

Tutorial

Prior information and uncertainty in inverse problems

John A. Scales* and Luis Tenorio†

ABSTRACT

Solving any inverse problem requires understanding the uncertainties in the data to know what it means to fit the data. We also need methods to incorporate data-independent prior information to eliminate unreasonable models that fit the data. Both of these issues involve subtle choices that may significantly influence the results of inverse calculations. The specification of prior information is especially controversial. How does one quantify information? What does it mean to know something about a parameter a priori? In this tutorial we discuss Bayesian and frequentist methodologies that can be used to incorporate information into inverse calculations. In particular we show that apparently conservative Bayesian choices, such as representing interval constraints by uniform probabilities (as is commonly done when using genetic algorithms, for example) may lead to artificially small uncertainties. We also describe tools from statistical decision theory that can be used to characterize the performance of inversion algorithms.

OVERVIEW

Solving an inverse problem means making inferences about physical systems from data. These inferences are based on mathematical representations of the systems, which we call models. Functionals of the models represent observable properties of the system such as the mass density as a function of space in the earth, the depth of continents, or the radius of the core–mantle boundary.

In formulating inverse problems and interpreting inversion estimates, we need to address the following questions.

- 1) How accurately are the data known? What does it mean to fit the data?

- 2) How accurately is the physical system modeled? Does the model include all the physical effects that contribute significantly to the data?
- 3) What is known about the model before the data are observed? What does it mean for a model to be reasonable?

We address these questions by presenting methods that can be used to study the performance of inversion estimates and methods to include prior information in the inversion process. We start by setting the general framework for the inverse problems considered. We then present two different approaches by which prior information can be included in geophysical inverse calculations: Bayesian and frequentist. These two approaches differ fundamentally in the means by which probability is introduced into the calculation. They also take fundamentally different approaches to the treatment of observed data and prior information. Bayesians define probabilities on the space of models (prior information is thus probabilistic), conditioned on the observed data. Frequentists assume a distribution prior to observing the data, which does not change once the data have been observed, and use deterministic prior information. Usually probability only enters the calculations via the data errors, which are assumed to have a random component.

The choice of prior probability model in Bayesian inference is not always clear even when the prior information is well defined. Our example shows how representing deterministic constraints probabilistically may inject information into the calculation that is not strictly required by the constraint. This problem is worse in high-dimensional spaces. We provide examples of inverse problems to illustrate some of the points raised in the tutorial.

Some notation

We use the symbol \mathbf{d} for data. Typically, this is an element of R^n , where n is the number of observations. The symbol \mathbf{m} is a model, typically an element of a linear vector space, usually

Published on Geophysics Online October 17, 2000. Manuscript received by the Editor October 28, 1999; revised manuscript received July 28, 2000.
 *Colorado School of Mines, Department of Geophysics and Center for Wave Phenomena, Golden, Colorado 80401. E-mail: jscales@mines.edu.
 †Colorado School of Mines, Department of Mathematical and Computer Sciences, Golden, Colorado 80401. E-mail: ltenorio@mines.edu.
 © 2001 Society of Exploration Geophysicists. All rights reserved.

infinite dimensional, such as the set of square integrable functions on R^3 . In other words, \mathbf{m} is a representation of the unknown physical process. $E(\cdot)$ and $\text{var}(\cdot)$ stand for expectation and variance operators, respectively.

An estimator of an unknown model \mathbf{m} (or a functional thereof) is a function $\hat{\mathbf{m}}$ of the data used to estimate the model. The estimate given the data \mathbf{d} is denoted as $\hat{\mathbf{m}}(\mathbf{d})$. When the dependence on \mathbf{d} is understood, it is denoted as $\hat{\mathbf{m}}$.

Models are usually parameterized so that estimating a model is equivalent to estimating its corresponding parameters \mathbf{m} . But clearly the choice of observables and parameterization is not unique. For instance, in problems of elasticity we may use the elastic stiffness tensor or the elastic compliance tensor. We can use wave speed or wave slowness.

A GENERAL STATEMENT OF THE INVERSE PROBLEM

As the result of some experiment, n data are collected. The data are related to the physical models through the forward modeling operator. This operator is a function g that maps models into data space. In practice the forward operator is always an approximation. In geophysics this is primarily because one cannot afford to model the true complexity of the earth. Even if this were possible, it might not be worth the effort given the instrument's resolution and the noise level in the data. This will be discussed later, but for now it suffices to be aware that there is a systematic error associated with using g . Let us represent it by an n -dimensional vector \mathbf{s} . Finally, there is an n -dimensional vector of random measurement errors, \mathbf{e} . Assuming additive errors, the connection between models and data can be written as

$$\mathbf{d} = g(\mathbf{m}) + \mathbf{e} + \mathbf{s}.$$

The goal is to estimate \mathbf{m} [or a functional $L(\mathbf{m})$], given a vector \mathbf{d} of measurements.

For example, suppose the forward operator, \mathbf{K} , is linear and the model is represented by an infinite sequence of coefficients, $\mathbf{m} = \{m_i\}$, with respect to some orthonormal basis. We model the data as

$$\mathbf{d} = \mathbf{K}\mathbf{m} + \mathbf{e}.$$

Since we have a finite amount of data, we can hope to estimate only finitely many m_i . Consider the vector containing the first ℓ coefficients, \mathbf{m}_ℓ , and the sequence \mathbf{m}_∞ containing the rest. We can write our model as

$$\mathbf{d} = \mathbf{K}_\ell \mathbf{m}_\ell + \mathbf{K}_\infty \mathbf{m}_\infty + \mathbf{e}$$

[e.g., Trampert and Snieder (1996)]. In this case we can consider the leftover $\mathbf{K}_\infty \mathbf{m}_\infty$ as a systematic error, \mathbf{s} . But since \mathbf{K} maps an infinite-dimensional space into a finite-dimensional one, the forward operator has a nontrivial kernel. So even in the absence of measurement and modeling errors, the forward operator will not be invertible and the set of models that predict the data equally well may be quite large. This in itself may not be a problem; the problem is when these equally predicting models yield wildly different values for the model functional we want to estimate. By including prior information, we attempt to constrain the range of feasible models and thus control the effects of those nullspace elements. We illustrate this later with an example. Also, even when there is a unique solu-

tion, it may be unstable to small perturbations in the data. In this case we may use some prior information to stabilize the solution.

Example: Estimating the derivative of a smooth function

We start with a simple example to illustrate the effects of noise and prior information in the performance of an estimator. Later we will introduce tools from statistical decision theory to study the performance of estimators given different types of prior information.

Suppose we have noisy observations of a smooth function f at the equidistant points $a \leq x_1 \leq \dots \leq x_n \leq b$:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where the errors ϵ_i are assumed to be *iid* $N(0, \sigma^2)$ —independent, identically distributed random variables normally distributed with mean 0 and variance σ^2 . We want to use these observations to estimate the derivative f' . We define the estimator

$$\hat{f}'(x_{m_i}) = \frac{y_{i+1} - y_i}{h}, \quad (2)$$

where h is the distance between consecutive points and $x_{m_i} = (x_{i+1} + x_i)/2$. To measure the performance of the estimator (2) we use the mean square error, which is the sum of the variance and squared bias. The variance and bias of equation (2) are

$$\text{var}[\hat{f}'(x_{m_i})] = \frac{\text{var}(y_{i+1}) + \text{var}(y_i)}{h^2} = \frac{2\sigma^2}{h^2},$$

$$\begin{aligned} \text{bias}[\hat{f}'(x_{m_i})] &\equiv E[\hat{f}'(x_{m_i}) - f'(x_{m_i})] \\ &= \frac{f(x_{i+1}) - f(x_i)}{h} - f'(x_{m_i}) \\ &= f'(\alpha_i) - f'(x_{m_i}) \end{aligned}$$

for some $\alpha_i \in [x_i, x_{i+1}]$ (by the mean value theorem). We need some information on f' to assess the size of the bias. Let us assume the second derivative is bounded on $[a, b]$ by M :

$$|f''(x)| \leq M, \quad x \in [a, b].$$

It then follows that

$$\begin{aligned} |\text{bias}[\hat{f}'(x_{m_i})]| &= |f'(\alpha_i) - f'(x_{m_i})| \\ &= |f''(\beta_i)(\alpha_i - \beta_i)| \leq Mh \end{aligned}$$

for some β_i between α_i and x_{m_i} . As $h \rightarrow 0$ the variance goes to infinity while the bias goes to zero. The mean square error (MSE) is bounded by

$$\begin{aligned} \frac{2\sigma^2}{h^2} \leq \text{mse}[\hat{f}'(x_{m_i})] &= \text{var}[\hat{f}'(x_{m_i})] + \text{bias}[\hat{f}'(x_{m_i})]^2 \\ &\leq \frac{2\sigma^2}{h^2} + M^2 h^2. \end{aligned} \quad (3)$$

It is clear that choosing the smallest h possible does not lead to the best estimate; the noise must be taken into account. The lowest upper bound is obtained with $h = 2^{1/4} \sqrt{\sigma/M}$. The larger the variance of the noise, the wider the spacing between the points.

We have used a bound on the second derivative to bound the mse. It is a fact that some type of prior information, in addition to model (1), is required to bound derivative uncertainties. Take any smooth function g which vanishes at the points x_1, \dots, x_n . Then, the function $\tilde{f} = f + g$ satisfies the same model as f , yet their derivatives could be very different. For example, choose an integer m and define

$$g(x) = \sin \left[\frac{2\pi m(x - x_1)}{h} \right].$$

Then, $f(x_i) + g(x_i) = f(x_i)$ and

$$\tilde{f}'(x) = f'(x) + \frac{2\pi m}{h} \cos \left[\frac{2\pi m(x - x_1)}{h} \right].$$

By choosing m large enough, we can make the difference, $\tilde{f}'(x_{m_i}) - f'(x_{m_i})$, as large as we want; without prior information the derivative cannot be estimated with finite uncertainty.

BAYESIAN AND FREQUENTIST METHODS OF INFERENCE

There are two fundamentally different meanings of the term probability in common usage (Scales and Snieder, 1997). If we toss a coin N times, where N is large, and see roughly $N/2$ heads, then we say the probability of getting a head in a given toss is about 50%. This interpretation of probability, based on the frequency of outcomes of random trials, is therefore called frequentist. On the other hand it is common to hear statements such as “The probability of rain tomorrow is 50%.” Since this statement does not refer to the repeated outcome of a random trial, it is not a frequentist use of the term probability. Rather, it conveys a statement of information (or lack thereof). This is the Bayesian use of probability. Both ideas seem natural to some degree, so it is perhaps unfortunate that the same term is used to describe them.

Bayesian inversion has gained considerable popularity in its application to geophysical inverse problems. The philosophy of this procedure is as follows. Suppose one knows something about a model before observing the data. This knowledge is cast in a probabilistic form and is called the prior probability model (prior means before the data have been observed). Bayesian inversion then provides a framework for combining the probabilistic prior information with the information contained in the observed data to update the prior information. The updated distribution is the posterior conditional model distribution given the data; it is what we know about the model after we have assimilated the data and the prior information. The point of using the data is that the posterior information should constrain the model more tightly than the prior model distribution.

However, the selection of a prior statistical model can in practice be somewhat shaky. For example, in a seismic survey we may have a fairly accurate idea of the realistic ranges of seismic velocity and density, and perhaps even of the vertical correlation length (if borehole measurements are available). However, the horizontal length scale of the velocity and density variation is to a large extent unknown. Given this, how can Bayesian inversion be so popular when our prior knowledge is often so poor? The reason is that, in practice, the prior model is used to regularize the posterior solution. Via a succession of different calculations, the characteristics of the prior model

are often tuned in such a way that the retrieved model has subjectively agreeable features. But logically, the prior distribution must be fixed beforehand. The features used to tune the prior should in fact be included as part of the prior information (Gouveia and Scales, 1997). So the practice of using the data to tune the prior suggests that the reason for the popularity of Bayesian inversion within the earth sciences is inconsistent with the underlying philosophy. A common attitude seems to be, “If I hadn’t believed it, I wouldn’t have seen it.”

Since Bayesian statistics relies completely on the specification of a prior statistical model, the flexibility taken in using the prior model as a knob to tune properties of the retrieved model is completely at odds with the philosophy of Bayesian inversion. One can, however, use an empirical Bayes approach to use data to help determine a prior distribution. But having used the data to select a prior, one must correct the uncertainty estimates so as not to be overconfident [see Carlin and Louis (1996)]. This correction is not usually done in geophysical Bayesian inversion.

Bayesian inversion in practice

Two important questions must be addressed in any Bayesian inversion. (1) How do we represent the prior information, both the prior model information and the description of the data statistics? (2) How do we summarize the posterior information? The second question is the easiest to answer, at least in principle. It is just a matter of applying Bayes’ theorem to compute the posterior distribution. We then use this distribution to study the statistics of different parameter estimates. For example, we can find credible regions for the model parameters given the data, or we can use posterior means as estimates and posterior standard deviations as error bars. However, very seldom are we able to compute all posterior estimates analytically; we often have to use computer-intensive approximations based on Markov Chain Monte Carlo methods [see, for example, Tanner (1993)]. Nevertheless, a complete Bayesian analysis may be computationally intractable.

The first question is a lot more difficult to answer. A first strategy is a subjective, Bayesian one: prior probabilities are designed to represent states of mind, prejudices, or prior experience. But depending on the amount and type of prior information, the choice of prior may or may not be clear. For example, if an unknown parameter μ must lie between a and b , are we justified to assume that μ has a uniform prior distribution on the interval $[a, b]$? This question is addressed later, but for now observe there are infinitely many probability distributions consistent with μ being in the interval $[a, b]$. To pick one may be an overspecification of the available information. Even an apparently conservative approach, such as taking the probability distribution that maximizes the entropy subject to the constraint that μ lies in the interval, may lead to pathologies in high-dimensional problems. This shows how difficult it may be to unambiguously select a prior statistical model. One way out of this dilemma is to sacrifice objectivity and presume that probability lies in the eye of the beholder. Of course, this means that our posterior probability will be different from yours.

A second approach attempts to make a somewhat more objective choice of priors by relying on theoretical considerations such as maximum entropy (Jaynes, 1982) or transformation invariance (Jaynes, 1968; Dawid, 1983), or by somehow

using observations to estimate a prior. This latter approach is the empirical Bayes mentioned earlier. For example, suppose one is doing a gravity inversion to estimate mass density in some reservoir. Suppose further a large number of independent, identically distributed laboratory measurements of density for rocks taken from this reservoir are available (a big if). One could use the measurements to estimate a probability distribution for mass density that could be used as a prior for the gravity inversion. This is the approach taken in Gouveia and Scales (1998), where in-situ (borehole) measurements are used to derive an empirical prior for surface seismic data.

An empirical Bayes analysis is basically an approximation to a full hierarchical Bayes analysis based on the joint probability distribution of all parameters and available data. In other words, in a full Bayesian analysis the prior distribution may depend on some parameters that in turn follow a second-stage prior. This latter prior can also depend on some third-stage prior, etc. This hierarchical model ends when all the remaining parameters are assumed known. We can use the empirical Bayes approach when the last parameters cannot be assumed to be known. Instead, we use the data to estimate the remaining parameters and stop the sequence of priors. We then proceed as in the standard Bayesian procedure. For an introduction to empirical and hierarchical Bayes methods, see Casella (1985), Carlin and Louis (1996), Gelman et al. (1997), and references therein. For a review on the development of objective priors, see Kass and Wasserman (1996).

A third strategy is to abandon Bayes altogether and use only deterministic prior information about models: wave speed is positive (a matter of definition), velocity is less than the speed of light (a theoretical prediction), the earth's mass density is less than 6 g/cm^3 (a combination of observation and theory that is highly certain). The inference problem is still statistical since random data uncertainties are taken into account. Essentially, the idea is to look at the set of all models that fit the data and then perform surgery on this set, cutting away those models that violate the deterministic criteria, e.g., have negative density. The result is a (presumably smaller and more realistic) set of models that fit the data and satisfy the prior considerations. We choose any model that fits the data to a desired level and satisfies the prior model constraints. Tikhonov's regularization (Philips, 1962; Tikhonov and Arsenin, 1977) is one way of obtaining an inversion algorithm by restricting the family of models that fit the data. For example, among all the models that fit the data, we choose one that has particular features—the smoothest, the shortest, etc. [e.g., Scales et al. (1990), Gouveia and Scales (1997)].

Bayes versus frequentist

In the Bayesian paradigm, probability distributions are the fundamental tools. Bayesians speak of the probability of a hypothesis given some evidence and conduct pre-data and post-data inferences. Frequentists, on the other hand, are more concerned with pre-data inference and run into difficulties when trying to give post-data interpretations to their pre-data formulation. In other words, uncertainty estimates such as confidence sets are based on the error distribution, which is assumed to be known a priori, and on hypothetical repetitions of the data gathering process. However, see Goutis and Casella (1995) for a discussion of frequentist post-data inference.

The choice of prior distributions is not always well defined. In this case it would seem more reasonable to follow a frequentist approach. But it may also be the case that the determinism that frequentists rely on in defining parameters may be ill defined. For instance, if we are trying to estimate the mass of the earth, is this a precisely defined, nonrandom quantity? Perhaps, but does the definition include the atmosphere? If so, how much of the atmosphere? If not, does it take into account that the mass is constantly changing (slightly) from, for example, micrometeorites? Even if you make the true mass of the earth well defined (it will still be arbitrary to some extent), it can never be precisely known any more than the temperature of an isolated gas.

So, which approach is better? Bayesians are happy to point to some well-known inconsistencies in the frequentist methodology and to difficulties frequentists face to use available prior information. Some Bayesians even go as far as claiming that anyone in her/his right frame of mind should be a Bayesian. Frequentists, on the other hand, complain about the sometimes subjective choice of priors and about the computational complexity of the Bayesian approach. In real down-to-earth data analysis we prefer to keep an open mind. Different methods may be better than others, depending on the problem. Both schools of inference have something to offer. For colorful discussions on the comparison of the two approaches, see Efron (1986) and Lindley (1975). Also see Rubin (1984) for ways in which frequentist methods can complement Bayesian inferences.

WHAT DIFFERENCE DOES THE PRIOR MAKE?

In a Bayesian calculation, whatever estimator we use depends on the prior and conditional distributions given the data. There is no clear, established procedure to check how much information a prior injects into the posterior estimates. [This is one of the open problems mentioned in Kass and Wasserman (1996).] In this example we compare the risks of the estimators.

To measure the performance of an estimator $\hat{\mathbf{m}}$ of \mathbf{m} , we define the loss function, $L(\mathbf{m}, \hat{\mathbf{m}})$, where L is a nonnegative function that is zero for the true model. That is, for any other model \mathbf{m}_1 , $L(\mathbf{m}, \mathbf{m}_1) \geq 0$ and $L(\mathbf{m}, \mathbf{m}) \equiv 0$. The loss is a measure of the cost of estimating the true model with $\hat{\mathbf{m}}$ when it is actually \mathbf{m} . For example, a common loss function is the square error: $L(\mathbf{m}, \hat{\mathbf{m}}) = (\mathbf{m} - \hat{\mathbf{m}})^2$. But there are other choices, such as ℓ_p -norm error: $L(\mathbf{m}, \hat{\mathbf{m}}) = |\mathbf{m} - \hat{\mathbf{m}}|^p$.

The loss, $L(\mathbf{m}, \hat{\mathbf{m}})$, is a random variable since $\hat{\mathbf{m}}$ depends on the data. We average over the data to obtain an average loss. This is called the risk of $\hat{\mathbf{m}}$ given the model \mathbf{m} :

$$R(\mathbf{m}, \hat{\mathbf{m}}) = E_P L(\mathbf{m}, \hat{\mathbf{m}}), \quad (4)$$

where P is the error probability distribution and E_P is the expectation with respect to this distribution. For square error loss the risk is the usual mean square error.

Bayes risk

The expected loss depends on the chosen model. Some estimators may have small risks for some models but not for others. To compare estimators, we need a global measure that takes all plausible models into account. A natural choice is to take

the expected value of the loss with respect to the posterior distribution, $p(\mathbf{m} | \mathbf{d})$, of the model given the data. This is called the posterior risk:

$$r_{\mathbf{m} | \mathbf{d}} = E_{\mathbf{m} | \mathbf{d}} L[\mathbf{m}, \hat{\mathbf{m}}(\mathbf{d})].$$

Alternatively we can take a weighted average of the risk [equation (4)] using the prior model distribution as weight function. This is the Bayes risk:

$$r_{\rho} = E_{\rho} R(\mathbf{m}, \hat{\mathbf{m}}),$$

where ρ is the prior model distribution. An estimator with the smallest Bayes risk is called a Bayes estimator. We use a frequentist approach to define the Bayes risk since we have not conditioned on the observed data. It makes sense, however, to expect good frequentist behavior if the Bayesian approach is used repeatedly with different data sets. In addition, and under very general conditions, minimizing the Bayes risk is equivalent to minimizing the posterior risk (Berger, 1985).

Let f denote the joint distribution of models and data. The distribution (marginal) of the data is obtained by integrating f over the models:

$$h(\mathbf{d}) = \int_{\mathcal{M}} f(\mathbf{m}, \mathbf{d}) d\mathbf{m},$$

where \mathcal{M} is the space of models. From Bayes' theorem, the conditional distribution of \mathbf{m} given \mathbf{d} is

$$p(\mathbf{m} | \mathbf{d}) = \frac{f(\mathbf{d} | \mathbf{m})\rho(\mathbf{m})}{h(\mathbf{d})},$$

where $\rho(\mathbf{m})$, the prior distribution, is the marginal of f with respect to \mathbf{m} . The conditional distribution, $p(\mathbf{m} | \mathbf{d})$, is the so-called Bayesian posterior distribution, which updates the prior information in view of the data.

One can define a number of reasonable estimators of \mathbf{m} based on $p(\mathbf{m} | \mathbf{d})$ —for example, the $\hat{\mathbf{m}}$ that maximizes $p(\mathbf{m} | \mathbf{d})$ (or that is close, in probability, to \mathbf{m}). Or one could compute the estimator that gives the smallest Bayes risk for a given prior and loss function. It can be shown [Lehmann (1983), p. 239] that, for square error loss function, the Bayes estimator is the posterior mean.

Here is a simple example of using a normal prior to estimate a normal mean. Assume n observations, $(d_1, d_2, \dots, d_n) = \mathbf{d}$, which are *iid* $N(\eta, \sigma^2)$. We want to estimate the mean, η , given that the prior, ρ , is $N(\mu, \beta^2)$. Up to a constant factor, the joint distribution of η and \mathbf{d} is [Lehmann (1983), p. 243]

$$f(\mathbf{d}, \eta) = \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (d_i - \eta)^2\right] \exp\left[-\frac{1}{2\beta^2} (\eta - \mu)^2\right].$$

The posterior mean is

$$\hat{\eta} = E(\eta | \mathbf{d}) = \frac{n\bar{\mathbf{d}}/\sigma^2 + \mu/\beta^2}{n/\sigma^2 + 1/\beta^2},$$

where $\bar{\mathbf{d}}$ is the arithmetic mean of the data. The posterior variance is

$$\text{var}(\eta | \mathbf{d}) = \frac{1}{n/\sigma^2 + 1/\beta^2}.$$

Note that the posterior variance is always reduced by the presence of a nonzero β . The posterior mean, which is the Bayes estimator for square error loss, can be written as

$$\hat{\eta}(\mathbf{d}) = \left[\frac{n/\sigma^2}{n/\sigma^2 + 1/\beta^2} \right] \bar{\mathbf{d}} + \left[\frac{1/\beta^2}{n/\sigma^2 + 1/\beta^2} \right] \mu.$$

We see that the Bayes estimator is a weighted average of the mean of the data and the mean of the Bayesian prior distribution; the latter is the Bayes estimator before any data have been observed. The Bayes risk is the integral, over the data, of the posterior variance of η . Since the posterior variance does not depend on \mathbf{d} , the Bayes risk is just the posterior variance. Note also that as $\beta \rightarrow 0$, or increasingly strong prior information, the estimate converges to the prior mean. As $\beta \rightarrow \infty$, or increasingly weak prior information, the Bayes estimate converges to the mean of the data. Also note that as $\beta \rightarrow \infty$, the prior becomes improper (not normalizable).

What is the most conservative prior?

Often, there is not enough information to choose a prior density for the unknown parameters or the information available is not easily translated into a probabilistic statement; yet we need a prior to be able to apply Bayes' theorem. In this case we try to find a noninformative, or conservative, prior that will let us conduct the Bayesian inference while injecting a minimum of artificial information, that is, information not justified by the physical process.

We have defined the Bayes risk, r_{ρ} , and the Bayes estimator for a given prior density. It stands to reason that the more informative the prior, the smaller its associated risk; we therefore say that a prior ρ is least favorable if $r_{\rho} \geq r_{\rho'}$ for any other prior, ρ' . A least favorable prior is associated with the greatest unavoidable loss.

In the frequentist approach the greatest unavoidable loss is associated with the maximum of the risk [equation (4)] over all possible models. An estimator that minimizes this maximum risk is called a minimax estimator. Under certain conditions the Bayes estimator corresponding to a least favorable prior actually minimizes the maximum risk [see Lehmann (1983)]. This is true, for example, when the Bayes estimator has a constant risk. In this sense we can think of a least favorable prior as being a route to the most conservative Bayesian estimator.

How does one find a conservative (noninformative) prior? There is no easy answer; even the terms conservative and noninformative are not well defined. One possibility is to define a measure of information (e.g., entropy) and determine a prior that minimizes/maximizes this measure (e.g., maximum entropy). We could also look for priors that are invariant under some family of transformations. But even the popular maximum entropy methods run into problems (Kass and Wasserman, 1996; Seidenfeld, 1987). For more information on noninformative priors, see Box and Tiao (1973) and Kass and Wasserman (1996).

EXAMPLE: A TOY INVERSE PROBLEM

We consider a simple example of estimating the mean η of a unit variance normal distribution, $N(\eta, 1)$, with an observation d from $N(\eta, 1)$, given that $|\eta|$ is known to be bounded by β .

Following Stark (1997), we use this as a model of an inverse problem with a prior constraint. Without the prior bound, d is an estimator of η , but we hope to do better (obtain a smaller risk) by including the bound information. How can we include this information in the estimation procedure? One possibility is to use a Bayesian approach and assign a prior distribution to η , which is uniform on $[-\beta, \beta]$. We will show that this distribution injects stronger information than might be evident.

Bayes risk

Start with an observation d from $N(\eta, 1)$ and suppose we know a priori that $|\eta|$ is bounded by β . We incorporate the bound by assigning to η a prior uniformly distributed on $[-\beta, \beta]$. The joint distribution of η and d is then

$$f(d, \eta) = \frac{1}{2\beta} \mathcal{I}_{[-\beta, \beta]} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(d - \eta)^2\right],$$

where $\mathcal{I}_{[-\beta, \beta]}(x) = 1$ for $x \in [-\beta, \beta]$ and zero otherwise.

We reproduce Stark’s Monte Carlo calculation of the Bayes risk for this problem. Figure 1 shows the Bayes risk, using a uniform prior on $[-\beta, \beta]$ and the minimax risk to be described next. As the constraint weakens (β increases), the Bayes risk gets closer to 1.

Flat prior is informative

We used the uniform distribution to soften (i.e., convert to a probabilistic statement) the constraint $|\eta| \leq \beta$. Now we want to measure the effect of this constraint softening. Have we included more information than we really had?

Given the observation d from $N(\eta, 1)$ and knowing that $|\eta| \leq \beta$, what is the worst risk (mean square error) we can hope to achieve with the best possible estimator without imposing a prior distribution on η ? In other words, we want to compute the minimax risk, $R(\beta)$, given the bound β :

$$R(\beta) = \min_{\delta} \max_{\eta \in [-\beta, \beta]} E_{\mathcal{P}}[\eta - \delta(d)]^2.$$

The value $R(\beta)$ is a lower bound for the maximum risk of any other estimator. Although its exact value is difficult to compute, $R(\beta) \leq \min\{\beta^2, 1\}$. In addition, Donoho et al. (1990) show that

$$\frac{4}{5} \frac{\beta^2}{\beta^2 + 1} \leq R(\beta).$$

Figure 1 shows upper and lower bounds for the minimax risk as a function of β . For $\beta \leq 3$ the Bayes risk is outside the minimax bounds. This is an artifact of the way we have softened the bound. In other words, the uniform prior distribution injects more information than the hard bound on η , as judged by comparing the most pessimistic frequentist risk with that of the Bayesian estimator. Also, $R(\beta) \rightarrow 1$ as $\beta \rightarrow \infty$. So as the bound weakens, the Bayes and minimax risk both approach 1.

**PRIORS IN HIGH-DIMENSIONAL SPACES:
THE CURSE OF DIMENSIONALITY**

As we have just seen, most probability distributions usually have more information than implied by a hard constraint. To say, for instance, that any model with $\|\mathbf{m}\| \leq 1$ is feasible is certainly not the same thing as saying that all models with $\|\mathbf{m}\| \leq 1$

are equally likely. And while we could look for the most conservative or least favorable such probabilistic assignment, Backus (1988) makes an interesting argument against any such probabilistic replacement in high- or infinite-dimensional model spaces. His point can be illustrated with a simple example. Suppose that all we know about an n -dimensional model vector \mathbf{m} is that its length, $m \equiv \|\mathbf{m}\|$, is less than some particular value—unity, for the sake of definiteness. In other words, suppose we know a priori that \mathbf{m} is constrained to be within the n -dimensional unit ball B_n . Backus considers various probabilistic replacements of this hard constraint; this is called softening the constraint. We could, for example, choose a prior probability on \mathbf{m} which is uniform on B_n , namely, the probability that \mathbf{m} will lie in some small volume, $\delta V \in B_n$, shall be equal to δV divided by the volume of B_n . Choosing this uniform prior on the ball, one can show that the expectation of m^2 for an n -dimensional \mathbf{m} is

$$E(m^2) = \frac{n}{n + 2},$$

which converges to 1 as n increases. Unfortunately, the variance of m^2 goes as $1/n$ for large n ; thus, we seem to have introduced a piece of information that was not implied by the original constraint—namely that for large n , the only likely vectors \mathbf{m} will have length equal to one. The reason for this apparently strange behavior has to do with the way volumes behave in high-dimensional spaces. The volume $V_n(R)$ of the R -diameter ball in n -dimensional space is

$$V_n(R) = C_n R^n,$$

where C_n is a constant that depends only on the dimension n , not on the radius [this is a standard result in statistical mechanics; e.g., Becker (1967)]. If we compute the volume $V_{\epsilon, n}$ of an n -dimensional shell of thickness ϵ just inside an R -diameter

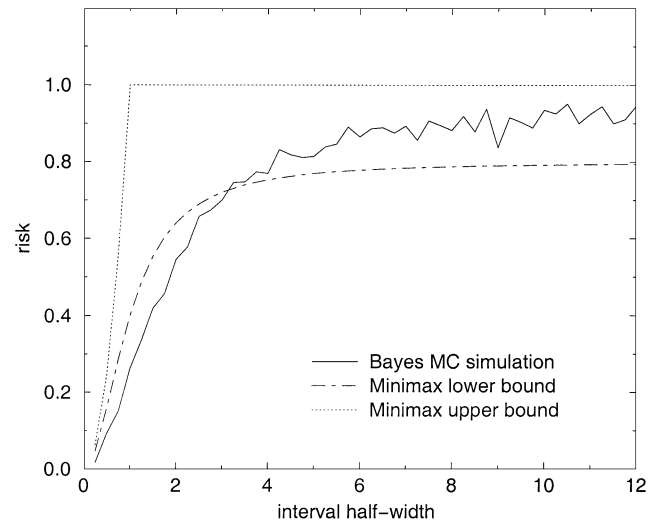


FIG. 1. For square error loss, the Bayes risk associated with a uniform prior is shown along with the upper and lower bounds on the minimax risk as a function of the size of the bounding interval $[-\beta, \beta]$. When β is comparable to or less than the variance (one in this case), the risk associated with a uniform prior is optimistic.

ball, we can see that

$$\begin{aligned}
 V_{\epsilon,n} &\equiv V_n(R) - V_n(R - \epsilon) = C_n(R^n - (R - \epsilon)^n) \\
 &= V_n(R) \left(1 - \left(1 - \frac{\epsilon}{R} \right)^n \right).
 \end{aligned}
 \tag{5}$$

Now, for $\epsilon/R \ll 1$ and $n \gg 1$, we have

$$V_{\epsilon,n} \approx V_n(R)(1 - e^{-n\epsilon/R}).$$

This says that as n gets large, nearly all of the volume of the ball is compressed into a thin shell just inside the radius.

But even this objection can be overcome with a different choice of probability distribution to soften the constraint. For example, choose m to be uniformly distributed on $[0, 1]$ and choose the $n - 1$ spherical polar angles uniformly on their respective domains. This probability is uniform on $\|\mathbf{m}\|$ but nonuniform on the ball. However, it is consistent with the constraint and has the property that the mean and variance of m^2 are independent of the dimension of the space (Scales, 1996).

So as Backus has said, we must be very careful in replacing a hard constraint with a probability distribution, especially in a high-dimensional model space. Apparently innocent choices may lead to unexpected behavior.

EXAMPLE: VERTICAL SEISMIC PROFILE

We now present a simple example related to the first question in the overview. We use a vertical seismic profile (VSP—used in exploration seismology to image the earth’s near surface) experiment to illustrate how a fitted response depends on the assumed noise level in the data. Figure 2 shows the geometry

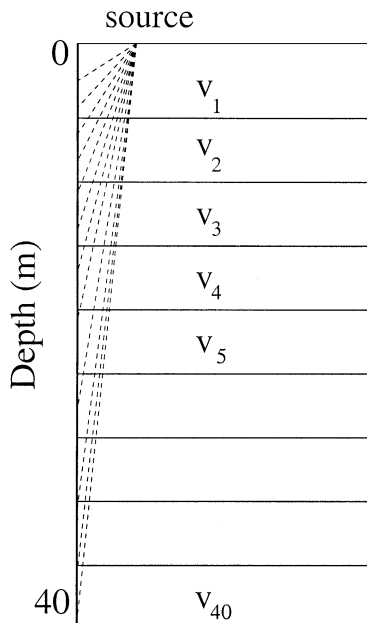


FIG. 2. Simple model of a vertical seismic profile (VSP). A source is at the surface of the earth near a vertical borehole (left side). A string of receivers is lowered into the borehole, recording the transit time of a downgoing compressional wave. These transit times are used to construct a best-fitting model. Here, v_i refers to the velocity in discrete layers, assumed to be constant. We ignore the discretization error in this calculation.

of a VSP. A source of acoustic energy is at the surface near a vertical borehole (left side). A string of receivers is lowered into a borehole, recording the transit time of the downgoing acoustic pulse. These transit times are used to construct a best-fitting model of the velocity $v(z)$ (or index of refraction) as a function of depth. There is no point in looking for lateral variations in v since the rays propagate nearly vertically. If the earth is not laterally invariant, this assumption introduces a systematic error into the calculation.

For each observation (and hence each ray) the forward problem is

$$t = \int_{\text{ray}} \frac{1}{v(z)} dl.$$

If the velocity model and the raypaths are known, then the traveltimes can be computed by integrating the velocity along the raypath.

The goal is to somehow estimate $v(z)$ or some functional of it. Unless v is constant, the rays will refract; therefore, the domain of integration depends on the unknown velocity. This makes the inverse problem mildly nonlinear. We will neglect nonlinearity in the present example by assuming the rays are straight lines.

How well a particular velocity model fits the data depends on how accurately the data are known. Roughly speaking, if the data are known very precisely, we will have to work hard to come up with a model that fits them to a reasonable degree. If the data are known only rather imprecisely, then we can fit them more easily. For example, in the extreme case of only noise, the mean of the noise is a reasonable fit to the data.

As a simple synthetic example we constructed a piecewise constant $v(z)$ with 40 layers and used 40 unknown layers to perform the reconstruction. We computed 78 synthetic traveltimes and contaminated them with uncorrelated Gaussian noise. The level of the noise is unimportant for our purposes; the point is that, given an unknown level of noise in the data, different assumptions about this noise will lead to different kinds of reconstructions. With the constant-velocity layers, the system of forward problems for all 78 rays reduces to

$$\mathbf{t} = \mathbf{K} \cdot \mathbf{s}, \tag{6}$$

where \mathbf{s} is the 40-dimensional vector of reciprocal layer velocity (slowness to seismologists) and \mathbf{K} is a matrix whose $i - j$ entry is the distance the i th ray travels in the j th layer. [See Bording et al. (1987) for the details behind this tomography calculation.] So the data mapping, g , is the inner product of the matrix \mathbf{K} and the slowness vector \mathbf{s} .

Let t_i^o be the i th observed traveltime, $t_i^c(\mathbf{s})$ the i th traveltime calculated through a given slowness model \mathbf{s} , and σ_i the standard deviation of the i th datum. If the true slowness is \mathbf{s}_o , then the following model of the observed traveltimes is assumed to hold:

$$t_i^o = t_i^c(\mathbf{s}_o) + \epsilon_i, \tag{7}$$

where ϵ_i is a noise term with zero mean and variance σ_i^2 . Our goal is to estimate \mathbf{s}_o . A standard approach to solve this problem is to determine slowness values \mathbf{s} that make a misfit function such as

$$\chi^2(\mathbf{s}) = \frac{1}{N} \sum_{i=1}^N \left(\frac{t_i^c(\mathbf{s}) - t_i^o}{\sigma_i} \right)^2, \tag{8}$$

smaller than some tolerance. Here, N is the number of observations. The symbol χ^2 is often used to denote this sum because when equation (7) holds and the noise is Gaussian and uncorrelated, $\chi^2(\mathbf{s}_o)$ is just an average of independent χ^2 -distributed variables.

We assume the number of layers is known—40 in this example—but this is usually not the case. Choosing too many layers may lead to an overfitting of the data. In other words we end up fitting noise-induced structures. Using an insufficient number of layers will not capture important features in the data. There are tricks and methods to try to avoid over- and underfitting. In this example we do not have to worry since we will be using simulated data. To determine the slowness values through equation (8), we have used a truncated singular value decomposition reconstruction [see Hansen (1998) for a definition and examples], throwing away all the eigenvectors in the generalized inverse approximation of \mathbf{s} that are not required to fit the data at the $\chi^2 = 1$ level. The resulting model is not unique, but it is representative of models that do not overfit the data (to the assumed noise level).

We consider the problem of fitting the data under two different assumptions about the noise. Figure 3 shows the observed and predicted data for models that fit the traveltimes on average to within 0.3 and 1.0 ms. Remember, the actual pseudo-random noise in the data is fixed throughout; all we are changing is our assumption about the noise, which is reflected in the data misfit criterion.

We refer to these as the optimistic (low-noise) and pessimistic (high-noise) scenarios. The smaller the assumed noise level in the data, the more the predicted data must follow the pattern of the observed data. It takes a complicated model to predict complicated data. Therefore, we should expect the best fitting model that produced the low-noise response to be more

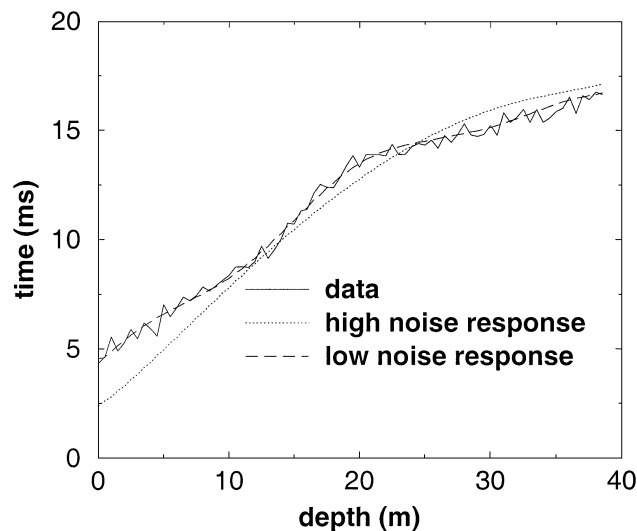


FIG. 3. Observed data (solid curve) and predicted data for two different assumed levels of noise. In the optimistic case (dashed curve) we assume the data are accurate to 0.3 ms. In the more pessimistic case (dotted curve) we assume the data are accurate to only 1.0 ms. In both cases the predicted traveltimes are computed for a model that just fits the data. In other words we perturb the model until the rms misfit between the observed and predicted data is about N times 0.3 or 1.0, where N is the number of observations. Here $N = 78$.

complicated than the model that produced the high-noise response. If the error bars are large, then a simple model will explain the data.

Now let us look at the models that actually fit the data to these different noise levels (Figure 4). It is clear that if the data uncertainty is only 0.3 ms, then the model predicts (or requires) a low-velocity zone. However, if the data errors are as much as 1 ms, then a very smooth response is enough to fit the data, in which case a low-velocity zone is not required. In fact, for the high-noise case essentially a linear velocity increase will fit the data, while for the low-noise case a rather complicated model is required. (In both cases because of the singularity of \mathbf{K} , the variances of the estimated parameters become very large near the bottom of the borehole.)

This example illustrates the importance of understanding the noise distribution to properly interpret inversion estimates. In this particular case, we didn't simply pull these standard deviations out of a hat. The low value (0.3 ms) is what we get if we assume the only uncertainties in the data are normally distributed fluctuations about the running mean of the traveltimes. However, nature doesn't know about traveltimes. Traveltimes are approximations to the true properties (i.e., finite bandwidth) of waveforms. Further, traveltimes themselves are usually assigned by a human interpreter looking at the waveforms. Based on these considerations, we might be led to conclude that a more reasonable estimate of the uncertainties for real data would be closer to 1 ms than 0.3 ms. In any event, the interpretation of the presence of a low-velocity zone should be viewed with some scepticism unless the smaller uncertainty level can be justified.

To summarize, this example shows it is impossible to know whether a certain model feature is resolved unless one understands the data uncertainties.

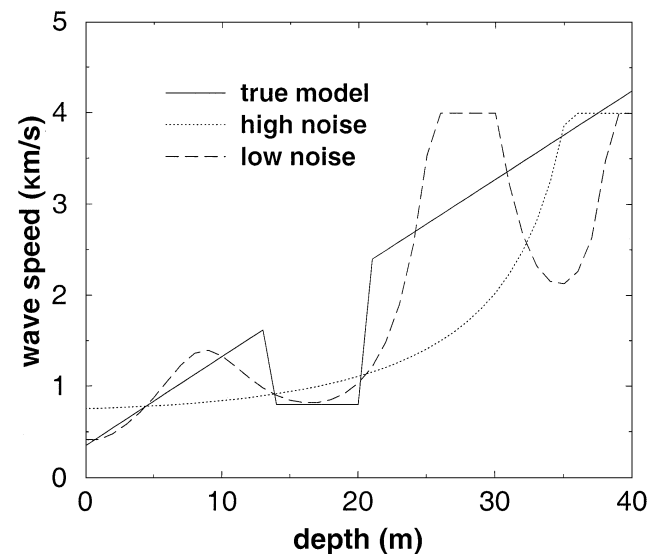


FIG. 4. The true model (solid curve) and the models obtained by a truncated SVD expansion for the two levels of noise, optimistic (0.3 ms, dashed curve) and pessimistic (1.0 ms, dotted curve). Both of these models just fit the data in the sense that we eliminate as many singular vectors as possible and still fit the data to within one standard deviation (normalized $\chi^2 = 1$). An upper bound of four has also been imposed on the velocity. The data fit is calculated for the constrained model.

SUMMARY

To solve any inverse problem we need an understanding of the uncertainties in the data. We also need methods to incorporate data-independent prior information. Both of these issues involve subtle choices that may significantly influence the results of inverse calculations. The specification of prior information is especially controversial. In this tutorial we have presented Bayesian and frequentist methodologies that can be used to incorporate information into inverse calculations.

We can choose among a variety of methods to obtain inversion estimates, but it is important to make sure the uncertainty estimates we use are not artificial. Whether one uses a frequentist or a Bayesian approach, it is always important to be aware of the model assumptions on which estimates rely. "The choice of models is usually a more critical issue than the differences between the results of various schools of formal inference" (Cox, 1981).

ACKNOWLEDGMENTS

We are grateful to A. Carnero, K. Osypov, P. Stark, and S. Treitel for helpful discussions and support. This tutorial is based on two lectures delivered at the 1998 Mathematical Geophysics Summer School in Stanford. This work was partially supported by the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, Colorado School of Mines.

REFERENCES

- Backus, G., 1988, Hard and soft prior bounds in geophysical inverse problems: *Geophys. J.*, **94**, 249–261.
- Becker, R., 1967, *Theory of heat*: Springer-Verlag.
- Berger, L. O., 1985, *Statistical decision theory and Bayesian analysis*: Springer-Verlag.
- Bording, R. P., Gersztenkorn, A., Lines, L. R., Scales, J. A., and Treitel, S., 1987, Applications of seismic travel time tomography: *Geophys. J. Roy. Astr. Soc.*, **90**, 285–303.
- Box, G. E. P., and Tiao, G. C., 1973, *Bayesian inference in statistical analysis*: John Wiley & Sons, Inc.
- Carlin, B. P., and Louis, T. A., 1996, *Bayes and empirical Bayes methods for data analysis*: Chapman & Hall.
- Casella, G., 1985, An introduction to empirical Bayes data analysis: *Am. Statistician*, **39**, 83–87.
- Cox, D. R., 1981, Theory and general principles in statistics: *J. Roy. Stat. Soc. A*, **144**, 289–297.
- Dawid, A. P., 1983, Invariant prior distributions, in Kotz, S., and Johnson, N. L., Eds. *Encyclopedia of statistical sciences*, John Wiley & Sons Inc., 228–236.
- Donoho, D. L., Liu, R. C., and MacGibbon, K. B., 1990, Mimimax risk over hyperrectangles, and implications: *Annals Stats.*, **18**, 1416–1437.
- Efron, B., 1986, Why isn't everyone a Bayesian: *Am. Stat.*, **40**, No. 1, 1–11.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., 1997, *Bayesian data analysis*: Chapman & Hall.
- Goutis, C., and Casella, G., 1995, Frequentist post-data inference: *Internat. Stat. Rev.*, **63**, No. 3, 325–344.
- Gouveia, W., and Scales, J. A., 1997, Resolution in seismic waveform inversion: Bayes vs. Occam: *Inverse Problems*, **13**, 323–349.
- 1998, Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis: *J. Geophys. Res.*, **103**, 2759–2779.
- Hansen, P. C., 1998, Rank-deficient and discrete ill-posed problems: *Monographs on Mathematical Modeling and Computation*: SIAM.
- Jaynes, E. T., 1968, Prior probabilities: *IEEE Trans. Syst., Science and Cyber.*, **4**, 227–241.
- 1982, On the rationale of maximum entropy methods: *Proceedings of IEEE*, **70**, 939–952.
- Kass, R., and Wasserman, L., 1996, The selection of prior distributions by formal rules: *J. Am. Stat. Ass.*, **91**, 1342–1370.
- Lehmann, E., 1983, *Theory of point estimation*: John Wiley & Sons, Inc.
- Lindley, D. V., 1975, The future of statistics—A Bayesian 21st century: *Supp. Adv. Appl. Prob.*, **7**, 106–115.
- Philips, D. L., 1962, A technique for the numerical solution of certain integral equations of the first kind: *J. Assoc. Comput. Mach.*, **9**, 84–97.
- Rubin, D. R., 1984, Bayesianly justifiable and relevant frequency calculations for the applied statistician: *Ann. Stats.*, **12**, No. 4, 1151–1172.
- Scales, J. A., 1996, Uncertainties in seismic inverse calculations, in Jacobson, B. H., Mosegaard, K., and Sibani, P., Eds., *Inverse methods*, Springer-Verlag, 79–97.
- Scales, J. A., and Snieder, R., 1997, To Bayes or not to Bayes?: *Geophysics*, **63**, 1045–1046.
- Scales, J. A., Docherty, P., and Gersztenkorn, A., 1990, Regularization of nonlinear inverse problems: Imaging the near-surface weathering layer: *Inverse Prob.*, **6**, 115–131.
- Seidenfeld, T., 1987, Entropy and uncertainty, in MacNeill, I. B., and Umphrey, G. J., Eds., *Foundations of statistical inference*: Reidel Publ. Co., 259–287.
- Stark, P. B., 1997, Does God play dice with the Earth? (And if so, are they loaded?): *Geosciences Meeting*, SIAM, oral presentation. Available at www.stat.berkeley.edu/users/stark/.
- Tanner, M. A., 1993, *Tools for statistical inference. methods for the exploration of posterior distributions and likelihood functions*: Springer-Verlag.
- Tikhonov, A., and Arsenin, V., 1977, *Solutions of ill-posed problems*: John Wiley & Sons, Inc.
- Trampert, J., and Snieder, R., 1996, Model estimations biased by truncated expansion: Possible artifacts in seismic tomography: *Science*, **271**, 1257–1260.