Center for Wave Phenomena 159 Printed 21 September 1994

An Example of Geologic Prior Information in a Bayesian Seismic Inverse Calculation

John A. Scales

Center for Wave Phenomena, Colorado School of Mines

Albert Tarantola

Institut de Physique du Globe de Paris, Université de Paris 6 et 7

ABSTRACT

All inverse problems require the specification of *a priori* information on the space of models. Typically, it is assumed that the *a priori* uncertainties can be described using Gaussian probabilities, whether explicitly through the use of some *a priori* covariances, or implicitly, with "regularized least squares". Our goal is to try to get beyond these assumptions and let realistic prior information speak for itself. To that end we show here a simple case study in the use of geologic information for seismic inversion. We estimate nonparametrically the Bayesian *a priori* distribution for layered earth models directly from a P-wave sonic log and give a numerical procedure for randomly sampling earth models following this distribution. This procedure is then modified in order to obtain earth models that sample the *a posteriori* distribution, which we regard as a complete solution of the inverse problem. Although the scope of this calculation is limited, the methods employed have a broad utility.

Key words: *a priori* information, inverse problems, Bayesian inference

1 INTRODUCTION

All inverse calculations must balance the extent to which models fit data with the extent to which they conform to our prejudices as to what makes a model good or bad. In some cases it is possible to find completely unrealistic models which nevertheless fit the data. The *a priori* information we use to judge the reasonableness of models comes in many forms, from the subjective wisdom of experts to quantitative data. By incorporating this prior

information into Bayesian probability distributions, we can tackle, in principle, completely general kinds of inverse problems without relying on the particular form of the distribution—as, for example, least squares makes Gaussian assumptions. In this paper we give a concrete example of the nonparametric estimation of a Bayesian prior for earth models derived from a P-wave sonic log. The ultimate goal is to do the seismic inverse problem for surface seismic data and, using the information derived from the sonic log, to constrain, in a Bayesian sense, the inferences we draw from the seismic data. We end up with an algorithm for sampling the *a priori* distribution associated with the well log. To solve the full inverse problem, i.e., to sample the Bayesian *a posteriori* distribution we must, in effect, select the models sampled from the prior distribution by how well they fit the data. We use the Metropolis-like procedure proposed by Mosegaard & Tarantola (1994). This procedure is guaranteed to converge to the posterior distribution.

1.1 Probabilistic Formulation of the Inverse Problem

Assume we are analyzing some physical system (in geophysics this is usually the Earth) that is described by some parameters. Any particular value of the parameters describing the system defines a *model* of the system. A generic model is represented by the symbol $\mathbf{m}$. The collection of all possible models defines the *model space.*

Probabilistic formulations of inverse theory consider probability densities in the model space. Two probability distributions are fundamental ingredients to the theory.

One probability distribution describes the *a priori* information we may have on the parameters describing our model. Here, *a priori* means information that is independent of the data set to be used to refine this information. This *a priori* probability density is denoted $\rho(\mathbf{m})$. The purpose of this paper is to give an example of its definition.

If an experiment produces some data, then a "likelihood function" has to be introduced (Tarantola, 1987) that measures the "degree of fit" between the data predicted by the model (using some physical theory) and the actually observed data. This likelihood function is denoted $L(\mathbf{m})$ and the most popular example is

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{p}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\text{obs}}|^p}{\sigma_i^p}\right) , \qquad (1)$$

where, if d_i^{obs} represents the *observed value* for the i-th datum, $g_i(\mathbf{m})$ represents the cal-

culated value corresponding to the model $\mathbf{m}$, and σ_i is an estimation of the uncertainty attached to the i-th observation. For $p = 2$ we have a Gaussian probability density (for independent uncertainties), while for $p = 1$ we have a Laplacian (double exponential) probability density. Equation (1) is only given as an example, and much more realistic expressions may be used that describe better experimental uncertainties, but this is outside the scope of this paper.

Once the probability density $\rho(\mathbf{m})$ and the likelihood function $L(\mathbf{m})$ have been defined, describing respectively the *a priori* information we have on model parameters and the likelihood of a model, the resulting *a posteriori* probability density is given by the expression (see Appendix A for a derivation)

$$\sigma(\mathbf{m}) = k\,\rho(\mathbf{m})\,L(\mathbf{m})\,, \tag{2}$$

where k is a normalization constant. The probability density $\sigma(\mathbf{m})$ combines the *a priori* information with the information coming from observations. The more the *a posteriori* uncertainties on model parameters [as described by $\sigma(\mathbf{m})$] have been reduced compared to the *a priori* uncertainties [as described by $\rho(\mathbf{m})$] the more successful the "inversion" of the data has been.

We can use the function $\sigma(\mathbf{m})$ to find out the probability that the true model is inside a particular region of the model space by integrating $\sigma(\mathbf{m})$ over the particular region. Neglecting the presence of $\rho(\mathbf{m})$ in Equation (2) and judging models only on how well they fit the data is called maximum likelihood estimation. On the other hand, neglecting $L(\mathbf{m})$ and sampling models only according to some prior distribution is akin to geostatistical simulation. The complete solution of the inverse problem combines these two features: sampling models according to a prior distribution and selecting them according to how well they fit the data.

With that brief introduction to the theory, we turn now to a simple example of the probabilistic approach to using prior information. For more details on the theoretical aspects of the problem see Tarantola (1987) and Mosegaard and Tarantola (1994). We will consider the problem of inverting reflection seismic data for subsurface elastic properties in an area where we have a single P-wave sonic log. We regard the sonic log as containing important information about the geology of earth models in the area even though it consists of in-situ measurements made on a much finer scale than the seismic wavelength. The question is, how

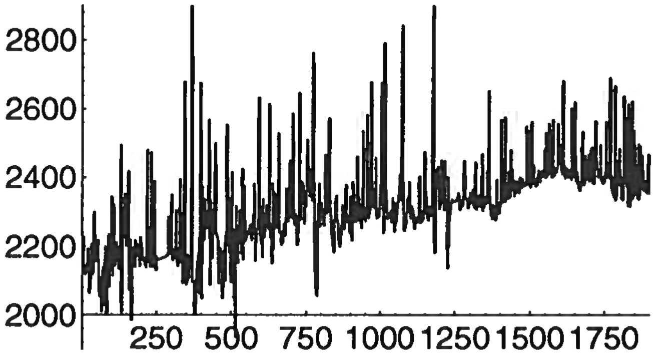

Figure 1. P-wave sonic log. The numbers on the abscissa refer to the samples, which were recorded every 30 cm. The wave speeds are in m/s.

to use this information? We will proceed by extracting the statistical properties from the log and then generating pseudo-random logs with these same statistical properties. By definition then, all of these pseudo-random models are *a priori* realistic in a geologic sense: this is how we sample the prior probability. Once this problem is solved, it is simple (although perhaps expensive) to use the likelihood function to sample the posterior probability.

2 CASE STUDY: EXTRACTING THE PRIOR FROM A SONIC LOG

Figure 1 shows an example of a P-wave sonic log. The wave speeds are in m/s and the samples were recorded every 30 cm. Our point of view is that the log represents a combination of gradual, deterministic processes and essentially random fluctuations. Thus if we subtract the smooth trend of the log (shown in Figure 2) from the original data, we will be left with an essentially random process, as shown in Figure 3. We will regard the trend of the log as being known exactly *a priori*. It could be asserted on geologic grounds or determined by some other geophysical technique such as travel time inversion. In any case, we will focus our attention on the fluctuating part of the log, which we regard as being a realization of a stationary stochastic process whose properties are to be determined. The assumption of stationarity can be relaxed, for example, by dividing the log into pieces associated with time or geologic horizons.

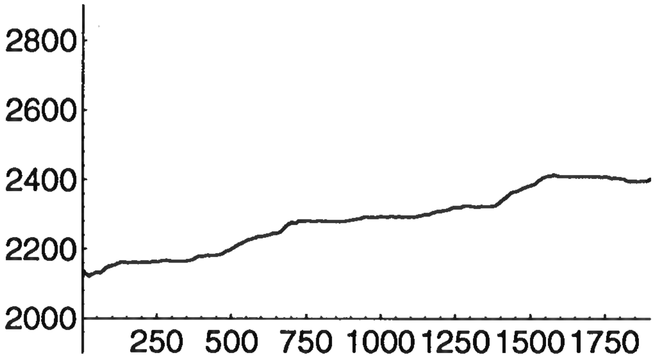

Figure 2. Trend of the log obtained with a sliding 150 sample alpha-trimmed mean filter.

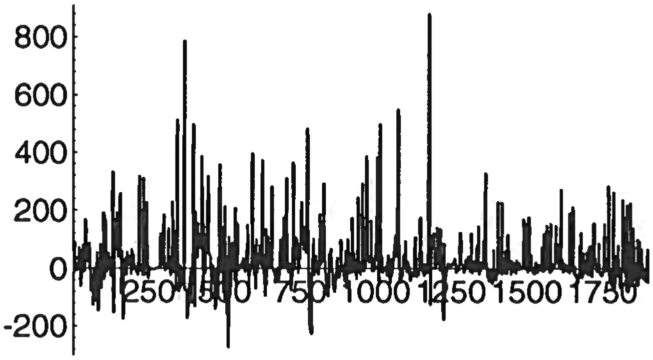

Figure 3. Fluctuating part of the log obtained by subtracting the trend from the original data.

2.1 Random Processes

Some basic principles of random functions are set forth in Appendix B; here we plan to keep the discussion as simple as possible. Consider a field $\xi(\mathbf{r})$ defined at every point $\mathbf{r}$ of the space. If what we have is, in fact, a *random field*, this means that at every point $\mathbf{r}$ of the space we have a random variable $\Xi(\mathbf{r})$. The notation

$$f_n(\xi_1, \xi_2, \ldots, \xi_m; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m)$$

represents the $n-$dimensional (joint) probability density for the random variables

$$\Xi(\mathbf{r}_1), \Xi(\mathbf{r}_2), \ldots, \Xi(\mathbf{r}_m).$$

To describe in all generality a random function requires, for any m, and for any points $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$, to give the $m-$dimensional joint probability density

$$f_m(\xi_1, \xi_2, \ldots, \xi_m; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m).$$

If the space has been discretized using, say n points, then we have only n random variables, and the most general case needs, at most an $n-$dimensional probability density.

Practically, we must rely on the fact that the $n-$dimensional distribution can be expressed in terms of the marginal distributions of the process. For example, the three dimensional distribution $f(x, y, z)$ can be written as $f(x|y, z)f(y, z)$, which can in turn be written as $f(x|yz)f(y|z)f(z)$ and so on by induction for the general n-dimensional distribution. For example, if a process is Markovian, this characterization in terms of marginals is especially simple since

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) = \frac{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) \dots f_2(\xi_{n-1}, \xi_n; \mathbf{r}_{n-1}, \mathbf{r}_n)}{f_1(\xi_2; \mathbf{r}_2)\ f_1(\xi_3; \mathbf{r}_3) \dots f_1(\xi_{n-1}; \mathbf{r}_{n-1})} \tag{3}$$

where f_1 and f_2 are, respectively the one- and two-dimensional marginal distributions associated with f_n. This brings out clearly the fact that for Markov fields the probability of a point depends only on the previous point.

2.2 Estimating the Marginal Distributions

For more general processes than Markovian ones, we need more of the marginal distributions, not just the one- and two-dimensional ones. But, in any case, the marginals can be estimated by making histograms of the data. The process of making these histograms begins by quantizing the log into some number of equal-length velocity intervals. Figure 4 shows how this is done.

Once the log is quantized, we just start counting. For example, to compute the two-dimensional marginals of length one (i.e., the probability density for the velocity values at two consecutive points) we count all the transitions from an interval i to an interval j which occur at adjacent sites along the log (note that we are using here the stationarity assumption). To compute all the two-dimensional distributions we must compute all transitions from an interval i to an interval j over non-adjacent sites as well. Figures 5 and 6 show second order histograms for the P-wave sonic log, the first for transitions of length 1, the second for transitions of length 5.

Clearly it would be impractical to attempt to compute all the marginal distributions for a process of several thousands of points. But at the same time we are not yet willing to restrict ourselves to parametric assumptions, such as Gaussian statistics, that would allow us to reduce the computational burden. Our compromise is to calculate enough of

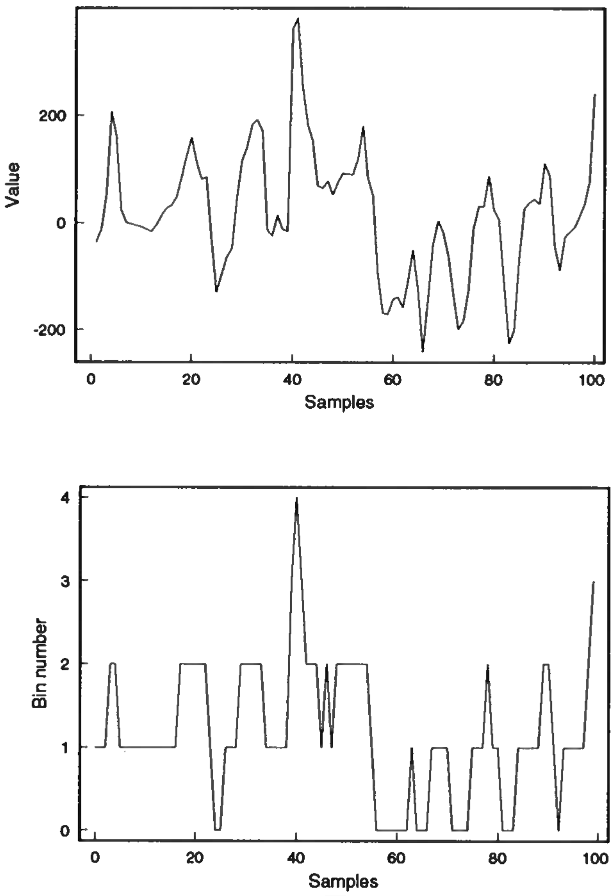

Figure 4. The data are quantized into a number of equal-length velocity intervals so that histograms can be made. Here we show the first 100 samples of the de-meaned sonic log and a toy quantization into 5 intervals. In practice we use a much larger number of intervals, from 100 to 500.

the marginals so as to convince ourselves that we have indeed captured the essence of the process. To quantify this, we use the Kolmogorov-Smirnov two-sample test and compare the original data with pseudo-random simulations calculated with a given number of marginals. The K-S test (described in Appendix C) involves comparing the cumulative distribution function of two different realizations (in this case the data and the simulation), the so-called null hypothesis being that the two were drawn from the same distribution. A high value of the K-S statistic says that the two realizations were very likely drawn from the same underlying distribution; this gives us confidence that the pseudo-random realizations have captured the essence of the data.

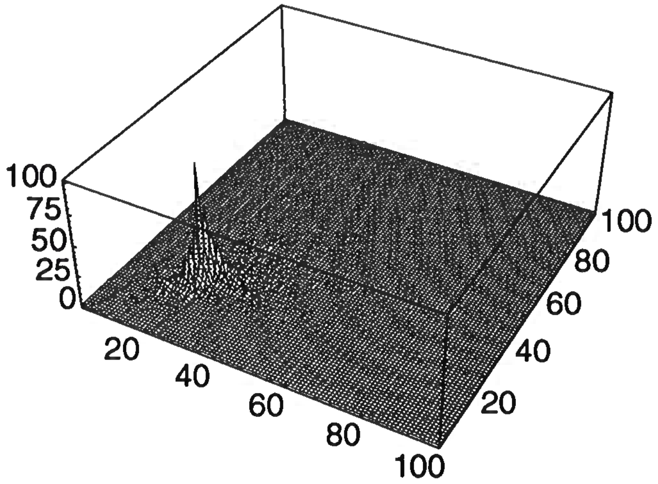

Figure 5. Histograms of two-point transitions of length 1 (i.e., pairwise transitions for adjacent sites along the log) for the P-wave sonic log. If for a given value of x and y axes, say (i, j), the count is n, that means that there are n sites along the log where the site itself and its neighbor have the quantized values i and j respectively.

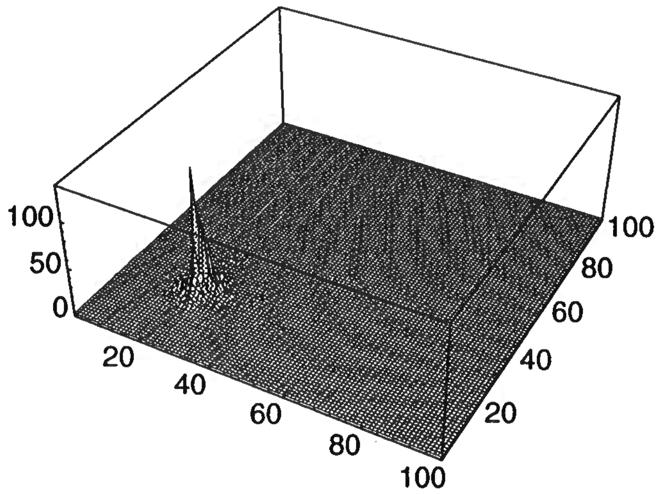

Figure 6. Histograms of two-point transitions of length 5 for the P-wave sonic log. These correspond to transitions from a site k to a site $k + 5$ on the log.

2.3 Sampling the Prior

Once we have computed a sufficient number of the marginals we must be able to sample the prior distribution. To *sample* a probability density means to produce a pseudo random "sample" (or *realization*) of the random variable. By definition, the probability for the produced "point" to belong to any volume of the space must equal the probability of the volume. The algorithms used to sample a probability produce, very often, consecutive samples that are

not independent (think of Brownian motion of a particle: if, in the long term, the particle may be at any point of the space with uniform probability, at close times the particle will be at close points). In such a situation, to obtain independent samples we simply have to "wait" until the algorithm loses the memory of the previous point.

Typically, a sampling algorithm produces a *test point* using some "background probability" (usually the uniform one) and then, uses some random criterion for "accepting" the test point or for "dropping" it. Obviously, if the background probability is very close to the probability we wish to sample, the test points will be often accepted. In this case, one says that the algorithm uses an "importance sampling method". Any nontrivial sampling algorithm, in fact, uses importance sampling, and this will be the case for the examples below.

To do our sampling we use the histograms directly. Suppose, for example, that we needed only the two-point histograms of length 1 (as would be the case if the underlying process were Markovian, since then the probability of a given point depends only on the probability of the previous point). Figure 7 shows a cartoon of this case. The values of the observed log are $\{\xi(\mathbf{r}_i)\}_{i=1}^{N}$. Let us refer to the values of a pseudo-random simulation as $\{\xi'(\mathbf{r}_i)\}_{i=1}^{N}$. We begin the construction of the simulation by assigning some value to $\xi'(\mathbf{r}_1)$, for example, this could be a likely value for the log based on the one-dimensional marginal. Now all we have to do is select $\xi'(\mathbf{r}_2)$ according to the histogram of length 1 transitions. Suppose $\xi'(\mathbf{r}_1) = 4$, as shown in Figure 7. Then the slice through the histogram associated with $i = 4$ is the conditional probability distribution of making the length 1 transition $4 \longrightarrow j$ for any j. To assign the value $\xi'(\mathbf{r}_2)$ in accordance with this conditional probability, we make, in effect, a weighted roulette wheel with a sector for each interval number. The size of each interval's sector is in proportion to its probability from the histogram. Then by spinning the roulette wheel we will select $\xi'(\mathbf{r}_2)$ with the appropriate probability. For Markov processes we would proceed in this way right down the log, filling in one value $\xi'(\mathbf{r}_i)$ after another.

But suppose our data are not Markovian? Suppose, for example, that the results of the K-S test say that we must include all two-point transitions out to some length ℓ. In this case we begin as before, filling out the first and second sites along the log. But now, to fill out the third site, we choose the roulette wheel associated with transitions of length two conditioned on the first point $\xi'(\mathbf{r}_1)$. Then we choose the roulette wheel associated with

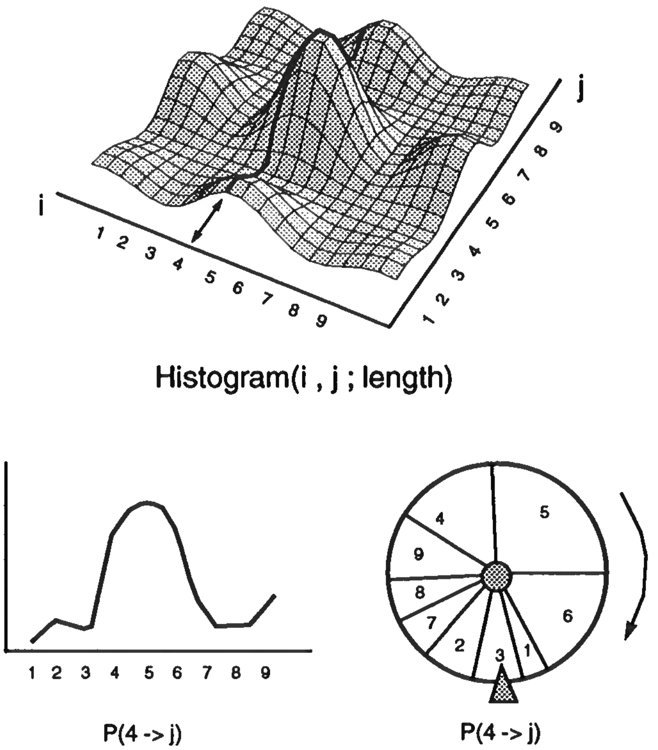

Figure 7. A weighted roulette wheel is used to sample the histograms. If the current site on the pseudo-random log has a value i and we need a length ℓ transition to fill in the next slot, we first select the length ℓ histogram. Then we extract the one-dimensional marginal associated with the value i. (In the example shown in the figure $i = 4$.) This gives us the observed probability of making a transition from a value i to all other values over a distance of ℓ sites.

length three transitions conditioned on $\xi'(\mathbf{r}_1)$, and so on until we have filled in the first $\ell+1$ sites on the log. Then we start over again at the site $\ell+2$. This procedure can be generalized to higher order histograms.

Figure 8 shows 15 such logs pseudo-randomly sampled from the second order histograms of the data in Figure 1. In this case we used all the histograms out to a length of 5 samples, which is estimated to be the average correlation length in the medium. Using histograms of this length we are able to routinely generate pseudo-random logs with a K-S statistic of better than 95%. In other words, the distributions associated with the two realizations, pseudo-random and data, are so close that K-S cannot tell whether they are the same or not. But it is important to emphasize, that the K-S test is used only as a "preprocessing" step. Once we have decided on the number of marginals to use, we sample them as described in the text. That means that in the course of the sampling, we may generate simulations with low *a priori* probability but these will be relatively rare occurrences.

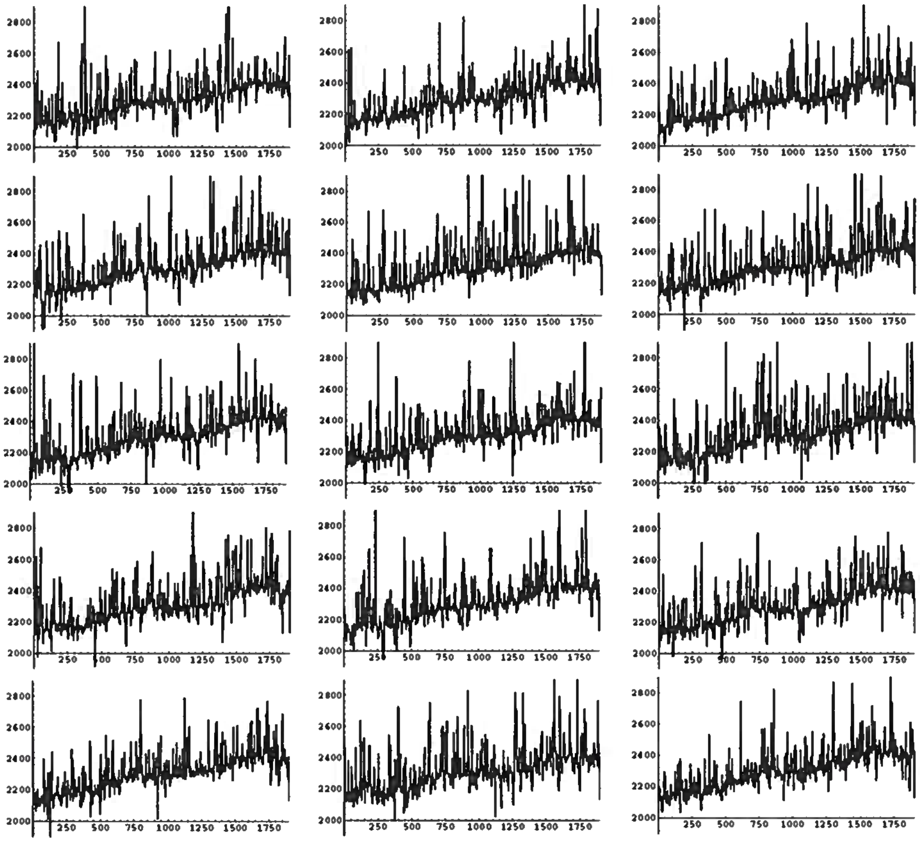

Figure 8. Pseudorandom logs obtained by sampling the *a priori* model distribution. Here we used all the second order histograms up to the average correlation length of the medium. Each of these pseudo-random logs is *a priori* feasible geologically. We have only to rank the models according to how well they fit the seismic data. We can imagine that these 15 models are 15 frames from a movie tour of the prior.

3 SOLVING THE INVERSE PROBLEM

Now that we have a procedure for sampling the prior, we can use these models to sample the posterior according to the algorithm of Mosegaard and Tarantola (1994). This involves accepting or rejecting models according to how well they fit the data, and represents a generalization of the usual Metropolis method for sampling the Gibbs-Boltzman distribution.

More precisely, let us assume that we are able to obtain samples $\mathbf{m}_1, \mathbf{m}_2, \ldots$, of the prior probability distribution $\rho(\mathbf{m})$. Then, if we wish to obtain samples of the posterior probability distribution $\sigma(\mathbf{m}) = k\,\rho(\mathbf{m})\,L(\mathbf{m})$ all that we need to do is to iterate the following procedure.

- Let $\mathbf{m}_i$ be the "current model."
- Use the rules that allow one to sample the prior probability distribution $\rho(\mathbf{m})$ to obtain a new model, say $\mathbf{m}'_i$.

- If $L(\mathbf{m}_i') \geq L(\mathbf{m}_i)$, take $\mathbf{m}_i'$ as new current model, i.e., make $\mathbf{m}_{i+1} = \mathbf{m}_i'$.
- If $L(\mathbf{m}_i') < L(\mathbf{m}_i)$, then decide randomly to take the model $\mathbf{m}_i'$ as new current model, or to destroy it, with a probability of taking it as new current model equal to the ratio $L(\mathbf{m}_i')/L(\mathbf{m}_i)$.

Mosegaard and Tarantola (1994) show that this procedure converges to the true *a posteriori* distribution.

The precise specification of L requires knowledge of the data and modeling uncertainties, as shown in Appendix A. We cannot say what it means to fit the data if we do not know the data uncertainites. This is obviously a crucial issue when inverting real data. But it is beyond the scope of this paper. So although we did not have to make a parametric assumption about the *a priori* distribution, we must make some choice for the data distribution. For this we make the simplest choice, gaussian, so that we can use a least squares criterion to measure data misfit. In other words, for purposes of this demonstration, we will generate synthetic data from the actual log and add a small amount of uncorrelated gaussian noise. Figure 9 shows the synthetic "data", a single zero-offset reflection seismogram, associated with the true model (top) and the same trace contaminated with a small amount (5%) uncorrelated gaussian noise. The seismograms are computed by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson (1967) and includes all multiples, but is limited to zero-offset, acoustic, marine data. Once we have the data trace, the likelihood function is the difference between this trace and the response of the sampled model, normalized by the known data variance

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\mathrm{obs}}|^2}{\sigma^2}\right) \tag{4}$$

where $g_i(\mathbf{m})$ are the synthetic data associated with a sampled model $\mathbf{m}$, d_i^{obs} are the "observed" data, and σ is the standard deviation of the errors.

3.1 Sampling the Posterior

Figure 10 shows a selection of models sampled according to the generalized Metropolis procedure of Mosegaard and Tarantola (1994). We can think of these models as 15 frames from a movie tour of the posterior. But how do these samples of the posterior actually "solve" an/the inverse problem? It is up to us to pose questions and use the posterior to

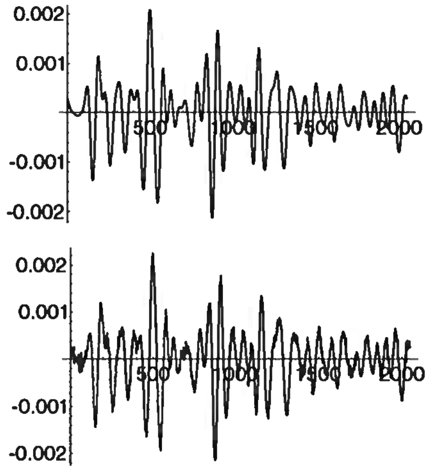

Figure 9. The top figure shows the "data" (i.e., model response) for the true log. The seismogram is obtained by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson's book (1967) and includes all multiples in a fixed time window, but is limited to zero-offset, acoustic, marine data. Below the noise free seismogram is the seismogram with 5% uncorrelated gaussian noise added. We take this to be the observed data for the inverse problem.

answer them probabilistically. A simple illustration of this is to investigate the extent to which a certain feature of our earth models is resolved by the seismic data. Figure 11 shows the distribution of P-wave reflectivity at two points along the log, one near the top of the log ($z = 10$, i.e., the tenth sample along the log) and one near the bottom ($z = 1800$). In both cases the posterior variance is visibly reduced relative to the prior variance. This is a quantitative measure of the extent to which the seismic data are able to resolve these parameters.

4 CAVEATS

There are a great many limitations to this work. Here we list some of the more obvious ones with evidence both for the prosecution (**con**) and the defense (**pro**).

con: The sample size is too small and therefore the *a posteriori* statistics unreliable.

pro: True. The calculations were done on a slow workstation. By moving to a faster workstation and optimizing the code, a factor of 10 speedup is easily obtainable. To move to 2-D, however, will require significantly greater computing resources.

con: The noise model is unrealistic.

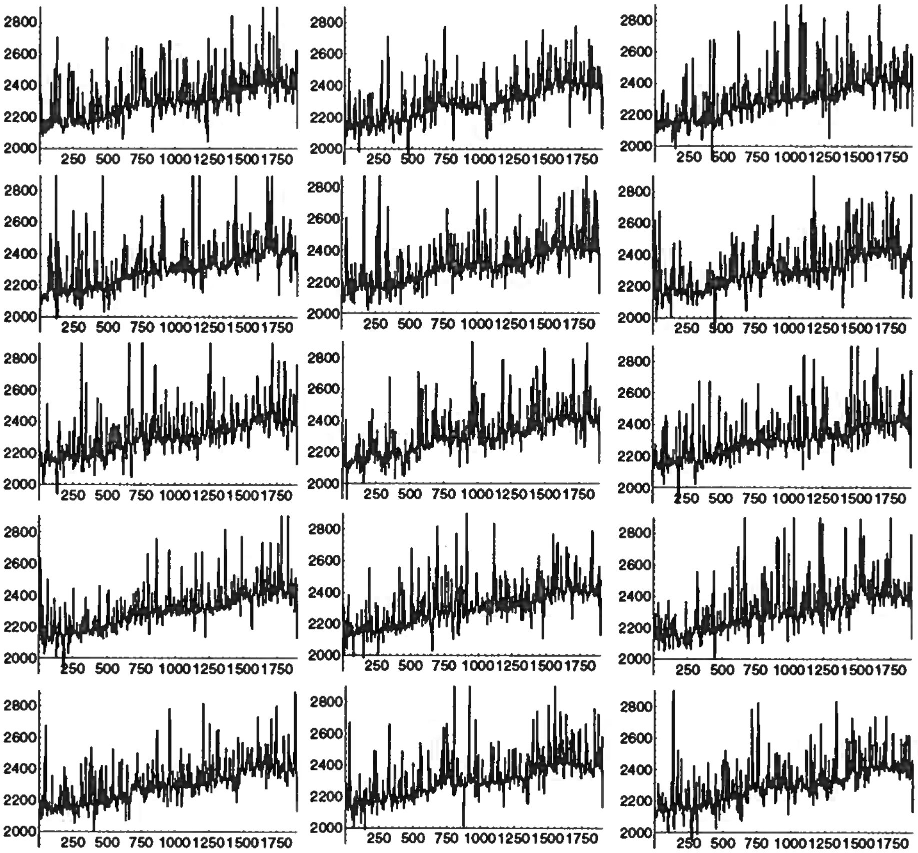

Figure 10. 1D earth models obtained by sampling the *a posteriori* distribution according to the Metropolis-like algorithm of Mosegaard and Tarantola (1994).

pro: True. Until we can estimate the noise directly from the data this will be a problem. In defense of our small *a priori* data error bars, we believe that the "noise" in exploration seismic data which can usefully be classified as gaussian (or some similar distribution) is very small. Far larger are the uncertainties due to modeling errors and, perhaps, coherent environmental noise, for which we have no realistic model.

con: None of the models in the posterior sample would generate seismograms that fit within the (tiny) error bars assigned to the data.

pro: True. This is a function of the small sample size. Should we allow the algorithm to run long enough, it will end by sampling a point inside this region and would remain there a long time, but this might take a prodigious number of iterations. On the other hand, the fact that the Monte Carlo scheme converges to the *a posteriori* distribution means that we can nevertheless compare the relative probability of different models. If the posterior variance of a parameter is smaller than the prior variance, then new information has been obtained,

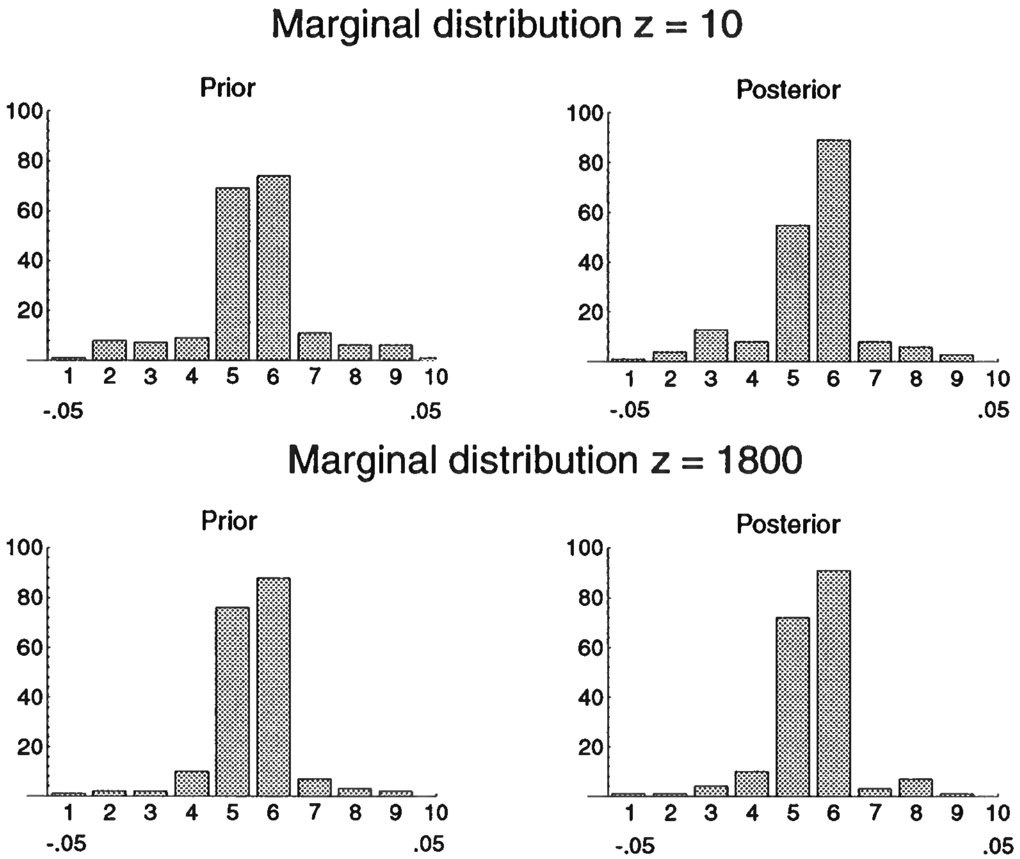

Figure 11. Prior versus posterior distribution of P-wave reflection coefficient at two points in the model. The z values shown refer to the sample number along the log. So these points are at extreme ends of the log. Near the upper surface, the posterior variance is noticeably smaller, indicating an increase in resolution. At the bottom of the log, however, the prior and posterior variances are virtually the same. So the seismic data have not resolved the deeper reflection coefficient at all.

even if the response of the posterior samples do not fit within the *a priori* error bars of the data.

A good simulation should answer positively to the following three questions:

(i) Have we chosen correctly the type of perturbations (of the current model)?

(ii) Is the size of the perturbations adapted to the size of the significant region of the model space?

(iii) Have we started at a point close enough to the region of significant probability so that we will enter it in a reasonable time?

At present, there remains a lot of work to be done in order to answer these questions rigorously.

5 CONCLUSIONS AND FUTURE WORK

We have shown a case study of the use of complex *a priori* information in a Bayesian inverse calculation. The problem we have examined is the inversion of reflection seismograms for 1D earth structure given a well log as *a priori* geologic information. We have developed a non-parametric technique for extracting the Bayesian *a priori* distribution from logs (or other similar data), as well as a method for sampling this *a priori* distribution. This algorithm allows us to generate as many pseudo-random earth models as we like, in accordance with the underlying *a priori* distribution. This sampling of the prior is then coupled with a Metropolis-like procedure to produce a sampling of the *a posteriori* distribution. Once we have this *a posteriori* distribution, which we regard as the ultimate solution of the inverse problem, it is up to us to pose proper questions and use the *a posteriori* distribution to answer them.

6 ACKNOWLEDGEMENTS

We thank Christophe Barnes, Jack Cohen, Philippe Doyen, and Martin Smith for useful discussions. The authors gratefully acknowledge the partial support of the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, Colorado School of Mines, the Shell Foundation, and the sponsors of the Geophysical Tomography Group, Institut de Physique du Globe de Paris, Université de Paris 6 et 7.

References

Duijndam, A. J. W. 1987. *Detailed Bayesian inversion of seismic data.* Ph.D. thesis, Technical University of Delft.

Jeffreys, R.C. 1983. *The logic of decision.* University of Chicago.

Mosegaard, K., & Tarantola, A. 1994. Monte Carlo sampling of solutions to inverse problems. *JGR: submitted.*

Press, William H., Flannery, Brian P., Teukolsky, Saul A., & Vetterling, William T. 1986. *Numerical recipes.* Cambridge University Press.

Pugachev, V. S. 1965. *Theory of random functions and its application to control problems.* Pergamon.

Robinson, E. A. 1967. *Multichannel Time Series Analysis with Digital Computer Programs.* Holden-Day.

Tarantola, A. 1987. *Inverse problem theory.* Elsevier.

APPENDIX A: THE BAYESIAN POSTERIOR PROBABILITY

As we have argued, the posterior probability density on the space of models must be the product of two terms: a term which involves the a priori probability on the space of models and a term which measures the extent of data fit

$$\sigma(\mathbf{d}) = k\rho(\mathbf{m})\ L(\mathbf{m}). \tag{A1}$$

L is called the likelihood function and depends implicitly on the data.

We now show how Equation (A1) follows logically from Bayes' theorem provided we generalize our notion of "data" to allow for the possibility that the data might be specified by probability distributions (Tarantola, 1987). To do so we make use of an idea due to Jeffreys (1983) (as described in (Duijndam, 1987)). We begin by using the notation common amongst Bayesians, then we show how this relates to the more standard inverse-theoretic notation in Tarantola (1987).

In this approach we assume that we have some prior joint distribution $p_0(\mathbf{m}, \mathbf{d})$. Further, we suppose that as the result of some observation, the marginal pdf of $\mathbf{d}$ changes to $p_1(\mathbf{d})$. We regard $p_1(\mathbf{d})$ as being the "data" in the sense that we often know the data only as a distribution, not exact numbers. In the special case where the data are exactly known, $p_1(\mathbf{d})$ reduces to a delta function $\delta(\mathbf{d} - \mathbf{d}_{obs})$.

How do we use this new information in the solution of the inverse problem? The answer is based upon the following assumption: whereas the information on $\mathbf{d}$ has changed as a result of the experiment, there is no reason to think that the conditional degree of belief of $\mathbf{m}$ on $\mathbf{d}$ has. I.e.,

$$p_1(\mathbf{m}|\mathbf{d}) = p_0(\mathbf{m}|\mathbf{d}). \tag{A2}$$

From this one can derive the posterior marginal $p_1(\mathbf{m})$:

$$
\begin{aligned}
p_1(\mathbf{m}) &\equiv \int_D p_1(\mathbf{m},\mathbf{d})\, d\mathbf{d} && \text{(A3)}\\
&= \int_D p_1(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} && \text{(A4)}\\
&= \int_D p_0(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} && \text{(A5)}\\
&= \int_D \frac{p_0(\mathbf{d}|\mathbf{m}) p_0(\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d} && \text{(A6)}\\
&= p_0(\mathbf{m}) \int_D \frac{p_0(\mathbf{d}|\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d}, && \text{(A7)}
\end{aligned}
$$

where D denotes the data space.

Switching now to the inverse-theoretic notation, let us regard $p_0(\mathbf{d})$ as being the non-informative prior distribution on data $\mu_D(\mathbf{d})$: this is what we know about the data **before** we've actually done this particular experiment. Further, we identify $p_1(\mathbf{m})$ as the posterior distribution on the space of models, $p_1(\mathbf{d})$ as the data errors, $p_0(\mathbf{m})$ as the prior distribution on the space of models, and $p_0(\mathbf{d}|\mathbf{m})$ as the modeling errors:

$$p_1(\mathbf{m}) \equiv \sigma(\mathbf{m})$$

$$p_1(\mathbf{d}) \equiv \rho_D(\mathbf{d})$$

$$p_0(\mathbf{m}) \equiv \rho_M(\mathbf{m})$$

$$p_0(\mathbf{d}|\mathbf{m}) \equiv \Theta(\mathbf{d}|\mathbf{m}),$$

then we arrive at precisely Equation (1.65) of Tarantola (1987)

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \int_D \frac{\rho_D(\mathbf{d})\Theta(\mathbf{d}|\mathbf{m})}{\mu_D(\mathbf{d})}\, d\mathbf{d} \tag{A8}$$

An important special case occurs when the modeling errors are negligible, i.e., we have a perfect theory. Then the conditional distribution $\Theta(\mathbf{d}|\mathbf{m})$ reduces to a delta function $\delta(\mathbf{d} - g(\mathbf{m}))$ where g is the forward operator. In this case, the posterior is simply

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \left[\frac{\rho_D(\mathbf{d})}{\mu_D(\mathbf{d})}\right]_{\mathbf{d}=g(\mathbf{m})} . \tag{A9}$$

APPENDIX B: ELEMENTS OF RANDOM FIELDS

Here we set forth the basic notations having to do with random functions. This section is an adaptation of Pugachev (1965). We will consider random fields in (3-D) physical space. Adaptation to other sorts of fields is straightforward.

$\Xi(\mathbf{r})$ will denote a random field. A realization of the random field will be denoted by $\xi(\mathbf{r})$. We may think of ξ as a physical parameter defined at every point of the space (like the velocity of seismic waves, the temperature, etc.). We will work inside a volume $\mathcal{V}$.

B1 n-dimensional joint probability densities

Let $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ be a set of n points inside $\mathcal{V}$. An expression like

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$$

will denote the n-dimensional (joint) probability density for the values $t(\mathbf{r}_1), t(\mathbf{r}_2), \ldots, t(\mathbf{r}_n)$. The notation $f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ may seem complicated, but it just indicates that for every different set of points $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ we have a possibly different probability density.

The random field $T(\mathbf{r})$ is completely characterized if, for any set of n points inside $\mathcal{V}$, the joint probability density $f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ is defined, and this for any value of n.

B2 Marginal and conditional probability

The definitions for marginal and conditional volumetric probabilities do not pose any special difficulty. As notations rapidly become intricate, let us only give the corresponding definitions for some particular cases, the generalization being straightforward.

If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *marginal* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$ is defined by

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) \,, \tag{A10}$$

where $dL(\xi_3)$ is the (1-D) volume element (it may be $d\xi_3$ or something different).

Let us now turn now to the illustration of the definition of conditional probability. If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random

field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *conditional* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$, given that the random field takes the value ξ_3 at point $\mathbf{r}_3$, is defined by

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{\int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)} , \tag{A11}$$

i.e., using equation A10,

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)} . \tag{A12}$$

Of particular interest will be the one- and the two-dimensional probability densities, denoted respectively $f_1(t; \mathbf{r})$ and $f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)$.

B3 Random fields defined by low order probability densities

First example: Independently distributed variables.

If

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$g(\xi_1, \mathbf{r}_1)\; h(\xi_2, \mathbf{r}_2) \ldots i(\xi_n, \mathbf{r}_n) , \tag{A13}$$

we say that the random field has independently distributed variables.

Second example: Markov random field.

For a Markov process,

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_2(\xi_n; \mathbf{r}_n | \xi_{n-1}; \mathbf{r}_{n-1})\; f_2(\xi_{n-1}; \mathbf{r}_{n-1} | \xi_{n-2}; \mathbf{r}_{n-2}) \tag{A14}$$

$$\ldots f_2(\xi_2; \mathbf{r}_2 | \xi_1; \mathbf{r}_1)\; f_1(\xi_1; \mathbf{r}_1) ,$$

where the vertical bar denotes conditional probability. This means that the value at a given point depends only on the value at the previous point. As

$$f_2(\xi_i; \mathbf{r}_i | \xi_{i-1}; \mathbf{r}_{i-1}) = \frac{f_2(\xi_{i-1}, \xi_i; \mathbf{r}_{i-1}, \mathbf{r}_i)}{f_1(\xi_{i-1}; \mathbf{r}_{i-1})} , \tag{A15}$$

we obtain

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = \tag{A16}$$

$$\frac{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) \ldots f_2(\xi_{n-1}, \xi_n; \mathbf{r}_{n-1}, \mathbf{r}_n)}{f_1(\xi_2; \mathbf{r}_2)\; f_1(\xi_3; \mathbf{r}_3) \ldots f_1(\xi_{n-1}; \mathbf{r}_{n-1})} .$$

This equation characterizes a Markov random field in all generality. It means that the random

field is completely characterized by 1-D and 2-D probability densities (defined at adjacent points).

Third example: Gaussian random field.

For a Gaussian random field, if we know the 2-D distributions, we know all the means and all the covariances, so we also know the n-dimensional distribution. It can be shown that a Gaussian process with exponential covariance is Markovian.

B4 Uniform random fields

A random field is uniform (i.e., stationary) in the strong sense if for any $\mathbf{r}_0$,

$$f_n(\xi_1,\xi_2,\ldots,\xi_n;\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n) =$$

$$f_n(\xi_1,\xi_2,\ldots,\xi_n;\mathbf{r}_0+\mathbf{r}_1,\mathbf{r}_0+\mathbf{r}_2,\ldots,\mathbf{r}_0+\mathbf{r}_n) \ .$$

Taking

$$\mathbf{r}_0 = -\mathbf{r}_1$$

gives then

$$f_n(\xi_1,\xi_2,\ldots,\xi_n;\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n) =$$

$$f_n(\xi_1,\xi_2,\ldots,\xi_n;\mathbf{0},\mathbf{r}_2-\mathbf{r}_1,\ldots,\mathbf{r}_n-\mathbf{r}_1) \ .$$

A distribution is said to be uniform in the weak sense if this expression holds only for $n=1$ and $n=2$.

As the random field is defined over the physical space, we prefer the term *uniform* to characterize what, in random fields defined over a time variable, is called *stationary*. This is entirely a question of nomenclature; we regard the terms as being interchangeable.

Example:

For the two-dimensional distribution,

$$f_2(\xi_1,\xi_2;\mathbf{r}_1,\mathbf{r}_2) = \Psi_2(\xi_1,\xi_2;\Delta\mathbf{r}) \ ,$$

with

$$\Delta\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1 \ .$$

Example:

For the three-dimensional distribution,

$$f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \Psi_3(\xi_1, \xi_2, \xi_3; \Delta\mathbf{r}_1, \Delta\mathbf{r}_2) \ .$$

with

$$\Delta\mathbf{r}_1 = \mathbf{r}_2 - \mathbf{r}_1$$

and

$$\Delta\mathbf{r}_2 = \mathbf{r}_3 - \mathbf{r}_1 \ .$$

Essentially a uniform random process is one whose properties do not change with space (or time if that is the independent variable).

APPENDIX C: THE KOLMOGOROV-SMIRNOV TEST

This brief discussion is taken directly from *Numerical Recipes* (1986). The two-sample Kolmogorov-Smirnov statistic tests the null hypothesis that two data sets are drawn from the same distribution. It is based on a comparison of the cumulative distribution functions (CDF) of the two data sets. One can imagine any number of comparisons between the two CDFs. K-S represents an especially simple one: it is defined as the maximum value of the absolute difference between the two CDFs. In symbols, the K-S statistic D is given by

$$D = \max_{-\infty < x < \infty} |S_{N_1}(x) - S_{N_2}(x)|$$

where $S_{N_1}(x)$ and $S_{N_2}(x)$ are the approximate CDFs for the two data sets. The key point, however, is that the distribution function for the K-S statistic itself (for the null-hypothesis that the data sets are drawn from the same distribution) can be calulated approximately. The significance function Q for this test is given by the following approximation:

$$Q_{K-S}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1} e^{-2j^2\lambda^2} .$$

A plot of this significance function is given in Figure A1.

For the two-sample test, the significance level for an observed value of D (as a disproof of the null-hypothesis) is given approximately by

$$\text{Probability}(D > \text{observed}) = Q_{K-S}\left(\sqrt{\frac{N_1 N_2}{N_1 + N_2}} D\right)$$

where N_2 and N_2 are the numbers of samples in the two data sets.

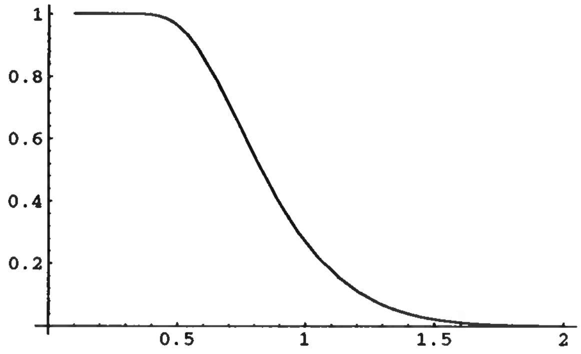

Figure A1. Kolmogorov-Smirnov significance function.

October 27, 1994

MEMO

TO: Consortium Sponsors, Inverse Methods for Complex Structures

FROM: Jo Ann Fink, Project Assistant JF

SUBJECT: CWP Research Reports

Enclosed are complimentary reprints of two CWP research reports that were recently published in GEOPHYSICS:

- CWP-114P, Deng, Lydia H., Acoustic-wave propagation in thin-layered media with steep reflectors.
- CWP-133P, Alkhalifah, T., and Larner, K., Migration error in transversely isotropic media.

Also enclosed is a complimentary copy of a paper co-written by John Scales and Albert Tarantola (Institut de Physique du Globe de Paris): An example of geologic prior information in a Bayesian Seismic inverse calculation (CWP-159).

In addition to the enclosed papers, we wish to inform you of the availability of a book co-authored by John Scales and Martin L. Smith (New England Research), *Introductory Geophysical Inverse Theory* (CWP-160). This book is available electronically via our anonymous ftp site 138.67.12.63. You may purchase a hard copy of the book by contacting the CWP office.

JF/bm

Enclosures (3)

CWP-159

From: Jo Ann Fink <jfink>
Date: Mon, 26 Sep 94 17:27:38 -0600
To: barbara
Subject: Re: John Scales & Tarantola's paper

CWP-159 distribution of one copy corner stapled with explanation that it is work he did with Tarantola and is being sent to a journal should do it. (1 copy for tech and cc list)

Sent with other reprints soon to arrive is what we agreed.

NOT YET SENT TO JOURNAL. 9.26.94

From: John Scales <jscales>
Date: Tue, 27 Sep 94 11:11:52 -0600
To: barbara
In-Reply-To: An example of geol... [from: Barbara McLenon]
Subject: An example of geol...

----- From: Barbara McLenon -----
| Have you already sent your paper with Tarantola to a journal?

No.

Scales-   cwP-159

From: John Scales <jscales>
Date: Wed, 21 Sep 94 15:02:37 -0600
To: jfink, barbara
Subject: cwp 159

I've put a copy of cwp 159 on Barbara's desk. A compressed postscript version is in my root directory. So if you want another hardcopy just type

```
zcat ~jscales/cwp-159.ps.Z | lpr
```

cwp-159 / not a proprietary paper
will, however, be distributed to sponsors.

From: John Scales <jscales>
Date: Mon, 19 Sep 94 13:03:34 -0600
To: jfink
Subject: cwp #
Cc: barbara

I need a cwp # for a new report that I'm just finishing up. I hope to have it for you later this week. It is "An Example of Geologic Prior Information for a Bayesian Seismic Inverse Calculation". The authors are Scales and Albert Tarantola.

If you could let me know soon what the number is I can insert it in a proposal I'm writing. Thanks.

An Example of Geologic Prior Information in a Bayesian Seismic Inverse Calculation

John A. Scales

Center for Wave Phenomena, Colorado School of Mines

Albert Tarantola

Institut de Physique du Globe de Paris, Université de Paris 6 et 7

ABSTRACT

All inverse problems require the specification of *a priori* information on the space of models. Typically, it is assumed that the *a priori* uncertainties can be described using Gaussian probabilities, whether explicitly through the use of some *a priori* covariances, or implicitly, with "regularized least squares". Our goal is to try to get beyond these assumptions and let realistic prior information speak for itself. To that end we show here a simple case study in the use of geologic information for seismic inversion. We estimate nonparametrically the Bayesian *a priori* distribution for layered earth models directly from a P-wave sonic log and give a numerical procedure for randomly sampling earth models following this distribution. This procedure is then modified in order to obtain earth models that sample the *a posteriori* distribution, which we regard as a complete solution of the inverse problem. Although the scope of this calculation is limited, the methods employed have a broad utility.

Key words: *a priori* information, inverse problems, Bayesian inference

1 INTRODUCTION

All inverse calculations must balance the extent to which models fit data with the extent to which they conform to our prejudices as to what makes a model good or bad. In some cases it is possible to find completely unrealistic models which nevertheless fit the data. The *a priori* information we use to judge the reasonableness of models comes in many forms, from the subjective wisdom of experts to quantitative data. By incorporating this prior

information into Bayesian probability distributions, we can tackle, in principle, completely general kinds of inverse problems without relying on the particular form of the distribution—as, for example, least squares makes Gaussian assumptions. In this paper we give a concrete example of the nonparametric estimation of a Bayesian prior for earth models derived from a P-wave sonic log. The ultimate goal is to do the seismic inverse problem for surface seismic data and, using the information derived from the sonic log, to constrain, in a Bayesian sense, the inferences we draw from the seismic data. We end up with an algorithm for sampling the *a priori* distribution associated with the well log. To solve the full inverse problem, i.e., to sample the Bayesian *a posteriori* distribution we must, in effect, select the models sampled from the prior distribution by how well they fit the data. We use the Metropolis-like procedure proposed by Mosegaard & Tarantola (1994). This procedure is guaranteed to converge to the posterior distribution.

1.1 Probabilistic Formulation of the Inverse Problem

Assume we are analyzing some physical system (in geophysics this is usually the Earth) that is described by some parameters. Any particular value of the parameters describing the system defines a *model* of the system. A generic model is represented by the symbol $\mathbf{m}$. The collection of all possible models defines the *model space.*

Probabilistic formulations of inverse theory consider probability densities in the model space. Two probability distributions are fundamental ingredients to the theory.

One probability distribution describes the *a priori* information we may have on the parameters describing our model. Here, *a priori* means information that is independent of the data set to be used to refine this information. This *a priori* probability density is denoted $\rho(\mathbf{m})$. The purpose of this paper is to give an example of its definition.

If an experiment produces some data, then a "likelihood function" has to be introduced (Tarantola, 1987) that measures the "degree of fit" between the data predicted by the model (using some physical theory) and the actually observed data. This likelihood function is denoted $L(\mathbf{m})$ and the most popular example is

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{p}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\text{obs}}|^p}{\sigma_i^p}\right) , \tag{1}$$

where, if d_i^{obs} represents the *observed value* for the i-th datum, $g_i(\mathbf{m})$ represents the cal-

culated value corresponding to the model $\mathbf{m}$, and σ_i is an estimation of the uncertainty attached to the i-th observation. For $p = 2$ we have a Gaussian probability density (for independent uncertainties), while for $p = 1$ we have a Laplacian (double exponential) probability density. Equation (1) is only given as an example, and much more realistic expressions may be used that describe better experimental uncertainties, but this is outside the scope of this paper.

Once the probability density $\rho(\mathbf{m})$ and the likelihood function $L(\mathbf{m})$ have been defined, describing respectively the *a priori* information we have on model parameters and the likelihood of a model, the resulting *a posteriori* probability density is given by the expression (see Appendix A for a derivation)

$$\sigma(\mathbf{m}) = k\,\rho(\mathbf{m})\,L(\mathbf{m})\,, \tag{2}$$

where k is a normalization constant. The probability density $\sigma(\mathbf{m})$ combines the *a priori* information with the information coming from observations. The more the *a posteriori* uncertainties on model parameters [as described by $\sigma(\mathbf{m})$] have been reduced compared to the *a priori* uncertainties [as described by $\rho(\mathbf{m})$] the more successful the "inversion" of the data has been.

We can use the function $\sigma(\mathbf{m})$ to find out the probability that the true model is inside a particular region of the model space by integrating $\sigma(\mathbf{m})$ over the particular region. Neglecting the presence of $\rho(\mathbf{m})$ in Equation (2) and judging models only on how well they fit the data is called maximum likelihood estimation. On the other hand, neglecting $L(\mathbf{m})$ and sampling models only according to some prior distribution is akin to geostatistical simulation. The complete solution of the inverse problem combines these two features: sampling models according to a prior distribution and selecting them according to how well they fit the data.

With that brief introduction to the theory, we turn now to a simple example of the probabilistic approach to using prior information. For more details on the theoretical aspects of the problem see Tarantola (1987) and Mosegaard and Tarantola (1994). We will consider the problem of inverting reflection seismic data for subsurface elastic properties in an area where we have a single P-wave sonic log. We regard the sonic log as containing important information about the geology of earth models in the area even though it consists of in-situ measurements made on a much finer scale than the seismic wavelength. The question is, how

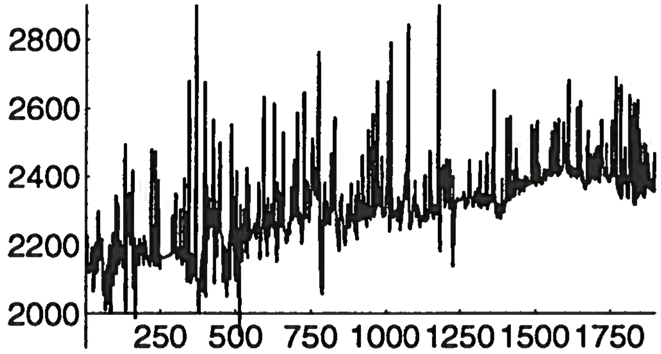

Figure 1. P-wave sonic log. The numbers on the abscissa refer to the samples, which were recorded every 30 cm. The wave speeds are in m/s.

to use this information? We will proceed by extracting the statistical properties from the log and then generating pseudo-random logs with these same statistical properties. By definition then, all of these pseudo-random models are *a priori* realistic in a geologic sense: this is how we sample the prior probability. Once this problem is solved, it is simple (although perhaps expensive) to use the likelihood function to sample the posterior probability.

2 CASE STUDY: EXTRACTING THE PRIOR FROM A SONIC LOG

Figure 1 shows an example of a P-wave sonic log. The wave speeds are in m/s and the samples were recorded every 30 cm. Our point of view is that the log represents a combination of gradual, deterministic processes and essentially random fluctuations. Thus if we subtract the smooth trend of the log (shown in Figure 2) from the original data, we will be left with an essentially random process, as shown in Figure 3. We will regard the trend of the log as being known exactly *a priori*. It could be asserted on geologic grounds or determined by some other geophysical technique such as travel time inversion. In any case, we will focus our attention on the fluctuating part of the log, which we regard as being a realization of a stationary stochastic process whose properties are to be determined. The assumption of stationarity can be relaxed, for example, by dividing the log into pieces associated with time or geologic horizons.

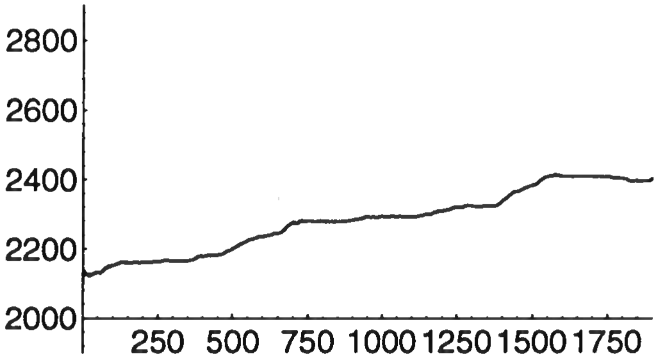

Figure 2. Trend of the log obtained with a sliding 150 sample alpha-trimmed mean filter.

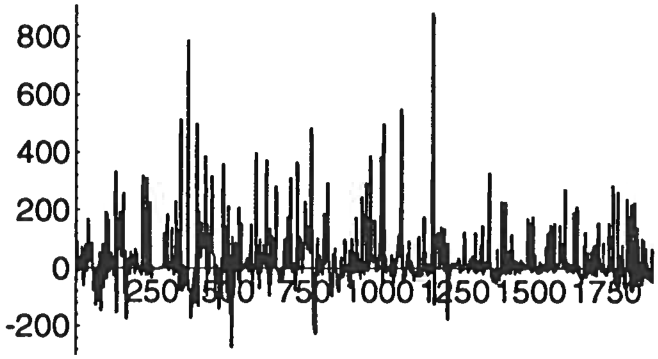

Figure 3. Fluctuating part of the log obtained by subtracting the trend from the original data.

2.1 Random Processes

Some basic principles of random functions are set forth in Appendix B; here we plan to keep the discussion as simple as possible. Consider a field $\xi(\mathbf{r})$ defined at every point $\mathbf{r}$ of the space. If what we have is, in fact, a *random field*, this means that at every point $\mathbf{r}$ of the space we have a random variable $\Xi(\mathbf{r})$. The notation

$$f_n(\xi_1, \xi_2, \ldots, \xi_m; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m)$$

represents the $n-$dimensional (joint) probability density for the random variables

$$\Xi(\mathbf{r}_1), \Xi(\mathbf{r}_2), \ldots, \Xi(\mathbf{r}_m).$$

To describe in all generality a random function requires, for any m, and for any points $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$, to give the $m-$dimensional joint probability density

$$f_m(\xi_1, \xi_2, \ldots, \xi_m; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m).$$

If the space has been discretized using, say n points, then we have only n random variables, and the most general case needs, at most an $n-$dimensional probability density.

Practically, we must rely on the fact that the $n-$dimensional distribution can be expressed in terms of the marginal distributions of the process. For example, the three dimensional distribution $f(x,y,z)$ can be written as $f(x|y,z)f(y,z)$, which can in turn be written as $f(x|yz)f(y|z)f(z)$ and so on by induction for the general n-dimensional distribution. For example, if a process is Markovian, this characterization in terms of marginals is especially simple since

$$f_n(\xi_1,\xi_2,\ldots,\xi_n;\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n)=\frac{f_2(\xi_1,\xi_2;\mathbf{r}_1,\mathbf{r}_2)\ldots f_2(\xi_{n-1},\xi_n;\mathbf{r}_{n-1},\mathbf{r}_n)}{f_1(\xi_2;\mathbf{r}_2)\ f_1(\xi_3;\mathbf{r}_3)\ldots f_1(\xi_{n-1};\mathbf{r}_{n-1})} \tag{3}$$

where f_1 and f_2 are, respectively the one- and two-dimensional marginal distributions associated with f_n. This brings out clearly the fact that for Markov fields the probability of a point depends only on the previous point.

2.2 Estimating the Marginal Distributions

For more general processes than Markovian ones, we need more of the marginal distributions, not just the one- and two-dimensional ones. But, in any case, the marginals can be estimated by making histograms of the data. The process of making these histograms begins by quantizing the log into some number of equal-length velocity intervals. Figure 4 shows how this is done.

Once the log is quantized, we just start counting. For example, to compute the two-dimensional marginals of length one (i.e., the probability density for the velocity values at two consecutive points) we count all the transitions from an interval i to an interval j which occur at adjacent sites along the log (note that we are using here the stationarity assumption). To compute all the two-dimensional distributions we must compute all transitions from an interval i to an interval j over non-adjacent sites as well. Figures 5 and 6 show second order histograms for the P-wave sonic log, the first for transitions of length 1, the second for transitions of length 5.

Clearly it would be impractical to attempt to compute all the marginal distributions for a process of several thousands of points. But at the same time we are not yet willing to restrict ourselves to parametric assumptions, such as Gaussian statistics, that would allow us to reduce the computational burden. Our compromise is to calculate enough of

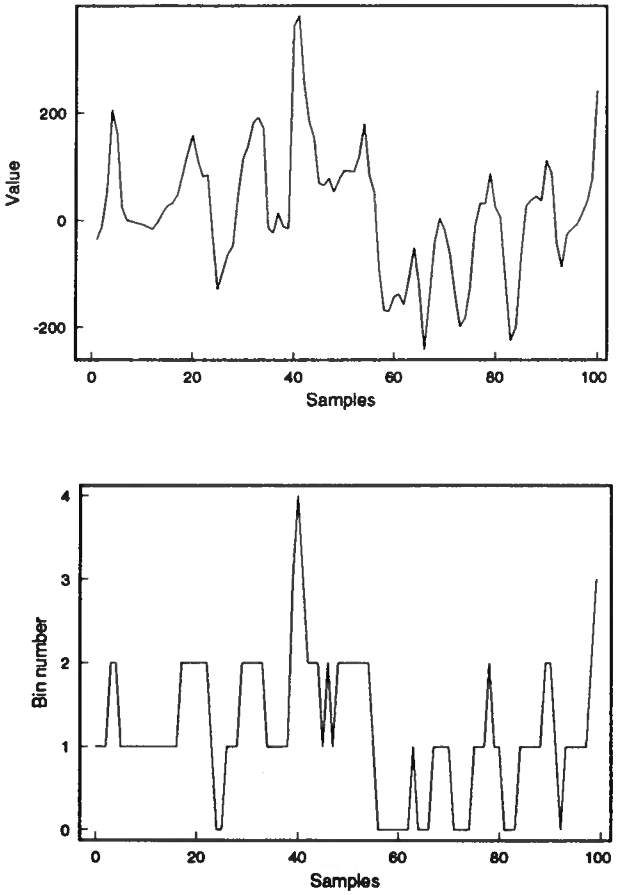

Figure 4. The data are quantized into a number of equal-length velocity intervals so that histograms can be made. Here we show the first 100 samples of the de-meaned sonic log and a toy quantization into 5 intervals. In practice we use a much larger number of intervals, from 100 to 500.

the marginals so as to convince ourselves that we have indeed captured the essence of the process. To quantify this, we use the Kolmogorov-Smirnov two-sample test and compare the original data with pseudo-random simulations calculated with a given number of marginals. The K-S test (described in Appendix C) involves comparing the cumulative distribution function of two different realizations (in this case the data and the simulation), the so-called null hypothesis being that the two were drawn from the same distribution. A high value of the K-S statistic says that the two realizations were very likely drawn from the same underlying distribution; this gives us confidence that the pseudo-random realizations have captured the essence of the data.

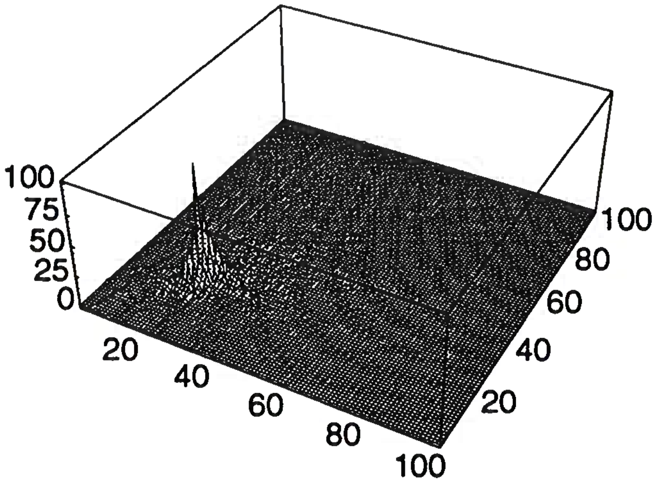

Figure 5. Histograms of two-point transitions of length 1 (i.e., pairwise transitions for adjacent sites along the log) for the P-wave sonic log. If for a given value of x and y axes, say (i, j), the count is n, that means that there are n sites along the log where the site itself and its neighbor have the quantized values i and j respectively.

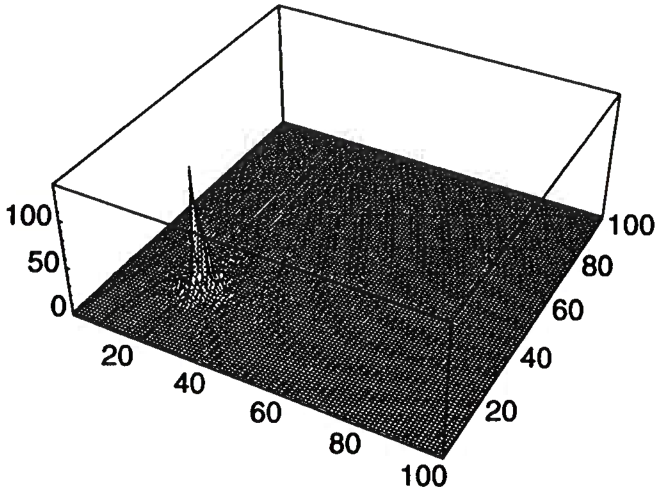

Figure 6. Histograms of two-point transitions of length 5 for the P-wave sonic log. These correspond to transitions from a site k to a site $k + 5$ on the log.

2.3 Sampling the Prior

Once we have computed a sufficient number of the marginals we must be able to sample the prior distribution. To *sample* a probability density means to produce a pseudo random "sample" (or *realization*) of the random variable. By definition, the probability for the produced "point" to belong to any volume of the space must equal the probability of the volume. The algorithms used to sample a probability produce, very often, consecutive samples that are

not independent (think of Brownian motion of a particle: if, in the long term, the particle may be at any point of the space with uniform probability, at close times the particle will be at close points). In such a situation, to obtain independent samples we simply have to "wait" until the algorithm loses the memory of the previous point.

Typically, a sampling algorithm produces a *test point* using some "background probability" (usually the uniform one) and then, uses some random criterion for "accepting" the test point or for "dropping" it. Obviously, if the background probability is very close to the probability we wish to sample, the test points will be often accepted. In this case, one says that the algorithm uses an "importance sampling method". Any nontrivial sampling algorithm, in fact, uses importance sampling, and this will be the case for the examples below.

To do our sampling we use the histograms directly. Suppose, for example, that we needed only the two-point histograms of length 1 (as would be the case if the underlying process were Markovian, since then the probability of a given point depends only on the probability of the previous point). Figure 7 shows a cartoon of this case. The values of the observed log are $\{\xi(\mathbf{r}_i)\}_{i=1}^{N}$. Let us refer to the values of a pseudo-random simulation as $\{\xi'(\mathbf{r}_i)\}_{i=1}^{N}$. We begin the construction of the simulation by assigning some value to $\xi'(\mathbf{r}_1)$, for example, this could be a likely value for the log based on the one-dimensional marginal. Now all we have to do is select $\xi'(\mathbf{r}_2)$ according to the histogram of length 1 transitions. Suppose $\xi'(\mathbf{r}_1) = 4$, as shown in Figure 7. Then the slice through the histogram associated with $i = 4$ is the conditional probability distribution of making the length 1 transition $4 \rightarrow j$ for any j. To assign the value $\xi'(\mathbf{r}_2)$ in accordance with this conditional probability, we make, in effect, a weighted roulette wheel with a sector for each interval number. The size of each interval's sector is in proportion to its probability from the histogram. Then by spinning the roulette wheel we will select $\xi'(\mathbf{r}_2)$ with the appropriate probability. For Markov processes we would proceed in this way right down the log, filling in one value $\xi'(\mathbf{r}_i)$ after another.

But suppose our data are not Markovian? Suppose, for example, that the results of the K-S test say that we must include all two-point transitions out to some length ℓ. In this case we begin as before, filling out the first and second sites along the log. But now, to fill out the third site, we choose the roulette wheel associated with transitions of length two conditioned on the first point $\xi'(\mathbf{r}_1)$. Then we choose the roulette wheel associated with

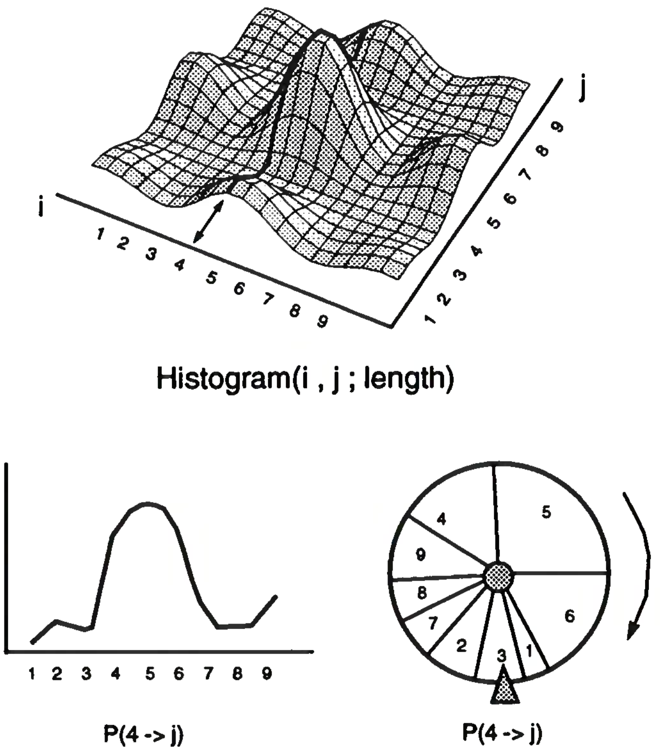

Figure 7. **A weighted roulette wheel is used to sample the histograms. If the current site on the pseudo-random log has a value i and we need a length ℓ transition to fill in the next slot, we first select the length ℓ histogram. Then we extract the one-dimensional marginal associated with the value i. (In the example shown in the figure $i = 4$.) This gives us the observed probability of making a transition from a value i to all other values over a distance of ℓ sites.**

length three transitions conditioned on $\xi'(\mathbf{r}_1)$, and so on until we have filled in the first $\ell+1$ sites on the log. Then we start over again at the site $\ell+2$. This procedure can be generalized to higher order histograms.

Figure 8 shows 15 such logs pseudo-randomly sampled from the second order histograms of the data in Figure 1. In this case we used all the histograms out to a length of 5 samples, which is estimated to be the average correlation length in the medium. Using histograms of this length we are able to routinely generate pseudo-random logs with a K-S statistic of better than 95%. In other words, the distributions associated with the two realizations, pseudo-random and data, are so close that K-S cannot tell whether they are the same or not. But it is important to emphasize, that the K-S test is used only as a "preprocessing" step. Once we have decided on the number of marginals to use, we sample them as described in the text. That means that in the course of the sampling, we may generate simulations with low *a priori* probability but these will be relatively rare occurrences.

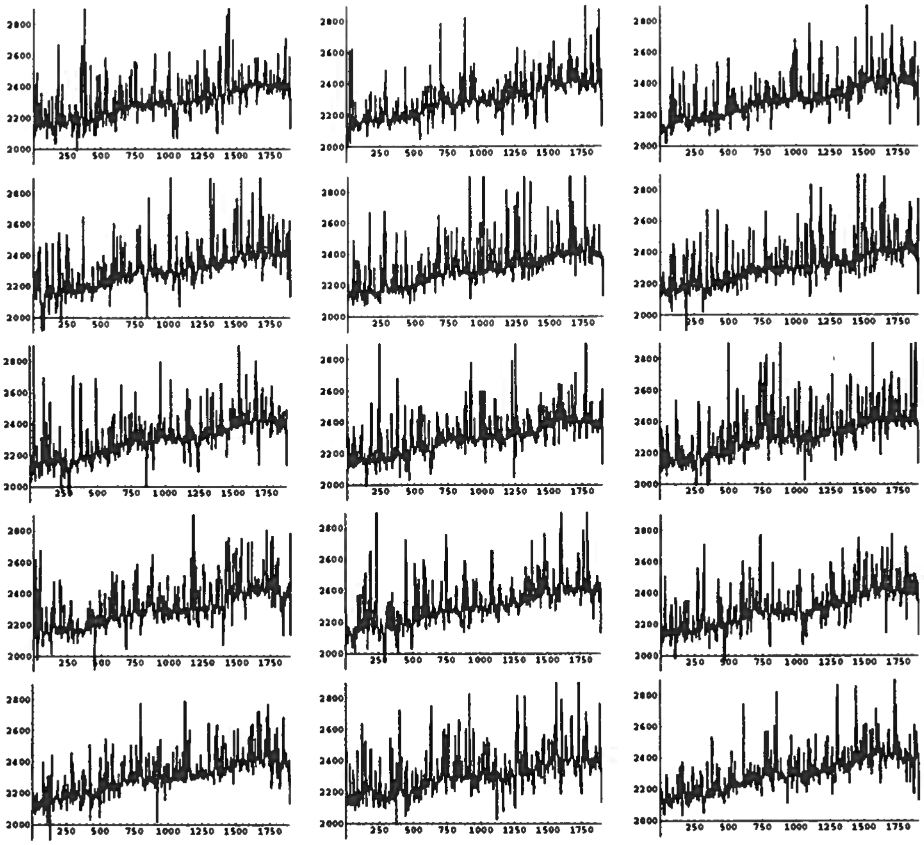

Figure 8. Pseudorandom logs obtained by sampling the *a priori* model distribution. Here we used all the second order histograms up to the average correlation length of the medium. Each of these pseudo-random logs is *a priori* feasible geologically. We have only to rank the models according to how well they fit the seismic data. We can imagine that these 15 models are 15 frames from a movie tour of the prior.

3 SOLVING THE INVERSE PROBLEM

Now that we have a procedure for sampling the prior, we can use these models to sample the posterior according to the algorithm of Mosegaard and Tarantola (1994). This involves accepting or rejecting models according to how well they fit the data, and represents a generalization of the usual Metropolis method for sampling the Gibbs-Boltzman distribution.

More precisely, let us assume that we are able to obtain samples $\mathbf{m}_1, \mathbf{m}_2, \ldots$, of the prior probability distribution $\rho(\mathbf{m})$. Then, if we wish to obtain samples of the posterior probability distribution $\sigma(\mathbf{m}) = k\, \rho(\mathbf{m})\, L(\mathbf{m})$ all that we need to do is to iterate the following procedure.

- Let $\mathbf{m}_i$ be the "current model."

- Use the rules that allow one to sample the prior probability distribution $\rho(\mathbf{m})$ to obtain a new model, say $\mathbf{m}'_i$.

- If $L(\mathbf{m}_i') \geq L(\mathbf{m}_i)$, take $\mathbf{m}_i'$ as new current model, i.e., make $\mathbf{m}_{i+1} = \mathbf{m}_i'$.
- If $L(\mathbf{m}_i') < L(\mathbf{m}_i)$, then decide randomly to take the model $\mathbf{m}_i'$ as new current model, or to destroy it, with a probability of taking it as new current model equal to the ratio $L(\mathbf{m}_i')/L(\mathbf{m}_i)$.

Mosegaard and Tarantola (1994) show that this procedure converges to the true *a posteriori* distribution.

The precise specification of L requires knowledge of the data and modeling uncertainties, as shown in Appendix A. We cannot say what it means to fit the data if we do not know the data uncertainites. This is obviously a crucial issue when inverting real data. But it is beyond the scope of this paper. So although we did not have to make a parametric assumption about the *a priori* distribution, we must make some choice for the data distribution. For this we make the simplest choice, gaussian, so that we can use a least squares criterion to measure data misfit. In other words, for purposes of this demonstration, we will generate synthetic data from the actual log and add a small amount of uncorrelated gaussian noise. Figure 9 shows the synthetic "data", a single zero-offset reflection seismogram, associated with the true model (top) and the same trace contaminated with a small amount (5%) uncorrelated gaussian noise. The seismograms are computed by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson (1967) and includes all multiples, but is limited to zero-offset, acoustic, marine data. Once we have the data trace, the likelihood function is the difference between this trace and the response of the sampled model, normalized by the known data variance

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\text{obs}}|^2}{\sigma^2}\right) \tag{4}$$

where $g_i(\mathbf{m})$ are the synthetic data associated with a sampled model $\mathbf{m}$, d_i^{obs} are the "observed" data, and σ is the standard deviation of the errors.

3.1 Sampling the Posterior

Figure 10 shows a selection of models sampled according to the generalized Metropolis procedure of Mosegaard and Tarantola (1994). We can think of these models as 15 frames from a movie tour of the posterior. But how do these samples of the posterior actually "solve" an/the inverse problem? It is up to us to pose questions and use the posterior to

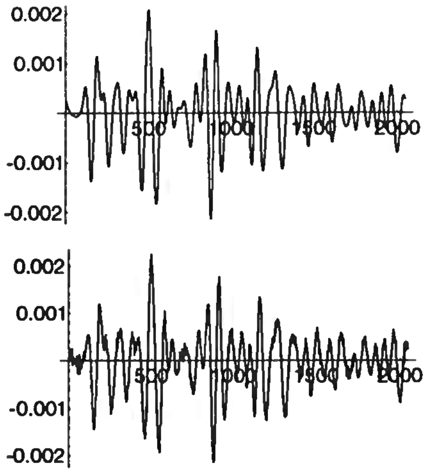

Figure 9. The top figure shows the "data" (i.e., model response) for the true log. The seismogram is obtained by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson's book (1967) and includes all multiples in a fixed time window, but is limited to zero-offset, acoustic, marine data. Below the noise free seismogram is the seismogram with 5% uncorrelated gaussian noise added. We take this to be the observed data for the inverse problem.

answer them probabilistically. A simple illustration of this is to investigate the extent to which a certain feature of our earth models is resolved by the seismic data. Figure 11 shows the distribution of P-wave reflectivity at two points along the log, one near the top of the log ($z = 10$, i.e., the tenth sample along the log) and one near the bottom ($z = 1800$). In both cases the posterior variance is visibly reduced relative to the prior variance. This is a quantitative measure of the extent to which the seismic data are able to resolve these parameters.

4 CAVEATS

There are a great many limitations to this work. Here we list some of the more obvious ones with evidence both for the prosecution (**con**) and the defense (**pro**).

con: The sample size is too small and therefore the *a posteriori* statistics unreliable.

pro: True. The calculations were done on a slow workstation. By moving to a faster workstation and optimizing the code, a factor of 10 speedup is easily obtainable. To move to 2-D, however, will require significantly greater computing resources.

con: The noise model is unrealistic.

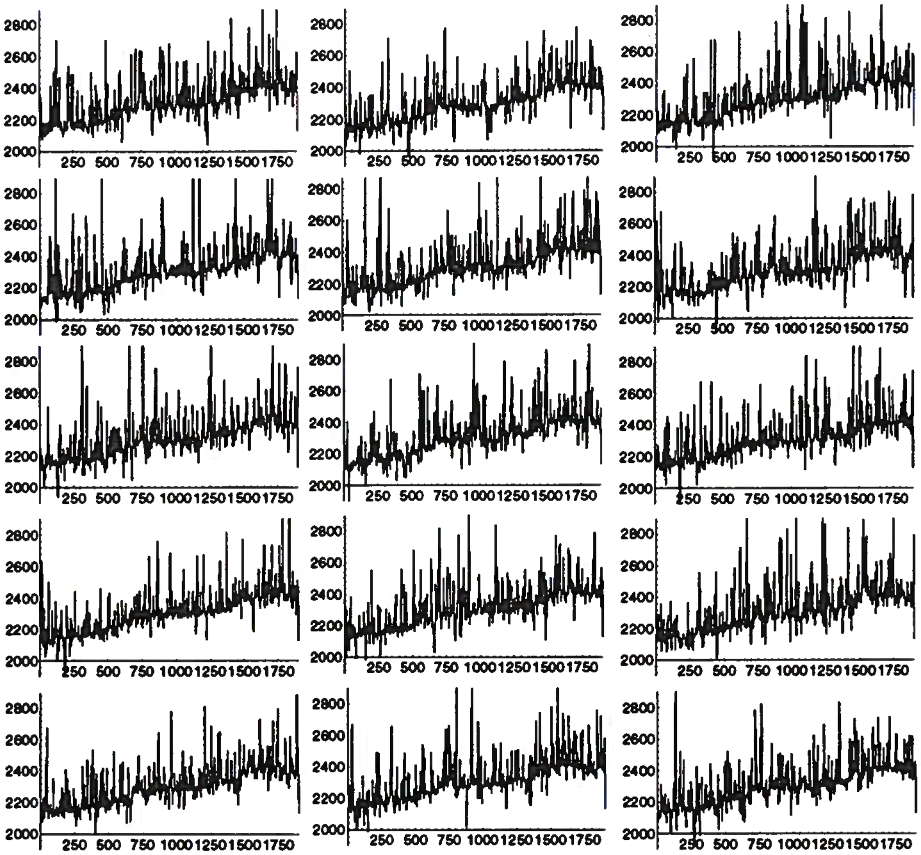

Figure 10. 1D earth models obtained by sampling the *a posteriori* distribution according to the Metropolis-like algorithm of Mosegaard and Tarantola (1994).

pro: True. Until we can estimate the noise directly from the data this will be a problem. In defense of our small *a priori* data error bars, we believe that the "noise" in exploration seismic data which can usefully be classified as gaussian (or some similar distribution) is very small. Far larger are the uncertainties due to modeling errors and, perhaps, coherent environmental noise, for which we have no realistic model.

con: None of the models in the posterior sample would generate seismograms that fit within the (tiny) error bars assigned to the data.

pro: True. This is a function of the small sample size. Should we allow the algorithm to run long enough, it will end by sampling a point inside this region and would remain there a long time, but this might take a prodigious number of iterations. On the other hand, the fact that the Monte Carlo scheme converges to the *a posteriori* distribution means that we can nevertheless compare the relative probability of different models. If the posterior variance of a parameter is smaller than the prior variance, then new information has been obtained,

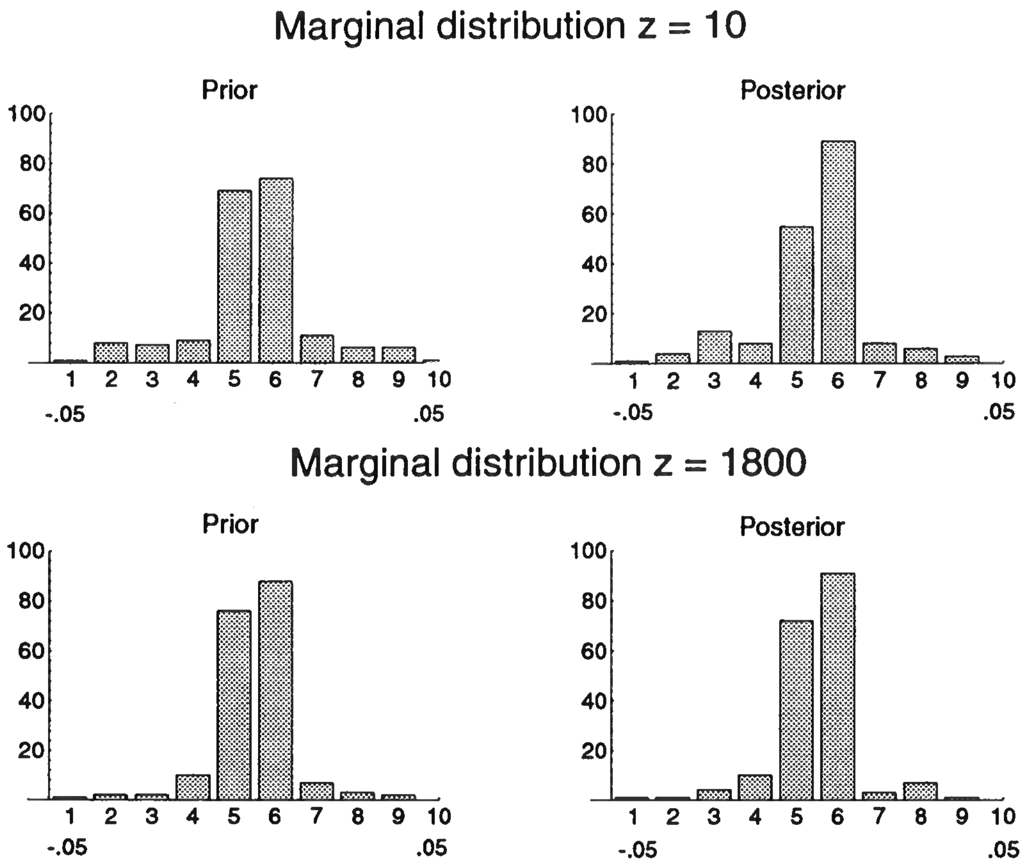

Figure 11. Prior versus posterior distribution of P-wave reflection coefficient at two points in the model. The z values shown refer to the sample number along the log. So these points are at extreme ends of the log. Near the upper surface, the posterior variance is noticeably smaller, indicating an increase in resolution. At the bottom of the log, however, the prior and posterior variances are virtually the same. So the seismic data have not resolved the deeper reflection coefficient at all.

even if the response of the posterior samples do not fit within the *a priori* error bars of the data.

A good simulation should answer positively to the following three questions:

(i) Have we chosen correctly the type of perturbations (of the current model)?

(ii) Is the size of the perturbations adapted to the size of the significant region of the model space?

(iii) Have we started at a point close enough to the region of significant probability so that we will enter it in a reasonable time?

At present, there remains a lot of work to be done in order to answer these questions rigorously.

5 CONCLUSIONS AND FUTURE WORK

We have shown a case study of the use of complex *a priori* information in a Bayesian inverse calculation. The problem we have examined is the inversion of reflection seismograms for 1D earth structure given a well log as *a priori* geologic information. We have developed a non-parametric technique for extracting the Bayesian *a priori* distribution from logs (or other similar data), as well as a method for sampling this *a priori* distribution. This algorithm allows us to generate as many pseudo-random earth models as we like, in accordance with the underlying *a priori* distribution. This sampling of the prior is then coupled with a Metropolis-like procedure to produce a sampling of the *a posteriori* distribution. Once we have this *a posteriori* distribution, which we regard as the ultimate solution of the inverse problem, it is up to us to pose proper questions and use the *a posteriori* distribution to answer them.

6 ACKNOWLEDGEMENTS

We thank Christophe Barnes, Jack Cohen, Philippe Doyen, and Martin Smith for useful discussions. The authors gratefully acknowledge the partial support of the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, Colorado School of Mines, the Shell Foundation, and the sponsors of the Geophysical Tomography Group, Institut de Physique du Globe de Paris, Université de Paris 6 et 7.

References

Duijndam, A. J. W. 1987. *Detailed Bayesian inversion of seismic data.* Ph.D. thesis, Technical University of Delft.

Jeffreys, R.C. 1983. *The logic of decision.* University of Chicago.

Mosegaard, K., & Tarantola, A. 1994. Monte Carlo sampling of solutions to inverse problems. *JGR: submitted.*

Press, William H., Flannery, Brian P., Teukolsky, Saul A., & Vetterling, William T. 1986. *Numerical recipes.* Cambridge University Press.

Pugachev, V. S. 1965. *Theory of random functions and its application to control problems.* Pergamon.

Robinson, E. A. 1967. *Multichannel Time Series Analysis with Digital Computer Programs.* Holden-Day.

Tarantola, A. 1987. *Inverse problem theory.* Elsevier.

APPENDIX A: THE BAYESIAN POSTERIOR PROBABILITY

As we have argued, the posterior probability density on the space of models must be the product of two terms: a term which involves the a priori probability on the space of models and a term which measures the extent of data fit

$$\sigma(\mathbf{d}) = k\rho(\mathbf{m})\ L(\mathbf{m}). \tag{A1}$$

L is called the likelihood function and depends implicitly on the data.

We now show how Equation (A1) follows logically from Bayes' theorem provided we generalize our notion of "data" to allow for the possibility that the data might be specified by probability distributions (Tarantola, 1987). To do so we make use of an idea due to Jeffreys (1983) (as described in (Duijndam, 1987)). We begin by using the notation common amongst Bayesians, then we show how this relates to the more standard inverse-theoretic notation in Tarantola (1987).

In this approach we assume that we have some prior joint distribution $p_0(\mathbf{m}, \mathbf{d})$. Further, we suppose that as the result of some observation, the marginal pdf of $\mathbf{d}$ changes to $p_1(\mathbf{d})$. We regard $p_1(\mathbf{d})$ as being the "data" in the sense that we often know the data only as a distribution, not exact numbers. In the special case where the data are exactly known, $p_1(\mathbf{d})$ reduces to a delta function $\delta(\mathbf{d} - \mathbf{d}_{obs})$.

How do we use this new information in the solution of the inverse problem? The answer is based upon the following assumption: whereas the information on $\mathbf{d}$ has changed as a result of the experiment, there is no reason to think that the conditional degree of belief of $\mathbf{m}$ on $\mathbf{d}$ has. I.e.,

$$p_1(\mathbf{m}|\mathbf{d}) = p_0(\mathbf{m}|\mathbf{d}). \tag{A2}$$

From this one can derive the posterior marginal $p_1(\mathbf{m})$:

$$\begin{aligned} p_1(\mathbf{m}) &\equiv \int_D p_1(\mathbf{m}, \mathbf{d})\, d\mathbf{d} && \text{(A3)} \\ &= \int_D p_1(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} && \text{(A4)} \\ &= \int_D p_0(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} && \text{(A5)} \\ &= \int_D \frac{p_0(\mathbf{d}|\mathbf{m}) p_0(\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d} && \text{(A6)} \\ &= p_0(\mathbf{m}) \int_D \frac{p_0(\mathbf{d}|\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d}, && \text{(A7)} \end{aligned}$$

where D denotes the data space.

Switching now to the inverse-theoretic notation, let us regard $p_0(\mathbf{d})$ as being the non-informative prior distribution on data $\mu_D(\mathbf{d})$: this is what we know about the data **before** we've actually done this particular experiment. Further, we identify $p_1(\mathbf{m})$ as the posterior distribution on the space of models, $p_1(\mathbf{d})$ as the data errors, $p_0(\mathbf{m})$ as the prior distribution on the space of models, and $p_0(\mathbf{d}|\mathbf{m})$ as the modeling errors:

$$p_1(\mathbf{m}) \equiv \sigma(\mathbf{m})$$

$$p_1(\mathbf{d}) \equiv \rho_D(\mathbf{d})$$

$$p_0(\mathbf{m}) \equiv \rho_M(\mathbf{m})$$

$$p_0(\mathbf{d}|\mathbf{m}) \equiv \Theta(\mathbf{d}|\mathbf{m}),$$

then we arrive at precisely Equation (1.65) of Tarantola (1987)

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \int_D \frac{\rho_D(\mathbf{d})\Theta(\mathbf{d}|\mathbf{m})}{\mu_D(\mathbf{d})}\, d\mathbf{d} \qquad \text{(A8)}$$

An important special case occurs when the modeling errors are negligible, i.e., we have a perfect theory. Then the conditional distribution $\Theta(\mathbf{d}|\mathbf{m})$ reduces to a delta function $\delta(\mathbf{d} - g(\mathbf{m}))$ where g is the forward operator. In this case, the posterior is simply

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \left[\frac{\rho_D(\mathbf{d})}{\mu_D(\mathbf{d})}\right]_{\mathbf{d}=g(\mathbf{m})} . \qquad \text{(A9)}$$

APPENDIX B: ELEMENTS OF RANDOM FIELDS

Here we set forth the basic notations having to do with random functions. This section is an adaptation of Pugachev (1965). We will consider random fields in (3-D) physical space. Adaptation to other sorts of fields is straightforward.

$\Xi(\mathbf{r})$ will denote a random field. A realization of the random field will be denoted by $\xi(\mathbf{r})$. We may think of ξ as a physical parameter defined at every point of the space (like the velocity of seismic waves, the temperature, etc.). We will work inside a volume $\mathcal{V}$.

B1 n-dimensional joint probability densities

Let $(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ be a set of n points inside $\mathcal{V}$. An expression like

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$$

will denote the n-dimensional (joint) probability density for the values $t(\mathbf{r}_1), t(\mathbf{r}_2), \dots, t(\mathbf{r}_n)$. The notation $f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ may seem complicated, but it just indicates that for every different set of points $(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ we have a possibly different probability density.

The random field $T(\mathbf{r})$ is completely characterized if, for any set of n points inside $\mathcal{V}$, the joint probability density $f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ is defined, and this for any value of n.

B2 Marginal and conditional probability

The definitions for marginal and conditional volumetric probabilities do not pose any special difficulty. As notations rapidly become intricate, let us only give the corresponding definitions for some particular cases, the generalization being straightforward.

If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *marginal* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$ is defined by

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)\,, \tag{A10}$$

where $dL(\xi_3)$ is the (1-D) volume element (it may be $d\xi_3$ or something different).

Let us now turn now to the illustration of the definition of conditional probability. If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random

field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *conditional* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$, given that the random field takes the value ξ_3 at point $\mathbf{r}_3$, is defined by

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{\int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)} , \tag{A11}$$

i.e., using equation A10,

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)} . \tag{A12}$$

Of particular interest will be the one- and the two-dimensional probability densities, denoted respectively $f_1(t; \mathbf{r})$ and $f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)$.

B3 Random fields defined by low order probability densities

First example: Independently distributed variables.

If

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$g(\xi_1, \mathbf{r}_1)\ h(\xi_2, \mathbf{r}_2) \ldots i(\xi_n, \mathbf{r}_n) , \tag{A13}$$

we say that the random field has independently distributed variables.

Second example: Markov random field.

For a Markov process,

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_2(\xi_n; \mathbf{r}_n | \xi_{n-1}; \mathbf{r}_{n-1})\ f_2(\xi_{n-1}; \mathbf{r}_{n-1} | \xi_{n-2}; \mathbf{r}_{n-2}) \tag{A14}$$

$$\ldots f_2(\xi_2; \mathbf{r}_2 | \xi_1; \mathbf{r}_1)\ f_1(\xi_1; \mathbf{r}_1) ,$$

where the vertical bar denotes conditional probability. This means that the value at a given point depends only on the value at the previous point. As

$$f_2(\xi_i; \mathbf{r}_i | \xi_{i-1}; \mathbf{r}_{i-1}) = \frac{f_2(\xi_{i-1}, \xi_i; \mathbf{r}_{i-1}, \mathbf{r}_i)}{f_1(\xi_{i-1}; \mathbf{r}_{i-1})} , \tag{A15}$$

we obtain

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = \tag{A16}$$

$$\frac{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) \ldots f_2(\xi_{n-1}, \xi_n; \mathbf{r}_{n-1}, \mathbf{r}_n)}{f_1(\xi_2; \mathbf{r}_2)\ f_1(\xi_3; \mathbf{r}_3) \ldots f_1(\xi_{n-1}; \mathbf{r}_{n-1})} .$$

This equation characterizes a Markov random field in all generality. It means that the random

field is completely characterized by 1-D and 2-D probability densities (defined at adjacent points).

Third example: Gaussian random field.

For a Gaussian random field, if we know the 2-D distributions, we know all the means and all the covariances, so we also know the n-dimensional distribution. It can be shown that a Gaussian process with exponential covariance is Markovian.

B4 Uniform random fields

A random field is uniform (i.e., stationary) in the strong sense if for any $\mathbf{r}_0$,

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_0 + \mathbf{r}_1, \mathbf{r}_0 + \mathbf{r}_2, \ldots, \mathbf{r}_0 + \mathbf{r}_n) \ .$$

Taking

$$\mathbf{r}_0 = -\mathbf{r}_1$$

gives then

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{0}, \mathbf{r}_2 - \mathbf{r}_1, \ldots, \mathbf{r}_n - \mathbf{r}_1) \ .$$

A distribution is said to be uniform in the weak sense if this expression holds only for $n = 1$ and $n = 2$.

As the random field is defined over the physical space, we prefer the term *uniform* to characterize what, in random fields defined over a time variable, is called *stationary*. This is entirely a question of nomenclature; we regard the terms as being interchangeable.

Example:

For the two-dimensional distribution,

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \Psi_2(\xi_1, \xi_2; \Delta\mathbf{r}) \ ,$$

with

$$\Delta\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1 \ .$$

Example:

For the three-dimensional distribution,

$$f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \Psi_3(\xi_1, \xi_2, \xi_3; \Delta\mathbf{r}_1, \Delta\mathbf{r}_2) \ .$$

with

$$\Delta\mathbf{r}_1 = \mathbf{r}_2 - \mathbf{r}_1$$

and

$$\Delta\mathbf{r}_2 = \mathbf{r}_3 - \mathbf{r}_1 \ .$$

Essentially a uniform random process is one whose properties do not change with space (or time if that is the independent variable).

APPENDIX C: THE KOLMOGOROV-SMIRNOV TEST

This brief discussion is taken directly from *Numerical Recipes* (1986). The two-sample Kolmogorov-Smirnov statistic tests the null hypothesis that two data sets are drawn from the same distribution. It is based on a comparison of the cumulative distribution functions (CDF) of the two data sets. One can imagine any number of comparisons between the two CDFs. K-S represents an especially simple one: it is defined as the maximum value of the absolute difference between the two CDFs. In symbols, the K-S statistic D is given by

$$D = \max_{-\infty<x<\infty} |S_{N_1}(x) - S_{N_2}(x)|$$

where $S_{N_1}(x)$ and $S_{N_2}(x)$ are the approximate CDFs for the two data sets. The key point, however, is that the distribution function for the K-S statistic itself (for the null-hypothesis that the data sets are drawn from the same distribution) can be calulated approximately. The significance function Q for this test is given by the following approximation:

$$Q_{K-S}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}.$$

A plot of this significance function is given in Figure A1.

For the two-sample test, the significance level for an observed value of D (as a disproof of the null-hypothesis) is given approximately by

$$\text{Probability}(D > \text{observed}) = Q_{K-S}\left(\sqrt{\frac{N_1 N_2}{N_1 + N_2}}D\right)$$

where N_2 and N_2 are the numbers of samples in the two data sets.

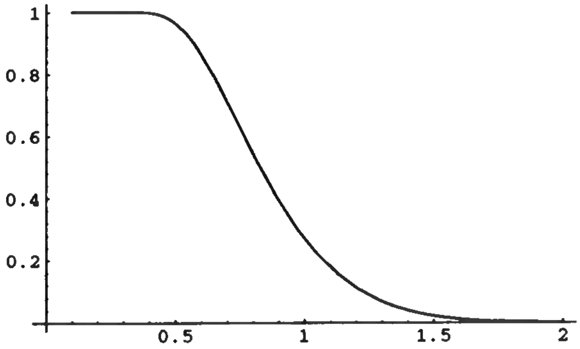

Figure A1. Kolmogorov-Smirnov significance function.

An Example of Geologic Prior Information in a Bayesian Seismic Inverse Calculation

John A. Scales

Center for Wave Phenomena, Colorado School of Mines

Albert Tarantola

Institut de Physique du Globe de Paris, Université de Paris 6 et 7

ABSTRACT

All inverse problems require the specification of *a priori* information on the space of models. Typically, it is assumed that the *a priori* uncertainties can be described using Gaussian probabilities, whether explicitly through the use of some *a priori* covariances, or implicitly, with "regularized least squares". Our goal is to try to get beyond these assumptions and let realistic prior information speak for itself. To that end we show here a simple case study in the use of geologic information for seismic inversion. We estimate nonparametrically the Bayesian *a priori* distribution for layered earth models directly from a P-wave sonic log and give a numerical procedure for randomly sampling earth models following this distribution. This procedure is then modified in order to obtain earth models that sample the *a posteriori* distribution, which we regard as a complete solution of the inverse problem. Although the scope of this calculation is limited, the methods employed have a broad utility.

Key words: *a priori* information, inverse problems, Bayesian inference

1 INTRODUCTION

All inverse calculations must balance the extent to which models fit data with the extent to which they conform to our prejudices as to what makes a model good or bad. In some cases it is possible to find completely unrealistic models which nevertheless fit the data. The *a priori* information we use to judge the reasonableness of models comes in many forms, from the subjective wisdom of experts to quantitative data. By incorporating this prior

ve only n random variables,
obability density.
onal distribution can be ex-
or example, the three dimen-
, which can in turn be written
-dimensional distribution. For
erms of marginals is especially

$$\frac{\xi_n; \mathbf{r}_{n-1}, \mathbf{r}_n)}{\xi_{n-1}; \mathbf{r}_{n-1})} \qquad (3)$$

onal marginal distributions asso-
Markov fields the probability of a

eed more of the marginal distribu-
, in any case, the marginals can be
ss of making these histograms begins
th velocity intervals. Figure 4 shows

g. For example, to compute the two-
ity density for the velocity values at two
an interval i to an interval j which occur
sing here the stationarity assumption).
ve must compute all transitions from an
s well. Figures 5 and 6 show second order
r transitions of length 1, the second for

to compute all the marginal distributions
it at the same time we are not yet willing
s, such as Gaussian statistics, that would
Our compromise is to calculate enough of

wave

log
tion
now
aps

ples
n of
ract
with
g as
l by
ocus
n of
n of
ime

, we can tackle, in principle, completely
the particular form of the distribution—
ptions. In this paper we give a concrete
an prior for earth models derived from
mic inverse problem for surface seismic
log, to constrain, in a Bayesian sense,
d up with an algorithm for sampling
. To solve the full inverse problem,
ve must, in effect, select the models
t the data. We use the Metropolis-
4). This procedure is guaranteed to

blem

hysics this is usually the Earth)
of the parameters describing the
presented by the symbol $\mathbf{m}$. The

bability densities in the model
ents to the theory.
mation we may have on the
rmation that is independent
priori probability density is
of its definition.
ction" has to be introduced
the data predicted by the
a. This likelihood function

(1)

$(\mathbf{m})$ represents the cal-

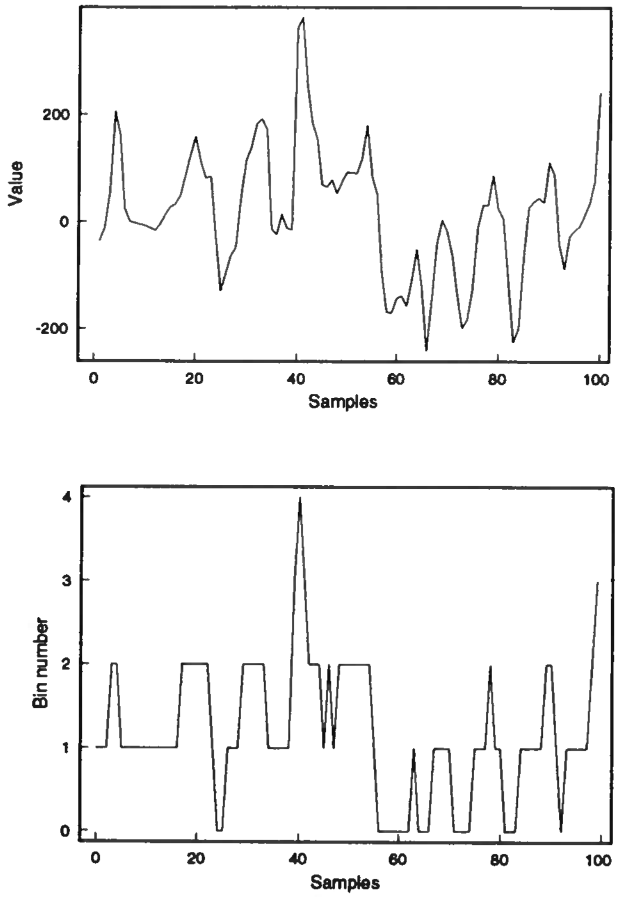

Figure 4. The data are quantized into a number of equal-length velocity intervals so that histograms can be made. Here we show the first 100 samples of the de-meaned sonic log and a toy quantization into 5 intervals. In practice we use a much larger number of intervals, from 100 to 500.

the marginals so as to convince ourselves that we have indeed captured the essence of the process. To quantify this, we use the Kolmogorov-Smirnov two-sample test and compare the original data with pseudo-random simulations calculated with a given number of marginals. The K-S test (described in Appendix C) involves comparing the cumulative distribution function of two different realizations (in this case the data and the simulation), the so-called null hypothesis being that the two were drawn from the same distribution. A high value of the K-S statistic says that the two realizations were very likely drawn from the same underlying distribution; this gives us confidence that the pseudo-random realizations have captured the essence of the data.

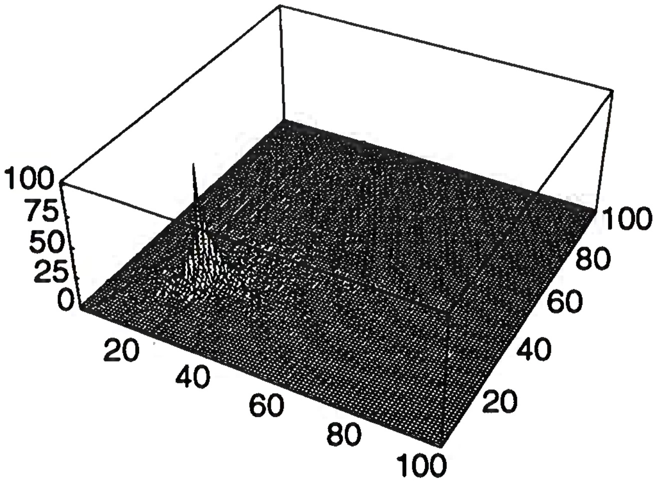

Figure 5. Histograms of two-point transitions of length 1 (i.e., pairwise transitions for adjacent sites along the log) for the P-wave sonic log. If for a given value of x and y axes, say (i, j), the count is n, that means that there are n sites along the log where the site itself and its neighbor have the quantized values i and j respectively.

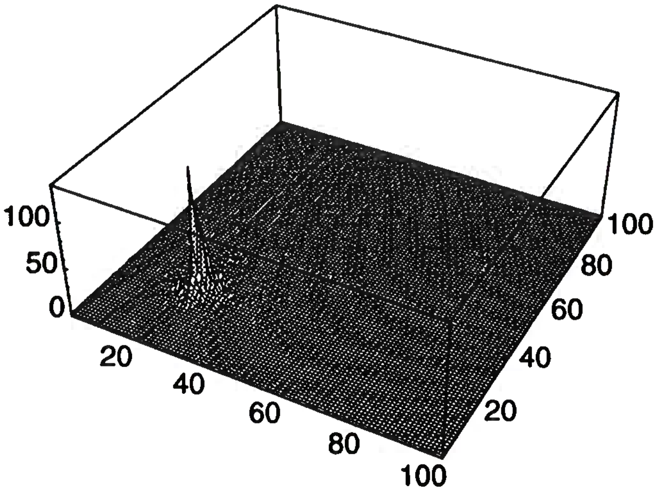

Figure 6. Histograms of two-point transitions of length 5 for the P-wave sonic log. These correspond to transitions from a site k to a site $k + 5$ on the log.

2.3 Sampling the Prior

Once we have computed a sufficient number of the marginals we must be able to sample the prior distribution. To *sample* a probability density means to produce a pseudo random "sample" (or *realization*) of the random variable. By definition, the probability for the produced "point" to belong to any volume of the space must equal the probability of the volume. The algorithms used to sample a probability produce, very often, consecutive samples that are

not independent (think of Brownian motion of a particle: if, in the long term, the particle may be at any point of the space with uniform probability, at close times the particle will be at close points). In such a situation, to obtain independent samples we simply have to "wait" until the algorithm loses the memory of the previous point.

Typically, a sampling algorithm produces a *test point* using some "background probability" (usually the uniform one) and then, uses some random criterion for "accepting" the test point or for "dropping" it. Obviously, if the background probability is very close to the probability we wish to sample, the test points will be often accepted. In this case, one says that the algorithm uses an "importance sampling method". Any nontrivial sampling algorithm, in fact, uses importance sampling, and this will be the case for the examples below.

To do our sampling we use the histograms directly. Suppose, for example, that we needed only the two-point histograms of length 1 (as would be the case if the underlying process were Markovian, since then the probability of a given point depends only on the probability of the previous point). Figure 7 shows a cartoon of this case. The values of the observed log are $\{\xi(\mathbf{r}_i)\}_{i=1}^{N}$. Let us refer to the values of a pseudo-random simulation as $\{\xi'(\mathbf{r}_i)\}_{i=1}^{N}$. We begin the construction of the simulation by assigning some value to $\xi'(\mathbf{r}_1)$, for example, this could be a likely value for the log based on the one-dimensional marginal. Now all we have to do is select $\xi'(\mathbf{r}_2)$ according to the histogram of length 1 transitions. Suppose $\xi'(\mathbf{r}_1) = 4$, as shown in Figure 7. Then the slice through the histogram associated with $i = 4$ is the conditional probability distribution of making the length 1 transition $4 \longrightarrow j$ for any j. To assign the value $\xi'(\mathbf{r}_2)$ in accordance with this conditional probability, we make, in effect, a weighted roulette wheel with a sector for each interval number. The size of each interval's sector is in proportion to its probability from the histogram. Then by spinning the roulette wheel we will select $\xi'(\mathbf{r}_2)$ with the appropriate probability. For Markov processes we would proceed in this way right down the log, filling in one value $\xi'(\mathbf{r}_i)$ after another.

But suppose our data are not Markovian? Suppose, for example, that the results of the K-S test say that we must include all two-point transitions out to some length ℓ. In this case we begin as before, filling out the first and second sites along the log. But now, to fill out the third site, we choose the roulette wheel associated with transitions of length two conditioned on the first point $\xi'(\mathbf{r}_1)$. Then we choose the roulette wheel associated with

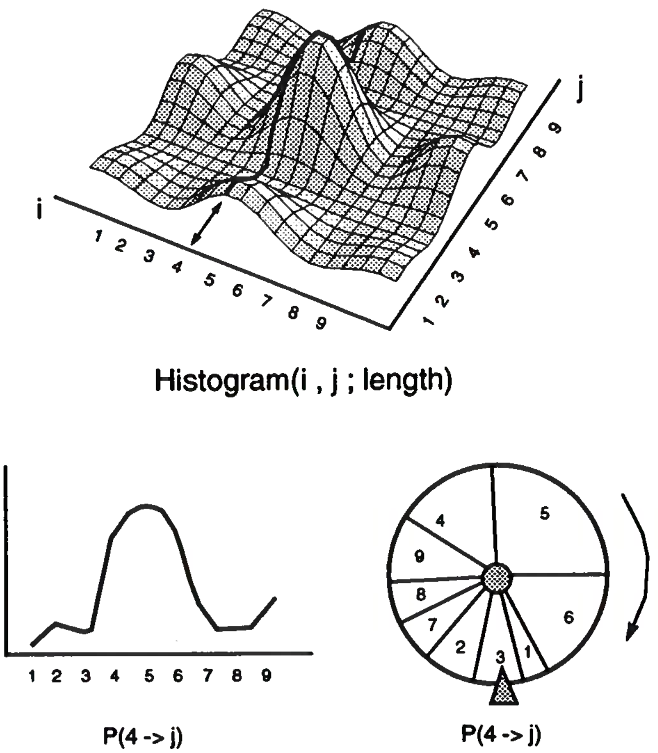

Figure 7. **A weighted roulette wheel is used to sample the histograms. If the current site on the pseudo-random log has a value i and we need a length ℓ transition to fill in the next slot, we first select the length ℓ histogram. Then we extract the one-dimensional marginal associated with the value i. (In the example shown in the figure $i = 4$.) This gives us the observed probability of making a transition from a value i to all other values over a distance of ℓ sites.**

length three transitions conditioned on $\xi'(\mathbf{r}_1)$, and so on until we have filled in the first $\ell + 1$ sites on the log. Then we start over again at the site $\ell + 2$. This procedure can be generalized to higher order histograms.

Figure 8 shows 15 such logs pseudo-randomly sampled from the second order histograms of the data in Figure 1. In this case we used all the histograms out to a length of 5 samples, which is estimated to be the average correlation length in the medium. Using histograms of this length we are able to routinely generate pseudo-random logs with a K-S statistic of better than 95%. In other words, the distributions associated with the two realizations, pseudo-random and data, are so close that K-S cannot tell whether they are the same or not. But it is important to emphasize, that the K-S test is used only as a "preprocessing" step. Once we have decided on the number of marginals to use, we sample them as described in the text. That means that in the course of the sampling, we may generate simulations with low *a priori* probability but these will be relatively rare occurrences.

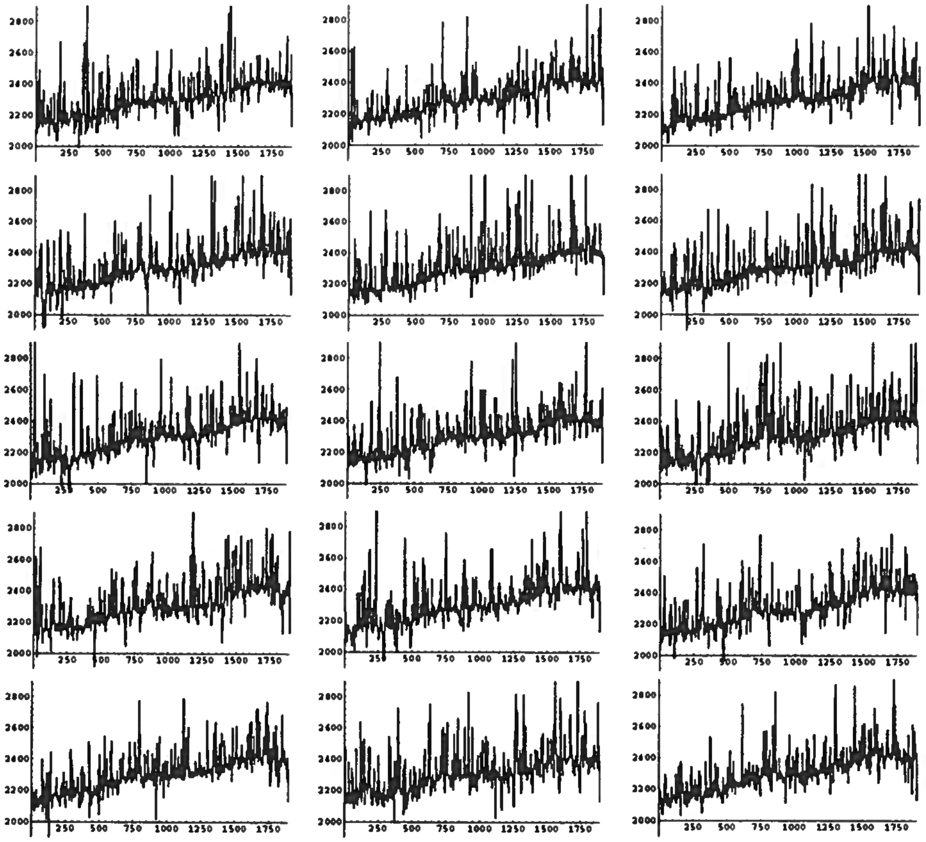

Figure 8. Pseudorandom logs obtained by sampling the *a priori* model distribution. Here we used all the second order histograms up to the average correlation length of the medium. Each of these pseudo-random logs is *a priori* feasible geologically. We have only to rank the models according to how well they fit the seismic data. We can imagine that these 15 models are 15 frames from a movie tour of the prior.

3 SOLVING THE INVERSE PROBLEM

Now that we have a procedure for sampling the prior, we can use these models to sample the posterior according to the algorithm of Mosegaard and Tarantola (1994). This involves accepting or rejecting models according to how well they fit the data, and represents a generalization of the usual Metropolis method for sampling the Gibbs-Boltzman distribution.

More precisely, let us assume that we are able to obtain samples $\mathbf{m}_1, \mathbf{m}_2, \ldots,$ of the prior probability distribution $\rho(\mathbf{m})$. Then, if we wish to obtain samples of the posterior probability distribution $\sigma(\mathbf{m}) = k\,\rho(\mathbf{m})\,L(\mathbf{m})$ all that we need to do is to iterate the following procedure.

- Let $\mathbf{m}_i$ be the "current model."
- Use the rules that allow one to sample the prior probability distribution $\rho(\mathbf{m})$ to obtain a new model, say $\mathbf{m}'_i$.

- If $L(\mathbf{m}_i') \geq L(\mathbf{m}_i)$, take $\mathbf{m}_i'$ as new current model, i.e., make $\mathbf{m}_{i+1} = \mathbf{m}_i'$.
- If $L(\mathbf{m}_i') < L(\mathbf{m}_i)$, then decide randomly to take the model $\mathbf{m}_i'$ as new current model, or to destroy it, with a probability of taking it as new current model equal to the ratio $L(\mathbf{m}_i')/L(\mathbf{m}_i)$.

Mosegaard and Tarantola (1994) show that this procedure converges to the true *a posteriori* distribution.

The precise specification of L requires knowledge of the data and modeling uncertainties, as shown in Appendix A. We cannot say what it means to fit the data if we do not know the data uncertainites. This is obviously a crucial issue when inverting real data. But it is beyond the scope of this paper. So although we did not have to make a parametric assumption about the *a priori* distribution, we must make some choice for the data distribution. For this we make the simplest choice, gaussian, so that we can use a least squares criterion to measure data misfit. In other words, for purposes of this demonstration, we will generate synthetic data from the actual log and add a small amount of uncorrelated gaussian noise. Figure 9 shows the synthetic "data", a single zero-offset reflection seismogram, associated with the true model (top) and the same trace contaminated with a small amount (5%) uncorrelated gaussian noise. The seismograms are computed by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson (1967) and includes all multiples, but is limited to zero-offset, acoustic, marine data. Once we have the data trace, the likelihood function is the difference between this trace and the response of the sampled model, normalized by the known data variance

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\text{obs}}|^2}{\sigma^2}\right) \tag{4}$$

where $g_i(\mathbf{m})$ are the synthetic data associated with a sampled model $\mathbf{m}$, d_i^{obs} are the "observed" data, and σ is the standard deviation of the errors.

3.1 Sampling the Posterior

Figure 10 shows a selection of models sampled according to the generalized Metropolis procedure of Mosegaard and Tarantola (1994). We can think of these models as 15 frames from a movie tour of the posterior. But how do these samples of the posterior actually "solve" an/the inverse problem? It is up to us to pose questions and use the posterior to

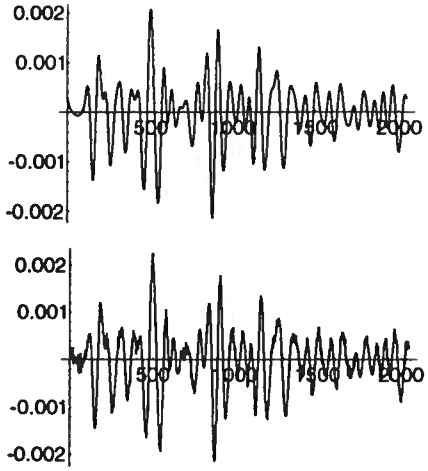

Figure 9. The top figure shows the "data" (i.e., model response) for the true log. The seismogram is obtained by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson's book (1967) and includes all multiples in a fixed time window, but is limited to zero-offset, acoustic, marine data. Below the noise free seismogram is the seismogram with 5% uncorrelated gaussian noise added. We take this to be the observed data for the inverse problem.

answer them probabilistically. A simple illustration of this is to investigate the extent to which a certain feature of our earth models is resolved by the seismic data. Figure 11 shows the distribution of P-wave reflectivity at two points along the log, one near the top of the log ($z = 10$, i.e., the tenth sample along the log) and one near the bottom ($z = 1800$). In both cases the posterior variance is visibly reduced relative to the prior variance. This is a quantitative measure of the extent to which the seismic data are able to resolve these parameters.

4 CAVEATS

There are a great many limitations to this work. Here we list some of the more obvious ones with evidence both for the prosecution (**con**) and the defense (**pro**).

con: The sample size is too small and therefore the *a posteriori* statistics unreliable.

pro: True. The calculations were done on a slow workstation. By moving to a faster workstation and optimizing the code, a factor of 10 speedup is easily obtainable. To move to 2-D, however, will require significantly greater computing resources.

con: The noise model is unrealistic.

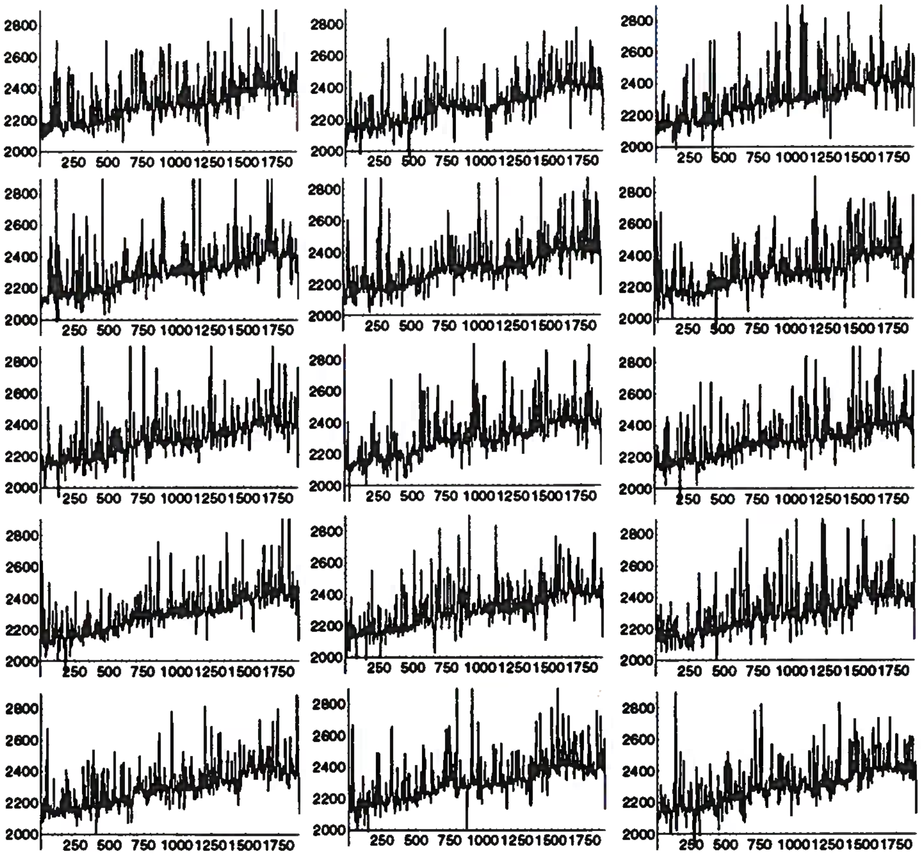

Figure 10. 1D earth models obtained by sampling the *a posteriori* distribution according to the Metropolis-like algorithm of Mosegaard and Tarantola (1994).

pro: True. Until we can estimate the noise directly from the data this will be a problem. In defense of our small *a priori* data error bars, we believe that the "noise" in exploration seismic data which can usefully be classified as gaussian (or some similar distribution) is very small. Far larger are the uncertainties due to modeling errors and, perhaps, coherent environmental noise, for which we have no realistic model.

con: None of the models in the posterior sample would generate seismograms that fit within the (tiny) error bars assigned to the data.

pro: True. This is a function of the small sample size. Should we allow the algorithm to run long enough, it will end by sampling a point inside this region and would remain there a long time, but this might take a prodigious number of iterations. On the other hand, the fact that the Monte Carlo scheme converges to the *a posteriori* distribution means that we can nevertheless compare the relative probability of different models. If the posterior variance of a parameter is smaller than the prior variance, then new information has been obtained,

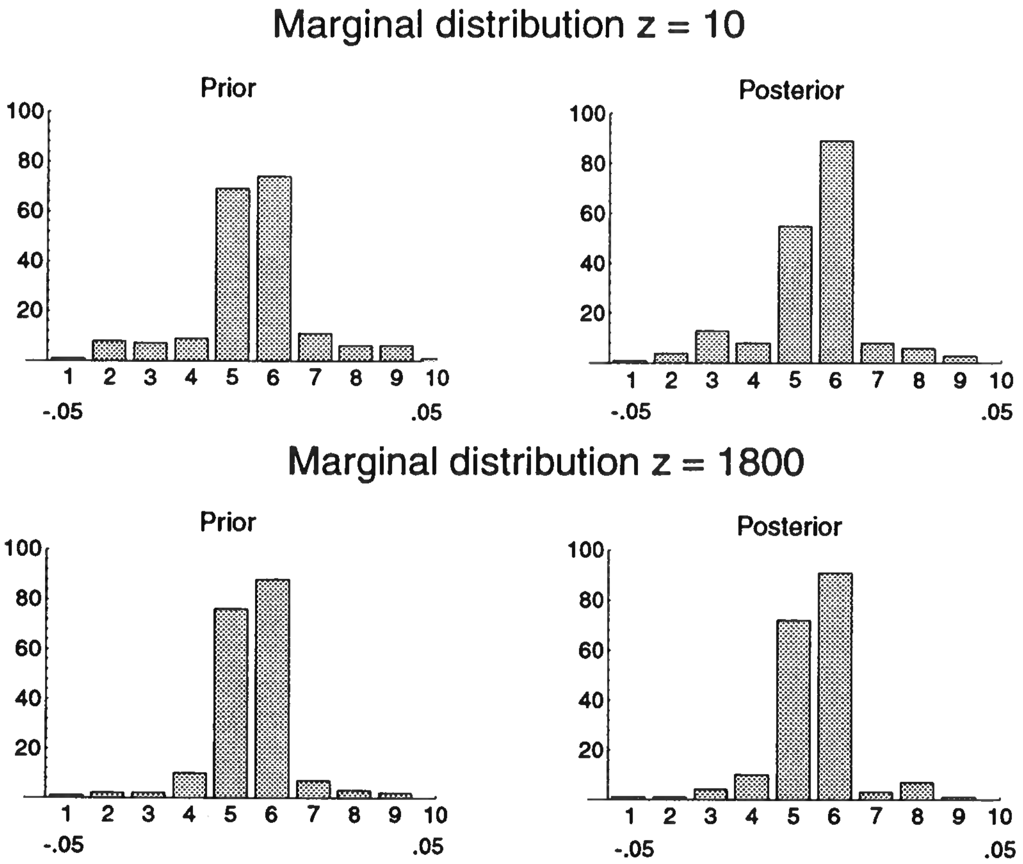

Figure 11. Prior versus posterior distribution of P-wave reflection coefficient at two points in the model. The z values shown refer to the sample number along the log. So these points are at extreme ends of the log. Near the upper surface, the posterior variance is noticeably smaller, indicating an increase in resolution. At the bottom of the log, however, the prior and posterior variances are virtually the same. So the seismic data have not resolved the deeper reflection coefficient at all.

even if the response of the posterior samples do not fit within the *a priori* error bars of the data.

A good simulation should answer positively to the following three questions:

(i) Have we chosen correctly the type of perturbations (of the current model)?

(ii) Is the size of the perturbations adapted to the size of the significant region of the model space?

(iii) Have we started at a point close enough to the region of significant probability so that we will enter it in a reasonable time?

At present, there remains a lot of work to be done in order to answer these questions rigorously.

5 CONCLUSIONS AND FUTURE WORK

We have shown a case study of the use of complex *a priori* information in a Bayesian inverse calculation. The problem we have examined is the inversion of reflection seismograms for 1D earth structure given a well log as *a priori* geologic information. We have developed a non-parametric technique for extracting the Bayesian *a priori* distribution from logs (or other similar data), as well as a method for sampling this *a priori* distribution. This algorithm allows us to generate as many pseudo-random earth models as we like, in accordance with the underlying *a priori* distribution. This sampling of the prior is then coupled with a Metropolis-like procedure to produce a sampling of the *a posteriori* distribution. Once we have this *a posteriori* distribution, which we regard as the ultimate solution of the inverse problem, it is up to us to pose proper questions and use the *a posteriori* distribution to answer them.

6 ACKNOWLEDGEMENTS

We thank Christophe Barnes, Jack Cohen, Philippe Doyen, and Martin Smith for useful discussions. The authors gratefully acknowledge the partial support of the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, Colorado School of Mines, the Shell Foundation, and the sponsors of the Geophysical Tomography Group, Institut de Physique du Globe de Paris, Université de Paris 6 et 7.

References

Duijndam, A. J. W. 1987. *Detailed Bayesian inversion of seismic data.* Ph.D. thesis, Technical University of Delft.

Jeffreys, R.C. 1983. *The logic of decision.* University of Chicago.

Mosegaard, K., & Tarantola, A. 1994. Monte Carlo sampling of solutions to inverse problems. *JGR: submitted.*

Press, William H., Flannery, Brian P., Teukolsky, Saul A., & Vetterling, William T. 1986. *Numerical recipes.* Cambridge University Press.

Pugachev, V. S. 1965. *Theory of random functions and its application to control problems.* Pergamon.

Robinson, E. A. 1967. *Multichannel Time Series Analysis with Digital Computer Programs.* Holden-Day.

Tarantola, A. 1987. *Inverse problem theory.* Elsevier.

APPENDIX A: THE BAYESIAN POSTERIOR PROBABILITY

As we have argued, the posterior probability density on the space of models must be the product of two terms: a term which involves the a priori probability on the space of models and a term which measures the extent of data fit

$$\sigma(\mathbf{d}) = k\rho(\mathbf{m})\ L(\mathbf{m}). \tag{A1}$$

L is called the likelihood function and depends implicitly on the data.

We now show how Equation (A1) follows logically from Bayes' theorem provided we generalize our notion of "data" to allow for the possibility that the data might be specified by probability distributions (Tarantola, 1987). To do so we make use of an idea due to Jeffreys (1983) (as described in (Duijndam, 1987)). We begin by using the notation common amongst Bayesians, then we show how this relates to the more standard inverse-theoretic notation in Tarantola (1987).

In this approach we assume that we have some prior joint distribution $p_0(\mathbf{m}, \mathbf{d})$. Further, we suppose that as the result of some observation, the marginal pdf of $\mathbf{d}$ changes to $p_1(\mathbf{d})$. We regard $p_1(\mathbf{d})$ as being the "data" in the sense that we often know the data only as a distribution, not exact numbers. In the special case where the data are exactly known, $p_1(\mathbf{d})$ reduces to a delta function $\delta(\mathbf{d} - \mathbf{d}_{obs})$.

How do we use this new information in the solution of the inverse problem? The answer is based upon the following assumption: whereas the information on $\mathbf{d}$ has changed as a result of the experiment, there is no reason to think that the conditional degree of belief of $\mathbf{m}$ on $\mathbf{d}$ has. I.e.,

$$p_1(\mathbf{m}|\mathbf{d}) = p_0(\mathbf{m}|\mathbf{d}). \tag{A2}$$

From this one can derive the posterior marginal $p_1(\mathbf{m})$:

$$\begin{aligned}
p_1(\mathbf{m}) &\equiv \int_D p_1(\mathbf{m},\mathbf{d})\, d\mathbf{d} && \text{(A3)}\\
&= \int_D p_1(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} && \text{(A4)}\\
&= \int_D p_0(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} && \text{(A5)}\\
&= \int_D \frac{p_0(\mathbf{d}|\mathbf{m}) p_0(\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d} && \text{(A6)}\\
&= p_0(\mathbf{m}) \int_D \frac{p_0(\mathbf{d}|\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d}, && \text{(A7)}
\end{aligned}$$

where D denotes the data space.

Switching now to the inverse-theoretic notation, let us regard $p_0(\mathbf{d})$ as being the non-informative prior distribution on data $\mu_D(\mathbf{d})$: this is what we know about the data **before** we've actually done this particular experiment. Further, we identify $p_1(\mathbf{m})$ as the posterior distribution on the space of models, $p_1(\mathbf{d})$ as the data errors, $p_0(\mathbf{m})$ as the prior distribution on the space of models, and $p_0(\mathbf{d}|\mathbf{m})$ as the modeling errors:

$$p_1(\mathbf{m}) \equiv \sigma(\mathbf{m})$$

$$p_1(\mathbf{d}) \equiv \rho_D(\mathbf{d})$$

$$p_0(\mathbf{m}) \equiv \rho_M(\mathbf{m})$$

$$p_0(\mathbf{d}|\mathbf{m}) \equiv \Theta(\mathbf{d}|\mathbf{m}),$$

then we arrive at precisely Equation (1.65) of Tarantola (1987)

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \int_D \frac{\rho_D(\mathbf{d})\Theta(\mathbf{d}|\mathbf{m})}{\mu_D(\mathbf{d})}\, d\mathbf{d} \tag{A8}$$

An important special case occurs when the modeling errors are negligible, i.e., we have a perfect theory. Then the conditional distribution $\Theta(\mathbf{d}|\mathbf{m})$ reduces to a delta function $\delta(\mathbf{d} - g(\mathbf{m}))$ where g is the forward operator. In this case, the posterior is simply

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \left[\frac{\rho_D(\mathbf{d})}{\mu_D(\mathbf{d})}\right]_{\mathbf{d}=g(\mathbf{m})} . \tag{A9}$$

APPENDIX B: ELEMENTS OF RANDOM FIELDS

Here we set forth the basic notations having to do with random functions. This section is an adaptation of Pugachev (1965). We will consider random fields in (3-D) physical space. Adaptation to other sorts of fields is straightforward.

$\Xi(\mathbf{r})$ will denote a random field. A realization of the random field will be denoted by $\xi(\mathbf{r})$. We may think of ξ as a physical parameter defined at every point of the space (like the velocity of seismic waves, the temperature, etc.). We will work inside a volume $\mathcal{V}$.

B1 n-dimensional joint probability densities

Let $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ be a set of n points inside $\mathcal{V}$. An expression like

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$$

will denote the n-dimensional (joint) probability density for the values $t(\mathbf{r}_1), t(\mathbf{r}_2), \ldots, t(\mathbf{r}_n)$. The notation $f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ may seem complicated, but it just indicates that for every different set of points $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ we have a possibly different probability density.

The random field $T(\mathbf{r})$ is completely characterized if, for any set of n points inside $\mathcal{V}$, the joint probability density $f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ is defined, and this for any value of n.

B2 Marginal and conditional probability

The definitions for marginal and conditional volumetric probabilities do not pose any special difficulty. As notations rapidly become intricate, let us only give the corresponding definitions for some particular cases, the generalization being straightforward.

If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *marginal* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$ is defined by

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)\,, \tag{A10}$$

where $dL(\xi_3)$ is the (1-D) volume element (it may be $d\xi_3$ or something different).

Let us now turn now to the illustration of the definition of conditional probability. If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random

field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *conditional* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$, given that the random field takes the value ξ_3 at point $\mathbf{r}_3$, is defined by

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{\int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)} , \tag{A11}$$

i.e., using equation A10,

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)} . \tag{A12}$$

Of particular interest will be the one- and the two-dimensional probability densities, denoted respectively $f_1(t; \mathbf{r})$ and $f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)$.

B3 Random fields defined by low order probability densities

First example: Independently distributed variables.

If

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) =$$
$$g(\xi_1, \mathbf{r}_1)\ h(\xi_2, \mathbf{r}_2) \dots i(\xi_n, \mathbf{r}_n) , \tag{A13}$$

we say that the random field has independently distributed variables.

Second example: Markov random field.

For a Markov process,

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) =$$
$$f_2(\xi_n; \mathbf{r}_n | \xi_{n-1}; \mathbf{r}_{n-1})\ f_2(\xi_{n-1}; \mathbf{r}_{n-1} | \xi_{n-2}; \mathbf{r}_{n-2}) \tag{A14}$$
$$\dots f_2(\xi_2; \mathbf{r}_2 | \xi_1; \mathbf{r}_1)\ f_1(\xi_1; \mathbf{r}_1) ,$$

where the vertical bar denotes conditional probability. This means that the value at a given point depends only on the value at the previous point. As

$$f_2(\xi_i; \mathbf{r}_i | \xi_{i-1}; \mathbf{r}_{i-1}) = \frac{f_2(\xi_{i-1}, \xi_i; \mathbf{r}_{i-1}, \mathbf{r}_i)}{f_1(\xi_{i-1}; \mathbf{r}_{i-1})} , \tag{A15}$$

we obtain

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) = \tag{A16}$$
$$\frac{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) \dots f_2(\xi_{n-1}, \xi_n; \mathbf{r}_{n-1}, \mathbf{r}_n)}{f_1(\xi_2; \mathbf{r}_2)\ f_1(\xi_3; \mathbf{r}_3) \dots f_1(\xi_{n-1}; \mathbf{r}_{n-1})} .$$

This equation characterizes a Markov random field in all generality. It means that the random

field is completely characterized by 1-D and 2-D probability densities (defined at adjacent points).

Third example: Gaussian random field.

For a Gaussian random field, if we know the 2-D distributions, we know all the means and all the covariances, so we also know the n-dimensional distribution. It can be shown that a Gaussian process with exponential covariance is Markovian.

B4 Uniform random fields

A random field is uniform (i.e., stationary) in the strong sense if for any $\mathbf{r}_0$,

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_0 + \mathbf{r}_1, \mathbf{r}_0 + \mathbf{r}_2, \ldots, \mathbf{r}_0 + \mathbf{r}_n) \ .$$

Taking

$$\mathbf{r}_0 = -\mathbf{r}_1$$

gives then

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{0}, \mathbf{r}_2 - \mathbf{r}_1, \ldots, \mathbf{r}_n - \mathbf{r}_1) \ .$$

A distribution is said to be uniform in the weak sense if this expression holds only for $n = 1$ and $n = 2$.

As the random field is defined over the physical space, we prefer the term *uniform* to characterize what, in random fields defined over a time variable, is called *stationary*. This is entirely a question of nomenclature; we regard the terms as being interchangeable.

Example:

For the two-dimensional distribution,

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \Psi_2(\xi_1, \xi_2; \Delta\mathbf{r}) \ ,$$

with

$$\Delta\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1 \ .$$

Example:

For the three-dimensional distribution,

$$f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \Psi_3(\xi_1, \xi_2, \xi_3; \Delta\mathbf{r}_1, \Delta\mathbf{r}_2) \ .$$

with

$$\Delta\mathbf{r}_1 = \mathbf{r}_2 - \mathbf{r}_1$$

and

$$\Delta\mathbf{r}_2 = \mathbf{r}_3 - \mathbf{r}_1 \ .$$

Essentially a uniform random process is one whose properties do not change with space (or time if that is the independent variable).

APPENDIX C: THE KOLMOGOROV-SMIRNOV TEST

This brief discussion is taken directly from *Numerical Recipes* (1986). The two-sample Kolmogorov-Smirnov statistic tests the null hypothesis that two data sets are drawn from the same distribution. It is based on a comparison of the cumulative distribution functions (CDF) of the two data sets. One can imagine any number of comparisons between the two CDFs. K-S represents an especially simple one: it is defined as the maximum value of the absolute difference between the two CDFs. In symbols, the K-S statistic D is given by

$$D = \max_{-\infty < x < \infty} |S_{N_1}(x) - S_{N_2}(x)|$$

where $S_{N_1}(x)$ and $S_{N_2}(x)$ are the approximate CDFs for the two data sets. The key point, however, is that the distribution function for the K-S statistic itself (for the null-hypothesis that the data sets are drawn from the same distribution) can be calulated approximately. The significance function Q for this test is given by the following approximation:

$$Q_{K-S}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}.$$

A plot of this significance function is given in Figure A1.

For the two-sample test, the significance level for an observed value of D (as a disproof of the null-hypothesis) is given approximately by

$$\text{Probability}(D > \text{observed}) = Q_{K-S}\left(\sqrt{\frac{N_1 N_2}{N_1 + N_2}}D\right)$$

where N_2 and N_2 are the numbers of samples in the two data sets.

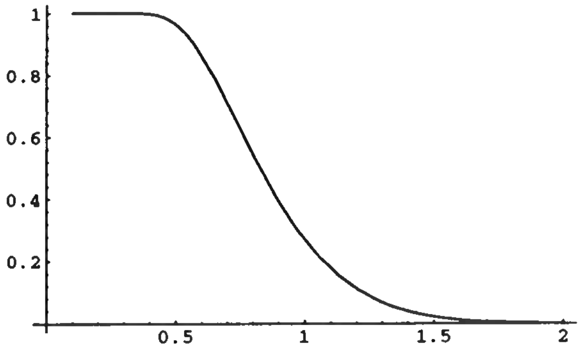

Figure A1. Kolmogorov-Smirnov significance function.

An Example of Geologic Prior Information in a Bayesian Seismic Inverse Calculation

John A. Scales

Center for Wave Phenomena, Colorado School of Mines

Albert Tarantola

Institut de Physique du Globe de Paris, Université de Paris 6 et 7

ABSTRACT

All inverse problems require the specification of *a priori* information on the space of models. Typically, it is assumed that the *a priori* uncertainties can be described using Gaussian probabilities, whether explicitly through the use of some *a priori* covariances, or implicitly, with "regularized least squares". Our goal is to try to get beyond these assumptions and let realistic prior information speak for itself. To that end we show here a simple case study in the use of geologic information for seismic inversion. We estimate nonparametrically the Bayesian *a priori* distribution for layered earth models directly from a P-wave sonic log and give a numerical procedure for randomly sampling earth models following this distribution. This procedure is then modified in order to obtain earth models that sample the *a posteriori* distribution, which we regard as a complete solution of the inverse problem. Although the scope of this calculation is limited, the methods employed have a broad utility.

Key words: *a priori* information, inverse problems, Bayesian inference

1 INTRODUCTION

All inverse calculations must balance the extent to which models fit data with the extent to which they conform to our prejudices as to what makes a model good or bad. In some cases it is possible to find completely unrealistic models which nevertheless fit the data. The *a priori* information we use to judge the reasonableness of models comes in many forms, from the subjective wisdom of experts to quantitative data. By incorporating this prior

information into Bayesian probability distributions, we can tackle, in principle, completely general kinds of inverse problems without relying on the particular form of the distribution—as, for example, least squares makes Gaussian assumptions. In this paper we give a concrete example of the nonparametric estimation of a Bayesian prior for earth models derived from a P-wave sonic log. The ultimate goal is to do the seismic inverse problem for surface seismic data and, using the information derived from the sonic log, to constrain, in a Bayesian sense, the inferences we draw from the seismic data. We end up with an algorithm for sampling the *a priori* distribution associated with the well log. To solve the full inverse problem, i.e., to sample the Bayesian *a posteriori* distribution we must, in effect, select the models sampled from the prior distribution by how well they fit the data. We use the Metropolis-like procedure proposed by Mosegaard & Tarantola (1994). This procedure is guaranteed to converge to the posterior distribution.

1.1 Probabilistic Formulation of the Inverse Problem

Assume we are analyzing some physical system (in geophysics this is usually the Earth) that is described by some parameters. Any particular value of the parameters describing the system defines a *model* of the system. A generic model is represented by the symbol $\mathbf{m}$. The collection of all possible models defines the *model space.*

Probabilistic formulations of inverse theory consider probability densities in the model space. Two probability distributions are fundamental ingredients to the theory.

One probability distribution describes the *a priori* information we may have on the parameters describing our model. Here, *a priori* means information that is independent of the data set to be used to refine this information. This *a priori* probability density is denoted $\rho(\mathbf{m})$. The purpose of this paper is to give an example of its definition.

If an experiment produces some data, then a "likelihood function" has to be introduced (Tarantola, 1987) that measures the "degree of fit" between the data predicted by the model (using some physical theory) and the actually observed data. This likelihood function is denoted $L(\mathbf{m})$ and the most popular example is

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{p}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\mathrm{obs}}|^p}{\sigma_i^p}\right) , \tag{1}$$

where, if d_i^{obs} represents the *observed value* for the i-th datum, $g_i(\mathbf{m})$ represents the cal-

culated value corresponding to the model $\mathbf{m}$, and σ_i is an estimation of the uncertainty attached to the i-th observation. For $p = 2$ we have a Gaussian probability density (for independent uncertainties), while for $p = 1$ we have a Laplacian (double exponential) probability density. Equation (1) is only given as an example, and much more realistic expressions may be used that describe better experimental uncertainties, but this is outside the scope of this paper.

Once the probability density $\rho(\mathbf{m})$ and the likelihood function $L(\mathbf{m})$ have been defined, describing respectively the *a priori* information we have on model parameters and the likelihood of a model, the resulting *a posteriori* probability density is given by the expression (see Appendix A for a derivation)

$$\sigma(\mathbf{m}) = k\, \rho(\mathbf{m})\, L(\mathbf{m})\,, \tag{2}$$

where k is a normalization constant. The probability density $\sigma(\mathbf{m})$ combines the *a priori* information with the information coming from observations. The more the *a posteriori* uncertainties on model parameters [as described by $\sigma(\mathbf{m})$] have been reduced compared to the *a priori* uncertainties [as described by $\rho(\mathbf{m})$] the more successful the “inversion” of the data has been.

We can use the function $\sigma(\mathbf{m})$ to find out the probability that the true model is inside a particular region of the model space by integrating $\sigma(\mathbf{m})$ over the particular region. Neglecting the presence of $\rho(\mathbf{m})$ in Equation (2) and judging models only on how well they fit the data is called maximum likelihood estimation. On the other hand, neglecting $L(\mathbf{m})$ and sampling models only according to some prior distribution is akin to geostatistical simulation. The complete solution of the inverse problem combines these two features: sampling models according to a prior distribution and selecting them according to how well they fit the data.

With that brief introduction to the theory, we turn now to a simple example of the probabilistic approach to using prior information. For more details on the theoretical aspects of the problem see Tarantola (1987) and Mosegaard and Tarantola (1994). We will consider the problem of inverting reflection seismic data for subsurface elastic properties in an area where we have a single P-wave sonic log. We regard the sonic log as containing important information about the geology of earth models in the area even though it consists of in-situ measurements made on a much finer scale than the seismic wavelength. The question is, how

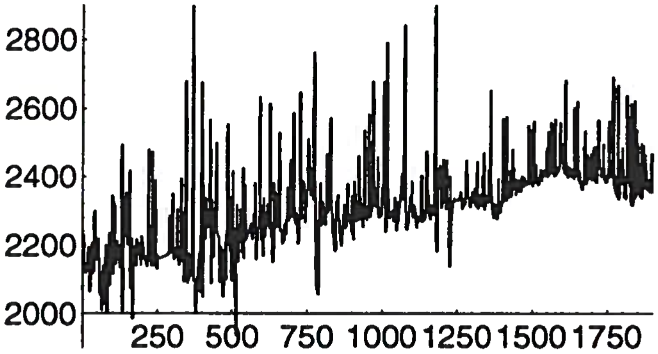

Figure 1. P-wave sonic log. The numbers on the abscissa refer to the samples, which were recorded every 30 cm. The wave speeds are in m/s.

to use this information? We will proceed by extracting the statistical properties from the log and then generating pseudo-random logs with these same statistical properties. By definition then, all of these pseudo-random models are *a priori* realistic in a geologic sense: this is how we sample the prior probability. Once this problem is solved, it is simple (although perhaps expensive) to use the likelihood function to sample the posterior probability.

2 CASE STUDY: EXTRACTING THE PRIOR FROM A SONIC LOG

Figure 1 shows an example of a P-wave sonic log. The wave speeds are in m/s and the samples were recorded every 30 cm. Our point of view is that the log represents a combination of gradual, deterministic processes and essentially random fluctuations. Thus if we subtract the smooth trend of the log (shown in Figure 2) from the original data, we will be left with an essentially random process, as shown in Figure 3. We will regard the trend of the log as being known exactly *a priori*. It could be asserted on geologic grounds or determined by some other geophysical technique such as travel time inversion. In any case, we will focus our attention on the fluctuating part of the log, which we regard as being a realization of a stationary stochastic process whose properties are to be determined. The assumption of stationarity can be relaxed, for example, by dividing the log into pieces associated with time or geologic horizons.

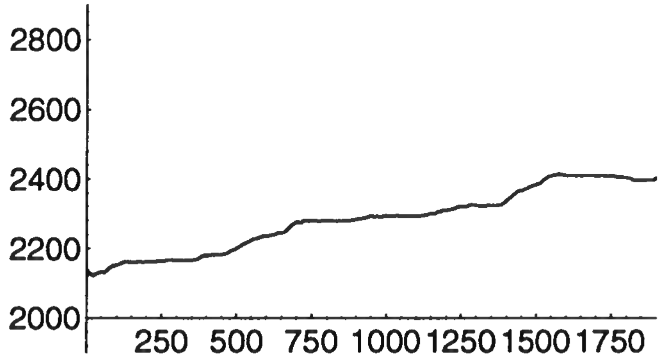

Figure 2. Trend of the log obtained with a sliding 150 sample alpha-trimmed mean filter.

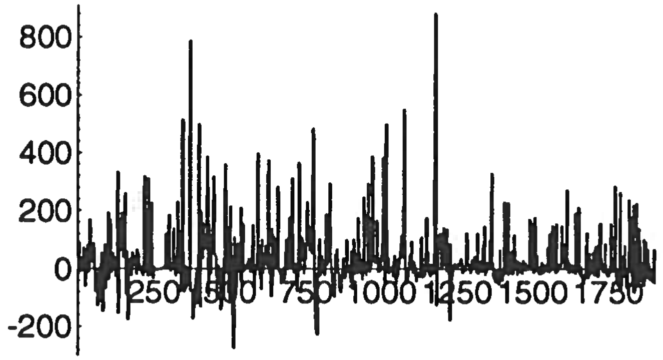

Figure 3. Fluctuating part of the log obtained by subtracting the trend from the original data.

2.1 Random Processes

Some basic principles of random functions are set forth in Appendix B; here we plan to keep the discussion as simple as possible. Consider a field $\xi(\mathbf{r})$ defined at every point $\mathbf{r}$ of the space. If what we have is, in fact, a *random field*, this means that at every point $\mathbf{r}$ of the space we have a random variable $\Xi(\mathbf{r})$. The notation

$$f_n(\xi_1, \xi_2, \ldots, \xi_m; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m)$$

represents the $n-$dimensional (joint) probability density for the random variables

$$\Xi(\mathbf{r}_1), \Xi(\mathbf{r}_2), \ldots, \Xi(\mathbf{r}_m).$$

To describe in all generality a random function requires, for any m, and for any points $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$, to give the $m-$dimensional joint probability density

$$f_m(\xi_1, \xi_2, \ldots, \xi_m; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m).$$

If the space has been discretized using, say n points, then we have only n random variables, and the most general case needs, at most an $n-$dimensional probability density.

Practically, we must rely on the fact that the $n-$dimensional distribution can be expressed in terms of the marginal distributions of the process. For example, the three dimensional distribution $f(x,y,z)$ can be written as $f(x|y,z)f(y,z)$, which can in turn be written as $f(x|yz)f(y|z)f(z)$ and so on by induction for the general n-dimensional distribution. For example, if a process is Markovian, this characterization in terms of marginals is especially simple since

$$f_n(\xi_1,\xi_2,\ldots,\xi_n;\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n) = \frac{f_2(\xi_1,\xi_2;\mathbf{r}_1,\mathbf{r}_2)\ldots f_2(\xi_{n-1},\xi_n;\mathbf{r}_{n-1},\mathbf{r}_n)}{f_1(\xi_2;\mathbf{r}_2)\ f_1(\xi_3;\mathbf{r}_3)\ldots f_1(\xi_{n-1};\mathbf{r}_{n-1})} \tag{3}$$

where f_1 and f_2 are, respectively the one- and two-dimensional marginal distributions associated with f_n. This brings out clearly the fact that for Markov fields the probability of a point depends only on the previous point.

2.2 Estimating the Marginal Distributions

For more general processes than Markovian ones, we need more of the marginal distributions, not just the one- and two-dimensional ones. But, in any case, the marginals can be estimated by making histograms of the data. The process of making these histograms begins by quantizing the log into some number of equal-length velocity intervals. Figure 4 shows how this is done.

Once the log is quantized, we just start counting. For example, to compute the two-dimensional marginals of length one (i.e., the probability density for the velocity values at two consecutive points) we count all the transitions from an interval i to an interval j which occur at adjacent sites along the log (note that we are using here the stationarity assumption). To compute all the two-dimensional distributions we must compute all transitions from an interval i to an interval j over non-adjacent sites as well. Figures 5 and 6 show second order histograms for the P-wave sonic log, the first for transitions of length 1, the second for transitions of length 5.

Clearly it would be impractical to attempt to compute all the marginal distributions for a process of several thousands of points. But at the same time we are not yet willing to restrict ourselves to parametric assumptions, such as Gaussian statistics, that would allow us to reduce the computational burden. Our compromise is to calculate enough of

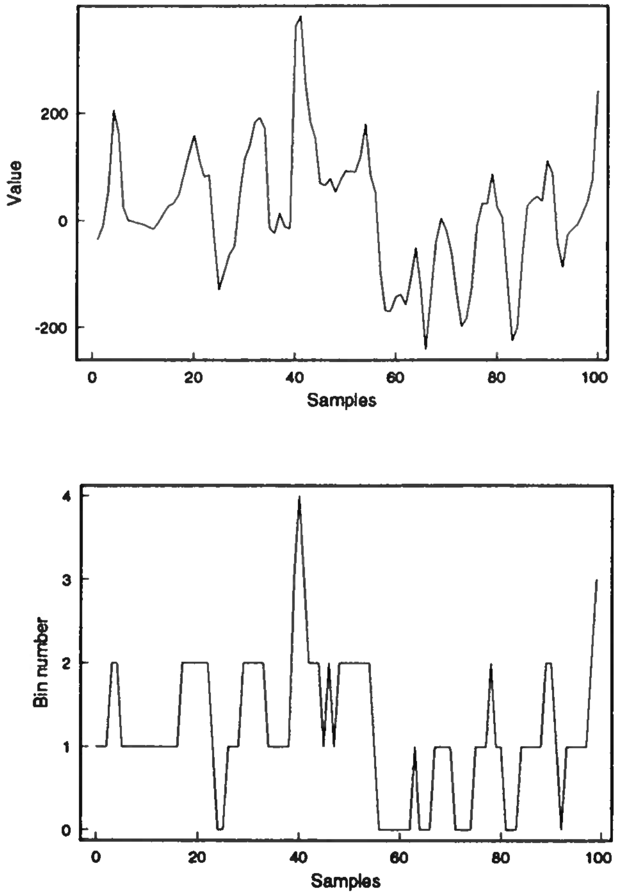

Figure 4. **The data are quantized into a number of equal-length velocity intervals so that histograms can be made. Here we show the first 100 samples of the de-meaned sonic log and a toy quantization into 5 intervals. In practice we use a much larger number of intervals, from 100 to 500.**

the marginals so as to convince ourselves that we have indeed captured the essence of the process. To quantify this, we use the Kolmogorov-Smirnov two-sample test and compare the original data with pseudo-random simulations calculated with a given number of marginals. The K-S test (described in Appendix C) involves comparing the cumulative distribution function of two different realizations (in this case the data and the simulation), the so-called null hypothesis being that the two were drawn from the same distribution. A high value of the K-S statistic says that the two realizations were very likely drawn from the same underlying distribution; this gives us confidence that the pseudo-random realizations have captured the essence of the data.

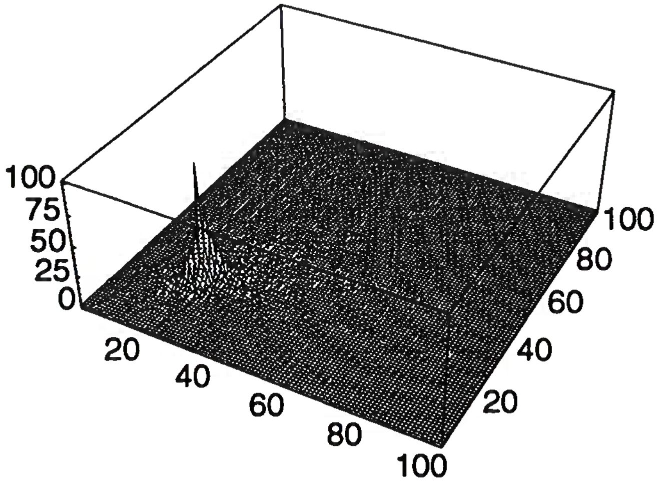

Figure 5. Histograms of two-point transitions of length 1 (i.e., pairwise transitions for adjacent sites along the log) for the P-wave sonic log. If for a given value of x and y axes, say (i, j), the count is n, that means that there are n sites along the log where the site itself and its neighbor have the quantized values i and j respectively.

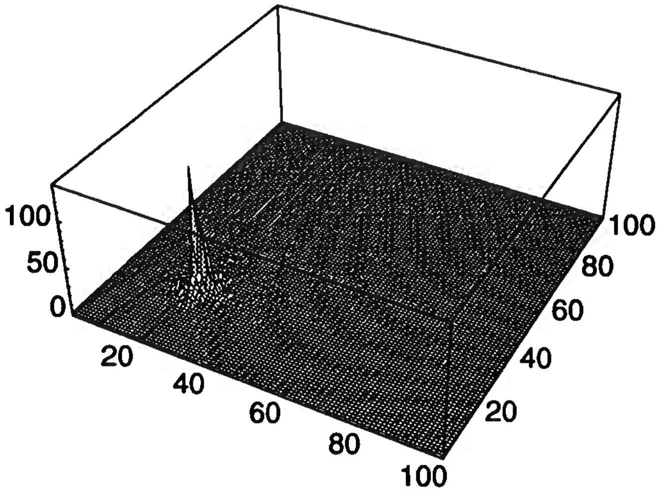

Figure 6. Histograms of two-point transitions of length 5 for the P-wave sonic log. These correspond to transitions from a site k to a site $k + 5$ on the log.

2.3 Sampling the Prior

Once we have computed a sufficient number of the marginals we must be able to sample the prior distribution. To *sample* a probability density means to produce a pseudo random "sample" (or *realization*) of the random variable. By definition, the probability for the produced "point" to belong to any volume of the space must equal the probability of the volume. The algorithms used to sample a probability produce, very often, consecutive samples that are

not independent (think of Brownian motion of a particle: if, in the long term, the particle may be at any point of the space with uniform probability, at close times the particle will be at close points). In such a situation, to obtain independent samples we simply have to "wait" until the algorithm loses the memory of the previous point.

Typically, a sampling algorithm produces a *test point* using some "background probability" (usually the uniform one) and then, uses some random criterion for "accepting" the test point or for "dropping" it. Obviously, if the background probability is very close to the probability we wish to sample, the test points will be often accepted. In this case, one says that the algorithm uses an "importance sampling method". Any nontrivial sampling algorithm, in fact, uses importance sampling, and this will be the case for the examples below.

To do our sampling we use the histograms directly. Suppose, for example, that we needed only the two-point histograms of length 1 (as would be the case if the underlying process were Markovian, since then the probability of a given point depends only on the probability of the previous point). Figure 7 shows a cartoon of this case. The values of the observed log are $\{\xi(\mathbf{r}_i)\}_{i=1}^{N}$. Let us refer to the values of a pseudo-random simulation as $\{\xi'(\mathbf{r}_i)\}_{i=1}^{N}$. We begin the construction of the simulation by assigning some value to $\xi'(\mathbf{r}_1)$, for example, this could be a likely value for the log based on the one-dimensional marginal. Now all we have to do is select $\xi'(\mathbf{r}_2)$ according to the histogram of length 1 transitions. Suppose $\xi'(\mathbf{r}_1) = 4$, as shown in Figure 7. Then the slice through the histogram associated with $i = 4$ is the conditional probability distribution of making the length 1 transition $4 \longrightarrow j$ for any j. To assign the value $\xi'(\mathbf{r}_2)$ in accordance with this conditional probability, we make, in effect, a weighted roulette wheel with a sector for each interval number. The size of each interval's sector is in proportion to its probability from the histogram. Then by spinning the roulette wheel we will select $\xi'(\mathbf{r}_2)$ with the appropriate probability. For Markov processes we would proceed in this way right down the log, filling in one value $\xi'(\mathbf{r}_i)$ after another.

But suppose our data are not Markovian? Suppose, for example, that the results of the K-S test say that we must include all two-point transitions out to some length ℓ. In this case we begin as before, filling out the first and second sites along the log. But now, to fill out the third site, we choose the roulette wheel associated with transitions of length two conditioned on the first point $\xi'(\mathbf{r}_1)$. Then we choose the roulette wheel associated with

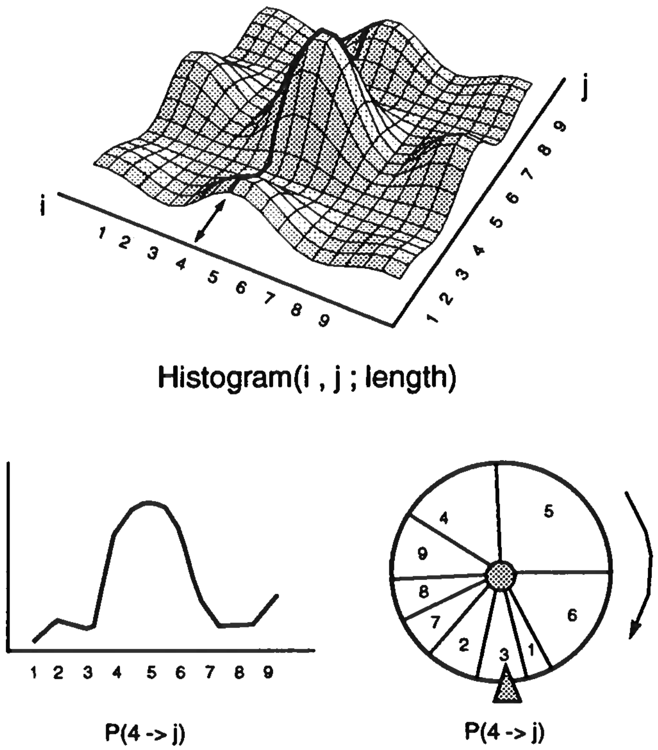

Figure 7. A weighted roulette wheel is used to sample the histograms. If the current site on the pseudo-random log has a value i and we need a length ℓ transition to fill in the next slot, we first select the length ℓ histogram. Then we extract the one-dimensional marginal associated with the value i. (In the example shown in the figure $i = 4$.) This gives us the observed probability of making a transition from a value i to all other values over a distance of ℓ sites.

length three transitions conditioned on $\xi'(\mathbf{r}_1)$, and so on until we have filled in the first $\ell + 1$ sites on the log. Then we start over again at the site $\ell + 2$. This procedure can be generalized to higher order histograms.

Figure 8 shows 15 such logs pseudo-randomly sampled from the second order histograms of the data in Figure 1. In this case we used all the histograms out to a length of 5 samples, which is estimated to be the average correlation length in the medium. Using histograms of this length we are able to routinely generate pseudo-random logs with a K-S statistic of better than 95%. In other words, the distributions associated with the two realizations, pseudo-random and data, are so close that K-S cannot tell whether they are the same or not. But it is important to emphasize, that the K-S test is used only as a “preprocessing” step. Once we have decided on the number of marginals to use, we sample them as described in the text. That means that in the course of the sampling, we may generate simulations with low *a priori* probability but these will be relatively rare occurrences.

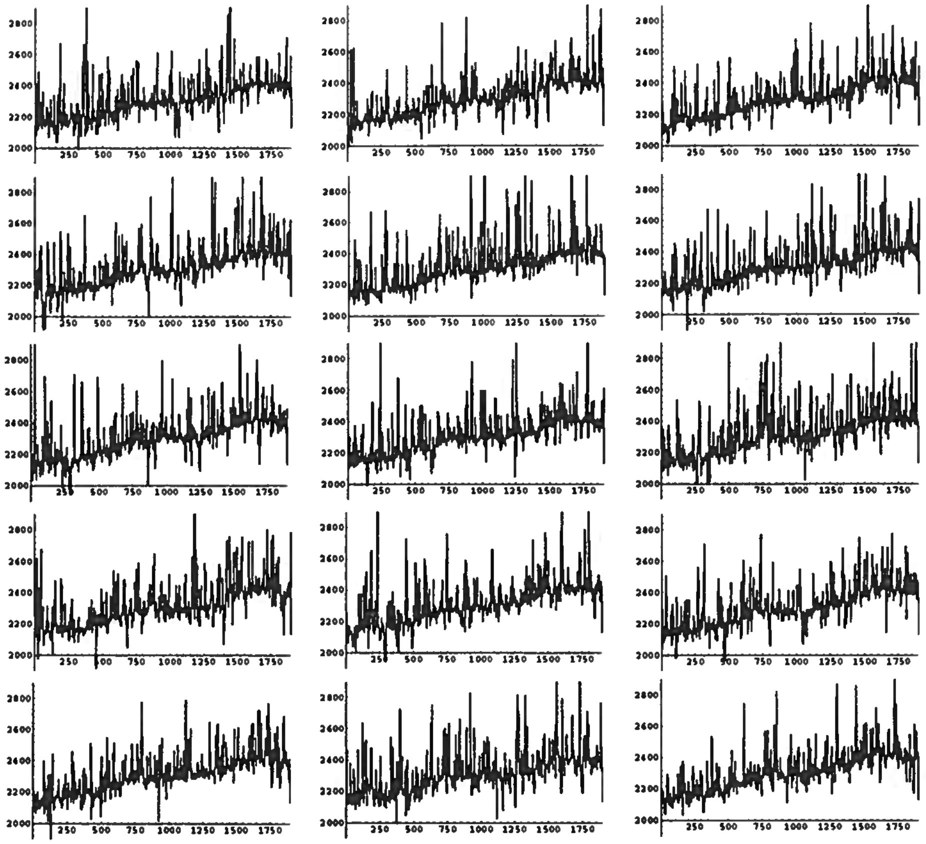

Figure 8. Pseudorandom logs obtained by sampling the *a priori* model distribution. Here we used all the second order histograms up to the average correlation length of the medium. Each of these pseudo-random logs is *a priori* feasible geologically. We have only to rank the models according to how well they fit the seismic data. We can imagine that these 15 models are 15 frames from a movie tour of the prior.

3 SOLVING THE INVERSE PROBLEM

Now that we have a procedure for sampling the prior, we can use these models to sample the posterior according to the algorithm of Mosegaard and Tarantola (1994). This involves accepting or rejecting models according to how well they fit the data, and represents a generalization of the usual Metropolis method for sampling the Gibbs-Boltzman distribution.

More precisely, let us assume that we are able to obtain samples $\mathrm{m}_1, \mathrm{m}_2, \ldots$, of the prior probability distribution $\rho(\mathrm{m})$. Then, if we wish to obtain samples of the posterior probability distribution $\sigma(\mathrm{m}) = k\, \rho(\mathrm{m})\, L(\mathrm{m})$ all that we need to do is to iterate the following procedure.

- Let m_i be the "current model."
- Use the rules that allow one to sample the prior probability distribution $\rho(\mathrm{m})$ to obtain a new model, say m'_i.

- If $L(\mathbf{m}'_i) \geq L(\mathbf{m}_i)$, take $\mathbf{m}'_i$ as new current model, i.e., make $\mathbf{m}_{i+1} = \mathbf{m}'_i$.
- If $L(\mathbf{m}'_i) < L(\mathbf{m}_i)$, then decide randomly to take the model $\mathbf{m}'_i$ as new current model, or to destroy it, with a probability of taking it as new current model equal to the ratio $L(\mathbf{m}'_i)/L(\mathbf{m}_i)$.

Mosegaard and Tarantola (1994) show that this procedure converges to the true *a posteriori* distribution.

The precise specification of L requires knowledge of the data and modeling uncertainties, as shown in Appendix A. We cannot say what it means to fit the data if we do not know the data uncertainites. This is obviously a crucial issue when inverting real data. But it is beyond the scope of this paper. So although we did not have to make a parametric assumption about the *a priori* distribution, we must make some choice for the data distribution. For this we make the simplest choice, gaussian, so that we can use a least squares criterion to measure data misfit. In other words, for purposes of this demonstration, we will generate synthetic data from the actual log and add a small amount of uncorrelated gaussian noise. Figure 9 shows the synthetic "data", a single zero-offset reflection seismogram, associated with the true model (top) and the same trace contaminated with a small amount (5%) uncorrelated gaussian noise. The seismograms are computed by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson (1967) and includes all multiples, but is limited to zero-offset, acoustic, marine data. Once we have the data trace, the likelihood function is the difference between this trace and the response of the sampled model, normalized by the known data variance

$$L(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\sum_i \frac{|g_i(\mathbf{m}) - d_i^{\mathrm{obs}}|^2}{\sigma^2}\right) \tag{4}$$

where $g_i(\mathbf{m})$ are the synthetic data associated with a sampled model $\mathbf{m}$, d_i^{obs} are the "observed" data, and σ is the standard deviation of the errors.

3.1 Sampling the Posterior

Figure 10 shows a selection of models sampled according to the generalized Metropolis procedure of Mosegaard and Tarantola (1994). We can think of these models as 15 frames from a movie tour of the posterior. But how do these samples of the posterior actually "solve" an/the inverse problem? It is up to us to pose questions and use the posterior to

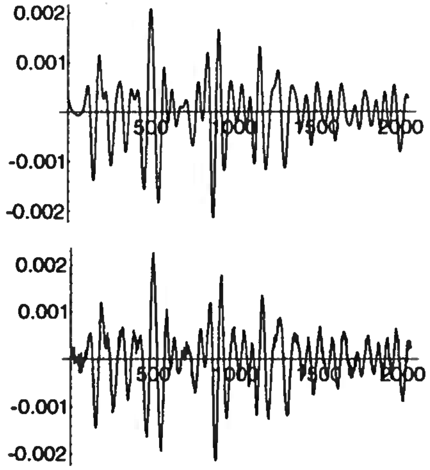

Figure 9. The top figure shows the "data" (i.e., model response) for the true log. The seismogram is obtained by convolving a 50 hz Ricker wavelet with the acoustic impulse response for the model. The algorithm for computing this impulse response is from Robinson's book (1967) and includes all multiples in a fixed time window, but is limited to zero-offset, acoustic, marine data. Below the noise free seismogram is the seismogram with 5% uncorrelated gaussian noise added. We take this to be the observed data for the inverse problem.

answer them probabilistically. A simple illustration of this is to investigate the extent to which a certain feature of our earth models is resolved by the seismic data. Figure 11 shows the distribution of P-wave reflectivity at two points along the log, one near the top of the log ($z = 10$, i.e., the tenth sample along the log) and one near the bottom ($z = 1800$). In both cases the posterior variance is visibly reduced relative to the prior variance. This is a quantitative measure of the extent to which the seismic data are able to resolve these parameters.

4 CAVEATS

There are a great many limitations to this work. Here we list some of the more obvious ones with evidence both for the prosecution (**con**) and the defense (**pro**).

con: The sample size is too small and therefore the *a posteriori* statistics unreliable.

pro: True. The calculations were done on a slow workstation. By moving to a faster workstation and optimizing the code, a factor of 10 speedup is easily obtainable. To move to 2-D, however, will require significantly greater computing resources.

con: The noise model is unrealistic.

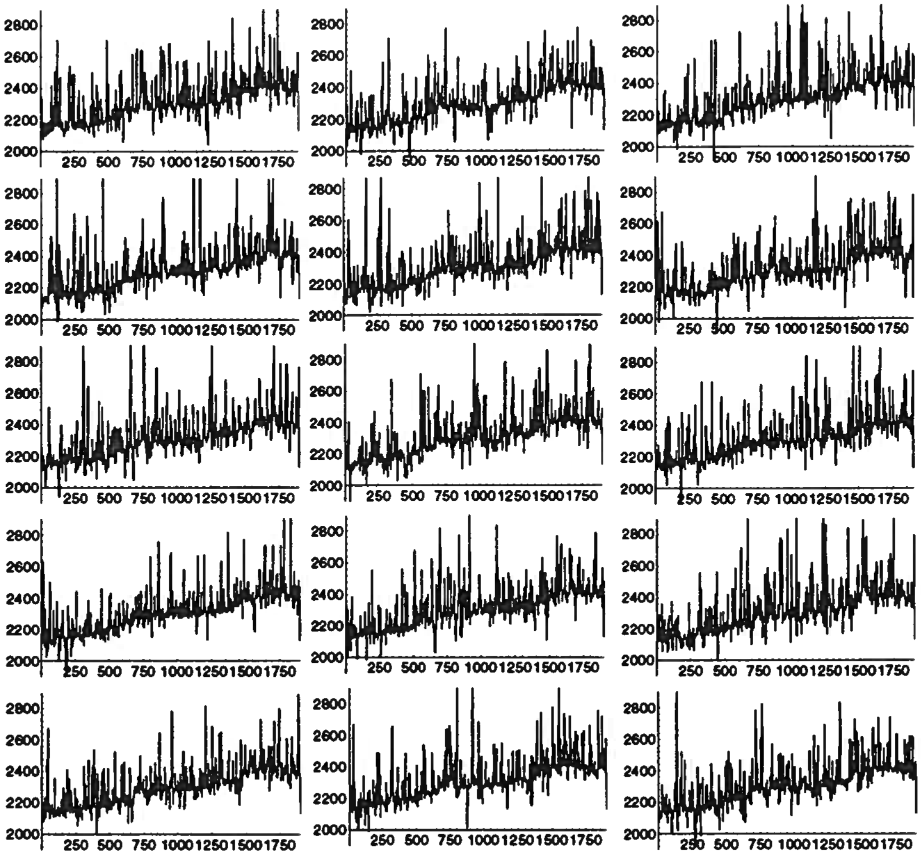

Figure 10. 1D earth models obtained by sampling the *a posteriori* distribution according to the Metropolis-like algorithm of Mosegaard and Tarantola (1994).

pro: True. Until we can estimate the noise directly from the data this will be a problem. In defense of our small *a priori* data error bars, we believe that the "noise" in exploration seismic data which can usefully be classified as gaussian (or some similar distribution) is very small. Far larger are the uncertainties due to modeling errors and, perhaps, coherent environmental noise, for which we have no realistic model.

con: None of the models in the posterior sample would generate seismograms that fit within the (tiny) error bars assigned to the data.

pro: True. This is a function of the small sample size. Should we allow the algorithm to run long enough, it will end by sampling a point inside this region and would remain there a long time, but this might take a prodigious number of iterations. On the other hand, the fact that the Monte Carlo scheme converges to the *a posteriori* distribution means that we can nevertheless compare the relative probability of different models. If the posterior variance of a parameter is smaller than the prior variance, then new information has been obtained,

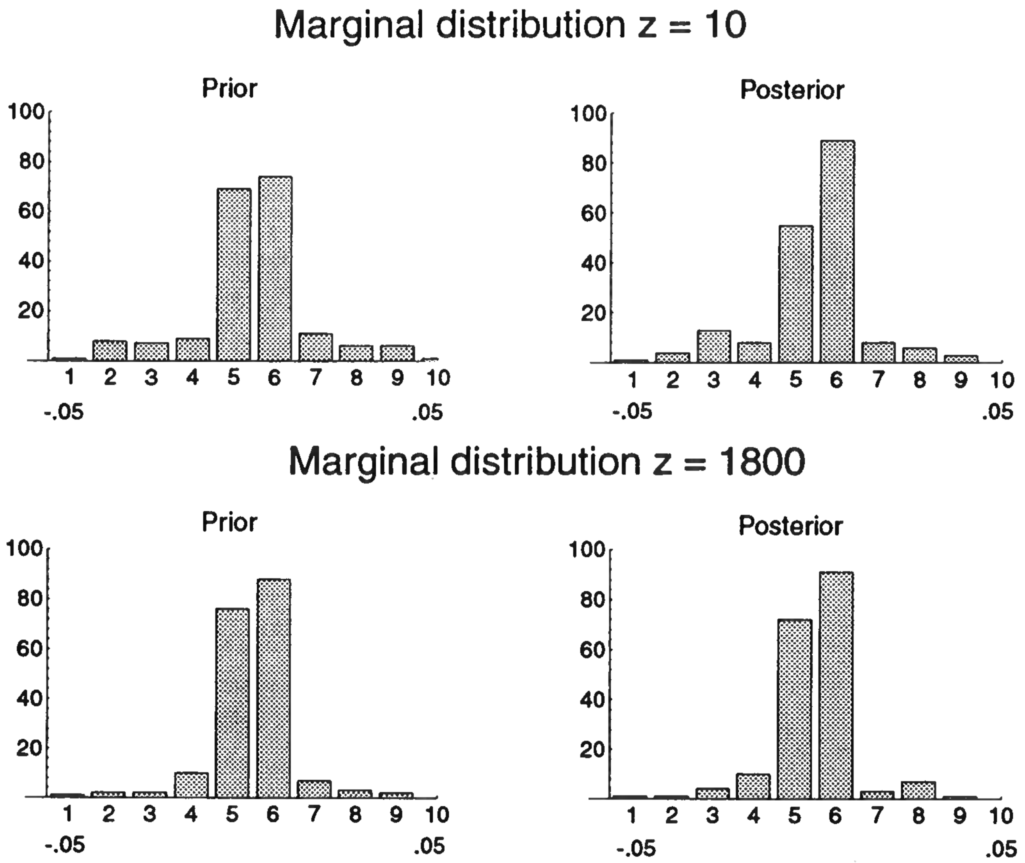

Figure 11. Prior versus posterior distribution of P-wave reflection coefficient at two points in the model. The z values shown refer to the sample number along the log. So these points are at extreme ends of the log. Near the upper surface, the posterior variance is noticeably smaller, indicating an increase in resolution. At the bottom of the log, however, the prior and posterior variances are virtually the same. So the seismic data have not resolved the deeper reflection coefficient at all.

even if the response of the posterior samples do not fit within the *a priori* error bars of the data.

A good simulation should answer positively to the following three questions:

(i) Have we chosen correctly the type of perturbations (of the current model)?

(ii) Is the size of the perturbations adapted to the size of the significant region of the model space?

(iii) Have we started at a point close enough to the region of significant probability so that we will enter it in a reasonable time?

At present, there remains a lot of work to be done in order to answer these questions rigorously.

5 CONCLUSIONS AND FUTURE WORK

We have shown a case study of the use of complex *a priori* information in a Bayesian inverse calculation. The problem we have examined is the inversion of reflection seismograms for 1D earth structure given a well log as *a priori* geologic information. We have developed a non-parametric technique for extracting the Bayesian *a priori* distribution from logs (or other similar data), as well as a method for sampling this *a priori* distribution. This algorithm allows us to generate as many pseudo-random earth models as we like, in accordance with the underlying *a priori* distribution. This sampling of the prior is then coupled with a Metropolis-like procedure to produce a sampling of the *a posteriori* distribution. Once we have this *a posteriori* distribution, which we regard as the ultimate solution of the inverse problem, it is up to us to pose proper questions and use the *a posteriori* distribution to answer them.

6 ACKNOWLEDGEMENTS

We thank Christophe Barnes, Jack Cohen, Philippe Doyen, and Martin Smith for useful discussions. The authors gratefully acknowledge the partial support of the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, Colorado School of Mines, the Shell Foundation, and the sponsors of the Geophysical Tomography Group, Institut de Physique du Globe de Paris, Université de Paris 6 et 7.

References

Duijndam, A. J. W. 1987. *Detailed Bayesian inversion of seismic data.* Ph.D. thesis, Technical University of Delft.

Jeffreys, R.C. 1983. *The logic of decision.* University of Chicago.

Mosegaard, K., & Tarantola, A. 1994. Monte Carlo sampling of solutions to inverse problems. *JGR: submitted.*

Press, William H., Flannery, Brian P., Teukolsky, Saul A., & Vetterling, William T. 1986. *Numerical recipes.* Cambridge University Press.

Pugachev, V. S. 1965. *Theory of random functions and its application to control problems.* Pergamon.

Robinson, E. A. 1967. *Multichannel Time Series Analysis with Digital Computer Programs.* Holden-Day.

Tarantola, A. 1987. *Inverse problem theory.* Elsevier.

APPENDIX A: THE BAYESIAN POSTERIOR PROBABILITY

As we have argued, the posterior probability density on the space of models must be the product of two terms: a term which involves the a priori probability on the space of models and a term which measures the extent of data fit

$$\sigma(\mathbf{d}) = k\rho(\mathbf{m})\, L(\mathbf{m}). \tag{A1}$$

L is called the likelihood function and depends implicitly on the data.

We now show how Equation (A1) follows logically from Bayes' theorem provided we generalize our notion of "data" to allow for the possibility that the data might be specified by probability distributions (Tarantola, 1987). To do so we make use of an idea due to Jeffreys (1983) (as described in (Duijndam, 1987)). We begin by using the notation common amongst Bayesians, then we show how this relates to the more standard inverse-theoretic notation in Tarantola (1987).

In this approach we assume that we have some prior joint distribution $p_0(\mathbf{m}, \mathbf{d})$. Further, we suppose that as the result of some observation, the marginal pdf of $\mathbf{d}$ changes to $p_1(\mathbf{d})$. We regard $p_1(\mathbf{d})$ as being the "data" in the sense that we often know the data only as a distribution, not exact numbers. In the special case where the data are exactly known, $p_1(\mathbf{d})$ reduces to a delta function $\delta(\mathbf{d} - \mathbf{d}_{obs})$.

How do we use this new information in the solution of the inverse problem? The answer is based upon the following assumption: whereas the information on $\mathbf{d}$ has changed as a result of the experiment, there is no reason to think that the conditional degree of belief of $\mathbf{m}$ on $\mathbf{d}$ has. I.e.,

$$p_1(\mathbf{m}|\mathbf{d}) = p_0(\mathbf{m}|\mathbf{d}). \tag{A2}$$

From this one can derive the posterior marginal $p_1(\mathbf{m})$:

$$\begin{aligned}
p_1(\mathbf{m}) &\equiv \int_D p_1(\mathbf{m},\mathbf{d})\, d\mathbf{d} &\text{(A3)}\\
&= \int_D p_1(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} &\text{(A4)}\\
&= \int_D p_0(\mathbf{m}|\mathbf{d}) p_1(\mathbf{d})\, d\mathbf{d} &\text{(A5)}\\
&= \int_D \frac{p_0(\mathbf{d}|\mathbf{m}) p_0(\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d} &\text{(A6)}\\
&= p_0(\mathbf{m}) \int_D \frac{p_0(\mathbf{d}|\mathbf{m})}{p_0(\mathbf{d})} p_1(\mathbf{d})\, d\mathbf{d}, &\text{(A7)}
\end{aligned}$$

where D denotes the data space.

Switching now to the inverse-theoretic notation, let us regard $p_0(\mathbf{d})$ as being the non-informative prior distribution on data $\mu_D(\mathbf{d})$: this is what we know about the data **before** we've actually done this particular experiment. Further, we identify $p_1(\mathbf{m})$ as the posterior distribution on the space of models, $p_1(\mathbf{d})$ as the data errors, $p_0(\mathbf{m})$ as the prior distribution on the space of models, and $p_0(\mathbf{d}|\mathbf{m})$ as the modeling errors:

$$p_1(\mathbf{m}) \equiv \sigma(\mathbf{m})$$

$$p_1(\mathbf{d}) \equiv \rho_D(\mathbf{d})$$

$$p_0(\mathbf{m}) \equiv \rho_M(\mathbf{m})$$

$$p_0(\mathbf{d}|\mathbf{m}) \equiv \Theta(\mathbf{d}|\mathbf{m}),$$

then we arrive at precisely Equation (1.65) of Tarantola (1987)

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \int_D \frac{\rho_D(\mathbf{d})\Theta(\mathbf{d}|\mathbf{m})}{\mu_D(\mathbf{d})}\, d\mathbf{d} \qquad \text{(A8)}$$

An important special case occurs when the modeling errors are negligible, i.e., we have a perfect theory. Then the conditional distribution $\Theta(\mathbf{d}|\mathbf{m})$ reduces to a delta function $\delta(\mathbf{d} - g(\mathbf{m}))$ where g is the forward operator. In this case, the posterior is simply

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \left[\frac{\rho_D(\mathbf{d})}{\mu_D(\mathbf{d})}\right]_{\mathbf{d}=g(\mathbf{m})} . \qquad \text{(A9)}$$

APPENDIX B: ELEMENTS OF RANDOM FIELDS

Here we set forth the basic notations having to do with random functions. This section is an adaptation of Pugachev (1965). We will consider random fields in (3-D) physical space. Adaptation to other sorts of fields is straightforward.

$\Xi(\mathbf{r})$ will denote a random field. A realization of the random field will be denoted by $\xi(\mathbf{r})$. We may think of ξ as a physical parameter defined at every point of the space (like the velocity of seismic waves, the temperature, etc.). We will work inside a volume $\mathcal{V}$.

B1 n-dimensional joint probability densities

Let $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ be a set of n points inside $\mathcal{V}$. An expression like

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$$

will denote the n-dimensional (joint) probability density for the values $t(\mathbf{r}_1), t(\mathbf{r}_2), \ldots, t(\mathbf{r}_n)$. The notation $f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ may seem complicated, but it just indicates that for every different set of points $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ we have a possibly different probability density.

The random field $T(\mathbf{r})$ is completely characterized if, for any set of n points inside $\mathcal{V}$, the joint probability density $f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ is defined, and this for any value of n.

B2 Marginal and conditional probability

The definitions for marginal and conditional volumetric probabilities do not pose any special difficulty. As notations rapidly become intricate, let us only give the corresponding definitions for some particular cases, the generalization being straightforward.

If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *marginal* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$ is defined by

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) \,, \tag{A10}$$

where $dL(\xi_3)$ is the (1-D) volume element (it may be $d\xi_3$ or something different).

Let us now turn now to the illustration of the definition of conditional probability. If $f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ is the 3-D joint volumetric probability for the values of the random

field at points $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ respectively, the *conditional* volumetric probability for the two points $\mathbf{r}_1$ and $\mathbf{r}_2$, given that the random field takes the value ξ_3 at point $\mathbf{r}_3$, is defined by

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{\int dL(\xi_3)\, f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)} , \tag{A11}$$

i.e., using equation A10,

$$f_3(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2 | \xi_3; \mathbf{r}_3) = \frac{f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)} . \tag{A12}$$

Of particular interest will be the one- and the two-dimensional probability densities, denoted respectively $f_1(t; \mathbf{r})$ and $f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2)$.

B3 Random fields defined by low order probability densities

First example: Independently distributed variables.

If

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) =$$

$$g(\xi_1, \mathbf{r}_1)\, h(\xi_2, \mathbf{r}_2) \dots i(\xi_n, \mathbf{r}_n) , \tag{A13}$$

we say that the random field has independently distributed variables.

Second example: Markov random field.

For a Markov process,

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) =$$

$$f_2(\xi_n; \mathbf{r}_n | \xi_{n-1}; \mathbf{r}_{n-1})\, f_2(\xi_{n-1}; \mathbf{r}_{n-1} | \xi_{n-2}; \mathbf{r}_{n-2}) \tag{A14}$$

$$\dots f_2(\xi_2; \mathbf{r}_2 | \xi_1; \mathbf{r}_1)\, f_1(\xi_1; \mathbf{r}_1) ,$$

where the vertical bar denotes conditional probability. This means that the value at a given point depends only on the value at the previous point. As

$$f_2(\xi_i; \mathbf{r}_i | \xi_{i-1}; \mathbf{r}_{i-1}) = \frac{f_2(\xi_{i-1}, \xi_i; \mathbf{r}_{i-1}, \mathbf{r}_i)}{f_1(\xi_{i-1}; \mathbf{r}_{i-1})} , \tag{A15}$$

we obtain

$$f_n(\xi_1, \xi_2, \dots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) = \tag{A16}$$

$$\frac{f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) \dots f_2(\xi_{n-1}, \xi_n; \mathbf{r}_{n-1}, \mathbf{r}_n)}{f_1(\xi_2; \mathbf{r}_2)\, f_1(\xi_3; \mathbf{r}_3) \dots f_1(\xi_{n-1}; \mathbf{r}_{n-1})} .$$

This equation characterizes a Markov random field in all generality. It means that the random

field is completely characterized by 1-D and 2-D probability densities (defined at adjacent points).

Third example: Gaussian random field.

For a Gaussian random field, if we know the 2-D distributions, we know all the means and all the covariances, so we also know the n-dimensional distribution. It can be shown that a Gaussian process with exponential covariance is Markovian.

B4 Uniform random fields

A random field is uniform (i.e., stationary) in the strong sense if for any $\mathbf{r}_0$,

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_0 + \mathbf{r}_1, \mathbf{r}_0 + \mathbf{r}_2, \ldots, \mathbf{r}_0 + \mathbf{r}_n) \ .$$

Taking

$$\mathbf{r}_0 = -\mathbf{r}_1$$

gives then

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) =$$

$$f_n(\xi_1, \xi_2, \ldots, \xi_n; \mathbf{0}, \mathbf{r}_2 - \mathbf{r}_1, \ldots, \mathbf{r}_n - \mathbf{r}_1) \ .$$

A distribution is said to be uniform in the weak sense if this expression holds only for $n = 1$ and $n = 2$.

As the random field is defined over the physical space, we prefer the term *uniform* to characterize what, in random fields defined over a time variable, is called *stationary*. This is entirely a question of nomenclature; we regard the terms as being interchangeable.

Example:

For the two-dimensional distribution,

$$f_2(\xi_1, \xi_2; \mathbf{r}_1, \mathbf{r}_2) = \Psi_2(\xi_1, \xi_2; \Delta\mathbf{r}) \ ,$$

with

$$\Delta\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1 \ .$$

Example:

For the three-dimensional distribution,

$$f_3(\xi_1, \xi_2, \xi_3; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \Psi_3(\xi_1, \xi_2, \xi_3; \Delta\mathbf{r}_1, \Delta\mathbf{r}_2) \ .$$

with

$$\Delta\mathbf{r}_1 = \mathbf{r}_2 - \mathbf{r}_1$$

and

$$\Delta\mathbf{r}_2 = \mathbf{r}_3 - \mathbf{r}_1 \ .$$

Essentially a uniform random process is one whose properties do not change with space (or time if that is the independent variable).

APPENDIX C: THE KOLMOGOROV-SMIRNOV TEST

This brief discussion is taken directly from *Numerical Recipes* (1986). The two-sample Kolmogorov-Smirnov statistic tests the null hypothesis that two data sets are drawn from the same distribution. It is based on a comparison of the cumulative distribution functions (CDF) of the two data sets. One can imagine any number of comparisons between the two CDFs. K-S represents an especially simple one: it is defined as the maximum value of the absolute difference between the two CDFs. In symbols, the K-S statistic D is given by

$$D = \max_{-\infty < x < \infty} |S_{N_1}(x) - S_{N_2}(x)|$$

where $S_{N_1}(x)$ and $S_{N_2}(x)$ are the approximate CDFs for the two data sets. The key point, however, is that the distribution function for the K-S statistic itself (for the null-hypothesis that the data sets are drawn from the same distribution) can be calulated approximately. The significance function Q for this test is given by the following approximation:

$$Q_{K-S}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}.$$

A plot of this significance function is given in Figure A1.

For the two-sample test, the significance level for an observed value of D (as a disproof of the null-hypothesis) is given approximately by

$$\text{Probability}(D > \text{observed}) = Q_{K-S}\left(\sqrt{\frac{N_1 N_2}{N_1 + N_2}} D\right)$$

where N_2 and N_2 are the numbers of samples in the two data sets.

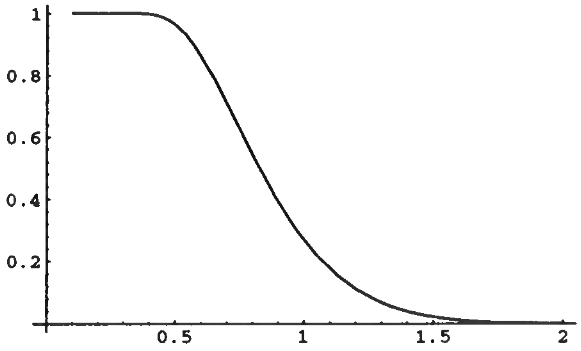

Figure A1. Kolmogorov-Smirnov significance function.