GEOPHYSICS, VOL. 66, NO. 1 (JANUARY-FEBRUARY 2001); P. 55–69, 4 FIGS.

Tutorial

A Bayes tour of inversion: A tutorial

Tadeusz J. Ulrych*, Mauricio D. Sacchi‡, and Alan Woodbury**

INTRODUCTION

It is unclear whether one can (or should) write a tutorial about Bayes. It is a little like writing a tutorial about the sense of humor. However, this tutorial is about the Bayesian approach to the solution of the ubiquitous inverse problem. Inasmuch as it is a tutorial, it has its own special ingredients. The first is that it is an overview; details are omitted for the sake of the grand picture. In fractal language, it is the progenitor of the complex pattern. As such, it is a vision of the whole. The second is that it does, of necessity, assume some ill-defined knowledge on the part of the reader. Finally, this tutorial presents our view. It may not appeal to, let alone be agreed to, by all.

Our presentation relies heavily on the work of others, and there is a wealth of literature on the subject of Bayesian inversion. Fortunately for us, this is not a review, and we are therefore less obligated to quote the literature. There are, however, must-read papers and books. We highly recommend Scales and Tenorio (2001), which contains many excellent references. Two articles by Duijdnam (1988a,b) are an excellent introduction and reference to seismic applications. Edwin Jaynes (1996) is one of our published heroes. We also frequently refer to works by Lupton (1993) and Sivia (1996).

An inherent feature of any inverse problem is randomness. As we will see, randomness may be associated with various parts of our quest, but there can certainly be no doubt that noise always associated with the observations is indeed random. Thus, our approach must be statistical in nature. Statistics, to many, imply probabilities. Probabilities, at least to us, imply Bayes. This is not the only view; in fact, we consider two quite different views. In the first, we consider the model parameters to be a realization of a random variable. In the second, we treat the parameters as nonrandom.

The concept of probability, for reasons that are most probably associated with early education, instill fear into the most lionhearted. In fact, it is probably not probability, per se, but the dreaded word statistics that is at the source, conjuring pictures of dreadfully complicated and never-ending mathematical formulae.

Probability theory and statistics are different. The former refers to the quest of predicting properties of observations from probability laws that are assumed known. The latter is, in a sense, the inverse. We observe data and wish to infer the underlying probability law. In general, inverse problems are more complex to solve than forward problems. They are often ill posed or nonunique. (In fact, Jaynes teaches us that all inverse problems that are overdetermined are badly posed.)

Having said this, we firmly believe that statistical fear is ill founded. After all, we use statistics and probability every day. Every decision we make has an element of chance or arbitrariness involved. In fact, it is our thesis that we live our lives while implementing Bayes' theorem. His theorem is our algorithm of decision-making.

A bit about Bayes

Thomas Bayes was born in London in 1702 into a religious atmosphere. His father, the Rev. Joshua Bayes, was one of the first six Nonconformist ministers to be ordained in England. Like his father, Thomas was ordained a Nonconformist minister and assisted his father until the late 1720s when he became a Presbyterian minister.

Bayes' theory of probability appeared posthumously in "Essay Towards Solving a Problem in the Doctrine of Chances," published in the *Philosophical Transactions of the Royal Society*

Manuscript received by the Editor November 9, 1999; revised manuscript received May 31, 2000.
*University of British Columbia, Department of Earth and Ocean Sciences, 2219 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada. E-mail: ulrych@geop.ubc.ca.
‡University of Alberta, Department of Physics, Avadh Bhatia Physics Laboratory, Edmonton, Alberta T6G 2J1, Canada. E-mail: sacchi@phys.ualberta.ca.
**University of Manitoba, Department of Civil and Geological Engineering, 15 Gillson Street, Winnipeg, Manitoba R3T 5V6, Canada. E-mail: woodbur@cc.umanitoba.ca.

of London in 1764. Thomas Bayes died in England in 1761, misunderstood by many but on a probabilistic par with an immortal, Pierre Simon Marquis de Laplace.

THE THEOREM

Consider two events, A and B. Let us designate the discrete probability of each event happening by $P(A)$ and $P(B)$. The probability of both A and B happening, designated as the joint probability $P(A, B)$ [or $P(A \text{ and } B)$], is given by

$$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A), \qquad (1)$$

where $P(A \mid B)$ and $P(B \mid A)$ are conditional probabilities. In other words, $P(A \mid B)$ is the probability of A happening, given that B has already happened. From equation (1) we obtain Bayes' theorem in its simplest form.

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}. \qquad (2)$$

There is no controversy whatsoever regarding equation (2). It follows logically, and so does Bayes' theorem, from the accepted axioms of probability theory. To gain deeper insight both into the controversy that we will elaborate on and into the use of this theorem, let us state it in a more general fashion. Let $\mathbf{m}^T = (m_1, m_2, \ldots, m_N)$ be a column vector of model parameters of length N. We make M observations concerning this model and obtain a data vector $\mathbf{d}^T = (d_1, d_2, \ldots, d_M)$. We write Bayes' theorem [equation (2)] in terms of probability distributions [or probability density functions (pdf)] as

$$p(\mathbf{m} \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid \mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}. \qquad (3)$$

We now discuss some salient points concerning equation (3). First, $p(\mathbf{m} \mid \mathbf{d})$ is the pdf we desire. It is the distribution of the model parameters posterior to the data $\mathbf{d}$, or the probability that the model is correct given a data set $\mathbf{d}$. In other words, solving for $p(\mathbf{m} \mid \mathbf{d})$ will answer the fundamental question, "What is the probability that the model is correct, given a certain set of observations (data)?"

Second, since the data have been measured, the denominator of equation (3)—the probability that data $\mathbf{d}$ is observed—is a constant and can be written as $p(\mathbf{d}) = \int p(\mathbf{d} \mid \mathbf{m})p(\mathbf{m})\, d\mathbf{m}$ to ensure that $p(\mathbf{m} \mid \mathbf{d})$ integrates to one as all legitimate pdf's should.

The value $p(\mathbf{d} \mid \mathbf{m})$ deserves special attention. It is called the likelihood function and has a long and important history in stochastic estimation problems. Before $\mathbf{d}$ has been observed, $p(\mathbf{d} \mid \mathbf{m})$ represents the pdf associated with possible data realizations for a fixed parameter vector. After observation, $p(\mathbf{d} \mid \mathbf{m})$ has a very different interpretation: it is the likelihood of obtaining the realization actually observed as a function of the parameter vector $\mathbf{m}$. Some authors call this function $L(\mathbf{d} \mid \mathbf{m})$; we call it $L(\mathbf{m})$ for short. It has the same form as $p(\mathbf{d} \mid \mathbf{m})$, but the interpretation is that $\mathbf{d}$ is fixed and $\mathbf{m}$ is variable. Examples follow.

Finally, $p(\mathbf{m})$ is at the center of any disputes that have arisen concerning the use of Bayes' theorem. It is the prior probability of the model vector. Possible questions that arise are "How do we convert any prior information that we might have (i.e., background geological information about an area such as logs, core samples, etc.,) into a pdf, and what are the ramifications associated with the particular choice?" More later. At this stage, we emphasize that the Bayesian approach is one where the model parameters are considered to be random and a prior probability is therefore appropriate if one has a Bayesian disposition.

A LITTLE ABOUT PRIORS

We will have much to say concerning priors throughout this article; but at the outset, we consider the important quest for priors that express complete ignorance so our estimates will not be biased by uncertain knowledge. We quote from Jeffreys (1939):

> Our first problem is to find a way of saying that the magnitude of a parameter is unknown, when none of the possible values need special attention. Two rules appear to cover the commonest cases. If the parameter may have any value in a finite range, or from $-\infty$ to $+\infty$, its prior probability should be taken as uniformly distributed. If it arises in such a way that it may conceivably have any value from 0 to ∞, the prior probability of its logarithm should be taken as uniformly distributed.

We have incorporated this quotation for three reasons: (1) because of Jeffreys' enormous contributions and stature, (2) because the uninformative prior is central in our problem, and (3) because the second part of this quotation defines the Jeffreys prior. Jeffreys, in considering the prior for the standard error σ, which can never be negative, invoked the uninformative prior $1/\sigma$ that is in standard use. Certainly, the notion of the uniform prior to describe complete ignorance for a parameter that can assume all values—the so-called location parameter—does not require detailed justification [although subtle arguments exist; c.f. Scales and Tenorio (2001)]. The case for the Jeffreys prior is perhaps less obvious. Sivia (1996) gives the following clear justification. Consider a parameter σ that can assume only positive values, often called a scale parameter. Let σ, for example, represent standard deviation. We are interested in assigning a pdf of $p(\sigma \mid I)$, where I represents complete prior ignorance with respect to σ. Clearly, our pdf must be invariant to the units of measurement of σ. For this to be so, it is required that

$$p(\sigma \mid I)\, d\sigma = p(\alpha\sigma \mid I)\, d(\alpha\sigma) = p(\alpha\sigma \mid I)\alpha\, d\sigma,$$

where α is a scale factor. For equality to hold, we must have

$$p(\sigma \mid I) \propto \frac{1}{\sigma}, \qquad (4)$$

since only then

$$k\frac{1}{\sigma}\, d\sigma = k\frac{1}{\alpha\sigma}\alpha d\sigma,$$

where k is a constant of proportionality.

A somewhat different but interesting derivation can be obtained as follows. Since $\sigma \geq 0$, we write

$$\sigma = e^{\beta}, \qquad (5)$$

where β is a parameter that can assume all values in $(-\infty, +\infty)$ with uniform probability. To find the pdf of σ, we compute the

Jacobian of the transformation and write

$$p(\sigma) = \left.\frac{p(\beta)}{e^{\beta}}\right|_{\beta=\ln\sigma}.$$

Assuming a uniform distribution for β,

$$p(\sigma) \propto \frac{1}{\sigma}.$$

Following equivalent logic, we deduce that

$$p(\log\sigma \mid I) = \text{constant},$$

confirming our intuition that magnitudes involve logarithms.

We notice immediately that the prior does not represent a proper pdf in the sense that it cannot be normalized. We consider this fact in more detail below; but at this stage we can state with confidence that this characteristic is of little concern, and we now have an uninformative prior for many important applications.

A SIMPLE EXAMPLE OR TWO

Let's illustrate the discussion thus far with examples based on some in Lupton (1993).

Consider the almost canonical problem of estimating the mean from n observations drawn from a Gaussian distribution designated by $N(\mu, \sigma^2)$, where μ is the actual mean and σ^2 the actual variance, assumed known. John Scales would call such an example a toy example (Scales and Tenorio, 2001). The beauty of toy examples is that they teach us much about the general problem in a very simple setting.

To use equation (3), we must first assign a prior density $p(\mathbf{m})$. We do this in the conventional manner by assuming that all mean values μ are equally probable and assign a uniform pdf. The likelihood function is the probability of obtaining the observed sample if we know that the mean was some particular value of μ. If the observations are independent and normally distributed, we obtain

$$p(\mathbf{d} \mid \mathbf{m}) = \frac{1}{(2\pi)^{n/2}\sigma^n}\prod_i e^{-(d_i-\mu)^2/2\sigma^2}. \tag{6}$$

Substituting our uniform prior into equation (3), it is clear that $p(\mathbf{m} \mid \mathbf{d}) \propto p(\mathbf{d} \mid \mathbf{m})$. Taking logarithms of $p(\mathbf{m} \mid \mathbf{d})$ to make life easier, differentiating, and setting the resulting expression to zero in the usual manner obtains the maximum a posteriori (MAP) estimate as

$$\hat{\mu} = \frac{1}{n}\sum_i d_i.$$

Lupton (1993) continues with this example, and we continue with him. Specifically, we now consider the estimation of the variance, assuming this time that μ is known. With equivalent assumptions on the prior distribution of σ^2, just a little algebra shows that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (d_i - \mu)^2. \tag{7}$$

Here, however, is the first point of contention. Equation (7) was derived under the hypothesis of a uniform prior, but variance cannot be negative. A possible prior in such a case is the Jeffreys prior we developed. Substituting equation (6) into equation (3) and proceeding as before yields

$$\hat{\sigma}^2 = \frac{1}{n+1}\sum_i (d_i - \mu)^2. \tag{8}$$

This result may appear somewhat disconcerting. After all, the estimator is biased (the denominator is $n+1$ rather than n for known μ). We will look at this result in more detail later when we introduce the concept of risk. For now, let's itemize and discuss our findings.

1) From a Bayesian perspective, the parameter vector is a realization of a random variable. As such, we can associate with it a prior probability that is essential in applying Bayes' theorem to find the posterior probability. Once we have computed $p(\mathbf{m} \mid \mathbf{d})$, we have computed everything we wish to know about the model.
2) The uniform prior (in some particular range) expresses our maximal ignorance concerning a variable that occupies that range. When positivity constraints apply to that variable, the uniform prior may no longer be justifiable and the Jeffreys prior should be considered.
3) Equation (8) shows clearly that the influence of the prior decreases with sample size. Lupton (1993) puts it nicely: the greater amount of information in the sample drowns out the information in the prior.
4) The role and the meaning of the likelihood are paramount. Certainly, the likelihood $L(\mathbf{m})$ has a meaning and use quite apart from the function it plays in Bayes' theorem. We will explore this issue next.

LIKELIHOOD AND THINGS

Likelihood explains the data, the fixed set of observations $\mathbf{d}$ actually obtained. In $L(\mathbf{m})$, $\mathbf{m}$ represents a vector of parameters that could have given rise to the observed data. If $\mathbf{m}$ is considered a random variable, it is naturally associated with a prior pdf. If it is not considered random, $p(\mathbf{m})$ plays no part and, having only $L(\mathbf{m})$ at our disposal, we obtain the well-known maximum likelihood estimates. In other words, if we adopt the view that the model vector is not random, Bayes' theorem does not apply. There is danger in using anything indiscriminately; this certainly applies to the use of the Bayesian approach.

The best vision of the likelihood function comes from its relationship to the ubiquitous method of least squares. The observed data $\mathbf{d}$ are modeled as the superposition of true signal $\mathbf{s}$ corrupted by additive noise $\mathbf{n}$, i.e., $\mathbf{d} = \mathbf{s} + \mathbf{n}$, where $\mathbf{s}$ is related to $\mathbf{m}$ by $f(\mathbf{m}) = \mathbf{s}$ and f represents some functional relationship. The likelihood function is constructed by taking the difference between observed data and the signal. This difference is the noise to which we can assign the most uninformative pdf that is consistent with the available information. The well-supported, common practice is to assign the Gaussian pdf. This assignment follows from the central limit theorem as well as from the principle of maximum entropy (Jaynes, 1982) that we describe later. Further, supposing that the noise values $(n_1, n_2, \ldots, n_N)$ are independent, we obtain

$$\begin{aligned} L(\mathbf{m}) &= \prod_{i=1}^{N}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp -\left(\frac{n_i^2}{2\sigma^2}\right)\right] \\ &= \sigma^{-N}\exp\left(\frac{1}{2\pi\sigma^2}\sum_{i=1}^{N}[d_i - f(m_i)]^2\right). \end{aligned} \tag{9}$$

Finding the model parameters that minimize the sum in the exponent of equation (9) is the ubiquitous method of least squares. Of course, in the underdetermined case we must use the singular value decomposition that allows construction of the smallest model (see related issues in Farqhuarson and Oldenburg, 1998). The procedure that, in the case of the Gaussion p.d.f. is equivalent, is to maximize $L(\mathbf{m})$, and represent the method of maximum likelihood. We see immediately that the latter method is much more flexible since it does not depend on assigning a Gaussian pdf.

Nonrandom model vector

Let us look in more detail at the fixed model vector approach. A logical thing to do might be to modify the Bayes procedure to eliminate the expectation (or average) over $p(\mathbf{m})$ since $\mathbf{m}$ is now a constant. For a start, we consider a mean squared error (MSE) criterion, written as

$$\mathrm{MSE}(\mathbf{m}) = \int_{-\infty}^{+\infty} (\hat{\mathbf{m}} - \mathbf{m})^2 p(\mathbf{d} \mid \mathbf{m})\, d\mathbf{d}, \qquad (10)$$

where we have taken expectations over the only random variable in the problem that is $\mathbf{d}$. We see immediately that a minimum MSE($\mathbf{m}$) is obtained when $\hat{\mathbf{m}} = \mathbf{m}$. True but hardly useful. We want to obtain an unbiased estimate, $E[\hat{\mathbf{m}}] = \mathbf{m}$, that at the same time has minimum variance. The way most people do that is to maximize the likelihood $L(\mathbf{m})$ [or, more often, the natural logarithm of $L(\mathbf{m})$, called $l(\mathbf{m})$] with respect to $\mathbf{m}$ to obtain the maximum likelihood estimate we mentioned previously. The literature abounds with details of the properties of the maximum likelihood estimator, its bias and variance for various probability distributions, and its relationship to the method of least squares. The important conclusion to be drawn in the present discussion stems from the relationship between the MAP and maximum likelihood estimators. Taking logarithms in equation (3), differentiating, and setting the result to zero, we obtain

$$\left.\frac{\partial}{\partial \mathbf{m}} l(\mathbf{m})\right|_{\mathbf{m}=\hat{\mathbf{m}}} = \left.\frac{\partial}{\partial \mathbf{m}} \ln p(\mathbf{d} \mid \mathbf{m})\right|_{\mathbf{m}=\hat{\mathbf{m}}} + \left.\frac{\partial}{\partial \mathbf{m}} \ln p(\mathbf{m})\right|_{\mathbf{m}=\hat{\mathbf{m}}} = 0. \qquad (11)$$

Clearly, the maximum likelihood estimator is equivalent to the MAP estimator when $\ln p(\mathbf{m}) = 0$. This implies that, in practice, a random hypothesis for the model parameters coupled with a uniform prior density is equivalent to the fixed vector hypothesis.

THE CONTROVERSY

The controversy that exists between the Bayesian and classical approaches arises because, as so eloquently stated by Frieden (1982), these are fundamentally different. The classical approach is not wrong. [J. R. Oppenheimer, paraphrased by Frieden (1982), states that "The word 'Classical' means only one thing in science: it's wrong!"] Both views are important, and both give answers to the problems that concern us.

The main aspect to the controversy lies in the use of prior information. The classical, or frequentist, view is that prior information makes no sense since probabilities can only be measured and not assigned. In this regard, the result obtained in equation (8) may indeed be gratifying to the frequentists. Here is an estimate of variance that differs from the accepted maximum likelihood estimate. But is it worse? Trade-offs exist in all walks of life. In parameter estimation, the most famous is the trade-off between bias and variance, expressed by the relation

$$\text{mean square error} = \text{variance} + \text{bias}^2.$$

It is interesting to see how this expression arises.

Writing $\hat{\Theta}$ to be an estimate of a parameter θ and using the definitions of bias (B) and variance $Var[\,\cdot\,]$,

$$B = E[\hat{\Theta}] - \theta,$$
$$Var[\hat{\Theta}] = E[(\hat{\Theta} - E[\hat{\Theta}])^2],$$

we anticipate the expression for MSE by computing the sum

$$\begin{aligned} Var[\hat{\Theta}] + B^2 &= E[(\hat{\Theta} - E[\hat{\Theta}])^2] + (E[\hat{\Theta}] - \theta)^2 \\ &= E[\hat{\Theta}^2] - (E[\hat{\Theta}])^2 + (E[\hat{\Theta}])^2 \\ &\quad -2E[\hat{\Theta}]\theta + \theta^2 \\ &= E[\hat{\Theta}^2] - 2E[\hat{\Theta}]\theta + \theta^2 \\ &= E[(\hat{\Theta} - \theta)^2] \\ &= \mathrm{MSE}. \end{aligned} \qquad (12)$$

One oft-cherished compromise between variance and bias is to minimize the MSE. Let us now look at the MSE($\hat{\sigma}^2$) associated with the Jeffreys prior.

We consider n normally distributed variates with known mean μ where $\mu = 0$. In this example, we are concerned with the estimation of the variance for known mean, not the estimation of the sample variance. The estimator $\hat{\sigma}^2$ is computed as

$$\hat{\sigma}^2 = \frac{1}{k} \sum_{i=1}^{n} x_i^2, \qquad (13)$$

where $k = n$ for the maximum likelihood estimator and $k = n + 1$ for the Jeffreys estimator. The simplest procedure to compute MSE($\hat{\sigma}^2$) is via the characteristic function $\phi(t)$, defined as

$$\begin{aligned} \phi(t) &= E[e^{itx}] \\ &= \int_{-\infty}^{+\infty} e^{itx} p(x)\, dx. \end{aligned}$$

If x is an $N(0, 1)$ variate, the characteristic function of x^2 is (Lupton, 1993)

$$\begin{aligned} \phi_{x^2}(t) &= E\left[e^{itx^2}\right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx^2 - x^2/2}\, dx \\ &= \frac{1}{(1 - 2it)^{1/2}}. \end{aligned}$$

The characteristic function of the estimator in equation (13), when x is distributed as $N(0, \sigma^2)$, is consequently

$$\phi_{\hat{\sigma}^2}(t) = \frac{1}{(1 - 2it\sigma^2/k)^{n/2}}. \qquad (14)$$

To find the mean and variance of the corresponding distribution, we remember that the moments about the origin μ_r' are

obtained from

$$\mu_r' = \left. \frac{d^r \phi}{d(it)^r} \right|_{t=0}.$$

Using equation (14) and performing the necessary algebraic steps, we obtain

$$E[\hat{\sigma}^2] = \mu_1' = \frac{n}{k}\sigma^2$$

and

$$Var[\hat{\sigma}^2] = \mu_2' - \mu^2 = 2\sigma^4 \frac{n}{k^2}.$$

Using these relations, we compute $\mathrm{MSE}(\hat{\sigma}^2)$ as

$$\mathrm{MSE}(\hat{\sigma}^2) = 2\sigma^4 \frac{n}{k^2} + \left(\frac{n}{k}\sigma^2 - \sigma^2\right)^2. \tag{15}$$

Minimizing equation (15) with respect to k obtains $k = n+2$. The minimum MSE estimator for this Gaussian example with known mean is

$$\hat{\sigma}^2 = \frac{1}{n+2}\sum_{i=1}^{n} x_i^2. \tag{16}$$

So, the maximum likelihood estimator is unbiased, the estimator of equation (16) is minimum MSE, and the Jeffreys estimate is somewhere in between. Not a bad choice, perhaps.

Our main reason for using the Bayesian philosophy is because it provides a particularly flexible approach to solving our underdetermined inverse problem. Bayes allows us to construct objective functions that provide particularly desirable flavors to the solutions we seek. Before describing our approach, let's review the steps required to achieve a Bayesian solution to an inverse problem.

INVERSION VIA BAYES

We use our previous notation to state the inverse problem. We have

$$f(\mathbf{m}) = \mathbf{d} = \mathbf{s} + \mathbf{n},$$

where $f(\mathbf{m})$ could represent a linear or nonlinear function. The likelihood, in light of the earlier discussion, is assumed Gaussian. We also assume some knowledge of the data covariance matrix $\mathbf{C}_d$ in terms of the data variances:

$$\mathbf{C}_d = \begin{bmatrix} \sigma_{d1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{d2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{dN}^2 \end{bmatrix}$$

Denoting the determinant of $\mathbf{C}_d$ by $|\mathbf{C}_d|$, we obtain the likelihood as

$$p(\mathbf{d}\,|\,\mathbf{m}) = \frac{1}{\left((2\pi)^N |\mathbf{C}_d|\right)^{1/2}} \exp -\left[\frac{1}{2}(\mathbf{d} - f(\mathbf{m}))^T \times \mathbf{C}_d^{-1}(\mathbf{d} - f(\mathbf{m}))\right]. \tag{17}$$

Now comes the question of assigning the prior probability of our model. We suppose in this example that we know the model is smooth in a first derivative sense. Clearly, this would be an inappropriate supposition for a problem concerning the earth's reflectivity. However, it might be appropriate if the problem were to find a source wavelet. For the smooth prior, we define $p(\mathbf{m})$ by

$$p(\mathbf{m}) = \left(\frac{\eta}{2\pi}\right)^{(M-1)/2} \exp\left[-\frac{\eta}{2}\mathbf{m}^T\mathbf{D}^T\mathbf{D}\mathbf{m}\right], \tag{18}$$

where $\mathbf{D}$ is the derivative matrix and is represented by

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & \vdots \\ 0 & -1 & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & -1 & 1 \end{bmatrix}$$

The parameter η, known as a hyperparameter, characterizes the distribution in that $1/\eta$ corresponds to the variance of the prior distribution. We now call upon Bayes:

$$p(\mathbf{m}\,|\,\mathbf{d}) = \frac{p(\mathbf{d}\,|\,\mathbf{m})p(\mathbf{m})}{\int p(\mathbf{d}\,|\,\mathbf{m})p(\mathbf{m})\,d\mathbf{m}}. \tag{19}$$

Although, the denominator of equation (19) is nothing more than a normalization factor, it plays an important role in the inversion. Specifically, we define $\Lambda(\mathbf{d}\,|\,\eta)$ as

$$\Lambda(\mathbf{d}\,|\,\eta) = \int p(\mathbf{d}\,|\,\mathbf{m})p(\mathbf{m}\,|\,\eta)\,d\mathbf{m}, \tag{20}$$

where $p(\mathbf{m})$ is rewritten as $p(\mathbf{m}\,|\,\eta)$ to indicate its dependence on η. In general, the inverse problem will be cast in terms of a vector $\boldsymbol{\eta}$ of hyperparameters. The value $\Lambda(\mathbf{d}\,|\,\eta)$ is known as the Bayesian likelihood, and we will see how it is used in the inversion. For the time being, we compute the following form for the posterior distribution from equations (17), (18), and (19):

$$p(\mathbf{m}\,|\,\mathbf{d}) = \frac{1}{\left((2\pi)^N |\mathbf{C}_d|\right)^{1/2}} \left(\frac{\eta}{2\pi}\right)^{(M-1)/2} \exp -\left[\frac{1}{2}\Phi(\mathbf{m}, \eta)\right],$$

where

$$\Phi(\mathbf{m}, \eta) = \| \mathbf{C}^{-1}(\mathbf{d} - f(\mathbf{m})) \|^2 + \eta \, \| \mathbf{D}\mathbf{m} \|^2 \tag{21}$$

and represents the objective, or cost, function for the problem. Inasmuch as the MSE is a trade-off between bias and variance, so our cost function is a trade-off between resolution and smoothness—a trade-off that is noise dependent.

We have built our first Bayesian objective function. We will see in the next section more examples of objective function building, an extremely useful approach in diverse applications. As is clear from equation (21), the hyperparameter η plays a central role. As its function would suggest, determining it is not an easy task. If η were known, the likeliest model parameters represented by $\mathbf{m}_{\mathrm{MAP}}$ may be found by minimizing $\Phi(\mathbf{m}, \eta)$.

Since the problem of minimizing $\Phi(\mathbf{m}, \eta)$ is nonlinear, it is linearized around a starting model and the minimization proceeds iteratively. We do not consider details here. We will, however, spend a little time looking into determining η.

Determining the hyperparameters

The approach we describe is based on the work of Hirotugu Akaike (1980) and on the Akaike's Bayesian information criterion (ABIC). The ABIC, in turn, is based on the Akaike information criterion (AIC). [For details, see Sakamoto et al. (1986) and Matsuoka and Ulrych (1986).] In essence, the AIC is based on the Kullback-Liebler information measure. Denoting the true model by $\mathbf{m}_t$ and the estimated model by $\hat{\mathbf{m}}$, the Kullback-Liebler measure defines the distance in probability space between the data distributions $p(\mathbf{d}\,|\,\mathbf{m}_t)$ and $p(\mathbf{d}\,|\,\hat{\mathbf{m}})$. In developing the AIC, Akaike transforms this measure into

$$\text{AIC} = -2\cdot\ln[p(\mathbf{d}\,|\,\hat{\mathbf{m}})]_{\max} + 2\cdot(\text{number of parameters}).$$

The first term in this expression is related to the sample variance and decreases with the number of parameters. The second is related to the fact that the error of fitting the parameters increases with their number. The minimum of the AIC lets us compute the appropriate number of parameters, a quest particularly arduous in problems such as fitting time series models. The ABIC is similar to the AIC in form and is computed in terms of the Bayesian likelihood defined in equation (20):

$$\text{ABIC} = -2\cdot\ln[\Lambda(\mathbf{d}\,|\,\eta)] + 2\cdot(\text{number of hyperparameters}). \quad (22)$$

The correct hyperparameters are evaluated at the minimum value of the ABIC.

The full Bayesian inversion proceeds as follows.

1) Evaluate the Jacobian matrix for the starting model vector $\mathbf{m}_k$.
2) Minimize $\Phi(\mathbf{m}\,|\,\eta)$ for a given value of η.
3) Evaluate the ABIC.
4) Repeat steps 2 and 3 until the optimum value of η that minimizes the ABIC is reached.
5) If convergence properties are satisfied, stop. Otherwise, return to step 1 with the updated model vector $\mathbf{m}_{k+1}$.

The convergence properties in step 5 are related to data uncertainties. An approach is to consider the rms misfit, given by $\sqrt{\|\mathbf{C}^{-1}(\mathbf{d}-f(\mathbf{m})\|^2/N}$.

PARAMETER ERRORS: CONFIDENCE AND CREDIBILITY INTERVALS

It is always important to say something about the confidence with which the model parameters that follow from our inverse computation are reported. Two issues need to be explored. The first entails the probability associated with the particular model we are seeking, given a perfectly noiseless set of observations. The second entails the confidence region for the parameters which results from the noise in the observations. Regarding the first issue, there are infinite possible solutions to our underdetermined inverse problem. Are all possible solutions equally probable? Certainly not. As an example, we consider the die problem, made famous by Jaynes, that has served us so well. This example is particularly suitable because it also allows us to introduce the concept of entropy.

Part I—Model probabilities

Here we consider that part of the problem not associated with errors in the observations. Some models that fit the data are more probable than others. Here is our view of why and how to obtain some measure of such probabilities. First, the canonical die problem.

The die problem.—This problem, originally proposed by Jaynes (1968), is an excellent illustration of underdetermined inversion, or inference, and clearly points out the role of model likelihood. A die of unknown prior probabilities $x_1, x_2, \ldots, x_M$ ($M=6$, of course, but we leave it at M for general use later) is thrown N times. The average number of spots $\bar{n}$, where $\bar{n}=(n_1+2n_2+\cdots+Mn_M)/N$, $N=\sum_{i=1}^{M} n_i$, is recorded (the individual occurrence numbers $n_1, n_2, \ldots, n_M$ are not known). Our problem is to estimate $\mathbf{q}^T=(q_1, q_2, \ldots, q_M)=(n_1/N, n_2/N \ldots, n_M/N)$, the vector of frequencies which describes the experiment.

From elementary principles, we can write the number of ways we can obtain the distribution $(n_1, n_2, \ldots, n_M)$. This number, W, is called the multinomial coefficient or multiplicity and is given by

$$W = \frac{N!}{n_1!n_2!\ldots n_M!}. \quad (23)$$

There is a direct relationship between multiplicity and entropy. We obtain it by means of Stirling's approximation for $\ln N!$:

$$\ln N! = \left(N+\frac{1}{2}\right)\ln N - N + \frac{1}{2}\ln(2\pi) + \frac{N}{12}.$$

This approximation is extraordinarily accurate. Even for $N=2$, for example, it is only in error by 0.05%. Using this relationship in the logarithm of equation (23), we obtain

$$\ln W = -N\sum_{M}\frac{n_i}{N}\ln\frac{n_i}{N} + \frac{1}{2}\left(\ln N - \sum_{M}\ln n_i\right) + \frac{i-M}{2}\ln 2\pi + \frac{1}{12}\left(\frac{1}{N} - \sum_{k}\frac{1}{n_i}\right). \quad (24)$$

Defining $H(\mathbf{q})$ to be the entropy associated with the distribution $\mathbf{q}$, $(q_i=n_i/N)$, we see that the first term in equation (24) is

$$-N\sum_{M} q_i \ln q_i = NH(\mathbf{q}).$$

As N increases, the multiplicity W increases exponentially as $e^{NH(\mathbf{q})}$, and the remaining terms in equation (24) very quickly become negligible. Clearly, if we are looking for a distribution that can be realized in the greatest number of ways, we should maximize the multiplicity. For large N, we should maximize the entropy; hence, the principle of maximum entropy. The principle of maximum entropy can be argued in a much more intuitive manner, beginning with the definition of information [see Jaynes (1982) for a review]. This development allows us to talk about probabilities of solutions in quantitative fashion.

Let us return to the die example. We use the principle of maximum entropy to solve the problem of estimating the unknown frequencies $q_i,\ i=1,6$ given the constraint that $\bar{n}=4.5$ (this is not a fair die, for which the value is 3.5). There is also

the constraint that $\sum_{i=1}^{6} q_i = 1$. We obtain the solution

$$\mathbf{q}_{\mathrm{ME}} = (0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475)$$

with entropy $H_{\mathrm{ME}} = 1.6136$, where ME is maximum entropy. Of course, the maximum entropy solution is not the only one possible. Consider another solution, $\mathbf{q}_{AS}$, that also exactly fits our constraints:

$$\mathbf{q}_{AS} = \left(\frac{3}{10}, 0, 0, 0, 0, \frac{7}{10}\right),$$

where $\mathbf{q}_{AS}$ has entropy $H_{AS} = 0.6110$. The question we posed above is, are both solutions equally plausible? Clearly not. Since when does throwing a die result in no fives coming up? We have deliberately chosen a solution that is implausible to stress the point.

But let us look at another possible solution for a more subtle comparison. We obtain this solution, $\mathbf{q}_{\mathrm{MRE}}$, by means of the principle of minimum relative entropy, which is an important extension of the principle of maximum entropy by Shore and Johnson (1980). [See Woodbury and Ulrych (1998) for a review of this versatile approach to the solution of underdetermined problems.] The $\mathbf{q}_{\mathrm{MRE}}$ solution is, for a uniform prior with a lower bound of zero and an unconstrained upper bound (Ulrych et al., 1990),

$$\mathbf{q}_{\mathrm{MRE}} = (0.0712, 0.0851, 0.1059, 0.1399, 0.2062, 0.3918),$$

with entropy $H_{\mathrm{MRE}} = 1.6091$.

On the other hand, if the upper bound is constrained to the reasonable value of one (Woodbury and Ulrych, 1998), we obtain the minimum relative entropy solution

$$\mathbf{q}_{\mathrm{MRE}} = (0.0678, 0.0826, 0.1055, 0.1453, 0.2232, 0.3755),$$

with entropy $H_{\mathrm{MRE}} = 1.6091$. This solution, interestingly enough, is close to the maximum entropy solution.

Now the question is, how probable is $\mathbf{q}_{\mathrm{MRE}}$ compared to the maximum entropy solution? We can quantify the results by means of the entropy concentration theorem (Jaynes, 1982).

The entropy concentration theorem.—Considering the random experiment described, let $\mathcal{C}$ be the subclass of all possible outcomes that could be observed in N trials which are compatible with the m linearly independent constraints:

$$\sum_{i=1}^{M} A_{ji} q_i = d_j \quad 1 \le j \le m \quad \text{and} \quad m < M. \tag{25}$$

A certain fraction F of outcomes in $\mathcal{C}$ will have entropy in the range

$$H_{\max} - \Delta H \le H(q_1 \ldots q_M) \le H_{\max}, \tag{26}$$

where $H_{\max}$ is obtained by maximizing equation (25) subject to equation (26). The entropy concentration theorem, formulated and proved by Jaynes (1982), states that, asymptotically, $2N\Delta H$ is distributed in $\mathcal{C}$ as χ_k^2 with k, the number of degrees of freedom, equal to $M - m - 1$. Specifically, denoting χ_k^2 at the $100P$ percent significance level by $\chi_k^2(P)$, we obtain

$$2N\Delta H = \chi_k^2(1 - F). \tag{27}$$

Equation (27) is a remarkable expression. It informs us of the percentage chance that the observed frequency distribution will have an entropy outside of the interval computed from equation (27). As shown by Jaynes (1982) in a number of examples, for large N the overwhelming majority of all possible distributions will have an entropy very close to $H_{\max}$.

We now use the entropy concentration theorem to judge the maximum entropy and maximum relative entropy solutions of the die problem. According to the entropy concentration theorem in an experiment consisting of 1000 trials, 99.99% of all outcomes allowed by the constraints have entropy in the range $1.609 \le H \le 1.614$ (Jaynes, 1982). Although maximum entropy is the more probable solution in terms of multiplicities, H_{MRE} is certainly well within the range of the most probable solutions.

We can also compare the multiplicities of the two solutions. For a large number of trials, say, $N = 1000$, we use our approximation to compute the ratio of multiplicities:

$$\frac{W_{\mathrm{ME}}}{W_{\mathrm{MRE}}} = e^{N(H_{\mathrm{ME}} - H_{\mathrm{MRE}})} = e^{10.2} = 26\,903.$$

This ratio tells us that for every way the maximum relative entropy solution can be realized, the maximum entropy solution can be realized in approximately 30 000 ways.

What does all this have to do with real earth inverse problems? We can repeat the die experiment 1000 times, but we do not have that luxury in dealing with the real earth. We do, however, have the luxury of imagination. If we imagine the earth from a random rather than a fixed viewpoint, we can think of the outcome of a measurement on the earth as the outcome of a random experiment. We measure our $\bar{n}$ and deduce the probability associated with the parameters that gave rise to our observation. We can then deduce all manner of model characteristics. We contend that a probabilistic approach is much more flexible than the fixed vector approach. In our view, we can and should say something about the likelihood of possible solutions.

A bit more about prior information.—We have addressed the probability of a particular outcome for an underdetermined problem. In obtaining the maximum relative entropy solution, we used a uniform prior to express our ignorance concerning the unknown distribution. We could have assumed some different prior distribution. Our point is that, having chosen a prior, there is no point in talking about the probability of the prior probability. If we are uncertain about the prior, we use the most uninformative, innocuous one we can. A prior should express knowledge that is not in doubt, e.g., density is positive, compressional wave velocity <10 000 m/s in our survey area, the Sun is bigger than Mars. If we are looking for a buried pipe, it seems appropriate to constrain the gravity inversion with a prior that expresses this structure. If we wish to regularize our inversion by computing the truncated singular value decomposition solution, we should—providing that we have a good a priori reason for choosing the smallest model as the appropriate solution.

Part II—Parameter uncertainties

We now come to the second issue involved with uncertainties of the inverse solution. All observations have errors. These errors propagate through the inversion and attack our parameters. We examine the behavior of such errors both in the fixed and random views that we have introduced.

Since, at this stage, we believe we have chosen an approach to the inversion that gives us the most probably correct result, we

consider the treatment of data errors only. For the fixed vector approach, the customary statistic is χ^2. It is well known, so we will not review it here. We do, however, examine the concept of confidence regions. Once again, we delve into Lupton (1993).

The frequentist approach to confidence intervals—for example, for determining the intervals associated with an estimate of μ when the sample is drawn from an $N(\sigma^2, \mu)$ population—is to compute the estimated mean $\bar{x}$ and to assign probability bounds such as $|\bar{x} - \mu| < 1.96\sigma/\sqrt{N}$ at the 95% confidence level. In this statement, μ is fixed and $\bar{x}$ is the random variable. Bayesians take a different view. Here, μ is the random variable. Given a uniform prior for μ, we can compute $p(\mu \mid \mathbf{x})$ using equation (3) and, hence, the associated confidence regions. In the Gaussian case, we get the same result as we do by means of the conventional approach, but the view is different.

We now complicate the situation by considering the case when σ^2 is unknown. Again, using equation (3) we compute the posterior pdf $p(\mu, \sigma \mid \mathbf{x})$ assuming a Gaussian likelihood, a uniform prior for μ, the Jeffreys prior for σ, and independence of μ and σ. To simplify matters, let s^2 indicate the sample variance, computed as

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

where $\bar{x}$ is the maximum likelihood estimate of the mean. By direct substitution, we can verify that

$$s^2 + (\bar{x} - \mu)^2 = \frac{1}{N}\sum_{i=1,n}^{N}(x_i - \mu)^2.$$

Hence, substituting into equation (3) we obtain

$$p(\mu, \sigma \mid \mathbf{x}) \propto \sigma^{-(N+1)} \exp -\left[\frac{N(s^2 + (\bar{x} - \mu)^2)}{2\sigma^2}\right].$$

We are interested in obtaining the distribution for $p(\mu \mid \mathbf{x})$. To do this, we integrate over σ. This is called marginalization. Integrating, we obtain

$$\begin{aligned} p(\mu \mid \mathbf{x}) &= \int_0^\infty p(\mu, \sigma \mid \mathbf{x})\, d\sigma \\ &\propto \int_0^\infty e^{-N(s^2+(\bar{x}-\mu)^2)/2\sigma^2} d\sigma. \end{aligned}$$

Following the procedure outlined in Lupton (1993),

$$p(\mu \mid \mathbf{x}) \propto \left[1 + \left(\frac{t^2}{\nu}\right)\right]^{-(\nu+1)/2},$$

where $\nu = N - 1$ and $t^2 = (N-1)[(\bar{x} - \mu)^2]/s^2$. This is the well-known t distribution. We are now free to evaluate any type of confidence regions for μ that we wish.

The procedure of obtaining the marginal density involves integration over the parameter that we are not interested in. Such parameters are called nuisance parameters and are dealt with in more detail below.

The title of this subsection contains the word credibility. We credit Press (1989) with this description and adopt it with a sigh of relief. Credibility intervals are to Bayesian estimation what confidence intervals are to likelihood estimation. We need this differentiation because, as we have seen, what is considered to be fixed and random in the two approaches differs completely. From now on, we consider parameter credibility regions.

A little about marginals.—We return to equation (1) with a modification that better encapsulates what we are doing. We considered events A and B; now let us substitute H for B and call it a hypothesis. Then we substitute D for A and call it data ($\mathbf{d}$ will be a realization from D). We now write Bayes' theorem [equation (2)] as

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}. \tag{28}$$

One other modification that extends the simple formulation is to introduce I into equation (28), where I is prior information:

$$P(H \mid D, I) = \frac{P(D \mid H, I)P(H \mid I)}{P(D)}.$$

We now have a more complete probabilistic assignment. Thus, for example, $P(H \mid D, I)$ is the probability of our hypothesis given the data and some prior information.

We now imagine a hypothesis of the form (Bretthorst, 1988)

$$H \equiv f(x) = \sum_{i=1}^{m} a_i \phi_j(x, \{k_x\}),$$

where $f(x)$ is some function of the spatial parameter x that is modeled in terms of the decomposition into m components consisting of m amplitudes and basis functions a_i and $\phi_i, i = 1, 2, \ldots, m$. Each ϕ_i in turn depends on a set of parameters—wavenumbers, for example—that we have designated as $\{k_x\}$. Imagine that, in our problem, what is of consequence to uncover is the set of wavenumbers $\{\phi_i\}$. The amplitudes are of less relevance. The Bayesian approach allows the formulation of the posterior probability so that it is independent of the a_i, or nuisance parameters. To see how this is done, we begin with a very simple problem with two parameters: k_x is desired, and a is the nuisance. In other words, we wish the posterior density $p(k_x \mid D, I)$. We first compute the joint posterior pdf of both k_x and a using Bayes' theorem:

$$p(k_x, a \mid D, I) = \frac{p(D \mid k_x, a, I)p(k_x, a \mid I)}{P(D \mid I)}.$$

We now integrate a out of the equation to obtain the marginal posterior pdf for k_x,

$$p(k_x \mid D, I) = \int p(k_x, a \mid I)\, da,$$

which expresses the information contained about k_x in the data and the prior information, regardless of the value of a. This apparently somewhat miraculous result is easily explained. Consider the joint pdf $p(x, y)$ of two random variables $\mathbf{x}$ and $\mathbf{y}$. The value $p(x)\Delta x$ is the probability of observing $\mathbf{x}$ in the interval $(x, x + \Delta x)$. The value of $\mathbf{y}$ is immaterial and may lie anywhere in the interval $(-\infty, +\infty)$. Hence,

$$\begin{aligned} &\lim_{\Delta x \to 0} p(x)\Delta x \\ &\quad = \text{probability}(x < \mathbf{x} \le x + \Delta x), (-\infty < \mathbf{y} \le +\mathbf{y}) \\ &\quad = \lim_{\Delta x \to 0} \int_x^{x+\Delta x} \int_{-\infty}^{+\infty} p(x, y)\, dx\, dy \\ &\quad = \lim_{\Delta x \to 0} \int_x^{x+\Delta x} \left[\int_{-\infty}^{+\infty} p(x, y)\, dy\right] dx. \end{aligned}$$

The integral inside the brackets is a function of x and is constant over the interval $(x, x + \Delta x)$ since $\Delta x \to 0$. Hence,

$$\lim_{\Delta x \to 0} p(x)\Delta x = \left[\int_{-\infty}^{+\infty} p(x, y)\, dy\right]\left[\lim_{\Delta x \to 0} \int_{x}^{x+\Delta x} dx\right]$$
$$= \Delta x \int_{-\infty}^{+\infty} p(x, y)\, dy. \qquad (29)$$

Therefore,

$$p(x) = \int_{-\infty}^{+\infty} p(x, y)\, dy.$$

We have illustrated the computation of a marginal distribution from a joint distribution of two variables. The strength of this approach is that it is quite general.

Parameter credibility intervals.—We have dealt with marginal distributions, nuisance parameters, and credibility intervals. It is time to put all this together for a more general treatment. We assume that, following the outline presented above, we have computed the a posteriori pdf for our model. We now wish to present credibility intervals associated with each parameter of the model.

The Bayesian approach has led us to an a posteriori pdf that, given the prior distribution, has allowed us to determine a complete statistical description of our model. To obtain the credibility intervals associated with individual model parameters, we obtain the marginal pdf's as follows. For a given parameter m_i,

$$p(m_i) = \int_{m_1} \cdots \int_{m_{i-1}} \int_{m_{i+1}} \cdots$$
$$\times \int_{m_M} p(\mathbf{m} \mid \mathbf{d})\, dm_1 \ldots dm_{i-1} dm_{i+1} \ldots dm_M.$$

We now obtain the means and variances of each marginal to characterize the statistics of the individual model parameters. Of course, as pointed out by Tarits et al. (1994), if the marginal pdf is not unimodal, the interpretation of the moments of the pdf is not simple.

COMPUTATIONAL TRACTABILITY AND MINIMUM RELATIVE ENTROPY

As pointed out by Tarits et al. (1994), the full Bayesian solution may not be tractable for a large number of model parameters, so the common practice is to settle for maximum likelihood or asymptotic solutions. We have developed a Bayesian-like approach: minimum relative entropy, introduced by Shore and Johnson (1980) and extended to the general linear inverse problem by Ulrych et al. (1990). Considerable advance in developing and applying this approach has been made and is summarized in Woodbury and Ulrych (1998). A very brief introduction is presented here for completeness.

We denote the posterior probability of the model by $q^{\dagger}(\mathbf{m})$ [to be differentiated from $p(\mathbf{m} \mid \mathbf{d})$] that we wish to determine from $p(\mathbf{x})$ and information in the form of some expected value constraints

$$\int q^{\dagger}(\mathbf{m}) f_j(\mathbf{m})\, d\mathbf{m} = \bar{f}_j, \qquad (30)$$

where $f_j(\mathbf{m})$ and $\bar{f}_j$, $j = 1, 2, \ldots, M$ are assumed known.

The constraints of equation (30) do not uniquely determine $q^{\dagger}(\mathbf{m})$, but they do restrict the allowable densities. The goal is to construct an estimate $q(\mathbf{m})$ of $q^{\dagger}(\mathbf{m})$ which satisfies the constraints, takes into account $p(\mathbf{m})$, and satisfies the axioms of consistent inference. The solution (Shore and Johnson, 1980) is to minimize $H(q, p)$, the entropy of $q(\mathbf{m})$ relative to $p(\mathbf{m})$, where

$$H(q, p) = \int q(\mathbf{m}) \ln\left[\frac{q(\mathbf{m})}{p(\mathbf{m})}\right] d\mathbf{m}, \qquad (31)$$

subject to the constraints.

The gist of the maximum relative entropy approach is to use first-moment constraints. Thus, in application to a linear problem where the data equations are $\mathbf{d} = \mathbf{F}\hat{\mathbf{m}}$, we treat the model estimate $\hat{\mathbf{m}}$ as the expectation of $\mathbf{m}$. Our data equations become

$$\mathbf{d} = \int q(\mathbf{m})[\mathbf{F}\mathbf{m}]\, d\mathbf{m}$$

and take the place of $\bar{f}_j$ in equation (30). By the same token, we treat the prior model parameters as expectations over the prior pdf that is determined by adopting the principle of maximum entropy. The maximum relative entropy approach is, in our opinion, very flexible and computationally much less demanding than the full Bayesian approach, yet it allows the incorporation of probabilistic constraints. Woodbury and Ulrych (1998) and Jacobs and van der Geest (1991) compare the Bayesian and maximum relative entropy approaches. The basic difference is that, as expressed by the latter authors, the maximum relative entropy posterior pdf is obtained from the minimization of the entropy functional constrained not by the full equation $\mathbf{d} = \mathbf{F}\mathbf{m}$ but by the first moments $E[\mathbf{d} = \mathbf{F}\mathbf{m}]$—a weakened form of the constraints. In fact, the maximum relative entropy solution coincides with the Bayesian solution when $H(q, p)$ in equation (31) is minimized over the M moments of the data.

A LAST WORD ABOUT PRIORS, INCLUDING RISK

We have seen the pivotal role played by the prior in the Bayesian inverse approach. It is therefore important to examine the effect of our choice on the answer. In particular, following Scales and Tenorio (2001), we examine the issue of the uninformative prior. To do this we must deal a little with the risk associated with an estimator. Our definitions follow Efron and Morris (1973), who inspired us to consider the issue of risk in association with the Stein estimator (Ulrych et al., 1999). First, we define loss and risk. For simplicity, we consider M samples, x_i, $i = 1, 2, \ldots, M$ of the random variable $\mathbf{x}$ distributed as $\mathbf{x}|\mu \overset{\text{ind}}{\sim} N(\mu, 1)$, and the associated maximum likelihood estimator of μ:

$$\boldsymbol{\delta}^0(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} x_i.$$

The loss L may be defined in the usual way as

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{k} (\hat{\mu}_i - \mu_i)^2,$$

where $\hat{\boldsymbol{\mu}}$ is the estimate of $\boldsymbol{\mu}$. The risk $\mathbb{R}(\cdot\,,\cdot)$ in turn is formally defined for the maximum likelihood estimator as

$$\mathbb{R}(\boldsymbol{\mu}, \boldsymbol{\delta}^0) = E_\mu \sum_{i=1}^{k} \left(\boldsymbol{\delta}^0(x_i) - \mu_i\right)^2,$$

where E_μ is the expectation over the distribution.

The value $\boldsymbol{\delta}^0$ is a very special estimator indeed. Statisticians have shown that, given $\mathbf{x}|\mu \overset{\text{ind}}{\sim} N(\mu, 1)$, $\boldsymbol{\delta}^0$ has lowest risk of any linear or nonlinear unbiased estimator. We have previously seen that $\boldsymbol{\delta}^0$ is the maximum likelihood estimator as well as the Bayesian estimator obtained with a uniform prior.

We have already met the risk in the guise of the MSE. Specifically, we dealt with the MSE associated with the Bayesian estimation of variance using the Jeffreys uninformative prior. Using the above notation, $\text{MSE}(\hat{\sigma}^2) = \mathbb{R}(\boldsymbol{\sigma}^2, \hat{\boldsymbol{\sigma}}^2)$. Bayes with the Jeffreys prior leads to an estimator that, at least for the case considered here, constitutes a trade-off between bias and variance.

Scales and Tenorio (2001) consider the following problem. Given a datum d from $N(\mu, 1)$ and the prior information that μ is bounded $(-b, b)$, what are the risks associated with a Bayesian and a frequentist approach to the estimation of μ where the Bayesian prior is taken to be uniform (maximally uninformative?) in $(-b, b)$. The answer is that the risk of Bayesian estimator is, in a certain range, lower than the lower bound risk of the minimax estimator that incorporates the constraint as a hard bound. Since the minimax risk is not based on any distribution, we can conclude that the uniform pdf has introduced information beyond the bounds. Putting it another way, whereas the frequentist approach (at least in this case) is truly noncommittal with respect to what we do not know, the Bayesian method of introducing a uniform prior is not—a very interesting observation.

BAYESIAN OBJECTIVE FUNCTIONS

The previous section drew attention to the information in the prior. In this section, the prior model is used in a very different sense. We will choose a set constrained to be of a certain quality by the prior distribution. In this approach, we do not pay heed to the truth of the prior. We choose it with an aim in mind—what John Scales would perhaps call a post prior. We illustrate the approach with the work of Sacchi et al. (1998), who consider the objective of obtaining an aperture-free Fourier transform.

A common approach to signal analysis and decomposition is based on mapping the data into a new domain where the support of each signal is considerably reduced; consequently, decomposition can be attained easily. This is applicable to the discrete Fourier transform. Since we are always concerned with a finite amount of data, the correct decomposition of events is complicated by sidelobe artifacts.

Often, our problem may be posed as a linear inverse where the correct regularization is crucial to the resolution of signals when the aperture is below the resolution limit. The regularization is obtained by incorporating into the problem a long-tailed prior using Bayes' rule.

For simplicity we start with the 1-D discrete Fourier transform since extensions to higher dimensions are straightforward. Consider an N-sample time or spatial series $x_0, x_1, x_2, \ldots, x_{N-1}$. The discrete Fourier transform of the discrete series is given by

$$X_k = \sum_{n=0}^{N-1} x_n\, e^{-i2\pi nk/N} \quad k = 0, \ldots, N-1.$$

Similarly, the inverse discrete Fourier transform is given by

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k\, e^{i2\pi nk/N} \quad n = 0, \ldots, N-1. \tag{32}$$

We wish to estimate M spectral samples where $M > N$. A standard approach to solving this problem is by means of zero padding. Defining a new time series consisting of the original series plus a zero extension for $n = N, \ldots, M-1$, we can estimate M spectral samples using the discrete Fourier transform. This procedure helps to remove ambiguities resulting from discretization of the Fourier transform. However, it does not reduce the sidelobes created by the temporal/spatial window or improve the resolution. Let us therefore consider the estimation of M spectral samples without zero padding. In other words, we want to estimate the discrete Fourier transform using only the available information. The underlying philosophy is similar to Burg's maximum entropy method (Burg, 1975), except that in the maximum entropy method the target is a power spectral estimate—a phaseless function.

To avoid biasing our results by the discretization, we also impose the condition $M \gg N$. Rewriting equation (32) as

$$x_n = \frac{1}{M} \sum_{k=0}^{M-1} X_k\, e^{i2\pi nk/M} \quad n = 0, \ldots, N-1$$

gives rise to a linear system of equations

$$\mathbf{y} = \mathbf{F}\mathbf{x},$$

where the vectors $\mathbf{y} \in \mathbf{R^N}$ and $\mathbf{x} \in \mathbf{C^M}$ denote the available information and the unknown discrete Fourier transform, respectively. We now have an underdetermined problem that must be suitably regularized to impose uniqueness. We examine two regularization strategies for clarity.

Zero-order quadratic regularization

In standard fashion, we assume data contaminated with noise which is distributed as $N(0, \sigma_n^2)$. Sacchi et al. (1998) first assumed that the samples of the discrete Fourier transform may be modeled with a Gaussian prior. After combining the likelihood with the prior probability of the model by means of Bayes' rule, the MAP solution is computed by minimizing

$$J_{\text{GG}}(\mathbf{x}) = \lambda\, \|\mathbf{x}\|_2^2 + \|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2, \tag{33}$$

where GG stands for the Gauss-Gauss model (Gaussian model and Gaussian errors). The scalar is the ratio of variances, $\lambda = \sigma_n^2/\sigma_x^2$. Equation (33) is the objective function of the problem. The first term represents the model norm, while the second term is the misfit function. The hyperparameter λ enables us to move our estimate along the trade-off curve. Taking derivatives and equating to zero yields

$$\hat{\mathbf{x}} = \left(\mathbf{F}^H\mathbf{F} + \lambda\mathbf{I}_M\right)^{-1}\mathbf{F}^H\mathbf{y}.$$

Following some algebra, this equation may be rewritten as

$$\hat{\mathbf{x}} = \left(\frac{1}{M} + \lambda\right)^{-1} \mathbf{F}^H \mathbf{y}. \tag{34}$$

The result is the discrete Fourier transform of x_n modified by a scale factor. The solution expressed by equation (34) becomes

$$\hat{X}_k = \frac{1}{1+\lambda M} \sum_{n=0}^{N-1} x_n \, e^{-i2\pi nk/(M-1)}. \tag{35}$$

Equation (35) represents the discrete Fourier transform of the windowed time series and is equivalent to padding with zeroes in the range $n = N, \ldots, M-1$. It is clear that the zero-order regularization yields a scaled version of the conventional discrete Fourier transform, and the associated periodogram will show the classical sidelobes artifacts due to truncation. Of course, this result is expected from our previous discussion of likelihood and the method of least squares. We do not wish to minimize the problem associated with choosing the appropriate value of λ. Estimating hyperparameters in a sensible and robust manner remains a central problem in inversion. We mention a nonoptimal approach below.

Regularization by the Cauchy-Gauss model

To obviate the sidelobe structure, Sacchi et al. (1998) propose a regularization derived from a pdf that mimics a sparse distribution of spectral amplitudes. A heavy-tailed distribution, like the Cauchy pdf, will induce a sparse model consisting of only a few elements different from zero. If the data consist of a few number of harmonics (1-D case) or a limited number of plane waves (2-D case), a sparse solution will help to attenuate sidelobe artifacts. The Cauchy pdf is given by

$$p(X_k \,|\, \sigma_c) \propto \frac{1}{\left(1 + \dfrac{X_k X_k^*}{2\sigma_c^2}\right)},$$

where σ_c is a scale parameter. When we combine the Cauchy prior with the data likelihood, the cost function becomes

$$J_{\mathrm{CG}}(\mathbf{x}) = S(\mathbf{x}) + \frac{1}{2\sigma_n^2}(\mathbf{y} - \mathbf{F}\mathbf{x})^H(\mathbf{y} - \mathbf{F}\mathbf{x}),$$

where the subscript CG stands for the Cauchy-Gauss model. The function $S(\mathbf{x})$, which is expressed by

$$S(\mathbf{x}) = \sum_{k=0}^{M-1} \log\left(1 + \frac{X_k X_k^*}{2\sigma_c^2}\right),$$

is the regularizer imposed by the Cauchy distribution and is a measure of the sparseness of the vector of spectral powers $P_k = X_k X_k^*$, $k = 0, \ldots, M-1$. The constant σ_c controls the amount of sparseness that can be attained by the inversion.

Taking derivatives of $J_{\mathrm{CG}}(\mathbf{x})$ and equating to zero yields

$$\hat{\mathbf{x}} = (\lambda \mathbf{Q}^{-1} + \mathbf{F}^H \mathbf{F})^{-1} \mathbf{F}^H \mathbf{y},$$

where $\lambda = \sigma_n^2/\sigma_c^2$ and $\mathbf{Q}$ is an $M \times M$ diagonal matrix with elements given by

$$Q_{ii} = 1 + \frac{X_i X_i^*}{2\sigma_c^2}, \quad i = 0, \ldots, M-1.$$

Sacchi et al. (1998) show that the solution may also be written as

$$\hat{\mathbf{x}} = \mathbf{Q}\mathbf{F}^H \left(\lambda \mathbf{I}_N + \mathbf{F}\mathbf{Q}\mathbf{F}^H\right)^{-1} \mathbf{y}. \tag{36}$$

The hyperparameters of the problem are fitted by Sacchi et al. (1998) by means of a χ^2 criterion. This approach is not optimum, as we have pointed out, but it certainly requires a much less intensive computational effort. Equation (36) is iteratively solved with the initial model being the discrete Fourier transform of the truncated signal. An excellent view of objective functions without appeal to Bayes is the work of Farquharson and Oldenburg (1998).

An example of this approach is illustrated in Figures 1–4. In the first example we attempt to determine a highly resolved power spectral estimate from irregularly sampled and noisy data. Figure 1a shows the continuous data without noise (solid line) and the noisy samples (squares) where the added noise was Gaussian with a standard deviation of 0.4. The sampled data clearly demonstrate a gapped appearance.

We used equation (32) to obtain the results illustrated in Figures 1b, 1c, and 2a. The former shows the reconstituted data (solid line) using the noisy, irregularly sampled input repeated as squares. Figure 1c shows the error panel. The computed high-resolution spectrum is shown in Figure 2a and is to be compared with Figure 2b, which shows the periodogram of the complete time series (the solid curve in Figure 1a).

Figures 3 and 4 illustrate the results of our Bayesian approach in a 2-D example. Specifically, we are interested in computing the 2-D spectrum of the vertical seismic profile shown in Figure 3.

Figure 4a illustrates the 2-D periodogram; Figure 4b demonstrates the resulting power spectrum of the discrete Fourier transform computed via the Cauchy-Gauss method. The enhancement in the resolution is very clear.

SUMMARY AND DISCUSSION

We have attempted to cogently summarize some of the concepts central to the Bayesian approach to inverse problems. Of course, the centerpiece is the prior distribution in function as well as in controversy. We follow the conclusion of Scales and Tenorio (2001), who pose the question "To Bayes or not to Bayes?" and answer it with a careful yes. We would perhaps be more affirmative and say yes by way of an honest appraisal of what we really know.

The Bayesian approach may be applied hierarchically (pure Bayes) or empirically (less pure Bayes). In the former, we are concerned about the information content in the prior and its effect upon our estimate. Thus, issues such as how informative an uninformative prior is, the risk of a Bayesian estimate, etc., are uppermost. In the empirical approach, these questions are less imperative.

Hierarchical issues

Let us return briefly to the Jeffreys prior. Certainly, $1/\sigma$ can hardly be construed as a pdf. It is, in fact, known as an improper prior since it is not normalizable. But how important is this fact? In practice, σ is always bounded; it is certainly greater than zero and less than infinity and can therefore be normalized. In any case, since the prior gets multiplied by the likelihood

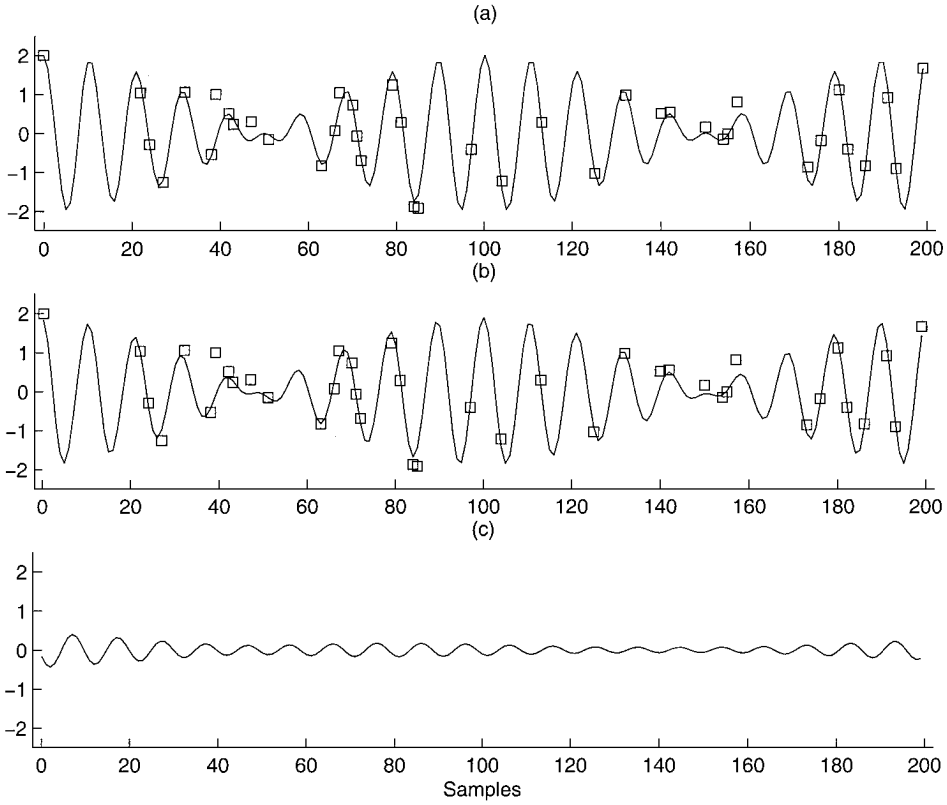

FIG. 1. (a) Input data. (b) Reconstructed data. (c) Error.

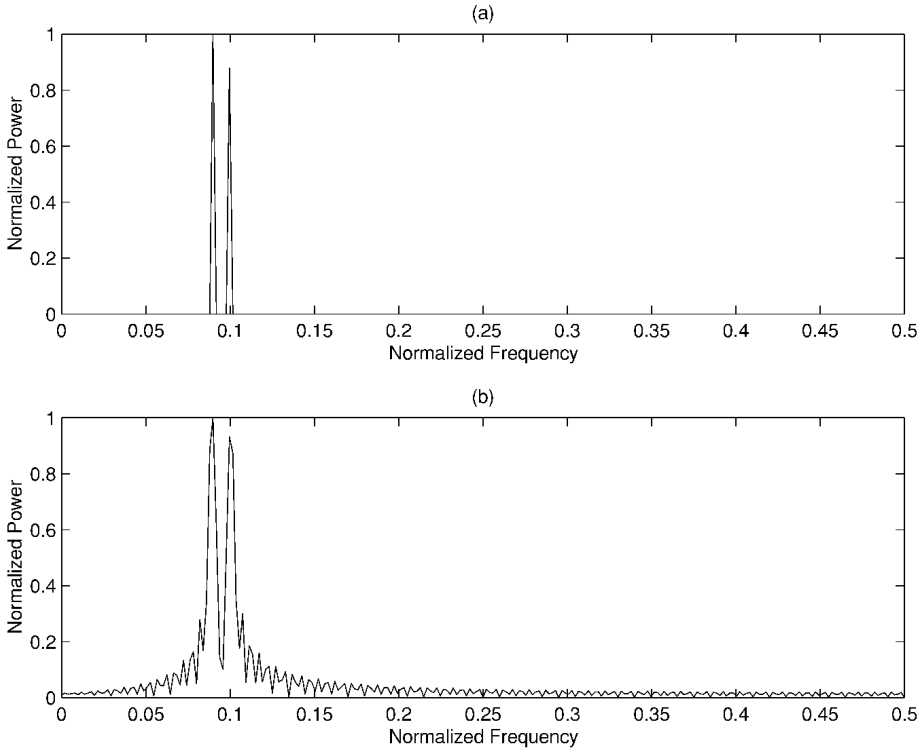

FIG. 2. (a) High-resolution spectrum. (b) Periodogram.

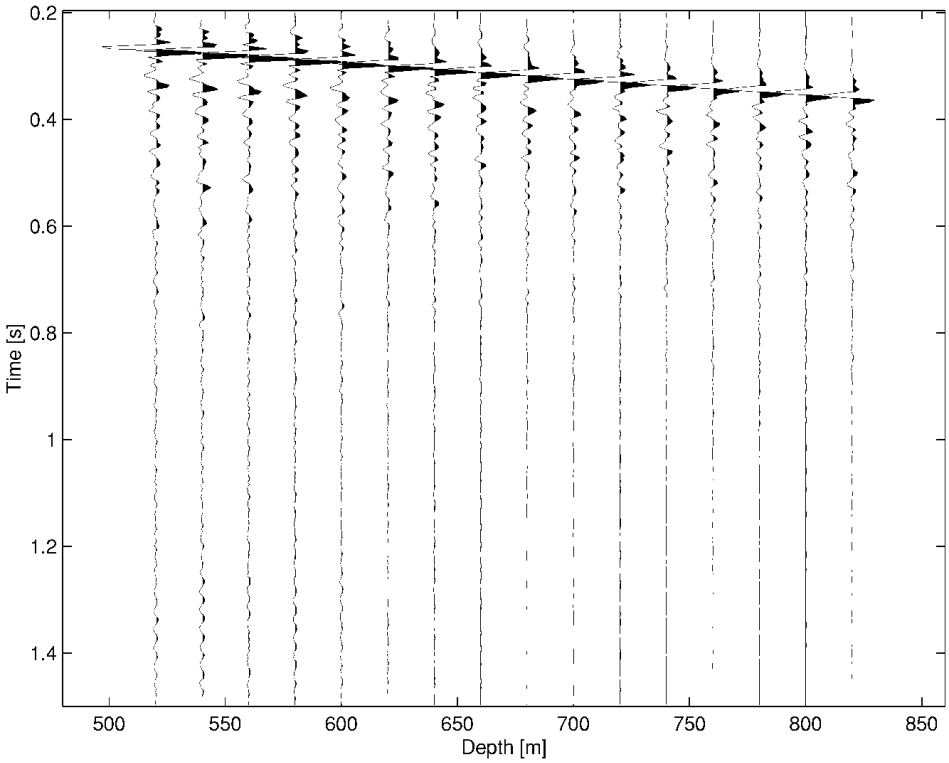

FIG. 3. Vertical seismic profile.

which certainly goes to zero as σ approaches zero or infinity, the bounds are not really significant.

Scales and Tenorio (2001) discuss two interesting issues. The first deals with the information contained in a seemingly uninformative prior which indicates that, from the point of view of the risk associated with the minimax and Bayes estimators for a particular problem, the Bayesian estimator is less noncommittal than the frequentist estimator. This is worth pondering. It is interesting that a uniform prior adds more information than a hard bound, particularly since incorporating the constraint via a distribution is referred to as softening the constraint. From a practical view, given all other uncertainties involved in a real data inversion problem, the differences incurred in the solution as a result of information possibly added because of the above effect are, in all probability, insignificant.

The second issue discussed by Scales and Tenorio (2001) is that of added difficulties when hard constraints are replaced by probability constraints that occur in higher dimensions. This topic is certainly worthy of further investigation.

Empirical issues

The idea behind the empirical Bayes approach is that the prior is based on information contained in the input data. The methodology for constructing objective functions from prior pdf's and Bayes' rule is not exactly empirical, but it certainly is not hierarchical. In this case, the truth of the prior is irrelevant. We use it to constrain the solution to a form that we know to be desired. We have illustrated this by computing a high-resolution discrete Fourier transform (Sacchi et al., 1998). Another application with which we had considerable success is the compensation of aperture effects in computing the Radon transform (Sacchi and Ulrych, 1995). We look for the transform that has a limited support in the Radon domain. The Bayesian approach here is, in some sense, empirical, in that the data tell us how sparse the model should be.

In our opinion, Bayes is to likelihood what likelihood is to least squares (or singular value decomposition, in this case). No one would argue that likelihood is not that much more flexible an approach than least squares. So we argue that Bayes, when used diligently, is much more flexible than plain old likelihood.

REFERENCES

Burg, J. P., 1975, Maximum entropy spectral analysis: Ph.D. thesis, Stanford Univ.

Bretthorst, G. L., 1988, Bayesian spectrum analysis and parameter estimation: Springer-Verlag New York Inc.

Duijndam, A. J. W., 1988a, Bayesian estimation in seismic inversion—Part I: Principles: Geophys. Prosp., **36**, 878–898.

———1988b, Bayesian estimation in seismic inversion—Part II: Uncertainty analysis: Geophys. Prosp., **36**, 899–918.

Efron, B., and Morris, C., 1973, Stein's estimation rule and its competitors—An empirical Bayes approach: J. Am. Stat. Assoc., **68**, 117–130.

Farquharson, C. G., and Oldenburg, D. W., 1998, Non-linear inversion using general measures of data misfit and model structure: Geophys. J. Internat., **134**, 213–227.

Frieden, B. R., 1987, Probability, statistical optics, and data testing: Springer-Verlag.

Jacobs, F. J., and van der Geest, P. A. G., 1991, Spiking band-limited traces with a relative-entropy algorithm: Geophysics, **56**, 1003–1014.

Jaynes, E. T., 1982, On the rationale of maximum entropy methods: Proc. IEEE, **70**, 939–952.

———1996, Probability theory—The logic of science: available at http://bayes.wustl.edu.

Jeffreys, H., 1939, Theory of probability: Oxford Univ. Press, Inc.

Lupton, R., 1993, Statistics in theory and practice: Princeton Univ. Press.

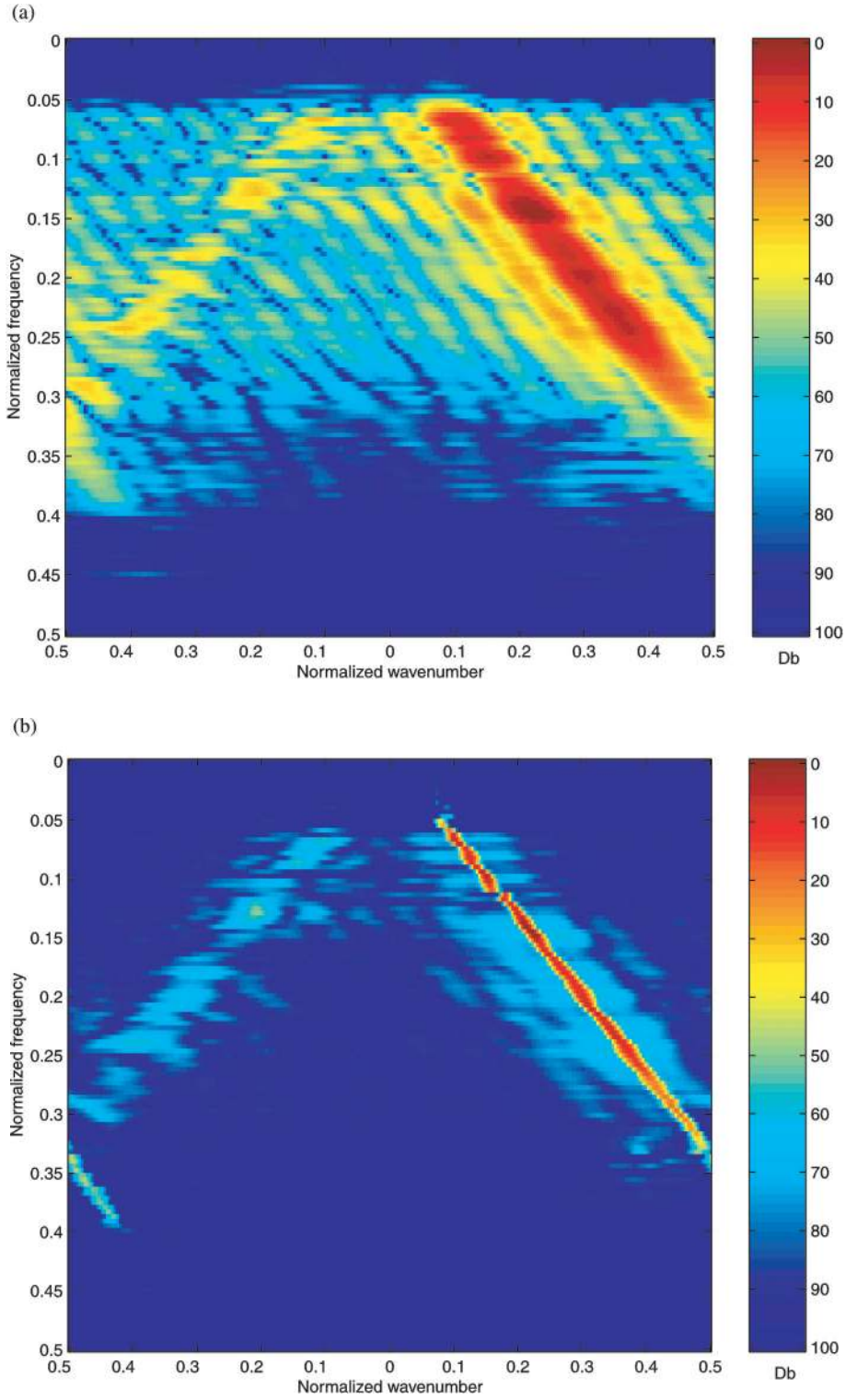

FIG. 4. (a) 2-D periodogram. (b) 2-D high-resolution spectrum.

Press, S. J., 1989, Bayesian statistics: Principles, models and applications: John Wiley & Sons, Inc.

Matsuoka, T., and Ulrych, T. J., 1986, Information theory measures with application to model identification: IEEE Trans. Acoust., Speech, Signal Processing, **ASSP-34**, 511–517.

Sacchi, M. D., and Ulrych, T. J., 1995, High-resolution velocity gathers and offset space reconstruction: Geophysics, **60**, 1169–1177.

Sacchi, M. D., Ulrych, T. J, and Walker, C., 1998, Interpolation and extrapolation using a high resolution discrete Fourier transform: IEEE Trans. Signal Proc., **46**, 31–38.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G., 1986, Akaike information criterion statistics: D. Reidel Publ. Co.
Scales, J. A., and Tenorio, L., 2001, Prior information and uncertainty in inverse problems: Geophysics, **66**, tentatively scheduled for March/April.
Shore, J. E., and Johnson, R. W., 1980, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy: IEEE Trans. Info. Theory, **IT-26**, 26–37.
Sivia, D. S., 1996, Data analysis: A Bayesian tutorial: Clarendon Press.
Tarits, P., Jouanne, V., Menvielle, M., and Roussignol, M., 1994, Bayesian statistics of non-linear inverse problems: Example of the magnetotelluric 1-D inverse problem: Geophys. J. Internat., **119**, 353–368.
Ulrych, T. J., Bassrei, A., and Lane, M., 1990, Minimum relative entropy inversion of 1D data with applications: Geophys. Prosp., **38**, 465–487.
Ulrych, T. J., Sacchi, M. D., and Graul, M., 1999, Signal and noise separation: Art and science: Geophysics, **64**, 1648–1656.
Woodbury, A. D., and Ulrych, T. J., 1998, Minimum relative entropy and probabilistic inversion in ground water hydrology: Stochastic Hydrology and Hydraulics, **12**, 317–358.