{64}$$

If $N \approx 10^6$ then the square root of N effect would imply that her height may be estimated to roughly a part in a thousand, clearly an absurd result. The reason is because the measurements are correlated. From the product rule one obtains

$$P(H|d_1 \ldots d_n I) = P(H|I)P(d_1|HI)P(H d_2 \ldots d_n|d_1 I). \tag{65}$$

So only the first data item may be assigned an independent probability. All the others must be assigned assuming the first data item known. But each person's opinion is based on news reports, papers, books, and by discussing her height with other people who all have access to basically the same information. All of the opinions are correlated: the data are not independent. In other words, ten million uninformed opinions are not as good as one expert opinion, a fact many politicians and pollsters have forgotten.

To determine whether one hypotheses is logically independent of another the only relevant question is to ask, would knowing the first hypothesis help to determine the other? If the answer to this is yes, the hypotheses are not logically independent and the rules of probability theory must be followed exactly to obtain a valid result. In this tutorial, logical independence will be assumed in many cases. In each case it will be pointed out when and why it is being used. However, in any given problem logical independence may or may not hold. Each case must be determined on its own merits and failure to resolve the issue correctly can lead to nonsense; not because probability theory is wrong, but because from a false hypothesis all conclusions follow, a simple fact of logic.

If logical independence is assumed, Eq. (62) may be factored to obtain

$$P(B\omega R_c|kI) = P(R_c|I)P(\omega|I)P(B_0|I)P(B_1|I)\cdots P(B_{N_k}|I). \tag{66}$$

Logical independence follows here for all the same reasons given earlier: the scatterers change intensity so rapidly, and in so unpredictable a manner, that knowledge of any one amplitude will not aid one in predicting the amplitudes of the others. Substituting the

factored prior back into the posterior probability for the known targets, Eq. (55), one obtains

$$\begin{aligned} P(k|DI) &\propto P(k|I)\int dBd\omega dR_c P(R_c|I)P(\omega|I) \\ &\times P(B_0|I)P(B_1|I)\cdots P(B_{N_k}|I)P(D|B\omega R_c kI) \qquad (1 \le k \le \ell-2) \end{aligned} \tag{67}$$

as the posterior probability for the known targets. None of the indicated probabilities may be further simplified. The next step in the calculation is to assign these probabilities numerical values and then perform the indicated integrals. These last two steps will be delayed until after the marginal likelihood for the unknown target has been simplified.

Apply Probability Theory Given The Unknown Target Model

Simplifying the marginal likelihood for the unknown target hypothesis is similar to what was done previously. The marginal likelihood given the unknown model is computed from the joint probability for the data and the nuisance parameters. For the unknown hypothesis the nuisance parameters are the amplitudes, B, the locations of the scatterers, $S \equiv \{S_1 \cdots S_{N_\nu}\}$, and the number of scatterers, N_ν. Applying the sum rule, the marginal likelihood is given by

$$P(D|[\ell-1]I) = \sum_{N_\nu=0} \int dBdSP(DBSN_\nu|[\ell-1]I) \tag{68}$$

where the target index k was replaced by $[\ell-1]$ to indicated that this is the posterior probability for the unknown hypothesis. The upper limit on this sum will be discussed when the prior probability for the number of scatterers is discussed. Also note that scatterer N_0 is the dc offset. Applying the product rule one obtains

$$P(D|[\ell-1]I) = \sum_{N_\nu=0} \int dBdSP(BSN_\nu|[\ell-1]I)P(D|BSN_\nu[\ell-1]I) \tag{69}$$

where $P(BSN_\nu|[\ell-1]I)$ is the joint prior probability for the parameters, and $P(D|BSN_\nu[\ell-1]I)$ is the likelihood of the data given those parameters. Using the logical independence assumption and substituting into the posterior probability for the unknown, one obtains

$$\begin{aligned} P(\ell-1|DI) &\propto P(\ell-1|I)\sum_{N_\nu=0}\int dBdSP(B_0|I)P(B_1R_cI)\cdots P(B_{N_\nu}|R_cI) \\ &\times P(N_\nu|I)P(S_1|I)\cdots P(S_{N_\nu}|I)P(D|BSN_\nu[\ell-1]I). \end{aligned} \tag{70}$$

The discussion on logical independence for the amplitudes given earlier applies equally well to the location of the scatterers. Because, the amplitudes cannot be predicted from first principles, knowing the amplitudes does not help in determining the location of the scatterers and conversely. The point has now been reached where these probabilities may not be further simplified. The next step in the calculation is to assign these probabilities numerical values and it is to this problem that we now turn.

3.4 Assign The Probabilities

The posterior probability for the hypothesis of interest has one of three different functional forms depending on the particular hypothesis, Eqs. (58,67,70). These three equations contain seven prior probabilities and three likelihood functions. The prior probabilities specify what was known about the various hypotheses before obtaining the data; while the likelihood functions tell us what was learned about the hypotheses from the data. These probabilities must be assigned to reflect the information actually available. Earlier, the principle of maximum entropy was used to assign three different probabilities: one when only the number of hypotheses, or range of values, was known. This lead to a uniform probability distribution. The other two cases assumed the first two moments of a probability distribution to be known and led to a Gaussian probability distribution. All three of these calculations will now be used to assign the indicated probabilities.

Assigning The Prior Probabilities

Of the seven prior probabilities that must be assigned, three of them have already been touched on. First, the prior probability for the targets, $P(k|I)$, represents what was known about the target before obtaining the data. In the numerical simulations that follow, the enumeration of possible targets is all that is assumed known. Using this information the principle of maximum entropy will assign a uniform prior probability. Because this prior appears in every target's posterior probability exactly one time, the prior range will cancel when the posterior probability is normalized. The other two prior probabilities discussed were those for the location and the orientation angle of the target, Eqs. (44,45). The remaining four priors that must be assigned are: $P(B_0|I)$, $P(B_l|I)$, $P(N_\nu|I)$, and $P(S_j|I)$. The first step in accomplishing this task is to state the information on which a given prior is to be based. In these four cases, the prior information will consist of the valid range of these parameters. This will result in assigning a uniform prior probability. However, care must be taken in assigning these priors because the three types of models have differing numbers and types of parameters and prior ranges are what sets the scale of comparison between the three types of models.

The prior probability for the constant dc offset, $P(B_0|I)$ is the simplest to address and will be taken first. The dc offset, like the prior probability for the target, occurs in every model exactly one time. Any constants that appear in each posterior probability the same number of times will cancel when the posterior probability is normalized. If a uniform prior probability is assigned, the prior range for B_0 will cancel. But note that it is the prior range that cancels, the integral over B_0 must still be over the valid ranges for this parameter. To specify this prior, the range of valid values must be given. In this calculation an approximation will be used that will simplify the results somewhat while introducing only a small error in the calculation. The approximation is that the integration ranges are wide compared to the expected value of the parameter. Consequently, when the integral over the dc offset is evaluated, the limits on the integral may be extended from minus to plus infinity. This amounts to ignoring a term contributed by an error function. But the error function goes to one so rapidly for large arguments that, for all practical purposes, the approximation is exact. Because the prior ranges cancel, it will not be specified other than saying it is uniform with wide bounds.

The next prior probability to be assigned is $P(B_l|I)$, the prior probability for an amplitude. What is known about the amplitudes? The amplitudes are bounded. The bounds are known based on the transmitted signal power, the distance to the target, the reflectivity of the target surface, the surface area of the scatterer, and the efficiency of the receiver. The amplitude must satisfy

$$0 \leq B_l \leq B_{max} \tag{71}$$

where B_{max} is the maximum signal intensity that one could obtain for any target. Using the principle of maximum entropy results in assigning a uniform prior probability given by

$$P(B_l|I) = \begin{cases} \dfrac{1}{B_{max}} & \text{If } 0 \leq B_l \leq B_{max} \\ 0 & \text{otherwise} \end{cases} . \tag{72}$$

To assign the prior probability for the unknown number of scatterers, $P(N_\nu|I)$, one must again state what is known. In this case, the unknown number of scatterers in a particular data set could range from one (this prior only occurs in models that have at least one scatterer) up to a maximum. But what is the maximum value? There are N data values, and if there were N scatterers, the data could be fit exactly by placing a scatterer at each data value and adjusting its amplitude. Because, no additional information is available about the number of scatterers, N may be taken as an upper bound. Using the principle of maximum entropy, one obtains

$$P(N_k|I) = \begin{cases} \dfrac{1}{N} & \text{If } 1 \leq N_k \leq N \\ 0 & \text{otherwise} \end{cases} \tag{73}$$

as the prior probability for the unknown number of scatterers.

Last, to assign the prior probability for the location of the scatterers, $P(S_l|I)$, one must again state what is actually known about their locations. The location of the target is known to within about 6 inches. The range window (the distance represented by the signature data) is centered on the middle of the target. So the scatterers must be somewhere within the data. If only this is known, then the principle of maximum entropy will again assign a uniform prior probability for the location of the scatterers:

$$P(S_l|I) = \begin{cases} \dfrac{1}{N} & \text{If } 1 \leq S_l \leq N \\ 0 & \text{otherwise} \end{cases} \tag{74}$$

where the range dimensions were taken to be unit steps.

All of the prior probabilities have now been assigned. These priors were uniform priors in the cases where only the valid range of values were known, and they were Gaussians when the prior information consisted of a (mean $\pm$ standard deviation) estimate of a parameter value. The only remaining probabilities that must be assigned are the three likelihoods, and, as it will turn out, these probabilities are also prior probabilities.

Assigning The Likelihoods

In the radar target identification problem there are three likelihoods: the likelihood of the data given the no-target hypothesis, $P(D|\ell B_0 I)$; the likelihood of the data given the known target hypothesis, $P(D|B\omega R_c k I)$; and the likelihood of the data given the unknown target hypothesis $P(D|BSN_k[\ell-1]I)$. To assign them, first note that the data D are not a single hypothesis, rather they represent a joint hypothesis: $D \equiv \{d_1, \ldots, d_N\}$. Applying the product rule and representing all of the given quantities as I', the likelihoods may be factored to obtain

$$P(d_1 \ldots d_N|I') = P(d_1|I')P(d_2 \ldots d_N|d_1 I'). \tag{75}$$

Probability theory tells one to assign the probability for the first data item given the parameters, and then assign the probability for the other data items assuming one knows the first data value. Probability theory automatically guards against the example mentioned earlier where assuming logical independence leads to nonsense. However, the designers of the radar take great care to insure that the errors in the data are independent. Given this is the case, the likelihoods may be factored to obtain

$$P(d_1 \ldots d_N|I') = P(d_1|I') \cdots P(d_N|I'). \tag{76}$$

The probability for the data is just the product of the probabilities for obtaining data items separately. Each of our model equations is of the form

$$n_j = d_j - L_k(r_j) \tag{77}$$

where $L_k(r_j)$ is the kth library model evaluated at position r_j. The probability for obtaining the data is just the probability that one should obtain a particular set of errors given that one knows the true signal $L_k(r_j)$.

Earlier it was shown that the results obtained using a Gaussian noise prior probability depend only on the first and second moments of the true noise values in the data. So if a Gaussian distribution is used for the prior probability for the noise, the results obtained will not depend on the underlying sampling distribution of the errors. But note that assigning a Gaussian noise prior probability in no way says the noise is Gaussian; rather, it says only that our estimates and the uncertainty in those estimates should depend only on the first and second moments of the noise. Notice that the Gaussian probability, Eq. (32), assumes the noise standard deviation is known, so σ must be added to the likelihoods in such a way as to indicate that it is known; this gives

$$P(D|\sigma I') = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{j=1}^{N} \frac{[d_j - L_k(r_j)]^2}{2\sigma^2}\right\}. \tag{78}$$

as the likelihood function. Using this equation as a prototype, the likelihood for the data given the known target hypothesis is given by

$$\begin{aligned} P(D|B\omega R_c \sigma k I) &= (2\pi\sigma^2)^{-\frac{N}{2}} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}[d_j - \sum_{l=0}^{N_k} B_l G(S_{kl}\cos(\phi_{kl}-\omega) - r_j + R_c)]^2\right\} \end{aligned} \tag{79}$$

where $(1 \leq k \leq \ell - 2)$ and, I' as been replaced by all of the given parameters. Similarly, the likelihood for the data given the unknown target hypothesis is given by

$$P(D|BSN_\nu\sigma[\ell-1]I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}[d_j - \sum_{l=0}^{N_\nu} B_l G(S_l - r_j)]^2\right\}. \tag{80}$$

Last, the likelihood for the data given the no-target hypothesis is given by

$$P(D|\ell B_0\sigma I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}[d_j - B_0]^2\right\}. \tag{81}$$

With the assignment of these likelihoods, all of the probabilities have now been assigned. The next task is to perform the indicated integrals and sums.

3.5 Evaluate The Integrals And Sums

All that remains to formally complete the problem is to apply the sum rule by evaluating the indicated integrals and sums. There are three types of hypotheses, so evaluating these integrals and sums must proceed in three steps. In these calculations only the multivariate Gaussian integrals may be evaluated in closed form. The remaining integrals must be evaluated numerically. Evaluating the multivariate Gaussian integrals for each of the three types of hypotheses is essentially identical. Consequently, the procedures needed will be demonstrated for the known targets and then the results will simply be given for the unknown and no-target hypotheses.

Evaluating The Integrals For The Known Targets

The posterior probability for the known target hypothesis is given by Eq. (67). The prior probability for the target hypothesis, $P(k|I)$, was assigned a uniform prior and because this term appears in all of the posterior probabilities its prior range cancels when these distributions are normalized. The prior probabilities for the range to the target, $P(R_c|I)$, and the orientation angle of the target, $P(\omega|I)$ are given by Eqs. (44,45). The prior probability for the dc offset was assigned a wide uniform prior and because this term also appears in every posterior probability exactly one time its prior range also cancels. The prior probabilities for the amplitudes, $P(B_l|I)$, are all given by Eq. (72). Last the likelihood function is given by Eq. (79). Gathering up these terms, the posterior probability for the known targets hypothesis is given by

$$\begin{aligned} P(k|DI) \propto & \int dB d\omega dR_c \\ & \times (2\pi\sigma_R^2)^{-\frac{1}{2}} \exp\left\{-\frac{[R_0 - R_c]^2}{2\sigma_R^2}\right\} \\ & \times (2\pi\sigma_\Omega^2)^{-\frac{1}{2}} \exp\left\{-\frac{[\Omega - \omega]^2}{2\sigma_\Omega^2}\right\} \qquad (1 \leq k \leq \ell - 2) \\ & \times \left(\frac{1}{B_{max}}\right)^{N_k} \\ & \times (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}[d_j - \sum_{l=0}^{N_k} B_l G(S_{kl}\cos(\phi_{kl} - \omega) - r_j + R_c)]^2\right\} \end{aligned} \tag{82}$$

where each of the terms has been intentionally separated so they may be more readily identified. Additionally, the notation should have been modified to indicate that σ_R^2, σ_Ω^2, σ, R_0, ω and B_{max} are known quantities. However, to compress the notation, these quantities have been incorporated into the general background information I.

There are $N_k + 3$ integrals that must be evaluated. Of these only the $N_k + 1$ amplitude integrals may be evaluated in closed form. These integrals are multivariate Gaussian integrals and any integral of this form may be evaluated in closed form. Designating the amplitude integrals as $p_B(R_c, \omega)$, the integrals that must be evaluated are given by

$$p_B(R_c,\omega) = \int dB \exp\left\{-\frac{1}{2\sigma^2}\left[d\cdot d - 2\sum_{l=0}^{N_k} B_l T_l + \sum_{l=0}^{N_k}\sum_{\eta=0}^{N_k} B_l B_\eta g_{l\eta}\right]\right\} \tag{83}$$

where

$$d\cdot d \equiv \sum_{j=1}^{N} d_j^2, \tag{84}$$

$$T_l \equiv \sum_{j=1}^{N} d_j G(S_{kl}\cos(\phi_{kl}-\omega) - r_j + R_c), \tag{85}$$

and

$$g_{l\eta} = \sum_{j=1}^{N} G(S_{kl}\cos(\phi_{kl}-\omega) - r_j + R_c) G(S_{k\eta}\cos(\phi_{k\eta}-\omega) - r_j + R_c). \tag{86}$$

There are a number of different ways to evaluate these integrals; one of the easiest to understand is to introduce a change of variables that makes the $g_{l\eta}$ matrix diagonal, then all integrals uncouple and each may be done separately. The new variables, $\{A_0 \cdots A_{N_k}\}$, are defined as

$$A_l = \sqrt{\lambda_l}\sum_{\eta=0}^{N_k} B_\eta e_{l\eta} \quad \text{and} \quad B_l = \sum_{\eta=0}^{N_k} \frac{A_\eta e_{\eta l}}{\sqrt{\lambda_\eta}} \tag{87}$$

where λ_η is the ηth eigenvalue of the $g_{l\eta}$ matrix and $e_{\eta l}$ is the lth component of its ηth eigenvector. The eigenvalues and eigenvectors have the property that

$$\sum_{\eta=0}^{N_k} g_{l\eta} e_{\ell\eta} = \lambda_\ell e_{\ell l} \tag{88}$$

from which the $p_B(R_c, \omega)$ integral may be rewritten as

$$p_B(R_c,\omega) = \int dA \lambda_0^{-\frac{1}{2}} \cdots \lambda_{N_k}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[d\cdot d - 2\sum_{l=0}^{N_k} A_l h_l + \sum_{l=0}^{N_k} A_l^2\right]\right\}, \tag{89}$$

where

$$h_l = \sum_{j=1}^{N} d_j H_l(r_j) \tag{90}$$

and

$$H_l(r_j) = \frac{1}{\sqrt{\lambda_l}} \sum_{\eta=0}^{N_k} e_{l\eta} G(S_{k\eta} \cos(\phi_{kl} - \omega) - r_j + R_c). \tag{91}$$

The model functions $H_l(r_j)$ are called orthonormal because they have the property

$$\sum_{j=1}^{N} H_l(r_j) H_\eta(r_j) = \delta_{l\eta} \tag{92}$$

where $\delta_{l\eta}$ is the Kronecker delta function.

This change of variables reduces the $p_B(R_c, \omega)$ integral to a series of independent Gaussian integrals. The integration limits are from zero to an upper bound. These limits are assumed so wide that the amount of probability contributed to the integral near the upper and lower bound is so small that the limits may be extended to plus and minus infinity and this extension will make only a negligible change in the results of the integral. Using this approximation one obtains

$$p_B(R_c, \omega) = (2\pi\sigma^2)^{\frac{N_k+1}{2}} \lambda_0^{-\frac{1}{2}} \cdots \lambda_{N_k}^{-\frac{1}{2}} \exp\left\{-\frac{d \cdot d - h \cdot h}{2\sigma^2}\right\}, \qquad (1 \le k \le \ell - 2) \tag{93}$$

where $h \cdot h$ is given by

$$h \cdot h = \sum_{l=0}^{N_k} h_l^2. \tag{94}$$

The quantity $h \cdot h$ plays the role of a sufficient statistic and summarizes all of the information in the data relevant to estimating the position and orientation angle of the target. Note that the sufficient statistic is a function of both R_c and ω even though this dependency has not been explicitly shown. Substituting $p_B(R_c, \omega)$ into the posterior probability for the known targets one obtains

$$\begin{aligned} P(k|DI) \quad &\propto \quad \frac{(2\pi\sigma^2)^{-\frac{N-N_k-1}{2}}}{2\pi\sigma_R\sigma_\Omega} \left(\frac{1}{B_{max}}\right)^{N_k} \\ &\times \quad \int d\omega dR_c \exp\left\{-\frac{[R_0 - R_c]^2}{2\sigma_R^2} - \frac{[\Omega - \omega]^2}{2\sigma_\Omega^2}\right\} \qquad (1 \le k \le \ell - 2) \\ &\times \quad \lambda_0^{-\frac{1}{2}} \cdots \lambda_{N_k}^{-\frac{1}{2}} \exp\left\{-\frac{d \cdot d - h \cdot h}{2\sigma^2}\right\}. \end{aligned} \tag{95}$$

The remaining two integrals must be evaluated numerically. In the numerical simulations, these integrals are approximated in a particularly simple way. Each integral is taken to be approximately the width of the integrand times its height. In this particular case this approximation is good enough because the data are extremely spiky. This results in an extraordinarily sharply peaked probability distribution. The widths are analogous to a prior penalty, and almost any values used for them will work (provided they are reasonable). Here reasonable means the widths must be within one or two orders of magnitude of the true values. Parameter estimates using probability theory as logic typically scale like one over root N, so the widths are easily set to the right order of magnitude.

Evaluating The Integrals For The Unknown Target

The process of evaluating the integrals for the unknown target hypothesis is essentially identical to what was done for the known target hypotheses. Consequently, only the results of these integrals are given. The posterior probability for the unknown target is given by

$$\begin{aligned} P(\ell-1|DI) &\propto \sum_{N_\nu=1}^{N} \int dBdS(N)^{-1}\,(N)^{-N_\nu}\left(\frac{1}{B_{max}}\right)^{N_\nu} \\ &\times (2\pi\sigma^2)^{-\frac{N-N_\nu-1}{2}}\lambda_0^{-\frac{1}{2}}\cdots\lambda_{N_\nu}^{-\frac{1}{2}}\exp\left\{-\frac{d\cdot d-h\cdot h}{2\sigma^2}\right\}, \end{aligned} \tag{96}$$

where the definitions of these quantities are analogous to those given in the preceding calculation. For example the sufficient statistic $h\cdot h$ is defined

$$h\cdot h=\sum_{l=0}^{N_\nu} h_l^2 \tag{97}$$

with

$$H_l(r_j)=\frac{1}{\sqrt{\lambda_l}}\sum_{\eta=0}^{N_\nu} e_{l\eta}G(S_{k\eta}-r_j) \tag{98}$$

and the eigenvalues and eigenvectors that appear in this calculation are formed from the interaction matrix associated with the unknown model function:

$$g_{l\eta}=\sum_{j=1}^{N} G(S_{kl}-r_j)G(S_{k\eta}-r_j). \tag{99}$$

For the known targets there was one sufficient statistic for each model, while here there is one for each value of the summation index N_ν. In principle, this is a long and tedious calculation. However, because of the spiked nature of the model function it is possible to implement this calculation using relatively simple approximations. In the numerical example two approximations were used: the sum was approximated by it largest term; while the integral was approximated by its height times its width. How these approximations worked in practice is the subject of the next Section.

Evaluating The Integrals For The No-target Model

The no-target model is particularly simple because the model contains only a single nuisance parameter. The posterior probability for the no-target model is given by

$$P(\ell|DI)\propto(N)^{-\frac{1}{2}}\left(\frac{1}{B_{max}}\right)(2\pi\sigma^2)^{-\frac{N-1}{2}}\exp\left\{-\frac{d\cdot d-N(\overline{d})^2}{2\sigma^2}\right\}, \tag{100}$$

where $\overline{d}$ is given by

$$\overline{d}=\frac{1}{N}\sum_{j=1}^{N} d_j. \tag{101}$$

With the completion of this integral the problem is formally completed. There are a number of integrals that must be evaluated numerically. In the case of the unknown, the entire calculation is so complicated that there is virtually no hope of implementing the exact calculation, and approximations must be used. However, unless one knows what one should aim for, it is hard to know how to make approximations, and this is one of the places where probability theory helps most. By telling one what to aim for, the problem is reduced to approximating the posterior probability to the best of one's ability. While making numerical approximations is difficult, it is less difficult than trying to guess the answer by intuition.

4 Numerical Methods

To demonstrate model selection, how to handle incomplete sets of hypotheses (the unknown hypothesis), and the feasibility of radar target identification, the identification calculations presented in this tutorial have been implemented in a numerical simulation. In this simulation there are three major routines: a data generation routine, an identification routine, and an output routine. In general terms the simulation is a loop. Each time through the loop a data set is generated, passed to the identification routines and the results of the simulation are written to an output file.

In this simulation there were 20 different hypotheses or targets; 18 known targets, one unknown, and one no-target hypothesis. The known target models were generated by a separate program and then used throughout the simulation. To do this the program used a random numbers generator to determine the angle and position of each scatter. The angular location of a scatter, ϕ_{kl}, was chosen to be between 0 and 2π, while radial location of a scatter, S_{kl}, was chosen to be between $-400 \leq R_c \leq 400$. There are 1024 data values, so the scatterers were chosen so that they always fit within the range window of the radar.

The data generation routine chooses one of the 20 targets at random. When it chooses the unknown target, the unknown is generated in a manner analogous to the known hypotheses. A uniform random number generator was used to randomly position between 3 and 10 scatterers within a range window of $-400 \leq R_c \leq 400$. Similarly, when the no-target model was chosen no scatterers were generated – only noise was placed in the simulated data. The amplitudes of the scatterers were set randomly. However, their amplitudes were scaled so that the *mean amplitude* to root-mean-square noise standard deviation was 20. The twenty target library is shown in Fig. 4 for one setting of the amplitudes, orientation angles, and positions of the targets.

In this simulation the real target identification problem was mimicked as closely as possible. To do this the data were generated and processed in a way that mimicked the effects encountered on a real radar. On a real radar, the radar will establish a track on a target, and only after the track has been established will the identification be attempted. In the process of tracking the target the radar will infer the vector position and orientation angle of the target. This information is available to the identification routines in the form of prior probabilities. Additionally, the amplitudes of the scatterers are extremely sensitive functions of the orientation angle of the target, changing by more than an order of magnitude for a change of only 0.1 degrees. For all practical purposes, this means the amplitude of a scatterer is completely unpredictable from one look to the next. To illustrate these effects, the same library targets (with a new unknown, and no-target model) were generated a second time. These targets are displayed in Fig. 5. While the positions of the

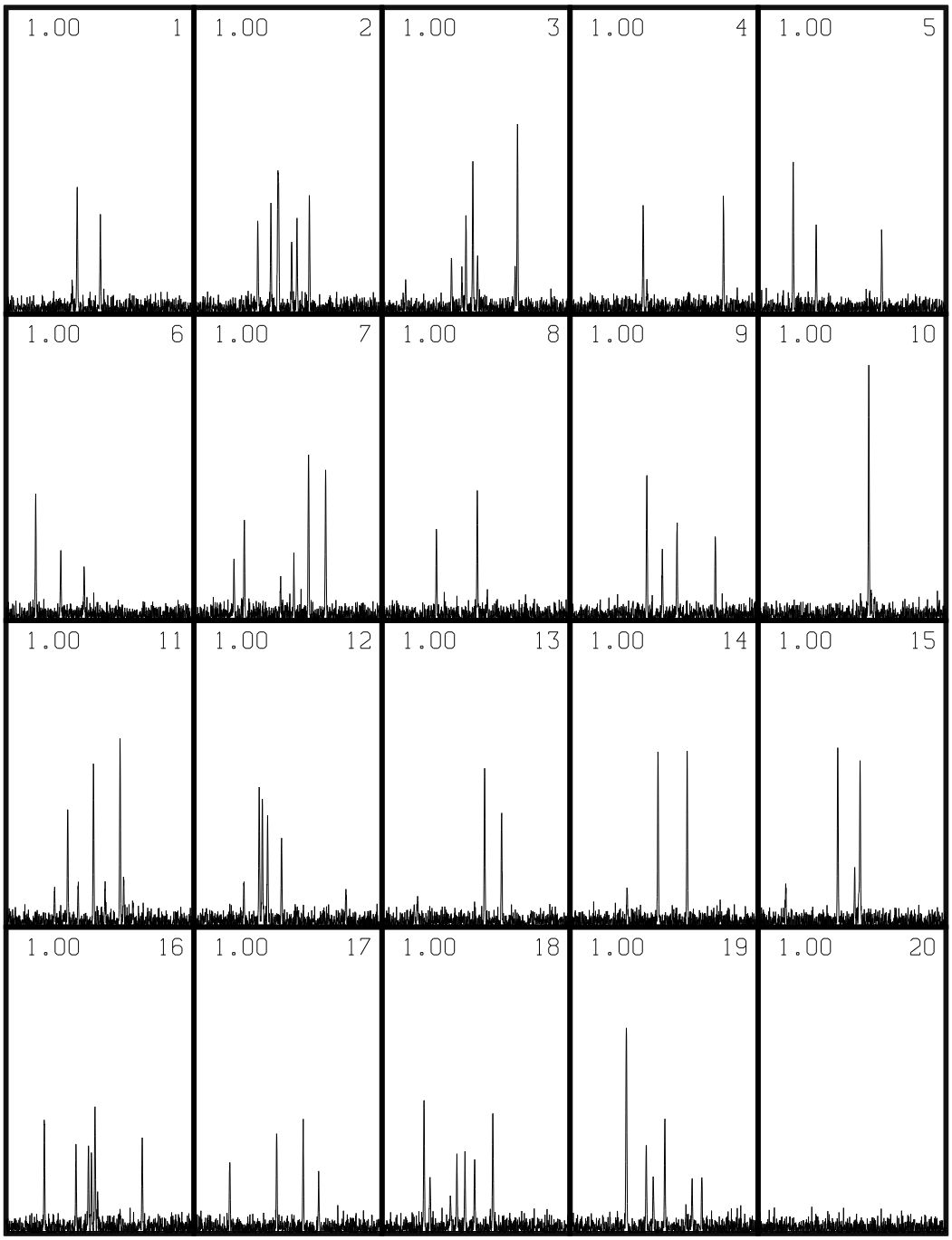

Fig. 4. The library used in the simulation consisted of 20 different targets. The first 18 of these are the known targets, corresponding to various types of aircraft. The amplitudes, location, and orientation angle of each known target is chosen randomly. Target 19 is the unknown. The locations, amplitudes, and number of scatterers for the this target are chosen randomly. Target 20 is the no-target model and contains no scatterers.

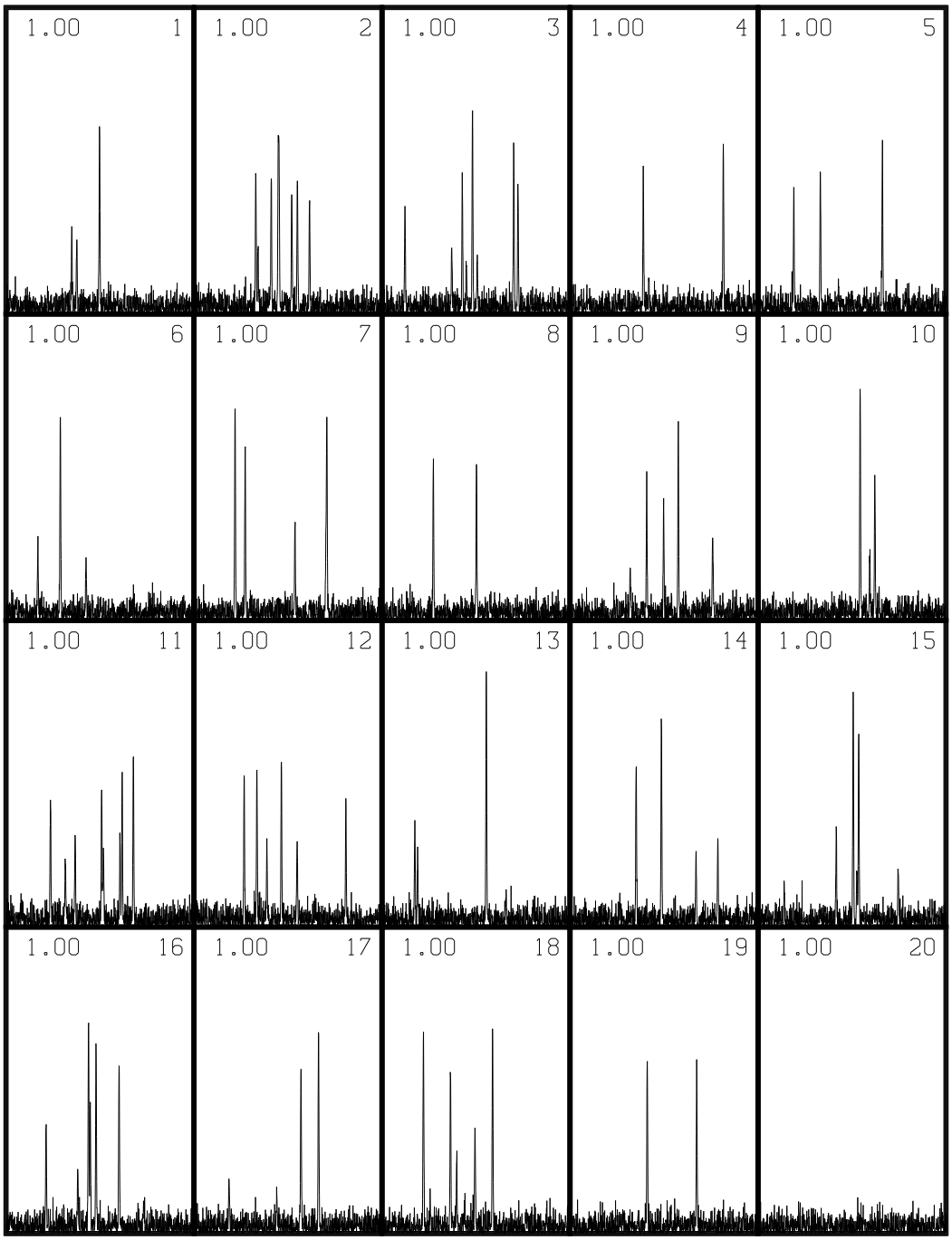

Fig. 5. The amplitudes of each scatterer vary rapidly as a function of the orientation angle of the target. Here is the same 20 targets as they might look at slightly different orientations angle. Try comparing these targets to those shown in Fig. 4 to see if you can tell that they are the same.

scatterers in Fig. 4 and Fig. 5 nearly overlap, the amplitudes are completely different. This effect is so striking that the author has difficulty telling that these are the same targets. The probability assigned to the *true* target is shown to three decimal places in the upper left hand corner of these figures. Note that in both Fig. 4 and Fig. 5 the correct target was identified to a probability of one to three decimal places in every case.

The simulation was run on 1000 simulated data sets, taking about 3 seconds for each simulation on an SGI Indigo. The first 20 simulated data sets are shown in Fig. 6. The full output from the simulation is shown in Table 1. This output consists of both a summary and detailed outputs. The summary output tells one the simulation number, i.e., 1, 2, 3 etc., the true target number, its probability, and the signal-to-noise ratio. The detailed output contains the unnormalized base 10 logarithm of the probability for each target. In the 1000 simulations the correct target was identified 999 times; there was only a single mis-identification. When the mis-identification was investigated, it was found that the generated target had most of its scatterers buried in the noise while two of them had exceptionally high signal-to-noise ratio. Under these conditions the unknown target *is* a better representation of the data than the true model. Thus the unknown target was understandably identified.

Table 1 illustrates very strongly why the unknown target hypothesis works. To understand it, look at the first simulation. The true target is number 5. The base 10 logarithm of its probability is 1991.3. Now look at the log probabilities for the other targets for this first simulation. The target with the second highest probability was the unknown, having a log probability of 1983.7, roughly seven orders of magnitude down from the true target. The target with the third highest probability is target 17, it has a log probability of 1456.0, more than 400 orders of magnitude down. Next, look at a second simulation, say simulation number 8. The true target is number 20, the no-target model, it's log probability is 145.49. The second highest log probability is again the unknown coming in at 139.83. Now examine all of the simulations in the table except simulation number 3. The unknown hypothesis is the second highest probability in every case! To understand this, note that the unknown target essentially fits all of the systematic detail in the data; its likelihood function is essentially identical to the likelihood of the true target (assuming the true target hypotheses is in the library) . But the unknown has many more parameters. In probability theory as logic these extra parameters carry a penalty in the form of the prior probabilities. The priors range for both the location and number of scatterers was $1/N$. If there were 3 scatterers on the target the unknown would have a prior penalty of $1/N^4$. The number of data values, N, was 1024 so the prior penalizes the unknown by a factor of approximately 10^{12}. This penalty is so large, that unless the true target is *not* present, the prior eliminates the unknown target from consideration. Now examine simulation number 3. The true target is the unknown. There is no known target present to prevent the unknown target from being identified.

5 Summary And Conclusions

To use probability theory as logic, one must relate the hypothesis of interest to the available evidence. This process is one of model building. While building the model one must state exactly what the hypotheses are and how they are related to the available evidence. In the case of radar target identification, this process has forced the radar target identification community to state exactly what is meant by the known, unknown and no-target

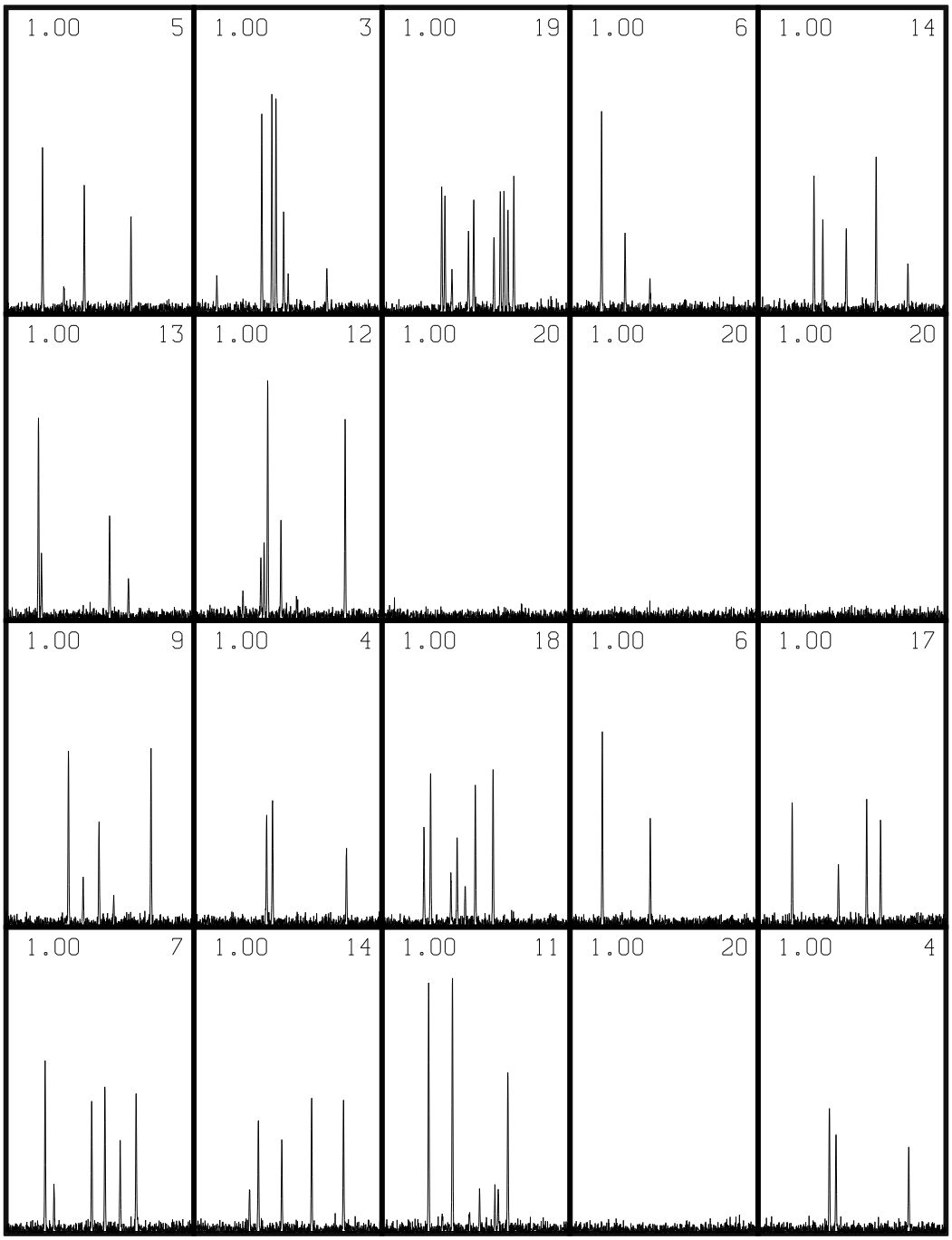

Fig. 6. This is the data generated in the first 20 simulations. The number in the upper right hand corner is the number of the true target. The number in the left hand corner is the probability assigned to this target. Note that the identification was perfect on these 20 targets.

Table 1: When Is The Unknown Target Identified?

```
1 True Target: 5, Its probability is:1.00; S/N=20.0}
  1  342.21      2  823.30      3  819.98      4  828.76      5  1991.3
  6  820.01      7  338.35      8  341.15      9  824.52     10  340.43
 11  337.96     12  338.88     13  1194.3     14  353.80     15  339.31
 16  336.93     17  1456.0     18  338.85     19  1983.7     20  346.57

2 True Target: 3, Its probability is:1.00; S/N=20.0}
  1  1758.0      2  1981.4      3  5028.1      4  1982.5      5  2026.9
  6  1983.0      7  655.49      8  627.86      9  2000.7     10  615.61
 11  889.01     12  1753.9     13  615.23     14  1754.5     15  1990.0
 16  2308.5     17  2031.3     18  3189.9     19  5015.5     20  615.86

}
3 True Target: Unknown(19), Its probability is:1.00; S/N=20.0}
  1  1475.6      2  1631.9      3  1780.4      4  752.65      5  1475.5
  6  751.85      7  1443.1      8  749.91      9  1331.7     10  1340.6
 11  1605.3     12  1298.8     13  1197.2     14  1881.0     15  752.00
 16  1244.8     17  1039.6     18  900.19     19  4168.9     20  750.25

4 True Target: 6, its probability is:1.00; S/N=20.0
  1  289.23      2  284.32      3  310.25      4  315.44      5  314.46
  6  1805.6      7  285.79      8  463.71      9  312.38     10  287.84
 11  285.03     12  286.36     13  1584.2     14  465.75     15  312.17
 16  285.62     17  309.19     18  286.62     19  1799.6     20  295.45

5 True Target: 14, its probability is:1.00; S/N=20.0
  1  615.56      2  596.31      3  599.18      4  391.18      5  967.78
  6  390.95      7  387.89      8  390.63      9  965.99     10  1158.6
 11  1141.0     12  822.83     13  1155.8     14  2287.1     15  389.47
 16  600.15     17  391.23     18  389.08     19  2277.0     20  395.64

6 True Target: 13, its probability is:1.00; S/N=20.0
  1  331.36      2  662.09      3  327.42      4  331.51      5  436.72
  6  1627.0      7  329.55      8  331.98      9  329.65     10  667.83
 11  327.76     12  663.45     13  2141.4     14  372.26     15  665.66
 16  327.18     17  2010.7     18  328.61     19  2133.4     20  337.16

7 True Target: 12, its probability is:1.00; S/N=20.0
  1  633.44      2  630.05      3  633.77      4  2359.5      5  539.65
  6  538.65      7  807.69      8  539.81      9  628.92     10  539.12
 11  2632.3     12  4219.2     13  540.15     14  644.69     15  539.97
 16  902.78     17  540.12     18  2448.6     19  4203.8     20  542.83
```

```
8 True Target: No Target(20), its probability is:1.00; S/N=20.0
  1  138.97      2  132.39      3  133.21      4  139.01      5  137.88
  6  139.23      7  134.28      8  138.86      9  136.85     10  138.45
 11  133.77     12  134.32     13  137.98     14  136.93     15  135.92
 16  133.91     17  138.39     18  134.35     19  139.83     20  145.49

9 True Target: No Target(20), its probability is:0.99999; S/N=20.0
  1  138.26      2  131.78      3  132.10      4  137.90      5  137.16
  6  137.82      7  133.13      8  137.71      9  136.05     10  136.61
 11  132.07     12  133.46     13  136.55     14  135.67     15  134.67
 16  131.81     17  137.37     18  132.92     19  138.88     20  144.26

10 True Target: No Target(20), its probability is:0.99999; S/N=20.0
  1  126.72      2  120.33      3  122.27      4  126.99      5  126.08
  6  127.59      7  122.20      8  126.86      9  124.99     10  126.00
 11  121.17     12  122.16     13  126.27     14  124.92     15  123.54
 16  122.42     17  125.93     18  122.54     19  127.77     20  133.32

11 True Target: 9, its probability is:1.00; S/N=20.0
  1  1358.1      2  1409.0      3  734.61      4  436.37      5  1355.3
  6  436.20      7  383.71      8  688.60      9  2703.1     10  394.53
 11  1349.7     12  1348.7     13  385.26     14  677.26     15  383.00
 16  677.75     17  397.55     18  682.65     19  2693.2     20  389.72

12 True Target: 4, its probability is:1.00; S/N=20.0
  1  272.47      2  756.10      3  761.39      4  1321.6      5  765.64
  6  767.68      7  267.81      8  272.65      9  764.52     10  270.85
 11  626.02     12  782.11     13  271.55     14  270.32     15  762.95
 16  625.45     17  757.66     18  627.50     19  1315.6     20  277.74

13 True Target: 18, its probability is:1.00; S/N=20.0
  1  1153.0      2  1156.6      3  613.99      4  769.26      5  561.42
  6  530.28      7  801.18      8  1766.5      9  1103.9     10  1261.6
 11  1240.7     12  1462.8     13  530.03     14  1828.7     15  528.62
 16  1188.0     17  1257.4     18  3248.9     19  3235.4     20  531.07

14 True Target: 6, its probability is:1.00; S/N=20.0
  1  267.83      2  261.75      3  592.98      4  589.68      5  593.95
  6  1789.7      7  265.15      8  268.20      9  595.83     10  268.10
 11  262.97     12  263.90     13  1441.5     14  266.55     15  594.71
 16  262.69     17  1774.2     18  265.63     19  1783.7     20  273.84
```

```
15 True Target: 17, its probability is:1.00; S/N=20.0
  1  331.89      2  327.58      3  427.78      4  332.51      5  1118.6
  6  430.92      7  329.53      8  799.51      9  801.72     10  799.39
 11  328.05     12  328.45     13  1130.8     14  674.66     15  430.44
 16  429.09     17  1730.6     18  853.47     19  1722.6     20  337.69

16 True Target: 7, its probability is:1.00; S/N=20.0
  1  557.89      2  1289.4      3  1138.5      4  560.31      5  559.63
  6  558.83      7  3525.4      8  598.40      9  560.97     10  758.86
 11  1816.4     12  1134.8     13  557.64     14  557.65     15  1191.6
 16  1952.9     17  557.83     18  1466.7     19  3511.5     20  561.05

17 True Target: 14, its probability is:1.00; S/N=20.0
  1  813.47      2  657.30      3  656.92      4  420.90      5  463.77
  6  420.31      7  1186.4      8  420.59      9  420.18     10  967.40
 11  1187.8     12  1350.8     13  962.76     14  2240.8     15  419.36
 16  658.86     17  420.70     18  418.35     19  2231.2     20  425.04

18 True Target: 11, its probability is:1.00; S/N=20.0
  1  639.06      2  675.18      3  641.23      4  2718.8      5  1425.5
  6  641.82      7  666.94      8  641.69      9  642.31     10  671.72
 11  5732.9     12  4731.0     13  1432.4     14  1430.0     15  663.53
 16  2759.3     17  669.44     18  2620.3     19  5704.7     20  642.51

19 True Target: No Target(20), its probability is:1.00; S/N=20.0
  1  130.83      2  123.28      3  124.80      4  130.47      5  129.77
  6  130.43      7  125.82      8  130.41      9  127.92     10  129.65
 11  125.02     12  125.97     13  129.78     14  128.61     15  127.31
 16  125.08     17  129.73     18  125.74     19  131.19     20  136.93

20 True Target: 4, its probability is:0.99997; S/N=20.0
  1  272.19      2  585.49      3  588.66      4  1329.5      5  588.29
  6  594.21      7  267.63      8  272.56      9  590.34     10  271.73
 11  775.71     12  993.07     13  270.58     14  270.30     15  589.80
 16  267.73     17  587.64     18  268.50     19  1325.0     20  277.63
```

Table 1 illustrates how the unknown target is identified. This is the detailed output from the first 20 simulations. The first line of each entry identifies the true target and its probability. Lines 2 thru 5 for each entry are the base 10 logarithm of the probability for each target. Target 19 is the unknown target. To see how, when and why the unknown is correctly identified, browse through this table and compare the log probability for the unknown to that of the true target. The log probability for the unknown is always less than the true target. However, it is always greater than any of the other targets. So the unknown is identified whenever the true target is not in the list of library targets.

hypotheses. In probability theory as logic there is *no such thing* as nonparametric statistics. Typically when this term is used, it is used to mean that the number of hypotheses grows very large. That is to say, the models are so general they can fit virtually any data. But this is not nonparametric statistics, indeed it is exactly the opposite: there are many more parameters than data. However, there are a few people who use the term to mean literally there are no models. Typically, the statistics advocated by these people cannot be derived from a correct application of the rules of probability theory and, at best, their results are intuitive and ad hoc.

Probability theory computes the probabilities for hypotheses. It computes the probability for parameters only in the sense that the parameter indexes a well defined hypothesis. Similarly, it test models only in the sense that models are statements of hypotheses. Thus there is no essential difference between model selection and parameter estimation. The differences are conceptual, not theoretical. These conceptual differences manifest themselves primarily in the prior probabilities. In parameter estimation it is often convenient and harmless to use improper priors (an improper prior is a function that is used as a prior probability that is not normalizable). It is convenient because improper priors often simplify the mathematics considerably, and harmless because the infinities so introduced cancel when the probabilities are normalized. Strictly speaking improper priors are not probabilities at all; rather they are the limit of a sequence of proper priors in the limit of infinite uncertainty in a hypothesis. As a limit, it must always be approached from well-defined finite mathematics to ensure one obtains a well behaved result. Use of an improper prior directly can and will result in disaster in model selection problems because the infinities don't generally cancel. For more on this point see Jaynes [11]. In parameter estimation, when using a uniform prior, the prior ranges cancel when the distribution is normalized. However, in model selection these prior ranges may or may not cancel. In the numerical simulation described in this tutorial, the prior range for the constant dc offset and which target was present canceled. The remaining prior ranges did not cancel, and so affect the results. These prior ranges essentially set the scale against which different models with differing parameterizations are compared. So it is vitally important that one think carefully about these quantities and set them based on the information one actually has.

The probability for a hypothesis C is computed conditional on the evidence $E_1 \cdots E_n$. This probability is given by $P(C|E_1 \cdots E_n)$. Every person who consistently follows the rules of probability theory will be lead to assign exactly the same probabilities conditional on that evidence. These probabilities are all of the form of prior probabilities. The distinction between data, strong prior information, weak prior information, and no prior information (which strictly speaking cannot exist in real problems) is purely artificial. Evidence is evidence and it is all used to assign prior probabilities! The principle of maximum entropy was used here to assign these priors because it assigns priors that are consistent with that evidence while remaining maximally uninformative. That is to say, the probabilities do not depend on things one does not know. This is particularly important when assigning the prior probability for the noise because it allows one to assign probabilities that depend only on what one actually knows about the true errors in the data and it renders the underlying sampling distribution of the noise completely irrelevant.

The calculations indicated by probability theory are often much too complicated to implement exactly. However, knowing what should be done enables one to reduce the

problem from one of guessing the answer to one of numerical approximation. This is a tremendous simplification that often leads to simple numerical algorithms which, although not exact, capture the essence of the probability theory calculation and enable one to solve problems that would otherwise prove impossible.

ACKNOWLEDGMENTS. The author would like to thank Dr. C. R. Smith, and Dr. Jeffrey J. Neil for their valuable comments on preliminary versions of this paper. In particular I would like to extend my deepest thanks to Dr. Tom Loredo for his comments on Section 2. Without his assistance Section 2 would have been a mere shadow of the final version. The encouragement of Professor J. J. H. Ackerman is greatly appreciated. Additionally, the author would like to acknowledge a vary large debt to Professor E. T. Jaynes for his help and guidance over the years. Last, this work was supported by the U. S. Army through the Scientific Services Program.

References

[1] William of Ockham, *ca* 1340.

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

[3] Jaynes, E. T., *JASA*, Sept. 1979, p. 740, review of "Inference, Methods, and Decision: Towards a Bayesian Philosophy of Science." by R. D. Rosenkrantz, D. Reidel Publishing Co., Boston.

[4] Gull, S. F., "Bayesian Inductive Inference and Maximum Entropy," in *Maximum Entropy and Bayesian Methods in Science and Engineering"* **1,** pp. 53-75, G. J. Erickson and C. R. Smith *Eds.,* Kluwer Academic Publishers, Dordrecht the Netherlands, 1988.

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

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

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

[8] Bretthorst, G. Larry, "Bayesian Analysis III: Spectral Analysis," *J. Magn. Reson.,* **88,** pp. 571-595 (1990).

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

[10] Zellner, A., *An Introduction to Bayesian Inference in Econometrics,* John Wiley and Sons, New York, 1971. Second edition (1987); R. E. Krieger Pub. Co., Malabar, Florida.

[11] Jaynes, E. T., "Probability Theory – The Logic of Science," in preparation. Copies of this TeX manuscript are available by anonymous FTP from "bayes.wustl.edu"

[12] 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.

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

[14] 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.

[15] 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 date 1814.

[16] Jaynes, E. T., "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics,* SSC-4, pp. 227-241 (1968); reprinted in [20].

[17] Shore J. E., R. W. Johnson, *IEEE Trans. on Information Theory,* **IT-26,** No. 1, pp. 26-37, 1981.

[18] Shore J. E., R. W. Johnson, *IEEE Trans. on Information Theory,* **IT-27,** No. 4, pp. 472-482, 1980.

[19] 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 [20].

[20] 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.

[21] 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 [20].

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

[23] Jaynes, E. T., 1989, "The Theory of Radar Target Discrimination," in MICOM Technical Report RD-AS-91-6, Feb. 1991.

[24] Bretthorst, G. Larry, "Radar Target Identification The Information Processing Aspects," Contract number DAAL03-92-c-0034, TCN 92060.