


Bayesian Inductive Inference Maximum Entropy & Neutron Scattering

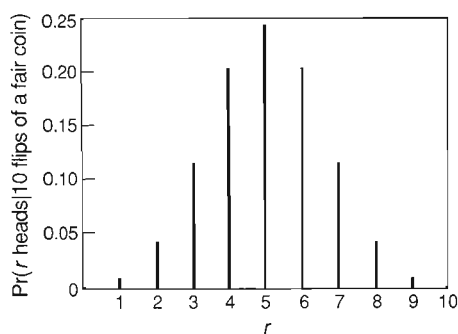
by Devinder Singh Sivia





The result of this experiment was inconclusive, so we had to use statistics. Such oft-heard statements reflect the “cookbook” approach to statistics that we are taught as undergraduates. Not satisfied with the maze of seemingly ad hoc statistical tests, many of us become inclined to avoid the subject as much as possible.

Fortunately, statistics does not have to be like that! A more logical and unified approach



DIRECT PROBABILITIES

Fig. 1. What is the probability of getting r heads in 10 flips of a fair coin, $\text{Pr}(r \text{ heads} | 10 \text{ flips of a fair coin})$? Deductive logic tells us that the probability in question is given by $N_{r \text{ heads}}/N$, where N is the number of all possible sequences of heads and tails that the 10 flips can generate and $N_{r \text{ heads}}$ is the number of those sequences that contain r heads (in any order). To obtain a numerical answer, note that $N = 2^{10}$ and that $N_{r \text{ heads}} = 10!/(10-r)!r!$. Thus $\text{Pr}(r \text{ heads} | 10 \text{ flips of a fair coin}) = [10!/(10-r)!r!]/2^{10}$.

to the whole subject is provided by the probability formulations of Bayes and Laplace. Bayes' ideas (published in 1763) were used very successfully by Laplace (1812) but were then allegedly discredited and largely forgotten until they were rediscovered by Jeffreys (1939). In more recent times they have been expounded by Jaynes and others. Here we present an introductory glimpse of the Bayesian approach. We then illustrate how Bayesian ideas, and developments such as the maximum entropy method, are affecting data analysis and thoughts on instrument design at the Manuel Lujan, Jr. Neutron Scattering Center (LANSCE).

Everyday games of chance are governed by deductive logic. For example, if we are told that a fair coin is flipped ten times, we can deduce accurately the chances that all ten flips produced heads, or that nine produced heads and one produced tails, ..., or that all ten flips produced tails (Fig. 1). Turning to neutron scattering, let's suppose we know the scattering law for a particular sample and the geometry of the diffractometer, the efficiencies of the detectors, and so on. Then we can predict the chances of observing a certain number of neutron counts in any given detector. These examples are in the realm of deductive logic, or pure mathematics: Given the rules of a "game," we can predict the chances of various outcomes.

Most scientists, however, are concerned with the more difficult inverse problem. Given that a coin of unknown origin was tossed ten times and the result was seven heads, was it a fair coin or a weighted one? Further, what is the best estimate of the bias-weighting of the coin and what is the confidence in the prediction? If we are now given more data on the coin, how should we incorporate the new information and how do our prediction and confidence level change? This type of problem is in the realm of inductive logic, plausible reasoning, or inference: Having seen the outcome of several "moves" in a game, we want to infer the rules governing that game. Returning to neutron scattering, let's suppose we have recorded so many neutron counts in various detectors and wish to infer the scattering law for the sample. Like all problems in inductive logic, this problem has no clear-cut answer. The most we can hope to do is make the "best" inference based on both the experimental evidence and any prior knowledge we have at hand, reserving the right to revise our position if new information comes to light. Around 500 B.C. Herodotus said much the same thing: "A decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it was the best one to make; and a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences."

Bayes' Theorem

Bayes' theorem, which was actually written down in its present-day form by Laplace and not Bayes, is the cornerstone of scientific inference. It provides the bridge between the inductive logic we require and the deductive logic we know how to use. Its status is somewhat akin to the position of Newton's second law of motion in mechanics: seemingly tame and innocuous, but powerful enough to analyze a wide variety of problems when the relevant details and assumptions are given. In mechanics we may be taught that $s = \frac{1}{2}gt^2$ is the relationship between the vertical distance s that a body falls under a gravitational field g after a time t when released from rest at $t = 0$. We may also be told that the speed of sound v through a gas with pressure P and density ρ is given by $v^2 = P/\rho$. Although these two formulae look quite different and apply to different situations, it is satisfying to know that both of them are derived from the same physical law: Force is equal to the rate of change of momentum. Similarly the Bayesian approach to probability and statistics provides the logical foundation for the conventional teaching of statistics we are given as undergraduates. A Bayesian analysis often leads us to use the same procedure as advocated by the "cookbook" school of statistics, but it forces us to state clearly the

assumptions (usually forgotten) made in going from the fundamental rule for inductive inference (Bayes' theorem) to the particular statistical prescription we use.

But what is Bayes' theorem? Simply stated, it says that the conditional probability of A (being true) given B (is true), written as $\Pr(A|B)$, is proportional to the conditional probability of B given A times the probability of A :

$$\Pr(A|B) \propto \Pr(B|A) \times \Pr(A). \quad (1)$$

Bayes' theorem is easy to prove for problems in which A and B are "macroscopic" events that can be realized in a large number of equally probable "microscopic" ways. In such problems the probability of an event is the number of ways in which the event can occur divided by the total number of possibilities. For example, suppose the space of "microscopic" possibilities is all the possible sequences of heads and tails that can occur if a fair coin is flipped ten times. Since the coin is fair, each of the possible sequences is equally probable. "Macroscopic" event A might then be the event that the total number of heads was less than four, and B might be the event that a head was obtained on the third and seventh tosses. Figure 2 shows a schematic representation of the space of all "microscopic" possibilities and the portions of that space occupied by realizations of event A and event B . Now, let N be the total number of possibilities, N_A be the number of possibilities resulting in event A , N_B be the number of possibilities resulting in event B , and N_{AB} be the number of possibilities resulting in both event A and event B . Then the probabilities of the various outcomes of interest become

$$\Pr(A) = N_A/N, \quad \Pr(B) = N_B/N, \quad \Pr(A|B) = N_{AB}/N_B, \quad \text{and} \quad \Pr(B|A) = N_{AB}/N_A.$$

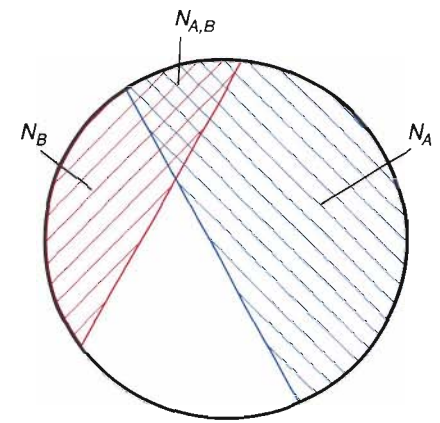
We can then write the probability of both A and B occurring, $\Pr(A, B)$, in two different ways:

$$\Pr(A, B) = N_{AB}/N = \Pr(A|B) \times \Pr(B) = \Pr(B|A) \times \Pr(A). \quad (2)$$

Bayes' theorem, as stated in Eq. 1, follows immediately from the two expressions for $\Pr(A, B)$ in Eq. 2, provided we associate $1/\Pr(B)$ in Eq. 2 with the proportionality constant in Eq. 1.

Although this proof is simple, the full implications of Bayes' theorem do not become apparent until we discover that the theorem applies equally well to cases in which A and B are any arbitrary propositions and the probabilities assigned to them represent merely our belief in the truths (or otherwise) of the propositions. This remarkable generalization, which is certainly not obvious, was proved by Cox (1946) while he was considering the rules necessary for logical and consistent reasoning.

Suppose we have a set of propositions. For example, a : It will rain tomorrow; b : King Harold died by being hit in the eye with an arrow during the battle of Hastings in 1066 A.D.; c : This is a fair coin; d : This coin is twice as likely to come up heads as tails; and so on. The minimum requirement for expressing our relative beliefs in the truth of the various propositions in a consistent fashion is that we rank them in a transitive manner. That is to say, if we believe proposition a more than b and b more than c , then we necessarily believe a more than c . Such a transitive ranking can easily be obtained by assigning a real number to each of the propositions in a manner so that the larger the numerical value associated with a proposition, the more we believe it. Cox went on to put forward two more axioms for logical, consistent reasoning: (1) If we first specify our degree of belief that A is true and then specify how much we believe B is true given that A is true, then we have implicitly defined our degree of belief for both A and B being true; and (2) If we specify how much we believe that A is true, then we have implicitly specified how much we believe that A is false. Cox showed that if we accept these remarkably mild desiderata, then

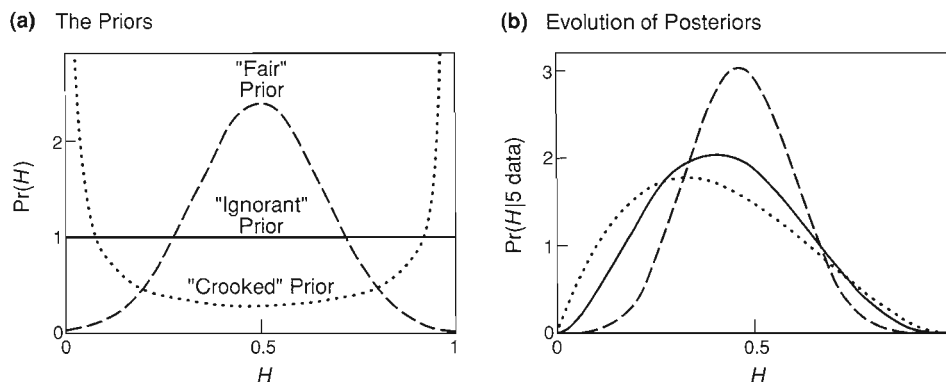


SAMPLE SPACE AND PROBABILITIES

Fig. 2. The sample space occupied by all N equally probable microscopic possibilities is depicted schematically here as a circle of area N . The microscopic possibilities result in various macroscopic events, such as A and B . The number of possibilities that result in A and the number of possibilities that result in B are represented by portions of the circle with areas N_A and N_B (hatched regions). The probability of A , $\Pr(A)$, is given by the fraction N_A/N ; similarly, $\Pr(B)$ is given by the fraction N_B/N . The probability of A and B , $\Pr(A, B)$, is given by $N_{A,B}/N$, where $N_{A,B}$, represented as an area of overlap between N_A and N_B , is the number of possibilities that result in both A and B .

INDIRECT PROBABILITIES: THE BIAS WEIGHTING OF A COIN

Fig. 3. (a) Shown here are three of the prior probability distributions that might be assigned to H , the bias weighting of a coin: the "ignorant" (or uniform) prior, which reflects the belief that all values of H ($0 \leq H \leq 1$) are equally probable; a "fair" prior, which reflects a belief that the coin is likely to have both a head and a tail and to be unbiased, or, in other words, a belief that the most likely value of H is 0.5; and a "crooked" prior, which reflects a belief that the coin is likely to be double-headed or double-tailed, or a belief that the most likely values of H are 1 or 0. The series of graphs in (b) shows how the posterior probability distributions corresponding to the priors in (a) evolve as the number of data increases. The data were generated by using a random-number generator in a computer and a value of 0.2 for the bias weighting. Note that, as the number of data increases, all the posteriors converge to a delta function centered at $H = 0.2$. In other words, as the experimental evidence increases, the assumptions embodied in the priors have less effect on our estimate of H .



there must be a mapping that transforms the real numbers we have associated with the various propositions (to express our beliefs in them) to another set of positive real numbers that obeys the usual rules of probability theory:

$$\Pr(A, B) = \Pr(A|B) \times \Pr(B) \text{ and } \Pr(A) + \Pr(\bar{A}) = 1,$$

where \bar{A} represents the proposition that A is false. In other words, any method of logical and consistent reasoning (no matter what the context) must be equivalent to the use of ordinary probability theory, where the probabilities represent our beliefs or state of knowledge about various propositions or hypotheses in the Bayes-Laplace-Jeffreys sense.

Bayes' theorem itself is just a simple corollary of these rules, but what does it really mean and why is it so powerful? Let us return to the coin-flipping problem as a concrete but simple example. Again we are told that a coin was flipped n times and came up heads r times, but we don't know whether the coin was fair. Our problem is to *infer* the coin's bias-weighting for heads, call it H . We will say that $H = 0$ represents a double-tailed coin (that is, a coin such that a head never appears), $H = 0.5$ represents a fair coin (that is, a coin such that its head is likely to come up as often as its tail), $H = 1$ represents a double-headed coin, and all other values of H (between 0 and 1) correspond to some intermediate bias-weighting.

To carry out the inference, we need to specify our beliefs in the set of propositions that, given the data, the value of H lies in a narrow range between h and $h + \delta h$, where h can take on values between 0 and 1. In terms of a probability distribution for H , $\Pr(H = h|\{\text{data}\})$, or simply $\Pr(H|\{\text{data}\})$, we write

$$\lim_{\delta h \rightarrow 0} \Pr(h \leq H \leq h + \delta h|\{\text{data}\}) = \Pr(H|\{\text{data}\})dh.$$

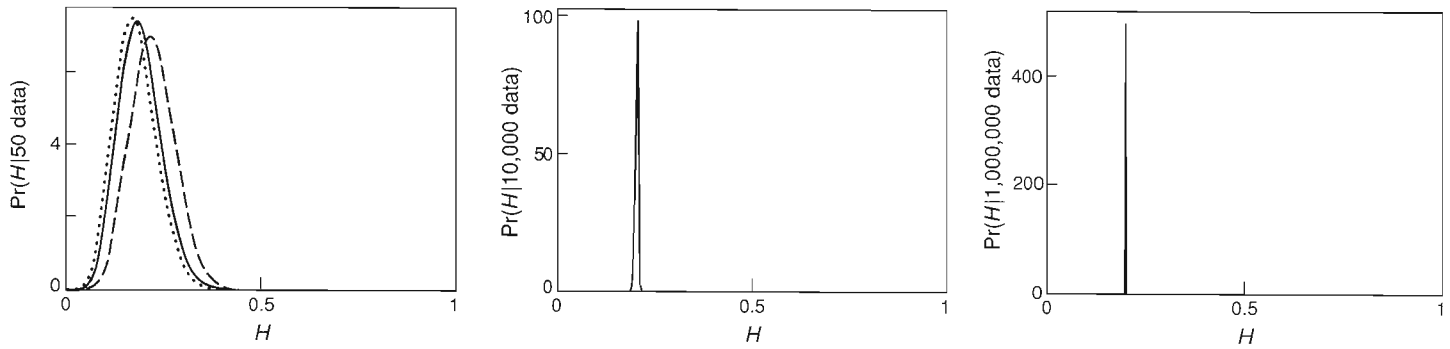
Thus $\Pr(H|\{\text{data}\})$, known as the posterior probability distribution (or simply the posterior), represents our state of knowledge about the bias-weighting for heads in light of the data. The value of h at which the posterior is a maximum gives our best estimate of the bias-weighting, and the spread of the posterior about the maximum gives our confidence in that estimate. If the posterior is sharply peaked, we are sure about our estimate; if it is broad, we are fairly uncertain about the true value of H .

In order to determine $\Pr(H|\{\text{data}\})$, we need to use Bayes' theorem,

$$\Pr(H|\{\text{data}\}) \propto \Pr(\{\text{data}\}|H) \times \Pr(H),$$

which relates the posterior to two other probability distributions, one of which can be "calculated" from the data and the other "guessed."

The probability distribution $\Pr(H = h)$, or simply $\Pr(H)$, which also is defined for $0 \leq h \leq 1$, represents our state of knowledge about the value of H *before* we are given the data. It is thus called the prior probability distribution (or simply the



prior). In the coin-tossing problem, if we are completely ignorant about the coin, we would assign a uniform prior, $\Pr(H) = \text{constant} = 1$ for all values of h between 0 and 1, to indicate that a priori all possible values of H are equally probable. If we do have other prior information, perhaps the results of previous data, then this information should be reflected in the nonuniform character of $\Pr(H)$. (Actually our statement of Bayes' theorem should read $\Pr(H|\{\text{data}\}, I) \propto \Pr(\{\text{data}\}|H, I) \times \Pr(H|I)$, where I represents other prior information or prior assumptions.) Figure 3a shows three possible assignments for $\Pr(H)$, each reflecting a different assumption about the coin: the uniform, or ignorant, prior; a prior that assumes the coin is most likely to be double-headed or double-tailed; and a prior that assumes the coin has a head and a tail and is probably fair.

Having specified our prior, we need now to consider the other probability distribution in Bayes' theorem, $\Pr(\{\text{data}\}|H)$, which reflects the nature of the "experiment." This probability distribution can be computed because it involves deductive logic. It is called the *likelihood* function because it tells us how likely it is that we would have obtained the data that we did if we had been given the value of H . For our problem we are told that a coin was flipped n times and came up heads r times. If we assume that the data are independent (that is, the outcome of one flip did not affect the result of another) and that the bias-weighting is H , then the likelihood function is simply a binomial distribution:

$$\Pr(\{\text{data}\}|H) = {}^nC_r \times H^r \times (1-H)^{n-r},$$

where ${}^nC_r = n!/r!(n-r)!$ is the number of ways of picking r objects (independent of order) from a choice of n . (Figure 1 shows such a binomial distribution.)

Multiplying $\Pr(H)$ and $\Pr(\{\text{data}\}|H)$, we obtain the posterior $\Pr(H|\{\text{data}\})$, which summarizes all that we can infer about the value of H given the data. Figure 3b shows how the posterior for each of the three priors in Fig. 3a changes as we are given more and more data. The data in this example were generated by using a random-number generator in a computer and setting H to 0.2. We find that as we obtain more data, we become more confident in our prediction for the inferred value of H (that is, the width of each posterior decreases) and our prior state of knowledge, as expressed in $\Pr(H)$, becomes less important (that is, no matter what our prior assumptions were, the posteriors converge to the same answer when enough data are available).

The power of Bayes' theorem is that it effectively provides the only consistent bridge between the inductive logic (or indirect probabilities) required for scientific inference and the deductive logic (or direct probabilities) that we know how to use. Generalizing, we see that Bayes' theorem encapsulates the process of "learning":

$$\Pr(\text{"hypothesis"}|\{\text{data}\}, I) \propto \Pr(\{\text{data}\}|\text{"hypothesis"}, I) \times \Pr(\text{"hypothesis"}|I),$$

where the “hypothesis” is the quantity that we wish to infer (the bias-weighting of a coin, for example, or the neutron scattering law for some sample) and I represents any prior knowledge we may have about the “hypothesis.” The prior probability distribution, $\Pr(\text{“hypothesis”}, I)$, reflects our knowledge (or ignorance) about the “hypothesis” before we obtained the data. This prior state of knowledge is modified by the likelihood function, $\Pr(\{\text{data}\}|\text{“hypothesis”}, I)$, which encodes the nature of the “experiment” and involves the use of deductive logic, to yield our posterior probability distribution, $\Pr(\text{“hypothesis”}|\{\text{data}\}, I)$, which represents our state of knowledge about the “hypothesis” after we have obtained the data. What we infer about some quantity of interest depends not only on the data we have but also on what we know or assume about it a priori! If the data are accurate, abundant, and sensitive to the quantity of interest, then the likelihood function will be sharply peaked and will dominate the posterior probability distribution. No matter what our prior state of knowledge, the data force us to the same conclusion. If the data are inaccurate, few in number, or insensitive to the quantity of interest, then the posterior will depend crucially on our prior. In other words, if the data do not tell us very much, then our state of knowledge after we have obtained the data will be governed largely by our state of knowledge (or ignorance) before the experiment.

Just as Newton’s second law of motion is central to all classical mechanics, Bayes’ theorem provides the fundamental rule for all logical and consistent inductive inference. Many statistical tests and procedures can be derived, justified, or at least understood from Bayes’ theorem when one states the relevant assumptions and details about the situation under consideration. Model fitting, least squares, maximum likelihoods, singular-value decomposition, the maximum entropy method, Tikhonov regularization, Fourier filtering, the χ^2 test, the F test, Student’s t test, and other statistical procedures for analyzing data can all be seen as suitable courses of action for different choices or assumptions about three things: the prior information I , which can even determine what we mean by “hypothesis”; the prior probability distribution, $\Pr(\text{“hypothesis”}|I)$; and the nature of the experiment, which is enshrined in the likelihood function, $\Pr(\{\text{data}\}|\text{“hypothesis”}, I)$.

The Maximum Entropy Method

The data-analysis method known as maximum entropy (MaxEnt) arises in the context of a specific but commonly occurring problem—that of making inferences about *positive* and *additive* distributions. The neutron scattering law $S(Q, E)$ for a sample is an example of such a positive and additive distribution. It is positive because $S(Q, E)dQdE$ is proportional to the number of neutrons scattered with momentum transfer between Q and $Q + dQ$ and energy transfer between E and $E + dE$. It is additive because the number of neutrons scattered into a large $\Delta Q \Delta E$ interval is equal to the sum of the neutrons scattered into the small $dQdE$ intervals that compose the large $\Delta Q \Delta E$ interval. Other examples of positive and additive distributions include probability distribution functions, the radio-frequency brightness function of an astronomical source, the electron density in a crystal, the intensity of incoherent light as a function of position in an optical image, and so on. (By contrast, the amplitude of incoherent light is positive but not additive.) Given only the information I that the quantity of interest is a positive and additive distribution f , what should we assign as the prior probability distribution $\Pr(f|I)$? The assignment of a prior is often a difficult problem. Bayes’ theorem tells us that the prior is a necessary and integral part of making a scientific inference, but the theorem does not tell us how to assign it. Methods that seem to avoid the use of a prior merely make an implicit choice (usually of a uniform distribution) rather than state an explicit choice. (Luckily, as mentioned above, the prior does not matter very much when we have “good” data.)

The choice of a prior usually involves somewhat obscure arguments and frequently involves a consideration of the allowed transformation groups that specify our ignorance about the quantity of interest. For example, consider the problem of estimating the length L of a biological molecule. What prior $\Pr(L)$ should be assigned to express complete ignorance about the value of L before we have carried out any measurements? Well, if we are really ignorant about the size of the molecule, then we should assign the same prior for the numerical value of L irrespective of whether we make the measurement in meters, inches, cubits, or whatever. The variable L would then be a so-called *scale parameter*. To express our complete ignorance about the value of a scale parameter, we say that the prior must be invariant under a change of scale in the measurement units. Mathematically we require that $\Pr(L)dL = \Pr(\beta L)d(\beta L)$ for all values of $\beta \geq 0$, leading us to the assignment $\Pr(L) \propto 1/L$, or a uniform prior for $\log L$: $\Pr(\log L) = \text{constant}$.

The appropriate prior for a positive and additive distribution is, again, not immediately obvious. Many different types of arguments, however, including logical consistency, information theory, coding theory, and combinatorial arguments, lead us to believe that the prior is of a rather special form:

$$\Pr(f|I, \alpha, m) \propto \exp[\alpha S(f, m)]. \quad (3a)$$

Here the (prior) information I assumes only that f is positive and additive, and S is the generalized Shannon-Jaynes entropy:

$$S(f, m) = \int \{f(x) - m(x) - f(x) \log[f(x)/m(x)]\} dx. \quad (3b)$$

In this general expression for entropy, $m(x)$ is a Lebesgue measure on x , the space of the distribution, and α is a dimensional constant (initially unknown). We will say more about what this entropic prior means (and the value of α) a little later, but let us continue by considering $m(x)$ further.

In the absence of any data, the posterior becomes directly proportional to the prior, and our best estimate of f is given by the maximum of the entropy function S , which occurs at $f(x) = m(x)$. The function $m(x)$ is therefore a *default model* (that is, the solution to which f will default unless the data say otherwise) and can be thought of as representing our prior state of knowledge, or ignorance, about f . The default model is usually taken to be uniform (that is, constant), but the use of a nonuniform $m(x)$ can be important for such difficult problems as protein crystallography or for introducing spatial correlations across the positive and additive distribution we want to infer. If we know that f is normalized, so that $\int f(x) dx$ is fixed, and if the Lebesgue measure is uniform ($m(x) = \text{constant}$), then the entropy formula above reduces to the form

$$S(f) = - \int f(x) \log[f(x)] dx,$$

which is the form of the entropy familiar from statistical mechanics.

The other quantity we need in order to make an inference about the distribution f is the likelihood function $\Pr(\{\text{data}\}|f, I)$. The likelihood function incorporates the information about the experiment, whether it is a neutron-scattering experiment, a nuclear-magnetic-resonance experiment, a radio-astronomy experiment, or whatever. It relates the quantity of interest to the data we have, thereby encoding details about the type of experiment and the accuracy of the measurements.

Let us consider the common case in which the data are independent (one measurement does not affect another) and are subject to additive Gaussian noise. The likelihood function then takes the form

$$\Pr(\{\text{data}\}|f, I) \propto \exp(-\frac{1}{2}\chi^2), \quad (4a)$$

where χ^2 is the familiar misfit statistic, which measures how well a trial distribution f fits the actual data:

$$\chi^2 = \sum_{k=1}^N \frac{(D_k - F_k)^2}{\sigma_k^2}. \quad (4b)$$

Here D_k is the k th datum (say the number of neutron counts in the k th bin), σ_k is the noise, or expected error, in that datum (for a neutron-scattering experiment, $\sigma_k = \sqrt{D_k}$), and F_k is the value for the k th datum that a trial distribution f would have produced in the absence of any noise. The noise in the neutron counts, though really described by a Poisson distribution, is approximated well by a Gaussian distribution because the number of counts is usually large (≥ 10). Thus the usual model fitting corresponds to assuming the likelihood function in Eq. 4 and maximizing that function to obtain the best “fit” to the data (that is, implicitly assuming a uniform prior so that the posterior becomes directly proportional to the likelihood function).

According to Bayes’ theorem we must combine Eq. 3, the entropic prior, with the likelihood function of Eq. 4 to find the posterior probability distribution for f :

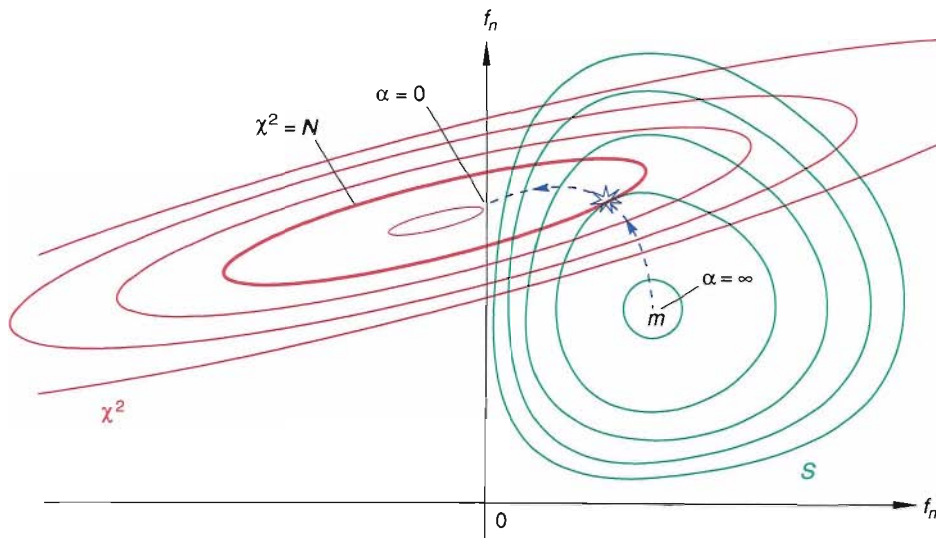
$$\Pr(f|\{\text{data}\}, I, \alpha, m) \propto \exp(\alpha S - \frac{1}{2}\chi^2).$$

Then, given the data and only the prior knowledge that f is a positive and additive distribution, our best estimate of f is given by the distribution that maximizes this posterior probability distribution. Since the exponential is a monotonic function, we obtain the solution by maximizing $\alpha S - \frac{1}{2}\chi^2$ (a general algorithm to do this is given in Bryan and Skilling 1984). This procedure can be interpreted as maximizing the entropy S subject to some constraint on the value of the misfit statistic χ^2 , where the initially unknown constant α is seen as a Lagrange multiplier. Hence the name *maximum entropy* method. The method is illustrated schematically in Fig. 4.

In past applications of the maximum entropy method, the constant α was chosen such that $\chi^2 = N$, where N is the number of data. This choice seems intuitively reasonable since any proposed distribution f should give data consistent with those actually measured, as defined by the constraint that $\chi^2 < \approx N$. The MaxEnt method was thus seen as choosing a distribution f that, while “fitting the data,” had the most entropy. More recent thinking (Skilling and Gull 1989), however, carries the Bayesian logic one step further: Since α is unknown, it becomes just one more parameter that needs to be estimated in the same sense that we are trying to estimate f . This approach, which leads to a slightly more complicated (but less ad hoc) criterion for the choice of α , has the advantage that the increased rigor allows us to automatically determine σ , the level of the noise, or expected error, in the measured data if it is not known. We leave these and other recent advances, including a discussion of practical reliability estimates of the inferred distribution f , to the avid reader (see Further Reading) and continue to pursue the more traditional approach to the MaxEnt method and its applications.

The Meaning of Maximum Entropy

Well, we have talked about the entropic prior, but what is its significance and what does it mean? To answer this question, we will use two very simple examples. The first, known as the *kangaroo problem*, is an example of having accurate but in-



THE MAXIMUM ENTROPY METHOD

Fig. 4. Suppose that we are trying to find the “best” estimate for some positive and additive distribution $f(x)$. Suppose further that the hypothesis space of f is defined by the values of f specified on a grid finely discretized with respect to x into N pixels. In other words, the hypothesis space of f is the N -dimensional space whose coordinate axes are the set $\{f_j\}$, where f_j is the value of f at pixel j . Shown here is a schematic two-dimensional section, namely the $f_m f_n$ plane, through the hypothesis space. Plotted (in red) are contours along which χ^2 , (twice) the logarithm of the likelihood function, is constant; the set $\{f_k\}$ for which $\chi^2 < \approx N$ (the number of data) compose the feasible set of distributions allowed by the data. Also plotted (in green) are contours along which the entropy S (the logarithm of the prior probability distribution) is constant; the entropy is a maximum at the default model $f = m$ (where m is a Lebesgue measure on the hypothesis space) and rapidly approaches $-\infty$ as any part of f becomes negative. The MaxEnt solution is that f for which the posterior probability distribution is maximum, that is, the f for which $\partial/\partial f_j(\alpha S - \frac{1}{2}\chi^2) = 0$. The blue line indicates the trajectory of the MaxEnt solution as the value of the Lagrange multiplier α goes from ∞ to 0; the blue star represents the traditional choice of α , which satisfies the condition that $\chi^2 = N$.

sufficient data. Nevertheless the problem is small enough that common sense tells us what constitutes a “sensible” solution. It will be shown that the MaxEnt choice, unlike several commonly used alternatives, concurs with our common sense. We will then use a second example, known as the *monkey argument*, to try to give a more general interpretation of the MaxEnt solution.

The Kangaroo Problem. We have said that in the MaxEnt method we choose, as our best estimate of a positive and additive distribution f , the f that agrees with the data and has the most entropy. This method of choosing a solution by maximizing some function of the desired distribution is known as regularization. The Shannon-Jaynes entropy is an example of a regularizing function, but several others are also commonly used. We will follow Gull and Skilling (1984) in using the kangaroo problem to demonstrate our preference for the choice of the Shannon-Jaynes entropy over the alternatives. The kangaroo problem, a physicists’ perversion of a formal mathematical argument (Shore and Johnson 1980), shows that the Shannon-Jaynes entropy is the only regularizing function that yields self-consistent results when the same information can be used in different ways (for example, the choice of coordinate system should not matter). The kangaroo problem is as follows.

Information: One-third of all kangaroos have blue eyes, and one-third of all kangaroos are left-handed.

Question: On the basis of this information alone, what proportion of kangaroos are both blue-eyed and left-handed?

Clearly, we do not have enough information to know the correct answer: All solutions of the type shown in the 2×2 contingency table of Fig. 5a agree with the data and thus constitute the *feasible set* of solutions. Without additional information, each solution is equally likely because they all fit the data exactly. Figure 5b shows three among the myriad of feasible solutions: namely, the one with no correlation between being blue-eyed and left-handed and the ones with the maximum positive and negative correlation. Although the data do not allow us to say which is the *correct* answer, our common sense compels us to choose the uncorrelated solution if we are forced to make a choice. That is to say, unless we have prior knowledge to the contrary, we do not expect that knowing the eye color of a kangaroo will tell us anything about whether the kangaroo is left-handed or right-handed. Thus our best estimate is that one-ninth of the kangaroos will be blue-eyed and left-handed.

Table 1 shows the results of selecting the solution by maximizing four commonly used regularizing functions. Note that the integral in the formula for the entropy, for example, has been replaced by a summation because the space of the distribution, x , is not continuous but discrete. In fact, it consists of just four pixels—the four boxes in the 2×2 contingency table. For this very simple example, where

TRUTH TABLES FOR THE KANGAROO PROBLEM

Fig. 5. (a) This truth table illustrates the general feasible solution to the kangaroo problem. That solution is obtained by letting $x \equiv f_1$ be the fraction of kangaroos that are blue-eyed and left-handed, where $0 \leq x \leq \frac{1}{3}$. Then the fractions corresponding to the other contingencies (f_2 , f_3 , and f_4) can be expressed in terms of x . (b) These truth tables illustrate three specific solutions derived by setting x to $\frac{1}{9}$, which corresponds to no correlation between being blue-eyed and left-handed, and by setting x to $\frac{1}{3}$ or 0, which correspond respectively to maximum positive or negative correlation between the traits.

(a) General Feasible Solution

		Left-Handed	
		T	F
Blue-Eyed	T	$f_1 = x$ $0 \leq x \leq \frac{1}{3}$	$f_2 = \frac{1}{3} - x$
	F	$f_3 = \frac{1}{3} - x$	$f_4 = \frac{1}{3} + x$

(b) Three Specific Solutions

		Left-Handed	
		T	F
Blue-Eyed	T	$f_1 = \frac{1}{9}$	$f_2 = \frac{2}{9}$
	F	$f_3 = \frac{2}{9}$	$f_4 = \frac{4}{9}$

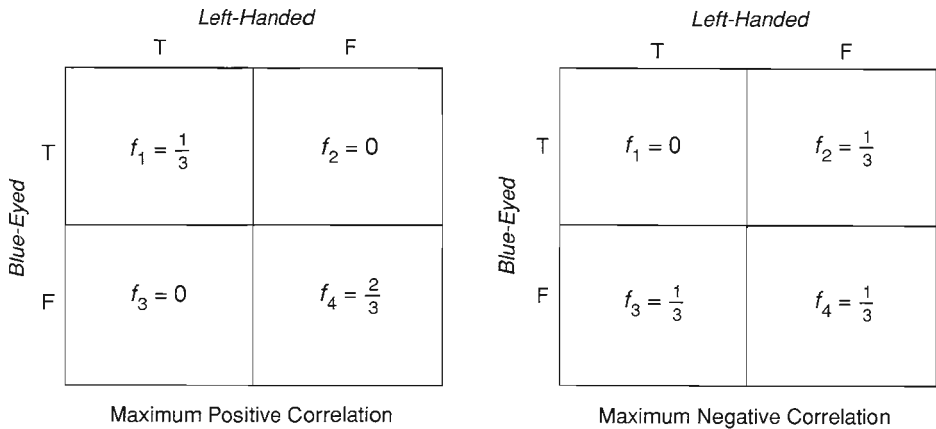
No Correlation

common sense tells us the “best” answer when faced with insufficient (but noise-free) data, it is only the Shannon-Jaynes entropy that yields a sensible answer! (Although we have considered only four regularizing functions, it can be shown that the Shannon-Jaynes entropy is the only one that has this desired property.)

Before going on to consider a more general interpretation of the MaxEnt choice, it is worth commenting on the frequently heard statement that in data analysis (or image reconstruction) positivity is the important constraint, not how you enforce it. For large problems that statement is very often true. Our small kangaroo problem, however, magnifies the differences among the regularizing functions and shows that we get more from MaxEnt than just positivity. The way we have set up the problem in Fig. 5a has the positivity constraint already built in, but it is still not sufficient to make a choice on the basis of the data we are given. The $\sum f^2$ regularizing function, for example, which for the kangaroo problem corresponds to the “Tikhonov with positivity” that some people seek, does not yield the same solution as our common sense—only MaxEnt does! Many general image-processing methods (both ad hoc and sound) often give similar results. The similarity merely reflects the fact that the prior probability distribution does not usually matter very much when the data are “good.” However, if we assume only that the quantity of interest is a positive and additive distribution and ask what is the appropriate choice for the prior, the answer is the entropic prior.

The Monkey Argument. Our common sense recommended the uncorrelated solution to the kangaroo problem because, intuitively, we knew that it was the most noncommittal choice. The data did not rule out correlation, but, without actual evidence, it was a priori more likely that the genes controlling handedness and eye color were on different chromosomes than on the same one. Crudely speaking, if we consider randomly scattering two genes among eight chromosomes, they are seven times more likely to land on different chromosomes than on the same one. Although we cannot usually appeal to specific knowledge such as what is known about genes and chromosomes, we can use the monkey argument (Gull and Daniell 1978) to see more generally that the MaxEnt choice is the one that is *maximally noncommittal* about the information we do not have.

The monkey argument can again be thought of as a physicists’ perversion of formal mathematical work, that of Shannon (1948) showing that entropy is a unique measure of “information content.” The words “information content” are being used here in the information-theory sense and have somewhat the opposite sense of their everyday use! We might better think of entropy as a measure of uncertainty (rather than as a measure of information) because uncertainty is closer to the idea of the lack of order that characterizes entropy. However, a system that has more entropy has a



greater degree of randomness, and its description requires more information (more bits in a computer). It is in this sense, then, that entropy is a measure of information.

The monkey argument presents the MaxEnt solution in graphic terms. Imagine a large team of monkeys who make images, or positive and additive distributions, by randomly throwing small balls of light at a rectangular grid. After a while, the grid is removed and replaced by another and so on. Eventually, the monkeys will generate all possible images, and many copies of each one. If we want an image of an object about which we have some experimental data, we can reject most of the monkey images because they give data that are inconsistent with the experimental measurements. Those images that are not rejected constitute the feasible set. If we are to select just one image from this feasible set as representing our best estimate of the object, the image that the monkeys generate most often would be a sensible choice. Because our hypothetical team of monkeys is presumed to have no particular bias, such a choice represents the image that is consistent with the measured data but, at the same time, is most noncommittal about the information we do not have. This preferred image is the MaxEnt solution, because the entropy is just the logarithm of the number of ways in which the image could have been generated (and, hence, the number of times it was).

Applications of MaxEnt at LANSCE

MaxEnt has been used successfully in image reconstruction in a wide variety of fields (see, for example, Gull and Skilling 1984). A small selection of its diverse applications, shown in Fig. 6, include forensic deblurring, radio astronomy, medical tomography, and nuclear-magnetic-resonance spectroscopy. We are now starting to use this powerful technique, and Bayesian ideas in general, to enhance the analysis of neutron-scattering data at LANSCE.

The Filter-Difference Spectrometer. The first example of the use of MaxEnt at LANSCE is the analysis of data from the Filter-Difference Spectrometer, or FDS. This example has the form of a standard convolution problem. That is, the data are related to the quantity of interest through a blurring process, so that they are a blurred (and noisy) version of what we want.

Our own eyes produce such a convolution, or blurring. Because the pupils of our eyes have a finite size, we do not see point sources of light as infinitesimal dots but as small fuzzy disks. (The angular size of the disk is roughly λ/d , where λ is the wavelength of the light and d is the diameter of the pupil.) If two point sources of light are so close that the disks overlap, we can no longer distinguish them as separate entities. Such blurring, response, resolution, or point-spread functions occur al-

Table 1

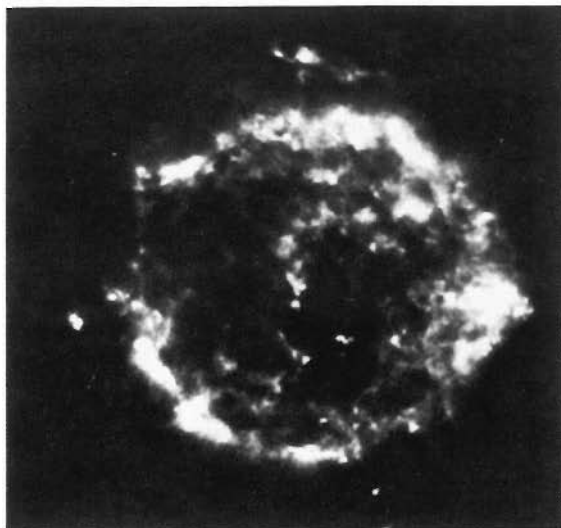
REGULARIZATION-FUNCTION SOLUTIONS OF KANGAROO PROBLEM

Listed here are values of x (fraction of kangaroos that are blue-eyed and left-handed) derived by maximizing four commonly used regularizing functions. Of the four only the Shannon-Jaynes entropy, $-\sum f_j \log f_j$, yields a value for x that agrees with our common sense, which tells us that, in the absence of relevant data, the two traits are most likely to be uncorrelated.

Regularizing Function	x	Correlation
$-\sum f_j \log f_j$	0.111..., or $\frac{1}{9}$	None
$-\sum f_j^2$	0.083..., or $\frac{1}{12}$	Negative
$\sum \log f_j$	0.13013	Positive
$\sum \sqrt{f_j}$	0.12176	Positive

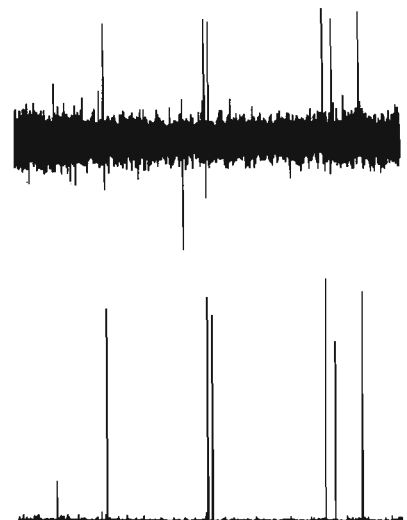


MaxEnt deblurring of a photographic image



MaxEnt x-ray tomograph of a human skull

MaxEnt image of radio-frequency (5-gigahertz) emissions from the supernova remnant Cassiopeia A



Comparison of (top) conventional Fourier reconstruction and (bottom) MaxEnt reconstruction of a nuclear-magnetic-resonance spectrum

SOME APPLICATIONS OF THE MAXIMUM ENTROPY METHOD

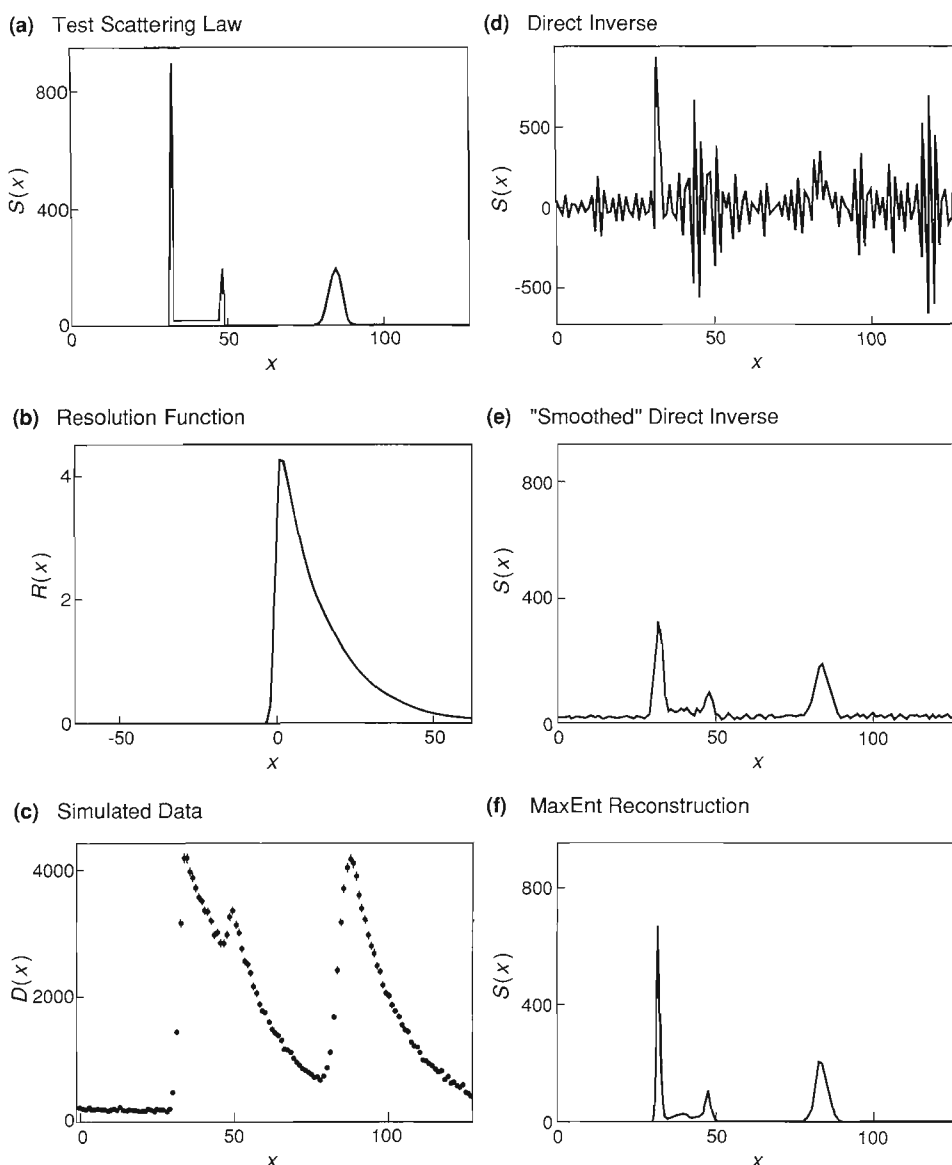
Fig. 6. The examples presented here, which are reproduced through the courtesy of J. Skilling and S. F. Gull, show how the maximum entropy method can be used to clarify the information extracted from a variety of data.

most universally in all experiments. The resolution functions of particular interest in neutron-scattering experiments arise from various aspects of the experimental setup, such as the finite size and temperature of the moderator and the finite angle of collimation of the neutron beam. In the case of the FDS, the major contribution comes from the transmission spectrum of the polycrystalline filters used to select for recording those inelastically scattering neutrons with certain final energies.

For those not familiar with the idea of a convolution, or the performance of MaxEnt, we start with a simple simulated example computed on a grid of 128 points. Suppose that the “true” object, or neutron scattering law, consists of two spikes on the left separated by a small plateau and a broader peak on the right, as shown in Fig. 7a. Also suppose that a noisy data set (Fig. 7c) is generated by first convolving the scattering law with a resolution function (Fig. 7b) that is similar to the transmission spectrum of the filters used in the FDS and then adding to the resulting blurred signal a small background count and random noise. In a convolution each point of the object (pixel) is replaced with a copy of the resolution function scaled by the “height” of the object at that point; the data are then the sum of all the scaled copies of the resolution function. As can be seen from Figs. 7a and 7c, a large single spike can give much the same data as a smaller broad peak. Mathematically, using matrix and vector notation, we can write the “experiment” as

$$\mathbf{d} = \mathbf{O} \cdot \mathbf{f} + \mathbf{b} \pm \sigma.$$

Here \mathbf{d} is the data vector, the matrix \mathbf{O} is the convolution operator ($O_{jk} = r_{k-j}$, where \mathbf{r} is the resolution function), \mathbf{f} is the scattering law, \mathbf{b} is the background, and σ is the root-mean-square value of the random noise ($\langle \sigma_k^2 \rangle = d_k$). Given the data set and a knowledge of the resolution function and background, we wish to infer the underlying scattering law. A simple way of performing the deconvolution is to ap-



DECONVOLUTION OF SIMULATED DATA

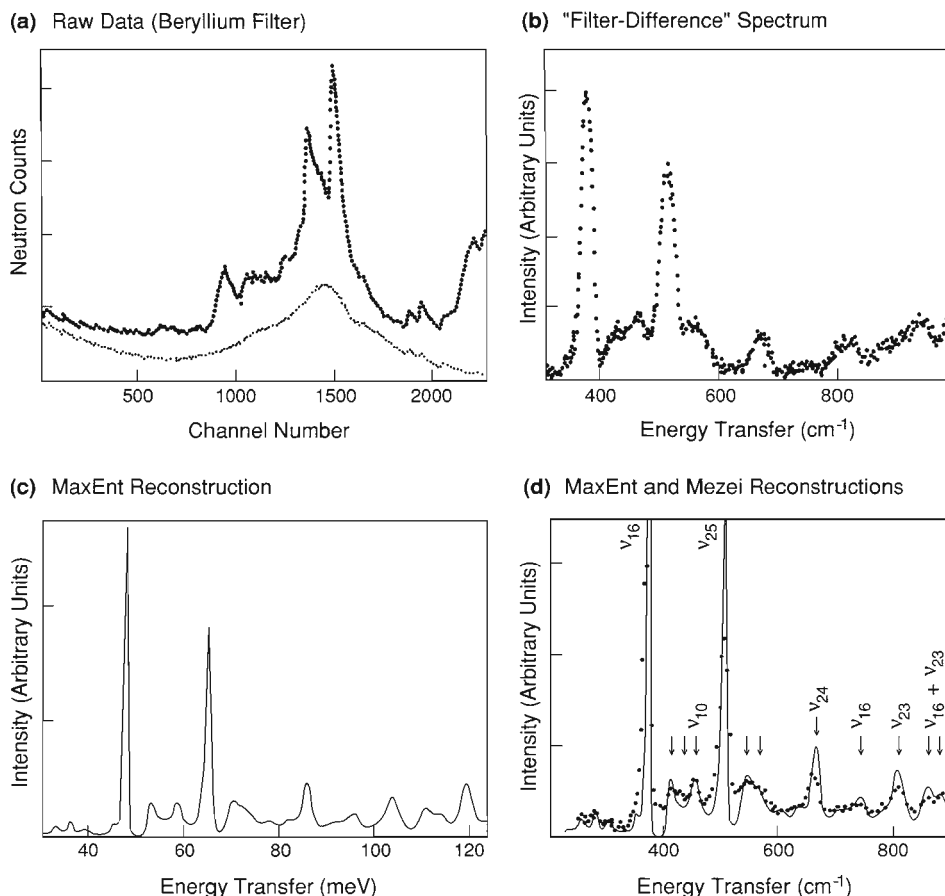
Fig. 7. The power of the maximum entropy method is illustrated by its application to a simulated data set. The simulated data set (c) was obtained by convolving a test scattering law (a) with an instrumental resolution function (b) and then adding a small background and random noise. (The instrumental resolution function shown in (b) is similar to the transmission spectrum of the filters used in the Filter-Difference Spectrometer at LANSCE). The series (d), (e), and (f) compares reconstructions, or deconvolutions, of the mock scattering law produced by three methods.

ply the inverse convolution operator O^{-1} to $\mathbf{d} - \mathbf{b}$ by using Fourier transforms. This procedure is equivalent to making the assumption that the prior is uniform, $\Pr(f) = \text{constant}$, and determining the *maximum-likelihood* solution. Unfortunately, the inverse solution may not exist. For example, the maximum-likelihood solution may not be unique because of missing data. Furthermore, even when the inverse does exist, it produces a reconstruction of the scattering law (Fig. 7d) that has a lot of high-frequency ringing (wiggles). To overcome this difficulty, it is common practice to use a smoothed (or slightly blurred) version of the direct inverse, a procedure known as *Fourier filtering* (Fig. 7e). In the grand scheme of things, Fourier filtering can be regarded as an example of *singular-value decomposition*. An alternative approach is to use the fact that the scattering law is a positive and additive distribution and hence choose an entropic prior ($\Pr(f|I) \propto \exp(\alpha S)$) and thus obtain the MaxEnt solution shown in Fig. 7f. We find that the maximum entropy method has suppressed the level of the artifacts without sacrificing as much detail in the reconstruction as does Fourier filtering.

Now, let us turn from simulated data to real data. The FDS is an instrument used to perform molecular rotational-vibrational spectroscopy with neutrons rather than with photons, as in infrared or Raman spectroscopy. Figure 8a shows data taken with a beryllium filter imposed between the sample and the detector. Those data

DECONVOLUTION OF INELASTIC-NEUTRON-SCATTERING DATA

Fig. 8. Shown in (a) are inelastic neutron-scattering data obtained with the Filter-Difference Spectrometer at LANSCE. Such data are the basis for deducing the energy levels of the molecular vibrations and rotations excited in a sample by the incident neutrons. (Here the sample is hexamethylene tetramine at 15 kelvins; its well-known rotational-vibrational spectrum is used to calibrate the energy-transfer values deduced from the recorded times of flight and the energy-cutoff points of the filters.) The raw data shown in (a) are a convolution of the true rotational-vibrational spectrum of the sample with the transmission spectrum of a beryllium filter located between the sample and the detector. (That transmission spectrum is similar to the resolution function shown in Fig. 7b.) Shown in (b) is the "filter-difference" spectrum, a hardware deconvolution of the data in (a) derived by subtracting the raw data in (a) from raw data obtained with a beryllium oxide filter. (The transmission spectrum of a beryllium oxide filter differs from that of a beryllium filter mainly in being slightly shifted in energy.) The filter-difference spectrum is inverted relative to the raw data plot because the abscissa in (a) is (essentially) the time of flight of the scattered neutrons whereas the abscissa in (b) (and (c) and (d)) is the energy transferred to the sample. Shown in (c) is the MaxEnt reconstruction of the data in (a). The MaxEnt reconstruction and a filtered inverse, or "Mezei," reconstruction (dots) are compared in (d).

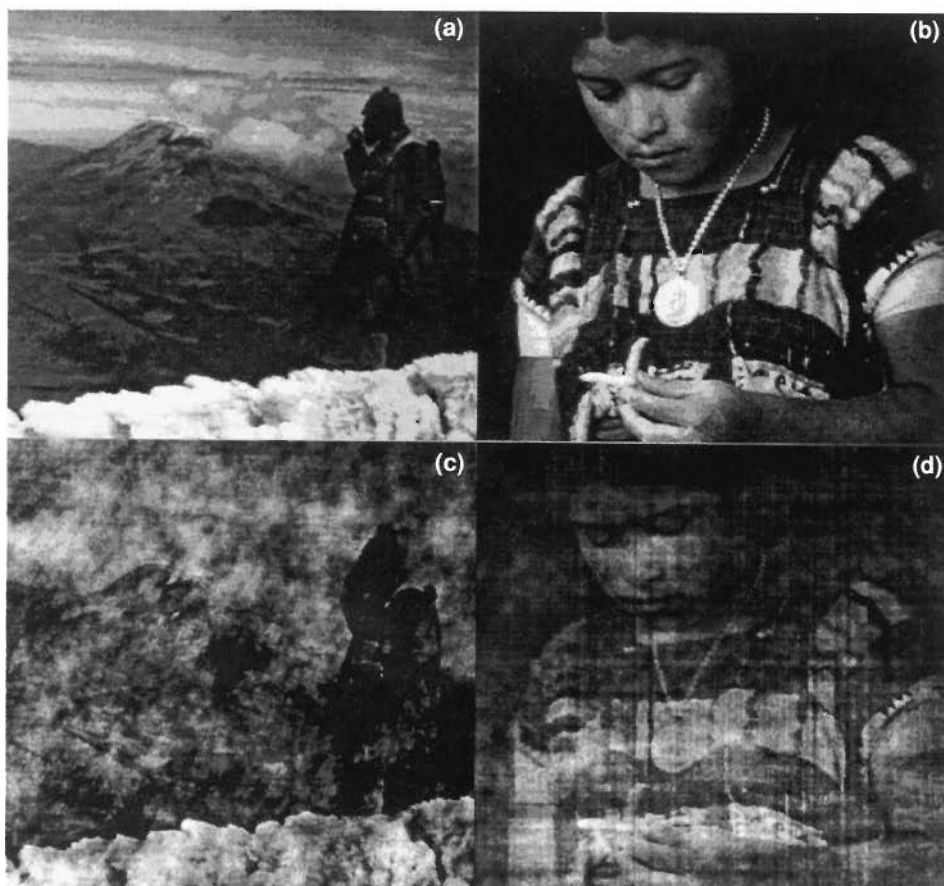


show the effects of the sharp edge and long decaying tail of the transmission spectrum of the filter (see Fig. 7b). The earliest method used to remove the blurring produced by such a resolution function is a hardware solution. Two data sets are collected, one consisting of the scattered neutrons transmitted through a beryllium filter and the other consisting of the scattered neutrons transmitted through a beryllium oxide filter. The transmission spectra of the two filters have almost the same shape, but their sharp energy cutoffs are slightly offset. Therefore, the data set obtained with one filter differs from the data set obtained with the other filter mainly in being shifted in energy by a small amount. When the two data sets are subtracted, the contributions from the long decaying tails (and background) tend to cancel, and only the significant features defined by the sharp rising edges remain. Figure 8a shows raw data obtained with only the beryllium filter plotted in data channels corresponding to increasing neutron time of flight. Figure 8b shows the corresponding "filter-difference" spectrum plotted as a function of energy transfer. The filter-difference spectrum is inverted relative to the data plot because increasing time of flight is equivalent to decreasing energy transfer.

Given only the data obtained with the beryllium filter and knowledge of the filter's transmission spectrum and the background, the deconvolution can be carried out mathematically (in software) by using the maximum entropy method. The MaxEnt reconstruction thus obtained is shown in Fig. 8c and is compared in Fig. 8d with a conventional reconstruction (due to Mezei) that can be interpreted as a filtered inverse. As expected, the MaxEnt reconstruction is an improvement over both the filter-difference and the Mezei deconvolutions in that it shows finer detail and fewer noise artifacts. The improvement is obvious but not dramatic because the data have good statistical accuracy. Noisier data causes the filtered inverse solution to deteriorate much more rapidly than the MaxEnt solution.

THE FOURIER PHASE PROBLEM

Fig. 9. Image (c) is a Fourier reconstruction obtained by using the Fourier phases of image (a) and the Fourier amplitudes of image (b); image (d) is a Fourier reconstruction obtained by using the Fourier phases of image (b) and the Fourier amplitudes of image (a). These two reconstructions demonstrate that most of the information in a Fourier transform is contained in the phases rather than in the amplitudes.

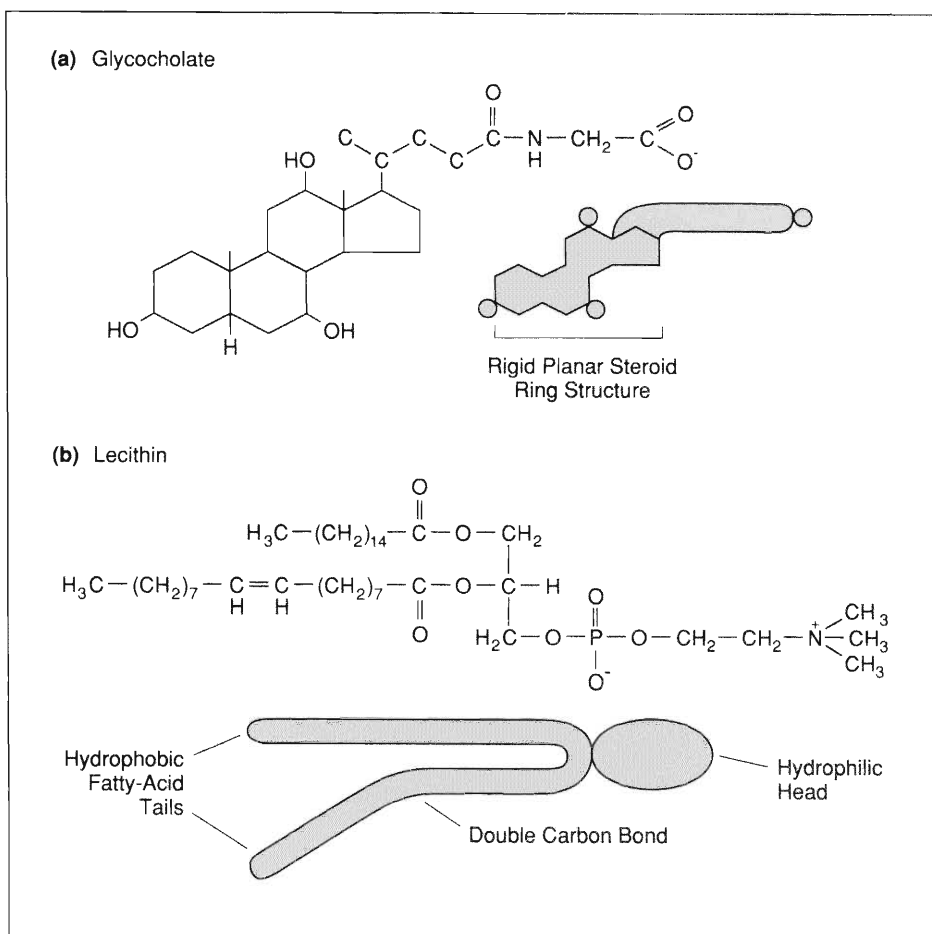


The Low-Q Diffractometer. The next application of the maximum entropy method involves the analysis of data that reflect the aggregation of biological macromolecules in solution. These data were taken on the Low-Q Diffractometer, a small-angle neutron-scattering (SANS) instrument useful for studying structures with dimensions ranging from 10 to 1000 angstroms. The spatial distribution of particles in a sample (including their size, shape, and location) is related to the neutron scattering law through a Fourier transform—in general, a complex quantity. (The elements of the O matrix for a Fourier transform are of the form $O_{jk} = \exp(i2\pi jk/N)$, where $i^2 = -1$ and N is the number of points in the discrete Fourier transform.) The neutron counts we measure are, of course, given by the Fourier intensities (or a blurred and noisy version thereof). We are thus brought face-to-face with the dreaded *Fourier phase problem*! The Fourier phase problem entails trying to make an inference about some quantity of interest given information about only the amplitudes (but not the phases) of its Fourier transform. It is a notoriously difficult problem, well known in x-ray crystallography, because the many local maxima of the likelihood function make it hard for us to find the global maximum of the posterior probability. The gravity of the situation is illustrated by Fig. 9. Luckily, we are not interested in determining the relative locations of the particles but only the number of particles of a given size and shape. Thus our problem is analogous to the problem in x-ray crystallography of determining not the electron-density map but only the autocorrelation (or Patterson) function, for which the Fourier intensities alone are sufficient.

The particles under study are involved in the digestion and transport of fats. My biologist colleague at LANSCE, Rex Hjelm, introduced me to the problem by saying: “Your body is mostly water. If you visit your favorite ice-cream parlor, then the fat in the ice cream will form a greasy blob at the bottom of your stomach and you will soon die!” He then told me that bile salts, produced in the liver, had hy-

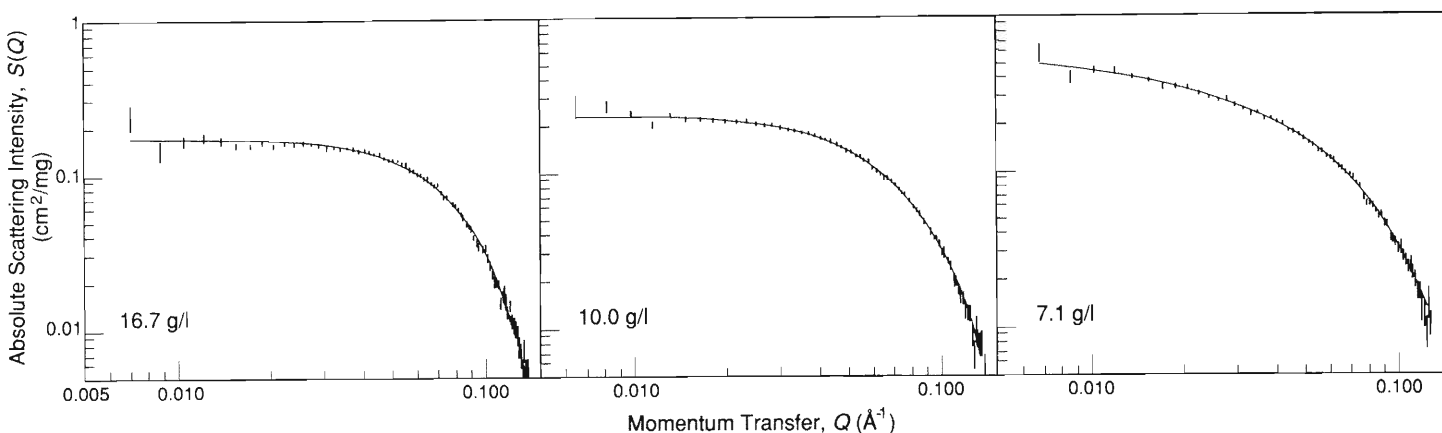
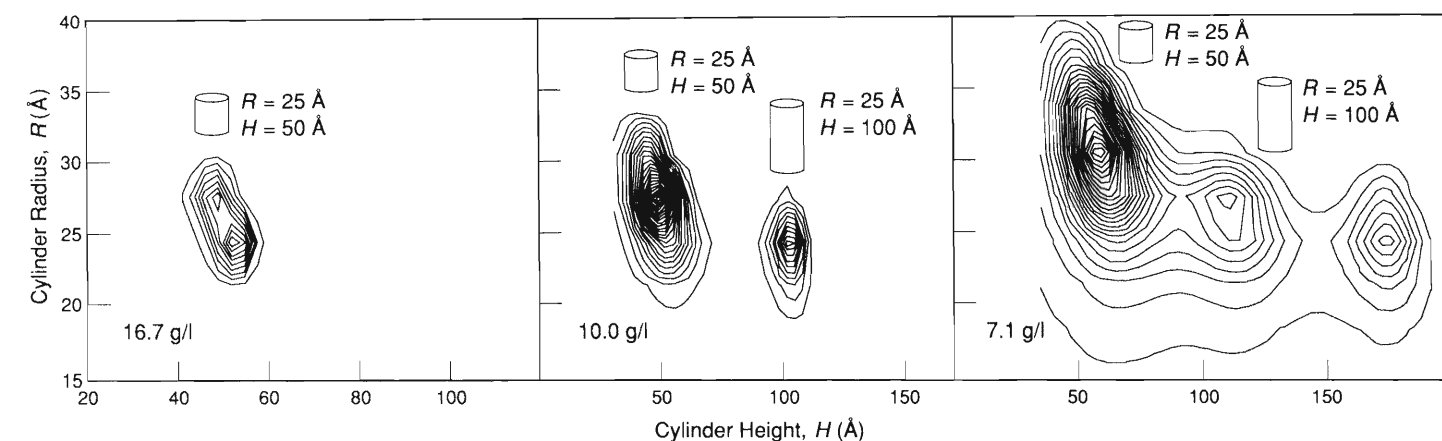
DIGESTION OF FATS

Fig. 10. The digestion of dietary fats begins with their emulsification by complexes of bile-salt and lecithin molecules. Bile salts (glycocholate and taurocholate) are polar derivatives of cholesterol. Shown in (a) are the structural formula and a schematic diagram of glycocholate. (Polar groups are denoted in the schematic diagram by circles.) Taurocholate differs in that the terminal carboxyl group (CO_2^-) is replaced by the group $\text{H}_2\text{C}-\text{SO}_3^-$. Although glycocholate, say, is itself an effective emulsifier, complexes of glycocholate and the lipid lecithin are, for reasons not yet known, even more effective. Shown in (b) are the structural formula and a schematic diagram of lecithin, or phosphatidyl choline. Aiding the digestion of fats is not the only physiological function of lecithin; it also is a major constituent of the lipid bilayers that compose biological membranes.



drophylic heads and hydrophobic tails (Fig. 10a). “So the body dumps in some bile salts to act as detergents,” I remarked, somewhat relieved. “No, that’s what an engineer would do!” came the reply. For reasons that we do not fully understand, nature uses a conglomerate of bile salts and the fat lecithin (Fig. 10b) to begin the digestion process.

An understanding of the action of bile salts in lipid digestion and in the transport of liver products such as cholesterol has potential applications in industrial processes and in the development of drug-delivery systems and model membranes. As a step in this direction, Hjelm et al. (1990) have been investigating the nature of particle growth in aqueous solutions of lecithin and the bile salt glycocholate. Figure 11a shows SANS data sets for three increasingly dilute solutions. Hjelm asked the following question: If I assume that the particles in the sample can be modeled as cylinders of uniform density, what is my “best” estimate of their size distribution, given the data and a knowledge of the experimental setup? Since SANS data are not sensitive to fine structure, the sharp edges of the cylinders are of little consequence; all that we are really assuming is that the particles are “blobs” of uniform density defined by a length and a diameter. Moreover, the fact that the distribution of particle sizes is a positive and additive quantity means that the relevant prior for the distribution of particle sizes is an entropic prior! Figure 11b shows the particle-size distributions derived by using MaxEnt on the data in Fig. 11a. The distribution for the highest lipid concentration indicates the presence of only a single type of particle, roughly globular, with a diameter of about 50 angstroms. As the sample is diluted, evidence for a second type of particle appears, a rod-like structure with a diameter of about 50 angstroms and a length of about 100 angstroms, or twice the original length. Even greater dilution leads to the appearance of even more elongated particles with

(a) Small-Angle Neutron-Scattering Data**(b) Cylinder-Size Distributions**

a length of about 170 angstroms, or three to four times the original length. These results lead us to believe that particle growth occurs through aggregation of preformed subunits with a size of about 50 angstroms (which corresponds nicely to the thickness of lecithin bilayers) rather than through the aggregation of individual bile-salt or lecithin molecules.

This example shows that the data need not bear any visual resemblance to the information extracted; in other words, MaxEnt is a method for data analysis, or scientific inference, and not just image enhancement.

The Constant-Q Spectrometer. Our last example involves data from the Constant-Q Spectrometer (CQS), an instrument designed to investigate phonons and magnons in single-crystal samples. The example illustrates a more advanced use of MaxEnt—*multichannel entropy*. This method is needed for convolution problems in which we want to determine not only the (sharp) scattering law of interest but also a broad, unknown background signal. We will begin with a simple simulation to illustrate multichannel entropy and then demonstrate its use on real data from the CQS.

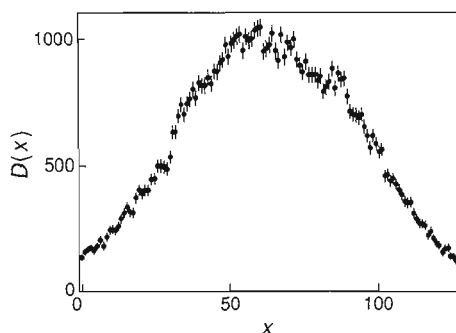
For our simulation we convolve the scattering law of Fig. 7a with the resolution function of Fig. 7b (scaled down by a factor of about 10) and then add a large background, assumed to be unknown, to generate the noisy data set shown in Fig. 12a. To analyze these data we use the technique of *two-channel entropy*. We assume that the unknown background b is also a positive and additive quantity and is fairly broad compared with the scattering law f . What we are attempting to do is an example of multichannel entropy because we are trying to make our best inference about several different “images” simultaneously. In this case we have only two channels: one for the background and the other for the scattering law. We set up two image channels,

PARTICLE GROWTH

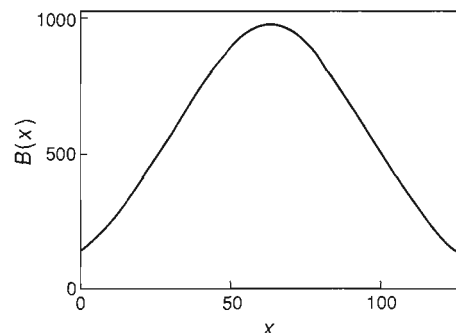
Fig. 11. Information about the sizes of glycocholate-lecithin complexes in an aqueous solution can be obtained by analysis of small-angle neutron-scattering data for the solution. Shown in (a) are such data (Hjelm et al. 1990) for increasingly dilute solutions. (The concentrations indicated are total concentrations of glycocholate plus lecithin.)

The corresponding particle-size distributions, shown in (b), were derived by assuming that the complexes are adequately represented by cylinders of radius R and height H and then using the maximum entropy method to determine the most probable distribution of cylinder sizes. Note that increasing dilution is accompanied by the appearance of populations of cylinders whose radii do not change significantly but whose heights increase by approximately integral multiples.

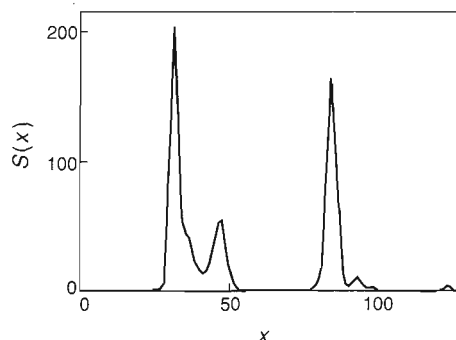
(a) Simulated Data



(b) MaxEnt Reconstruction: Background



(c) MaxEnt Reconstruction: Signal



TWO-CHANNEL DECONVOLUTION OF SIMULATED DATA

Fig. 12. (a) Two-channel entropy is an example of an advanced use of the maximum entropy method. It allows deconvolution of data into two components, such as a scattering law with sharp features and a relatively featureless background. Application of the two-channel entropy method to the simulated data in (a), which were generated by convolving the scattering law and the resolution function shown in Fig. 7a and Fig. 7b and then adding a large, unknown background, yields the background and scattering-law reconstructions shown, respectively, in (b) and (c).

f_1 and f_2 . One channel is allowed to have only broad structure (by construction); the other is permitted the full resolution of the 128-pixel grid. We also arrange the problem so that the “entropic cost” of putting structure in the broad channel is very low relative to the cost of putting structure in the high-resolution channel. What do we mean by entropic cost? Recall that the absolute maximum of entropy occurs when f is the same as the default model m . But as f deviates from m , in order to become consistent with the data, the entropy decreases, and that decrease in entropy is what we mean by entropic cost. Thus by making the entropic cost of putting structure in the broad channel relatively low, we ensure that if a broad distribution can account for the data, it will appear in the broad channel. If sharp structure is required, it can appear only in the high-resolution channel. We identify the high-resolution channel with the scattering law and the broad channel with the unknown background. Carrying out this procedure (for details see Sivia 1990), we obtain the MaxEnt reconstructions for the background and scattering law shown in Figs. 12b and 12c. Although the image of Fig. 12c is not as good as that of Fig. 7f, it is still a very impressive reconstruction in light of the given data (compare Fig. 7c with Fig. 12a!).

Finally, we show the application of this two-channel entropy algorithm to real data on the inelastic scattering of neutrons from phonons and magnons in a sample of iron. The data (Yethiraj et al. 1990) are shown in Fig. 13a as a function of the experimental variables: time of flight and detector angle. The data suffer from a combination of broadening and an unknown background signal (in addition to \sqrt{N} noise) that obscures the scattering law of interest. The MaxEnt reconstruction of the signal, or high-resolution, channel (Fig. 13b) shows a dramatic improvement in both the detail seen in the scattering law and in the reduction of background artifacts. When the scattering law is plotted in terms of the physically meaningful coordinates of energy and momentum transfer (Fig. 13c), we can easily identify the dispersion curves for the magnon and phonon excitations characteristic of iron.

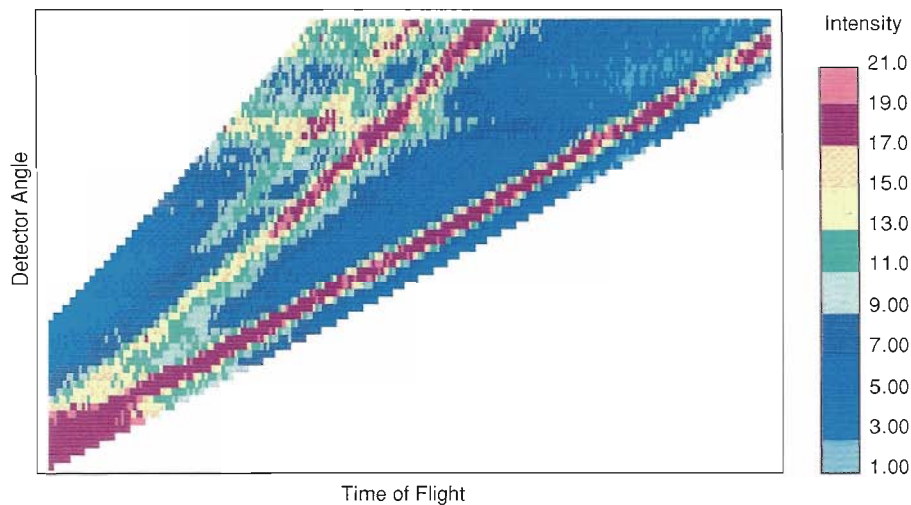
Instrument Design

The examples of the use of MaxEnt given in the last section are all cases of doing the “best” with the data we have. Usually that is all we can do. The instrumentation and hardware already exist at facilities like LANSCE, and often the only freedom a user has to improve the quality of the data is to increase its statistical accuracy by collecting data for a longer time. Let us suppose, however, that we are going to build a new facility, or just a new spectrometer. How should we design it to get the “best” data? This is an important question since a new facility can cost a hundred million dollars or more, and even a single spectrometer can cost a million or two!

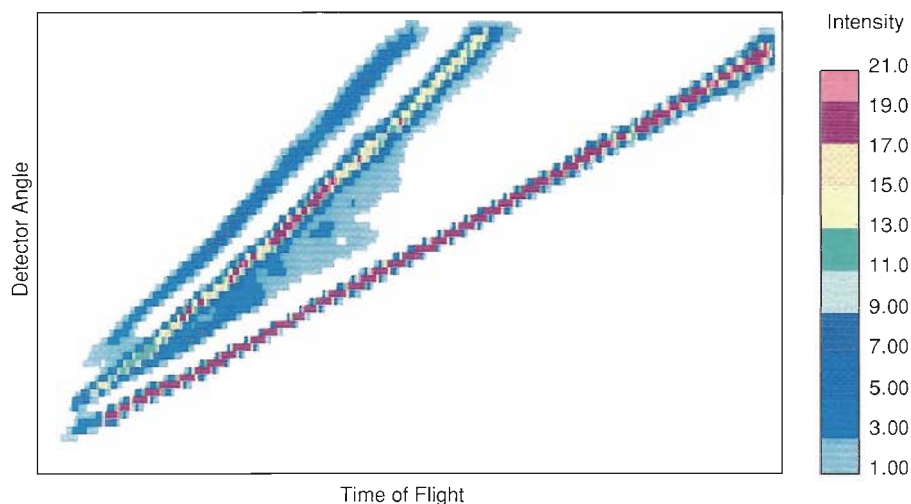
Silver, Sivia, and Pynn (1989) have addressed this question from a heuristic viewpoint and have also suggested a quantitative answer based on elementary signal-to-noise ratio arguments from a power-spectrum error analysis. They posed the following question: Given that the neutron-scattering data are usually a blurred and noisy version of the scattering law we want, what are the optimal characteristics of the instrumental resolution (blurring) function? Conventional wisdom suggests that the most important characteristic of the resolution function is its width: the wider the resolution function, the poorer the quality of the data in the sense that it is more difficult to determine reliably the underlying scattering law. Such thinking is based on a visual, or “what-you-see-is-what-you-get,” consideration of the data. A more formal analysis based on statistical inference, or image processing, leads to the conclusion that the overall shape of the resolution function is more important than its width.

We now outline the formal Bayesian approach to the question of instrument design; the algebra is presented in Sivia (1990). We will cast the problem in the same way as did Silver et al. and arrive at the same results; what we add here is the Bayesian rationale for their results. The real advantage of the Bayesian approach is

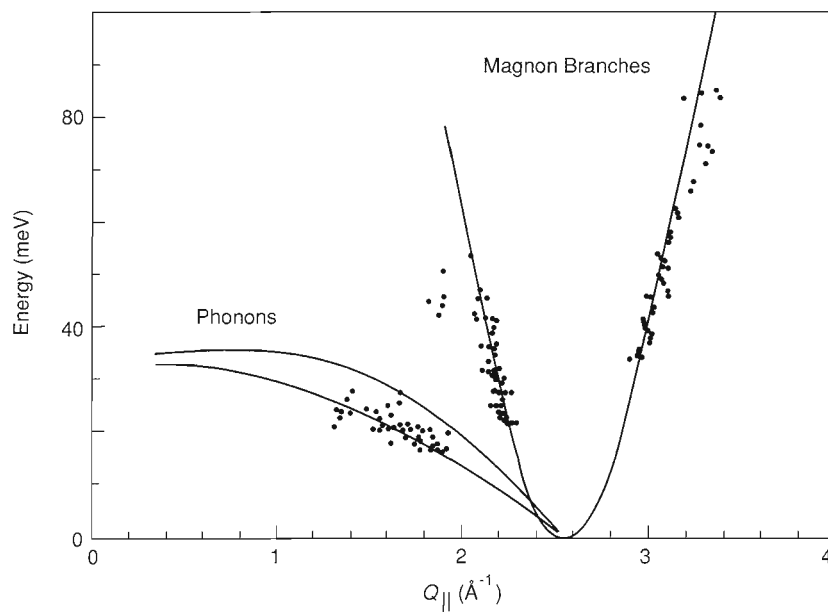
(a) Inelastic Scattering Data for Iron



(b) MaxEnt Reconstruction

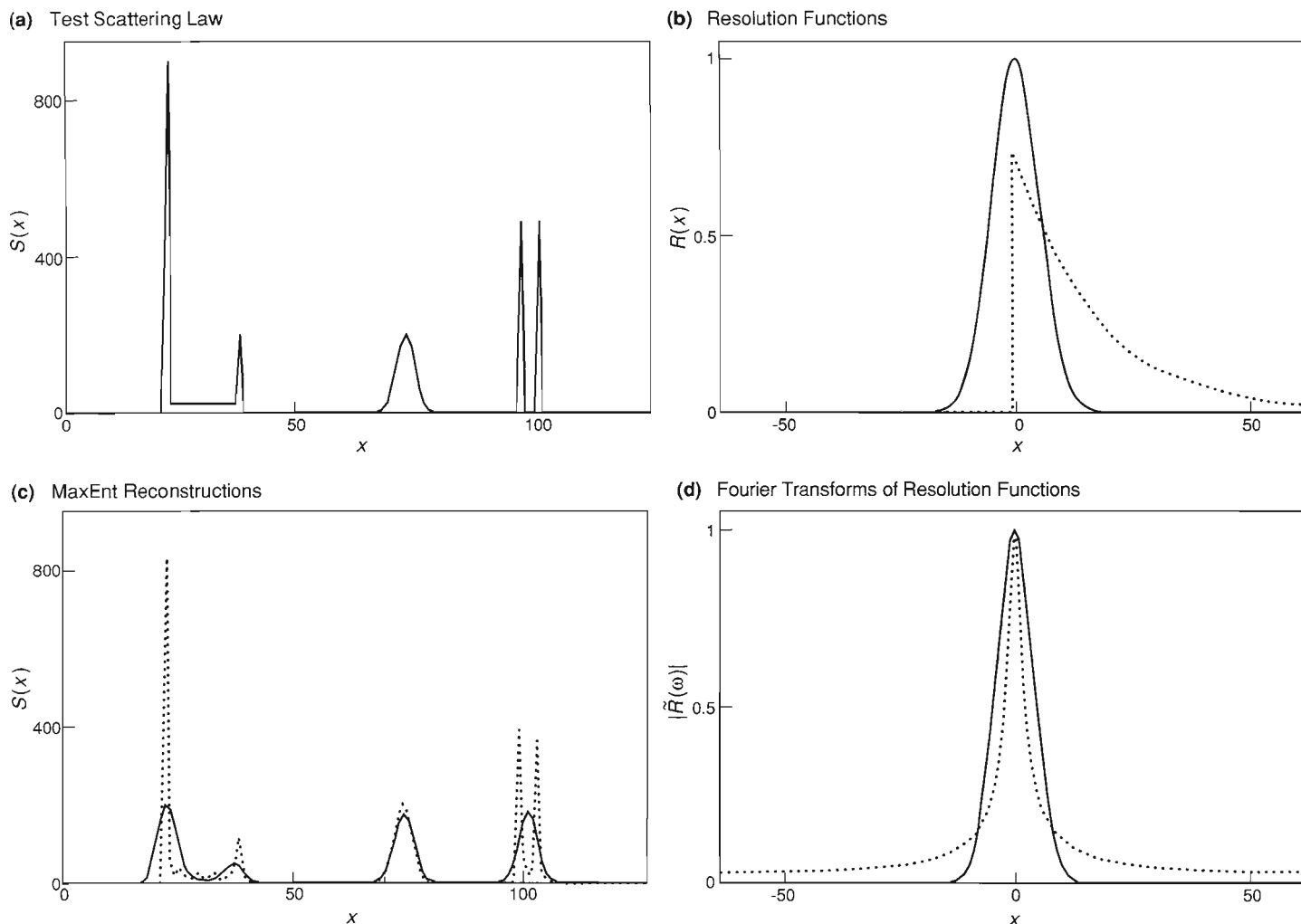


(c) Magnon and Phonon Dispersion Curves



TWO-CHANNEL DECONVOLUTION OF INELASTIC-NEUTRON-SCATTERING DATA

Fig. 13. Application of the two-channel entropy algorithm to the inelastic-neutron-scattering data for iron shown in (a) (Yethiraj et al. 1990) yields the deconvolved “signal” channel shown in (b). Transformation of the lines in (b) to the physically meaningful coordinates of energy transfer and Q_{parallel} (the component of the momentum transfer parallel to the incident neutron beam) reveals both branches of a magnon and some phonons (c).



THE FIGURE-OF-MERIT PROBLEM: RESOLUTION FUNCTIONS WITH THE SAME FWHM

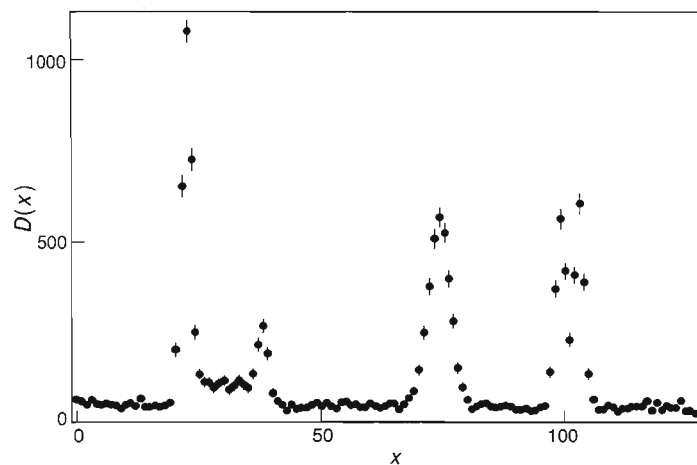
Fig. 14. Application of the maximum entropy method to data sets obtained by convolving the test scattering law in (a) with one or the other of the two resolution functions in (b) yields the reconstructions shown in (c). Both resolution functions have the same full width at half maximum (FWHM) and the same integrated intensity and hence have the same conventional figure of merit. Nevertheless, the reconstruction corresponding to the sharp-edged resolution function more nearly matches the original scattering law than does the reconstruction corresponding to the Gaussian resolution function. Also shown, in (d), is the Fourier transform of each resolution function. As discussed in the text, the Fourier transform of a resolution function, and not its full width at half maximum, is most relevant to defining a versatile figure of merit.

its generality; an almost identical analysis can be used to address questions about experimental design in many other contexts (not just convolutions). Moreover, our conclusions are relevant not only to neutron-scattering experiments but also to any other type of experiment involving some element of an instrumental resolution.

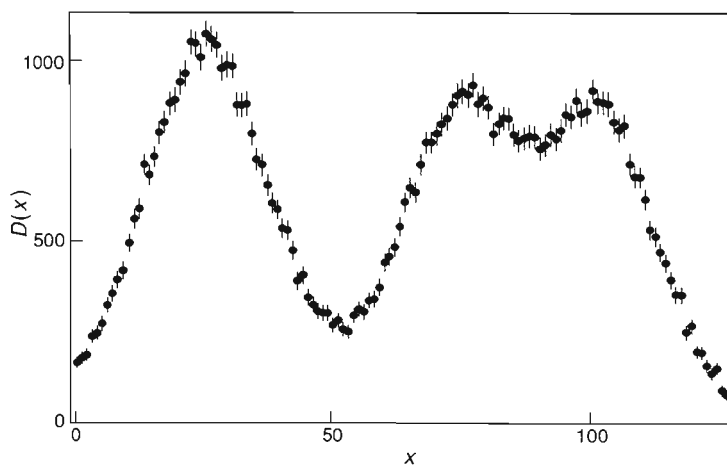
The first step in any data analysis is the formulation of the precise question we wish to answer. Formally, we must define the space of possible answers, or choose the hypothesis space. In the case of neutron scattering, we may say that we wish to know the scattering law of our sample, but how is the scattering law to be described? If we know (or assume) that the scattering law consists of a single Lorentzian, for example, then we have a three-dimensional hypothesis space defined by the position, height, and width of the Lorentzian. If, on the other hand, we have no functional form for the scattering law, then we might digitize it into a large number M of pixels, whereupon we have an M -dimensional hypothesis space defined by the flux in each pixel. However, the fact that our best estimate of the scattering law depends not only on the data but also on our choice of hypothesis space limits our ability to provide a universal figure of merit for instrument design. Nevertheless, we will be able to suggest at least a versatile figure of merit, one that is meaningful for many types of problems. But first let's analyze the problem using Bayesian logic.

Once we have chosen the hypothesis space, we can assign a probability distribution over it to indicate our relative beliefs in the various possible scattering laws. The assignment we make before conducting the experiment is, of course, the prior, and Bayes' theorem tells us how the prior is modified by the experimental data, through the likelihood function, to yield the posterior. We also know that the position

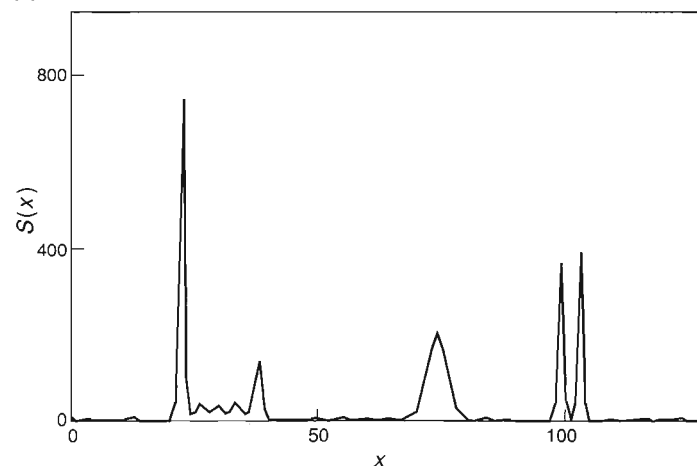
(a) Simulated Data (Narrow Gaussian)



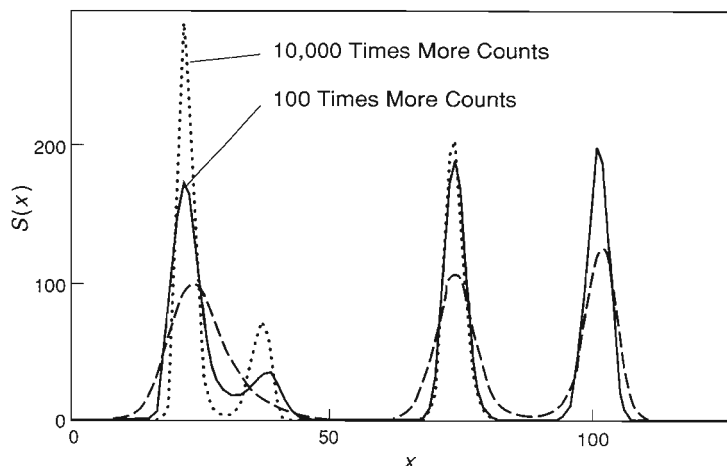
(b) Simulated Data (Wide Gaussian)



(c) MaxEnt Reconstruction (Narrow Gaussian)



(d) MaxEnt Reconstruction (Wide Gaussian)



of the maximum in the posterior gives us our best estimate of the scattering law and that the width, or spread, of the bump in the posterior around the maximum gives us a measure of the reliability of our estimate. Both, of course, depend on our choice of hypothesis space and on our assignment of the prior probability distribution, but they also depend on the data. The question of how to optimize instrument design can thus be stated as follows: How should we choose the instrumental parameters so that the resulting data give us the most reliable estimate of the scattering law?

Since Bayes' theorem tells us that the data affect our estimate of the scattering law only through the likelihood function, we need to look at its sharpness, or spread. The sharper the likelihood function, the greater the "information content" of the experiment in the sense that the data impose a more severe constraint on what the scattering law could be.

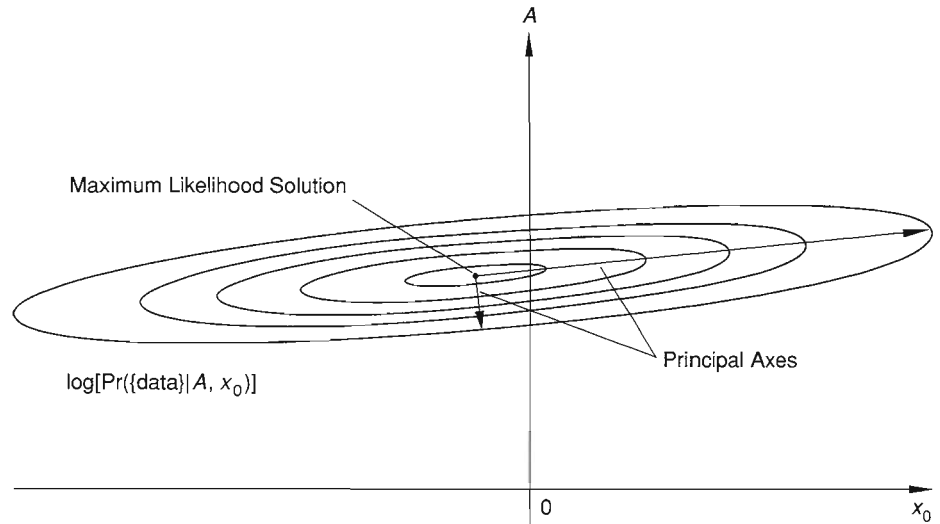
Let us begin by considering a very simple situation. Suppose we know that the scattering law consists of a single delta-function excitation, $A\delta(x - x_0)$, of known magnitude A and unknown position x_0 . In other words, we have a one-dimensional hypothesis space defined by x_0 . Suppose also that the experimental data are the result of a convolution between this scattering law and a Gaussian resolution function $T \exp(-x^2/2w^2)$. The height T of this Gaussian resolution function is determined by the length of time for which the data are collected, and its width w is some function of the instrumental parameters, such as flight-path length and collimation angle. The question now is: What restrictions do the data impose on the value of x_0 ? The width of the likelihood bump, viewed in the one-dimensional space of x_0 , gives us the uncertainty in x_0 , δx_0 , allowed by the data. After some algebra we find that δx_0

THE FIGURE-OF-MERIT PROBLEM: RESOLUTION FUNCTIONS OF SIMILAR SHAPE

Fig. 15. Shown in (a) and (b), respectively, are noisy data obtained by convolving the scattering law shown in Fig. 14a with a narrow Gaussian resolution function and another Gaussian resolution function ten times wider. Deconvolution of (a) and (b) yields (c) and the dashed curve in (d), respectively. Also shown in (d) are deconvolutions of data sets with 100 times (solid curve) and 10,000 times (dotted curve) the number of counts shown in (b). Contrary to conventional wisdom, increasing the number of data by a factor of 100 does not compensate (in terms of recovering sharp structure) for an increase in FWHM by a factor of 10.

A LIKELIHOOD FUNCTION IN A TWO-DIMENSIONAL HYPOTHESIS SPACE

Fig. 16. Suppose that the scattering law of interest is characterized by specific values of only two parameters, A and x_0 . (An example of such a scattering law is a delta function of unknown amplitude and position.) The likelihood function $\Pr(\{\text{data}\}|A, x_0)$ is then a bump in a two-dimensional hypothesis space. Shown here schematically in that hypothesis space are contours along which the logarithm of $\Pr(\{\text{data}\}|A, x_0)$ is constant. The shape of the likelihood function can be described either by a covariance matrix or by the eigenvectors and eigenvalues of the logarithm of the likelihood function. The elements of the covariance matrix tells us the expected uncertainties allowed by the data in our estimates of A and x_0 and how our estimate of one affects our estimate of the other. The eigenvectors, which specify the directions of the principal axes of the likelihood bubble, tell us which properties, or linear combinations, of A and x_0 can be determined independently. The eigenvalues, which are proportional to the widths of the likelihood function in the directions of the eigenvectors, tell us how reliably each independent property can be estimated.



depends on the instrumental design, enshrined in the resolution-function parameters T and w , in the following manner: $\langle(\delta x_0)^2\rangle \propto w/T$. The inverse of this quantity can be used as a figure of merit and has been quoted in the neutron-scattering literature:

$$\text{Conventional Figure of Merit} = \frac{\text{Total Number of Neutrons}}{(\text{FWHM})^2} \propto \frac{T}{w},$$

where FWHM is the full width of the resolution function at half maximum and the total number of neutrons detected is proportional to Tw .

We now show, by means of the examples presented in Figs. 14 and 15, that, although the conventional figure of merit is the correct answer to the question posed above, it is quite unsuitable for general use. Figure 14 presents the MaxEnt reconstructions derived from two data sets obtained by convolving a test object with one or the other of two resolution functions. Even though the resolution functions have identical figures of merit according to the equation above, the reconstruction from the data set obtained by convolution with the sharp-edged resolution function is clearly far superior to the reconstruction from the data set obtained by convolution with the Gaussian resolution function. But the figure of merit above was based on a Gaussian resolution function, you might complain, and so is not valid here. Figure 15 counters that argument by showing the MaxEnt reconstructions derived from two data sets obtained by convolving a test object with one or the other of two Gaussian functions whose FWHMs differ by a factor of 10. According to conventional thinking, the figures of merit can be equalized by increasing the total number of counts for the wide Gaussian by a factor of 100. But Fig. 15 shows instead that, to recover the sharpest features, the number of neutrons counts must be increased by many orders of magnitude!

Next, we move on to consider a slightly more complicated case. Let the situation be exactly the same as before, except that now the scattering law is known to consist of a single delta function of not only unknown position but also unknown magnitude. That is, we have a two-dimensional hypothesis space, defined by the magnitude A and position x_0 of the delta function. Again, we want to know what restrictions the data impose on the value of A and x_0 . The likelihood function is now a bump in a two-dimensional space, as illustrated schematically in Fig. 16. To describe the shape of this probability bubble, we need at least three numbers: two for the width in each of the two dimensions and one for the orientation. One way of specifying these numbers is to give the so-called *covariance matrix*, a symmetric 2×2 matrix whose elements tell us the expected uncertainty in the position, $\langle(\delta x_0)^2\rangle$, the expected uncertainty in the magnitude, $\langle(\delta A^2)\rangle$, and how the uncertainty in one af-

fects the uncertainty in the other, $\langle \delta x_0 \delta A \rangle$. After doing some algebra, we find that the correlation term is zero, $\langle \delta x_0 \delta A \rangle = 0$. In other words, the reliability with which we can estimate the position of the delta function has no bearing on the reliability with which we can estimate its magnitude. Thus, in terms of the general schematic picture of Fig. 16, the principal axes of the likelihood probability bubble should lie along the A and x_0 directions. We also find that the instrumental parameters T and w affect the reliability of the inferred magnitude and position of the delta function as follows:

$$\langle (\delta x_0)^2 \rangle \propto w/T \quad \text{and} \quad \langle (\delta A)^2 \rangle \propto 1/Tw.$$

This raises a fundamental question: What do we mean by a figure of merit? The formulae above say that to improve our estimate of the position of the delta function, we should make the width of the Gaussian resolution function as narrow as possible, but to improve our estimate of its magnitude, we should make the resolution function as wide as possible!

We can, of course, keep working through specific problems, but we will only come up with the conclusion that different questions, or different choices of hypothesis space, have different answers. So let us try to ask a *generalized* question. We accept that it will not give the exact answer in every specific case but hope that it will yield a sensible figure of merit for a wide range of situations.

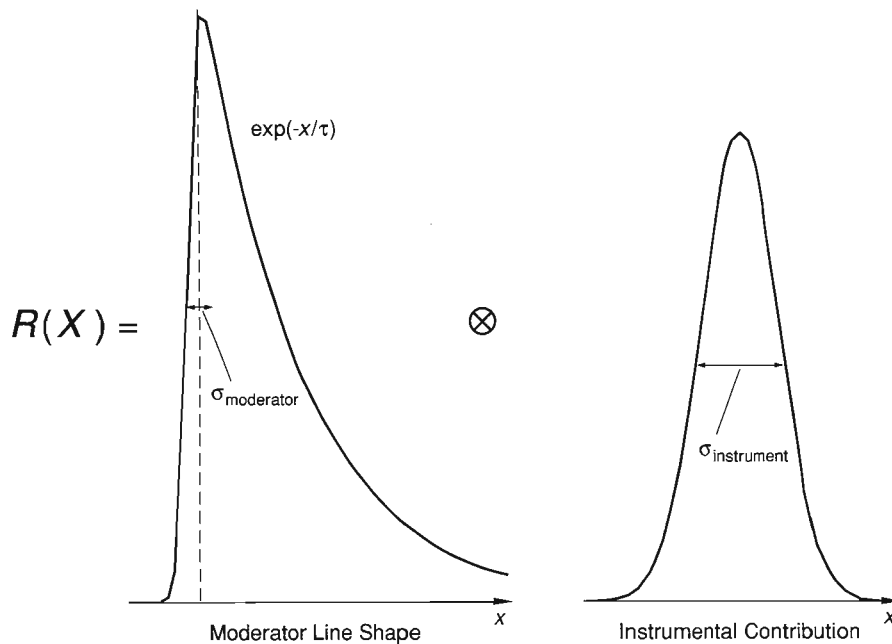
Let us say that the experimental parameters (moderator material, moderator temperature, flight-path length, collimation angle, and so on) all combine to give some resolution function $R(x)$ (not necessarily Gaussian). The question we will ask is: Given that the data are the result of a convolution between the sample scattering law $S(x)$ and the resolution function $R(x)$, how reliably can we estimate the scattering law assuming no particular functional form for $S(x)$?

Since we do not have a functional form for the scattering law, as we did before, an obvious hypothesis space to choose is the one defined by the values of $S(x)$ specified on a grid finely digitized in x . That is, we have an M -dimensional hypothesis space, where M is very large. The likelihood function is now a bump in a multi-dimensional space, and we can consider Fig. 16 as a schematic two-dimensional slice through that space if the axis labels are changed to read $S(x_i)$ and $S(x_j)$ instead of A and x_0 . The spread of this multi-dimensional probability bubble about its maximum will, of course, give us a measure of how well the data constrain the permissible scattering laws. However, since the likelihood bubble is, in general, skew with respect to our $\{S(x_j)\}$ axes, its width is difficult to describe. It is convenient, therefore, to rotate our coordinate axes from the original $\{S(x_j)\}$ axes to another set of axes that lie along the principal axes of the probability bump; the spread of the bubble is then given simply by its widths along the new coordinate axes. These principal axes are vectors in the coordinates $\{S(x_j)\}$ and hence represent relative pixel heights in our digitized x coordinate—they are discretized functions of x . Formally, the principal axes are called *eigenvectors* or, if we go to the continuum limit by making the digitized grid infinitesimally fine, *eigenfunctions*.

The eigenfunctions define the natural hypothesis space for our problem because they represent the properties of the scattering law that can be estimated independently of each other. If we write the required scattering law as a linear combination of the eigenfunctions, $S(x) = \sum a_j \eta_j(x)$, where $\eta_j(x)$ are the eigenfunctions and a_j are coefficients (or parameters) that are now to be determined from the data, then we find that the reliability of our estimate of one parameter does not affect the reliability of our estimate of another; that is, the covariance matrix is diagonal ($\langle \delta a_i \delta a_j \rangle = 0$). The widths of the likelihood function along the principal directions, δa_j , tell us the reliability with which the eigenfunction properties of the scattering law can be estimated; the widths are related to the so-called *eigenvalues* λ by $\langle (\delta a_j)^2 \rangle = 2/\lambda_j$.

RESPONSE MATCHING

Fig. 17. The resolution function $R(x)$ appropriate to a neutron-scattering experiment at a spallation source is a convolution (\otimes) of two response functions: the moderator line shape and an instrumental contribution. The moderator line shape (the time spectrum of neutrons exiting the moderator) has a sharp leading edge, which can be regarded as the rising edge of a narrow Gaussian with a FWHM, $\sigma_{\text{moderator}}$, determined by the moderator material, and a long tail that decays roughly exponentially with a decay constant τ determined by the "poison" added to the moderator. The instrumental contribution is roughly Gaussian with a FWHM of $\sigma_{\text{instrument}}$. The analysis presented in the text indicates that $\sigma_{\text{instrument}}$ should probably be matched to $\sigma_{\text{moderator}}$ (rather than to the FWHM of the moderator line shape as a whole) to obtain the "best" $R(x)$.



If we were to carry out the algebra for our problem, making suitable (usually reasonable) assumptions to obtain an analytic solution, we would find that the eigenfunctions $\eta_\omega(x)$ and their corresponding eigenvalues λ_ω are given by

$$\eta_\omega(x) = \cos(\omega x) \text{ and } \sin(\omega x)$$

$$\text{and } \lambda_\omega = \frac{2}{\sigma^2} |\tilde{R}(\omega)|^2,$$

where $\tilde{R}(\omega)$ is the Fourier transform of the resolution function $R(x)$ and σ^2 is a measure of the average number of counts in the data. This solution tells us that if we do not have a functional form for the scattering law, then we should express it in terms of a Fourier series (a sum of sine and cosine functions). The advantage of doing so is that the reliability with which we can estimate one Fourier coefficient will not affect the accuracy with which we can determine another—it is an uncorrelated space. Since the reliability with which we can estimate any Fourier coefficient is inversely proportional to the corresponding eigenvalue, $\langle (\delta a_\omega)^2 \rangle = 2/\lambda_\omega$, we can use λ_ω as a figure of merit for inferring structure in the scattering law with detail $\delta x \approx 1/\omega$.

The implications of this analysis for instrument design are as follows.

- A versatile figure of merit depends largely on the Fourier transform of the resolution function rather than on its full width at half maximum. This result is illustrated in Fig. 14: The two resolution functions in Fig. 14b have the same full width at half maximum and the same integrated intensity, but, as shown in Fig. 14d, the Fourier transform of the one with the sharp edge does not decay as rapidly with increasing frequency ω as the Fourier transform of the Gaussian resolution function. Resolution functions that have sharp features, therefore, allow high-frequency information to be recovered reliably from the data. An electrical engineer would say that the figure of merit is governed by the *bandwidth* of the resolution function.
- The figure of merit for a given resolution function is not constant but depends on the amount of detail required in the inferred scattering law.
- The background signal has not been forgotten; it enters the figure of merit through the dependence on average number of counts, or σ^2 . Any long decaying tail of the resolution function reduces the figure of merit in the same way that the background does, since such a tail adds to the average number of counts but does not contribute to the Fourier term $\tilde{R}(\omega)$ at high frequency.

Since the resolution function in neutron scattering depends on details of the spectrometer and moderator, our results suggest a potential revision of ideas on the design of neutron-scattering facilities. Take, for example, the *matching* of resolution elements on a neutron spectrometer at an accelerator-based source, which is illustrated in Fig. 17. The resolution function for an experiment is the resultant of a convolution between a roughly Gaussian instrumental contribution (flight-path length, collimation angle, and so on), and the moderator line shape (the time spectrum of the pulse of neutrons leaving the moderator). The moderator line shape has a sharp rising edge, the sharpness of which is governed by the moderator material, and a long decaying tail, the decay of which is governed by the “poison” added to the moderator. The question is how to choose the width of the instrumental component so as to get the “best” resultant resolution function. Conventional wisdom recommends that we should make the width of the Gaussian-like instrumental contribution comparable to the width of the moderator line shape. The analysis above, however, suggests that following this advice could seriously impair our ability to infer (reliably) the scattering law at high resolution and that we should probably match the width of the instrumental component to the narrow width of the sharp leading edge of the moderator line shape. How such considerations translate into the optimal choice of collimation angle, of flight-path length, of moderator material, and of a moderator “poison” is the subject of ongoing research. ■

Further Reading

Bayesian Methods: Historical Background, Fundamentals, and Tutorials

T. Bayes. 1763. *Philosophical Transactions of the Royal Society of London*, 330–418.

P. S. de Laplace. 1812. *Theorie Analytique des Probabilités*. Paris.

Harold Jeffreys. 1939. *Theory of Probability*. Oxford: Oxford University Press. (A paperback version of the third (1961) edition of this work was issued in 1983.)

R. T. Cox. 1946. Probability, frequency and reasonable expectation. *American Journal of Physics* 14: 1–13.

E. T. Jaynes. 1983. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz. Dordrecht, The Netherlands: D. Reidel Publishing Company. (Reprinted in 1989 as a Pallas Paperback by Kluwer Academic Publishers, Dordrecht, The Netherlands.)

E. T. Jaynes. 1986. Bayesian methods: An introductory tutorial. In *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice. Cambridge: Cambridge University Press.

Maximum Entropy

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 623–656.

S. F. Gull and G. J. Daniell. 1978. Image reconstruction from incomplete and noisy data. *Nature* 272: 686–690.

John E. Shore and Rodney W. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory* IT-26: 26–37.

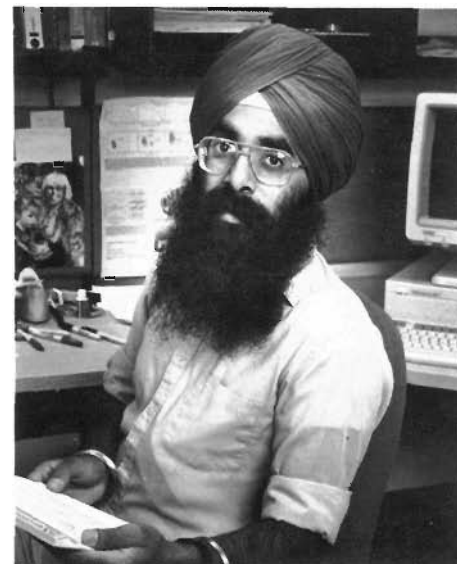
S. F. Gull and J. Skilling. 1984. Maximum entropy method in image processing. *IEE Proceedings* 131, Part F: 646–659.

J. Skilling and R. K. Bryan. 1984. Maximum entropy image reconstruction: General algorithm. *Monthly Notices of the Royal Astronomical Society* 211: 111–124.

Some Recent Developments on Maximum Entropy

Stephen F. Gull. 1989. Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods (Cambridge, 1988)*, edited by J. Skilling. Dordrecht, The Netherlands: Kluwer Academic Publishers.

John Skilling. 1989. Classic maximum entropy. In *Maximum Entropy and Bayesian Methods (Cambridge, 1988)*, edited by J. Skilling. Dordrecht, The Netherlands: Kluwer Academic Publishers.



Devinderjit Singh Sivia has been a postdoctoral fellow at Los Alamos National Laboratory since 1988, holding a joint appointment with the Theoretical Division and the Manuel Lujan, Jr. Neutron Scattering Center. He received his B.A. in the Natural Sciences Tripos at Cambridge University, in 1984, and continued his studies there with the Radio-Astronomy Group at the Cavendish Laboratory, receiving his Ph.D. in 1988. He has been working on the applications of maximum entropy and Bayesian methods, both as a graduate student and a postdoctoral fellow, in a wide variety of fields including x-ray crystallography, very-long-baseline interferometry, “phaseless holography,” neutron scattering, medical imaging, and theoretical condensed-matter physics.

Acknowledgments

The work reported here involved close collaboration with several colleagues, including Richard Silver, Roger Pynn, Peter Vorderwisch, Rex Hjelm, and Mohana Yethiraj. I am also greatly indebted to Steve Gull and John Skilling, who brought me up on a good dose of MaxEnt and Bayesian ideas. This research was supported by the Office of Basic Energy Sciences of the U.S. Department of Energy.

R. K. Bryan. 1990. Solving oversampled data problems by maximum entropy. In *Maximum Entropy and Bayesian Methods (Dartmouth College, 1989)*, edited by P. Fougere. Dordrecht, The Netherlands: Kluwer Academic Publishers.

S. Sibisi. 1990. Quantified MaxEnt: An NMR application. In *Maximum Entropy and Bayesian Methods (Dartmouth College, 1989)*, edited by P. Fougere. Dordrecht, The Netherlands: Kluwer Academic Publishers.

J. Skilling. 1990. Quantified maximum entropy. In *Maximum Entropy and Bayesian Methods (Dartmouth College, 1989)*, edited by P. Fougere. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Examples of the Use of MaxEnt at LANSCE and Related Papers

B. N. Brockhouse, H. E. Abou-Helal, and E. D. Hallman. 1967. Lattice vibrations in iron at 296°K. *Solid State Communications* 5: 211–216.

J. W. Lynn. 1975. Temperature dependence of the magnetic excitations in iron. *Physical Review B* 11: 2624–2637.

A. D. Taylor, E. J. Wood, J. A. Goldstone, and J. Eckert. 1984. Lineshape analysis and filter difference method for a high intensity time-of-flight inelastic neutron scattering spectrometer. *Nuclear Instruments and Methods in Physics Research* 221: 408–418.

T. J. Newton. 1985. Blind deconvolution and related topics. Ph.D. thesis, Cambridge University.

R. A. Robinson, R. Pynn, and J. Eckert. 1985. An improved constant-Q spectrometer for pulsed neutron sources. *Nuclear Instruments and Methods in Physics Research* A241: 312–324.

F. Mezei and P. Vorderwisch. 1989. Spectroscopy with asymmetric resolution functions: Resolution improvement by an on-line algorithm. *Physica B* 156 and 157: 678.

P. A. Seeger, R. P. Hjelm, Jr., and M. J. Nutter. 1989. The low-Q diffractometer at the Los Alamos Neutron Scattering Center. *Molecular Crystals and Liquid Crystals* 180A: 107–117.

R. P. Hjelm, P. Thiyagarajan, D. S. Sivia, P. Linder, H. Alken, D. Schwahn. 1990. Small-angle neutron scattering from aqueous mixed colloids of lecithin and bile salts. Accepted for publication in *Colloid and Polymer Science*.

Devinderjit Singh Sivia. 1990. Applications of maximum entropy and Bayesian methods in neutron scattering. In *Maximum Entropy and Bayesian Methods (Dartmouth College, 1989)*, edited by P. Fougere. Dordrecht, The Netherlands: Kluwer Academic Publishers.

D. S. Sivia, P. Vorderwisch, and R. N. Silver. 1990. Deconvolution of data from the filter difference spectrometer: From hardware to maximum entropy. Accepted for publication in *Nuclear Instruments and Methods in Physics Research*.

M. Yethiraj, R. A. Robinson, D. S. Sivia, J. W. Lynn, and H. A. Mook. A neutron scattering study of the magnon energies and intensities in iron. Submitted to *Physical Review B*.

Optimal Instrument Design

A. Michaudon. 1963. The production of moderated neutron beams from pulsed accelerators. *Journal of Nuclear Energy Parts A/B* 17: 165–186.

D. H. Day and R. N. Sinclair. 1969. Neutron moderator assemblies for pulsed thermal neutron time-of-flight experiments. *Nuclear Instruments and Methods* 72: 237–253.

C. G. Windsor. 1981. *Pulsed Neutron Scattering*. London: Taylor and Francis Ltd.

R. N. Silver, D. S. Sivia, and R. Pynn. 1989. Information content of lineshapes. In *Advanced Neutron Sources 1988*, edited by D. K. Hyer. Bristol: Institute of Physics.

D. S. Sivia, R. N. Silver, and R. Pynn. 1990. Optimization of resolution functions for neutron scattering. *Nuclear Instruments and Methods in Physics Research* A287: 538–550.

The quotation that opens this article appears in *Statistics for Nuclear and Particle Physicists* by Louis Lyons (Cambridge University Press, 1986).