Prior Information and Uncertainty in Inverse Problems

John A. Scales (1,2) and Luis Tenorio (3)

(1) Department of Geophysics
(2) Center for Wave Phenomena
(3) Deptartment of Mathematical and Computer Sciences
Colorado School of Mines,
Golden CO 80401, USA

Contents

1 Overview

Solving an inverse problem means making inferences about physical systems from real data. These inferences are based on mathematical representations of the systems; we call these representations models. Functionals of the models represent observable properties of the system; for example, the mass density as a function of space in the Earth, the depth of the continents, the radius of the core-mantle boundary.

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

I. How accurately are the data known? That is, what does it mean to "fit the data"?

II. How accurately is the physical system modeled? Does the model include all the physical effects that contribute significantly to the data?

III. What is known about the model parameters independently of the data? In other words, what does it mean for a model to be reasonable or unreasonable?

The note is organized as follows. In Section 2 we set the general framework for the inverse problems that will be considered. In Section 3 we 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 the observed data and prior information. Bayesians introduce probabilities on the space of models (prior information is thus probabilistic) and condition on the observed data. Frequentists, on the other hand, assume a distribution prior to observing the data, which does not change once the data have been observed, and use deterministic prior information; probability enters the calculations via the data errors, which are assumed to have a random component. The choice of prior probability model for the Bayesian inference is not always clear even when the prior information is well defined; this is discussed in Section 4 and exemplified with a toy problem in Section 5. The example illustrates how representing deterministic constraints probabilistically may inject information into the calculation that is not strictly required by the constraint. This problem becomes worse in high-dimensional spaces. In Sections 7 and 8 we provide two examples of inverse problems based on real data to illustrate the points raised in the note.

1.1 Some Notation

Here is a short summary of the notation that will be used henceforth.

> $\mathcal{M}$: The space of Earth models. A linear vector space, usually infinite dimensional. E.g., $\mathcal{L}^2(R^3)$.
>
> $\mathcal{D}$: The space of measurements. In practice always finite dimensional. Typically R^n, where n is the number of observations.
>
> $\mathbf{D}$: Data random variable, usually vector valued.
>
> $\mathbf{d}$: Values taken on by $\mathbf{D}$, the data. $\mathbf{d} \in \mathcal{D}$. Where there is no danger of confusion we will use $\mathbf{D}$ and $\mathbf{d}$ interchangeably.
>
> $\mathbf{m}$: A model or parameter. $\mathbf{m} \in \mathcal{M}$.
>
> $E(\,)$, $var(\,)$: Expectation and variance operators.

Models are usually parameterized so that estimating a model is equivalent to estimating its corresponding parameters. But, clearly, the choice of observables and parameterization is not unique. For instance, in problems of elasticity we can use the elastic stiffness tensor of the elastic compliance tensor. We can use wave-speed, or wave slowness. An estimator of a model $\mathbf{m}$ (or a functional thereof) is a function $\delta : \mathcal{D} \rightarrow \mathcal{M}$. The estimator may just estimate the parameters that characterize the model. The estimate given the data $\mathbf{d}$ is denoted as $\delta(\mathbf{d})$.

2 A General Statement of the Inverse Problem

As the result of some measurement, or set of measurements, a number n of data are collected. The amount of data is always finite because instruments have finite bandwidth. We usually take the data space $\mathcal{D}$ to be R^n. The data are related to the physical models through the forward modeling operator (also known as the data mapping). The forward operator is a function that maps vectors from model space into data space

$$g : \mathcal{M} \to \mathcal{D}.$$

In practice the forward operator g is always an approximation. In geophysics this is primarily because one cannot afford to represent the true complexity of the Earth on computer. Even if this were possible it might not be worth the effort given the instrument's resolution and noise level in the data. This will be discussed in detail in the note, but for now it suffices to be aware that there is an error associated with using g. Since this will be an error associated with the ability to predict the measurement, it is an n-dimensional vector $\mathbf{f}$. Finally, there will be an n-dimensional vector of random measurement errors, $\mathbf{e}$. So the connection between models and data can be written as:

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

Given measurements $\mathbf{d}$, the goal is to estimate the model $\mathbf{m}$ (or a functional $L(\mathbf{m})$). A function of the data that is used to estimate the model $\mathbf{m}$ (i.e., the inversion algorithm) is called an estimator of $\mathbf{m}$.

Note that since g maps an infinite dimensional space into a finite dimensional space, the data mapping has a non-trivial kernel. So, even in the absence of measurement and modeling errors the data mapping will not be invertible and the set of models that predict the data equally well may be quite large. This in itself in not a problem, the problem is when these equally predicting models yield wildly different values for the model functional we are interested in. By including prior information we attempt to constrain the range of feasible models and thus control the effect of those null elements. This is illustrated with an example in Section 2.1. Even when there is a unique solution, it may be unstable to small perturbations in the data. In this case we may also use some prior information to stabilize the solution.

2.1 Example: Estimating the derivative of a smooth function

To motivate the discussion let us consider a simple example which shows the necessity of including prior information. Later, in Section 4, we will introduce tools from statistical decision theory that allow us to quantify the influence of different types of prior information.

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

$$f_i = f(x_i) + \epsilon_i, \;\; i = 1, ..., n, \tag{1}$$

where the errors ϵ_i are assumed to be *iid* $N(0, \sigma^2)$[1]. We want to use these observations to estimate the derivative f'. We define the estimator

$$\hat{f}'(x_{m_i}) = \frac{f_{i+1} - f_i}{h}, \tag{2}$$

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 squared error (MSE). The variance and bias of (2) are

$$\mathrm{Var}[\hat{f}'(x_{m_i})] = \frac{2\sigma^2}{h^2},$$

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

for some $\alpha_i \in [x_i, x_{i+1}]$. We need more information to assess the size of the bias. Let us assume that the second derivative is bounded on $[a, b]$

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

It then follows that

$$|\mathrm{Bias}[\hat{f}'(x_{m_i})]| = |f''(\beta_i)(\alpha_i - \beta_i)| \leq Mh,$$

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

$$\frac{2\sigma^2}{h^2} \leq \mathrm{MSE}[\hat{f}'(x_{m_i})] = \mathrm{Var}[\hat{f}'(x_{m_i})] + \mathrm{Bias}[\hat{f}'(x_{m_i})]^2 \leq \frac{2\sigma^2}{h^2} + M^2h^2. \tag{3}$$

[1] Independent, identically distributed random variables, normally distributed with mean 0 and variance σ^2.

It is clear that choosing the smallest h possible does not lead to the best estimate; the noise has to be taken into account. In fact, the lowest upper bound is obtained for $h = 2^{1/4}\sqrt{\sigma/M}$. The larger the variance of the noise, the wider the spacing between the points. But, do we really need to assume any more prior information, in addition to model (1), to bound the MSE? We do. Take any smooth function g which vanishes at the points $x_1, ..., x_n$. 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 we can not estimate the derivative with finite uncertainty.

3 Bayesian and Frequentists Methods of Inference

There are two fundamentally different meanings of the term “probability” in common usage (Scales & 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 trails, 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 using the data. This knowledge is cast in a probabilistic form and is called the prior probability model. Prior means before the data have been recorded; i.e., information that is independent of the data to be recorded. If one has a set of data whose statistical characteristics are known (e.g., the data covariance matrix for Gaussian errors), then Bayesian inversion provides a framework for combining the probabilistic prior model information with the information contained in the observed data in order to refine the prior distribution. The updated distribution is the posterior model distribution given the data; it is what we know after we have assimilated the data and the prior information. The point of using the data is that the posterior model information hopefully constrains the model more tightly than the prior model distribution.

However, the notion of prior model statistics can in practice be somewhat shaky. As an example, consider a seismic survey. In such a situation one may have a fairly accurate idea of the ranges of seismic velocity and density that are realistic, and perhaps even the vertical correlation length (if bore-hole 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 for this is that in practice the formalism of prior model statistics is used to regularize the posterior solution. But logically, the prior distribution must be known a priori, in which case it cannot matter whether it regularizes the problem (Gouveia & Scales, 1997). In practice, 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. Note that in such an approach the prior model statistics are used a posteriori to tune the retrieved model!

Bayesian statistics relies completely on the specification of prior model statistics, i.e. on the knowledge that one has of the model before using the recorded data. The flexibility taken in using the prior model statistics as a knob to tune properties of the retrieved model therefore is completely at odds with the philosophy of Bayesian inversion. This does not mean that there is anything wrong with Bayesian inversion, but it does suggest 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.”

3.1 Bayesian Inversion in Practice

In practice it is difficult to honestly use Bayes' Theorem to solve realistic inverse problems. On the one hand the information at our disposal that could be regarded as being known a priori is highly varied and often difficult to quantify objectively. On the other hand a complete Bayesian analysis may be computationally intractable.

There are two important questions that have to be addressed in any Bayesian inversion:

- How do we represent the prior information? This applies both to the prior model information and to the description of the data statistics.
- 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 simply use posterior means as estimates and posterior standard deviations as "error bars". However, very seldom will we be able to compute all the posterior estimates analytically; we often have to use computer-intensive approximations based on Markov Chain Monte Carlo methods (see for example, Tanner (1993)).

The first question is a lot more difficult to answer. In practice there are essentially three strategies. The 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 a parameter μ must lie between a and b, is it justified to assume that μ has a uniform prior distribution on the interval $[a, b]$? We will address this question in an example below, but for now simply observe that there are infinitely many probability distributions consistent with this statement. To pick one may be an over-specification 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 prescribe the statistical properties of the prior model.

One way out of this dilemma is to ignore it and presume that "probability lies in the eye of the beholder", but this means that our posterior probability will be different from yours.

A second approach attempts to make a more objective choice of priors by relying on theoretical considerations such as maximum entropy, transformation invariance (Jaynes invariant prior); or by somehow using a large number of observations to estimate a prior. This latter approach is sometimes called *empirical Bayes.* For example, suppose one is doing a gravity inversion to estimate mass density in some reservoir. Suppose further that there are available a large number of independent, identically distributed laboratory measurements of density for rocks taken from this reservoir (a big if!). Then 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 & Scales, 1998), where in-situ (bore-hole) measurements are used as an empirical prior for surface seismic data. The empirical Bayes analysis can be seen as an approximation to a full *hierarchical Bayes* analysis based on the joint probability distribution of all parameters and available data. For an introduction to empirical and hierarchical models see, for example, Gelman et al. (1997) and references therein. For a review on the development of objective priors see Kass & Wasserman (1996).

3.2 Bayes vs Frequentist

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 25 $\mathrm{g/cm^3}$ (a combination of observation and theory that is almost certainly true). 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. Then perform surgery on this set, cutting away those models that violate the deterministic criteria, e.g., have negative density. The result will be a (presumably smaller) set of models that fit the data and satisfy the prior considerations. In this framework no particular model in the constrained set has special significance. We have no way of saying that model $\mathbf{m}_1$ is more likely than model $\mathbf{m}_2$, since in this frequentist approach we have not defined a probability distribution on the models themselves. All we do is choose any model that fits the data to a desired level and satisfies the prior model constraints. Tikhonov's regularization is one way of obtaining an inversion algorithm by restricting the family of models that fit the data; e.g., among all the models that fit the data, you choose one that has particular features, the smoothest, the shortest, etc. (For instance, see Scales et al. (1990) and Gouveia & Scales (1997)).

In the Bayesian paradigm, probability distributions are the fundamental tools. Bayesians regard it as meaningful to speak of the probability of a hypothesis given some evidence, and are able to conduct pre-data and post-data inferences. Frequentists 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 & Casella (1995) for a review of methodologies that have been developed to do frequentist post-data inference.

We have seen that 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 happen that parameters are not well defined. For instance, is the "true mass of the earth" a meaningful expression? 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, e.g., micrometeorites? Even if you make the "true mass of the Earth" well-defined (it will still be arbitrary to some extent), it can never precisely known any more than the temperature of an isolated gas can be.

So, which approach is better? Bayesians are happy to point out some well known inconsistencies in the frequentist methodology. 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 prior and 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). See also Rubin (1984) for ways in which frequentist methods can be used to complement Bayesian inferences.

4 What Difference Does the Prior Make?

In a Bayesian calculation, whatever we are estimating depends on the prior and conditional distributions given the data. As far as we know, there is no established procedure to check how much information the prior injects into the posterior estimates. In this example we will compare the *risks* of the estimators. To measure the performance of an estimator $\delta(\mathbf{d})$ of $\mathbf{m}$ we define the loss function $L(\mathbf{m}, \delta(\mathbf{d}))$; L should always be non-negative and should be zero for the true model. That is, for any $\mathbf{s} \in \mathcal{M}$ $L(\mathbf{s}, \mathbf{m}) \geq 0$ and $L(\mathbf{m}, \mathbf{m}) \equiv 0$. The loss is a measure of the cost of estimating the true model to be $\delta(\mathbf{d})$ when it is actually $\mathbf{m}$. For example, a common loss function is the squared error: $L(\mathbf{m}, \delta(\mathbf{d})) = (\mathbf{m} - \delta(\mathbf{d}))^2$. But there are other choices like ℓ_p-norm error.

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

$$\textbf{Risk} \quad R(\mathbf{m}, \delta) = E_P[L(\mathbf{m}, \delta(\mathbf{d}))] \tag{4}$$

where P is the probability distribution of the errors. For squared error loss the risk is the usual mean squared error.

4.1 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 the models into account. The Bayesian risk is defined as the expected value of the risk over the model distribution; in other words, it is a weighted average of the risk using the model distribution as weight function:

$$\textbf{Bayes Average Risk} \quad r_\rho = E_{\rho,P}[L(\mathbf{m}, \delta(\mathbf{d}))],$$

where ρ is the prior distribution on the models. An estimator with the smallest Bayesian risk is called a *Bayes estimator*. Note that we have used a frequentist approach to define the Bayes risk, since we have not conditioned on the observed data; this goes against the principles of some Bayesians. It does make sense, however, to expect good frequentist behavior if the Bayesian approach is to be used repeatedly with different data sets.

Denote by f the joint distribution on models and data. The marginal of f with respect to the data is obtained by integrating f over the models:

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

From Bayes' theorem, the conditional probability on $\mathbf{m}$ 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 a priori distribution, is the marginal of f with respect to $\mathbf{m}$. The conditional probability $p(\mathbf{m}|\mathbf{d})$ is the so-called Bayesian posterior probability, expressing the idea that $p(\mathbf{m}|\mathbf{d})$ assimilates the data and prior information.

One can define a number of reasonable estimators of $\mathbf{m}$ from $p(\mathbf{m}|\mathbf{d})$. For example, the $\mathbf{m}$ that maximizes $p(\mathbf{m}|\mathbf{d})$ (or $P[\|\delta(\mathbf{d}) - \mathbf{m}\| < c]$ for some $c > 0$). Or one could compute the estimator that gives the smallest Bayes risk for a given prior ρ and loss function L. We have the following theorem:

Theorem For squared error loss function the Bayes estimator is the posterior mean. Lehmann (1983), page 239.

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

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

The posterior mean is given by

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

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

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

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

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

In this case 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 recorded. The Bayes risk is the integral over the data of the posterior variance of a. Since the posterior variance does not depend of $\mathbf{d}$, the Bayes risk is just the posterior variance. Note also that as $\beta \to 0$, increasingly strong prior information, the estimate converges to the prior mean. As $\beta \to \infty$, increasingly weak prior information, the Bayes estimate converges to the mean of the data. Also note that as $\beta \to \infty$ the prior is improper (not normalizable).

4.2 What is the Most Conservative Prior?

It often happens that there is not enough information to construct a prior density for the unknown variables, or that 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 prior that will allow us to conduct the Bayesian inference while injecting a minimum of artificial information; that is, information which is not justified by the process. When no prior information is available we call conservative priors noninformative.

We have defined the Bayes risk r_ρ for a Bayes estimator δ_ρ given a prior ρ. It stands to reason that the more informative the prior the smaller its associated risk; we therefore say that the 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 risk (4) over all the possible models. An estimator that minimizes the maximum risk is called a *minimax estimator.* Lehmann (1983) shows that under certain conditions the estimator corresponding to a least favorable prior actually minimizes the maximum risk. 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? Again, there is no easy answer, even the terms ‘conservative’ and ‘noninformative’ are not well defined. But one can, for example, define a measure of information (e.g. entropy) and determine a prior which minimizes/maximizes this measure (e.g. maximum entropy); or one could define properties that noninfomative priors are expected to have (e.g., invariance). See Kass & Wasserman (1996).

5 Example: A Toy Inverse Problem

In Section 8 we shall discuss a cosmological problem where we explore the sensitivity of some unknown parameters to different amounts of prior information. Here we consider a simple example of estimating the mean a of a unit variance normal distribution $N(a, 1)$ with an observation d from $N(a, 1)$ given that $|a|$ is known to be bounded by b. These two problems are actually related. Following Stark (1997), we will use this as a model of an inverse problem with a prior constraint. Without the prior bound, d is an estimator of a 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 a which is uniform on $[-b, b]$. We will show that this uniform distribution injects stronger information than might be evident.

5.1 Bayes Risk

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

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

where $\mathcal{I}_{[-b,b]}$ is the indicator function for the interval $[-b, b]$; i.e.,

$$\rho(a) = \frac{1}{2b}\mathcal{I}_{[-b,b]}(a).$$

Stark (1997) shows the results of a Monte Carlo calculation of the Bayes risk for this problem. We will reproduce this calculation, although it is possible to compute the Bayes estimate and risk via numerical integration. Figure 1 shows the results of this Monte Carlo calculation of Bayes risk with a uniform prior on $[-b, b]$ compared with the minimax risk to be described next. As the constraint weakens (b increases) the Bayes risk gets closer to 1. (The dashed and dotted curves in this figure will be explained in the next section.)

5.2 The Flat Prior is Informative

We have used the uniform distribution to "soften" (i.e., convert to a probabilistic statement) the constraint $a \in [-b, b]$. Now we want to measure the effect of this softening of the constraint. Have we included more information than we really had?

Given the observation d from $N(a, 1)$ and knowing that $|a| \leq b$, what is the worst risk (mean squared error) we may hope to achieve with the *best* estimator without imposing a prior distribution on a? In other words we want to compute the minimax risk $R(b)$ given the bound b

$$R(b) = \min_\delta \max_{a \in [-b,b]} E_P(a - \delta(d))^2.$$

$R(b)$ is a lower bound for the maximum risk and an upper bound for the risk of any other estimator. Although it is difficult to compute its exact value, it is easy to see that $R(b) \leq \min\{b^2, 1\}$. Donoho et al. (1990) show that

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

Figure 1 shows upper and lower bounds for for the minimax risk as a function of b. Note that for $b \leq 3$ 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 a, as judged by comparing the most pessimistic frequentist risk with that of the Bayesian estimator. For $b > 4$ the comparison is not easy but it can also be shown that $R(b) \to 1$ as $b \to \infty$. So, as the bound weakens the Bayes and minimax risk both approach 1.

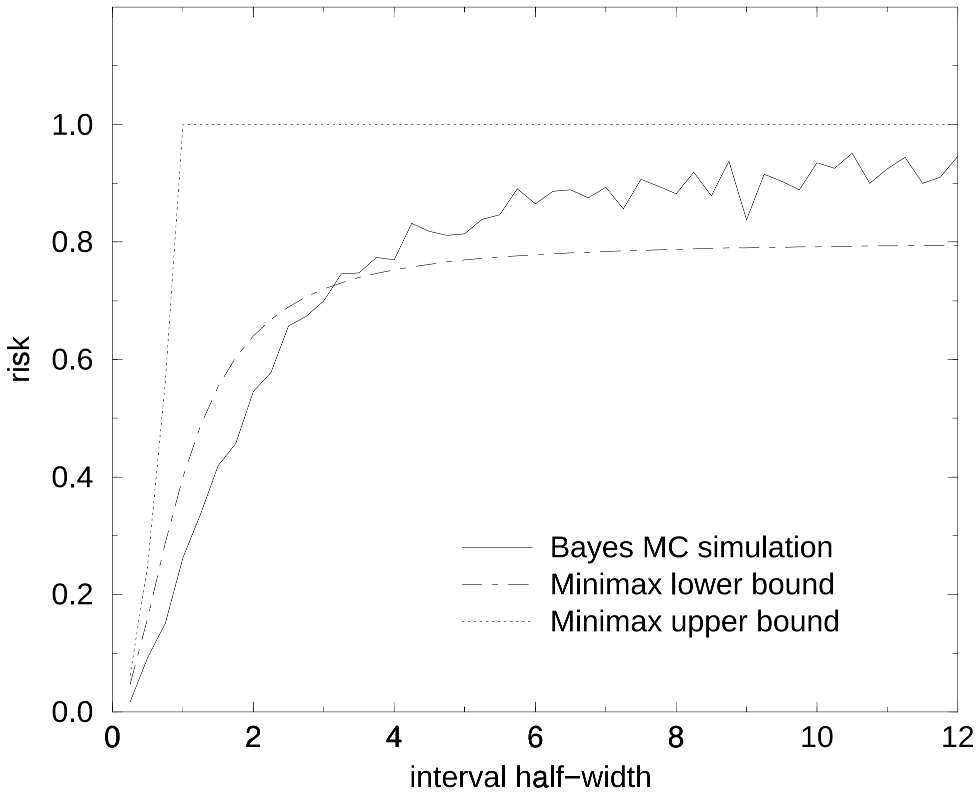

Figure 1: For squared-error loss, the Bayes risk associated with a uniform prior is shown along with the upper and lower bounds on the minimax risk as a function of the size of the bounding interval $[-b, b]$. When b is comparable to or less than the variance (1 in this case), the risk associated with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

6 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, it is not difficult to 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, and 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. If we compute the volume of an $n-$dimensional shell of thickness ϵ just inside an $R-$diameter ball we can see that:

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

where C_n depends only on the dimension. Now for $\epsilon/R \ll 1$ and $n \gg 1$ we have

$$V_\epsilon \approx V(R)\left(1 - e^{-n\epsilon/R}\right).$$

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 non-uniform on the ball. However it is consistent with the constraint and has the property that the mean and variance of m^2 is independent of the dimension of the space.

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. For more information on non-informative priors see Box and Tiao (1973) and Kass & Wasserman (1996).

7 Example: Vertical Seismic Profile

We now present a simple example related to Question I in Section 1. 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 of a VSP. A source of acoustic energy is at the surface near a vertical bore-hole (left side). A string of receivers is lowered into a bore-hole, recording the transit time of the down-going acoustic pulse. These transit times are used to construct a "best-fitting" model of the velocity (or index of refraction) as a function of depth $v(z)$. 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

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

If the velocity model $v(z)$ and the ray paths are known, then the travel time can be computed by integrating the velocity along the ray path.

The goal is to somehow estimate $v(z)$ (or some functional of $v(z)$), or to estimate confidence intervals for $v(z)$. Unless v is constant, the rays will refract and 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 that the rays are straight lines.

How well a particular $v(z)$ 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 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 travel times and contaminated them with uncorrelated Gaussian noise. The level of the noise doesn't matter for the present 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

$$\mathbf{t} = J \cdot \mathbf{s} \tag{6}$$

where $\mathbf{s}$ is the 40-dimensional vector of reciprocal layer velocity (*slowness* to seismologists) and J 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 J and the slowness vector $\mathbf{s}$.

Let t_i^o be the $i-$th observed travel time, $t_i^c(\mathbf{s})$ is the i-th travel time calculated through a given slowness model $\mathbf{s}$, and σ_i is the standard deviation of the i-th datum. If the true slowness is $\mathbf{s}_o$, then the following model of the observed travel times 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

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

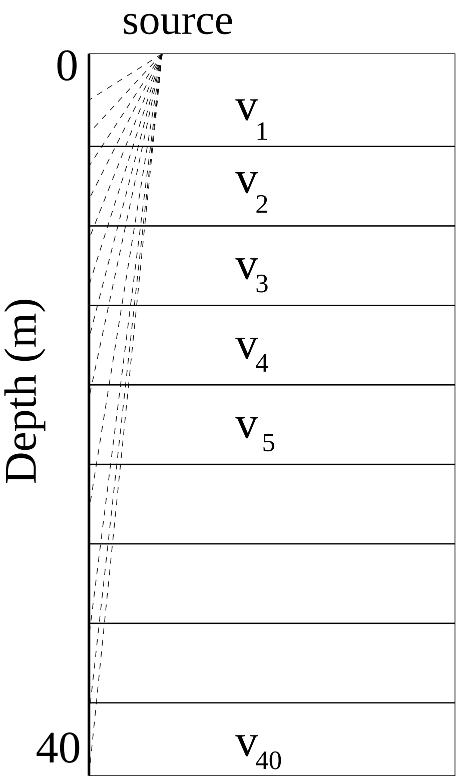

Figure 2: Simple model of a vertical seismic profile (VSP). A source is at the surface of the Earth near a vertical bore-hole (left side). A string of receivers is lowered into the bore-hole, recording the transit time of a down-going compressional wave. These transit times are used to construct a "best-fitting"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

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

We have assumed that 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 over-fitting 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 under-fitting. In the present example we do not have to worry since we will be using simulated data. To determine the slowness values through (8) we have used a truncated SVD reconstruction, 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 over-fit the data (to the assumed noise level).

We will 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 travel times on average to within 0.3 ms 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. You can clearly see that 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 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; these are shown in 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 $v(z)$ 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 J, the variances of the estimated parameters become very large near the bottom of the borehole.)

Hopefully 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 hat. The low value (0.3 ms) is what you happen to get if you assume that the only uncertainties in the data are normally distributed fluctuations about the running mean of the travel times. However, keep in mind that nature doesn't really know about travel times. Travel times are approximations to the true properties (i.e., finite bandwidth) of waveforms. Further, the travel times themselves are usually assigned by a human interpreter looking at the waveforms. Based on these considerations, one 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.

8 Example: Cosmic Microwave Background

The cosmic microwave background (CMB) is the radiation left over from the Big Bang. Through the CMB we see the universe as it was only 300,000 years after the Big Bang: Studying the CMB is like doing archaeology at a cosmological scale. In 1992 cosmologists established the existence of small directional variations in the CMB temperature (Smoot et. al (1992)). Had these variations not been found, cosmologists would have required completely new theories to explain the large-scale structure we see in the Universe today. Interested readers may consult Silk (1997) for a gentle introduction to cosmology.

For the moment it suffices to think of the CMB as a temperature function defined on the unit sphere. Let $T(\hat{r})$ be the CMB temperature in the direction $\hat{r}$ in the sky. We assume that T is a square-integrable function; the model space is $\mathcal{M} = \mathcal{L}^2(S^2)$. Any function T in $\mathcal{M}$ has a spherical harmonic representation

$$T(\hat{r}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell,m} Y_{\ell,m}(\hat{r}),$$

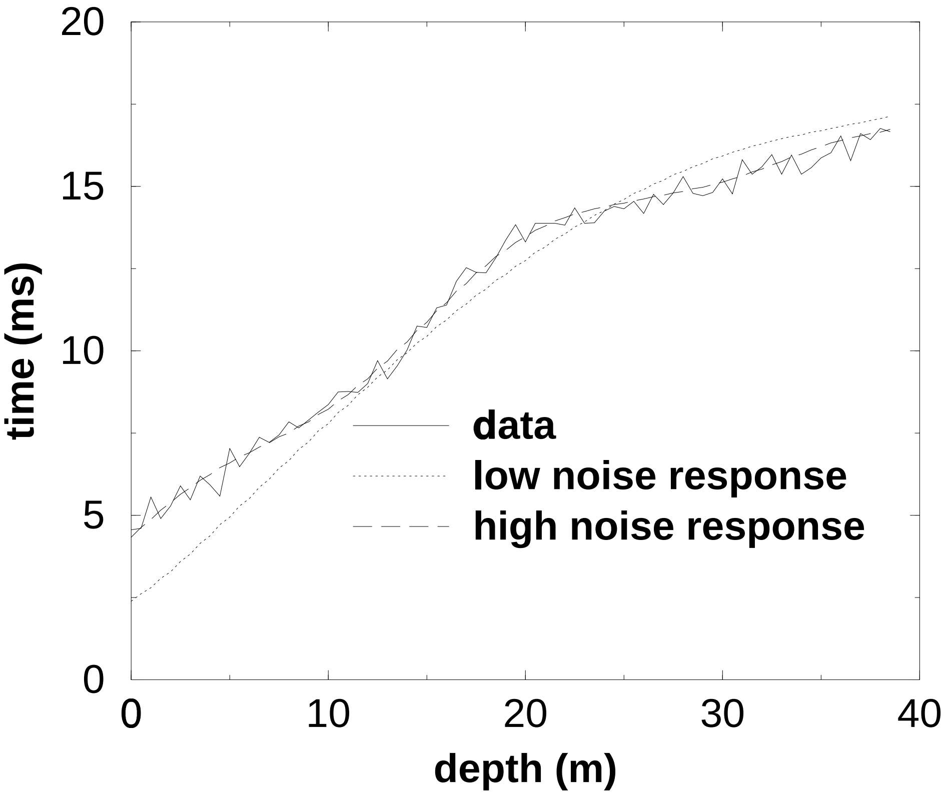

Figure 3: Observed data (solid curve) and predicted data for two different assumed levels of noise. In the optimistic case (dashed curve) we assume the data are accurate to 0.3 ms. In the more pessimistic case (dotted curve), we assume the data are accurate to only 1.0 ms. In both cases the predicted travel times 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 $N = 78$.

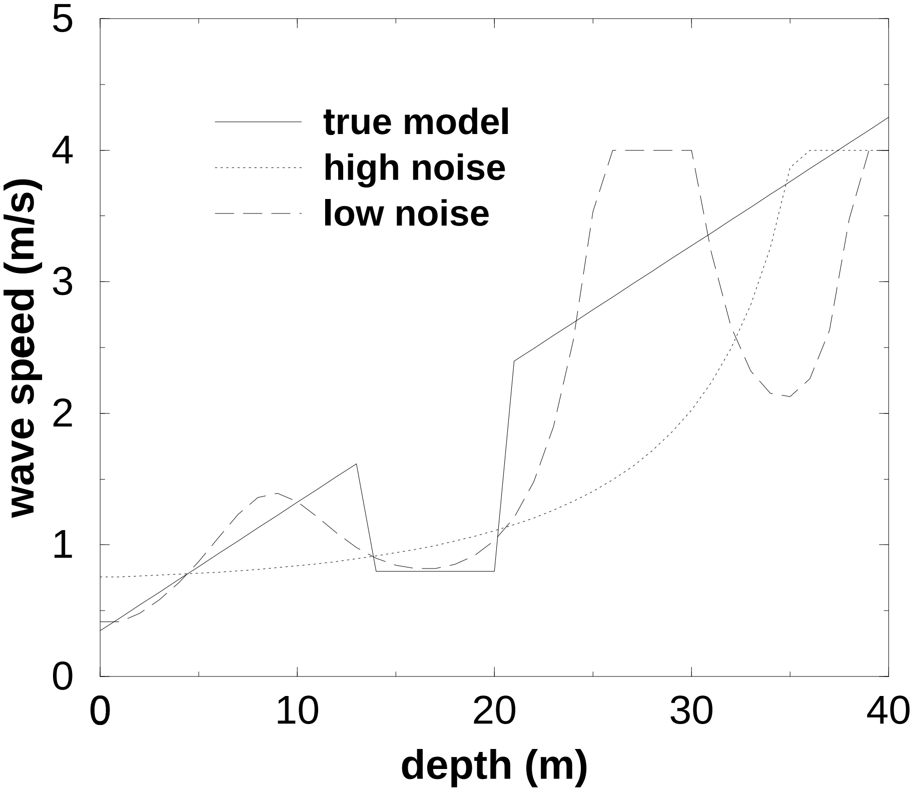

Figure 4: The true model (solid curve) and the models obtained by a truncated SVD expansion for the two levels of noise, optimistic (0.3 ms, dashed curve) and pessimistic (1.0 ms, dotted curve). Both of these models *just* fit the data in the sense that we eliminate as many singular vectors as possible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

where $Y_{\ell,m}$ is the $m-$th spherical harmonic of degree ℓ. Any element of the model space can be identified with its sequence of harmonic coefficients $\mathbf{a} = \{a_{\ell,m}\}$. If the CMB were uniform then $a_{\ell,m} = 0$ for $\ell > 0$. To determine the existence of small variations in the CMB is equivalent to finding non-zero harmonic coefficients for $\ell > 0$.

We now formulate the estimation of CMB harmonic coefficients as an inverse problem. Similar questions arise in geomagnetism when trying to estimate the Gauss coefficients of the Earth's magnetic field, or in the estimation of harmonic coefficients of the core mantle boundary topography (Stark (1992)). We model CMB data as

$$\mathbf{T} = \mathbf{K}\mathbf{a} + \mathbf{z}, \tag{9}$$

where $\mathbf{z}$ is a vector of uncorrelated Gaussian noise and $\mathbf{K}$ is the data mapping representing, for example, the radiometer beam smoothing. $\mathbf{K}$ is a linear operator from the space of square-integrable sequences to R^n, our measurement space. Because of kinematic effects the cosmologically interesting coefficients are those of degree $\ell \geq 2$. An interesting question is thus to estimate the five coefficients of degree $\ell = 2$ (*quadrupole* coefficients). For simplicity, assume that we want to estimate a single coefficient. Since any coefficient is a linear functional L of the model $\mathbf{a}$, the question is then to estimate $L(\mathbf{a})$ given the data (9). It sounds simple enough but there are important difficulties in practice: We do not have data uniformly spread on the sphere, and even when we do, since Galactic emissions dominate the microwave radiation, we have to neglect data near the Galactic plane. As a consequence, there are infinitely many models $\mathbf{a}_o$ satisfying $\mathbf{K}(\mathbf{a}_o) = 0$, where $L(\mathbf{a}_o)$ can take arbitrarily large values. A similar problem arises in the estimation of harmonic coefficients of the core mantle boundary where the gaps are in the distribution of rays which reflect off this boundary.

Since the spherical harmonics are no longer orthonormal on the cut sky, a simple least-squares fit to a truncated expansion will yield misleading coefficient estimates. So, how can we estimate the coefficients? To start with, we can assume a mild bound on the CMB energy and use minimax methods similar to those described in Section 5 to answer the following question: Among all the possible estimators (of a quadrupole coefficient), what is the size of the smallest $1 - \alpha$ confidence interval that covers the quadrupole coefficient of any model satisfying the energy bound, given the Galactic cut, noise level, and geometry of the observations in the sky? In other words, what is the best we can do with the information we have. It turns out that even with the best available data (NASA's COBE DMR[2] data) the confidence intervals are almost 10 times larger than the minimum required to obtain cosmologically interesting conclusions (Tenorio et al. (1998)). Therefore we need more prior information.

We have been naturally lead to Question III in Section 1: what prior information can we use? So far we have considered the CMB temperature as fixed in our sky but cosmologists view our Universe as only a realization of an infinite number of plausible universes. Cosmologists think of the CMB temperature T as a random field on the sphere with the measured CMB being just a noisy version of one of its realizations. In the current most popular theories T is modeled as a homogeneous Gaussian random field: the $\{a_{\ell,m}\}$ are independent zero-mean Gaussian random variables whose variance depends only on ℓ, i.e. $\text{var}(a_{\ell,m}) = \sigma_\ell^2$.

Now we have a multivariate normal distribution for the data given the model $\mathbf{a}$, and a physically motivated multivariate normal prior on $\mathbf{a}$. Both distributions are centered at zero. If the σ_ℓ are known, then it is straightforward to compute the posterior mean and variance of a harmonic coefficient given the data (e.g., Lindley & Smith (1972)). Many frequentists would not object to such approach (Bayes' formula is, after all, a theorem in probability) unless, of course, they do not believe in the cosmological models accepted by most astrophysicists. The problem is that the σ_ℓ are unknown: they are determined by two unknown cosmological parameters which we denote by Q (quadrupole amplitude normalization) and η (spectral index). What can we do then? Here is where frequentists and Bayesians start to disagree. One approach, which has been followed by cosmologists, is to use previously estimated values of Q and η as if they were the true values. In this case the posterior quadrupole uncertainties may be too optimistic since they do not consider the uncertainty in the estimated values of Q, η. A good first step is to study the sensitivity of the posterior estimates to a "reasonable" range of values of Q and η. We have done this in Tenorio et al. (1998). But again we have a subjective choice of what reasonable means. We chose a region defined by a bound on the CMB energy and found that the variability over

[2] DMR stands for the differential microwave radiometer that was on board NASA's cosmic background explorer (COBE) satellite

this region is large enough that needs to be accounted for. We then used a prior joint distribution for Q, η. Unfortunately, the choice of this prior is no longer physically motivated.

In summary, the CMB example shows how prior information is included in the quadrupole estimation problem. We first realized that there was insufficient information to estimate quadrupole coefficients to the required degree of accuracy. We then added prior information based on physical theories and finally concluded that estimates to the required accuracy depend on somewhat subjective choices of prior information. As pointed out before, we could have just performed a least-squares (LS) fit to a spherical harmonic expansion truncated at $\ell = 2$, but the uncertainty of these estimates would have been far too optimistic. For example, Figure 5 shows estimates of $Q_{\rm rms} = \sqrt{(a_{2,-2}^2 + \ldots + a_{2,2}^2)/4\pi}$ using LS and posterior means of the $a_{\ell,m}$ for different truncations $\ell = \ell_{\rm max}$. The uncertainty of the LS estimate grows with $\ell_{\rm max}$. This does not happen with the posterior mean of $Q_{\rm rms}$ because it is constrained by the prior cosmological model.

We can use any method we want to obtain estimates of some unknown parameter, but we have to make sure that the uncertainty estimates we use are not artificial, as in with the quadrupole LS estimates. 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)).

References

Backus, G. 1988. Hard and soft prior bounds in geophysical inverse problems. *Geophysical Journal*, **94**, 249–261.

Bording, R. P., Gersztenkorn, A., Lines, L. R., Scales, J. A., & Treitel, S. 1987. Applications of seismic travel time tomography. *Geophysical Journal of the Royal Astronomical Society*, **90**, 285–303.

Box, G. E. P, & Tiao, G. C. 1973. *Bayesian Inference in Statistical Analysis.* Wiley.

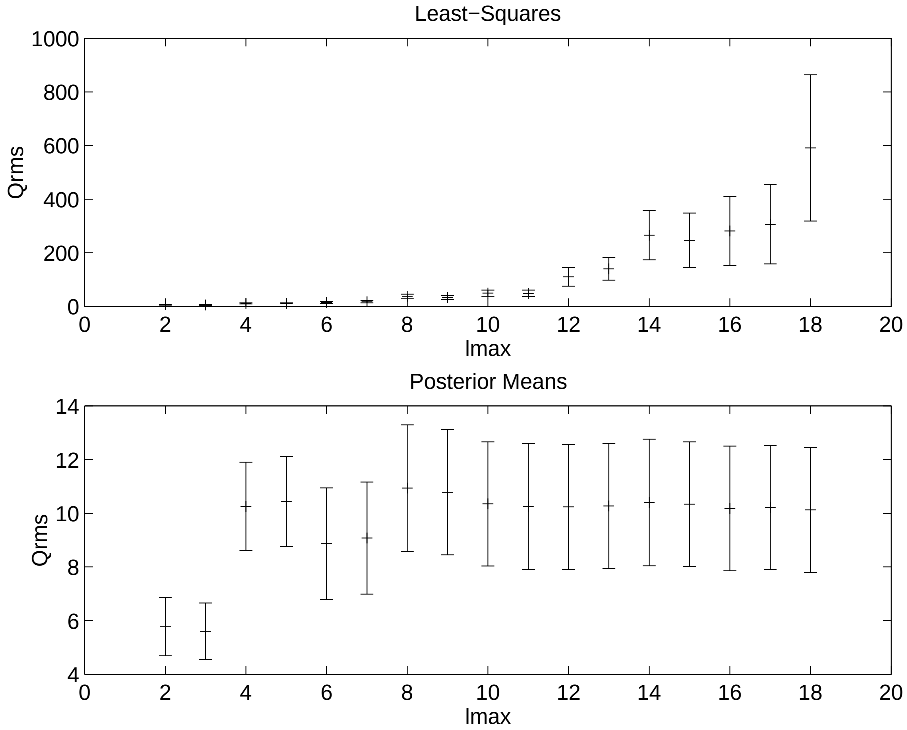

Figure 5: Least-squares and Bayesian estimates of Q_{rms} (μK) as a function of truncation order ℓ_{max}. The fits are done to the 53 GHz DMR sky map using a 20° Galactic cut. The top panel shows the Q_{rms} obtained from LS estimates of the a_{2m}. The lower plot shows estimates based on the posterior distribution of Q_{rms} with n and Q fixed to 1 and 15.3 μK, respectively.

Cox, D. R. 1981. Theory and general principles in statistics. *Journal of the Royal Statistical Society A*, **144**, 289–297.

Donoho, D. L., Liu, R. C., & MacGibbon, K. B. 1990. Mimimax risk over hyperrectangles, and implications. *Ann. Stat.*, **18**, 1416–1437.

Efron, B. 1986. Why isn't everyone a bayesian. *American Statistician*, **40**(1), 1–11.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 1997. *Bayesian Data Analysis*. Chapman & Hall.

Goutis, C., & Casella, G. 1995. Frequentist post-data inference. *International Statistical Review*, **63**(3), 325–344.

Gouveia, W., & Scales, J. A. 1997. Resolution in seismic waveform inversion: Bayes vs Occam. *Inverse Problems*, **13**, 323–349.

Gouveia, W., & Scales, J. A. 1998. Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis. *Journal of Geophysical Research*, **103**, 2759–2779.

Kass, R., & L., Wasserman. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1342–1370.

Lehmann, E. 1983. *Theory of point estimation.* Wiley.

Lindley, D. V. 1975. The future of statistics–A Bayesian 21st century. *In: Proceedings of the Conference on Directions for Mathematical Statistics* .

Lindley, D. V., & Smith, A. F. M. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, **1**, 1–18.

Rubin, D. R. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**(4), 1151–1172.

Scales, J. A., & Snieder, R. 1997. To Bayes or not to Bayes? *Geophysics*, **63**, 1045–1046.

Scales, J. A., Docherty, P., & Gersztenkorn, A. 1990. Regularization of nonlinear inverse problems: imaging the near-surface weathering layer. *Inverse Problems*, **6**, 115–131.

Silk, J. 1997. *A Short History of the Universe.* Scientific American Library.

Smoot, G. F., *et al.* 1992. Structure in the COBE DMR First Year Maps. *Astrophysical Journal*, **396**, L1–L5.

Stark, P. B. 1992. Minimax confidence intervals in geomagnetism. *Geophysical Journal International*, **108**, 329–338.

Stark, P. B. 1997. *Does God play dice with the Earth? (And if so, are they loaded?).* http://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.

Tenorio, L., Stark, P. B., & Lineweaver, C. H. 1998. Bigger uncertainties and the Big Bang. *Submitted.*