Monte Carlo Analysis of Inverse Problems

Klaus Mosegaard

Monte Carlo Analysis of Inverse Problems

Cover: Part of a painting by Henning Mosegaard
Print: J & R Frydenberg A/S – København N.

Printed in Denmark 2006
ISBN 87-991228-0-4

Denne afhandling er i forbindelse med de nedenfor anførte offentliggjorte afhandlinger af Det Naturvidenskabelige Fakultet ved Københavns Universitet antaget til offentligt at forsvares for den naturvidenskabelige doktorgrad.

København, den 2. december 2005.
Flemming Nicolaisen
Kst. dekan

Forsvaret finder sted Fredag den 17. marts 2006 klokken 13.15 i auditoriet, Rockefellerkomplekset, Juliane Maries Vej 30.

Contents

Preface

The field of probabilistic Monte Carlo analysis of inverse problems was largely developed during the 1990's and early 2000's as a result of new advances in computer technology. Albert Tarantola and myself played an active role in this development, beginning with studies and experiments with 'simulated annealing' optimization for model construction, and followed by introduction of methods for sample-based probabilistic analysis of inverse problems. In this fertile period, the new methods were extensively tested on intermediate-scale geophysical inverse problems, ranging from seismic studies of the lithosphere and geomagnetic studies of fluid flow in the earth's core, to paleoclimate variations in Greenland, and seismic investigation of the Moon's interior. One of the latest developments is the introduction of an invariant formulation of inverse problems which in a consistent way clarifies what is meant by the essential concept of volume, and hence 'sampling density', in a model space.

The present booklet is an attempt at summarizing the theoretical and numerical achievements obtained in this field until now, with emphasis on my own contributions. Over the years, an understanding of many of the fundamental problems of sampling strategies for inversion has emerged, and in the present summary it is my hope to give an overview of the key findings, which is accessible by non-specialists in the field. My point of departure will be 11 papers of my own, published during the 1990s and 2000s in leading scientific journals and books (see Appendix B), and an emphasis on mapping out the development of key ideas in the field, avoiding excessive mathematical detail. The interested reader will find details and formal derivations in the papers of Appendices B1-B11.

Much of the work was done in collaboration with many colleagues and students in Copenhagen, Aarhus, Paris, San Diego, Cambridge UK, Canberra and Denver. I am especially indebted to (in alphabetic order of their first name) Albert Tarantola, Amir Khan, Bjarne Andresen, Camilla Rygaard-Hjalsted, David Snyder, Dorthe Dahl-Jensen, Egon Nørmark, Helle Wagner, Jacob Mørch Pedersen, Malcolm Sambridge, Peter Salamon, Peter Vestergaard, Satish Singh and Thomas Mejer Hansen. All these have people have contributed with great insight and enthusiasm to the 'Monte Carlo project' which in many respects has been a highly multidisciplinary activity. I also wish to thank Dr. Niels Højerslev from my institute for his uninterrupted moral support during the preparation of this booklet.

Inverse problems and sampling algorithms

Chance is always powerful. Let your hook be always cast; in the pool where you least expect it, there will be a fish.

Ovid

The Inverse Problem

As the so-called *highly nonlinear inverse problem* is a focal point of this thesis, it is natural to start with a definition of an inverse problem. An inverse problem arises when theoretical physics provides us with a mathematical relation (which we will refer to as the *forward relation*)

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) \tag{1}$$

between observable parameters (data) $\mathbf{d}$ and unobservable parameters (model parameters) $\mathbf{m}$ of the same physical system, and when we wish to infer the model parameters from the data. Data are contaminated by noise whose statistical properties we know (at least approximately). The notion of an 'inverse problem' is essentially a generalization of the notion of 'measurement': we are 'measuring' the model parameters by collecting and analyzing data. However, it is customary to use the term 'measurement' when the relation between data and model parameters is simple, and reserve the term 'inverse problem' to cases where the relation is complex.

In a general, probabilistic formulation of inverse problems [1], the solution to the problem can (under mild assumptions[1]) be expressed as a probability density function (hereafter simply called a *probability density*) σ_m over the model space:

$$\sigma_m(\mathbf{m}) = \rho_m(\mathbf{m})L(\mathbf{m}) \tag{2}$$

[1]We assume here that data $\mathbf{d}$ depend explicitly on the model $\mathbf{m}$, that is, there exists a function $\mathbf{f}$ such that $\mathbf{d} = \mathbf{f}(\mathbf{m})$ (For a treatment of implicitly defined problems, see [38]). Furthermore, we assume that data uncertainties are independent of model uncertainties.

Here, $L(\mathbf{m})$ is a *likelihood function* describing the degree of fit between observed data and data computed from the model[2] $\mathbf{m}$. L is given by

$$L(\mathbf{m}) = \rho_d(\mathbf{f}(\mathbf{m})) \tag{3}$$

where $\rho_d(\mathbf{d})$ is the (prior) probability density describing the noisy data - a probability density centered at $\mathbf{d}_{obs}$ and with a dispersion describing the measurement uncertainties. In (2), $\rho_m(\mathbf{m})$ is the *prior* probability density describing the information we have on $\mathbf{m}$ before the inverse problem is solved.

Occasionally, an inverse problem is simplistically viewed as an optimization problem, in which case a *misfit function*, e.g., the squared Euclidian distance $S(\mathbf{m}) = \| \mathbf{d}_{obs} - \mathbf{f}(\mathbf{m}) \|^2$ between observed and modeled data, is used as an objective function to be minimized. Alternatively, one can maximize the objective functions $-S(\mathbf{m})$ or $\exp(-S(\mathbf{m}))$.

If the inverse problem is linear, and the noise- (data-) and prior probability densities are Gaussian, then the posterior probability density is also Gaussian. A vast literature exists about linear Gaussian problems, which to a large extent are susceptible to detailed mathematical analysis. Slight deviations from linearity and Gaussianity can be dealt with by linearization, and most of the current literature on 'non-linear' inverse problems is based on this method.

In this dissertation we shall focus on a class of inverse problems where the posterior probability density deviates so strongly from Gaussianity that it is unfeasible to use the theory of Gaussian linear problems as an approximation. The non-Gaussianity of the posterior probability density stems either from nonlinearity of the relation between data and model parameters, or from non-Gaussianity of the noise- and/or the prior probability density. Although the term *highly non-Gaussian problems* would be the most accurate name for this class of problems, we shall instead use the less precise, but widely used, name *highly non-linear problems*. Besides being the most popular name, it refers to the perhaps most common reason for non-Gaussianity, namely non-linearity.

Ideas behind the Monte Carlo method

The main theme in the following will be analysis of highly non-linear inverse problems using Monte Carlo sampling methods. As a prepa-

[2] In inverse problem theory the set (or vector) of model parameters $\mathbf{m}$ is often called 'the model'. This is in contrast to the most common usage elsewhere in the literature, where the term 'model' stands for the relationship (1).

ration for this, let us briefly consider Monte Carlo methods in a more general context.

The idea of using random numbers to solve mathematical problems of a non-random nature seems paradoxical and inefficient, but the idea originated long before the advent of the digital computer. In the 18th century Buffon [2] described an experiment where a needle of length p is repeatedly thrown on a wooden floor with distance l between its (equispaced) cracks. He calculated that if $n(N)$ is the number of tosses where the needle crosses a crack out of N tosses in total, then

$$\frac{n(N)}{N} \to \frac{2}{\pi}\frac{p}{l} \quad \text{for} \quad N \to \infty. \tag{4}$$

This experiment can be used as a way of estimating π, and therefore became one of the first so-called Monte Carlo[3] calculations reported in the literature. This way of calculating π is, in many ways, a typical Monte Carlo algorithm: Using a random input (initial conditions of the needle tossed in the experiment), an estimate of a non-random number (here π) is asymptotically obtained. The convergence of the algorithm is extremely slow - a property unfortunately shared by many other Monte Carlo algorithms.

Buffon's algorithm was perhaps only a curiosity, but later - about a hundred years ago - it was realized that integrals such as

$$I = \int_{\mathcal{X}} h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}\,, \tag{5}$$

where $f(\mathbf{x})$ is a non-negative valued integrable function, could, in principle, be evaluated numerically by generating a large number of random realizations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of $\mathbf{x}$ using $f(\mathbf{x}) / \int_{\mathcal{X}} f(\mathbf{x}) d\mathbf{x}$ as a probability density [3]. The sum

$$I \approx \frac{1}{N} \sum_{n=1}^{N} h(\mathbf{x}_n). \tag{6}$$

could then be used as an estimate of the integral. This observation was of little practical value at that time because of the time consumption required to generate large samples from a probability density, but the advent of the digital computer changed the situation completely. The fact that the integral (5) over a high-dimensional space $\mathcal{X}$ is more efficiently calculated by the above Monte Carlo algorithm (once statistical independent realizations of the probability density are given) than

[3]The name 'Monte Carlo' was invented by Metropolis and Ulam (1949) in jest, of course referring to the city of the famous casinos.

by any other method [4], means that Monte Carlo integration methods have become important in numerical analysis.

Sampling a probability density

An important assumption behind the above mentioned Monte Carlo integration method is that *independent* realizations[4] from the considered probability density are available. As we shall see in the following sections, the problem of generating such statistically independent realizations may be difficult to solve, in some cases even practically unsolvable.

The difficulties mainly arise from the fact that, in many applications, the probability density to be sampled is not fully known. Often, an explicit closed-form expression for the probability density is not available, and the only way to acquire information about f is by picking one point $\mathbf{x} \in \mathcal{X}$ at a time, and evaluate $f(\mathbf{x})$ in the selected point using some numerical algorithm. We shall in the following use the terminology that '$f(\mathbf{x})$ can only be evaluated point-wise' for this important case.

It is quite easy to design a so-called 'perfect' sampler, that is, a method that draws perfectly independent sample points according to a probability density which can only be evaluated point-wise. A simple method is available if a number $M \geq \max_{\mathbf{x}} f(\mathbf{x})$ is given: Let each iteration consist of two steps, where in the first step a point $\mathbf{x}$ is drawn uniformly at random from $\mathcal{X}$, and in the second step an acceptance probability of $f(\mathbf{x})/M$ is used to decide if the point is accepted. However, this method is usually extremely inefficient, even when $M = \max_{\mathbf{x}} f(\mathbf{x})$. The reason is that in many practical applications, where the space $\mathcal{X}$ is high dimensional, the probability density $f(\mathbf{x})$ is near-zero almost everywhere in $\mathcal{X}$. For this reason, the waiting time before the first iteration step is successful in finding a point with a high value of $f(\mathbf{x})$ is extremely large, and the algorithm is very slow. This method is only of practical interest in low-dimensional spaces.

During the Los Alamos Project in the 1940s, where the first nuclear bomb was developed, a much more efficient method was developed by Metropolis and co-workers [5]. The method was designed to estimate thermodynamic averages for statistical mechanical systems with many degrees of freedom. Their problem was to estimate integrals of the

[4] We are here using the term 'realization' for individual points generated according to a probability density, and, as it is customary in the statistics literature, the term 'sample' denotes a collection of such points. Note that in some of the papers in *Appendix B* the term 'sample' is synonymous with the term 'realization'.

form (5) where $f(\mathbf{x})$ is the Gibbs-Boltzmann probability density over the phase space $\mathcal{X}$. The main problem here is again that $f(\mathbf{x})$ can only be evaluated point-wise.

The idea behind the Metropolis Algorithm is simple. To sample a probability density f, the algorithm picks, in each iteration, the next realization from pair of points $\mathbf{x}_1$ and $\mathbf{x}_2$ (the current point and a 'candidate' point, respectively) with probabilities proportional to $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$, respectively. This ensures, in the long run, a correct balance between sample densities in all points visited by the algorithm, and therefore, asymptotically, produces a sample of $f(\mathbf{x})$, even if $f(\mathbf{x})$ is unnormalized and can only be evaluated point-wise. More precisely, the algorithm is, for our purposes, defined in the following way:

Algorithm 1 ***(Metropolis).*** *Given a (possibly unnormalized) probability density $f(\mathbf{x})$ over the manifold $\mathcal{X}$, a random function $V(\mathbf{x})$ which samples a constant probability density if applied iteratively:*

$$\mathbf{x}^{(n+1)} = V\left(\mathbf{x}^{(n)}\right), \tag{7}$$

and a random function $U(0,1)$ generating a uniformly distributed random number from the interval $[0,1]$. The random function W, which iteratively operates on the current *parameter vector $\mathbf{x}^{(n)}$ and produces the next parameter vector $\mathbf{x}^{(n+1)}$:*

$$\mathbf{x}^{(n+1)} = W\left(\mathbf{x}^{(n)}\right) = \begin{cases} V\left(\mathbf{x}^{(n)}\right) \text{ if } U(0,1) \leq \min\left[1, \frac{f\left(V\left(\mathbf{x}^{(n)}\right)\right)}{f\left(\mathbf{x}^{(n)}\right)}\right] \\ \mathbf{x}^{(n)} \text{ otherwise} \end{cases}, \tag{8}$$

asymptotically samples the probability density $C f(\mathbf{x})$, where C is a normalization constant.

Algorithm 1 works under rather mild assumptions about V, namely that V is *irreducible* and *aperiodic*. See [6, 7] for a definition of these properties. The word "asymptotically" means in this case that the set of points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ visited in n successive steps by the algorithm converges towards a sample of f as n goes to infinity.

A generalization of the Metropolis Algorithm was introduced by Geman and Geman [8]. In each iteration of their so-called *Gibbs Sampler* the next realization is picked from a large collection of points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ (including the current point) with probabilities proportional to $f(\mathbf{x}_1)$,

$f(\mathbf{x}_2)$, $f(\mathbf{x}_3), \ldots$, respectively. This again ensures, in the long run, a correct balance between sample densities in all regions of $\mathcal{X}$ visited by the algorithm, and therefore, asymptotically, produces a sample of $f(\mathbf{x})$. Due to the correspondence between the ideas behind the Gibbs sampler and the ideas behind the Metropolis algorithm, it is natural to regard the Gibbs Sampler as a member of the 'Metropolis family'.

The Metropolis algorithms and its relatives are so-called *Markov chains*. For a Markov chain, the probability of visiting a point in $\mathcal{X}$ in a given iteration depends only on the point visited in the previous iteration (and not on any of the earlier iterations). In this sense a Markov chain has the shortest possible 'memory', and this is an advantage from the point of view of simplicity and required memory for the algorithm. However, the fact that a Markov chain algorithm discards information on the probability density to be sampled, obviously limits its efficiency.

Metropolis samplers and inverse problems

Some 40 years after the birth of the Metropolis Algorithm a seminal paper by Kirkpatrick and co-workers [9] on *Simulated Annealing* stimulated further studies of the possibilities and limitations of the algorithm [6].

Simulated annealing is a Metropolis Algorithm, designed to sample the Gibbs-Boltzmann probability density

$$p_B(\mathbf{x}) = \frac{\exp\left(-\frac{E(\mathbf{x})}{T}\right)}{Z(T)}. \tag{9}$$

(where $1/Z(T)$ is a normalization constant). For a constant value of the temperature parameter T this simulates a statistical mechanical system in equilibrium with a heat bath of temperature T, but the idea behind simulated annealing is to lower the temperature T gradually and slowly from a high value to near-zero. After this 'annealing' process, the Gibbs-Boltzmann probability density approximates a delta function at the global minimum for $E(\mathbf{x})$ (if it is unique). In other words, the annealing has solved the optimization problem of finding the global minimum for E.

The potential of simulated annealing to solve highly complex combinatorial optimization problems was a main reason for the almost explosive increase in interest in the algorithm during the 1980s. At that time, much work in highly nonlinear inverse problems was focussed on the 'model construction problem', that is, the problem of finding

at least one solution that fits the available data within the noise. This can be regarded as an optimization problem: Minimize a 'misfit function' $S(\mathbf{m})$ that measures the difference between data computed from a given model $\mathbf{m}$ and the measurements (observed data) $\mathbf{d}_{obs}$. Overfitting (the fact that a perfect solution to this problem will yield an unrealistic model that also fits the noise) can be dealt with by regularization, that is, allowing only models that are 'smooth' or 'simple' in some sense. Rothman [10, 11] was the first to apply simulated annealing to a model construction problem, namely the highly nonlinear problem of residual statics estimation in reflection seismology. The problem essentially deals with estimation of seismic wave propagation velocities in the near-surface layers from industrial seismic reflection data. The problem is known to be a hard optimization problem, and hence a stern test of the simulated annealing approach.

Rothman's work, and further experimentation with large, complex optimization problems in the late 80s, revealed how difficult it was to use simulated annealing in practice. Kirkpatrick et al.'s suggestion [9], that the temperature parameter is lowered as a decreasing exponential, worked well in some cases, but not for the hardest optimization problems. Investigations [12] have shown that for such problems, even very slow exponential annealing resulted in a solution that corresponded, not to the global minimum, but rather a local minimum for the objective function. In many cases the local minimum was far away from the global solution.

In the late 1980s a group of researchers at San Diego State University, University of Copenhagen and University of Heidelberg developed a new method, *simulated annealing at constant thermodynamic speed*, rooted in finite-time thermodynamics, for calculating near-optimal annealing temperature schedules [13, 14]. The idea was to find ways of lowering the temperature, such that the 'numerical system' (viewed as a statistical mechanical system) was kept at a constant distance from equilibrium during annealing, minimizing the risk of settling into a spurious local minimum in the process. The results of this development were promising, and new studies [15, 16] were initiated to investigate how this method could help solving hitherto unmanageable inverse problems - the so-called 'highly non-linear' problems.

Model construction: Locating acceptable solutions in limited time

Seek not the things that are too hard for thee, neither search the things that are beyond thy strength.

Apocrypha

The idea of computing acceptable solutions to (highly) non-linear inverse problems by Monte Carlo methods is not new. Amongst the most famous early examples of such calculations we find investigations by Keilis-Borok and Yanovskaya [17] and Press [18, 19, 20]. Based on these early experiences, Monte Carlo had the reputation of being extremely inefficient, but around 1990 the new simulated annealing approach based on finite-time thermodynamics gave some hope that reasonable results could be obtained, even for many-parameter inverse problems. A key idea in the new method was to seek the best possible solutions within the finite time given by limited computer resources.

The first investigations [15, 16] focussed on a type of inverse problems known from industrial reflection seismology. Assuming from geological knowledge that the uppermost kilometers of a sedimentary basin consists of almost horizontally stratified layers, and assuming that each layer is almost homogeneous, it is possible to parametrize such a model using relatively few parameters. The model is, at each surface location, characterized by the location (depth) of the base of each layer, and the wave reflection coefficient at each layer interface. Two situations were investigated, one with only one surface location and 28 model parameters in total [15], and one with 71 surface locations simultaneously, invoking 1136 model parameters [16]. For both problems it was assumed that data is given by convolving the *reflectivity* (reflection coefficient as a function of depth at each surface point) with a known seismic source function (the *wavelet*).

Data were seismograms, one for each surface location. In the one-location example, the seismogram consisted of 50 samples, and in the 71-location example, the seismograms each consisted of 100 samples. In both cases the data sampling interval was 0.004 s.

The two studies represent the first, full-scale analysis of inverse problems using simulated annealing at constant thermodynamic speed. For the first time, realistic inverse problems were treated fully as a statistical mechanical problem, characterized by its 'thermodynamic properties': a density of states, a heat capacity, and a relaxation time.

Estimation of heat capacity and relaxation time takes place before annealing is initiated, as a necessary preparation for design of a cooling temperature schedule for constant speed annealing [14, 21]. The drawback of this is, of course, some computational overhead, but the advantage is a considerable overall gain in efficiency of the annealing process.

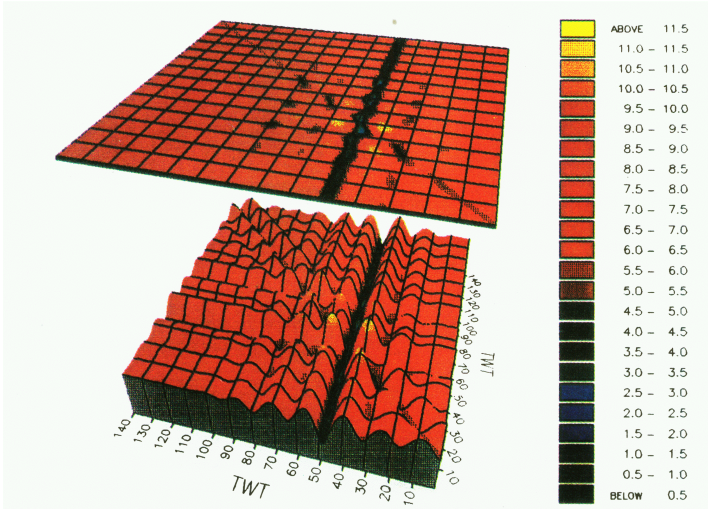


Figure 1: Misfit function surface for a simple, single-seismogram model optimization problem [15, 16]. The misfit surface is shown for a two-dimensional cut through the parameter space. The independent parameters in the considered plane are the depths (two–way travel times) of two reflectors.

It can be demonstrated that the above-mentioned seismic problem is indeed highly non-linear [16] (Figure 1), and inspection of evolution of the model during the annealing process revealed many similarities between the numerical system, and a crystal being annealed through its melting point [10, 11, 16] (Figure 2).

One of the important things we can learn from the first inversions using simulated annealing at constant thermodynamic speed [15, 16] is that *prior* (in this case 'thermodynamic') information about the structure of the misfit function $S(\mathbf{m})$, or the corresponding probability den-

sity $f(\mathbf{m}) \propto \exp(-S(\mathbf{m}))$, if used properly, can have a dramatic effect on the efficiency of the Monte Carlo search. In fact, it can be argued that the most important single factor in determining the difficulty of a sampling problem (of which simulated annealing is a special case) is the information we have on the misfit function, or corresponding probability density. If we know little in advance, the sampling problem is hard. If we have comprehensive information, the sampling problem is easy.

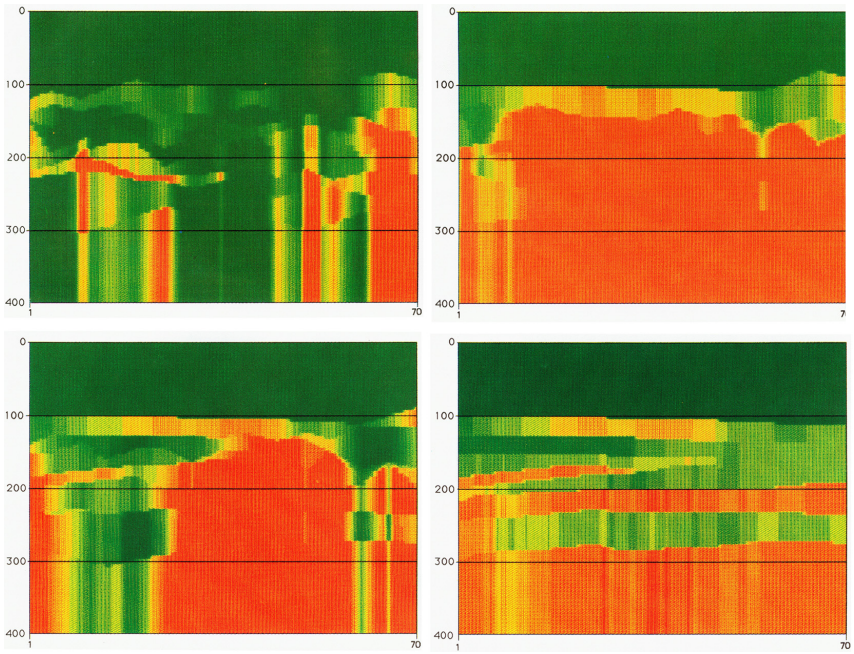


Figure 2: Sample subsurface models obtained at successively lower temperatures. The plots show the acoustic impedance as a function of two–way travel times.

Considering the difficulty of gaining information about the structure of the misfit function for a particular problem, it is of great practical interest to study to which extent similar problems have similar misfit functions. In fact, some highly non-linear inverse problems of practical importance appear again and again, but in a slightly different form each time. One example is the solution with constant speed annealing of the residual statics problem. It has been shown [23] that there is a simple way of transforming the thermodynamic properties of one residual statics problem into the thermodynamic properties of another residual

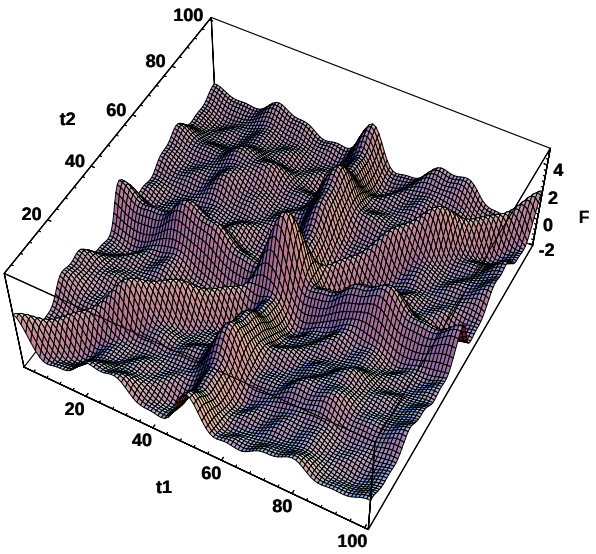


Figure 3: The objective function (to be maximized) for a residual statics problem, seen on a 2-dimensional intersection through the model space [22]

statics problem. In fact, it is possible to find an approximate 'master' temperature schedule giving near optimal results for a wide range of such problems if they are properly scaled beforehand [23]. The successful result of this study indicates that most residual statics problems have objective functions (called *stack power functions*, see Figure 3) with statistically similar structure.

Resolution analysis of highly non-linear inverse problems

We think in generalities, but we live in detail.

A. N. Whitehead

Sampling the posterior probability density

Sampling may be the only way to characterize the posterior probability density for a highly nonlinear inverse problem if there is no explicit, closed-form expression available for the prior probability density $\rho_m(\mathbf{m})$ and/or the relation between data and model. In some cases these expressions exist, but we do not wish make use of them, perhaps because they are not mathematically simple [24]. In the worst case we only have

1. A 'black box' algorithm that is able to sample $\rho_m(\mathbf{m})$ (but not necessarily able to evaluate $\rho_m(\mathbf{m})$ for a given $\mathbf{m}$).
2. A 'black box' algorithm that is able to compute $L(\mathbf{m})$ for a given model $\mathbf{m}$ (that is, $L(\mathbf{m})$ can only be evaluated point-wise).

In this case we may resort to a characterization of the posterior probability density $\sigma_m(\mathbf{m}) = L(\mathbf{m})\rho_m(\mathbf{m})$ by a sample of independent realizations. This can be obtained by the following algorithm [25], which is an extended version of the Metropolis Algorithm:

Algorithm 2 ***(Extended Metropolis).*** *Given a random function $V(\mathbf{m})$ which samples the prior probability density $\rho_m(\mathbf{m})$ if applied iteratively:*

$$\mathbf{m}^{(n+1)} = V\left(\mathbf{m}^{(n)}\right), \tag{10}$$

and a random function $U(0,1)$ generating a uniformly distributed random number from the interval $[0,1]$. The random function W, which iteratively operates on the current *parameter vector $\mathbf{m}^{(n)}$ and produces the next parameter vector $\mathbf{m}^{(n+1)}$:*

$$\mathbf{m}^{(n+1)} = W\left(\mathbf{m}^{(n)}\right) = \begin{cases} V\left(\mathbf{m}^{(n)}\right) \text{ if } U(0,1) \leq \min\left[1, \frac{L\left(V\left(\mathbf{m}^{(n)}\right)\right)}{L\left(\mathbf{m}^{(n)}\right)}\right] \\ \mathbf{m}^{(n)} \text{ else} \end{cases}, \tag{11}$$

asymptotically samples the posterior probability density $\sigma(\mathbf{m}) = CL(\mathbf{m})\rho(\mathbf{m})$, *where* C *is a normalization constant.*

Algorithm 2 works if V is *irreducible* and *aperiodic*.

The extended Metropolis algorithm permits a dramatic time saving under certain circumstances [25]. In its basic formulation (above) there is already a time saving in the fact that only models that are accepted a priori (by V) proceed to the usually very time consuming misfit calculation needed to evaluate $L(\mathbf{m})$.

Cascaded Metropolis

In some applications we are inverting many different types of data simultaneously, as for instance in geophysics, where we may attempt a joint inversion of earth tides, free oscillations, and body-wave seismic data. Typically, data uncertainties are independent amongst these data sets, and the total likelihood can be expressed as a product

$$L(\mathbf{m}) = L_1(\mathbf{m})\, L_2(\mathbf{m})\, \ldots \tag{12}$$

of partial likelihoods, one for each data type. Using the original Metropolis algorithm (8) directly on $\sigma(\mathbf{m})$ would require solving the full forward problem (usually the most time-consuming part of the algorithm) to every model proposed by the prior random walk. Instead, we can use the extended Metropolis algorithm (11) in cascade: A basic, extended Metropolis algorithm that samples the product of $\rho_m(\mathbf{m})$ and $L_1(\mathbf{m})$ can be used as a prior random walk V_1 in a new, extended Metropolis algorithm which, through evaluation of $L_2(\mathbf{m})$ samples the product $\rho_m(\mathbf{m})L_1(\mathbf{m})L_2(\mathbf{m})$. In turn, this random walk can be used as a prior random walk V_2 in another, extended Metropolis algorithm to sample $\rho_m(\mathbf{m})L_1(\mathbf{m})L_2(\mathbf{m})L_2(\mathbf{m})$, and so on, until the posterior probability density that takes into account the total data set is sampled [25, 26].

The practical consequences of the above procedure are important. A model proposed by the a priori sampler V, proceeds to evaluation of misfit with respect to the the first data subset. The proposed model may then be accepted or rejected. If it is rejected (typically when there is a large misfit) there is no need to calculate the misfit with respect to the other data subsets. A new model can now be proposed by the a priori sampler V. In general, each time a model is rejected at some stage of the algorithm, it returns to the lower level and propose a new model. If

misfit calculations for the likelihoods $L_1(\mathbf{m})L_2(\mathbf{m})\ldots$ are arranged in order of increasing expense, this cascaded extended Metropolis algorithm may be much more efficient than using the Metropolis algorithm directly to the total data set.

Resolution analysis

Linear resolution analysis

The term *resolution* in inverse theory embraces the ability of an inversion method to reveal structures in the true model, using the given data. The traditional resolution concept grew out of linear inverse theory [27, 28, 29] where resolution was measured by the 'width' (appropriately defined) of a model obtained from synthetic data which are computed from a delta model (zero everywhere, except at a single point). In other words, if a true model $\mathbf{m}_{true}$ is a delta function, and $\mathbf{d}$ is *noise free* data related to the true model through the linear relation

$$\mathbf{d} = \mathbf{F}\mathbf{m}_{true} \tag{13}$$

and we are solving the inverse problem by means of a linear operator $\mathbf{H}$ to obtain an estimated model

$$\mathbf{m}_{est} = \mathbf{H}\mathbf{d} \tag{14}$$

then, if $\mathbf{m}_{true}$ is a delta model, then the width of $\mathbf{m}_{est}$ is a measure of the resolution. The broader $\mathbf{m}_{est}$, the poorer the resolution.

The traditional resolution concept deals with possible nonuniqueness of solutions to the linear inverse problem $\mathbf{d} = \mathbf{F}\mathbf{m}$, caused by underdetermination. *Regularization* (picking an acceptable model out of a multitude of solutions with a reasonable data fit) is built into the inversion operator $\mathbf{H}$, but the price to be paid is a reduced resolution.

Although data uncertainties may strongly influence the ability of an inversion method to reveal structure in the true model, the traditional resolution concept does not take uncertainties into account. Uncertainties are dealt with separately.

Highly non-linear resolution analysis

In a purely probabilistic formulation of the inverse problem, combined with a model sampling approach [1, 24, 25], the exhaustive solution to the inverse problem is not one model, but the posterior probability density (2). The likelihood function measures the data misfit, and

carries information about the data noise characteristics, whereas the prior probability density replaces the traditional regularization as the factor allowing only reasonable solutions and removing or reducing non-uniqueness. In the a highly non-Gaussian ('non-linear') inverse

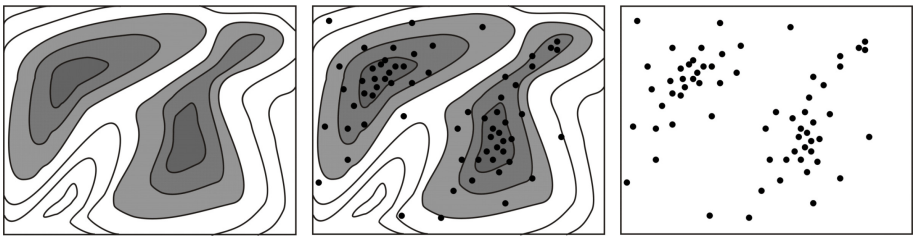


Figure 4: Complex probability densities can be represented by a collection of points sampled from the probability density (a *sample*).

problem, the posterior probability density will be poorly represented by a mean model and a covariance matrix. We may therefore resort to the simplest possible representation, namely by a sample (a number of independent realizations) (Figure 4) [24, 25]. From a sample of the posterior probability density it is then our task to extract useful information about the solution to the inverse problem.

At this point it should be noted that, in order to represent the probability density satisfactorily, the sample must have a certain size. The least acceptable size of the sample depends on the structure of the posterior probability density and the dimension of the model space. But it also depends on the demands of the further analysis. If only a mean model is required we have a situation known from classical Monte Carlo integration theory [30, 31]. In this case a smaller sample is required than if we wish to estimate the probability of the existence of finer details in the model.

In probabilistic, sample-based analysis of highly non-linear inverse problems, resolution analysis and uncertainty analysis are inseparable, and the combined investigation is simply termed *resolution analysis* [25].

Before sample-based resolution analysis of a highly non-linear inverse problem can take place, we must make sure that our model is not constrained through an arbitrary oversimplification of the model. For this reason we overparameterize the system, i.e., we make the model sufficiently 'fine grained'[5] [24]. After this, the analysis can proceed

[5]If a prior for the parameterization is available (e.g., favoring models described by

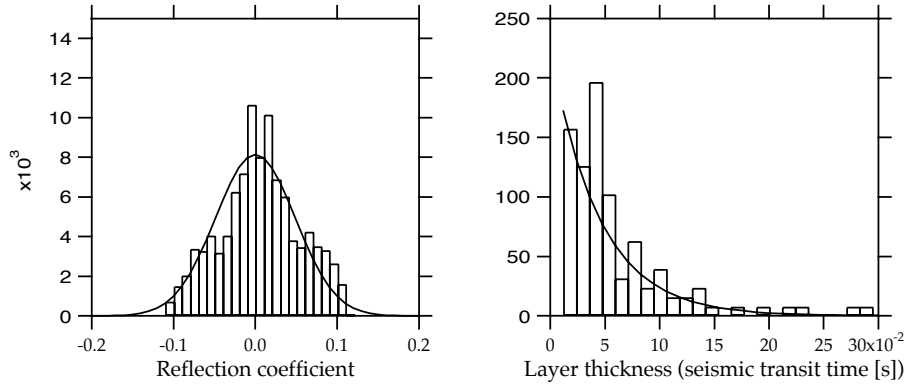


Figure 5: A priori information used in a study of seismic reflections from the deep lithosphere [32, 24]. Left: Reflection coefficients distribution obtained from studies of rock samples in the field [33, 34]. The solid curve is a Gaussian probability density with standard deviation $\sigma = 0.047$. Right: Layer thickness distribution as a function of one-way travel time, derived from field observations. The solid curve is an exponential probability density with mean $\mu = 1/\lambda$, where $\lambda = 22.5\ s^{-1}$.

through the following steps [24, 25, 36]:

1. Collect a large sample of models from the posterior probability density.

2. Compare all models in the sample with the aim of recognizing structures repeated in a large number of realizations. Alternatively, identify realizations which contain structures of particular interest to the application from which the inverse problem originates.

3. Compute the fractions of all realizations that contain the considered structures. The fractional occurrence of a structure approximates the posterior probability that the structure is present (per construction of the sample).

One important difference between traditional resolution and uncertainty analysis, and the above procedure, is the pattern recognition step

few parameters), it is possible to determine the parameterization as part of the analysis [35].

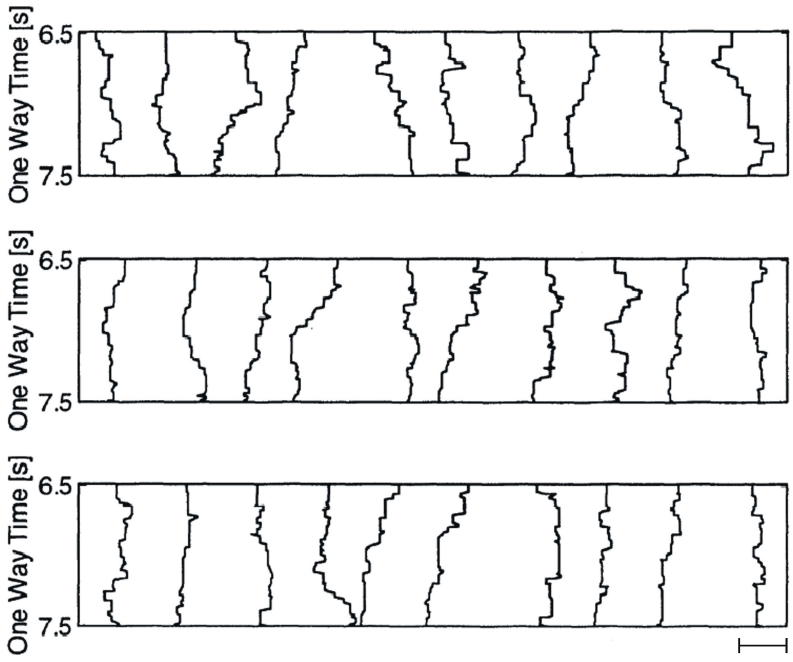


Figure 6: A selection of typical a priori models generated assuming the prior probability densities shown in Figure 5. Each curve shows an acoustic impedance variation as a function of depth (one-way wave propagation time). The models are all statistically equivalent and have equal likelihoods of existence [32, 24]. The bar in the lower-right corner indicates an acoustic impedance contrast of 10.000 (g/cm^3) (m/s).

mentioned in (2). A simple approach where models are displayed side by side, or replayed sequentially on a screen (a method nicknamed 'the movie philosophy'), has been suggested [25]. The idea is here to exploit the pattern recognition abilities of the human eye/brain to discover structure that occurs often in the sample, and therefore has a high posterior probability (is 'well resolved'). Future application of pattern recognition methods may automatize this part of the analysis. Pattern recognition, whether 'manual' or automatic, can be aided by a preprocessing of the sample models. One simple kind of preprocessing is to filter the models before inspection [25].

An example of application of the extended Metropolis algorithm for resolution analysis of a highly non-linear inverse problem is shown in Figures 5, 6 and 7. In this study of seismic reflections from the deep

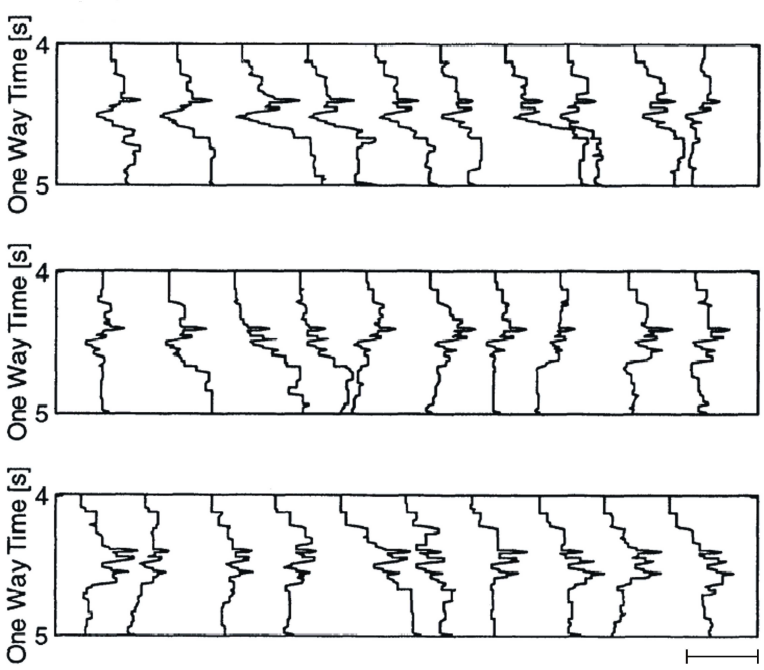


Figure 7: The results of the inversion on the Moho data set: A selection of a posteriori models [32, 24]. Each curve shows a possible acoustic impedance variation as a function of depth (one-way wave propagation time). The bar in the lower-right corner indicates an acoustic impedance contrast of 10.000 (g/cm^3) (m/s).

lithosphere [32, 24], the prior sampler V is based on parametric probability densities, derived as approximations to histograms of reflection coefficients and layer thicknesses (Figure 5). The histograms were obtained from laboratory- and field measurements on rocks from the island of Rum, Scotland, and the Great Dyke in Zimbabwe. Output from running V is shown in Figure 6 which gives an impression of the 'soft constraints' imposed by the prior. The final output of the extended Metropolis (Figure 7) is a sample of the posterior probability density, which – loosely speaking – can be thought of as models that are consistent with the prior, but also able to fit the data (in this case reflection seismograms) within the noise.

Inspection of the posterior sample in Figure 7 reveals that part of the structure appears persistently in the models, namely the oscillating structure around 4.5 s. The persistence reflects a high posterior prob-

ability for the existence of this structure. In fact, the posterior probability of the presence of the oscillating feature can be estimated as the fractional occurence of the feature in a large posterior sample. One can conclude that this feature is 'well resolved', in contrast to structure in, e.g, the lower part of Figure 7 which changes significantly from model to model, reflecting a low probability.

Some specific inverse problems

Try to put well in practice what you already know. In so doing, you will, in good time, discover the hidden things you now inquire about.

Rembrandt

When applying sampling-based, probabilistic inversion to real-world situations, we face the problem of how to establish the prior information, and how we obtain the noise distribution needed to evaluate the likelihood $L(\mathbf{m})$.

To see how this problem has been solved in practice, we shall consider four different major geophysical inverse problems. The primary scientific results obtained from the analysis of the particular problems can be found in the original papers [32, 37, 38, 39, 40]. Here we shall only address the diversity of ways by which prior and noise distributions were obtained in the four studies.

A brief summary of four problems

1. A seismic study of deep lithosphere structures [32, 24]. Seismic near-normal-incidence reflection data from the DRUM (Deep Reflections from the Upper Mantle) reflection profile from the north of Scotland were inverted with the purpose of estimating acoustic impedance variations (reflection coefficients) in the deep lithosphere. Data were in the form of seismograms recorded at 10 surface locations, each with a time duration of 2.028 s, and a sampling interval of 0.008 s. The unknown model parameters were reflection coefficients of 128 layer interfaces in the considered target interval in the lithosphere.

2. A study of paleotemperatures from temperatures measured in boreholes in the Greenland ice cap [37]. The temperature profiles measured in the borehole of the Greenland Ice Core Project (GRIP), and in the Dye 3 borehole 865 kilometers farther south, was inverted with the aim of recovering a 50,000-year-long temperature history at GRIP, a 7000-year history at Dye3, and a terrestrial heat flow density. Data were in the form of temperatures measured at 61 depth levels in the boreholes, which has a depth of 3080 m and 2037 m, respectively. The unknown model parameters were the terrestrial heat flow density, and

temperatures at 125 times with exponentially decreasing time intervals (25 ky 450000 years ago and 10 years at present).

3. An analysis of the motion of the earth's core fluid based on the secular variation of the geomagnetic field [38, 39]. Geomagnetic data measured at or near the Earth's surface were inverted with the purpose of recovering the dynamics of the horizontal fluid flow structures on a surface at the core-mantle boundary (CMB), in particular upwelling and down-welling of flow crossing the CMB. Data were 195 coefficients in a spherical harmonic expansion of the geomagnetic field, and 195 expansion coefficients describing the secular variation of the geomagnetic field. The unknown model parameters were 390 expansion coefficients describing the fluid flow velocity field at CMB.

4. A study of the Lunar interior from the Apollo seismic data [40, 6]. As part of the Apollo project, seismic stations were deployed at five locations on the Moon, and operated simultaneously as a four-station seismic array, from April 1972 until 30 September 1977. Data from the experiment was inverted to obtain information about Lunar structure and moonquake hypocenter locations. Data were available as *P*- and *S* wave arrival times from 177 lunar seismograms. The unknown parameters of the spherically symmetric model were P- and S-velocities in 56 layers, the depth to 56 layer boundaries, and near-surface velocity corrections for P- and S-arrivals at 26 seismic stations, meteorite- and spacecraft impact points. In addition to these velocity parameters there were 230 unknown hypocenter coordinates and -times.

Establishing the noise distribution for the four problems

In a probabilistic formulation of inverse problems we assume full knowledge of the statistical properties of the noise. That this, to some extent, is an Achilles' heel of this approach is reflected in the difficulties we face when trying to establish the noise distribution in concrete cases.

Practically, there are two typical ways of choosing an appropriate noise distribution. The first method is illustrated in Problem 2, where the noise is assumed to originate only from the measuring instrument. The instrument builder has given a noise *variance* only, and this (together with the assumption that the measurements were statistically independent) provides the basis for assuming that the noise distribution is an isotropic Gaussian in the data space, with zero mean and with the given variance.

The other typical way of choosing an appropriate noise distribution

is to estimate it directly from the data. This approach was followed in problems 1, 3 and 4. In all three cases, the method was essentially the same: A smooth or simplified (regularized), "physically realistic" model was fitted to the observations, and the difference between the observations and the data computed from the model was assumed to be noise. The reason for choosing this modelling approach was *not* lack of information on instrumental noise. The reason was the unimportance of instrumental noise, compared to other noise sources. In Problem 1, for instance, the main source of noise came from unknown structure (deviations from horizontal stratification) which was difficult or impossible to embrace in the forward calculations.

Establishing the prior

The prior probability density is always the focal point of discussions when comparing probabilistic/Bayesian approaches to other methods. However, we shall avoid this discussion here, accept the fully probabilistic approach, and focus on ways of establishing an appropriate prior.

Amongst the four sample problems, we can distinguish three different ways of choosing appropriate prior probability densities: The first method is illustrated in Problem 2 and 4. Here, the priors are chosen as the probabilistic equivalents of 'hard constraints', namely as constant distributions over a certain interval (a 'box') in the model space. This reflects a desire from the analysts to use a 'neutral prior', essentially only preventing unacceptable models to be considered. This kind of prior is somewhat *ad-hoc* and belongs to the most criticized by adversaries of the probabilistic method.

A much more widely accepted approach is seen in Problem 1, where geologically realistic a priori information on reflection coefficients and layer thicknesses was established through laboratory analysis of rocks sampled in the field, and incorporated into the inversion (Figure 5 and 6). In this way subjectivity in the choice of prior was minimized. In fact, here the inverse problem can rightly be viewed as one where two independent data sets – geological field data and seismic data – are combined.

A third method is seen in Problem 3. Here, the assumption that the Coriolis force approximately balances the horizontal pressure gradient in the earth's core (approximate *geostrophy*), was used to define the (Gaussian) prior. This represents another way of avoiding subjectivity, namely by invoking physical arguments in the choice of prior.

Consistent and invariant solutions to inverse problems

Consistency is the foundation of virtue.

Francis Bacon

A probabilistic formulation of an inverse problem, or any other physical problem, is meaningless without the notion of a *homogeneous probability density*, assigning equal probabilities to 'equal sized' regions in the space. If we cannot define a homogeneous probability density, the act of assigning (usually) different weights to different regions of the space through the definition of a probability density, would lose its meaning.

In many practical applications of probabilistic/Bayesian inverse theory, the desire to use a 'neutral' prior in the calculations leads to the use of a prior that is constant over $\mathcal{M}$. However, it should be realized that, after a transformation of parameters, a constant prior would usually transform into a non-constant prior, and that would under certain circumstances reveal an inconsistency [41, 42, 1]. If some of the original parameters as well as some of the transformed parameters are basic physical parameters (lengths, positions, times, resistivities, frequencies, etc.), each of these parameters would demand its own homogeneous probability density. At the same time, the coordinate transformation between old and new parameters dictates that the probability densities of old and new parameters are linked through a Jacobian transformation. Hence, constant homogeneous probability densities for old parameters may in this way lead to accepting non-constant homogeneous probability densities for new parameters, and vice versa. A simple example is an experiment where we measure oscillations of a physical system. If we describe the oscillations by their *frequency*, and assign a constant prior to the frequency parameter, we must accept (by virtue of the Jacobian transformation) that the prior of the oscillation *period* is non-constant. On the other hand, accepting a constant homogeneous probability density for the period leads to a non-constant homogeneous probability density for the frequency. Since we are not willing to accept that some fundamental parameters take priority over others, we are facing an inconsistency.

Inconsistency can be avoided if a metric, and hence distances and volumes, is defined over the manifold $\mathcal{M}$ such that any physical parameter and its equivalent counterpart (to which it is in $1-1$ correspondence) have homogeneous probability densities of the same form [1]. In practice, special attention is required for physical parameters that are inherently positive (termed the Jeffreys parameters [1]). Such parameters appear in pairs (Frequency - period, conductivity - resistivity, velocity - slowness, temperature - thermodynamic parameter, etc.), each of which consists of parameters that are mutually reciprocal. It can be shown [1] that if we define the distance between two different states S_1 and S_2 of the physical system, characterized by two values s_1 and s_2 of a Jeffreys parameter, respectively, in the following way

$$D(S_1, S_2) = \left| \log \frac{s_2}{s_1} \right| \tag{15}$$

then the homogeneous probability density for the Jeffreys parameter has the same functional from as the homogeneous probability density for its reciprocal parameter, namely

$$f(x) = \frac{k}{x}. \tag{16}$$

where k is a constant[6].

The need to introduce a non-trivial metric in the model space in order to avoid inconsistency has profound consequences for probabilistic solution of inverse problems. Probabilistic inversion relies on the notion of *conditional probability* since the solution is established through a conditioning of the prior data probability density to the submanifold defined by the forward relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ (see (3)). However, it can be shown [1] that the traditional notion of a conditional probability density is not invariant under reparameterization (change of variables). The upper part of Figure 8 illustrates the traditional definition, where the curve, on which the probability density is conditioned, is given as the limit of a region in the space. In the traditional definition the limit is 'vertical', and the conditional probability (as a function of the horizonal coordinate) is given as the vertical average of the original 2-D probability density. The lower part of Figure 8 shows a different, orthogonal limit (with respect to the metric of the data-model product

[6]Note that homogeneous probability densities over unbounded physical spaces are usually not normalizable. They are therefore *measures*, rather than probability densities, and cannot be used to compute absolute probabilities. They can, however, be used to assign equal weights to equal volumes.

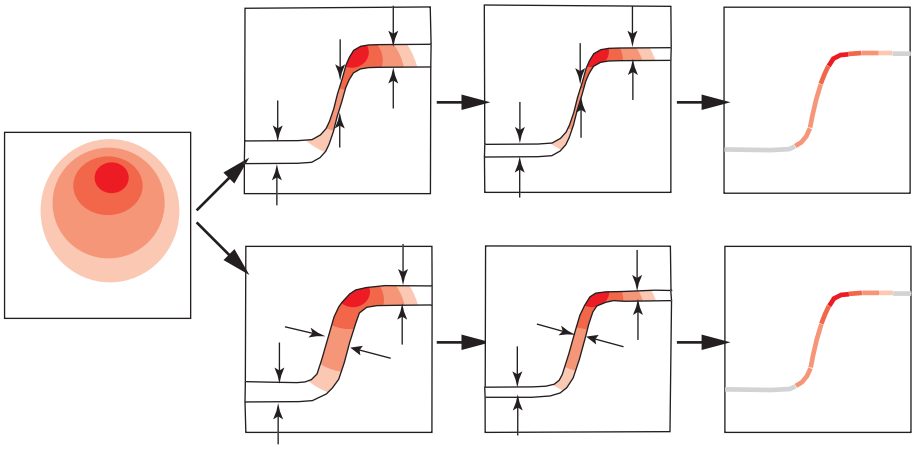


Figure 8: Definition of conditional probability. To the left a 2D probability density, and to the right two ways of defining a region of the space whose limit is a given curve. At the top is the 'vertical' limit, while at the bottom is the normal (or orthogonal) limit. Each possible limit defines a different 'induced' or 'conditional' probability density. Only the orthogonal limit gives a definition that is invariant under a change of variables.

space) which in a similar way induces its own conditional probability. Only the orthogonal limit gives a definition that is invariant under any change of variables [1].

The implications of introducing an invariant definition of conditional probability densities is that the likelihood function (3) is now modified to

$$L(\mathbf{m}) \;=\; \left(\frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_d(\mathbf{d})}} \; \frac{\sqrt{\det \left(\mathbf{g}_m(\mathbf{m}) + \mathbf{F}^T(\mathbf{m})\, \mathbf{g}_d(\mathbf{d})\, \mathbf{F}(\mathbf{m}) \right)}}{\sqrt{\det \mathbf{g}_m(\mathbf{m})}} \right) \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \;. \tag{17}$$

Here, the matrix of partial derivatives $\mathbf{F} = \mathbf{F}(\mathbf{m})$, with components $F_{i\alpha} = \partial f_i / \partial m_\alpha$, appears. Contrary to many 'nonlinear' formulations of inverse problems, the partial derivatives $\mathbf{F}$ are needed even if we use a Monte Carlo method [1].

The invariant and consistent formulation of inverse problems also has important implications for the formulation of deterministic methods used for solution of weakly nonlinear problems. This is, however, beyond our main theme, for which reason the reader is referred to [1] for further details.

Difficulties and limitations

Man needs difficulties; they are necessary for health.

Carl Jung

Optimizing the proposal probability density

Of utmost importance for the practical applicability of the Metropolis Algorithm is the question of how its computational efficiency can be maximized. A definitive answer to this question has not yet been found, although some practical advice is available in the literature [43]. Some insight into the problem can be gained by reformulating Algorithm 1, the basic Metropolis Algorithm for sampling a probability density $p(\mathbf{x})$, in the following way [21]:

1. **The exploration step:** Propose a "candidate" point $\mathbf{x}_i$ using a so-called *proposal probability density* $U(\mathbf{x}_i|\mathbf{x}_j)$, where $\mathbf{x}_j$ is the currently visited point. The proposal probability density is symmetric:

$$U(\mathbf{x}_i|\mathbf{x}_j) = U(\mathbf{x}_j|\mathbf{x}_i), \tag{18}$$

   but otherwise it can be chosen arbitrarily, in the sense that its form is chosen before running the algorithm, and is, in principle, independent of any knowledge about the probability density $p(\mathbf{x})$. $U(\mathbf{x}_i|\mathbf{x}_j)$ embodies the "strategy" by which the algorithm explores $\mathcal{X}$, when sampling the probability density $p(\mathbf{x})$.

2. **The exploitation step:** Decide if the candidate point should be accepted as the next sample point. The Metropolis acceptance probability is given by

$$p_{accept} = \begin{cases} \frac{p(\mathbf{x}_i)}{p(\mathbf{x}_j)} & \text{for } p(\mathbf{x}_i) \leq p(\mathbf{x}_j) \\ 1 & \text{otherwise} \end{cases} \tag{19}$$

   If the candidate point is rejected, the current point is repeated (thus counting one more time).

The proposal probability density $U(\mathbf{x}_i|\mathbf{x}_j)$ determines the search strategy of the algorithm, so the problem of maximizing efficiency can be formulated in the following way:

Given a measure of dissimilarity $D(p_1, p_2)$ in the space of probability probability densities over $\mathcal{X}$, and a (usually small) positive number M. Find a proposal probability density $U(\mathbf{x}_i|\mathbf{x}_j)$ minimizing the expected number of iterations needed to obtain

$$D(s, p) \leq M, \tag{20}$$

where $s(\mathbf{x})$ is the sampling probability density and $p(\mathbf{x})$ is the target (equilibrium) probability density.

When $\mathcal{X}$ is a discrete space with relatively few points, and the algorithm has a completely known transition probability matrix $\mathbf{P} = \{P(\mathbf{x}_i|\mathbf{x}_j)\}$, giving the conditional probability of jumping to $\mathbf{x}_i$, given $\mathbf{x}_j$ is the current point, convergence speed may be estimated through an eigenvalue analysis of $\mathbf{P}$ [44]. However, given that a complete knowledge of $\mathbf{P}$ is rare in most problems of practical importance, one may resort to an approximate eigenvalue analysis by lumping points of similar values of $p(\mathbf{x})$ into a small number of "states" between which transition probabilities can be estimated numerically [14].

Lacking theoretical guidance, it has become common practice to tune $U(\mathbf{x}_i|\mathbf{x}_j)$ empirically [43]. A first demand on an acceptable proposal probability density is that it must keep the *settling time* for the algorithm (often referred to as the *burn-in period* or *mixing time*) at a minimum. Loosely speaking, the settling time is the time it takes for the algorithm to evolve from a typical initial state where it samples low-probability points, until it reaches a region of high-probability points in the space (Figure 9). Once the Metropolis algorithm has completed its settling process, its outputs (coordinates of sampled points, frequency of accepted models, etc.) usually become approximately stationary. Experience has shown that, in this situation, a frequency of accepted models of $25\% - 50\%$ is optimal, since it indicates (again loosely speaking) that the jumps in the space, performed in each step of the Metropolis algorithm, are small enough to avoid too many rejections, but large enough to ensure that a large portion of the space is sampled in a reasonable time [45]. By studying the autocorrelation functions of algorithm outputs, it is possible to estimate a minimum time separation between approximately independent sample points, generated by the algorithm [43].

A recent addition to Monte Carlo theory is the development of the so-called *exact sampling methods*. These Markov-chain-based methods

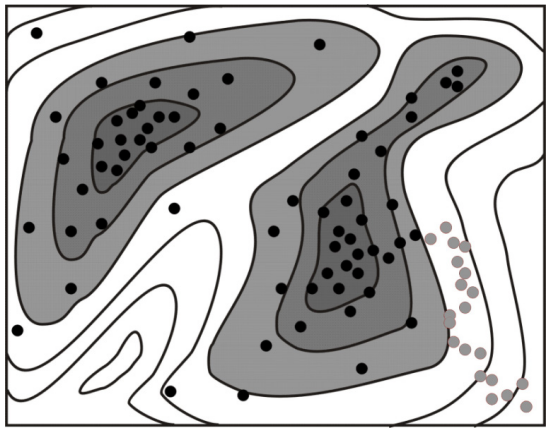


Figure 9: Settling is the process of going from a typical starting point of low probability to a region of high probability (grey dots). After this process the algorithm collects its proper sample points (black dots).

guarantee that perfectly independent sample points can be drawn from the target probability density in finite time, making convergence assessments irrelevant [46]. However, it is still unclear whether these methods will be available for the kind of large-scale continuous problems we are considering here. Exact sampling is currently a fertile area of research, but the methods proposed so far are often difficult to implement and far from automatization.

Hard inverse problems

A *hard* (or *exponential-time*) computational problem is a problem where the time needed to obtain a solution increases at least exponentially with the size of the input (number of input parameters) [47]. Similarly, we can define a *hard inverse problem* as a problem whose solution time increases faster than exponentially with the *dimension* of the model space (the number of unknown parameters).

Being an exponential-time problem is (as suggested by the terminology) a property that is independent of the algorithm used for the analysis. In fact, we follow the convention used in the theory of algorithmic complexity, and say that a problem is 'exponential-time' (or hard) when a best conceivable algorithm used on the 'worst case' version of the problem is exponential-time [47].

Full probabilistic analysis of an inverse problem, using Monte Carlo

sampling or any other conceivable method, requires that we can collect sufficient information about the posterior probability density to ensure that an approximate reconstruction of the density is possible (Figure 4). If it is known a priori that we are sampling, e.g., a Gaussian posterior, an efficient sampling scheme can be designed [48]. However, if for instance the structure of the posterior is only known to be smooth in some sense, sampling can be a huge computational task.

The difficulty of solving highly nonlinear inverse problems is illustrated by three of the four specific problems, described in an earlier section (page 29-30). Since it is debatable whether Problem 2 is rightly labeled 'highly nonlinear', we will not consider it here.

In Tables 1 and 2 the size, number of iterations, and an assessment of the complication of Problems 1, 3 and 4 are summarized. In all

	# data	# parameters	# iterations	sample size
1. Lithosphere seis.	2560	128	100000	1000
3. Core fluid	390	390	400000	200
4. Lunar seis.	354	450	100000	1000

Table 1: Number of data and model parameters, number of iterations performed by the Monte Carlo samplers, and sample size used in the final analysis of the posterior probability densities.

	Prior	Likelihood	Posterior
1. Litho..	Pronounced non-Gauss	Moderately non-Gauss	Pron. non-Gauss
2. Core..	Gauss	Pronounced non-Gauss	Pron. non-Gauss
3. Lunar..	Constant over interval	Pronounced non-Gauss	Pron. non-Gauss

Table 2: Qualitative characteristics of the prior, the likelihood, and the posterior for the four inverse problems. The table should be interpreted with caution, as the description of the characteristics is semi-subjective.

three problems, the amount of computational work reached a practical maximum for a work station-sized computer at the time. As can be seen in Table 1, sample sizes exceeding the number of model parameters were only obtained in Problems 1 and 4. Since any N points distributed in $\mathbb{R}^M$ will always be located in a linear subspace of dimension less than or equal to $N-1$, we would expect that $M+1$ sample points, at the very least, are required to characterize a probability density in an M-dimensional model space. This observation leads us to conclude

that the sample obtained in Problem 3 is likely to be inadequate. Unfortunately, a further reduction in the number of model parameters in Problem 3 would have resulted in a too coarse model, so there was no way to improve the ratio between sample size and number of model parameters.

For Problems 1 and 4, however, it was possible (but only just possible) to obtain a satisfactory sample within the practical limits. Considerable experimentation was done in order to determine the maximum number of model parameters for which a satisfactory sample could be obtained. During this experimentation a 'brick wall effect' was observed, namely that the problem was practically solvable up to a certain number of unknown model parameters, but required excessive computer resources if only a few more model parameters were added.

For certain problems, it is likely that the brick wall effect can be attributed to an exponential increase in the number of secondary maxima for the posterior probability density [22] [Appendix A]. Some inverse problems in seismology are likely to suffer from this difficulty. In analysis of inverse problems involving near-vertical incidence reflection seismic data, where either short-wavelength acoustic impedances and large-scale variations in propagation velocities, or reflection coefficients and reflector depths are sought, it can be shown that the posterior probability density is highly oscillatory along directions of changing depth- or velocity parameters [15, 16, 36]. This behavior may result in the presence of multiple maxima for the probability density (see Figure 1). For such problems, an increase in the number of model parameters, due to an extension of the Earth model, is likely to result in an exponential increase in the number of secondary maxima. However, a full clarification of these issues must await further research.

Intuitively, one would expect that problems with none or few local maxima are comparatively easy to solve [22]. This statement, however, is apparently not true, since the posterior probability densities for the practically hard problems 1 and 2 are unlikely to possess a large number of maxima. Interestingly, it has been demonstrated that determination of the volume of an (unknown) n-gon shaped region[7] is a hard problem [49]. This problem is similar to the problem of characterizing a probability density which is constant over an n-gon shaped region, and zero elsewhere.

[7] An n-gon is a generalization of a polygon to an n-dimensional space. Volume determination algorithms visit one point at a time, and test if each point is inside or outside the unknown region. Such algorithms can be random or non-random.

Summary, conclusions and future perspectives

Reasoning draws a conclusion, but does not make the conclusion certain, unless the mind discovers it by the path of experience.

Roger Bacon

Probability theory provides a convenient framework for analysis of inverse problems, and in this formulation the solution is not a single model, but rather a probability density – the so-called posterior probability density – which combines data information, theoretical information and prior information.

When the posterior probability density is strongly non-Gaussian, either because of a non-linear relation between data and model, or because the noise probability density or prior probability density is non-Gaussian, the problem is often insusceptible to analytical treatment. For this reason we may resort to an algorithmic approach where information about the posterior probability density is gradually build up through sampling.

The conceptually most natural way of sampling the posterior probability density is by *random sampling* where points in the model space are generated according to the probability density. This is the idea behind the Metropolis algorithm and its extensions, which is the backbone of the work presented in this dissertation.

The Metropolis Algorithm is a rather efficient 'all-purpose' random sampling method, which had a renaissance in the beginning of the 80s where it had a dominant role in the newly discovered simulated annealing optimization method. The discovery of the simulated annealing method stimulated research in highly nonlinear inverse problems, where simulated annealing served as a way of finding best-fitting solutions. This research took analogies between simulated annealing and statistical mechanics even further, and resulted in improvements in efficiency which, for the first time, made the solution of intermediate-size, highly nonlinear problems practical.

In the 1990s a further step towards exploiting the full potential of random sampling methods were taken. Instead of focusing on locating

best-fitting solutions, the goal was now an approximate *characterization* of the posterior probability density through a large sample of models. Combined with strategies for translating sample properties into model properties, this method enabled an analysis of the resolution and noise propagation even for highly non-linear inverse problems.

Practical experience with this method over the last 10 years has been satisfactory, but it has also revealed a serious difficulty. Most of the considered problems were practically solvable up to a certain size (measured by the number of unknown model parameters), but above this, even a small increase in size led to a substantial increase in computation time needed to obtain satisfactory solutions. This indicated that some of the considered inverse problems are *hard*, that is, requires computation times that increase at least exponentially with the number of parameters.

To further investigate this, a branch of ongoing research is centered around the complexity of sampling problems. The aim of this research, which focuses on random sampling as well as its contrast, *nonprobability sampling*, is to identify hard inverse problems, and also to find ways of testing whether a problem is hard or not, before undertaking time-consuming computations.

The often complex structure of the posterior probability density for highly non-linear inverse problems, and problems with noise- and prior probability densities which are far from Gaussian, makes them difficult to analyze. Although random sampling strategies are natural means of solving such problems, they are probably not the fastest. Other alternatives have demonstrated their efficiency. Important examples are the genetic algorithms [50, 51], which are search algorithms based on ideas from biological evolution theories. Genetic algorithms are examples of nonprobability sampling, since they are not built to sample the given probability density. Instead, they aim at producing samples of 'large probability'. However, nonprobability sampling algorithms like the genetic algorithm *can* be used to generate random samples, but only if their output is appropriately post-processed.

A recent example of a nonprobability sampling algorithm developed for analysis of non-linear inverse problems is the *Neighborhood sampling algorithm* [52, 53, 54]. This algorithm is interesting in that, in each step, it uses previous samples to generate a rough (in fact, piecewise constant) approximation to the posterior probability density, and in this way paves the way for further sampling of high-probability regions in the model space. In this way it contrasts with the 'short-

memory' Metropolis approach, which is a Markov chain method where any step of the algorithm depends only on the last sample point obtained. There is no doubt that the neighborhood algorithm will be a pioneer for a new class of sampling methods with long memories, taking advantage of information from all previous sample points.

Two complications are important for the use of sampling methods in probabilistic inversion. The first complication is essentially the danger of conflicting definitions of prior information on fundamental physical parameters that are interrelated (e.g., conductivity and resistivity). To avoid potential conflicts between such a pair of inherently positive, mutually reciprocal, equally worthy, and fundamental physical parameters, one is lead to accept a non-constant probability density to represent the 'least informative prior' (the homogeneous probability density). If a metric is defined in the model space, the resulting definition of volume enables us to define the homogeneous probability density as one that assigns equal probabilities to equal-sized volumes. At the same time, the metric allows us to solve another problem which is encountered in probabilistic formulations of inverse problems, using the notion of 'conditional probability'. When a classical definition of this notion is used in the definition of the likelihood function, the inversion result may depend on the chosen parameterization. This unwanted problem can be removed by introducing an invariant definition of conditional probability, using orthogonal distances in the combined data/model-space metric when calculating limit functions.

The ability to incorporate complex, realistic prior information into the inversion is one of the attractions of the probabilistic approach, combined with sampling strategies. The way prior information about a physical system is obtained is often by studying a number of realizations of the system, observed in nature [32]. In the most complex of these situations, the problem of 'learning' the probabilistic rules behind such 'examples' is in itself a challenge. If the prior information is to be extracted from, say, a number of geological cross sections through the earth, recent advances in geostatistics [55, 56] may provide methods that could be integrated into the inversion. Combination of probabilistic, sample-based inversion with modern geostatistics is one of the promising future areas in this field of research.

References

[1] Mosegaard, K., and Tarantola, A., 2002: Probabilistic Approach to Inverse Problems: Chapter for the International Handbook of Earthquake and Engineering Seismology p 237-265. (Academic Press, ISBN 0-12-440652-1) Published for the International Association of Seismology and Physics of the Earth Interior (IASPEI). *

[2] Buffon, G., 1777. Essai d'arithmetique morale.

[3] Housholder, A.S. (ed), 1951. Monte Carlo Method (Mathematics Series 12): National Bureau of Standards, Washington DC.

[4] Fishman, G. S., 1996. Monte Carlo. Concepts, Algorithms, and Applications: Springer, New York.

[5] Metropolis, N., Rosenbluth, M.N., Rosenbluth, A.W., Teller, A.H. and Teller, E., 1953. Equation of state calculations by fast computing machines: J. Chem. Phys. 21, pp. 1087-1092.

[6] Mosegaard, K., and Sambridge, M., 2002. Monte Carlo analysis of inverse problems: Inverse Problems 18, pp. R29-R54. *

[7] Kaipio, J.P., Kolehmainen, V., Somersalo, E., and Vauhkonen, M., 2000. Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography: Inverse Problems 16 (5), pp. 1487-1522.

[8] Geman, S. and Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, pp. 721-741.

[9] Kirkpatrick, S.C., Gelatt, D. and Vecchi M.P., 1983. Optimization by simulated annealing: Science 220, pp. 671-680.

[10] Rothman, D.H., 1985. Nonlinear inversion statistical mechanics, and residual statics corrections: Geophysics 50, pp. 2784-2796.

[11] Rothman, D.H., 1986. Automatic estimation of large residual statics corrections: Geophysics 51, 332-346.

[12] Pedersen, J.M., 1990. Simulated annealing and finite-time thermodynamics: Ph.D. dissertation, University of Copenhagen, Physics Institute.

[13] Nulton, J.D. and Salamon P., 1988. Statistical mechanics of combinatorial optimization: Phys. Rev. A 37, pp. 1351-1356.

[14] Andresen, B., Hoffman K.H., Mosegaard K., Nulton J., Pedersen J.M. and Salamon P., 1988. On lumped models for thermodynamic properties of simulated annealing problems: J. Phys. France 49, 1485.

[15] Mosegaard, K., and Vestergaard, P.D., 1991. A simulated annealing approach to seismic model optimization with sparse prior information: Geophysical Prospecting 39, pp. 599-611. *

[16] Vestergaard, P.D., and Mosegaard, K., 1991. Inversion of post-stack seismic data using simulated annealing: Geophysical Prospecting 39, p. 613-624. *

[17] Keilis-Borok V.I. and Yanovskaya T.B., 1967. Inverse problems of seismology: Geophys. J. 13, pp. 223-234.

[18] Press F., 1968. Earth models obtained by Monte Carlo inversion: J. Geophys. Res. 73, pp. 5223-5234.

[19] Press F., 1970a. Earth models consistent with geophysical data: Phys. Earth Planet. Inter. 3, pp. 3-22.

[20] Press F., 1970b. Regionalized Earth models: J. Geophys. Res. 75, pp. 6575-6581.

[21] Sambridge, M., and Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems: Reviews of Geophysics 40, 3. September 2002, pp. (3-1)-(3-29).

[22] Hongling L.D., and J. A. Scales, 1999. Estimating the topography of multi-dimensional fitness functions: Cent. for Wave Phe- nomena Tech. Rep. 208, Colo. Sch. of Mines, Golden.

[23] Nørmark, E., and Mosegaard, K., 1993. Residual statics estimation: scaling temperature schedules using simulated annealing: Geophysical Prospecting 41, pp. 565-578. *

[24] Mosegaard, K., 1998. Resolution Analysis of General Inverse Problems through Inverse Monte Carlo Sampling: Inverse Problems 14, pp. 405-426. *

[25] Mosegaard, K., and Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems: Journal of Geophysical Research 100, B7, pp. 12431-12447. *

[26] Bosch M., 1999. Lithologic tomography: From plural geophysical data to lithology estimation: Journal of Geophysical Research 104, B1, pp. 749-766.

[27] Backus, G.E., and Gilbert, J.F., 1967. Numerical applications of a formalism for geophysical inverse problems, Geophys. J.R. Astron. Soc., 13, pp. 247-276.

[28] Backus, G.E., and Gilbert, J.F., 1968. The resolving power of gross Earth data, Geophys. J. R. Astron. Soc., 16, pp. 169-205.

[29] Backus, G.E., and Gilbert, J.F., 1970. Uniqueness in the inversion of inaccurate gross Earth data: Philos. Trans. R. Soc. London, Ser. A., 266, pp. 123-192.

[30] Bakhvalov, N.S., 1977. Numerical Methods: Mir publishers, Moscow.

[31] Hammersley, J. M., and Handscomb, D. C., 1964. Monte-Carlo methods, Chapman and Hall.

[32] Mosegaard, K., Singh, S.C., Snyder, D., and Wagner, H., 1997. Monte Carlo Analysis of seismic reflections from Moho and the W-reflector: Journal of Geophysical Research B, 102, pp. 2969-2981. *

[33] Singh, S. C., and D. P. McKenzie, 1993. Layering in the lower crust: Geophys. J. Int. 113, pp. 622-628.

[34] Weibe, R.A., 1993. The Pleasant Bay layered gabbro-diorite, coastal Maine: ponding and crystallization of basaltic injections into a silicic magma chamber: J. Petrology 34, pp. 461-489.

[35] Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem: Geophys. J. Int. 151 (3), pp. 675-688.

[36] Koren, Z., Mosegaard, K., Landa, E., Thore, P., and Tarantola, A., 1991. Seismic background velocity estimation and model error analysis by simulated annealing: Journal of Geophysical Research 96, B12, pp. 20289-20299.

[37] Dahl-Jensen, D. , Mosegaard, K., Gundestrup, N., Clow, G. D., Johnsen, S. J., Hansen, A. W., and Balling, N., 1998. Past temperatures directly from the Greenland Ice Sheet: Science, Oct. 9, pp. 268-271. *

[38] Mosegaard, K., and Rygaard-Hjalsted, C., 1999. Probabilistic analysis of implicit inverse problems: Inverse Problems 15, pp. 573-583. *

[39] Rygaard-Hjalsted, C., Mosegaard, K., and Olsen, N., 2000. Resolution studies of Fluid Flow Models Near the Core-Mantle Boundary through Bayesian Inversion of Geomagnetic Data, in: Methods and Applications of Inversion: Proc. of the IIC98 Conference, Copenhagen 1998 (eds. Hansen, P.C., Jacobsen, B.H., and Mosegaard, K., pp. 255-275.

[40] Khan, A., and Mosegaard, K., 2002. An Enquiry into the Lunar Interior: A Non-Linear Inversion of the Apollo Lunar Seismic Data: Journal of Geophysical Research (Planets), Vol 107, no. E6, pp. (3-1)-(3-18). *

[41] Tarantola, A., and Valette, B., 1982a. Inverse Problems = Quest for Information, J. Geophys., 50, 159-170.

[42] Tarantola, A., 1987. Inverse problem theory; methods for data fitting and model parameter estimation, Elsevier.

[43] Hastings, W., K., 1970. Monte Carlo sampling methods using Markov Chain and their applications, Biometrika 57, p. 97-109.

[44] Mosegaard, K., 1988. Stochastic model optimization in reflection seismology: Ph.D. thesis, Københavns Universitet.

[45] Gelman, A., Roberts, G.O., and Gilks, W.R., 1996. Efficient Metropolis Jumping Rules, in Bayesian Statistics 5, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, pp. 599-608, Clarendon press, Oxford.

[46] Propp, J.G., and Wilson, D.B., 1996. in Random Structures and Algorithms, Volume 9, Issue 1-2, pp. 223-252, John Wiley & Sons, Inc.

[47] Papadimitriou, C.H., and Steiglitz, 1998. Combinatorial Optimization: Algorithms and Complexity. Mineola, NY: Dover Publications, Inc.

[48] Gouveia, W. and J.A. Scales, 1997. Resolution of seismic waveform inversion: Bayes versus Occam: Inverse Problems, 13, pp. 323-349.

[49] Khachiyan, L.G., 1989. The problem of computing the volume of polytopes is np-hard: Uspekhi Mat. Nauk 44, pp. 199-200.

[50] Sen, M. K., and P. L. Stoffa, 1992. Rapid sampling of model space using genetic algorithms: Geophys. J. Int., 108, pp. 281-292.

[51] Lomax, A., and R. Snieder, 1995. Identifying sets of acceptable solutions to nonlinear geophysical inverse problems which have complicated misfit functions: Nonlinear Proc. Geophys., 2, pp. 222-227.

[52] Sambridge, M., 1998. Exploring multi-dimensional landscapes without a map: Inverse Problems 14, No. 3, pp. 427-440.

[53] Sambridge, M., 1999a. Geophysical inversion with a Neighborhood algorithm - I. Searching a parameter space: Geophysical Journal International 138, pp. 479-494.

[54] Sambridge, M., 1999b. Geophysical inversion with a Neighborhood algorithm - II. Appraising the ensemble: Geophysical Journal International 138, pp. 727-746.

[55] Srivastava, M., 1994. The visualization of spatial uncertainty: in Stochastic Modeling and Geostatistics (eds. Yarus J.M. and Chambers R.L.) AAPG Computer Applications in Geology, no. 3, pp. 339-346.

[56] Wang, L., 1996. Modeling complex reservoir geometries with multiple-point statistics: Mathematical Geology, 28, pp. 895-908.

[57] Girosi, F., and Poggio, T., 1990. Networks and the best approximation property: Biological Cybernetics 63, pp. 169-176.

Dansk Sammendrag

Sandsynlighedsteori udgør en bekvem ramme for analyse af inverse problemer, og i denne formulering er løsningen ikke en enkelt model, men en sandsynlighedsfordeling – den såkaldte a posteriori fordeling – som kombinerer data, teoretisk information og a priori information.

Når a posteriori fordelingen er stærkt ikke-Gaussisk, enten på grund af en ikke-lineær relation mellem data og model, eller fordi støjfordelingen eller a priorifordelingen er ikke-Gaussisk, er problemet ofte utilgængeligt for analytisk behandling. Af denne grund er vi henvist til en algoritmisk tilgangsvinkel, hvor information om a posteriori fordelingen gradvist opbygges gennem sampling.

Den begrebsmæssigt mest direkte måde at sample a posteriori fordelingen på, er ved tilfældig sampling, hvor punkter i modelrummet genereres i overensstemmelse med fordelingen. Dette er netop ideen bag Metropolisalgoritmen og dens udvidelser, som er ryggraden i det arbejde, der præsenteres i denne afhandling.

Metropolisalgoritmen er en ret effektiv allround samplingmetode, som fik en renæssance i begyndelsen af 80erne, hvor den havde en dominerende rolle i den nyopdagede metode *simuleret udglødning*. Opdagelsen af simuleret udglødning stimulerede forskningen i stærkt ikke-lineære inverse problemer, hvor metoden kunne bruges til at finde optimale løsninger. Denne forskning bragte analogierne mellem simuleret udglødning og statistisk mekanik et skridt videre, og dette resulterede i forbedringer i effektiviteten, som for første gang muliggjorde løsning af mellemstore, stærkt ikke-lineære inverse problemer.

I 90erne blev der taget yderligere skridt til udnyttelse af samplingalgoritmers potentiale ved løsningen af inverse problemer. I stedet for at fokusere på lokalisering af optimale løsninger, var målet nu en omtrentlig karakterisering af a posteriori fordelingen ved hjælp af en stor stikprøve fra modelrummet. Kombineret med strategier til oversættelse af stikprøveegenskaber til egenskaber ved løsningen, muliggjorde denne metode en analyse af opløsningsevne og støjpropagering, selv for stærkt ikke-lineære problemer.

Praktisk erfaring med denne metode har i de sidste 10 år været tilfredsstillende, men den har også afsløret en alvorlig vanskelighed. De fleste af de undersøgte problemer var praktisk løselige op til en vis størrelse (målt ved antallet af ubekendte parametre), men over denne størrelse førte selv en lille øgning af problemstørrelsen til en betydelig øgning i den regnetid, der var påkrævet for at opnå acceptable løsninger. Dette indikerede, at nogle af de betragtede inverse problemer er 'svære',

dvs kræver regnetider, som vokser eksponentielt med antallet af modelparametre.

For at undersøge dette nærmere, er en gren af den igangværende forskning koncentreret om kompleksiteten af samplingproblemer. Formålet med denne forskning, som fokuserer på probabilistiske såvel som ikke-probabilistiske samplingmetoder, er at identificere svære inverse problemer, og endvidere at finde måder, hvorpå det kan testes om et problem er svært eller ikke, før tidrøvende beregninger påbegyndes.

Den komplekse struktur af a posteriori-fordelingen for stærkt ikke-lineære inverse problemer og for problemer med stærkt ikke-Gaussiske støj- og a priorifordelinger gør den vanskelig at analysere. Skønt probabilistiske samplingstrategier er naturlige måder at løse disse problemer på, er de muligvis ikke de hurtigste. Andre metoder har demonstreret deres effektivitet. Vigtige eksempler er genetiske algoritmer, der er søgealgoritmer, som baserer sig på analogier til biologiske evolutionsteorier. Genetiske algoritmer [50, 51] er eksempler på ikke-probabilistiske samplingmetoder, eftersom de ikke er bygget til at sample en given sandsynlighedsfordeling. I stedet søger de at producere nærtoptimale modeller. Ikke-probabilistiske samplingalgoritmer som de genetiske algoritmer *kan* dog bruges til at generere samples fra en given fordeling, men kun hvis deres output efterbehandles på passende vis.

Et nyere eksempel på ikke-probabilistisk sampling, udviklet til analyse af ikke-lineære inverse problemer, er *Neighborhood sampling algoritmen* [52, 53, 54]. Denne algoritme er interessant, idet den i hvert skridt bruger tidligere stikprøver til at generere en grov (faktisk stykkevist konstant) approksimation til a posteriorifordelingen, og på denne måde bereder vejen for yderligere sampling af høj-sandsynlighedsområder i modelrummet. På denne måde står metoden i kontrast til Metropolis algoritmen, der som Markovkæde har kortest mulig 'hukommelse', hvor ethvert trin af algoritmen kun afhænger af den sidste stikprøve. Der er ingen tvivl om, at Neighborhood algoritmen will blive pioner for en ny klasse af samplingmetoder med lang hukommelse, som drager fordel af information fra alle tidligere samplepunkter.

To komplikationer af stor vigtighed for anvendelsen af samplingmetoder kan opstå når et inverst problem formuleres sandsynlighedsteoretisk. Den første komplikation består i faren for modstridende definitioner af a priori information om fundamentale fysiske parametre, som er indbyrdes relaterede (fx elektrisk ledningsevne og elektrisk modstand). For at undgå potentielle konflikter mellem sådanne par af positive, indbyrdes reciprokke, ligeværdige, og fundamentale fysis-

ke parametre, føres vi til at acceptere en ikke-konstant fordeling som den 'mindst informative' a priori fordeling (den homogene sandsynlighedsfordeling). Hvis man definerer en metrik i modelrummet, vil den resulterende definition af volumen muliggøre definition af en homogen sandsynlighedsfordeling som den, der tilordner lige store sandsynligheder til lige store volumener. På samme tid tillader metrikken os at løse et andet problem, som man støder på i sandsynlighedsteoretiske formuleringer af inverse problemer, hvor begrebet 'betinget sandsynlighed' indgår. Når den klassiske definition af dette begreb bruges i definitionen af likelihoodfunktionen, vil inversionsresultaterne nemlig afhænge af den valgte parametrisering. Dette uønskede problem kan fjernes ved en invariant definition af betinget sandsynlig, hvor man bruger ortogonale afstande i metrikken i det kombinerede data/modelrum, når der beregnes grænsefunktioner.

Mulighederne for at indbygge kompleks og realistisk a priori information ind i en inversion, er et af de tiltrækkende elementer i den sandsynlighedsteoretiske formulering, når den er kombineret med samplingstrategier. A priori information om et fysisk system indsamles ofte ved at studere en række realisationer af systemet i naturen [32]. I de mest komplekse af disse situationer er selve det at 'lære' den bagvedliggende sandsynlighedsmodel for sådanne eksempler i sig selv en stor udfordring. Hvis a priori informationen fx skal udledes af et antal geologiske tværsnit gennem jorden, kan nyere bidrag til geostatistisk teori [55, 56] muligvis vise vejen til metoder, som kan integreres med inversionen. Kombination af sandsynlighedsteoretisk, samplingbaseret inversion med moderne geostatistik er et af de lovende, fremtidige forskningsområder.

Appendices

Appendix A: On the complexity of characterizing a probability density

We shall consider the family of probability densities which in the neural network literature is termed 'mixture distributions'. This family of probability densities is particularly interesting because it provides good approximations to a wide range of multivariate probability densities seen in practice [57]. Analyzing this family essentially means analyzing all smooth multivariate probability densities, and hence most such probability densities arising in physical problems.

Consider in $\mathbb{R}^n$ an n-dimensional cube $\mathcal{M}_n$ of edge length L, and the class $\mathcal{F}_n$ of normalizable (but not neccesarily normalized) probability densities defined over $\mathcal{M}_n$, given by

$$f(\mathbf{x}) = \sum_{k=1}^{K(n)} u_k \, \phi_k(\mathbf{x}) \tag{21}$$

where the functions $\phi_1, \phi_2, \ldots, \phi_{K(n)}$ is a set of non-negative-valued, smooth 'basis functions' defined over the n-dimensional manifold $\mathcal{M}_n$. Assume that all basis functions have the same 'shape', but are centered differently. In other words, for any two different basis functions ϕ_k and ϕ_l centered at $\mathbf{x}_k$ and $\mathbf{x}_l$, respectively, we have:

$$\phi_k(\mathbf{x}_k + \mathbf{x}) = \phi_l(\mathbf{x}_l + \mathbf{x}) \; . \tag{22}$$

Furthermore, assume that the base functions are 'localized' in the sense that there exists a number $R > 0$ such that, for all $k = 1 \ldots K(n)$, the base function $\phi_k(\mathbf{x})$ is zero outside a ball of radius R, centered at $\mathbf{x}_k$.

Since the center $\mathbf{x}_k$ of each of the $K(n)$ basis functions $\phi_k(\mathbf{x})$ are given by n parameters, we can conclude that (for basis functions with a fixed 'shape') $\mathcal{F}_n$ is a manifold of dimension

$$\nu(n) \; = \; K(n) \, (1 + n) \, . \tag{23}$$

We shall now evaluate the number of sample points needed to approximate $f(\mathbf{x})$ in the worst case. To do this, we shall use a non-trivial result, proved by Brouwer around 1910 (see Brouwer, 1976):

Theorem 1 (Invariance of dimension). *No bijective, continuous mapping exists between an m-dimensional manifold and an $(m + h)$ -dimensional manifold $(h > 0)$.*

In the special case where we were to reconstruct the particular mixture probability density $f \in \mathcal{F}_n$ exactly, Brouwer's theorem would require that f should be evaluated in at least $\nu(n)$ different points. However, if we only want to *approximate* f, we might expect to do that from only $\nu(n) - 1$ (or fewer) function evaluations. To show that an acceptable approximation to f cannot be guaranteed in this case, assume that f is a (to us unknown) superposition of 'non-overlapping' basis functions. Under these assumptions, the determination of f can essentially be decomposed into $K(n)$ independent determinations of basis functions. If only $\nu(n) - 1$ (or fewer) function evaluations are available, this means that at least one of the basis functions is undetermined. In fact, there exists in this case a manifold of dimension at least 1 containing significantly different solutions to the f reconstruction problem. Consequently, $\nu(n) - 1$ function evaluations cannot provide even a reasonable approximation to f. We are forced to conclude that any sampling algorithm requires at least $\nu(n)$ function evaluations to approximate a 'worst-case' mixture probability density f. In short, the complexity of characterizing this type of mixture probability density is $\nu(n)$.

The remaining issue is the factor $K(n)$ in equation 23. If $K(n)$ (the number of basis functions needed to represent f) increases at least exponentially with n, then the problem of approximating the mixture distribution from a sample is hard, and will remain practically unsolvable in high-dimensional parameter spaces. It is likely that this situation occurs when solving certain highly nonlinear inverse problems in seismology, but further analysis of concrete problems are needed to demonstrate this.

Appendix B1–B11: Papers included in this dissertation

A note to the reader

Two of the papers in this collection are review papers, which, in the nature of the case, introduce some redundancy into the material. They are, however, included for the following reasons:

The paper [24] (Appendix B6) is included because it provides additional insight into the technical details of the lithosphere-seismic inversion, presented in a companion paper [32] (Appendix B5). Furthermore, it emphasizes for the first time, I believe, the idea that the difficulty of the so-called nonlinear problems stems from our lack of information about the problem structure, rather than on the actual complexity of the posterior probability density for the problem.

The paper [6] (Appendix B11) is included because it provides further technical details on the inversion of the Lunar seismic data, which are not available in the original paper [40] (Appendix B10). Last, but not least, it contains a more satisfactory historic account of the development of Monte Carlo methods, than is presented in the first chapters of this dissertation. Readers who are looking for even more detail should consult [21].

The book chapter [1] started out as a review, but turned (despite some resistance from the editor!) into an independent publication where, to my knowledge, the problem of invariance of an inverse problem is treated for the first time.

Included papers

B1. **Mosegaard, K.**, and Vestergaard, P.D., 1991. A simulated annealing approach to seismic model optimization with sparse prior information: Geophysical Prospecting 39, pp. 599-611.

B2. Vestergaard, P.D., and **Mosegaard, K.**, 1991. Inversion of post-stack seismic data using simulated annealing: Geophysical Prospecting 39, pp. 613 -624.

B3. Nørmark, E., and **Mosegaard, K.**, 1993. Residual statics estimation: scaling temperature schedules using simulated annealing: Geophysical Prospecting 41, pp. 565-578.

B4. **Mosegaard, K.**, and Tarantola, A., 1995, Monte Carlo sampling of solutions to inverse problems: Journal of Geophysical Research 100, B7, pp. 12431-12447.

B5. **Mosegaard, K.**, Singh, S.C., Snyder, D., and Wagner, H., 1997. Monte Carlo Analysis of seismic reflections from Moho and the W-reflector: Journal of Geophysical Research B, 102, pp. 2969-2981.

B6. **Mosegaard, K.**, 1998. Resolution Analysis of General Inverse Problems through Inverse Monte Carlo Sampling: Inverse Problems 14, pp. 405-426.

B7. Dahl-Jensen, D., **Mosegaard, K.**, Gundestrup, N., Clow, G. D., Johnsen, S. J., Hansen, A. W., and Balling, N., 1998. Past temperatures directly from the Greenland Ice Sheet: Science, Oct. 9: pp. 268-271.

B8. **Mosegaard, K.**, and Rygaard-Hjalsted, C., 1999. Probabilistic analysis of implicit inverse problems: Inverse Problems 15, pp. 573-583.

B9. **Mosegaard, K.**, and Tarantola, A., 2002. Probabilistic Approach to Inverse Problems: Chapter for the International Handbook of Earthquake and Engineering Seismology, pp. 237-265., Academic Press, ISBN 0-12-440652-1. Published for the International Association of Seismology and Physics of the Earth Interior (IASPEI).

B10. Khan, A., and **Mosegaard, K.**, 2002. An Enquiry into the Lunar Interior: A Non-Linear Inversion of the Apollo Lunar Seismic Data: Journal of Geophysical Research (Planets) Vol 107, no. E6, pp. (3-1)-(3-18).

B11. **Mosegaard, K.**, and Sambridge, M., 2002. Monte Carlo analysis of inverse problems: Inverse Problems 18, pp. R29-R54.

Errata for Appendices B1-B11

B3. Page 567, line 12 from below: *marginal* should read *conditional*.

B5. Page 2975, Figure 4*b*: The unit on the abscissa should read 30×10^{-2}.

B5. Page 2975, figure caption for Figure 4*b*: $\lambda = 225.0\ s^{-1}$ should read $\lambda = 22.5\ s^{-1}$.

B6. Page 407, line 7 from below: The right-hand-side of the equation is missing the factor $1/N$.

B6. Page 419, Figure 4*b*: The unit on the abscissa should read 30×10^{-2}.

B6. Page 414, line 5 and 11 from below: $\rho_{\mathcal{D}}(\mathbf{m})$ should read $\sigma_{\mathcal{D}}(\mathbf{m})$.

Geophysical Prospecting **39,** 599–611, 1991

A SIMULATED ANNEALING APPROACH TO SEISMIC MODEL OPTIMIZATION WITH SPARSE PRIOR INFORMATION[1]

KLAUS MOSEGAARD[2] and PETER D. VESTERGAARD[3]

ABSTRACT

MOSEGAARD, K. and VESTERGAARD, P.D. 1991. A simulated annealing approach to seismic model optimization with sparse prior information. *Geophysical Prospecting* **39,** 599–611.

It is well known that seismic inversion based on local model optimization methods, such as iterative use of linear optimization, may fail when prior information is sparse. Where the seismic events corresponding to reflectors of interest remain to be identified, a global optimization technique is required.

We investigate the use of a global, stochastic optimization method, that of simulated annealing, to solve the seismic trace inversion problem, in which the two-way traveltimes and reflection coefficients are to be determined. The simulated annealing method is based on an analogy between the model-algorithm system and a statistical mechanical system. We exploit this analogy to produce improved annealing schedules. It is shown that even in cases of virtually no prior information about two-way traveltimes and reflection coefficients, the method is capable of producing reliable results.

INTRODUCTION

The seismic trace inversion problem can be formulated as a non-linear search for an acoustic impedance function that is stepwise constant in depth, and whose seismic response, modelled by means of the convolutional model, is as close as possible to the measured response in the least-squares sense. Cooke and Schneider (1983) found that for this problem, a traditional, local optimization method fails if not provided

[1] Received August 1989, revision accepted November 1990, last material received January 1991.

[2] Geophysical Institute, University of Copenhagen, Haraldsgade 6, 2200 Copenhagen N, Denmark.

[3] Imperial College, Department of Geology, Royal School of Mines, Prince Consort Road, London SW7 2BP, U.K. Formerly Ødegaard and Danneskiold-Samsøe Aps, Copenhagen.

with an initial guess for which the two-way traveltimes to the layer interfaces are closer to the true values than the side lobes of the corresponding seismic events.

In cases of good well control, detailed prior information about the subsurface model may be available, and a good initial guess may be made. In such cases, traditional trace inversion methods may prove successful. However, in many exploration problems and in the early phases of oil and gasfield developments, the well control is sparse and a good initial guess may not be available. In such cases the trace inversion problem becomes a global model optimization problem, as many secondary minima for the misfit function exist.

Global optimization problems are much more difficult to solve than local optimization problems, since the local geometry of the misfit function surface in the model space does not directly contain information about the direction to follow in a search for the global minimum. We have therefore applied a stochastic optimization technique, that of simulated annealing, in a global search for the optimal subsurface model. The usefulness of this method, when applied to global optimization problems in geophysics, has already been demonstrated by previous authors. The work by Rothman (1985, 1986) on the residual statics estimation problem was the first published application to geophysical inverse problems. Later contributions have been made by Jakobsen, Mosegaard and Pedersen (1988) and Landa, Beydoun and Tarantola (1989). The experience of these authors and others who have tried to apply simulated annealing to geophysical inverse problems of a realistic size, is that the method is very difficult to use. The main problems are: to discover the best annealing temperature schedule or the 'critical' temperatures, and how many times the annealing should be performed in order to arrive at a useful solution. Consequently, a high degree of experimentation has been an important characteristic of simulated annealing work. The aim of our work has been to investigate the efficiency and accuracy of a recently developed implementation of simulated annealing applied to a realistic, seismic trace inversion problem. This method, which is called 'simulated annealing at constant thermodynamic speed' (Nulton and Salamon 1988; Andresen *et al.* 1988), replaces the random experimental approach with a systematic approach that takes advantage of statistical information about the model-algorithm system, acquired during the annealing. As a standard of reference in our investigation, we have used a more primitive, stochastic model optimization method, the so-called iterative improvement, when investigating the efficiency and accuracy of the implementation of simulated annealing.

Theory and Implementation

Simulated annealing is a statistical technique for finding near-optimal solutions to complex optimization problems. In this technique, the state ω of the system being optimized is identified with the state of a statistical mechanical system, the objective function $E(\omega)$ being minimized is identified with the physical energy, and the optimization process is controlled by a parameter T which can be identified with the physical temperature. The system to be optimized is allowed to 'equilibrate' by

applying a set of moves, i.e. a set of system perturbations, and accepting or rejecting the moves according to the Metropolis algorithm (Metropolis *et al.* 1953):

if $E_{\text{attempted}} \leq E_{\text{current}}$, accept the move;

if $E_{\text{attempted}} > E_{\text{current}}$, accept the move with probability

$$P_{\text{accept}} = \exp\left(-\frac{(E_{\text{attempted}} - E_{\text{current}})}{T}\right). \tag{1}$$

Under the following rather mild assumptions, it can be shown that the system tends towards 'thermal equilibrium' when iterated at any temperature (Hammersley and Handscomb 1964). (1) Any state of the system to be optimized can be reached from any other state of the system, using the prescribed move class. (2) There must be non-zero probability of staying in the current state in a given 'move'.

In thermal equilibrium at temperature T, the states ω of a hypothetical, large statistical ensemble of systems, identical to the considered system, are distributed according to the Boltzmann distribution

$$P_{\text{B}}(\omega) = \frac{\exp\left(-\frac{E(\omega)}{T}\right)}{Z(T)}, \tag{2}$$

where the partition function

$$Z(T) = \sum_{\omega} \exp\left(-\frac{E(\omega)}{T}\right). \tag{3}$$

Assuming that only one global minimum exists for E, the equilibrium ratio between the probability p_0 that the system is in the configuration ω_0, representing the global minimum for the objective function E_0, and the probability p_ω that the system is in any other state ω, corresponding to a value E_ω for the objective function, is

$$\frac{p_0}{p_\omega} = \exp\left(-\frac{E_0 - E_\omega}{T}\right), \tag{4}$$

showing that

$$\frac{p_0}{p_\omega} \geq 1, \tag{5}$$

and

$$\frac{p_0}{p_\omega} \to \infty \qquad \text{as} \qquad T \to 0. \tag{6}$$

The limit (6) can be explained as follows: consider the family of Boltzmann distributions (2), parametrized by the temperature parameter T, and consider the probabilities p_0 and p_ω that the equilibrated system is in the global minimum for the objective function (the ground state) and in the fixed state $\omega \neq \omega_0$, respectively. If

we consider the probabilities as functions of the temperature T, the limit (6) is valid. Since

$$\forall \omega: p_\omega \leq 1, \tag{7}$$

and

$$\sum_\omega p_\omega = 1, \tag{8}$$

for a system with a finite number of states (6) yields

$$p_0 \to 1 \qquad \text{as} \qquad T \to 0. \tag{9}$$

Therefore, optimization of our system can be achieved by attaining equilibrium at a low value of T.

The limit (9) suggests the following simple algorithm, which is known as 'simulated annealing' (Kirkpatrick, Gelatt and Vecchi 1983). (1) Distribute a number of copies of the model-algorithm system uniformly over the state space. This uniform distribution corresponds to a Boltzmann distribution at $T = \infty$. (2) Decrease T gradually from infinity to zero over a large number of steps, and let the system equilibrate approximately at each step. After this process is terminated, we have for each system, $p_0 \approx 1$.

A serious problem with this algorithm is that at low values of T, approximate equilibrium can only be attained in a large number of steps. It is therefore necessary to run the algorithm when the system is out of equilibrium, and in this case, (9) does not apply.

For non-equilibrium systems with ground state probability $p_0(T)$, we have in general,

$$p_0(T) \to \pi_0 \leq 1 \qquad \text{as} \qquad T \to 0, \tag{10}$$

where the probability π_0 depends on the way the temperature decreases with time. We must now find the optimal annealing schedule $T(t)$, satisfying the constraints

$$T(0) \gg 1 \tag{11}$$

and

$$T(t_{\max}) = 0, \tag{12}$$

where t is the time measured in number of Metropolis moves.

The optimal annealing schedule $T(t)$ should maximize π_0 for a given, finite number of iterations $t_{\max}$ (a given run time).

A solution has been proposed by Nulton and Salamon (1988). They suggest that the optimal annealing schedule keeps a constant difference between the (non-equilibrium) mean value $\langle E \rangle$ of the objective function of the system, and the mean value $\langle E \rangle_{\text{eq}}$ which the objective function would have had, if the system were in equilibrium at the considered temperature. The distance should be measured in units of the standard deviation $\sigma_E(T)$ of the fluctuating objective function, i.e.

$$\frac{\langle E\rangle - \langle E\rangle_{eq}}{\sigma_E(T)} = v, \tag{13}$$

where v is a constant named 'the thermodynamic speed'.

Combining (13) with the differential equation for simple, thermodynamic relaxation

$$\frac{d\langle E\rangle}{dt} = -\frac{\langle E\rangle - \langle E\rangle_{eq}}{\varepsilon(T)}, \tag{14}$$

where $\varepsilon(T)$ is the relaxation time of the system at temperature T, and the relationship between $\sigma_E(T)$ and the heat capacity $C(T)$ of the system at temperature T is given by

$$C(T) = \frac{\sigma_E^2(T)}{T^2}, \tag{15}$$

the following first-order differential equation in $T(t)$ is obtained

$$\frac{dT}{dt} = -\frac{vT}{\varepsilon(T)\sqrt{C(T)}}. \tag{16}$$

Simulated annealing using a schedule satisfying (16) is denoted 'simulated annealing at constant thermodynamic speed'. This implementation of simulated annealing is superior to previously used implementations (Salamon *et al.* 1988; Jakobsen *et al.* 1988).

The constant thermodynamic speed v in (16) is adjusted such that the temperature schedule resulting from the integration of (16) satisfies (11) and (12).

Andresen *et al.* (1988) describe methods by which $C(T)$ and $\varepsilon(T)$ can be estimated from statistical information about the system, collected during the annealing process. They suggest that a transition frequency matrix $\mathbf{Q}$, for attempted transitions, is formed during the annealing process. For each attempted transition, the values of the objective function, $E_{current}$ and $E_{attempted}$, for the current model and the trial model respectively, are saved. The ijth element of the current $\mathbf{Q}$ can then be determined since

$$Q_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}, \tag{17}$$

where n_{ij} is the total number of attempted transitions (since the first iteration) between models having values of E between E_i and $E_i + \delta E$, and models for which E lies between E_j and $E_j + \delta E$. Here, δE is the constant difference between two successive, preselected levels, E_k and E_{k+1} of the objective function. A temperature dependent matrix $\mathbf{G}(T)$ can now be formed by multiplying elements of $\mathbf{Q}$ corresponding to $j > i$ with Boltzmann factors

$$\exp\left(-\frac{(E_j - E_i)}{T}\right), \tag{18}$$

and adjusting the diagonal elements of $\mathbf{G}(T)$ so as to keep the row sums equal to 1.

$\mathbf{G}(T)$ is a stochastic matrix (its row sum is equal to 1), and its largest eigenvalue is therefore 1. Let $\mathbf{p}$ be the corresponding, normalized eigenvector.

The temperature-dependent variance $\sigma_E^2(T)$ of the fluctuating objective function can now be estimated as

$$\sigma_E^2(T) = \langle E^2(T) \rangle - \langle E(T) \rangle^2, \tag{19}$$

where

$$\langle E^n \rangle = \sum_i (E_i)^n p_i. \tag{20}$$

The heat capacity is now calculated from (15).

If the second largest eigenvalue of $\mathbf{G}(T)$ is $\lambda_2(T)$, the relaxation time can be estimated as

$$\varepsilon(T) = -\frac{1}{\ln(\lambda_2)}. \tag{21}$$

Computer Simulations

We used simulated annealing at constant thermodynamic speed. We ran 50 system copies in parallel, each one having a unique starting model and a unique random number sequence.

Running several copies of the model-algorithm system enabled an assessment of the distribution of the computed model parameters. We used the dispersion of the final acoustic impedance values of the ten best models found, as a measure of the quality of the model optimization methods. This dispersion reflects the degree of non-uniqueness of the inverse problem, the influence of noise on the model estimate, and possible errors due to lack of convergence of the search algorithm.

In the present implementation of seismic trace inversion, the subsurface model was parametrized by the two-way traveltimes of the reflectors, and the acoustic impedances of the homogeneous layers between the reflectors. In the simulated annealing optimization, each move in the model space consisted of perturbing a randomly selected two-way traveltime or reflection coefficient, which obeyed the constraints imposed by the *a priori* information. Synthetic seismic traces were modelled by convolving subsurface reflectivities with a known, highly oscillatory wavelet, and the objective function used in the optimization was the energy of the difference between the true data and the modelled data. After the model optimization process, the surface acoustic impedance and the reflection coefficient series were mapped into acoustic impedance as a function of two-way traveltime.

Test Based on Real Well Data

The simulated annealing inversion method was compared with iterative improvement by solving a trace inversion problem in which the data consisted of a synthetic trace generated from well data. In the inversion we searched for an acoustic impedance function that was stepwise constant in depth and had a limited number of discontinuities, i.e. a 'blocked' impedance function.

Iterative improvement is one of the well-established traditional optimization

methods, and is therefore well suited as a standard of reference for the results obtained by simulated annealing. The iterative improvement method is very simple: a random move (obeying the move class specifications) in the model space is accepted only if it results in an improvement (a decrease) in the value of the objective function. It should be noted that iterative improvement is a local optimization method, and hence it is likely to converge towards local minima when used on the considered, global optimization problem.

The subsurface model used in the generation of the test data was derived from the sonic and the density logs of the onshore well Løgumkloster-1, situated in South Jutland in Denmark. Based on calibrated velocity and density logs, and a seismic wavelet extracted from a seismic line close to the well, a synthetic trace was generated (Fig. 2), using the detailed impedance function from the well (Fig. 1).

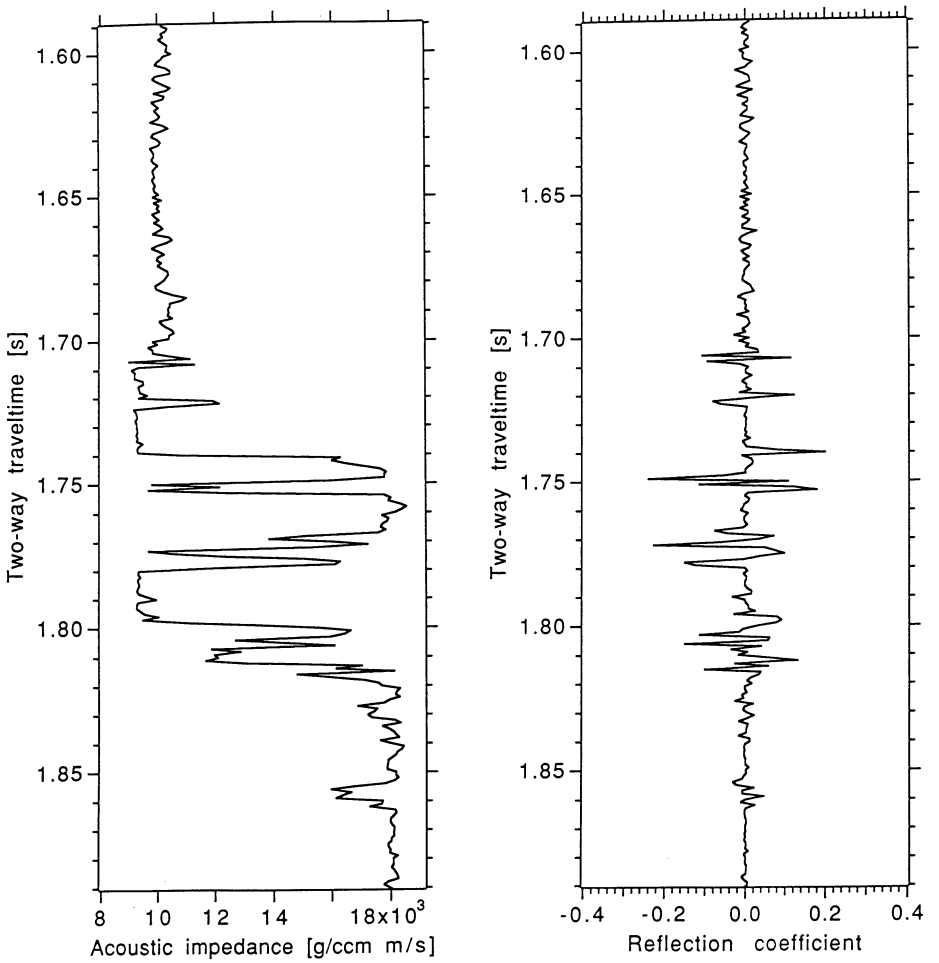


FIG. 1. The computed acoustic impedance (left) and reflection coefficients (right) versus two-way traveltime in the target zone.

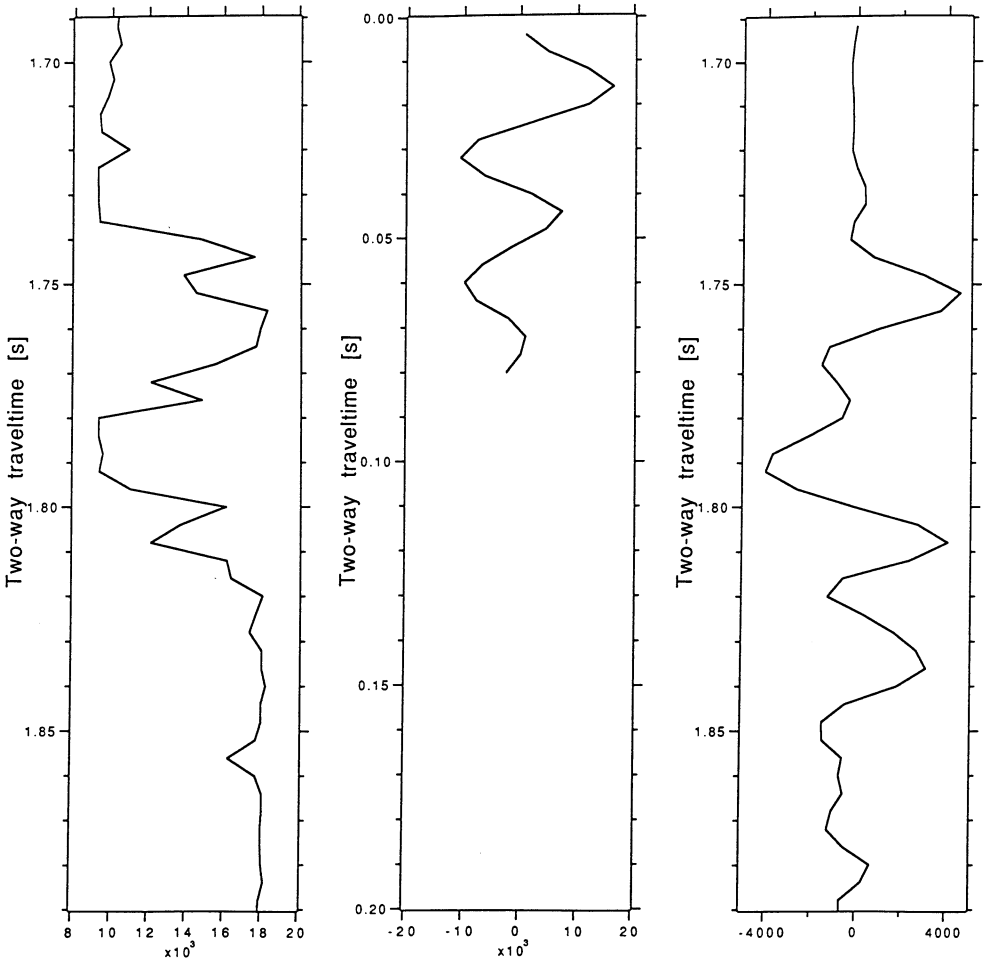


FIG. 2. The detailed acoustic impedance for the target zone (left), the estimated wavelet (centre), and the synthetic trace generated from the impedance log (right).

A manually blocked model for the Zechstein sequence is shown in Fig. 3. It is seen that ten homogeneous layers are sufficient to approximate the actual impedance log in this case. However, in order to simulate a realistic situation, we over-parametrized our model, assuming that it consisted of 15 homogeneous layers. The goal of the inversion was to reproduce approximately the well data from the seismic data and from the prior geological knowledge, without using the well log information.

The *a priori* information in the considered inverse problem is assumed to be sparse. The reflection coefficients are only known to be between -0.4 and 0.4, and there are no restrictions on the two-way traveltimes for the layer interfaces, except that they fall within the considered target zone.

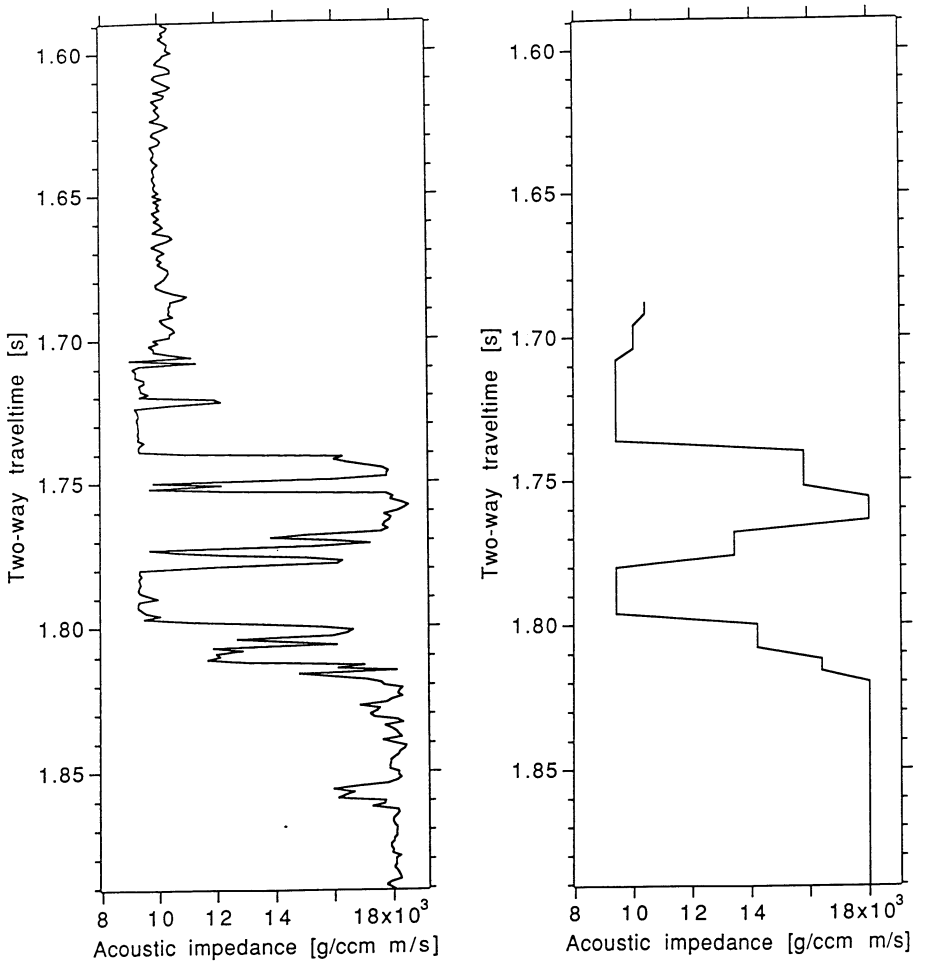


FIG. 3. The computed acoustic impedance (left) and the interpreter's blocking of the acoustic impedance (right) for the well Løgumkloster-1, resulting in a ten-layer model in the target zone.

The sparse prior information leads to multiple local minima for the objective function, typically appearing at parameter values representing cycle skips.

A 200 ms target zone from the synthetic reflection data set is inverted for a 15-layer impedance model. An important exploration problem to be solved in this area is to discriminate between salt layers and porous carbonate layers, on the basis of their acoustic impedance. In the data set used, a salt layer is found in the interval between 1.780 and 1.790 s, whilst the interval between 1.800 and 1.815 s is occupied by a porous carbonate.

An equal number of iterations were allocated to the two methods. First, 50 iterative improvement runs were performed, starting at different points, randomly

distributed in the model space. The individual runs were terminated when no significant changes in the value of the objective function had taken place within 500 iterations. The purpose of this terminating criterion was to optimize the use of the iterative improvement technique so that no time should be wasted by iterating after convergence to a local minimum had occurred. A total of 47 800 iterations were performed in this way.

Secondly, 50 simulated annealing runs were performed, all using the same number of iterations (namely 956), so that the total number of iterations was 47 800. Hence, no attempt was made to optimize the number of runs (or, equivalently, the number of iterations per run) performed by the annealing algorithm with the 47 800 iterations.

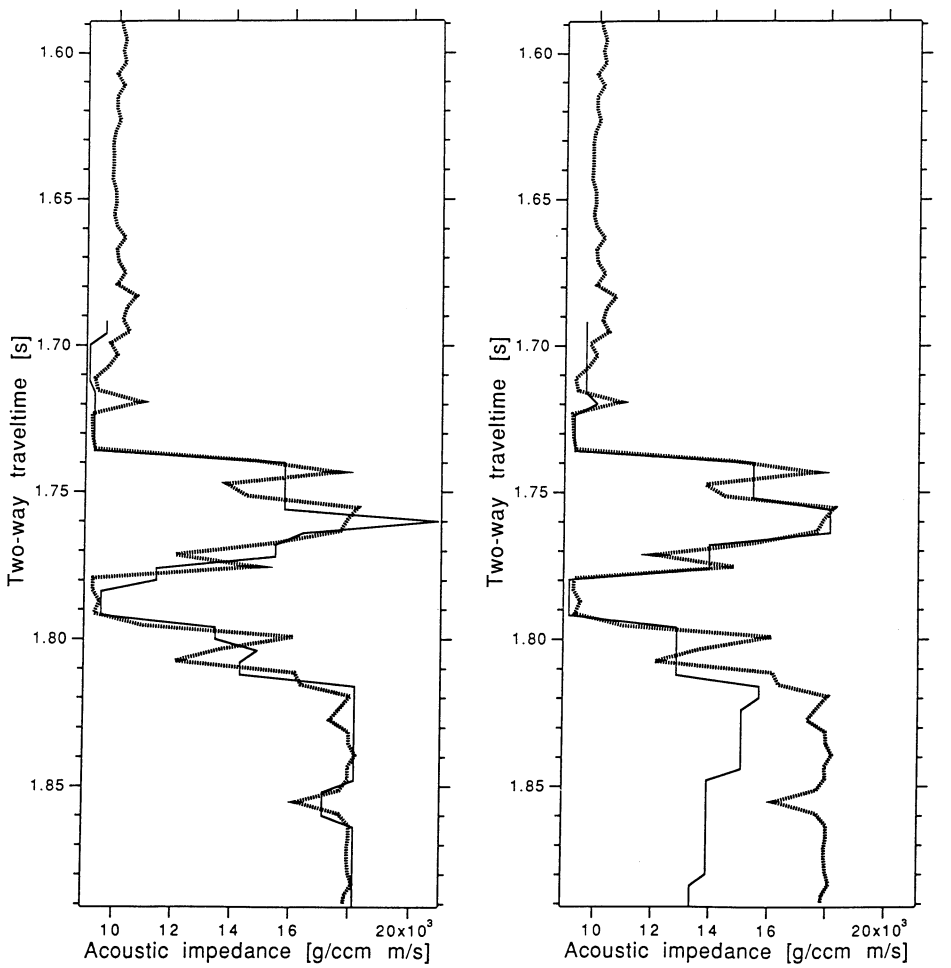


FIG. 4. The best acoustic impedances found by simulated annealing (left), and by iterative improvement (right). The true model is shown dotted.

In order to assess the performance of the two considered model optimization methods, we have chosen to display the ten best models obtained from each method, together with the true well data.

In this example, the chosen, blocked model parametrization is incapable of explaining all the data, due to the convolutional noise generated by the fine detail in the logs. This means that even the optimal solution has a finite, positive error energy. The best result obtained by simulated annealing is, however, in very good agreement with the well data (Fig. 4). In contrast to this, the best model obtained by iterative improvement fails to resemble the well data in the deeper part of the target zone. In particular, salt and porous carbonate are unlikely to be distinguished by inspection of the iterative improvement results. The near-optimal simulated annealing

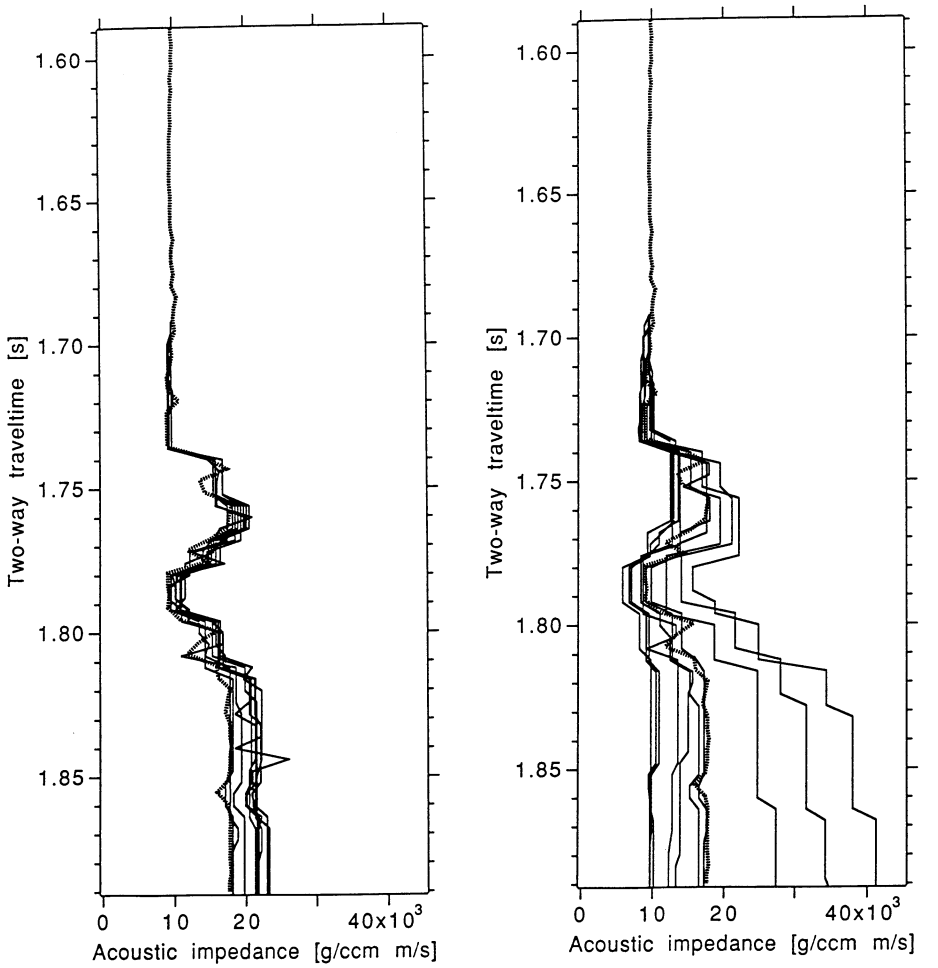


FIG. 5. The ten best impedance models obtained by simulated annealing (left) and by iterative improvement (right). The true model is shown dotted.

solution assigns separate impedance values to the two rock types, and hence enables an unambiguous rock identification.

To illustrate the uncertainties in the inversion procedures, the ten best computed models from the two methods are shown in Fig. 5. For the simulated annealing method, all the models shown are close to the true model, and the estimated impedances in the salt layer and in the porous carbonate layer are clearly separated. The best models from the iterative improvement method exhibit a gross divergence with increasing two-way traveltime, indicating that little confidence can be attached to the result of iterative improvement in this example.

Conclusion

We have investigated the use of a global, stochastic inverse method, i.e. simulated annealing, to solve the seismic trace inversion problem. The inverse problem was formulated as a search for a weakly constrained, blocked impedance function. The seismic data used were generated synthetically by convolving a known, highly oscillatory wavelet with the true reflectivity function from an onshore well located in South Jutland, Denmark. Hence, convolutional noise from a large number of thin layers in the subsurface was present in the data.

The simulated annealing algorithm employed in the present study was based on recent improvements by Nulton and Salamon (1988) and Andresen *et al.* (1988), in which statistical information about the system to be optimized is used to improve the performance of the algorithm. A traditional, stochastic model optimization method, iterative improvement, was used as a standard of reference when investigating the accuracy of the simulated annealing approach. 50 copies of the system, differing only in the starting models and the random sequences used, were run in parallel for both types of optimization. The dispersion of the ten best models obtained, and the similarity between these models and the true model were used as a measure of the performance of the algorithms.

The result of the investigation was that the best solutions found by the improved simulated annealing schedule displayed a significantly lower dispersion than the solutions found by iterative improvement, using the same total number of iterations. Furthermore, the best simulated annealing solutions were closest to the true model. The significance of these results was emphasized by the fact that only for iterative improvement was an attempt made to optimize the computations with respect to the number of iterations performed in each run.

The particular exploration problem of discriminating between salt and low porosity carbonate on the basis of the acoustic impedances of these rocks was solved by the simulated annealing approach, but remained unsolved by the iterative improvement algorithm.

Acknowledgements

This work was sponsored by The Danish Ministry of Energy under contract No. 1313/87-12. Klaus Mosegaard was partly sponsored by The Danish Natural Science Research Council. We thank Dr Daniel Rothman for his criticism and valuable

suggestions. We are indebted to Dr Jacob M. Pedersen, Ødegaard & Danneskiold-Samsøe, for assistance in the processing of our data.

REFERENCES

ANDRESEN, B., HOFFMANN, K.H., MOSEGAARD, K., NULTON, J.D., PEDERSEN, J.M. and SALAMON, P. 1988. On lumped models for thermodynamic properties of simulated annealing problems. *Journal de Physique* **49**, 1485–1492.

COOKE, D.A. and SCHNEIDER, W.A. 1983. Generalized linear inversion of reflection seismic data. *Geophysics* **48**, 665–676.

HAMMERSLEY, J.M. and HANDSCOMB, D.C. 1964. Monte Carlo Methods. In: Monographs on Statistics and Applied Probability. D.R. Cox and D.V. Hinkley (eds). Chapman and Hall.

JAKOBSEN, M.O., MOSEGAARD, K. and PEDERSEN, J.M. 1988. Global model optimization in reflection seismology by simulated annealing. In: *Model Optimization in Exploration Geophysics* **2**, p. 361. Proceedings of the 5th International Mathematical Geophysics Seminar, Berlin 1987, A. Vogel (ed.). Friedr. Vieweg & Son, Braunschweig, Wiesbaden.

KIRKPATRICK, S., GELATT, C.D. and VECCHI, M.P. 1983. Optimization by simulated annealing. *Science* **220**, 671–680.

LANDA, E., BEYDOUN, W. and TARANTOLA, A. 1989. Reference velocity model estimation from prestack waveforms: coherency optimization by simulated annealing. *Geophysics* **54**, 984–990.

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H. and TELLER, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **1**, 1087–1092.

NULTON, J.D. and SALAMON, P. 1988. Statistical mechanics of combinatorial optimization. *Physical Review* A **37**, 1351–1356.

ROTHMAN, D.H. 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics* **50**, 2797–2807.

ROTHMAN, D.H. 1986. Automatic estimation of large residual static corrections. *Geophysics* **51**, 332–346.

SALAMON, P., NULTON, J.D., HARLAND, J.R., PEDERSEN, J.M., RUPPEINER, G. and LIAO, L. 1988. Simulated annealing with constant thermodynamic speed. *Computer Physics Communications* **49**, 423–428.

Geophysical Prospecting **39**, 613–624, 1991

INVERSION OF POST-STACK SEISMIC DATA USING SIMULATED ANNEALING[1]

PETER D. VESTERGAARD[2] and KLAUS MOSEGAARD[3]

ABSTRACT

VESTERGAARD, P.D. and MOSEGAARD, K. 1991. Inversion of post-stack seismic data using simulated annealing. *Geophysical Prospecting* **39**, 613–624.

Model-based inversion of seismic reflection data is a global optimization problem when prior information is sparse. We investigate the use of an efficient, global, stochastic optimization method, that of simulated annealing, for determining the two-way traveltimes and the reflection coefficients.

We exploit the advantage of an ensemble approach to the inversion of full-scale target zones on 2D seismic sections.

In our ensemble approach, several copies of the model-algorithm system are run in parallel. In this way, estimation of true ensemble statistics for the process is made possible, and improved annealing schedules can be produced.

It is shown that the method can produce reliable results efficiently in the 2D case, even when prior information is sparse.

INTRODUCTION

Automatic inversion schemes for the reconstruction of subsurface structures from seismic reflection data are used more and more frequently in the oil industry for detailed studies of oil and gasfields during the development phases. In cases of good well control such methods have often produced satisfactory predictions concerning the lithological columns seen in wells drilled at later stages.

However, it is frequently observed that surprisingly large errors in the prediction of reflector locations and acoustic impedance values occur in cases where the well

[1] Received November 1989, revision accepted December 1990, last material received January 1991.

[2] Imperial College, Department of Geology, Royal School of Mines, Prince Consort Road, London SW7 2BP, U.K. Formerly Ødegaard and Danneskiold-Samsøe ApS, Copenhagen.

[3] Geophysical Institute, University of Copenhagen, Haraldsgade 6, 2200 Copenhagen N, Denmark.

control is sparse or in cases where the correlation between seismic events and nearby wells is made difficult by fault zones, thinning of beds, local disappearance of impedance contrasts or by the presence of noise. Under such circumstances the prior information about the subsurface structure in the zone of interest is very limited and it is not possible to put strong constraints on the solution to the inverse problem.

An analysis of the principles behind the presently available inverse methods reveals that these techniques all belong to a category called 'local optimization methods'. A characteristic property of these algorithms is that they systematically adjust the subsurface model in such a way that the misfit function (measuring the misfit between synthetic data and actual data) decreases monotically. This property would have been desirable if the misfit function possessed only one minimum.

However, since seismic data is of a highly oscillating nature, the misfit function generally has a very large number of minima (Fig. 1). Moreover, secondary minima representing low values of the misfit function often correspond to subsurface models that are quite different from the true model. It is therefore imperative that a local model optimization method uses a starting model that is 'connected' to the optimal solution by a path along which the misfit function decreases monotonically. In practice, the only way to ensure that this is the case is to use data from a nearby well and provide a starting model that is very close to the optimal model.

These considerations lead to the conclusion that local model optimization algorithms are likely to fail in the previously mentioned cases of limited well control. In such cases, the probability that a starting model is sufficiently close to the optimal model is small and the corresponding probability that a local optimization method will be attracted by an irrelevant minimum for the misfit function is high.

The solution to the above-described problems is to employ a 'global' optimization method. Global optimization methods are capable of searching for the optimal subsurface model with only a small risk of being trapped by irrelevant minima for the misfit function. Global optimization methods are typically statistical techniques.

A prototype development and implementation of a global, full-scale, seismic model optimization program for inversion of seismic profiles is presented. This program is based on the global optimization method 'simulated annealing' and is aimed at inversion of selected parts of migrated, seismic profiles with the purpose of producing geological cross-sections showing the acoustic impedance and location of layer interfaces in the subsurface.

Classical simulated annealing is known to be a rather inefficient Monte Carlo technique, only applicable in cases where a very large number of iterations can be performed within the available computer resources. However, in the present implementation we employ a recent version of simulated annealing (Nulton and Salamon 1988; Andresen *et al.* 1988) in which we are able to extract important statistical information about the structure of the optimization problem during the computations. As a result, we have been able to speed up the algorithm significantly. The efficiency is improved by a factor of 7 to 100, and the fact that highly accurate results can be produced in a limited time has made the algorithm interesting from a practical point of view.

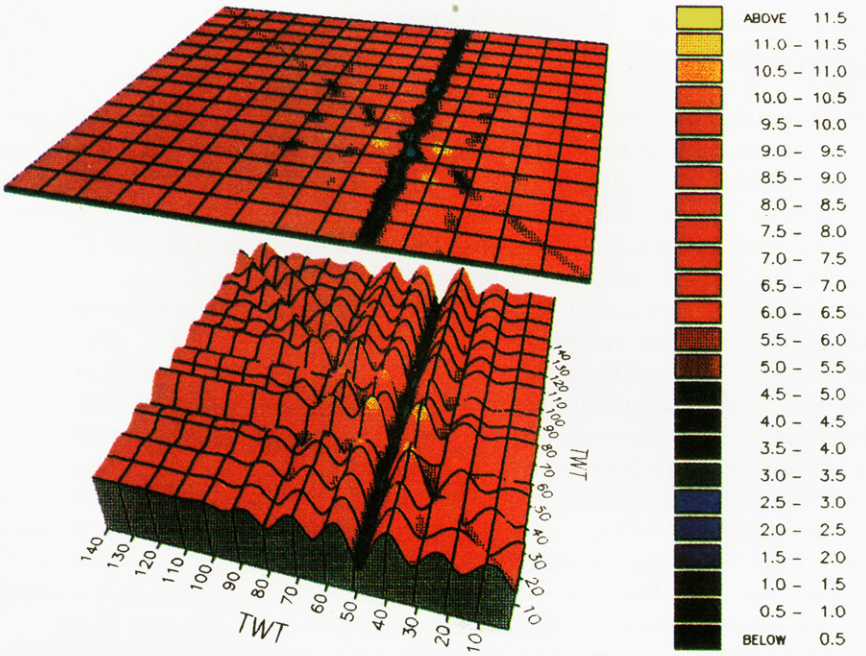


FIG. 1. Misfit function surface for a simple, single trace model optimization problem. The misfit surface is shown for a 2D cut through the parameter space. The independent parameters in the considered plane are two-way traveltimes of two reflectors.

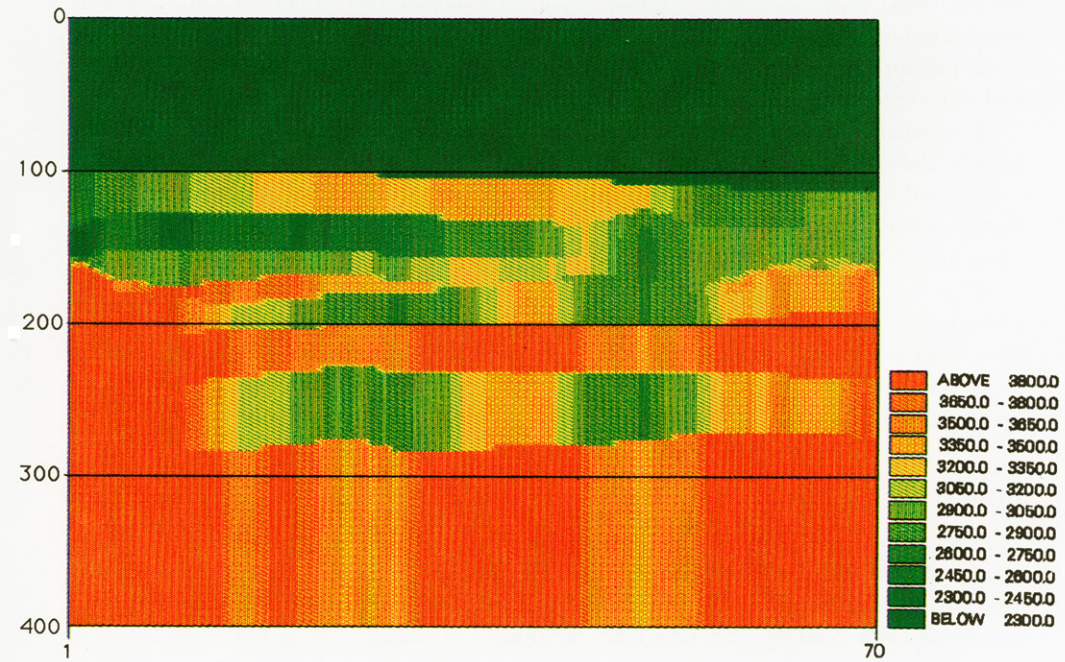


FIG. 2. Subsurface model used to generate the synthetic test data shown in Fig. 3.

Seismic Interpretation and Inversion

Interpretation

Interpretation of seismic data is an extremely complex process in which quantitative as well as qualitative information from several sources is compiled, weighted and combined into a geological model, which is displayed in such a way that it throws light on the most important aspects of the considered exploration problem.

The compilation of information includes all kinds of relevant data. One source of quantitative data is recently and previously recorded seismic data, and possibly also other kinds of geophysical data such as gravity and magnetic measurements.

Another source of information is qualitative, *a priori* geological models for the considered area, derived from the known, or partly known, geology of the considered area or from neighbouring areas, or from remote geological provinces that are expected to be geologically similar. Well data from the area or from adjacent areas can also be used. Quantitative well data include measurements of a number of physical parameters including mechanical ones. Qualitative well data includes geological descriptions of cuttings, small pieces of subsurface rock cut loose by the drill bit and brought up to the surface by the mud flow.

A third, not very often appreciated but extremely important, source of information in the interpretation problem is our knowledge of the theoretical connection between the seismic data and the mechanical properties of the subsurface. Without this information, an interpretation of the data would be impossible. Part of our theoretical knowledge of wave generation and propagation is applied during the conventional data processing. However, even successfully processed data are still dominated by at least one residual source or wave propagation effect: the wavelet. It is the interpreter's knowledge of the effect of this wavelet that initially determines his ability to resolve fine details in the subsurface.

The seismic wavelet gives rise to two important problems in data interpretation. Firstly, the oscillatory appearance of the wavelet makes traveltime determination of events ambiguous, and therefore serious reflector dislocations may occur in the produced model. Secondly, interference between events results in distortion of their apparent traveltimes and amplitudes.

The first dislocation effect can be removed in the vicinity of wells, where the reflectivity derived from the well data can be correlated directly with the seismic data, and a one-to-one correspondence between reflectors and reflections can be established. The second interference effect typically remains unsolved by the interpreter, due to the qualitative nature of the interpretation process.

Inversion

The above-mentioned interference effect can be solved by traditional, local optimization techniques such as steepest descent search, conjugate gradient search, etc. The dislocation problem, however, must be solved in advance. In case of sparse or absent well control, this can only be done by means of a global optimization technique.

In order to remove or avoid seismic reflector dislocations, it is natural to simulate a technique which is used for removal of crystallographic dislocations, namely chemical annealing. In this technique, the crystalline material is melted and subsequently cooled slowly through its melting point, allowing large, highly ordered crystals to grow. If annealing is performed sufficiently slowly, the crystalline material eventually consists of one crystal without dislocations. In this state the crystalline material has the lowest possible internal energy.

As pointed out by Mosegaard and Vestergaard (1991), the analogy between the seismic inversion process and chemical annealing can be reinforced in the following way: the subsurface models can be identified with the atomic configurations of the crystalline material. The misfit function used in the seismic model optimization as a measure of the difference between synthetic data, computed from a trial model, and the observed data, can be identified with the energy of the crystal. Furthermore, random changes in the sub-surface model during a stochastic search can be performed in a way that is analogous to the random movements of atoms in the melted material or in the crystal lattice. By this analogy, a gradual decrease in the average size of the 'thermal movements' of the model from large values down to zero is likely to result in a settling into a subsurface model possessing a low value of the misfit function. Such a model is exactly what we wish to find.

The technical details of this algorithm, which is known as 'simulated annealing' (Kirkpatrick, Gelatt and Vecchi 1983) are somewhat more involved than the above exposition suggests. The interested reader may consult the review by Mosegaard and Vestergaard (1991).

We have used a recently developed simulated annealing method (Nulton and Salamon 1988; Andresen *et al.* 1988). This method needs statistical information about the system to be optimized, in order to extract optimal annealing schedules. Reliable statistical information would be difficult to obtain from annealing with a single copy of the model-algorithm system, since the convergence property of the simulated annealing algorithm often results in sampling of a limited part of the model space, typically concentrated around the misfit minimum found by the algorithm. In order to reduce this problem, we run a number of copies of the annealing at the same time. These copies of the model-algorithm system share the same temperature schedule, but they use different random number sequences, and their initial states are distributed according the prior knowledge. Hence, their time evolution is different, and they sample widely different parts of the model space.

Calculation of the Misfit

In the present model optimization problem, the misfit function S is the error trace energy

$$S(\mathbf{r}, \tau) = \sum_{n=0}^{N} (s(\mathbf{r}, \tau, n) - d(n))^2, \tag{2}$$

where $d(n)$ is the nth data sample, $s(\mathbf{r}, \tau, n)$ is the nth sample of the modelled trace, $\mathbf{r}$ and τ are vectors of reflection coefficients and two-way traveltimes, respectively,

and $N + 1$ is the number of data samples. $s(\mathbf{r}, \tau, n)$ is obtained from the convolutional model of the seismic trace:

$$s(\mathbf{r}, \tau, n) = \sum_{k=1}^{K} r_k w_k(n - \tau_k), \tag{3}$$

where τ_k is the two-way traveltime of the kth reflector, w_k is the wavelet corresponding to the kth reflection, and K is the number of reflectors considered.

However, for the considered type of global model optimation problems, the time required to obtain a near-optimal solution by simulated annealing grows rapidly with the number of parameters to be determined. Therefore it is desirable to reduce the number of parameters to be optimized by means of simulated annealing.

In the present problem, it is observed that the assumed dependence of the modelled trace $s(\mathbf{r}, \tau, n)$ on the reflection coefficients r_k is linear. Hence, it is possible to restrict the simulated annealing optimization to the two-way traveltime parameters τ_k only, and perform a simple, linear optimization of the reflection coefficients as part of the misfit calculations. We therefore redefine the misfit function as:

$$E(\tau) = \min_{\forall r} S(\mathbf{r}, \tau). \tag{4}$$

In the following, we assume that the wavelets $w_k(n)$ are non-zero only in the interval $0 \leq n \leq N_w$. For transient wavelets, this situation can always be obtained by introducing an appropriate time shift. We also assume that none of the individual reflection events are clipped at the end of the modelled trace. In other words, $0 \leq \tau_k \leq N - N_w$ for all k.

The partial derivatives of S with respect to the reflection coefficients r_k are given by

$$\frac{\partial S}{\partial r_k} = \sum_{n=0}^{N} 2(s(\mathbf{r}, \tau, n) - d(n))w_k(n - \tau_k) = 2 \sum_{n=0}^{N} e(\mathbf{r}, \tau, n)w_k(n - \tau_k), \tag{5}$$

where $e(\mathbf{r}, \tau, n) = s(\mathbf{r}, \tau, n) - d(n)$ is the error trace. The optimal values of r_k must satisfy

$$\frac{\partial S}{\partial r_k} = 0 \tag{6}$$

for all k. This leads to the following system of linear equations:

$$\sum_{n=0}^{N} \left[\left[\sum_{i=1}^{K} (r_i w_i(n - \tau_i)) - d(n) \right] w_k(n - \tau_k) \right] = 0 \tag{7}$$

for $k = 1, \ldots, K$. This system of equations is equivalent to the system

$$\sum_{i=1}^{K} r_i \left[\sum_{n=0}^{N} w_i(n - \tau_i)w_k(n - \tau_k) \right] = \sum_{n=0}^{N} d(n)w_k(n - \tau_k) \tag{8}$$

or

$$\sum_{i=1}^{K} r_i \left[\sum_{n=0}^{N_w} w_i(n + (\tau_k - \tau_i))w_k(n) \right] = \sum_{n=0}^{N} d(n)w_k(n - \tau_k). \tag{9}$$

If we assume that the wavelet is the same, say $w(n)$, for all reflections, the equations reduce to

$$\sum_{i=1}^{K} r_i R_{ww}(\tau_k - \tau_i) = R_{dw}(-\tau_k), \tag{10}$$

where

$$R_{ww}(\tau) = \sum_{n=0}^{N_w} w(n)w(n+\tau) \tag{11}$$

is the autocorrelation of the wavelet, and

$$R_{dw}(\tau) = \sum_{n=0}^{N} d(n)w(n+\tau) \tag{12}$$

is the cross-correlation between the data and the wavelet. From (10), optimal reflection coefficients r_i can be found, if the two-way traveltimes τ_k are given. Equations of the form (10) are known in filter theory as the 'normal equations' and they can be solved very efficiently by the Wiener–Levinson algorithm. In each iteration, a new set of two-way traveltimes is selected to become candidate destinations for the next move in the two-way traveltime parameter space (having half the dimensions of the combined traveltime–reflection coefficient space). The misfit E is now calculated after having minimized S with respect to the reflection coefficients.

A necessary condition for the system of equations (10) to have a unique solution is that all the K two-way times τ_k are different (no reflectors coincide). This condition must be satisfied by each perturbation applied to the subsurface models. Moreover, reflectors are not allowed to be too close, since numerical instabillity will occur when the equations (10) are near singular. Geological situations such as layer pinchouts must therefore be treated separately.

Another problem to be mentioned is that a straight linear optimization of the reflection coefficients for given traveltimes may yield reflection coefficients that violate the constraints imposed by the prior information. This problem can be avoided by performing a constrained, linear optimization.

A Synthetic Example

The synthetic data test is based on the synthetic seismic response for a model of thin layers (Fig. 2). The model consists of a layered sequence between a low velocity half-space above and a high velocity half-space below. The acoustic impedance of the layers varies laterally, and two layers pinchout from left to right. The data set (Fig. 3) is generated by convolving the model reflectivity with a 40 ms long, bandpass filtered wavelet. The target zone is a 400 ms time window over 71 common depth points, which is a realistic size for many practical applications of target zone oriented inversion.

In this numerical example, the prior knowledge is sparse. The two-way traveltimes for the layer interfaces are limited to intervals that are slightly longer than one

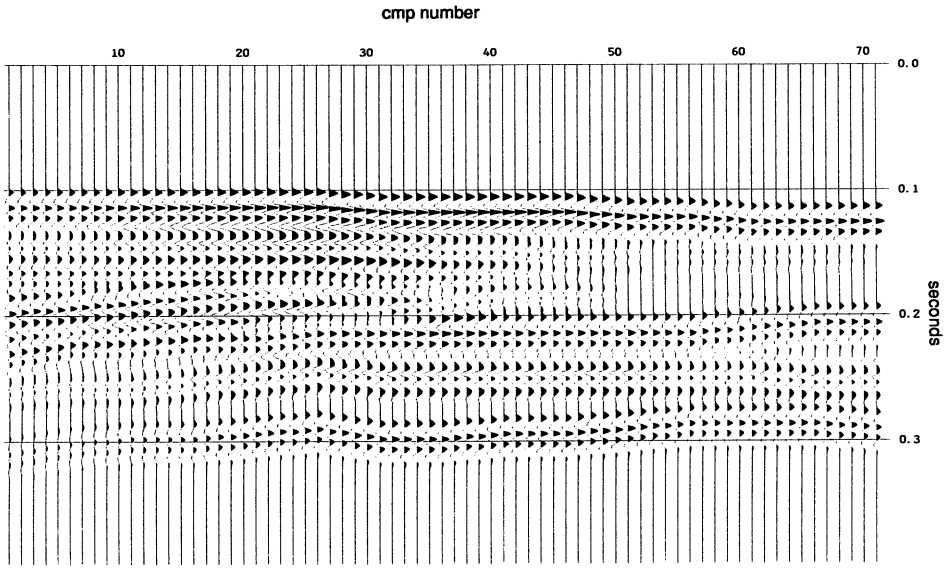


FIG. 3. Synthetic data set used in the test example.

wavelet-length, and the amplitudes of the reflections are constrained to be between -0.2 and 0.2. A weak, lateral smoothness constraint is applied to the reflectors during the optimization. Even when the wavelet is assumed to be known, these wide limits on the parameters impose a very large number of local minima on the misfit function for the model optimization problem.

Five independent copies of the model-algorithm system were allowed to perform 5000 iterations each. Two of these copies settled into a near-optimal solution. In order to illustrate how the solution to the model optimization problem was formed during the most successful of these annealing processes, a number of 'snapshots' of intermediate subsurface models are shown for increasing iteration numbers, corresponding to a temperature variation from a very high value (effectively infinity) down to zero. The main point to notice is how the models in the ensemble gradually change from being typical samples from the *a priori* distribution (model parameters uniformly distributed over the parameter intervals), to models that reflect the information contained in the seismic data, under the limitations imposed by the prior knowledge.

The first subsurface model in Fig. 4 is the result of a substantial number of iterations at a very high temperature. In this model, the seismic dislocations are very large, and the resulting model for the acoustic impedance is far from the optimal model. The illustrated model gives an impression of the weak model constraints used in this annealing run.

In the first part of the annealing, an initial ordering of the models takes place. Figure 5 shows a model after the first part of the annealing has taken place. It can be seen that a layered structure is growing in the upper part of the target zone, corresponding to the first ordering of a crystal structure in the physical analogy.

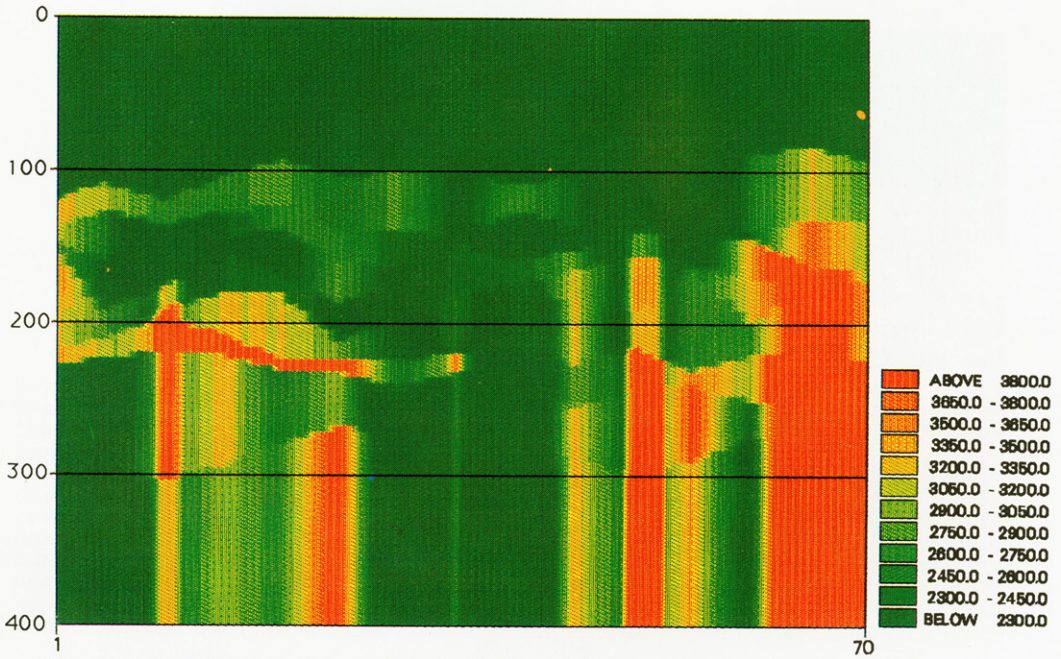


FIG. 4. A typical subsurface model obtained at infinite temperature. The plot shows the acoustic impedance as a function of two-way traveltimes.

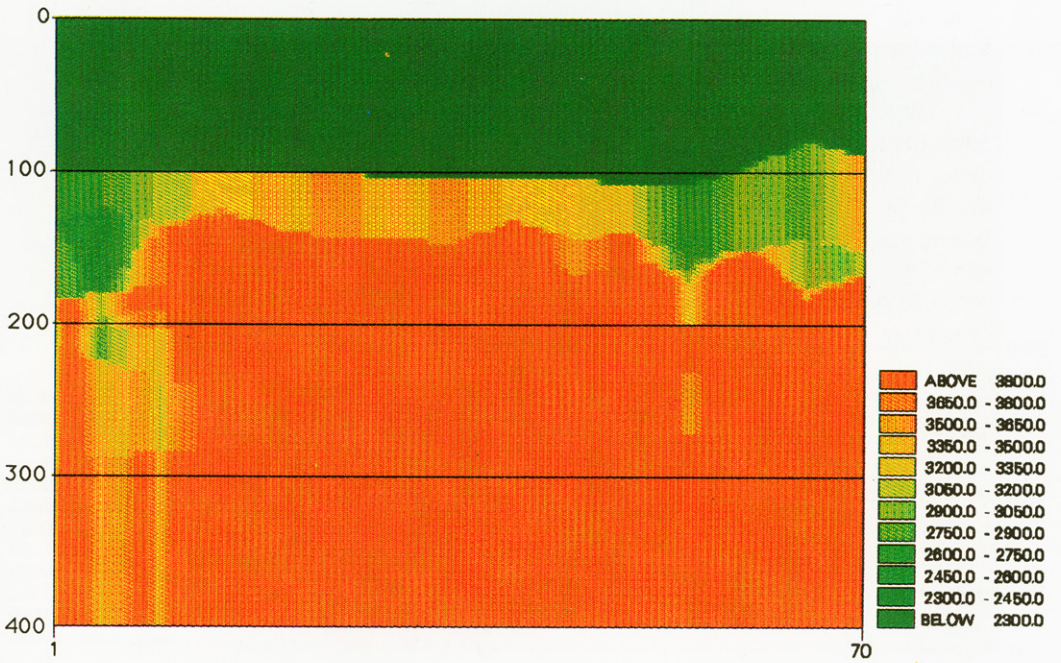


FIG. 5. A subsurface model from the ensemble, picked after the first part of the simulated annealing has taken place.

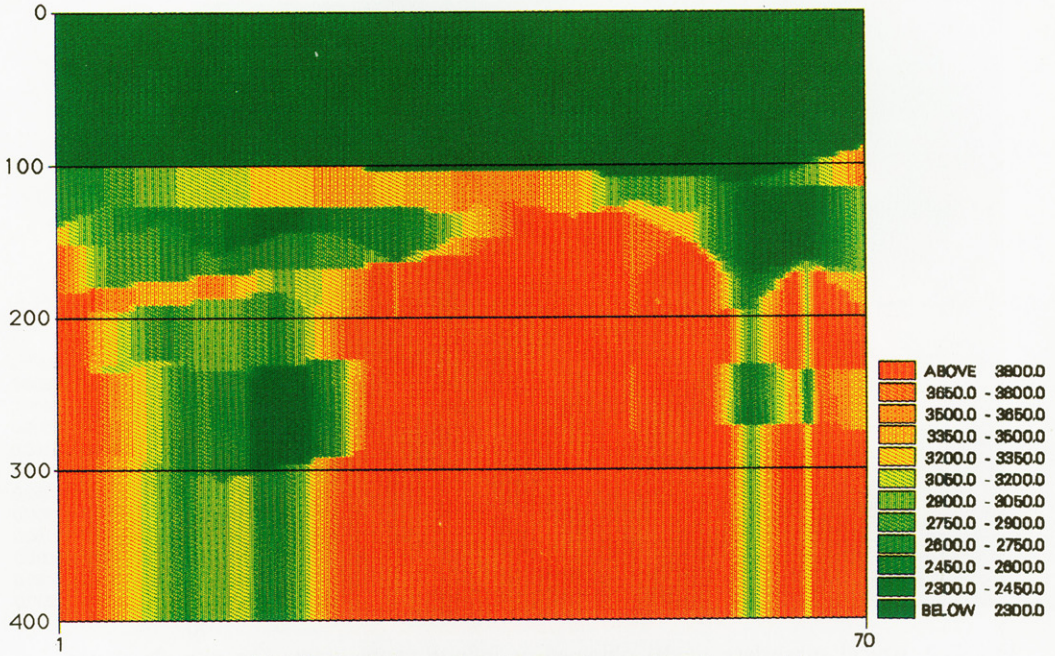


FIG. 6. An intermediate temperature subsurface model.

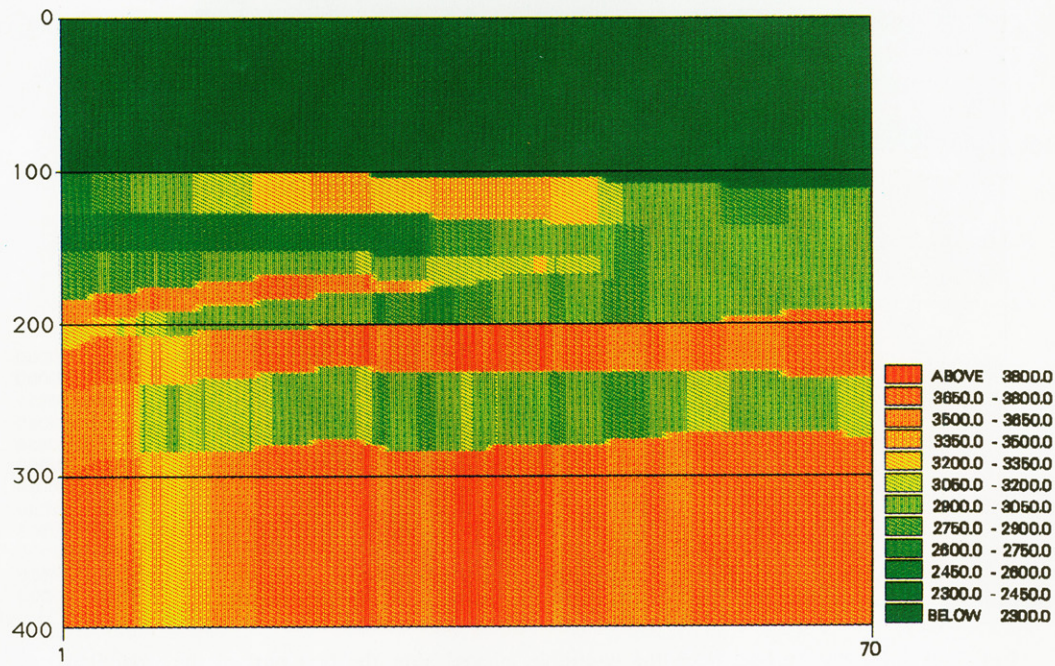


FIG. 7. The best subsurface model found close to zero temperature.

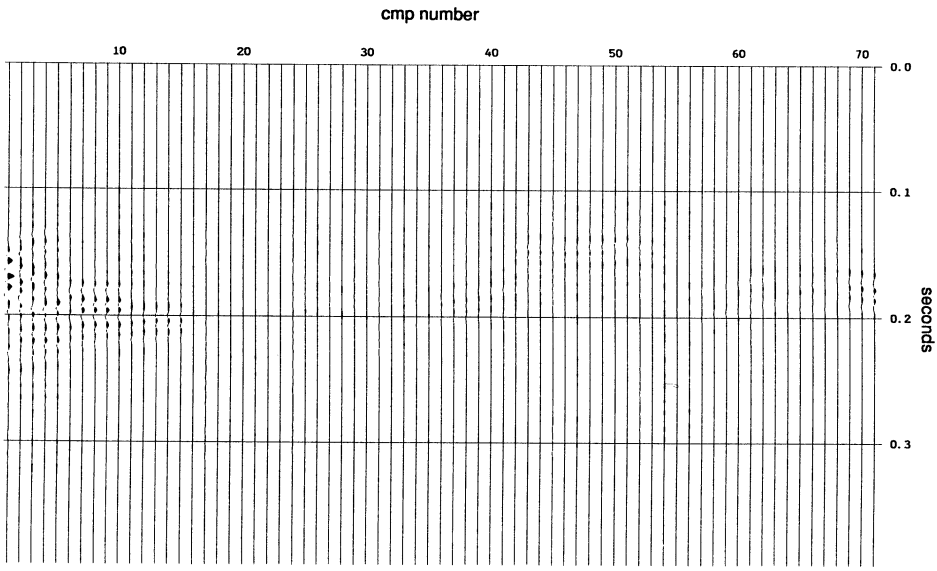


FIG. 8. Error traces for the model shown in Fig. 7.

After a further lowering of the temperature, models like that shown in Fig. 6 can be found in the ensemble. An increasing ordering of the model into a layered structure is seen. This illustrates how the influence of the data on the models increases as the annealing progresses. The major part of the target zone is not yet resolved at this intermediate temperature.

When the annealing temperature approaches zero, the models look like that shown in Fig. 7. These models are near optimal in the misfit sense, that is, the total error trace energy (Fig. 8) is small, compared to the total trace energy of the data. At a temperature close to zero, the influence of the data on the model is very strong, yielding a highly ordered subsurface model. The differences between the models in the ensemble reflect the limited resolution in the data, the non-uniqueness of the inverse problem, and possible imperfect convergence of the algorithm. It is seen that the actual target in this model, namely the pinch-out, located approx. 175 ms below the top of the target zone, developing from CDP 10 to CDP 40, is resolved at this point. However, due to end-effects, the error traces build up to the left of this zone.

If the annealing is terminated by a number of iterations at zero temperature (corresponding to a local optimization starting from the best model obtained from the annealing), the true model is reached, since the data used in this example is noise-free.

CONCLUSIONS

The process of seismic interpretation is made difficult by two main types of distortion, both caused by the seismic wavelet: the oscillatory appearance of the individual reflection events and interference between different events. The former effect

results in a serious ambiguity in event identification, that is, in establishing a one-to-one correspondence between geological layer interfaces and features observed in the seismic data. The latter effect is responsible for minor errors in the estimated two-way traveltimes, and errors in reflection strengths observed in the seismic data.

Post-stack, sparse spike inversion methods provide quantitative methods for determining two-way times and reflection coefficients from carefully processed seismic data. If the event identification problem can be solved by comparing the seismic data with synthetic seismograms, calculated from well data, the remaining problem of removing interference effects can be solved by means of traditional, local optimization methods. However, if sufficient well data are not available, the event identification problem can only be solved quantitatively by means of a global optimization technique.

Global optimization methods are typically stochastic, and at present the most efficient is simulated annealing. In the present work, a recently developed, improved version of simulated annealing has been shown to produce near-optimal solutions to a seismic model optimization problem of a realistic size and complexity. An ensemble, consisting of several copies of the model-algorithm system, sharing the same temperature schedule but using different random number sequences, is used to collect statistical information about the model optimization problem, and efficient annealing temperature schedules are produced.

In order to reduce the computational workload considerably, the optimization of the reflection coefficients, which turns out to be a simple linear optimization, can be done separately as part of the misfit calculations. This reduces the dimensionality of the parameter space, in which the stochastic optimization is performed, to half the original dimensions. Consequently, the resulting number of accessible model configurations for the stochastic search decreases drastically, and the average time taken by the algorithm before a near-optimal model is located, is greatly reduced.

Acknowledgements

The authors are indebted to Jacob Mørch Pedersen, Peter Salamon, Bjarne Andresen and the simulated annealing group at the University of Copenhagen, whose comments and suggestions were invaluable. This work was sponsored by The Danish Ministry of Energy under contract No. 1313/87-12. Klaus Mosegaard was partly supported by the Danish Natural Science Research Council.

References

Andresen, B., Hoffman, K.H., Mosegaard, K., Nulton, J.D., Pedersen, J.M. and Salamon, P. 1988. On lumped models for thermodynamic properties of simulated annealing problems. *Journal de Physique* **49**, 1485–1492.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science* **220**, 671–680.

Mosegaard, K. and Vestergaard, P.D. 1991. A simulated annealing approach to seismic model optimization with sparse prior information. *Geophysical Prospecting* **39**, 599–611.

Nulton, J.D. and Salamon, P. 1988. Statistical mechanics of combinatorial optimization. *Physical Review* A **37**, 1351–1356.

Geophysical Prospecting **41,** 565–578, 1993

RESIDUAL STATICS ESTIMATION: SCALING TEMPERATURE SCHEDULES USING SIMULATED ANNEALING[1]

E. NØRMARK[2] and K. MOSEGAARD[3]

ABSTRACT

NØRMARK, E. and MOSEGAARD, K. 1993. Residual statics estimation: scaling temperature schedules using simulated annealing. *Geophysical Prospecting* **41,** 565–578.

Linearized residual statics estimation will often fail when large static corrections are needed. Cycle skipping may easily occur and the consequence may be that the solution is trapped in a local maximum of the stack-power function. In order to find the global solution, Monte Carlo optimization in terms of simulated annealing has been applied in the stack-power maximization technique. However, a major problem when using simulated annealing is to determine a critical parameter known as the temperature.

An efficient solution to this difficulty was provided by Nulton and Salamon (1988) and Andresen *et al.* (1988), who used statistical information about the problem, acquired during the optimization itself, to compute near optimal annealing schedules.

Although theoretically solved, the problem of finding the Nulton–Salamon temperature schedule often referred to as the schedule at constant thermodynamic speed, may itself be computationally heavy. Many extra iterations are needed to establish the schedule.

For an important geophysical inverse problem, the residual statics problem of reflection seismology, we suggest a strategy to avoid the many extra iterations. Based on an analysis of a few residual statics problems we compute approximations to Nulton–Salamon schedules for almost arbitrary residual statics problems. The performance of the approximated schedules is evaluated on synthetic and real data.

INTRODUCTION

When making reflection seismic surveys on land, different thicknesses and velocities of the surface layers will induce different delays on the seismic recordings, which

[1] Received April 1992, revision accepted November 1992.

[2] Department of Earth Sciences, Geophysical Laboratory, University of Aarhus, Finlandsgade 8, 8200 Aarhus N, Denmark.

[3] Geophysical Institute, University of Copenhagen, Haraldsgade 6, 2200 Copenhagen N, Denmark.

may generate false structures and reduce the quality of the common midpoint (CMP) stack. To compensate for these timing errors, static timeshifts of the observations are made. If the *a priori* information in terms of the field statics is insufficient, residual statics are estimated by automatic static correction procedures.

Usually, residual statics estimation is approached by a technique described by Taner, Koehler and Alhilali (1974), sometimes referred to as the traveltime picking method. This procedure is based on estimating timeshifts for all the individual traces in the CMP gathers and later resolving the time lags into surface-consistent source and receiver statics by least-squares fitting. Ronen and Claerbout (1985) have suggested an alternative statics estimation technique maximizing the stack-power S(**x**), where $\mathbf{x} = (x_1, x_2, \ldots, x_M)$ is the vector of the M source and receiver static parameters. In this method surface-consistent statics are estimated directly by evaluating one static parameter at a time.

The change in stack-power $\Delta S(\delta x_j)$ for a perturbation δx_j of the jth static parameter is

$$\Delta S(\delta x_j) = 2 \sum_{c \in \mathbf{C}_j} \phi^{pf}_{c(\delta x_j)}, \tag{1}$$

where $\phi^{pf}_{c(\delta x_j)}$ is the cross-correlation between the trace f affected by the jth static parameter and the partial stack p of the cth CMP gather. The partial stack is the stack of all traces in the CMP gather except the trace being studied. $\mathbf{C}_j$ is the subset of CMP gathers affected by the static parameter x_j.

For further details about the traveltime picking method and the stack-power maximization technique see Nørmark (1993).

Normally, the solution is approached by local optimization in both residual statics estimation techniques. The permitted timeshifts are evaluated and the static displacement giving the best correlation between the actual trace and a constructed reference trace (normally the partial stack) is applied to the data. However, a major problem is that cycle skipping often occurs when large residual statics, compared to the dominating period in the data, are estimated. The consequence may easily be that the solution is trapped in a local maximum of the objective function. This presents a highly non-linear inverse problem, which requires a global optimization technique to solve it.

Global Optimization by Simulated Annealing

Residual statics estimation considered as a global optimization problem was first treated by Rothman (1985). The optimization problem was approached by a Monte Carlo optimization technique, rooted in statistical mechanics, called simulated annealing. Simulated annealing is a numerical technique that resembles chemical annealing, in the way crystals are grown from a melt. This process is characterized by the fact that if the melt is carefully cooled in a critical temperature interval, a regular crystal with a minimum of energy is formed, whereas if the temperature is lowered too quickly, glass may be the result. A similarity between a thermodynamic description of such processes and the task of combinatorial optimization of a non-

linear function of high dimension was discovered by Kirkpatrick, Gelatt and Vecchi (1983). Based on this analogy they introduced simulated annealing as a tool to locate near-optimum solutions to global optimization problems.

Minimization of a parameter-dependent objective function $E(\mathbf{x})$ by simulated annealing is accomplished by making random perturbations of the parameters (corresponding to the state space coordinates in the physical problem) according to the Metropolis algorithm. The Metropolis algorithm states that, in each iteration, a random perturbation of a parameter, causing a change in the objective function (the energy) given by $\Delta E(\delta x_j)$, is

(1) accepted if $\Delta E(\delta x_j) \leq 0$,
(2) rejected with the probability $P(\delta x_j) = \exp(-\Delta E(\delta x_j)/T)$ if $\Delta E(\delta x_j) > 0$

(Metropolis *et al.* 1953). The parameter T corresponds to the temperature in thermodynamics, and for convenience is also named so in this context. When T is infinitely large, pure random perturbations of the parameters are made, whereas at $T = 0$ only perturbations decreasing the energy are accepted. Simulated annealing is initiated at high temperature. By slowly lowering the temperature and perturbing the parameters according to the Metropolis algorithm, the global minimum is reached with a high probability. Experimental evidences show the simulated annealing is much more efficient in locating the global minimum than the crude Monte Carlo optimization see e.g. Jakobsen, Mosegaard and Pedersen (1987).

Traditionally, only minimization problems are considered in simulated annealing. To keep the same convention in the residual statics estimation problem, the negative stack-power is minimized, which is equivalent to maximizing the stack-power. Thus $E(\mathbf{x}) = -S(\mathbf{x})$.

In the stack-power maximization problem, Rothman (1986) applies a modified version of the Metropolis algorithm, known as the heat bath method, which is claimed to be more efficient for problems where the energy evaluations are computationally inexpensive. In this method, the random trials are chosen according to the marginal Gibbs–Boltzmann probability distribution

$$P(\delta x_j) = \frac{\exp(\Delta E(\delta x_j)/T)}{\sum_{h=1}^{N} \exp(\Delta E(\delta x_h)/T)}, \tag{2}$$

where N is the number of possible static corrections for each parameter. $\Delta E(\delta x_j)$ (or $-\Delta S(\delta x_j)$) is evaluated using (1). By repeating the random perturbations according to the transition probabilities above (at constant temperature) and taking all static parameters once in each iteration, Rothman (1986) shows that eventually this will lead to a Boltzmann distribution of the available states, just as a repetition of the Metropolis algorithm will do.

The temperature schedule chosen is crucial for the performance of the optimization algorithm, in the sense that lowering the temperature too slowly will be a waste of computer time and cooling the system too fast will most likely result in a solution trapped in a local minimum. The temperature schedule suggested by

Rothman (1986) takes the form

$$T = \begin{cases} a^k T_0, & a^k T_0 > T_{\min}, \\ T_{\min}, & \text{otherwise}, \end{cases}$$

where k is the iteration number and α is a constant (set at 0.99). Firstly, a few iterations (controlled by T_0) are made with exponential cooling and afterwards a constant temperature $T_{\min}$ is applied at which essentially all iterations are made. The establishment of the above schedule is based on experiments.

Nulton and Salamon (1988) and Andresen *et al.* (1988) described a method by which near-optimum annealing temperature schedules could be produced. Their idea was to extend the analogy between the Monte Carlo optimization algorithms and statistical mechanics. Nulton and Salamon (1988) defined the heat capacity and the relaxation time for a problem and used principles from finite-time thermodynamics to design annealing schedules with minimum 'entropy' production i.e. annealing schedules at constant thermodynamic speed. Andresen *et al.* (1988) provided a method for the numerical estimation of constant speed schedules. In their method, attempted energy transitions are monitored during annealing, and from this information they estimate heat capacity and relaxation time for the considered problem. (See Appendix A). Constant speed schedules can then be calculated.

The Nulton–Salamon method has proved its efficiency in many cases (see e.g. Mosegaard and Vestergaard (1991)), but there are two important, practical problems in using this method.

Firstly, the Andresen *et al.* (1988) procedure is rather difficult to implement and requires a great deal of experience to use it. Secondly, the amount of extra iterations needed to give useful information about the heat capacity and relaxation time of a problem, following Andresen *et al.* (1988), may be so large, that the expected gain in computational efficiency is significantly reduced.

In order to overcome these problems in large residual statics estimation, we shall in the following suggest a strategy in which we compute a Nulton–Salamon schedule for a single (or a few) representative residual statics problems and apply a simple scaling procedure to approximate Nulton–Salamon schedules for other residual statics problems.

The Objective Function

Let us first consider the trivial problem of transferring temperature schedules from one problem to another, when the optimization problems are defined in the same parameter space and when the energies are linearly related. Then for two problems (1 and 2) the energies are related by

$$E_1(\mathbf{x}) = \alpha E_2(\mathbf{x}) + \beta, \tag{3}$$

where α and β are constants. Considering the same transition of state in both problems we have

$$\Delta E_1 = \alpha \, \Delta E_2 .$$

If we apply simulated annealing in the two problems with temperature schedules $T_1(t)$ and $T_2(t)$ respectively, so that

$$T_1(t) = \alpha T_2(t),$$

it is obvious that the optimizations become identical, because the parameter perturbations follow the same probability distributions in the two cases (see (2)).

In order to study whether a linear transformation of the temperature schedules is approximately valid for a broader class of residual statics estimation problems, not necessarily defined in parameter spaces with the same dimension and not necessarily satisfying (3), the following considerations are made:

Let the CMP gather c consist of M traces identified by index i. The individual traces are assumed to carry the signals s_{cit} and the noise n_{cit}, and are displaced by varying timeshifts $k(ci)$. t indicates the sample number. The CMP stack g_{ct} can be expressed as

$$g_{ct} = g_{ct}^{\text{signal}} + g_{ct}^{\text{noise}},$$

where

$$g_{ct}^{\text{signal}} = \sum_{i=1}^{M} \delta_{tk(ci)} * S_{\text{cit}} \qquad \text{and} \qquad g_{ct}^{\text{noise}} = \sum_{i=1}^{M} \delta_{tk(ci)} * n_{\text{cit}}.$$

$\delta_{tk(ci)}$ is Kronecker's delta given by

$$\delta_{tk(ci)} = \begin{cases} 1, & t = k(ci), \\ 0, & t \neq k(ci), \end{cases}$$

describing the static displacements.

If it is assumed that the signals are uncorrelated with the noise, the power of the CMP stack $S_c = \sum_t (g_{ct})^2$ can be approximated by

$$S_c \approx S_c^{\text{signal}} + S_c^{\text{noise}}.$$

Let us subdivide the stack-power range into intervals. The ith interval contains stack-powers between S_i and $S_i + \delta S$, where δS is the interval width. During a run, we can now form a matrix $\mathbf{A} = \{a_{ij}\}$ where a_{ij} is the number of transitions that could have taken place from the ith stack-power level to the jth stack-power level. In every iteration, we start at a certain stack-power level i and are allowed to perform a transition to any state that can be reached by only changing one static parameter. In such an iteration, all the reachable statics contribute to $\mathbf{A}$. The ith row in the matrix $\mathbf{A}$ will, after normalization, contain the distribution of potential stack-power transitions starting in stack-power level i. Let $d(S_i)$ be the standard deviation of this distribution.

By assuming that the stack-power perturbations have a Gaussian distribution, we obtain

$$d^2 \approx d_{\text{signal}}^2 + d_{\text{noise}}^2,$$

where d_{signal} and d_{noise} are the standard deviations for the noise and the signals respectively.

As will be seen the magnitude of the potential stack-power perturbations is proportional to the average stack-power associated with the static parameters. That is

$$d(\langle S_c \rangle) = aC\langle S_c \rangle + b,$$

where $\langle S_c \rangle$ is the average stack-power per CMP, and C is the average number of members in $\mathbf{C}_j$, the subset of the CMP gather over which the stack-power changes are evaluated (see (1)). a and b are constants.

In order to verify this relationship, and in order to estimate a and b, two optimization experiments were made, one on noise-free data and one on pure noise. 6-fold synthetic seismic data are used on which simulated annealing is applied. In Fig. 1 the standard deviations d of the potential stack-power perturbations are shown as a function of the stack-power for both experiments. Both d and S are normalized with the maximum stack-power being observed.

It can be seen that a linear relationship between stack-power and the standard deviation of the potential stack-power perturbations is a good approximation, although slight deviation from this relationship is observed at low stack-power, especially for the data containing pure noise. In both cases the intersections of the linear fits with the horizontal axes are approximately equal to the minimum stack-power being observed. Thus

$$d_{\text{signal}} \approx a_{\text{signal}} C(\langle S_c \rangle^{\text{signal}} - \langle S_c \rangle^{\text{signal}}_{\text{min}})$$

and

$$d_{\text{noise}} \approx a_{\text{noise}} C(\langle S_c \rangle^{\text{noise}} - \langle S_c \rangle^{\text{noise}}_{\text{min}}).$$

In the experiments it was found that $a_{\text{signal}} \approx a_{\text{noise}}$.

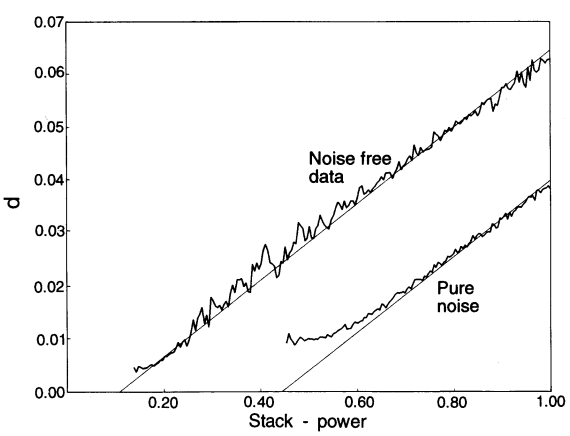


FIG. 1. Standard deviation of the potential stack-power perturbations d for experiments on noise-free data and on pure noise. The stack-power perturbations are mapped as a function of the stack-power itself. Both the stack-power and the standard deviations are normalized by the maximum stack-power. The linear fits are indicated by thin lines.

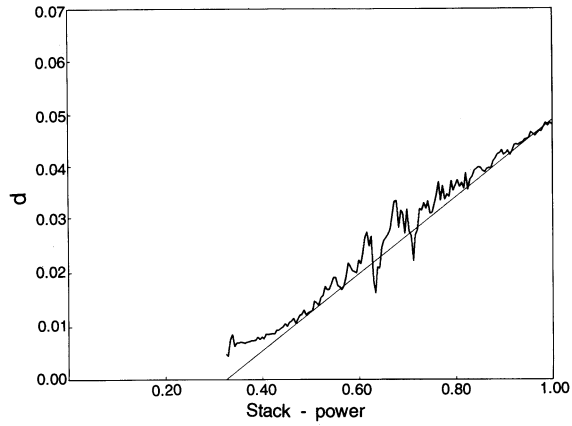


FIG. 2. The standard deviation d of the potential stack-power perturbations for data with a signal-to-noise ratio of 1 on the unstacked data. As in Fig. 1 the stack-power perturbations are mapped as function of the stack-power. The straight line indicates the predicted stack-power perturbation.

If it is assumed that signal and noise are uncorrelated, we have

$$S_c^{\text{signal}} \approx (1 + P^{-2})^{-1} S_c \quad \text{and} \quad S_c^{\text{noise}} \approx (1 + P^2)^{-1} S_c,$$

where $P = (S_c^{\text{signal}})^{1/2}/(S_c^{\text{noise}})^{1/2}$ is the signal-to-noise ratio of the CMP stack.

In the general case of noise-contaminated data, an estimate of d is given by

$$\begin{aligned} d^2 &\approx (a_{\text{signal}} C(\langle S_c \rangle^{\text{signal}} - \langle S_c \rangle_{\min}^{\text{signal}}))^2 + (a_{\text{noise}} C(\langle S_c \rangle^{\text{noise}} - \langle S_c \rangle_{\min}^{\text{noise}}))^2 \\ &\approx ((1 + P^{-2})^{-1} a_{\text{signal}} + (1 + P^2)^{-1} a_{\text{noise}})^2 (C(\langle S_c \rangle - \langle S_c \rangle_{\min}))^2. \end{aligned}$$

Since a is similar for both the signal and the noise experiments, P vanishes and d is approximated by

$$d = aC(\langle S_c \rangle - \langle S_c \rangle_{\min}). \tag{4}$$

In order to confirm (4), an experiment on noise-contaminated data is made, similar to those experiments made on noise-free data and on pure noise. The signal-to-noise ratio is 1 on the unstacked data within the cross-correlation window. In Fig. 2 the observed standard deviations of the potential stack-power perturbations are shown, together with the standard deviations predicted from (4). It can be seen that the estimated and the observed deviations are in good agreement with each other.

NORMALIZING THE TEMPERATURE SCHEDULES

From (2) it seems promising to obtain approximate Nulton–Salamon schedules by normalizing the temperatures by the magnitude of the expected energy perturbations, which we evaluated as the standard deviation of the potential stack-power

changes d. The aim of this section is to estimate the temperature schedules for a small, but hopefully reasonably representative, group of residual statics estimation problems, and normalize them to form schedules for other residual static problems.

According to (4), which has been verified experimentally, d varies approximately linearly with the stack-power. Let us consider two different static problems (1 and 2). During annealing, the Nulton–Salmon schedule $T_1(t)$ for problem 1 will increase the stack-power from $S_{1\,\min}$ to $S_{1\,\max}$ in a given number of iterations. The schedule $T_2(t)$ for problem 2 should increase the stack-power for that problem from $S_{2\,\min}$ to $S_{2\,\max}$ over the same number of iterations. Let us assume further that the stack-power functions (in the parameter space) for the two problems are different realizations of the same stochastic process, except for a linear transformation such as (3). In any iteration, we therefore require that the typical Gibbs–Boltzmann probabilities of (2) are approximately the same for both problems. This will be the case if the relationship

$$\frac{d_1}{T_1} = \frac{d_2}{T_2}$$

is satisfied. Here, d_1 and d_2 are the standard deviations of the potential stack-power perturbations for problems 1 and 2 respectively. It is readily seen that the above realization can be satisfied by the scaling

$$\frac{d_{1\,\max}}{T_1} = \frac{d_{2\,\max}}{T_2},$$

where $d_{1\,\max}$ and $d_{2\,\max}$ are d_1 and d_2 at the maximum stack-powers $S_{1\,\max}$ and $S_{2\,\max}$ respectively.

Normally, $d_{\max}$ is inaccessible, but according to (4) one may normalize by the stack-power instead. We use $C(\langle S_c\rangle_{\max} - \langle S_c\rangle_{\min})$ as a scaling factor. a vanished as it was found to be almost problem independent. In most residual static problems $\langle S_c\rangle_{\max}$ and $\langle S_c\rangle_{\min}$ are unknown, but a reasonable guess is usually possible.

In order to estimate the validity of the scaling suggested above, a number of schedules are estimated on synthetic examples with different numbers of parameters, varying signal-to-noise ratios, different stack-folds and different maximum static corrections. A schedule based on a real seismic data set is also included. Seismic data with between 6- and 12-fold coverages are employed. All examples have a fairly high signal-to-noise ratio. The input data are normalized in such a way that the mean power of the unstacked traces is equal to all examples. In all experiments a number of static parameters are fixed to the assumed solution at the end of the profile before initiating the optimization. This makes it easier for the algorithm to start aligning the seismic events. In all optimization problems the static parameters are taken in random order. The resulting temperature schedules are shown in Fig. 3a. No normalization has been applied. When normalizing by $C(\langle S_c\rangle_{\max} - \langle S_c\rangle_{\min})$ as in Fig. 3b, it is seen that the schedules have become much closer. It will now be investigated whether cooling curves normalized in this way can be used for the construction of the master schedule, from which approximate Nulton–Salamon schedules for other static problems can be derived.

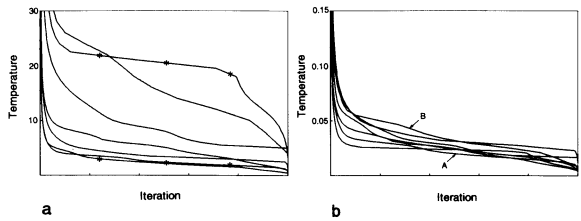


FIG. 3. (a) Temperature before normalization; (b) after normalization has been applied.

The data examples A and B, providing schedules with the lowest and highest normalized temperatures, are analysed. Example A is constructed from a synthetic seismic data set carrying a minimum phase signal and example B originates from real seismic data on which random statics are applied. In the latter case it was decided to use data where originally no significant statics problems existed. This makes an evaluation of the static solution easier. Figure 4a shows the initial states for the optimization and Fig. 4b shows the true solutions. In Fig. 4c, two examples of carrying out local optimizations are given. Almost all experiments with local optimization gave solutions trapped in local maxima of the stack-power function.

In order to demonstrate the possible consequence of using non-normalized temperature schedules estimated from another experiment, the schedules with highest and lowest temperatures (indicated by * in Fig. 3a) have been applied on data set A.

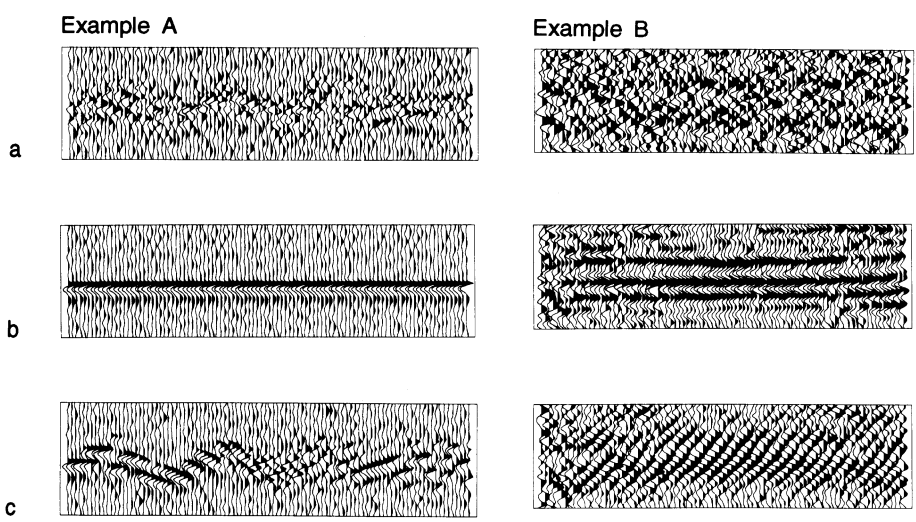


FIG. 4. Local optimization on two data sets. A is constructed from synthetic seismic data and B originates from a real seismic data set. (a) The initial state of the optimization; (b) the true solution; (c) examples of making local optimization by taking the static parameters in random order.

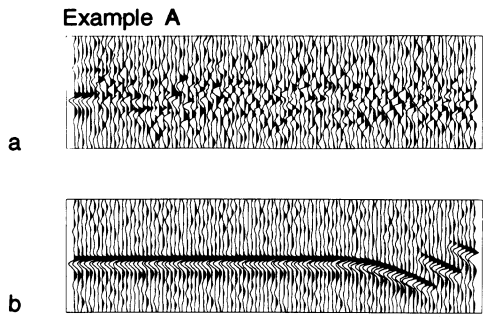


FIG. 5. The outcome of simulated annealing using (unnormalized) temperature schedules estimated from other experiments. (a) A schedule is used with temperatures that are generally too high. No order in the seismic traces has been achieved. (b) The temperatures are too low for the present data. The solution is trapped in a local maximum of the stack-power function.

In Fig. 5 the outcome of simulated annealing with 600 iterations is illustrated. (One iteration refers here to a perturbation of all static parameters). It is observed that when using a temperature schedule that is too high, no order in the seismograms has been detected (Fig. 5a). By using a schedule with temperatures that are too low, convergence is achieved after 530 iterations (Fig. 5b). Apparently, too rapid cooling has taken place through the most critical temperatures and the solution is trapped in a local optimum. This clearly demonstrates the need for normalizing the temperature schedules.

The normalized schedules in Fig. 3b still show a rather large scattering of the temperatures. In order to study the significance of these variations and thus the validity of a master temperature schedule based on an average of these cooling curves, the following experiments are made. First, simulated annealing on both data sets A and B is made using their own temperature schedules. 600 iteratons are made on 5 copies of the same data set. Representative solutions are shown in Fig. 6a. It is observed that the local maxima of the stack-power have now been avoided. By repeating these two experiments and changing their temperature schedules the importances of the differences between the schedules can be studied. The cooling curves are based on the normalized temperatures, which are scaled to match the actual data. The outcome of the optimizations is shown in Fig. 6b. Practically, the same solutions are obtained as when using their own temperature schedules. Only a minor effect of employing another (normalized) schedule than its own, is observed on the static corrections as well as on the stack-power function. For instance, on the synthetic data example, long periodic statics are still left by using a schedule estimated from another data set. However, one cannot expect to resolve long-periodic static by residual statics estimation (Wiggins, Larner and Wisecup 1976).

These experiments demonstrate that a temperature schedule for a given problem can successfully be estimated from another experiment, provided a proper normalization has been applied. They also show that slight variations on the schedules do not give rise to any significant variations in the appearance of the stacked data, at

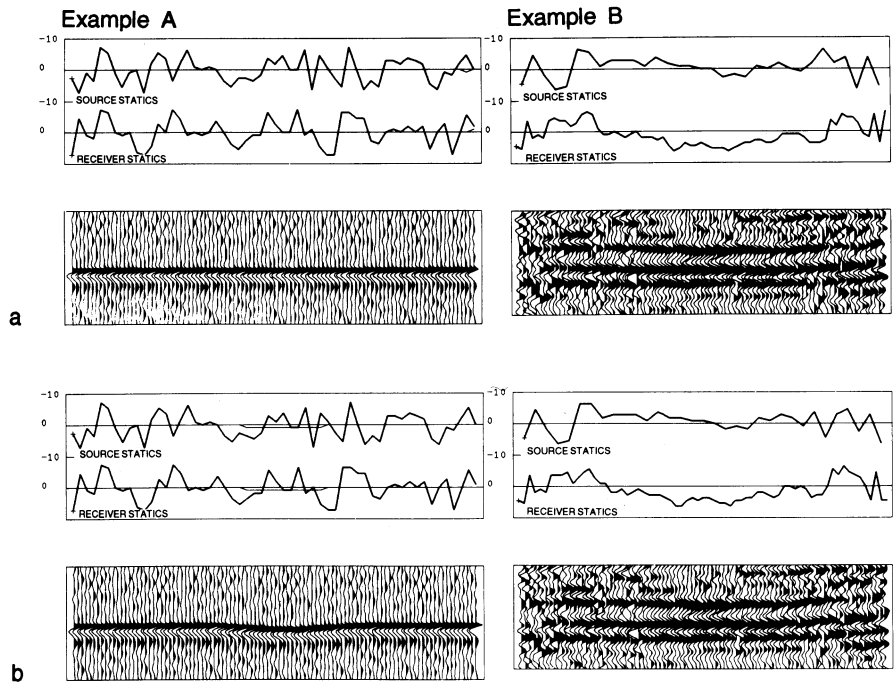


FIG. 6. (a) Simulated annealing using their own temperature schedules. (b) The result of changing their normalized temperature schedules. Before the schedules are used they are adapted to the actual data. The statics are given above the seismograms. The thin line in example A indicates the difference between the true and the estimated statics.

least not for the data examples studied in this context. Yet, the results of the optimization are not completely insensitive to such fluctuations. In order to construct a master schedule we suggest following an average schedule that is the mean value of the estimated master schedule.

DISCUSSION

Since the change in the energy for a static parameter x_j is only influenced by the CMP gathers in $\boldsymbol{C}_j$, we can expect that the temperatures are independent of the size of the seismic profile. However, when treating larger profiles local order may start to occur in different parts of the section, without global consistency. The consequence may be that more iterations are needed to find the global solution and that the temperature schedule must be modified accordingly. Yet, in this context only minor profiles, extending over a few spread lengths, have been treated. End effects causing decreasing coverage in the end of the profile and the resulting difficulties in estimating static parameters at such places have also been ignored.

Generally, it was found that normalized temperature schedules are fairly insensitive to variations in the input data. Experiments confirm that changing the stack-

fold does not have a significant effect on the normalized cooling schedules. However, only data of low coverages, i.e. up to 14-fold, have been examined. Using data of higher coverages will in itself make it more difficult to find a solution, because order in the seismograms is harder to achieve. Experiments also showed that changing the length of the cross-correlation window does not have any effect on the schedules. Modifying the maximum allowable static correction will also have no significant effect on the temperature schedules.

Our analysis of the temperature scaling problem shows that the temperatures generally decrease with increasing noise level due to the fact that the difference between $\langle S_c \rangle_{\max}$ and $\langle S_c \rangle_{\min}$ is small for noisy data. If statics estimation is made on data with a very high noise level, it may be hard to find any correlation at all between the seismic traces. In such cases it may be fruitful to scale the temperature schedules as if the data were pure noise.

Experimentally, it was discovered that when the source spacing deviates significantly from the receiver spacing the overall character of the temperature schedules may change. The numbers of traces in the source and the receiver gathers are then different, and consequently different magnitudes of the energy perturbations of the source and the receiver parameters can be expected, which may have a significant influence on the temperature schedule.

Conclusion

By studying the objective function of the residual statics estimation problem, it has been shown that the magnitude of the energy perturbations can be estimated and used to normalize the temperature schedules. Generally, it was found that normalized schedules are fairly consistent, and can, for a larger class of residual statics estimation problems, form a master temperature schedule from which approximate Nulton–Salamon schedules can be found. Our master schedules are based on a few static problems. A larger variety of problems should be included in order to determine a practically applicable master temperature schedule.

The present technique of calculating temperature schedules on the representative group of problems, studying the characteristics of the objective function and normalizing the temperatures in order to estimate a master schedule, could be used as the model for other non-linear optimization problems solved by simulated annealing.

Appendix A
Annealing with Constant Thermodynamic Speed

The philosophy of the present technique is to determine the temperature schedule keeping the same 'distance' to equilibrium mean energy $\langle E(T) \rangle_{\text{eq}}$ during the optimization. The 'distance' called the thermodynamic distance v, is defined as

$$v = \frac{\langle E(T) \rangle - \langle E(T) \rangle_{\text{eq}}}{\sigma(E_{\text{eq}}(T))},$$

where $\langle E(T) \rangle$ is the (non-equilibrium) mean energy of the states at temperature T and $\sigma(E(T)_{eq})$ is the standard deviation of energy fluctuations. (Nulton and Salamon 1988).

By adopting two other concepts from thermodynamics, the heat capacity $C(T)$ and the relaxation time $\varepsilon(T)$, and keeping v constant, the temperature schedule can be estimated as

$$\frac{\delta T}{\delta t} = -\frac{vT}{\varepsilon(T)\sqrt{C(T)}}, \tag{A1}$$

where t is the time measured in units of iterations. This first-order differential equation defines the temperature schedule for what is called annealing with constant thermodynamic speed. (See also Mosegaard and Vestergaard (1991)).

$C(T)$ and $\varepsilon(T)$ are determined by a method based on statistics of the energy transitions, (Andresen *et al.* 1988). The states of the system are lumped by the energy and the number of attempted moves from one energy interval i to another energy interval j is recorded in a matrix $\mathbf{Q} = \{Q_{ij}\}$, which is normalized to form a transition matrix. Using attempted moves means that both accepted perturbations and perturbations being rejected, according to the Metropolis algorithm, are employed. By transforming $\mathbf{Q}$ to the temperature-dependent transition probability probability matrix $\mathbf{Q}(T)$ and calculating its largest and second largest eigenvalues and the corresponding eigenvectors, an estimate of the density of states (from which the heat capacity can be defined) and the relaxation time can be obtained. See e.g. Mosegaard and Vestergaard (1991).

References

Andresen, B., Hoffmann, K.H., Mosegaard, K., Nulton, J., Pedersen, J.M. and Salmon, P. 1988. On lumped models for thermodynamical properties of simulated annealing problems. *Journal Physique (France)* **49**, 1485–1492.

Jakobsen, M.O., Mosegaard, K. and Pedersen, J.M. 1987. Model optimization in exploration geophysics 2: Global model optimization in reflection seismology simulated annealing. In: *Proceedings of the 5th International Mathematical Geophysics Seminar, Berlin*, 361–381. Vieweg & Sohn, Braunschweig.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science* **220**, 671–680.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal Chemical Physics* **21**, 1087–1092.

Mosegaard, K. and Vestergaard, P.D. 1991. A simulated annealing approach to seismic model optimization with sparse prior information. *Geophysical Prospecting* **39**, 599–611.

Nørmark, E. 1993. Residual statics estimation by stack-power maximization in the frequency domain. *Geophysical Prospecting* **41**, 551–563.

Nulton, J. and Salamon, P. 1988. Statistical mechanics of combinatorial optimization. *Physical Review* **A37**, 1351–1356.

Ronen, J. and Claerbout, J.F. 1985. Surface-consistent residual statics estimation by stack-power maximization. *Geophysics* **50**, 2759–2767.

ROTHMAN, D.H. 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics* **50**, 2784–2796.

ROTHMAN, D.H. 1986. Automatic estimation of large statics corrections. *Geophysics* **51**, 332–346.

TANER, M.T., KOEHLER, F. and ALHILALI, K.A. 1974. Estimation and correction of near-surface time anomalies. *Geophysics* **39**, 441–463.

WIGGINS, R.A., LARNER, K.L. and WISECUP, R.D. 1976. Residual statics analysis as a general linear inverse problem. *Geophysics* **42**, 922–938.

Monte Carlo sampling of solutions to inverse problems

Klaus Mosegaard
Niels Bohr Institute for Astronomy, Physics and Geophysics, Copenhagen
Albert Tarantola
Institut de Physique du Globe, Paris

This is a typeset LaTeX version of the paper originally published in
Journal of Geophysical Research, Vol. 100, No., B7, p 12,431–12,447, 1995.

Abstract

Probabilistic formulation of inverse problems leads to the definition of a probability distribution in the model space. This probability distribution combines a priori information with new information obtained by measuring some observable parameters (data). As, in the general case, the theory linking data with model parameters is nonlinear, the a posteriori probability in the model space may not be easy to describe (it may be multimodal, some moments may not be defined, etc.). When analyzing an inverse problem, obtaining a maximum likelihood model is usually not sufficient, as we normally also wish to have information on the resolution power of the data. In the general case we may have a large number of model parameters, and an inspection of the marginal probability densities of interest may be impractical, or even useless. But it is possible to pseudorandomly generate a large collection of models according to the posterior probability distribution and to analyze and display the models in such a way that information on the relative likelihoods of model properties is conveyed to the spectator. This can be accomplished by means of an efficient Monte Carlo method, even in cases where no explicit formula for the a priori distribution is available. The most well known importance sampling method, the Metropolis algorithm, can be generalized, and this gives a method that allows analysis of (possibly highly nonlinear) inverse problems with complex a priori information and data with an arbitrary noise distribution.

Introduction

Inverse problem theory is the mathematical theory describing how information about a parameterized physical system can be derived from observational data, theoretical relationships between model parameters and data, and prior information. Inverse problem theory is largely developed in geophysics, where the inquiry is how to infer information about the Earth's interior from physical measurements at the surface. Examples are estimation of subsurface rock density, magnetization, and conductivity from surface measurements of gravity or electromagnetic fields. An important class of complex inverse problems is found in seismology, where recorded seismic waves at the Earth's surface or in boreholes are used to compute estimates of mechanical subsurface parameters.

In what follows, any given set of values representing a physical system, we call a model. Every model $\mathbf{m}$ can be considered as a point in the model space $\mathcal{M}$. We will define different probability densities over $\mathcal{M}$. For instance, a probability density $\rho(\mathbf{m})$ will represent our a priori information on models, and another probability density, $\sigma(\mathbf{m})$ will represent our a posteriori information, deduced from $\rho(\mathbf{m})$ and from the degree of fit between data predicted from models and actually observed data. In fact, we will use the expression $\sigma(\mathbf{m}) = k\rho(\mathbf{m})L(\mathbf{m})$ [see *Tarantola*, 1987], where $L(\mathbf{m})$, the likelihood function, is a measure of the degree of fit between data predicted from the model $\mathbf{m}$ and the observed data (k is an appropriate normalization constant). Typically, this is done through the introduction of a misfit function $S(\mathbf{m})$, connected to $L(\mathbf{m})$ through an expression like $L(\mathbf{m}) = k\exp(-S(\mathbf{m}))$.

In seismology, the misfit function usually measures the degree of misfit between observed and computed seismograms as a function of the subsurface model parameters. It usually has many secondary minima. In terms of the probability density in the model space, we deal typically with a (possibly degenerate) global maximum, representing the most likely solution, and a large number of secondary maxima, representing other possible solutions. In such cases, a local search for the maximum likelihood solution using, for instance, a gradient method, is very likely to get trapped in secondary maxima. This problem is avoided when using a global search method. A global search is not confined to uphill (or downhill) moves in the model space and is therefore less influenced by the presence of local optima. Some global methods are not

influenced at all.

The simplest of the global search methods is the exhaustive search. A systematic exploration of the (discretized) model space is performed, and all models within the considered model subspace are visited. Although this method may be ideal for problems with low dimensionality (i.e., with few parameters), the task is computationally unfeasible when problems with many model parameters are considered.

When analyzing highly nonlinear inverse problems of high dimensionality, it is therefore necessary to severely restrict the number of misfit calculations, as compared to the exhaustive search. One way to do this is to use a Monte Carlo search, which consists of a (possibly guided) random walk in the model space. A Monte Carlo search extensively samples the model space, avoids entrapment in local likelihood maxima, and therefore provides a useful way to attack such highly nonlinear inverse problems.

In resolution studies, the advantages of Monte Carlo methods become even more significant. Resolution analysis carried out by means of local methods gives erroneous results due to the inherent assumption that only one minimum for the misfit function exists. However, a Monte Carlo method can take advantage of the fact that all local likelihood maxima will be sampled, provided a sufficient number of iterations are performed.

Early geophysical examples of solution of inverse problems by means of Monte Carlo methods, are given by *Keilis-Borok and Yanovskaya* [1967] and *Press* [1968, 1971]. Press made the first attempts at randomly exploring the space of possible Earth models consistent with seismological data. More recent examples are given by *Rothman* [1985, 1986], who nicely solved a strongly nonlinear optimization problem arising in seismic reflection surveys, and *Landa et al.* [1989], *Mosegaard and Vestergaard* [1991], *Koren et al.*, [1991], and *Cary and Chapman* [1988], who all used Monte Carlo methods within the difficult context of seismic waveform fitting. Cary and Chapman and Koren et al. described the potential of Monte Carlo methods, not only for solving a model optimization problem but also for performing an analysis of resolution in the inverse problem.

The idea behind the Monte Carlo method is old, but its actual application to the solution of scientific problems is closely connected to the advent of modern electronic computers. J. von Neumann, S. Ulam and E. Fermi used the method in nuclear reaction studies, and the name "the Monte Carlo method" (an allusion to the famous casino) was first used by *Metropolis and Ulam* [1949]. Four years later, Metropolis et al. [1953] introduced an algorithm, now known as the Metropolis algorithm, that was able to (asymptotically) sample a space according to a Gibbs-Boltzmann distribution. This algorithm was a biased random walk whose individual steps (iterations) were based on very simple probabilistic rules.

It is not difficult to design random walks that sample the posterior probability density $\sigma(\mathbf{m})$. However, in cases where $\sigma(\mathbf{m})$ has narrow maxima. these maxima (which are the most interesting features of $\sigma(\mathbf{m})$) will be very sparsely sampled (if sampled at all). In such cases, sampling of the model space can be improved by importance sampling, that is, by sampling the model space with a probability density as close to $\sigma(\mathbf{m})$ as possible. *Cary and Chapman* [1988] used the Monte Carlo method to determine $\sigma(\mathbf{m})$ for the refraction seismic waveform inversion problem, where the travel times were used as data, as well as waveforms, and the model parameters were the depths as a function of velocity. They improved the sampling of the model space by using a method described by *Wiggins* [1969, 1972] in which the model space was sampled according to the prior distribution $\rho(\mathbf{m})$. This approach is superior to a uniform sampling by crude Monte Carlo. However, the peaks of the prior distribution are typically much less pronounced than the peaks of the posterior distribution. Moreover, the peaks of the two distributions may not even coincide. It would therefore be preferable to draw sample models from the model space according to a probability distribution which is close to the posterior distribution $\sigma(\mathbf{m})$, the idea being to use a probability distribution that tends to $\sigma(\mathbf{m})$ as iterations proceed.

Geman and Geman [1984] discussed an application of simulated annealing to Bayesian image restoration. For their particular inverse problem, a two-dimensional deconvolution problem, they derived an expression for the posterior distribution from (1) the prior distribution, (2) a model of the convolutional two-dimensional image blurring mechanism, and (3) the parameters of the Gaussian noise model. By identifying this posterior distribution with a Gibbs-Boltzmann distribution, they performed a maximum a posteriori estimation in the model space, using a simulated annealing algorithm. In their paper, they mention the possibility of using the simulated annealing algorithm, not only for maximum a posteriori estimation but also to sample the model space according to the posterior distribution. However, they did not pursue this possibility further, nor did they describe how to extend this idea to inverse problems in general.

Marroquin et al. [1987] adopted an approach similar to that of Geman and Geman. However, they used the Metropolis algorithm to generate the posterior distribution, from which they computed model estimates. One of the problems raised by these authors was that their Bayesian approach requires an explicit formula for the a priori distribution.

Recent examples of using Bayes theorem and the Metropolis algorithm for generating a posteriori probabilities for an inverse problem are given by *Pedersen and Knudsen* [1990] and *Koren et al.* [1991].

In the present paper we will describe a method for random sampling of solutions to an inverse problem. The solutions are sampled at a rate proportional to their a posteriori probabilities, that is, models consistent with a priori information as well as observations are picked most often, whereas models that are in incompatible with either a priori information or observations (or both) are rarely sampled.

In brief our sampling algorithm can be described as consisting of two components. The first component generates a priori models, that is, models sampled with a frequency distribution equal to the a priori probability distribution in the model space. This is accomplished by means of a random walk, a kind of "Brownian motion" in the model space. The second component accepts or rejects attempted moves of the a priori random walk with probabilities that depend on the models ability to reproduce observations. Output from the combined algorithm consists of a collection of models that passed the test performed in the second component. This collection of models is shown to have a frequency distribution that is (asymptotically) proportional to the a posteriori probability distribution in the model space.

It is an important property of our method that in contrast to usual Bayesian inverse calculations, the a priori distribution need not be given by an explicit formula. In fact, the first component of our algorithm may consist of a large number of mutually dependent sub-processes, each of which generates part of the a priori models.

The definition of which models are accessible from a given model is an essential ingredient of the method, from a practical point of view. We will "jump" from a model to a neighboring model. But, what is a neighbor? The theory to be developed below is independent of the particular choice of model perturbations to be considered, but, as illustrated below, a bad definition of model neighborhood may lead to extremely inefficient algorithms.

Probabilistic Formulation of Inverse Problems

Parameters Taking Continuous Values

The "forward problem" is the problem of predicting (calculating) the "data values" $\mathbf{d}_{\rm cal} = \{d^1_{\rm cal}, d^2_{\rm cal}, \ldots\}$ that we should observe when making measurements on a certain system. Let the system be described (parameterized) by a parameter set $\mathbf{m} = \{m^1, m^2, \ldots\}$. One generally writes as

$$\mathbf{d}_{\rm cal} = g(\mathbf{m}) \tag{1}$$

the generally nonlinear, mapping from the model space $\mathcal{M}$ into the data space $\mathcal{D}$ that solves the forward problem.

In its crudest formulation, the "inverse problem" consists of the following question: An actual measurement of the data vector $\mathbf{d}$ gave the value $\mathbf{d}_{\rm obs} = \{d^1_{\rm obs}, d^2_{\rm obs}, \ldots\}$. Which is the actual value of the model parameter vector $\mathbf{m}$?

This problem may well be underdetermined, due to lack of significant data or due to experimental uncertainties. It can also be overdetermined, if we repeat similar measurements. Usually, it is both. A better question would have been: What information can we infer on the actual value of the model parameter vector $\mathbf{m}$?

The "Bayesian approach" to inverse problems, describes the "a priori information" we may have on the model vector, by a probability density $\rho(\mathbf{m})$. Then, it combines this information with the information provided by the measurement of the data vector and with the information provided by the physical theory, as described for instance by equation (2), in order to define a probability density $\sigma(\mathbf{m})$ representing the "a posteriori information". This a posteriori probability density describes all the information we have. It may well be multimodal, not have a mathematical expectation, have infinite variances, or some other pathologies, but it constitutes the complete solution to the inverse problem.

Whatever the particular approach to the problem may be [e.g., *Backus*, 1970a,b,c; *Tarantola and Valette*, 1982a; *Tarantola*, 1987], we end up with a solution of the form

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

where k is an appropriate normalization constant. The a posteriori probability density $\sigma(\mathbf{m})$ equals the a priori probability density $\rho(\mathbf{m})$ times a "likelihood function" $L(\mathbf{m})$ which, crudely speaking, measures the fit between observed data and data predicted from the model $\mathbf{m}$ (see an example below).

As an example, when we describe experimental results by a vector of observed values $\mathbf{d}_{\rm obs}$ with Gaussian experimental uncertainties described by a covariance matrix $\mathbf{C}$, then

$$L(\mathbf{m}) = k \exp\left[\frac{1}{2}\big(\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\rm obs}\big)^t \mathbf{C}^{-1}\big(\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\rm obs}\big)\right]. \tag{3}$$

If, instead, we describe experimental uncertainties using a Laplacian function, where $d^i_{\rm obs}$ are the "observed values" and σ^i are the estimated uncertainties, then

$$L(\mathbf{m}) = k \exp\left[-\sum_i \frac{|g^i(\mathbf{m}) - d^i_{\rm obs}|}{\sigma^i}\right]. \tag{4}$$

As a last example (to be used below), if the measured data values $d^i_{\rm obs}$ are contaminated by statistically independent, random errors ε_i given by a double Gaussian probability density function,

$$f(\varepsilon) = k\left[a \exp\left(-\frac{\varepsilon^2}{2\sigma_1^2}\right) + b \exp\left(-\frac{\varepsilon^2}{2\sigma_2^2}\right)\right], \tag{5}$$

then

$$L(\mathbf{m}) = k \prod_i \left[a \exp\left(-\frac{(g^i(\mathbf{m}) - d^i_{\text{obs}})^2}{2\sigma_1^2} \right) + b \exp\left(-\frac{(g^i(\mathbf{m}) - d^i_{\text{obs}})^2}{2\sigma_2^2} \right) \right] . \quad (6)$$

These three examples are very simplistic. While in this paper we show the way to introduce realistic a priori information in the model space, we do not attempt to advance in the difficult topic of realistically describing data uncertainties.

Discretization of Parameters

So far, the theory has been developed for parameters that although finite in number may take continuous values. Then, at any point $\mathbf{m}_i$ we can define a probability density $f(\mathbf{m}_i)$, but not a probability, which can only be defined for a region of the space:

$$P(\mathbf{m} \in \mathcal{A}) = \underbrace{\int dm^1 \int dm^2 \cdots}_{\mathcal{A}} f(\mathbf{m}). \quad (7)$$

Here, $m^1, m^2 \ldots$ denote the different components of the vector $\mathbf{m}$.

For numerical computations, we discretize the space by defining a grid of points, where each point represents a surrounding region $\Delta m^1 \Delta m^2 \ldots$, small enough for the probability densities under consideration to be almost constant inside it. Then, when we say "the probability of the point $\mathbf{m}_i$" we mean "the probability of the region $\Delta m^1 \Delta m^2 \ldots$ surrounding the point $\mathbf{m}_i$". In the limit of an infinitely dense grid and assuming a continuous $f(\mathbf{m})$, "the probability of the point $\mathbf{m}_i$ " tends to

$$f_i = f(\mathbf{m}_i) \Delta m^1 \Delta m^2 \ldots . \quad (8)$$

The discrete version of equation (2) is then

$$\sigma_i = \frac{\rho_i L(\mathbf{m}_i)}{\sum_j \rho_j L(\mathbf{m}_j)} , \quad (9)$$

where

$$\sigma_i = \sigma(\mathbf{m}_i) \Delta m^1 \, \Delta m^2 \ldots , \quad (10)$$

and

$$\rho_i = \rho(\mathbf{m}_i) \Delta m^1 \, \Delta m^2 \ldots . \quad (11)$$

For simplicity, we will rather write

$$\sigma_i = \frac{\rho_i \, L_i}{\sum_j \rho_j \, L_j} , \quad (12)$$

where we use the notation

$$L_i = L(\mathbf{m}_i) \quad (13)$$

(note that $\Delta m^1 \, \Delta m^2 \ldots$ does not enter into the definition of L_i).

Once the probability (12) has been defined, we could design a method to sample directly the posterior probability σ_i (and, in fact, the methods below could be used that way). But any efficient method will proceed by first sampling the prior probability ρ_i . It will then modify this sampling procedure in such a way that the probability σ_i is eventually sampled. This, after all, only corresponds to the Bayesian viewpoint on probabilities: one never creates a probability ex nihilo but rather modifies some prior into a posterior.

Monte Carlo Sampling of Probabilities

Essentially, the sampling problem can be stated as follows: given a set of points in a space, with a probability p_i attached to every point i , how can we define random rules to select points such that the probability of selecting point i is p_i ?

Terminology

Consider a random process that selects points in the model space. If the probability of selecting point i is p_i , then the points selected by the process are called "samples" of the probability distribution $\{p_i\}$. Depending on the random process, successive samples $i, j, k, \ldots$ may be dependent or independent, in the sense that the probability of sampling k may or may not depend on the fact that i and j have just been sampled.

An important class of efficient Monte Carlo (i.e., random) sampling methods is the random walks. The possible paths of a random walk define a graph in the model space (see Figure 1). All models in the discrete model space are nodes of the graph, and the edges of the graph define the possible steps of the random walk. The graph defines the "neighborhood" of a model as the set of all models directly connected to it. Sampling is then made by defining a random walk on the graph: one defines the probability P_{ij} for the random walker to go to point i if it currently is at the neighboring point j . P_{ij} is called the "transition probability". (As, at each step, the random walker must go somewhere, including the possibility of staying at the same point, P_{ij} satisfies $\sum_i P_{ij} = 1$.) For the sake of mathematical simplicity, we shall always assume that a graph connects any point with itself: staying at the point is considered as a "transition" (a"step"), and the current point, having been reselected, contributes with one more sample.

Consider a random walk, defined by the transition probabilities $\{P_{ij}\}$, and assume that the model where it is initiated is only known probabilistically: there is a probability q_i that the random walk is initiated at point i . Then, when the number of steps tends to infinity, the

probability that the random walker is at point i will converge to some other probability p_i [*Feller*, 1970]. We say that $\{p_i\}$ is an "equilibrium probability distribution" of

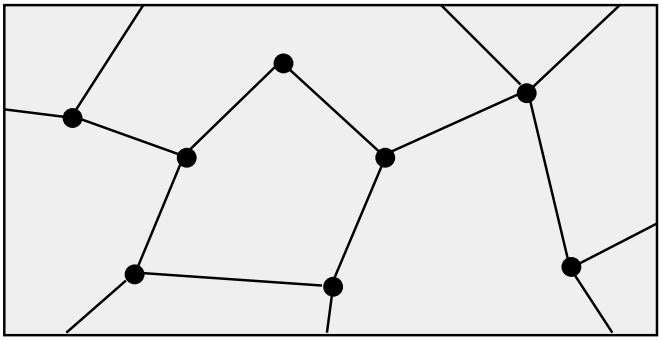


Figure 1: Part of a graph in the model space. The graph defines the possible steps of a random walk in the space. The random walk obeys some probabilistic rules that allow it to jump from one model to a connected model in each step. The random walker will, asymptotically, have some probability, say p_i , to be at point i at a given step. The neighborhood of a given model is defined as the models to which a random walker can go in one step, if it starts at the given model. Thus a neighborhood is defined solely through the graph and does not need to be a metric concept.

$\{P_{ij}\}$. (Then, $\{p_i\}$ is an eigenvector with eigenvalue 1 of $\{P_{ij}\}$: $\sum_j P_{ij} p_j = p_i$.) If the random walk always equilibrates at the same probability $\{p_i\}$, independent of the initial probability $\{q_i\}$, then there is only one equilibrium probability $\{p_i\}$. (Then, $\{p_i\}$ is a unique eigenvector of $\{P_{ij}\}$.) This is the case if the graph is "connected", that is, if it is possible to go from any point to any other point in the graph (in a sufficient number of steps) [*Feller*, 1970].

Many random walks can be defined that have a given probability distribution $\{p_i\}$ as their equilibrium probability. Some random walks converge more rapidly than others to their equilibrium probability. Successive models $i, j, k, \ldots$ obtained with a random walk will, of course, not be independent unless we only consider models separated by a sufficient number of steps. Instead of letting p_i represent the probability that a (single) random walker is at point i (in which case $\sum_i p_i = 1$), we can let p_i be the number of "particles" at point i . Then, $\sum_i p_i$ represents the total number of particles. None of the results presented below will depend on the way $\{p_i\}$ is normalized.

If, at some moment, the probability for the random walker to be at a point j is p_j and the transition probabilities are P_{ij} , then $f_{ij} = P_{ij} p_j$ represents the probability that the next transition will be from j to i while P_{ij} is the conditional probability of going to point i if the random walker is at j , f_{ij} is the unconditional probability that the next step will be a transition to i from j .

When p_i is interpreted as the number of particles at point i , f_{ij} is called the "flow", as it can be interpreted as the number of particles going to point i from point j in a single step. (The flow corresponding to an equilibrated random walk has the property that the number of particles p_i at point i is constant in time. Thus that a random walk has equilibrated at a distribution $\{p_i\}$ means that in each step, the total flow into a given point is equal to the total flow out from the point. Since each of the p_i particles at point i must move in each step (possibly to point i itself), the flow has the property that the total flow out from point i and hence the total flow into the point must equal p_i : $\sum_j f_{ij} = \sum_k f_{ki} = p_i$.) The concept of flow is important for designing rules that sample probabilities (see Appendix A).

Naïve Walks

Consider an arbitrary (connected) graph, as the one suggested in Figure 1, and denote by n_i the number of neighbors of point i (including the point i itself). Consider also a random walker that performs a "naïve random walk". That is, when he is at some point j , he moves to one of j's neighbors, say neighbor i , chosen uniformly at random (with equal probability). It is easy to prove (see Appendix B) that the random walk, so defined equilibrates at the probability distribution given by $p_i = n_i / \sum_j n_j$, i.e., with all points having a probability proportional to their number of neighbors.

Uniform Walks

Consider now a random walker that when he is at some point j , first chooses, uniformly at random, one of j's neighbors, say neighbor i , and then uses the following rule to decide if he moves to i or if he stays at j :

1. If $n_i \leq n_j$ (i.e., if the "new" point has less neighbors than the "old" point (or the same number), then always move to i .
2. If $n_i > n_j$ (i.e., if the "new" point has more neighbors than the "old" point), then make a random decision to move to i , or to stay at j , with the probability n_j / n_i of moving to i .

It is easy to prove (see Appendix B) that the random walk so defined equilibrates at the uniform probability, i.e., with all points having the same probability. This method of uniform sampling was first derived by *Wiggins* [1969].

The theory developed so far is valid for general, discrete (and finite) spaces, where the notion of metric is not necessarily introduced. In the special case of metric, Euclidean spaces, it is possible to choose Cartesian

coordinates, and to define the points in the space, where the random walk will be made, as a standard Cartesian grid of points. Let us, for instance, choose a graph as the one indicated in Figure 2. Then, away from the boundaries, the rule above degenerates into a (uniform) random choice of one of the $2N+1$ neighbors that any point has (including itself) in a space of dimension N . It can be shown (see Appendix B) that the walks so defined produce symmetric flows.

Modification of Random Walks

Assume that some random rules are given that define a random walk having $\{\rho_i\}$ as its equilibrium probability (uniform or not). How can the rules be modified so that the new random walk equilibrates at the probability.

$$\sigma_i = \frac{\rho_i L_i}{\sum_j \rho_j L_j} \ ? \tag{14}$$

Consider the following situation. Some random rules define a random walk that samples the prior probability $\{\rho_i\}$. At each step, the random walker is at point j , and

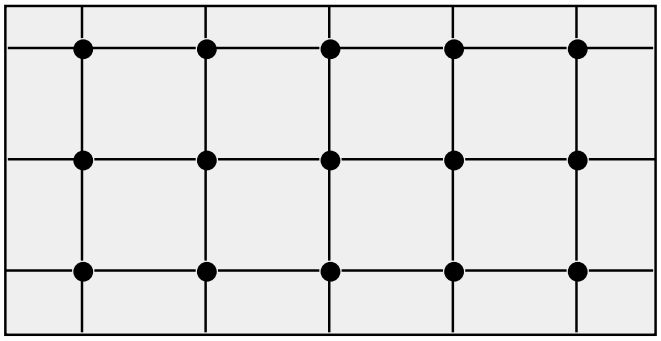


Figure 2: Part of a Cartesian graph in an Euclidean space. In this case, the definition of rules that sample points with the uniform probability is trivial.

an application of the rules would lead to a transition to point i . If that "proposed transition" $i \leftarrow j$ was always accepted, then the random walker would sample the prior probability $\{\rho_i\}$. Let us, however, instead of always accepting the proposed transition $i \leftarrow j$, sometimes thwart it by using the following rule to decide if he is allowed to move to i or if he is forced to stay at j :

1. If $L_i \geq L_j$ (i.e., if the "new" point has higher (or equal) likelihood than the "old" point), then accept the proposed transition to i .
2. If $L_i < L_j$ (i.e., if the "new" point has lower likelihood than the "old" point), then make a random decision to move to i , or to stay at j , with the probability L_i/L_j of moving to i .

Then it can be proved (see Appendix C) that the random walker will sample the posterior probability defined by equation (14). This modification rule, reminiscent of the Metropolis algorithm, is not the only one possible (see Appendix C).

To see that our algorithm degenerates into the Metropolis algorithm [*Metropolis et al.*, 1953] when used to sample the Gibbs-Boltzmann distribution, put $q_j = \exp(-E_j/T)/\sum_i \exp(-E_i/T)$, where E_j is an "energy" associated to the j-th point in the space and T is a "temperature". The summation in the denominator is over the entire space. In this way, our acceptance rule becomes the classical Metropolis rule: point i is always accepted if $E_i \leq E_j$, but if $E_i > E_j$, it is only accepted with probability $p_{ij}^{\text{acc}} = \exp(-(E_i - E_j)/T)$. Accordingly, we will refer to the above acceptance rule as the "Metropolis rule".

As an example, let us consider the case of independent, identically distributed Gaussian uncertainties. Then the likelihood function describing the experimental uncertainties (equation (3)) degenerates into

$$L(\mathbf{m}) = k \exp\left(-\frac{S(\mathbf{m})}{s^2}\right) \ , \tag{15}$$

where

$$S(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N} \left(g^i(\mathbf{m}) - d^i_{\text{obs}}\right)^2 \tag{16}$$

is the misfit function, $\mathbf{m}$ is a model vector, $\mathbf{d}$ is a data vector, $\mathbf{g}(\mathbf{m})$ is the forward modeling function, and s^2 is the total "noise" variance. In this example, s^2 is the same for all N data values. The acceptance probability for a perturbed model becomes in this case

$$P_{\text{accept}} = \begin{cases} 1 & \text{if } S(\mathbf{m}_{\text{new}}) \leq S(\mathbf{m}_{\text{old}}) \\ \exp(-\frac{\Delta S}{s^2}) & \text{if } S(\mathbf{m}_{\text{new}}) > S(\mathbf{m}_{\text{old}}) \end{cases} , \tag{17}$$

where

$$\Delta S = S\big(\mathbf{m}_{\text{new}}\big) - S\big(\mathbf{m}_{\text{old}}\big) \ . \tag{18}$$

This means that the perturbation is accepted if the perturbed model improves the data fit, and has a probability of being accepted of $P_{\text{accept}} = \exp(-\Delta S/s^2)$ if it degrades the data fit. From (17) we see that in the case of uniform a priori distribution, our algorithm becomes identical to the traditional Metropolis algorithm by identifying the misfit function S with the thermodynamic energy E and by identifying the noise variance s^2 with (k times) the thermodynamic temperature T .

Starting a Random Walk

We have just shown how a random walk sampling some prior probability $\{\rho_i\}$ can be modified by the Metropolis rule to sample the posterior probability $\{\sigma_i\}$. This

procedure is very suitable for solution of inverse problems. Usually, we will define some probabilistic rules that, when applied directly, would generate models $\mathbf{m}_1, \mathbf{m}_2, \ldots$ that, by definition, would be samples of the prior probability $\{\rho_i\}$. The application of the Metropolis rule defined above will modify this random walk in the model space so that it produces samples of the posterior probability $\{\sigma_i\}$ instead.

The fact that we have a random walk that samples the prior does not imply that we have an expression that allows us to calculate the value of the prior probability ρ_i of any model $\mathbf{m}_i$. The numerical example below gives an example of this. Of course, using the random walk that samples the prior and making the histograms of the models selected would be a numerical way of obtaining the value of the prior probability ρ_i for every model $\mathbf{m}_i$, but this is not a question that normally arises.

Using random rules that, if unmodified, generate samples of the prior and using the Metropolis rule to modify this random walk in order to sample the posterior corresponds to the Bayesian way of modifying a prior probability into a posterior. This approach will usually lead to efficient random walks, since the algorithm only explores the (usually) very limited subset of models that are consistent with our a priori information.

It often happens that we have data of different nature, as for instance in geophysics, when we have gravity, magnetic, or seismic data. Then, typically, data uncertainties are independent, and the total likelihood of a model, $L(\mathbf{m})$, can be expressed as a product of partial likelihoods: $L(\mathbf{m}) = L_1(\mathbf{m})L_2(\mathbf{m})\ldots$, one for each data type. Using the Metropolis rule directly to the total likelihood $L(\mathbf{m})$ would force us to solve the full forward problem (usually the most time-consuming part of the algorithm) to every model proposed by the prior random walk. Instead, we can use the Metropolis rule in cascade: If the random walk sampling the prior is modified first by considering the partial likelihood; $L_1(\mathbf{m})$, then we define a random walk that samples the product of the prior probability density $\rho(\mathbf{m})$ and $L_1(\mathbf{m})$. In turn, this random walk can be modified by considering the partial likelihood $L_2(\mathbf{m})$, and so on, until the posterior probability density that takes into account the total data set is sampled. Practically this means that, once a model is proposed by the rules sampling the prior, the forward problem is solved for the first data subset. The proposed model may then be accepted or rejected. If it is rejected by the Metropolis rule (typically when there is a large misfit between the synthetic data and the observed data for this first data subset), then there is no need to solve the forward problem for the other data subsets, and the rules sampling the prior have to propose a new model. More generally: Each time the Metropolis rule rejects a model at some stage of the algorithm, we go back to the lower level and propose a new model. When the solution of the forward modeling is inexpensive for certain data subsets, using this "cascade rule" may render the algorithm much more efficient than using the Metropolis rule to the total data set.

If, for some reason, we are not able to directly design a random walk that samples the prior, but we have an expression that gives the value of the prior probability ρ_i for any model $\mathbf{m}_i$ (an example is given by expression (19) below), we can, for instance, start a random walk that samples the model space with uniform probability (see the section on uniform walks). Using the Metropolis rules given above but replacing the likelihood values L_i by the prior probabilities ρ_i , we will obviously produce a random walk that samples the prior (the product of a constant times ρ_i equals ρ_i). Then, in cascade, we can use the Metropolis rule, with the likelihood values L_i , to modify this random walk into a random walk that samples the posterior probability $\sigma_i = \text{const}\,\rho_i L_i$.

A second option is to modify directly a uniform random walk (using the Metropolis rule above but with the product $\rho_i L_i$ instead of L_i) into a walk that directly samples the posterior, but this results, generally, in an inefficient random walk.

Multistep Iterations

An algorithm will converge to a unique equilibrium distribution if the graph that describes the move of a random walker in a single iteration is connected [*Feller*, 1970]. Often, it is convenient to split up an iteration in a number of steps, having its own graph and its own transition probabilities. A typical example is a random walk on a set of discrete points in an N-dimensional Euclidean space, as the one suggested in Figure 2. In this case the points are located in a regular grid having N mutually perpendicular axes, and one is typically interested in dividing an iteration of the random walk into N steps, where the nth move of the random walker is in a direction parallel to the nth axis.

The question is now: if we want to form an iteration consisting of a series of steps, can we give a sufficient condition to be satisfied by each step such that the complete iteration has the desired convergence properties? It is easy to see that if the individual steps in an iteration all have the same distribution $\{p_i\}$ as an equilibrium distribution (not necessarily unique), then the complete iteration also has $\{p_i\}$ as an equilibrium distribution. (The transition probability matrix for a complete iteration is equal to the product of the transition probability matrices for the individual steps. Since the vector of equilibrium probabilities is an eigenvector with eigenvalue 1 for each of the step transition probability matrices, it is also an eigenvector with eigenvalue 1, and hence the equilibrium distribution, for the transition probability matrix

for the complete iteration.) If this distribution is to be the unique equilibrium distribution for the complete iteration, then the graph of the complete iteration must be connected. That is, it must be possible to go from any point to any other point by performing iterations consisting of the specified steps.

If the steps of an iteration satisfy these sufficient conditions, there is also another way of defining an iteration with the desired, unique equilibrium distribution. Instead of performing an iteration as a series of steps, it is possible to define the iteration as consisting of one of the steps, chosen randomly (with any distribution having nonzero probabilities) among the possible steps (see Appendix D). Of course, a step of an iteration can, in the same way, be built from substeps and in this way acquire the same (not necessarily unique) equilibrium distribution as the substeps.

Sampling the a Priori Probability Density

We have previously assumed that we were able to sample the a priori probability density $\rho(\mathbf{m})$. Let us see how this can be achieved.

There are two ways of defining the a priori probability distribution:

1. By defining a (pseudo) random process (i.e., a set, of pseudo random rules) whose output is models assumed to represent pseudo random realizations of $\rho(\mathbf{m})$
2. By explicitly giving a formula for the a priori probability density $\rho(\mathbf{m})$.

Let us see an example of each.

First Example

From nearby wells we may have found that in a certain area of locally horizontal stratification, the distribution of layer thicknesses is approximately an exponential distribution, and the mass densities in the layers follow a log-normal distribution. Hence we can decide to generate one dimensional Earth models for mass density by the following random walk in the model space:

In each iteration:

1. Select a layer uniformly at random.
2. Choose a new value for the layer thickness according to the exponential distribution.
3. Choose a value for the mass density inside the layer, according to the log-normal distribution.

If we decide to discretize the model at constant Δz intervals, $\mathbf{m} = \{\rho(z_1), \rho(z_2), \ldots\}$ will have some probability distribution (representing our a priori knowledge) for the parameters $\{\rho(z_1), \rho(z_2), \ldots\}$ which we may not need to characterize explicitly.

In this example, the pseudo random procedure produces, by its very definition, samples $\mathbf{m}_1, \mathbf{m}_2, \ldots$ of the a priori probability density $\rho(\mathbf{m})$. These samples will be the input to the Metropolis decision rule. We recommend in particular this way of handling the a priori information, as it allows arbitrarily complex a priori information to enter the solution to an inverse problem. For an example of this procedure, see the section on numerical example.

Second Example

We may choose the probability density

$$\rho(\mathbf{m}) = k \exp\left(- \sum_{\alpha} \frac{|m^{\alpha} - m^{\alpha}_{\text{prior}}|}{\sigma^{\alpha}} \right) , \qquad (19)$$

where m^{α} represent components of the vector $\mathbf{m}$.

In this example, where we only have an expression for $\rho(\mathbf{m})$, we have to generate samples from this distribution. This can be done in many different ways. One way is to start with a naïve walk, as described above, and then use the Metropolis rule to modify it, in order to sample $\rho(\mathbf{m})$.

Sampling the a Posteriori Probability Density

In the previous section we described how to perform a random walk in the model space producing samples $\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \ldots$ of the a priori probability $\rho(\mathbf{m})$. In order to obtain samples of the a posteriori probability $\sigma(\mathbf{m}) = k\rho(\mathbf{m})L(\mathbf{m})$ we simply need to use the results given in the section on modification of random walks: if $\mathbf{m}_j$ is the "current point" and if the random walk sampling the prior would move from point $\mathbf{m}_j$ to point $\mathbf{m}_i$ (and whatever the used rules may be), accept the move if $L(\mathbf{m}_i) \geq L(\mathbf{m}_j)$, and decide randomly to accept or reject the move if $L(\mathbf{m}_i) < L(\mathbf{m}_j)$, with a probability $P = L(\mathbf{m}_i)/L(\mathbf{m}_j)$ of accepting the move.

Numerical Example

We now illustrate the theory developed in this paper with the inversion of gravity data. This is a classical example for testing any theory of inversion, and similar examples are given by *Dorman* [1975], *Parker* [1977] and *Jackson* [1979].

As the relationship between mass density and gravity data is strictly linear, one may wonder why we should illustrate a Monte Carlo method, with its inherent ability

to solve nonlinear problems, with the gravity inversion example. The reason is that our major concern is not the possibility of solving nonlinear problems, but the possibility of using, in standard geophysical inverse problems, realistic a priori information in the model space and realistic description of data uncertainties. This is what forces us to leave the comfortable realm of least squares and related methods and to develop the notions described here. It should be noted that the complex a priori knowledge used in this example renders the a posteriori distribution non-Gaussian.

The Problem

We consider a subsurface with a vertical fault, extending from the surface to infinite depth, as depicted in Figure 3. At the left of the fault the medium is homogeneous, while

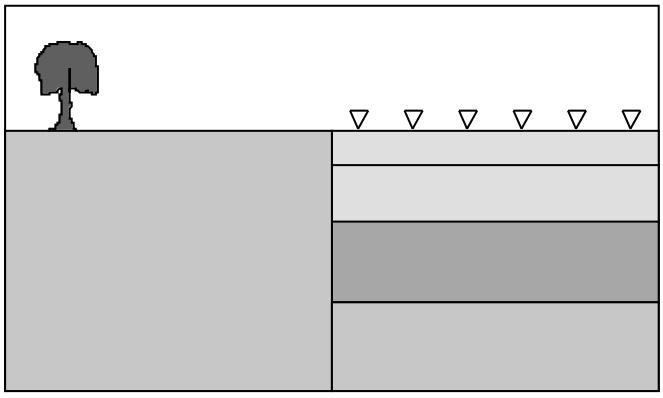


Figure 3: The geological model considered in our numerical example.

at the right of the fault the medium is depth dependent and characterized by a vertical profile of mass density $\rho(z)$.

The contrasts of mass density across the vertical fault produce a gravity anomaly at the surface. Let us assume that we have observed the horizontal gradient of the vertical component of the gravity at 20 equispaced points to the right of the fault, the first point being located 2 km from the fault, and the last point being located 40 km from the fault. The forward problem of computing the data values $d_i = d(x_i)$ from the density contrast function is solved by

$$d(x) = \frac{\partial g}{\partial x}(x) = 2G \int_0^\infty dz \frac{z \Delta\rho(z)}{z^2 + x^2} \ , \qquad (20)$$

where x is the horizontal distance from the fault, z is the depth, $g(x)$ is the vertical component of the gravity, $\Delta\rho(z)$ is the horizontal density contrast across the fault at depth z , and G is the gravitational constant.

The a Priori Information

Let us assume that in addition to the "hard" model constraints described above, we have the following a priori knowledge about the subsurface structure: The density of the rock to the left of the vertical fault is known to be 2570 kg/m^3. To the right of the fault is a stack of (half) layers, and we have the a priori information that the thicknesses ℓ_i of the layers are distributed according to the exponential probability density

$$f(\ell) = \frac{1}{\ell_0} \exp\left(-\frac{\ell}{\ell_0}\right) \ , \qquad (21)$$

where ℓ_0 , the mean layer thickness, has the value $\ell_0 = 4$ km.

Independently of the thickness of the layers, the mass density for each layer follows an empirical probability density, displayed in Figure 4. To simplify the calculation, the stack of layers is assumed to have a total thickness of 100 km, resting on a homogeneous basement having the same mass density as the half space at the left of the fault (2570 kg/m^3), and the top layer is truncated (eroded) at the surface.

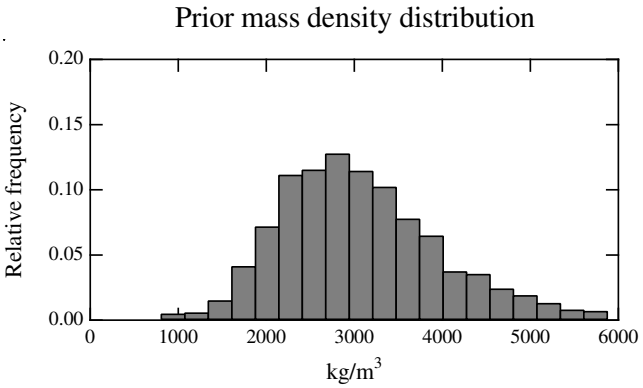


Figure 4: The a priori probability density function for the mass density inside each layer. The a priori probability density function for the thickness of each layer is an exponential function.

True Model, Experimental Uncertainties, and Observed Data Values

The measured data is assumed to be the response of a "true model" (Figure 5). The exact data corresponding to the true model are shown in Figure 6. The measured data values are assumed to be contaminated by statistically independent, random errors ε_i modeled by the sum of

two Gaussian probability density functions,

$$f(\varepsilon) = \frac{a}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{\varepsilon^2}{2\sigma_1^2}\right) + \frac{(1-a)}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{\varepsilon^2}{2\sigma_2^2}\right) , \quad (22)$$

where we have chosen the constants $\sigma_1 = 0.25\,10^{-9} s^{-2}$, $\sigma_2 = 1.25\,10^{-9} s^{-2}$, and $a = 0.25$ (see Figure 7).

The simulated observations, which are formed by summing the "true" data and the simulated noise, are displayed in Figure 6. Then, the likelihood function $L(\mathbf{m})$, measuring the degree of fit between synthetic and observed data is the one given by equation (6).

The Sampling Algorithm

The prior random walk. Let us now describe how our algorithm works. First, we define the graph in the model space that will guide our random walk. To ensure efficiency of the algorithm, it is important that very few of the possible steps in the model space lead to a radical change in the synthetic data generated from these models.

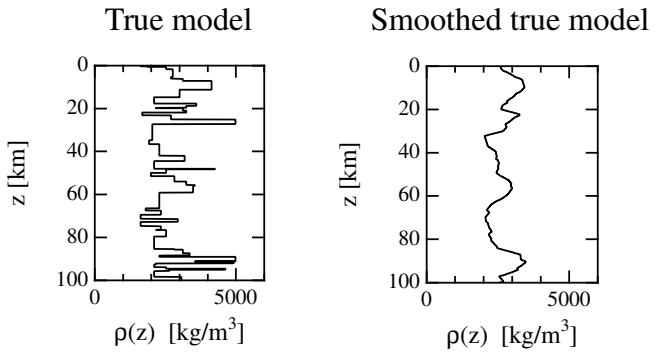


Figure 5: The true model used to generate synthetic data.

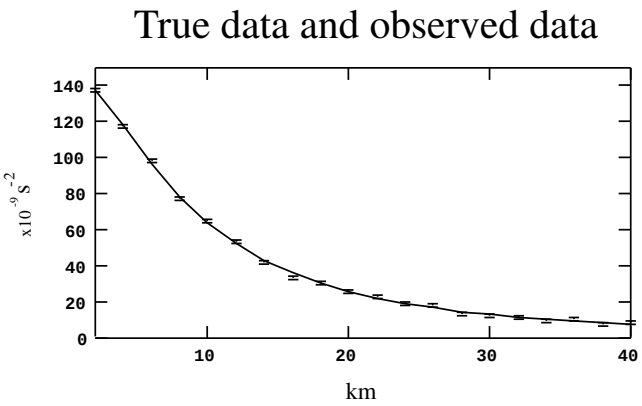


Figure 6: Synthetic data (solid line) used for the inversion, generated from the "true model" of Figure 5, and the "observed data" (points with error bars), equal to the "true data" plus some noise.

A simple way of sampling the a priori probability in the model space would be to use a random walk that generates successive models totally independently. To generate a new model, we could, for instance, pseudo- randomly generate layer thicknesses $\ell_1, \ell_2, \ldots$ from bottom to top, according to the exponential distribution given by equation (21), until they add up to the 100 km of total thickness ("eroding", if necessary, the top layer). Then we could pseudorandomly generate, inside each layer, the corresponding value for the mass density, according to the empirical distribution displayed in Figure 4. However this would produce a radical change in the synthetic data in each step of the random walk, and therefore it would be a very inefficient algorithm. The reason is that if the current model is one having a high posterior probability, a radical change would most likely lead to one of the very abundant models having a low posterior probability and would therefore be rejected by the algorithm.

Another way to produce samples of the a priori probability in the model space could be the following: Given a sample of the prior (i.e., given a model), we could produce another sample by, for instance, randomly choosing a layer and replacing its thickness by a new thickness

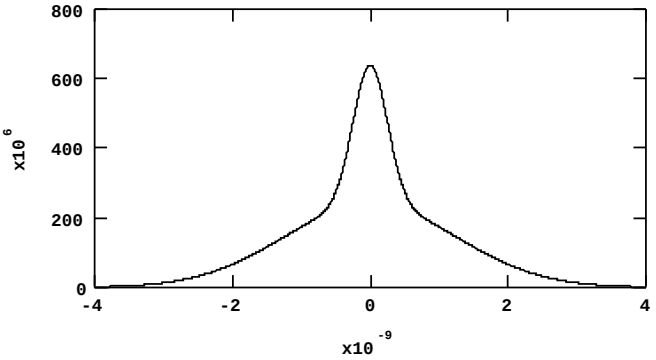


Figure 7: The arbitrary function used to model data uncertainties, as a sum of two Gaussians.

drawn from the exponential distribution given by equation (21) or by replacing its mass density by a new mass density drawn from the empirical distribution displayed in Figure 4.

It is obvious that iterating this procedure, we would always produce models whose layer thicknesses and mass densities are distributed properly; i.e., we would produce samples of the prior probability in the model space. Successive models will be "close" in some sense, but our numerical experimentation has shown that they are still too far apart: when testing models produced by this prior random walk by the likelihood function $L(\mathbf{m})$ (see be-

low), the probability of being accepted as samples of the a posteriori probability is extremely low. The reason is that when perturbing one layer thickness, all the layers above are shifted (remember that we go from bottom to top), and this strongly changes the synthetic data.

Therefore we decided to define the neighbors of a model as the models we can get, not by changing the thickness of a layer but by creating or destroying a new interface in the model (in a way described below). Then, all the other layers remain intact, and we only make a small perturbation in the synthetic data.

More precisely, the neighbors of a model are the models we can get by performing one of the following three perturbations:

(1) changing the mass density in one layer,

(2) adding a new layer boundary and assigning mass densities to the layers above and below it, or

(3) removing one layer boundary and assigning a mass density to the new compound layer.

To complete the description of our algorithm, we will now specify the random rules used by the random walk on the graph.

In each iteration it is first decided which kind of model perturbation step should be performed next. Performing a "pure" layer density perturbation has the same probability (0.5) as performing a layer boundary perturbation (removing or adding a boundary).

In case of a step involving a pure layer mass density perturbation, a layer is selected uniformly at random and a (new) density is chosen for that layer according to the density histogram of Figure 4.

In case of a layer boundary perturbation step we face the problem of adding or removing layer boundaries in such a way that if the step was iterated alone, it would leave the (a priori) distribution of models unchanged. In particular, the exponential layer thickness distribution $(1/\ell_0)\exp(-\ell/\ell_0)$ should be maintained. There is a simple solution to this problem: we exploit the fact that (approximately) exponentially distributed layer thicknesses can be obtained by assuming that the probability that a layer interface is present at a given depth (sample point) is equal to $(40\mathrm{m}/\ell_0) = 0.01$ and independent of the presence of other layer interfaces.

A layer boundary perturbation step therefore works as follows. First, we select one of the 2500 discrete points of the current mass density function, uniformly at random. We then randomly decide if there should exist a layer boundary at that point or not. The probability for the point to be a layer boundary is 0.01.

In case this operation creates a new layer boundary, we generate a mass density for the layers above and below the new layer boundary according to the a priori probability distribution shown in Figure 4.

In case this operation removes a layer boundary, we generate a mass density for the new compound layer (consisting of the layers above and below the removed layer boundary) according to the a priori probability distribution.

This exactly corresponds to the a priori information we wanted to input to our problem: the random walk in the model space so defined is sampling the probability density describing our a priori information.

The posterior random walk. Let us now describe how the above prior random walk is modified into a new random walk, sampling the posterior distribution.

Every time a model perturbation is attempted by the prior random walk, the gravity response is computed from the perturbed layer sequence $\mathbf{m}_{\,\mathrm{pert}}$ by summing up the contributions from the layers in the interval between 0 km depth and 100 km depth. The contribution from a homogeneous half layer is given by

$$G\Delta\rho\log\left(\frac{D^2+x^2}{d^2+x^2}\right) \tag{23}$$

where d is the depth to the top of the homogeneous half layer, D is the depth to the bottom of the half layer, $\Delta\rho$ is the layer density, and x is the horizontal distance to the edge of the half layer.

From the computed gravity response $\mathbf{g}(\mathbf{m}_{\,\mathrm{pert}})$ and the observed gravity response $\mathbf{d}_{\mathrm{obs}}$ the value of the likelihood function $L(\mathbf{m}_{\,\mathrm{pert}})$ is computed using equation (6). The attempted perturbation is now accepted or rejected according to the Metropolis rule, using the likelihoods $L(\mathbf{m}_{\mathrm{cur}})$ and $L(\mathbf{m}_{\,\mathrm{pert}})$ of the current and perturbed models, respectively (see the section on sampling the a posteriori probability density).

This completes the description of the algorithm used in our numerical example. There are, however, a few remaining issues concerning the use of its output models. Most importantly, we want independent samples from the a posteriori distribution.

If independent sample models are required, one has to wait some time between saving the samples. In practice, a single test run of, say, 1000 iterations is performed, and the value of the likelihood function is recorded for the current model of each iteration. After some iterations the likelihood has risen from the usually very low value of the initial model to a rather stable "equilibrium level", around which it fluctuates during the remaining iterations. By calculating the autocorrelation function for the equilibrium part of this series of likelihood values, it is possible to estimate the waiting time (in iterations) between statistically independent likelihood values. This waiting time a very rough measure of the minimum waiting time between statistically independent model samples from $\sigma(\mathbf{m})$. The waiting time between saving model samples in our computations is 100 iterations. A discussion

of the validity of the above measure is beyond the scope of this paper. It shall, however, be noted that the described method is only approximate and that the crucial problem of estimating how many iterations are needed to yield a sufficient number of samples (to characterize a given inverse problem) is still unsolved.

Making of a Movie

First, the comparison between computed and observed data is "turned off", so as to generate a sample of models representing the a priori probability. This has two purposes. First, it allows us to make statistics and to verify that the algorithm is working correctly. More importantly, it allows us to really understand which sort of a priori information we are inputting to the problem. Figure 8, for instance, shows 30 of the models representing the a priori probability distribution, of the many tens of thousands generated. We call this figure a "movie", as this is the way the whole set of generated models is displayed on a computer screen. These 30 models give an approximate idea of the sort of a priori information used. Of course, more models are needed if we want a more accurate representation of the a priori probability.

We may not be interested in the models per se but only in smooth Earth models (for instance, if we know that only smooth properties are resolved by the data). The movie of Figure 8 then easily becomes the smooth movie displayed in Figure 9 (where the density at each point is arbitrarily chosen to be a simple average over 250 points surrounding it).

"Turning on" the comparison between computed and observed data, i.e., using the Metropolis rule, the random walk sampling the prior distribution is modified and starts sampling the posterior distribution. Figure 10 shows a movie with some samples of the posterior distribution, and Figure 11 shows the smoothed samples.

Let us first concentrate on the a posteriori movie of Figure 10. It is obvious that many different models are possible. This is no surprise, as gravity data do not constrain strongly the Earth model. But it is important to look at Figure 12. We display the a priori and the a posteriori data movie, i.e., the synthetic data corresponding to models of the a priori random walk in the model space and the synthetic data corresponding to models of the a posteriori random walk in the model space, when the Metropolis rule is biasing the prior random walk towards the posterior. Even though the models in the posterior movie of Figure 10 are quite different, all of them predict data that, within experimental uncertainties, are models with high likelihood: gravity data alone can not have a preferred model.

Let us now analyze the smoothed models of Figure 11. They do not look as "random" as the models without smoothing: they all have a zone of high-density contrast centered around 10 km depth, which is a "structure" resolved by the data.

Answering Questions

From the viewpoint defended here, there are no well-posed questions or ill-posed questions, but just questions that have a probabilistic answer.

Making histograms. We may be interested in the value of the mass density at some depth, say z_0. Each of

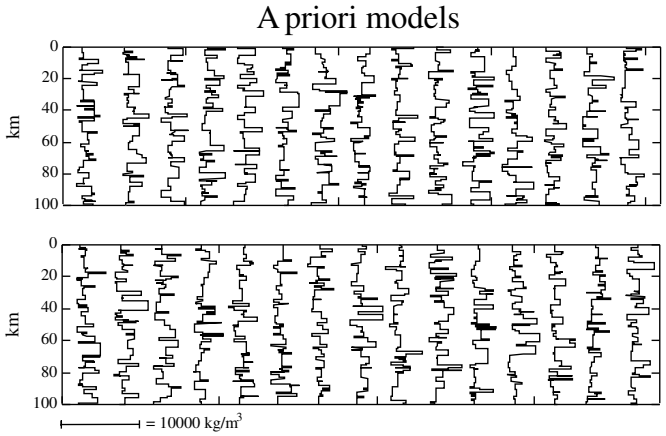


Figure 8: Some images of a movie representing the a priori probability density.

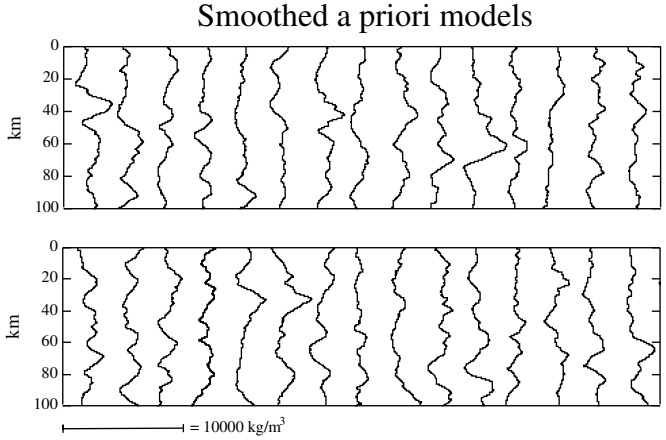


Figure 9: Same as Figure 8 but with the models smoothed.

our many samples (of both the a priori and the a posteriori probability in the model space) has a particular value of the mass density at z_0. The histogram of these values clearly represents the marginal probability distribution for the mass density at that point.

Figures 13 and 14 show both the prior and posterior histograms for the mass density at 2 km, 10 km and 80

km depth, respectively. In particular, we see, when comparing the prior and posterior histograms at 2 km depth, that the mass density to some extent has been resolved there: the histogram has been slightly "narrowed". This is not the case at 80 km depth. Instead of the value of the mass density at some particular depth, we may be interested in the average mass density between, say, z_1 and z_2 . Taking this average for all our samples gives the histogram shown at the bottom of Figure 14.

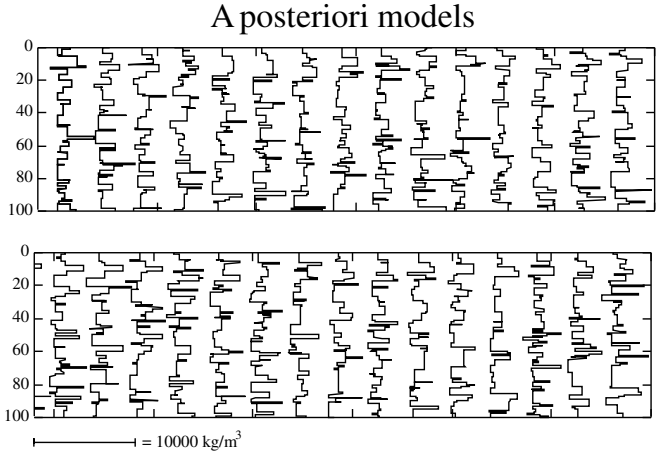


Figure 10: Some images of a movie representing the a posteriori probability density.

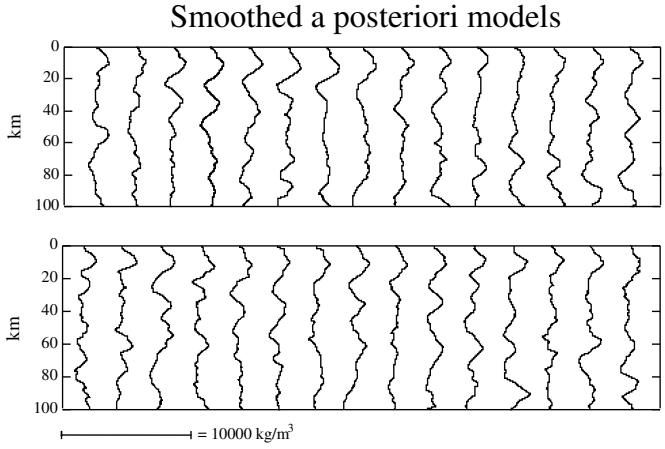


Figure 11: Same as Figure 10, smoothed. The smoothed models do not look as "random" as the models without smoothing (Figure 10): they all have a "bump" at about 10 km depth, which is a "structure" resolved by the data.

Computing central estimators, or estimators of dispersion. Central estimators and estimators of dispersion are traditional parameters used to characterize simple probability distributions. It is well known that while mean values and standard deviations are good measures

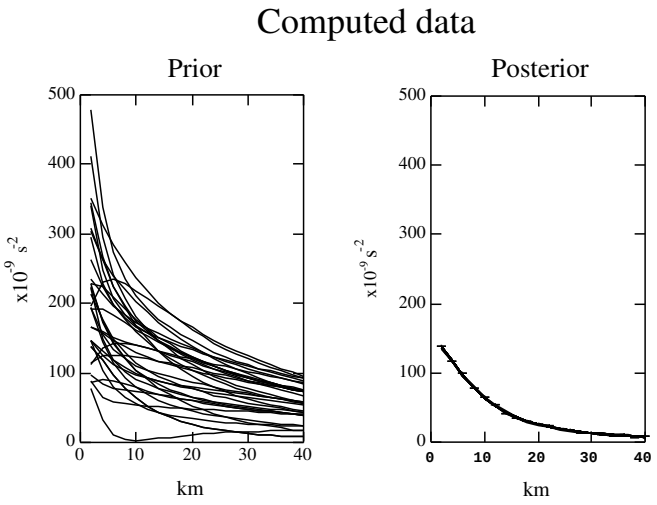


Figure 12: The a priori and a posteriori data movie.

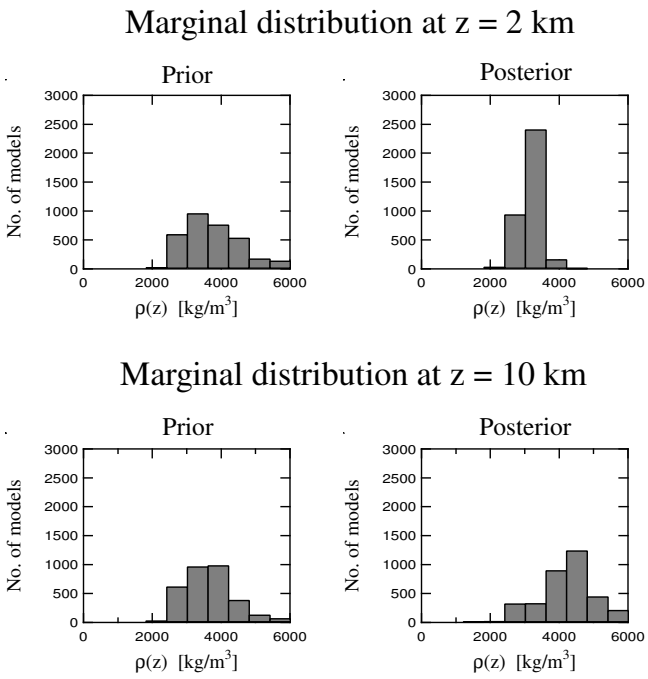


Figure 13: Prior and posterior histograms for the mass density respectively at 2 km and 10 km. When comparing the prior and posterior histograms at 2 km depth, we see that the mass density has been quite well resolved there. the histogram has been considerably "narrowed".

for Gaussian functions, median values and mean deviations are better adapted to Laplacian (double exponential) functions. We can compute both estimators (or any other), as we are not dependent on any particular assumption.

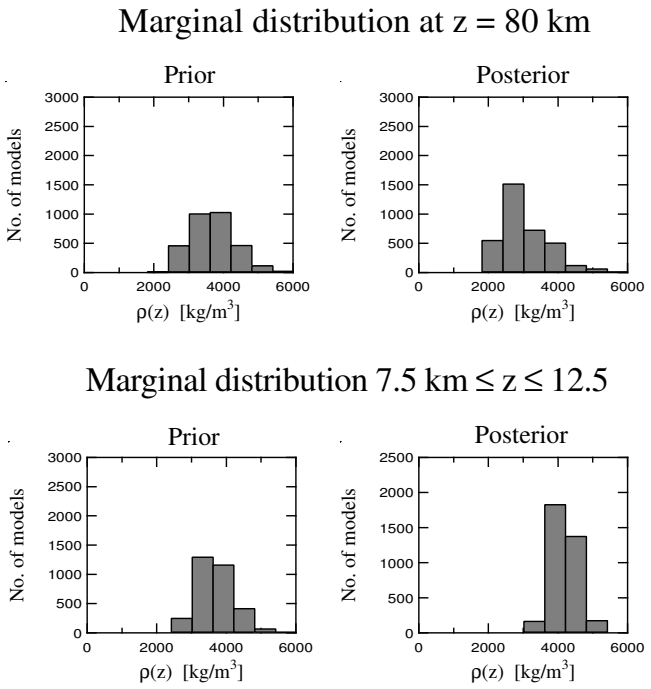


Figure 14: Prior and posterior histograms for the mass density at 80 km depth, and average mass density between 7.5 km and 12.5 km. The mass density at 80 km depth has been less well "resolved" that at 2 km depth (see Figure 13).

Figure 15 shows the mean value for the mass density, plus and minus the standard deviation, and the median, plus and minus the mean deviation for both the a priori and the a posteriori movie. Again, these plots represent the mean and median (and corresponding deviations) of the a priori and a posteriori probability distributions in the model space. Notice that the mean and the median a posteriori models both show the zone of high density contrast centered around 10 km depth, characteristic of the true model of Figure 5, a feature well resolved by our data.

Computing correlations. We may also ask how correlated are the mass density values at different depth locations. From our movies, we can, for instance, easily compute the covariance function $C(z, z')$. The correlation function is given by $c(z, z') = C(z, z')/(\sigma(z)\sigma(z'))$, where $\sigma(z)$ is the standard deviation at point z (just estimated). The correlation function, taking its values in the interval $(-1, +1)$, has a simpler interpretation than the covariance function.

We have chosen to compute the correlation between a point arbitrarily chosen at $z_0 = 10$ km and all other points, i.e., the function $c(z_0, z)$. The result is displayed in Figure 16.

Notice that correlations in the a priori probability distribution decay approximately exponentially, and that they are all positive. In the a posteriori probability distribution, anticorrelations appear. This means, roughly speaking, that if the mass density of any particular realization is in error at 10 km depth, it is likely that it will also be in error, but with opposite sign, in the layers just above and below 10 km.

The approximate exponential decay of the correlation in the prior probability results from the exponential prior probability chosen for the layer thicknesses. The anticorrelations appearing in the posterior probability describe the uncertainty in our posterior models due to the type of information brought by the gravity data.

Discussion

All the results presented in Figures 8 and 9, and the left parts of Figures 13 to 16 concern the a priori movie (i.e., they correspond to the sampling of the model space according to the a priori probability density). Should we at this point decide that we are not representing well enough our a priori information or that we are inputting a priori information that we do not actually have, it would be time to change the way we generate pseudorandom models. If the a priori movie is acceptable, we can "switch on" the synthetic data calculation, and the filter described above, to generate samples of the a posteriori probability distribution, i.e., to produce the a posteriori movie.

It should be properly understood in which way the feature at 10 km depth is "resolved" by the data. None of the models of the a posteriori movie shows a clear density bump at 10 km depth, as the considered inverse problem has a highly nonunique solution (i.e., many different models fit the data and are in accordance with the a priori information). From the a posteriori movie we can not conclude that the true model does have the bump, as many models without it are acceptable. Simply, models with the bump, and arbitrary "high frequencies" superimposed, have a greater chance of being accepted.

General Considerations

There are two major differences between our Metropolis rule (for solving inverse problems) and the original Metropolis algorithm. First, it allows an introduction of non-uniform a priori probabilities. Moreover, an explicit expression for the a priori probabilities is unnecessary: an algorithm that samples the model space according to the prior is sufficient. Second, our Metropolis rule is valid

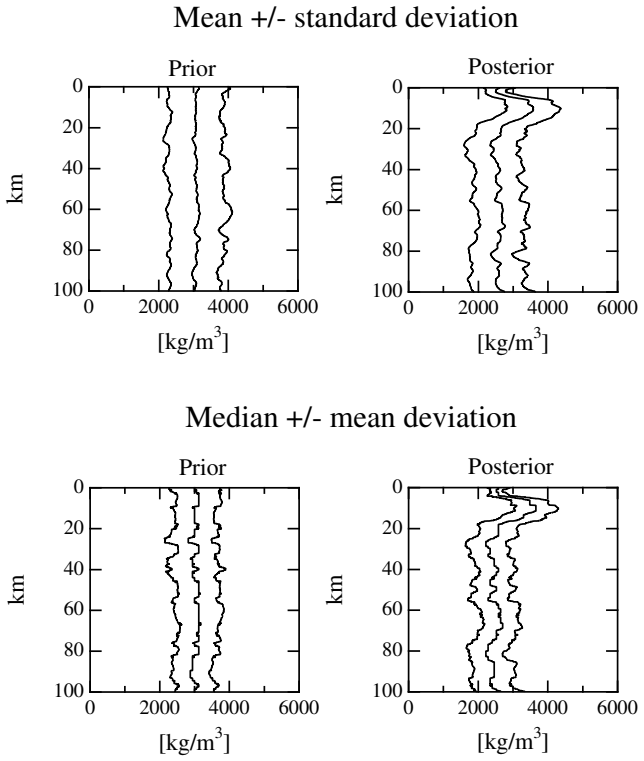


Figure 15: Mean value for the mass density, plus and minus the standard deviation, and the median, plus and minus the mean deviation for both, the a priori and the a posteriori movie. These represent the mean and median (and corresponding deviations) of the a priori and a posteriori probability distributions in the model space.

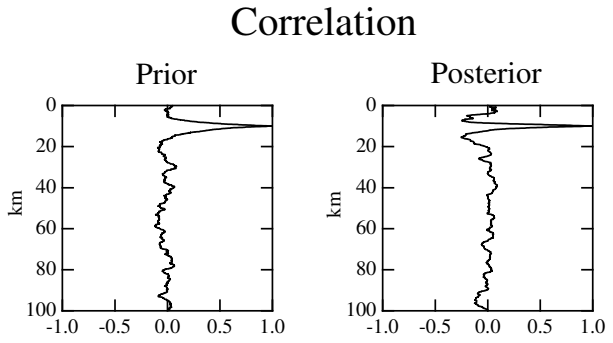


Figure 16: The (left) a priori and (right) a posteriori correlation functions $c(z_0, z)$ for $z_0 = 10$ km. Notice the anticorrelations appearing in the posterior correlation function.

for an arbitrary probability (i.e., it is not linked to the Gibbs-Boltzmann distribution).

Our algorithm has been developed for sampling of discrete spaces according to given probabilities. However, it can be used for optimization. The Metropolis algorithm is already used in simulated annealing [*Kirkpatrick et al.*, 1983], where the desired distribution is changed during the process, starting with a uniform distribution and ending with a near-delta distribution, centered at the optimal solution. We could also find the "best model" by artificially using in the equations values for the experimental uncertainties that tend to zero. However, we do not recommend paying any interest to this concept of "best model".

The method developed above is independent of the way probabilities have been normalized. This is important, as many interesting properties of a probability distribution can be inferred from a random walk, even before the walk has been so extensive that it allows an effective estimation of the denominator of equation (14).

Although we have designed a sampling algorithm (and given proof of its convergence to the desired distribution), we have only addressed heuristically the difficult problem of designing efficient algorithms. It can be shown that the Metropolis rule is the most efficient acceptance rule of the kind we consider (see Appendix C), but the acceptance rule is only part of the efficiency problem: defining the graph (i.e., how the models can be perturbed) is a nontrivial task, and we have only shown an example of it, having no general theory to propose.

Conclusion

We have described a near-neighbor sampling algorithm (random walk) that combines prior information with information from measurements and from the theoretical relationship between data and model parameters. The input to the algorithm consists of random models generated according to the prior distribution $\rho(\mathbf{m})$ and the corresponding values of the likelihood function that carries information from measurements and the theoretical data/model relationship. Output from the algorithm are pseudo-random realizations of the posterior distribution $\sigma(\mathbf{m})$. We applied the algorithm to a highly nonunique, linear inverse problem, to show the method's ability to extract information from noisy data.

The a posteriori distribution contains all the information about the parameterized physical system that can be derived from the available sources. Unfortunately, this distribution is multidimensional and is therefore impossible to display directly.

It is important to direct future efforts toward the development of methods for analyzing and displaying key properties of a posteriori distributions of highly nonlinear inverse problems. For this class of inverse problems, the a posteriori distributions are typically multimodal, and traditional techniques for analyzing error and resolution properties of unimodal a posteriori distributions

break down. There is no known way of understanding uncertainties in the result of a highly nonlinear inverse problem. Here, we have defined the a posteriori probability density $\sigma(\mathbf{m})$, which contains all the information, but how to extract it? Clearly, computing standard deviations or covariances may be meaningless if the posterior probability density is far from Gaussian, which is always the case for highly nonlinear problems. Also, an extensive exploration of the model space can not be made if the space is of high dimension, as, for instance, in the problem of interpretation of seismic reflection data.

In that problem, each model is usually represented by an image. Using the methods described above, we should start by generating pseudo random models with the prior distribution $\rho(\mathbf{m})$. The movie should show models that, on the grounds of our prior information, are more or less likely. In geophysics, this is the right time for a geologist to tell us if he agrees with the movie or if, on the contrary, he sees too many unlikely or too few likely models. When the geologist is satisfied, we now can turn to look at the data, and to run the Metropolis rule, using data misfits, to converge to the posterior probability distribution $\sigma(\mathbf{m})$. The movie is now showing only models which are likely after examination of prior evidence and of geophysical data.

It must be understood that this point of view is much more general than the usual one. For instance, imagine a problem where certain parameters can be resolved deterministically and other parameters can only be resolved statistically. This is the case, for instance, when inverting seismograms to obtain earth models. The major impedance contrasts, for instance, can be deterministically resolved from reflected energy. However, imagine that our space of admissible models contains models with very fine layering, much finer than the seismic wavelength. The position of these very fine layers can not be resolved deterministically, but, as some properties of the seismograms (coda amplitude decay, etc.) do contain information on the average density of fine layers, models (with fine layering) compatible with this information should be generated. Those fine layers could of course not be located individually, but if the data, say, perfectly resolve the average density of a series of layers, all the selected models should display the same average density of these layers. A simple illustration of this possibility has been made here with the "bump" in our mass density models.

From the final collection of models we can start posing questions. Ask for instance for any particular property of the model, for instance, the depth of a particular layer, the smoothed matter density distribution, etc. We have now many examples of that property. It may happen that all the models give the same value for it: the property is well constrained by the data. Some, using old terminology, would say that asking for that property is a "well-posed question". On the contrary it may happen that all the models give absolutely different answers to the question.

In general, we are able to estimate statistics on that property and give answers with a clear probabilistic meaning. In almost all the interesting cases, those statistics will not follow the nice bell-shaped Gaussian distribution, but this should not be an obstacle to a proper analysis of uncertainties. We are well aware of the often tremendous computational task imposed by this approach to inversion. However, the alternative may be an uncertain estimation of uncertainties.

Appendix A: Design of Random Walk With a Desired Equilibrium Distribution

The design of a random walk that equilibrates at a desired distribution $\{p_i\}$ can be formulated as the design of an equilibrium flow having a throughput of p_i particles at point i . The simplest equilibrium flows are symmetric, that is, they satisfy $f_{ij} = f_{ji}$: the transition $i \leftarrow j$ is as likely as the transition $i \rightarrow j$. It is easy to define a symmetric flow on any graph, but it will in general not have the required throughput of p_j particles at point j . This requirement can be satisfied if the following adjustment of the flow is made: first, multiply all the flows f_{ij} with the same positive constant c . This constant must be small enough to assure that the throughput of the resulting flows cf_{ij} at every point j is smaller than its desired probability p_j . Finally, at every point j , add a flow f_{jj} , going from the point to itself, such that the throughput at j gets the right size p_j . Neither the flow scaling nor the addition of f_{jj} will destroy the equilibrium property of the flow. In practice, it is unnecessary to add a flow f_{jj} explicitly, since it is implicit in our algorithms that if no move away from the current point takes place, the move goes from the current point to itself. This rule automatically adjusts the throughput at j to the right size p_j .

Appendix B: Naïve and Uniform Random Walks

Naïve Walks

Consider two arbitrary neighbors, i and j , having n_i and n_j neighbors, respectively, and a random walk with the simple transition probabilities $p_{ji} = 1/n_i$ and $p_{ij} = 1/n_j$ (choosing one of the neighbors, as the next point, uniformly at random). If we want the equilibrium flow to be symmetric, $p_{ji}q_i = p_{ij}q_j$, which is satisfied if $q_i = n_i$.

Furthermore, the above probabilities make all the flows $f_{ji} = p_{ji}q_i$ equal to unity. So, the total throughput through point i is $\sum_k f_{ik} = \sum_j f_{ji} = n_i = q_i$. Hence $q_i = n_i$ must be the equilibrium distribution for the random walk.

Uniform Walks

The rules for the uniform walk follows now directly from applying the Metropolis rule (see later) to the above random walk. The Metropolis acceptance probabilities are $p_{ji}^{\rm acc} = \min(v_j, v_i)/v_i$, where $v_i = 1/q_i$ and $v_j = 1/q_j$ are the "modification probabilities".

Appendix C: Modifying a Random Walk by Introduction of an Acceptance Rule

Consider a random walk P_{ij} with equilibrium distribution ρ_i and equilibrium flow f_{ij} . We can multiply f_{ij} with any symmetric flow ψ_{ij}, where $\psi_{ij} \leq L_j$, for all i and j , and the resulting flow $\varphi_{ij} = f_{ij}\psi_{ij}$ will also be symmetric and hence an equilibrium flow. The transition probabilities of a "modified" algorithm with flow φ_{ij} and equilibrium probability σ_j is obtained by dividing φ_{ij} with the product probability $\sigma_j = \rho_j L_j$. This gives the transition probability: $P_{ij}^{\rm modified} = f_{ij}\psi_{ij}/\rho_j L_j = P_{ij}\psi_{ij}/L_j$, which is equal to the product of two factors: the initial transition probability, and a new probability: the acceptance probability $P_{ij}^{\rm acc} = \psi_{ij}/L_j$. If we choose to multiply f_{ij} with the symmetric flow $\psi_{ij} = \min(L_i, L_j)$, we obtain the Metropolis acceptance probability $P_{ij}^{\rm metrop} = \min(L_i, L_j)/L_j$, which is one for $L_i \geq L_j$, and equals L_i/L_j when $L_i < L_j$. Choosing, instead, $\psi_{ij} = L_i L_j/(L_i + L_j)$, we get the "logistic rule" with acceptance probability $P_{ij}^{\rm log} = L_i/(L_i + L_j)$. The simplest algorithm can be derived from $\psi_{ij} = \min_i(L_i)$, giving the acceptance probability $P_{ij}^{\rm evap} = \min_i(L_i)/L_j$. The acceptance rule for this constant flow we call the "evaporation rule", as the move by a random walker away from the current point depends only on the desired probability at that point and that this recalls the behavior of a water molecule trying to evaporate from a hot point. A last example appears by choosing $\psi_{ij} = L_i L_j$, which gives the acceptance probability $P_{ij}^{\rm cond} = L_i$. We refer to this acceptance rule as the "condensation rule", as it recalls the behavior of a water molecule trying to condensate at a cold point. The efficiency of an acceptance rule can be defined as the sum of acceptance probabilities for all possible transitions. The acceptance rule with maximum efficiency is obtained by simultaneously maximizing ψ_{ij} for all pairs of points j and i . Since the only constraint on ψ_{ij} (except for positivity) is that ψ_{ij} is symmetric and $\psi_{kl} \leq L_l$, for all k and l , we have $\psi_{ij} \leq L_j$ and $\psi_{ij} \leq L_i$. This means that the acceptance rule with maximum efficiency is the Metropolis rule, where $\psi_{ij} = \min(L_i, L_j)$.

Appendix D: An Iteration Consisting of a Randomly Chosen Step

In this case, the transition probability matrix for the iteration is equal to a linear combination of the transition probability matrices for the individual steps. The coefficient of the transition probability matrix for a given step is the probability that this step is selected. Since the vector of desired probabilities is an equilibrium distribution (eigenvector with eigenvalue 1) for each of the step transition probability matrices, and since the sum of all the coefficients in the linear combination is equal to 1, it is also an equilibrium distribution for the transition probability matrix for the complete iteration. This equilibrium distribution is unique, since it is possible, following the given steps, to go from any point to any other point in the space.

Acknowledgments. We thank Zvi Koren and Miguel Bosch for helpful discussions on different aspects of Monte Carlo optimization. This research has been supported in part by the Danish Energy Ministry, the Danish Natural Science Research Council (SNF), the French Ministry of National Education (MEN), the French National Research Center (CNRS,INSU) and the sponsors of the Groupe de Tomographic Gèophysique (Amoco, CGG, DIA, Elf, IFP, Schlumberger, Shell, Statoil).

References

[1] Backus, G., Inference from inadequate and inaccurate data, I, Proc. Natl. Acad. Sci. U.S.A., 65 (I), 1–105, 1970a.

[2] Backus, G., Inference from inadequate and inaccurate data, 11, Proc. Natl. Acad. Sci. U.S.A., 65 (2), 281–287, 1970b.

[3] Backus, G., Inference from inadequate and inaccurate data, 111, Proc. Natl. Acad. Sci. U.S.A., 67 (I), 282–289, 1970c.

[4] Cary, P.W., and C.H. Chapman, Automatic 1-D waveform inversion of marine seismic refraction data, Geophys. J. R. Astron. Soc., 93, 527–546, 1988.

[5] Dorman, L.M., The gravitational edge effect, J, Geophys. Res., 80, 2949–2950, 1975.

[6] Feller, W., An Introduction to Probability Theory and Its applications? New York 1970.

[7] Geman, S., and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intel., PAMI-6, 721–741, 1984.

[8] Jackson, D.D., The use of a priori data to resolve non- uniqueness in linear inversion, Geophys. J. R. Astron. SOC., 57, 137–157, 1979.

[9] Keilis-Borok, V.J., and T.B. Yanovskaya, Inverse problems in seismology (structural review), Geophys. J. R. Astron. SOC., 13, 223–234, 1967.

[10] Kirkpatrick, S., C.D. Gelatt, Jr., and M.P. Vecchi, Optimization by simulated annealing, Science, 220, 671–680, 1983.

[11] Koren, Z., K. Mosegaard, E. Landa, P. Thore, and A. Tarantola, Monte Carlo estimation and resolution analysis of seismic background velocities, J. Geophys. Res., 96, 20289–20299, 1991.

[12] Landa, E., W. Beydoun, and A. Tarantola, Reference velocity model estimation from prestack waveforms: Coherency optimization by simulated annealing, Geophysics, 54, 984–990, 1989.

[13] Marroquin, J., S. Mitter, and T. Poggio, Probabilistic solution of ill-posed problems in computational vision, J. Am. Stat. Assoc., 82, 76–89, 1987.

[14] Metropolis, N., and S.M. Ulam, The Monte Carlo method, J. Am. Stat. Assoc., 44, 335–341, 1949.

[15] Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys., 1, (6), 1087–1092, 1953.

[16] Mosegaard, K., and P.D. Vestergaard, A simulated annealing approach to seismic model optimization with sparse prior information, Geophys. Prospect., 39, 599–611, 1991.

[17] Parker, R.L., Understanding Inverse Theory, Annu. Rev. Earth Planet. Sci., 5, 35–64, 1977.

[18] Pedersen, J.B., and 0. Knudsen, Variability of estimated binding parameters, Biophys. Chem., 36, 167–176, 1990.

[19] Press, F., Earth models obtained by Monte Carlo inversion, J. Geophys. Res., 73, 5223–5234, 1968.

[20] Press, F., An introduction to Earth structure and seismotectonics, Proc. of the Int. Sch. Phys. Enrico Fermi, 209–241, 1971.

[21] Rothman, R.H., Nonlinear inversion, statistical mechanics, and residual statics estimation, Geophysics, 50, 2797–2807, 1985.

[22] Rothman, D.H., Automatic estimation of large residual statics corrections, Geophysics, 51, 332–346, 1986.

[23] Tarantola, A., Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation, Elsevier, New York, 1987. Tarantola, A., and B. Valette, Inverse problems = Quest for information, J. Geophys., 50, 159–170, 1982a.

[24] Tarantola, A., and B. Valette, Generalized nonlinear inverse problems solved using the least squares criterion, Rev. Geophys., 20, 219–232, 1982b.

[25] Wiggins, R.A., Monte Carlo inversion of body wave observations, J. Geophys. Res., 74, 3171–3181, 1969.

[26] Wiggins, R.A., The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure, Rev. Geophys., 10, 251–285, 1972.

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 102, NO. B2, PAGES 2969–2981, FEBRUARY 10, 1997

Monte Carlo analysis of seismic reflections from Moho and the W reflector

Klaus Mosegaard

Dept. of Geophysics, Niels Bohr Institute for Astronomy, Physics and Geophysics, Copenhagen Denmark

Satish Singh and David Snyder

British Institutions Reflection Profiling Syndicate, Bullard Laboratories, University of Cambridge Cambridge, England

Helle Wagner

Dept. of Geophysics, Niels Bohr Institute for Astronomy, Physics and Geophysics, Copenhagen Denmark

Abstract. Near-normal-incidence reflections have been used to image the Moho and the W reflector structure in the lithosphere, offshore northern Scotland. To determine the impedance variations at these reflectors, we use a Monte Carlo technique which allows incorporation of geologically realistic a priori information as well as an extensive exploration of the model space, after testing it on a synthetic data set. The method is based on Bayesian inversion theory. The modeled Moho consists of a series of layers with a total thickness of $\sim 2.4 \pm 0.3$ km with an overall positive impedance contrast. Inversion of the W reflector results in a model of five-seven layers with a total thickness of about 3.7 ± 0.6 km and mostly nonpositive impedance contrasts. The implied fine-scale impedance structure of the Moho is consistent with the broader velocity structure determined from previous wide-angle reflection/refraction profiles. However, the overall nonpositive impedance contrast at the W reflector requires a structure which is overlain or underlain by a broad increase in velocity in order to match amplitudes of reflected phases observed at large offsets interpreted previously to originate at a similar depth.

Introduction

In the last few decades, deep seismic reflection profiling has provided spectacular images of the continental lithosphere worldwide, of which subhorizontal reflections from the lower crust, reflections from the Moho, and reflections from the upper mantle have been particularly notable. Some reflections have been associated with known structures; for example, bright reflections near the Moho depth have been associated with the crust-mantle transition zone, and dipping reflections in the crust have been associated with near surface faults and known subduction zones [*Klemperer and Hobbs,* 1991; *Clowes et al.,* 1992; *BABEL Working Group,* 1993; *Zhao et al.,* 1993; *Clowes and Green,* 1994]. Many features, mainly subhorizontal ones, remain incompletely understood. Although various models have been proposed for these features, one ultimately requires outcrops, drill holes, or the physical properties (density and seismic velocity) of these reflectors to choose from among these models. Deep seismic reflections, as such, are incapable of providing direct information on physical properties. Instead, one derives some estimates of the physical properties from the seismic record using a modeling strategy.

Conventional seismic data processing allows a comparatively fast but rough analysis of large volumes of seismic data. It is essentially based on a linear model of the seismic trace, the so-called convolutional model, in which the seismic trace $s(t)$ is approximated by a convolution of the subsurface reflectivity $r(t)$ and a source wavelet $w(t)$:

$$s(t) = w(t) * r(t)$$

According to the convolution theorem, one consequence of this approximate model is that only those frequencies that are present in the source wavelet can be retrieved from the reflectivity series of a recorded seismic trace. Seismic sources used in conventional deep seismic marine experiments typically have 5 to 80-Hz bandwidths [*Hobbs and Snyder,* 1993]. Due to attenuation, a seismic wavelet at a given two-way time will have little energy above 60 Hz, and the convolutional model alone cannot

Paper number 96JB02566.
0148-0227/97/96JB-02566$09.00

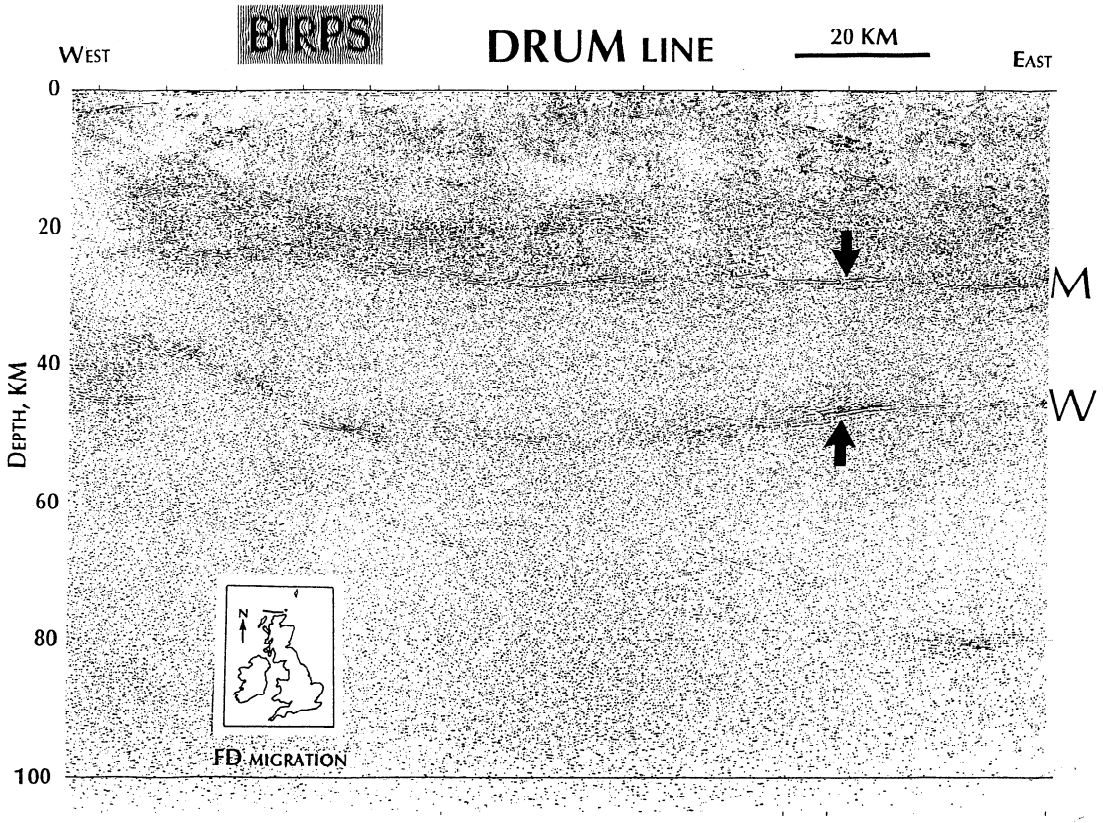


Figure 1. The DRUM seismic section showing the Moho (M) and W (W) reflectors and the location of the data used in the Monte Carlo inversion. The section was migrated using a two-dimensional velocity field derived from nearby refraction results [*Snyder and Flack,* 1990] and then depth converted using the same velocities. The Flannan reflector can also be observed, dipping from 30-km depths at the western edge of the profile to 80-km depths in the east.

provide any information about the reflectivity at spatial frequencies equivalent to ≥ 60 Hz at that two-way time. At frequencies less than 5 Hz, reflected seismic energy is normally absent or obscured by ship noise. Consequently, the convolutional model is unable to predict any low-degree trend or nonzero average value of the reflectivity.

If one is interested in extracting information outside the limited passband of the source wavelet, it is necessary to incorporate a priori information about the subsurface, abandon the convolutional model and replace it with the correct relationship between seismic data and Earth model. In this paper, we propose a framework under which variations in impedance inside and outside the frequency passband of the seismic wavelet can be estimated by nonlinear inversion using prior information on the model in terms of probabilities. A Monte Carlo inversion technique [*Mosegaard and Tarantola,* 1995] provides models which fit the observed data and estimates errors and resolution of these models. We apply this framework to shot records from the DRUM (Deep Reflections from the Upper Mantle) reflection profile from the north of Scotland to estimate impedance variations at the Moho and at the W reflector (Figures 1 and 2) and to provide estimates of the resolution of these impedances. The DRUM line was designed to investigate the reflectivity of the lower continental lithosphere and was recorded for 30 s two-way travel time (TWT), corresponding to about 110-km depth [*McGeary and Warner,* 1985]. Various dipping reflections are observed in the upper crust (0 to 5 s) and a series of subhorizontal reflections in the lower crust (6 to 9 s), the base of which, at ~ 9 s TWT, coincides with the Moho defined by nearby refraction observations [*Barton,* 1992]. Three strong reflectors occur in the mantle: a subhorizontal reflector around 13-15 s TWT (the W reflector), an easterly dipping reflector (the Flannan reflector) recorded from 9 s down to at least 27 s and possibly 30 s, and a 15-km-long banded zone at 23 s (Figure 1) [*McGeary and Warner,* 1985]. Derivation of the physical properties of these deep reflectors from the seismic records is critical to understanding the formation and evolution of the continental lithosphere around the British Isles.

A Probabilistic Formulation of the Problem

Analysis of the resolution of subsurface structures from seismic data requires a probabilistic formulation of the inverse problem. Due to the often strongly non-

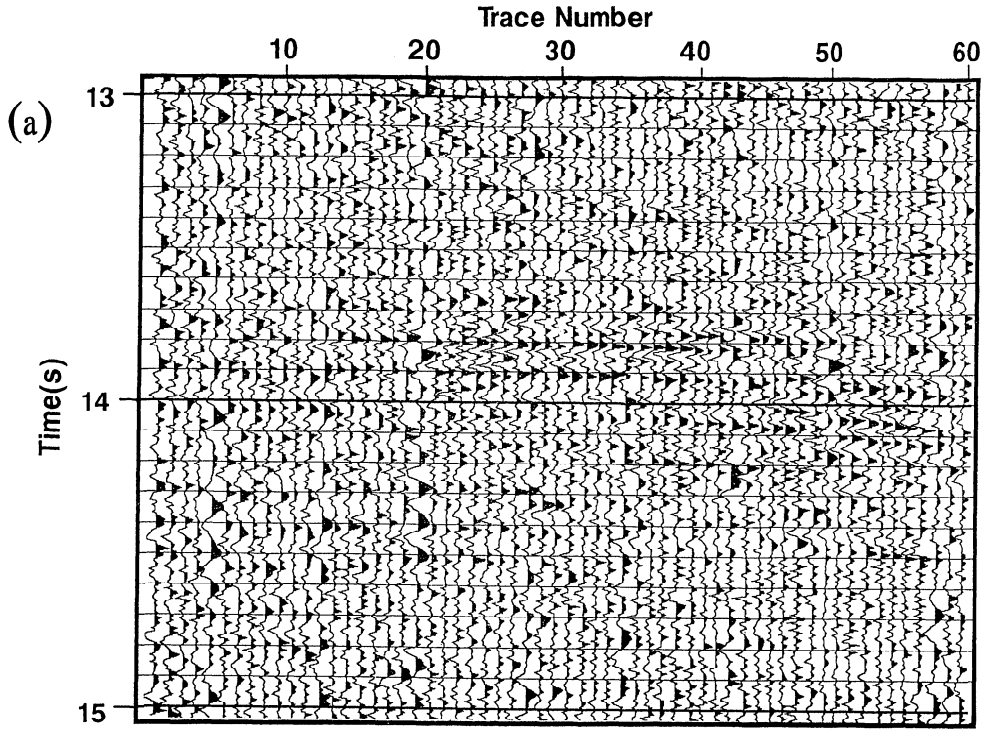


Figure 2. The data used for inversion. (a) The shot gather from 13 to 15 s at shot point 3564 of the DRUM profile, a TWT zone for the W reflector. These 60 traces have increasing offset from left (209 m) to right (3210 m). These traces were summed in groups of six to produce 10 traces with enhanced signal-to-noise ratios for the inversion. (b) The summed traces over the time range of 8-10 s that contains the Moho reflection at about 8.9 s. (c) The summed traces over the the time range of 13-15 s that contains the W reflection at about 13.8 s. (d) Amplitude spectrum of data with Moho reflection. (e) Amplitude spectrum of data with W reflection. (f) Estimated marginal noise distribution for a single data sample (Moho data). The solid curve is a Gaussian distribution. (g) Estimated noise autocorrelation (Moho data).

linear relationship between the subsurface impedance model and seismic data, the distribution of errors in the observed data is mapped into the model space as a complex error distribution. Among the important pathologies of this relationship is an inherent nonuniqueness of models that fit the data. Further complexity is added to the model distribution when we, on geological grounds, must introduce complex, data-independent a priori information to further weigh models based on their geological feasibility.

We use a Bayesian formulation of the inverse problem. In this formulation, the state of information about the subsurface after incorporation of both a priori information and data information is completely described by the a posteriori probability density $\sigma(\mathbf{m})$ over the model space [*Tarantola and Valette*, 1982]. From $\sigma(\mathbf{m})$ it is possible to calculate the probability that the true model belongs to a given class $\mathcal{A}$ of models:

$$P(\mathbf{m} \text{ belongs to } \mathcal{A}) = \int_{\mathcal{A}} \sigma(\mathbf{m}) d\mathbf{m},$$

The a posteriori probability density is the complete solution to the inverse problem [*Tarantola and Valette*, 1982]. It contains all the available prior information such as the approximate sizes of reflection coefficients and layer thicknesses, and all the data information, as "seen through the glasses" of the theoretical relationship between model and data in the form of the wave equation. In the Bayesian analysis, all the input information is preserved and no uncontrolled subjective bias is introduced.

The a posteriori probability distribution for the inverse problem is given by

$$\sigma(\mathbf{m}) = \frac{\rho(\mathbf{m})\ L(\mathbf{m})}{\int_M dm^1 dm^2 \ldots\ \rho(\mathbf{m})\ L(\mathbf{m})} . \qquad (1)$$

[*Tarantola and Valette*, 1982]. The a posteriori probability density $\sigma(\mathbf{m})$ equals (except for a normalization constant) the a priori probability density $\rho(\mathbf{m})$ times a likelihood function $L(\mathbf{m})$ measuring the fit between observed data and synthetic data calculated from the model $\mathbf{m}$.

The likelihood function is typically of the form

$$L(\mathbf{m}) = C \exp\left[-S(\mathbf{m})\right]$$

where C is a constant and $S(\mathbf{m})$ is a misfit function. $S(\mathbf{m})$ measures the difference between the observed

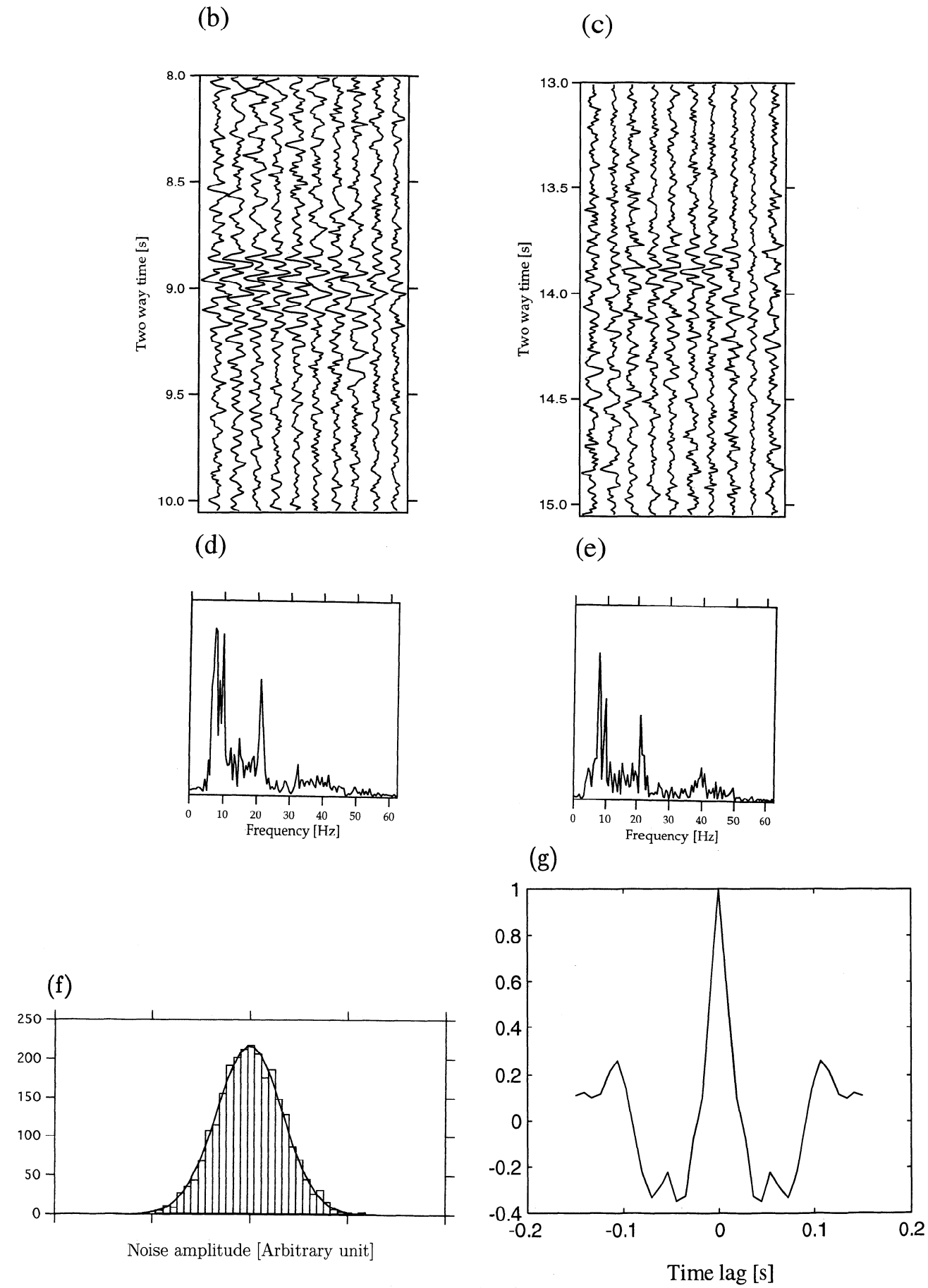


Figure 2. (continued)

data and synthetic data calculated from the model $\mathbf{m}$. If, for instance, the observational noise consists of independent Gaussian errors, $S(\mathbf{m})$ is proportional to the sum of squared differences between observed and calculated data values (the L_2 misfit), and the likelihood function is Gaussian. If, instead, the noise consists of independent Laplace distributed errors, $S(\mathbf{m})$ is the L_1 misfit (where squares are replaced by absolute values), and the likelihood is Laplacian.

Monte Carlo Sampling of the A Posteriori Distribution

Real, quantitative geological a priori knowledge cannot be described by means of simple mathematical expressions for $\rho(\mathbf{m})$. Such knowledge is often available as statistical information: histograms giving the occurrence frequency of, for instance, certain lithologies or physical rock parameters, observed in outcrops or nearby wells.

Mosegaard and Tarantola [1995] provide a method for Bayesian Monte Carlo inversion that overcomes this problem. This method has two major advantages, as compared to previously published methods [e.g., *Stoffa and Sen,* 1991]. First of all, the method is exact in the sense that it will provably sample the posterior probability density. Second, it allows incorporation of arbitrarily complex statistical a priori information into the inversion.

The algorithm consists essentially of two interacting parts. The first part is an a priori model generator that is able to produce random subsurface models, consistent with the available a priori knowledge. Consistency here means that the generated models have exactly the same statistical properties as those we have obtained from observations in the real Earth. The models produced by the a priori model generator are fed into the second part, an algorithm that decides if the a priori model can pass a data fitting test.

One iteration of the algorithm runs as follows: First, given the current model, a new model is chosen by the a priori model generator (by perturbing the current model), and the probability for a given new model to be chosen is proportional to its a priori probability. The new model $\mathbf{m}_{\text{new}}$ is now accepted or rejected according to the following rule:

1. If the value of the likelihood $L(\mathbf{m}_{\text{new}})$ of the new model is larger than or equal to the likelihood $L(\mathbf{m}_{\text{cur}})$ of the current model, the model $\mathbf{m}_{\text{new}}$ is accepted with probability 1.

2. If the value of the likelihood $L(\mathbf{m}_{\text{new}})$ is smaller than the likelihood $L(\mathbf{m}_{\text{cur}})$, the model $\mathbf{m}_{\text{new}}$ is accepted (as the next "current model") only with probability

$$P_{\text{accept}} = \frac{L(\mathbf{m}_{\text{new}})}{L(\mathbf{m}_{\text{cur}})}.$$

If the new model is rejected, the current model also becomes the next current model.

The series of "current models" produced by this two-part algorithm are, asymptotically, samples from the a posteriori distribution $\sigma(\mathbf{m})$ [*Mosegaard and Tarantola,* 1995]. After a large number of iterations, the number of times a given model $\mathbf{m}$ occurs in the collection of "current models" is approximately proportional to $\sigma(\mathbf{m})$. A large collection of such samples from $\sigma(\mathbf{m})$ provides the raw material from which various characteristics of the models can be obtained [*Mosegaard and Tarantola,* 1995]. A set of statistically independent samples of the accepted models allows structures in the subsurface that are well-resolved to be distinguished from those that are ill-resolved. A well-resolved structure will appear in most of the accepted models, whereas an ill-resolved structure will appear in only a few models. The probability that a certain structure exists is roughly proportional to its frequency of occurrence in the set of a posteriori samples. Therefore approximate a posteriori probability distributions for model parameters can be represented by normalized histograms of the parameters values. The peak of the histogram will be at the most probable model parameter, and the deviation from this peak will provide a measure of uncertainty.

Monte Carlo Sampling of Lithosphere Reflectivity

We use the Bayesian Monte Carlo algorithm described above to generate reflectivity models for selected parts of the DRUM reflection profile. We have concentrated our efforts only at a location (Figure 1) where the Moho and W reflectors are brightest and subhorizontal as the algorithm requires large computation time. We have analyzed a total of 60 unprocessed traces in the interval 8.0 s - 10.0 s TWT bracketing the Moho and the interval 13.0 - 15.0 s covering the "W reflector." We chose the 60 traces from one shot gather (3564) with the best signal to noise ratio among 10 neighboring shot records. The receiver group spacing was 50 m along a 3000-m-long streamer. The raw shot gather shows coherent energy between 13.7 s and 14.2 s TWT for the W reflector (Figure 2a). To increase the signal to noise ratio, we applied a correction for dip move out and stacked six adjacent traces, yielding 10 traces for the inversion (Figures 2b and 2c). The normalized frequency spectra show dominant energy between 6 and 30 Hz (Figure 2d and 2e). Since the Moho and W reflectors are subhorizontal and are at large depths, the incident seismic waves are essentially vertically traveling plane waves. It is therefore possible to perform a rather careful calculation of synthetic seismograms using a fast one-dimensional propagator matrix method [*Haskell,* 1953].

Wavelet Estimation

The DRUM profile was shot by GECO (Geophysical Company of Norway) using an 8536-cubic-inch air gun array at 7.5 m below the sea surface [*McGeary and Warner,* 1985]. A simulated far-field source signature

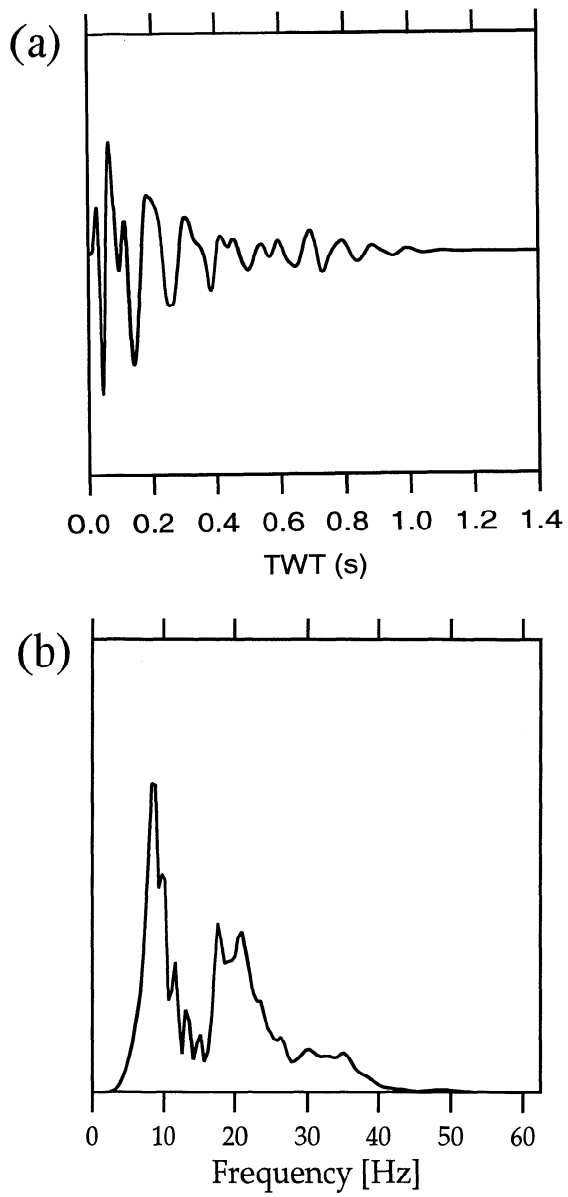


Figure 3. (a) The source wavelet used in the inversion. This trace represents a simulated far-field source signature for the entire air gun array provided by the contractor (GECO), convolved with the receiver ghost, recording filter, and reverberations within the 55-m-thick water layer. The wavelet was then propagated to the appropriate two-way travel time using a quality factor of 500 to approximate effects of attenuation. (b) The associated frequency spectrum.

for the air gun configuration used in the experiment, provided by GECO, was used in this study. Since the streamer was at 17-m depth, the source wavelet was convolved with the appropriate receiver ghost and was then filtered using a minimum phase band-pass filter, 5.3 Hz at 18 dB/octave to 45 Hz at 72 dB/octave, the filter used in recording the data. The effect of reverberations in the water layer at both the source and receiver ends of the ray path was included in the source wavelet. This estimate used a water depth of 55 m measured by an echo sounder during the survey, and a water-seabed reflection coefficient of 0.5 and a water velocity of 1.5 km/s determined using refracted arrivals from the seabed that were observed in shot gathers. The resulting source wavelet was propagated down to 8 s for the Moho and to 13 s for the W reflector, and the effect of attenuation was incorporated using a quality factor of 500 [*Hobbs and Snyder,* 1993] (Figure 3).

The size of the inverse problem was in this way reduced by focusing only on two zones, each of 2 s (TWT), containing the Moho or W reflections. In the calculation of synthetic seismograms it was assumed that the response from each interval considered depended on the overlaying layers (the overburden) only via the propagation effects built into the source wavelet. We neglected in this approximation all multiples in the earth that involve reflectors in the overburden. These multiples are considered negligible due to the rather small reflection coefficients (no more than 0.1) generally assumed to be present in the lithosphere [*Holbrook et al.,* 1992] and the presence of attenuation in the Earth.

A Priori Information

Reflections from deep in the lithosphere require large impedance contrasts. The simplest approach is to treat each reflection as originating at the boundary of a layer overlying a half space and then estimate reflectivity, which could be due to either a positive or negative impedance contrast. Such an approach is appropriate when a reflection is sharp and distinct and the data have no noise. In practice, reflections consist of a series of events, which cannot be distinguished, and often contain considerable noise. In such cases, this simple model is inappropriate and additional information is required.

To obtain appropriate a priori information on the structure of deep reflections, we considered the possible geological processes which might be responsible for these reflectors. These are widely considered to include layered mafic intrusions, shear zones with possible metasomatic residuals enhancing reflectivities, or alternating types of metamorphic rocks [*Mooney and Meissner,* 1992; *Clowes and Green,* 1994]. We have chosen one of these possibilities as a basis for generation of a priori models: layered igneous intrusions in a large magma chamber formed by fractional crystalization of magma derived from the mantle [*Weibe,* 1993]. Physical properties of rocks found in the igneous intrusive complexes of Rum, an island west of Scotland, and Great Dyke in Zimbabwe form the basis of our quantitative a priori information [*Singh and McKenzie,* 1993]. We have only used that part of the information from the igneous complexes that is expected to be fairly independent of the depth of burial, that is, reflection coefficients and layer thicknesses. In other words, there are no constraints on the absolute acoustic impedance values in our calculations. Any other plausible geolog-

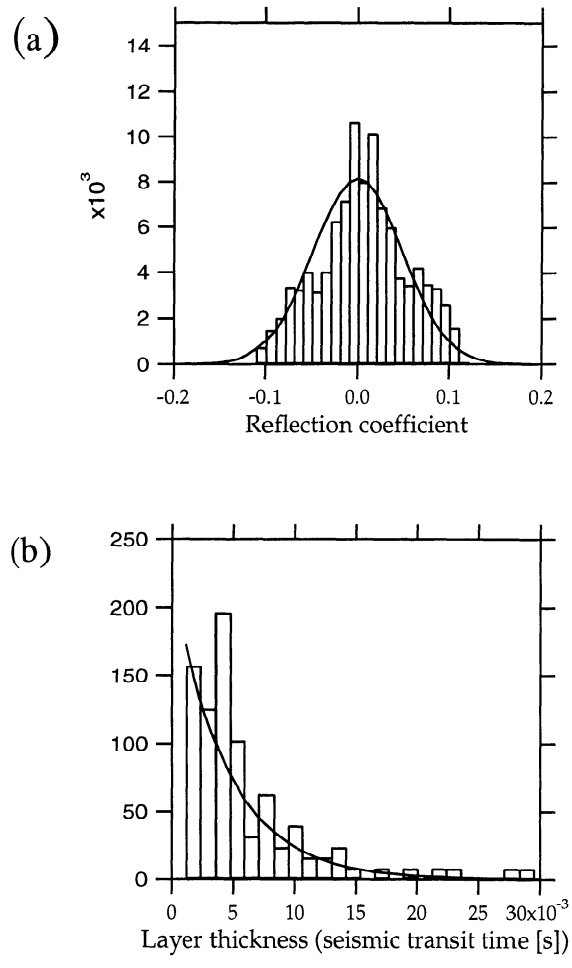


Figure 4. A priori information used by the a priori model generator in the first stage of the Bayesian inversion scheme. (a) Reflection coefficients distribution obtained from studies of the igneous intrusions of Rum, Scotland, and the Great Dyke, Zimbabwe [*Singh and McKenzie,* 1993], that are consistent with other studies such as those of the Pleasant Bay layered gabbro-diorite [*Weibe,* 1993]. The solid curve has a Gaussian distribution with standard deviation $\sigma = 0.047$. (b) Layer thickness distribution as a function of one-way travel time derived from observations of the Rum and Great Dyke intrusions. The solid curve has an exponential distribution with $\lambda = 225.0\ s^{-1}$.

ical model could have been used as prior information, but the igneous intrusions have been more thoroughly mapped on the surface and relevant statistics are available with greater detail than for either shear zones or metamorphic layering.

Histograms (Figure 4) illustrate the distributions of reflection coefficients and the layer thicknesses for these intrusive complexes that guided our a priori model generator. The thicknesses of the layers have been converted from meters into seconds using the velocity in each layer derived from the modal composition [*Singh and McKenzie,* 1993]. In order to retain only the general features of the reflectivity and thickness histograms, they were approximated with a Gaussian and an exponential distribution, respectively. In this way we avoided detailed histogram structure that is only characteristic of the specific igneous intrusion complexes considered in this study. All a priori models generated by our algorithm are consistent with the approximate distributions; that is, histograms of reflection coefficients and layer thicknesses produced from these models are statistically equivalent to corresponding histograms from the igneous complexes. The combined a priori distribution used here allows models with reflection coefficients ranging from about -0.1 to +0.1 and thicknesses from about 10 m to 2000 m.

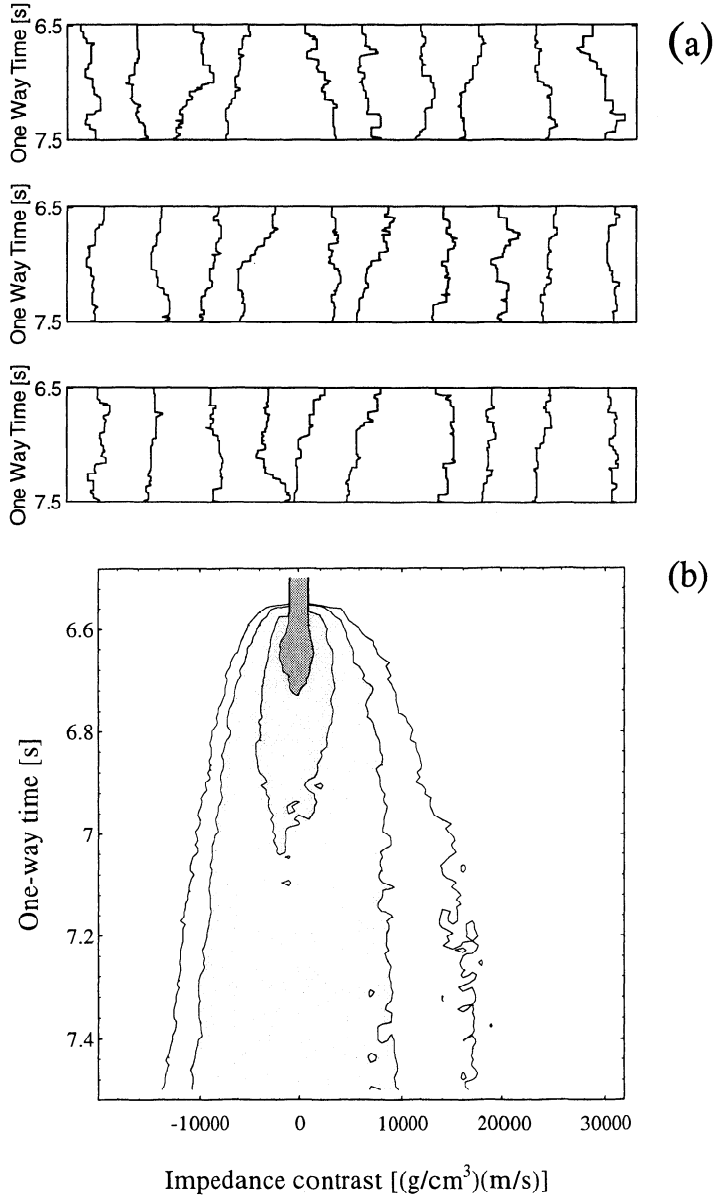


Figure 5. (a) A selection of typical a priori models generated assuming the distributions shown in Figure 4. These models are all statistically equivalent and have equal likelihoods of existence. (b) Marginal a priori impedance contrast distributions for all one-way times in a 1-s interval.

Figure 5 shows a selection of a priori impedance models generated using the statistics derived from Figure 4 for the igneous intrusions. All the a priori models are different, but they all have identical statistics. These impedance models have been calculated from their corresponding reflectivity models, and for comparison, all models have the same impedance value, 19,000 m/s(g/cm^3), at the top. Under this assumption, the weakness of the constraints on the absolute acoustic impedance values can be observed directly, since unrealistically high and low values of impedance occur (Figure 5). Consequently, our a priori information will provide no strong impedance constraints on the results of our calculations.

Model Parameterization and Calculation of Synthetic Seismograms

The complexity of the a priori information means that $\rho(\mathbf{m})$ and hence $\sigma(\mathbf{m})$ deviate strongly from a Gaussian distribution. Consequently, the inverse problem is not easily analyzed by conventional, linearized methods. In particular, a correct, nonlinear error and resolution (nonuniqueness) analysis may only be possible through a Monte Carlo approach.

We employed a Monte Carlo technique for inversion, and therefore a fast computation of synthetic seismograms was required in order to be efficient. Synthetic seismograms were computed using a propagator matrix method [*Ganley*, 1981]. All multiple reflections, absorption, and dispersion within the target zones were incorporated. Dispersion was derived from Q, using a dispersion model given by *Futterman* [1962].

To further speed up the Monte Carlo inversion, model parameterization consisted of reflection coefficients specified as a function of one-way travel time. As the relation between reflectivity and data is only moderately nonlinear, this parameterization (and the fact that the noise is Gaussian) makes $L(\mathbf{m})$ become close to a Gaussian. Consequently, the inverse problem is easier to solve. Each model consists of 128 reflection coefficients with a sampling interval of 0.008 s.

Data Noise and the Likelihood Function $L(\mathbf{m})$

Each of the data sets considered consists of 10 neighboring traces of 2.048-s length (Figure 2b and 2c). Because coherent events in the data sets are approximately horizontal and laterally invariant, we estimated the incoherent noise in the data by first finding one of the best fitting horizontally stratified models for the data set. The (10 identical) vertical incidence traces computed from this model were then subtracted from the 10 data traces to form corresponding error traces, the 10 noise traces. These noise traces were all fairly similar, stationary signals, and the distribution of noise values was very close to a Gaussian. Since the data noise has a Gaussian distribution, we define the likelihood function as

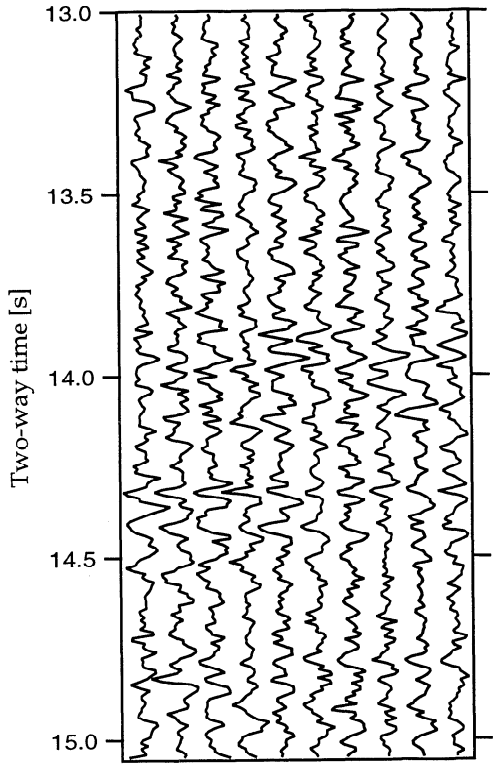


Figure 6. Noise contaminated, synthetic data used to test the algorithm. The noise (and signal to noise ratio) is equal to the noise extracted from the Moho data.

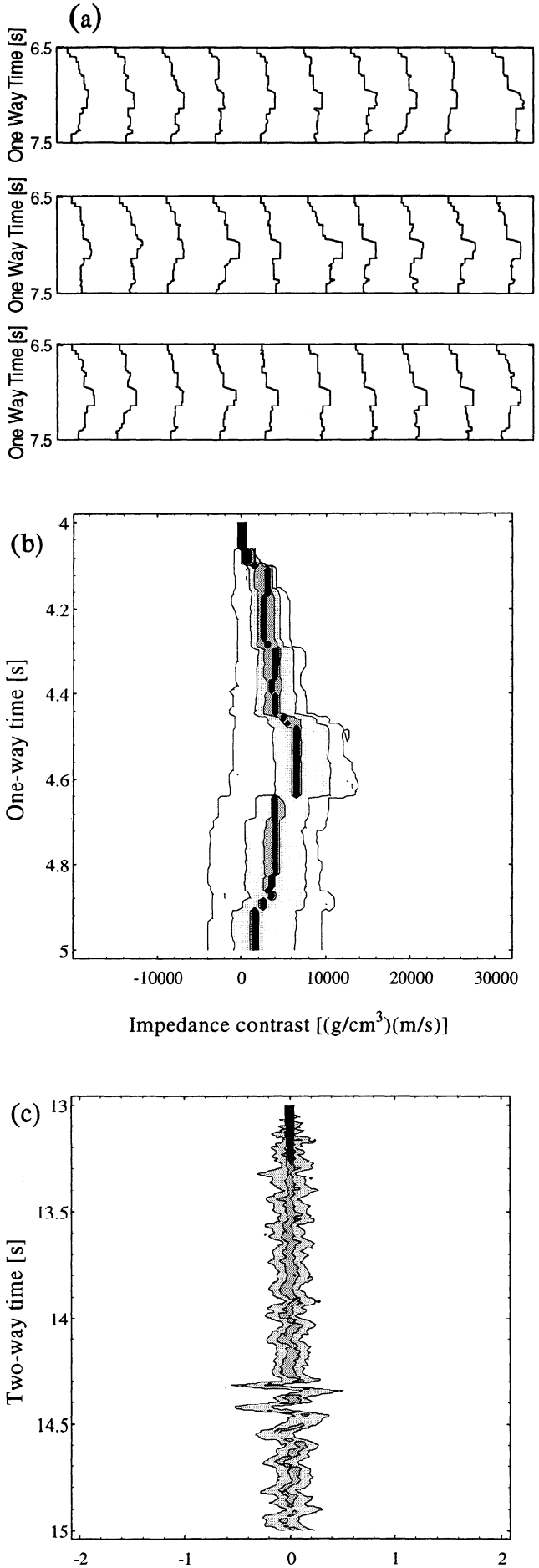
$$L(\mathbf{m}) = \exp\left\{-\frac{1}{2}\left[\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}}\right]^T \mathbf{C}_d^{-1}\left[\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}}\right]\right\}, \tag{2}$$

where $\mathbf{C}_d$ is the covariance matrix for the noise. For the Moho data and W reflector data we found signal to noise (amplitude) ratios of 1.9 and 1.4, respectively. The estimated marginal distribution of a single noise value is shown in Figure 2f, and the estimated noise autocorrelation (normalized to one at zero lag) is shown in Figure 2g.

Synthetic Example

Our algorithm was calibrated and tested on synthetic data from a known subsurface model. The synthetic test was designed to mimic a situation similar to the analysis of the Moho reflector. The DRUM source wavelet was used, and the noise extracted from the Moho reflector data was added to the synthetic traces to form an artificial, noisy data set (Figure 6).

We ran 100,000 iterations of the Monte Carlo inversion on the synthetic data set. The models were generated according to the a priori information derived from the igneous complexes of Rum and Great Dyke (Figure 4). The second part of the Monte Carlo algorithm, accepting or rejecting models proposed by the a priori model generator, used a likelihood function defined by (2) with a noise variance corresponding to a signal to noise ratio of 1.9, the signal to noise ratio of the Moho reflector data.



In each iteration, all 128 reflection coefficients are considered by the algorithm. Each of the reflection coefficients are perturbed by the a prior model generator, and the perturbed model is accepted or rejected according to the rule given above. The series of "current models" produced by this algorithm are samples from the a posteriori distribution $\sigma(\mathbf{m})$. The strategy of perturbing only one reflection coefficient at a time preserves most characteristics of the current model, which may have fitted the data well. This strategy is efficient, since we want to visit many models with a good data fit, but it provides models that are successively correlated. This is a problem since error and resolution analysis requires a collection of statistically independent models from the a posteriori distribution. A smaller set of models chosen from among the accepted models in such a way that they are sufficiently separated in time (iterations) constitutes such a set of independent models. Here, as in the analyses of the field data below, we chose to save only every hundredth model. This waiting time of 100 iterations between accepted models was found by analyzing the fluctuations of $L(\mathbf{m})$ as the iterations proceeded. Inspection of the autocorrelation function for these fluctuations showed that accepted models separated by a hundred iterations were unlikely to be correlated. Thus we have 1000 independent samples of the a posteriori distribution from 100,000 sampled models.

The results are shown in two ways. Figure 7a shows the original model used to generate the synthetic data in the upper left corner together with a selection of a posteriori models, randomly selected from the 1000 a posteriori models selected by the algorithm. Since these models are randomly chosen from the 1000 independent samples of the posteriori models, they roughly approximate the a posteriori distribution. In Figure 7c, the distribution of synthetic data is shown. All the synthetic traces are statistically indistinguishable from the reference trace calculated from the original impedance model, in that their deviations from this noise trace are within the noise level.

The variations in the model that are obtained by the Monte Carlo inverion, and hence permitted by the a priori information and the data, are evident. To interpret this output correctly, it should be remembered that the method is designed to produce particular model features with a frequency proportional to their a poste-

Figure 7. The results of the inversion on the synthetic data. (a) The true model (the curve at the upper leftmost corner) and a selection of a posteriori models. (b) Marginal a posteriori impedance contrast distributions for all one-way times between 6.5 and 7.5 s. (c) Marginal a posteriori data distributions for all two-way times between 13.0 and 15.0 s. Since all impedance models are fixed at 19,000 g/cm^3(m/s) near 13.0 s, these distributions approach a delta function at the top of the figure.

riori probability. This means that the probability of a feature that exists in the impedance model is roughly proportional to the number of times the feature occurs on Figure 7a. If it appears on almost all the a posteriori models, it is well resolved. It is clear from Figure 7a that, for instance, the high impedance zone between 6.950 and 7.150 s is well resolved. The actual impedance contrast value between the interior of the zone and its surroundings (above and below) shows some variation between the a posteriori models and is therefore poorly resolved.

Of particular interest to this study is the overall impedance contrast, that is, the increase in impedance from the top to the bottom of the zone. In Figure 7b, the marginal a posteriori distributions for impedance contrasts at all the considered depths are shown. It is clear from this figure that the magnitude of the overall change in impedance is poorly resolved. However, the polarity (sign) of the overall impedance contrast is well resolved. As seen from the impedance contrast distribution at 7.5 s one-way time, it has a very high probability of being positive, in agreement with the original model (shown in black in the figure).

Moho

A similar analysis of the actual Moho data from the DRUM profile results in the a posteriori models shown in Figure 8a and 8b, respectively. Due to the fact that the estimated wavelet may be in error by an unknown, constant scaling factor k, the only information we can extract from the data is the ratio between impedance contrast $\Delta I(T)$ and k. The impedance contrast is the impedance at one-way time T minus impedance at the top of the target zone. See the Appendix for further details on this problem. Figure 8a shows this ratio as a function of one-way travel time in the target zone. Figure 8b shows estimated marginal a posteriori distributions for impedance contrasts at all depths in the considered interval.

An inspection of the a posteriori Moho models (Figure 8a) shows a well-resolved layered sequence having a thickness of about 0.3 ± 0.04 s one-way time, equivalent to approximately 2.4 ± 0.6 km. Impedances alternately increase and decrease within the series, but overall the series has a cumulative increase. The observed, well-resolved, positive polarity of the overall impedance contrast for this interval (Figure 8b) is consistent with previous refraction results that used diving rays and wide-angle reflections to model observed phases [*Barton*, 1992].

W Reflector

The calculated a posteriori models for the W reflector zone are shown in Figure 9a. Some well-resolved, consistent features are observed in the interval between 6.8 s and 7.3 s one-way time. Over this short interval, with an approximate thickness of 3.7 ± 0.6 km, the

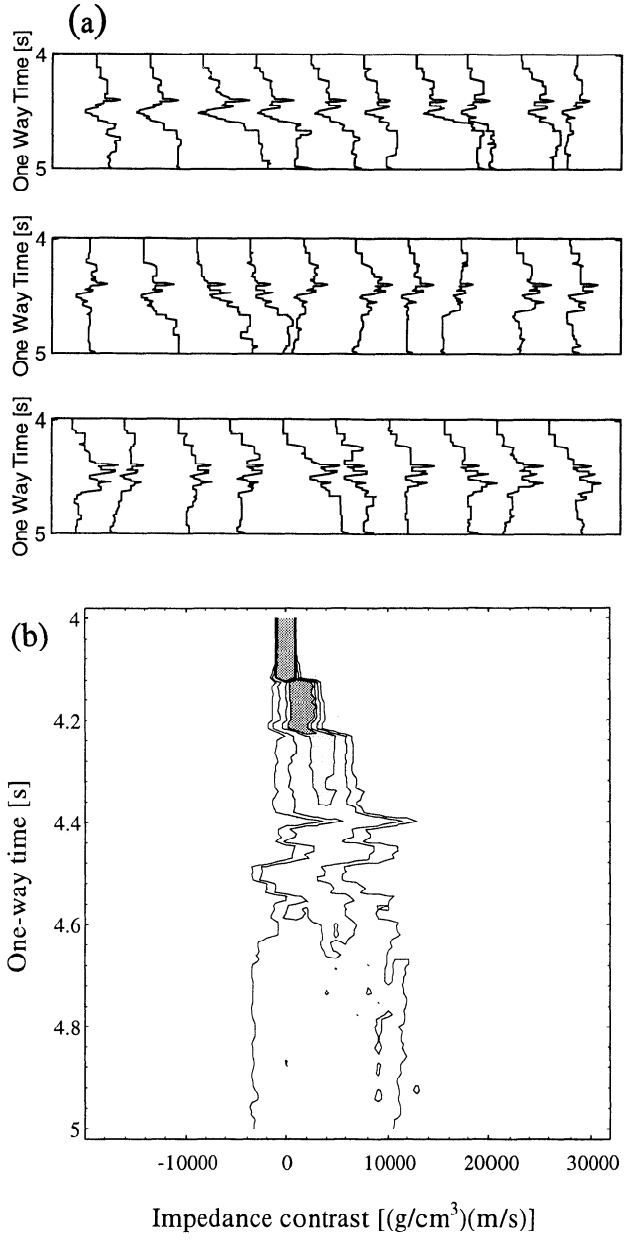


Figure 8. The results of the inversion on the Moho data set. (a) A selection of a posteriori models. (b) Marginal a posteriori impedance contrast distributions for all one-way times between 4.0 and 5.0 s.

overall change in acoustic impedance is poorly resolved (Figure 9b). The consistency of the models between 6.8 and 6.9 s indicates a high probability of negative impedance contrasts over this zone about 1.25 km thick. The greater variability of a posteriori models below 6.9 s indicates some probability for both positive and negative contrasts deeper in the interval. The polarity of the overall impedance contrast is not as well resolved as over the Moho interval described above. The marginal impedance contrast histogram at 7.5 s one-way time (Figure 9b) reveals that the a posteriori probability of a nonpositive polarity, approximated by the portion of this histogram around and to the left of the origin, is

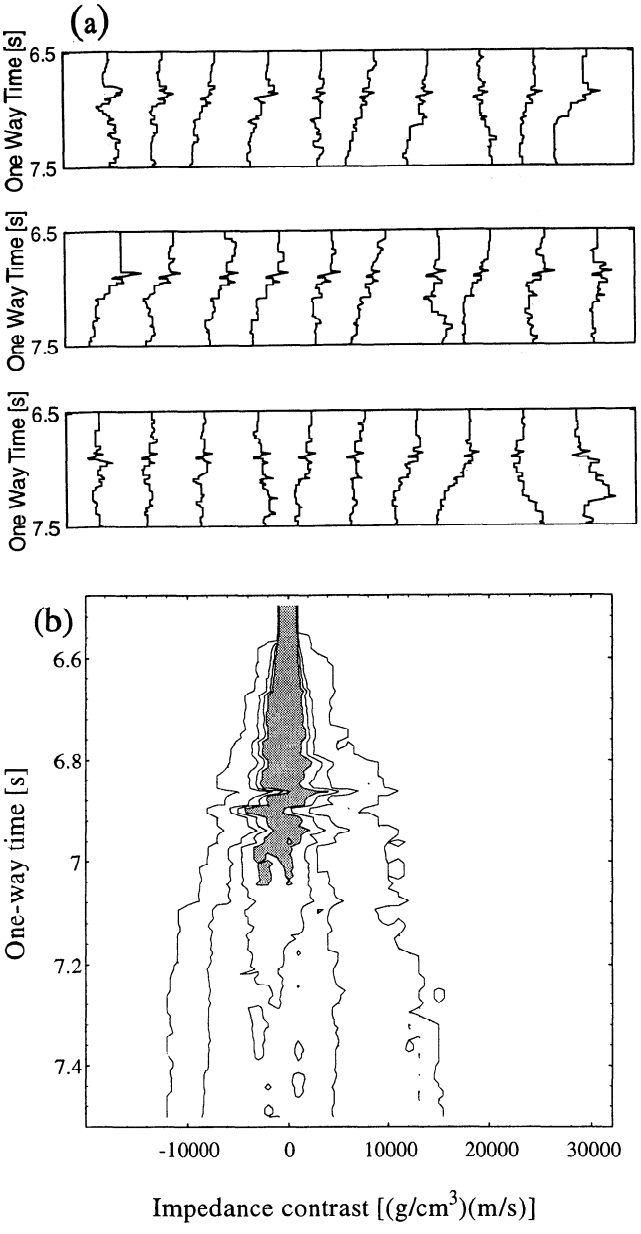


Figure 9. The results of the inversion on the W reflector data set. (a) A selection of a posteriori models. (b) Marginal a posteriori impedance contrast distributions for all one-way times between 6.5 and 7.5 s.

significantly larger than the corresponding probability of a positive polarity.

This model with an upper layer of negative impedance contrasts overlying a layer of more ambiguous but generally nonpositive impedance contrasts cannot be simply reconciled with previous estimates of velocities at a similar depth that were obtained from previous seismic surveys in the vicinity of this area [*Barton*, 1992; *Faber and Bamford*, 1979]. The LISP-B (Lithospheric Seismic Profile in Britain) seismic records showed high-amplitude P wave phases that arrived closely after first arrivals from rays diving into the uppermost mantle. These observed high-amplitude phases can be generated from post critical reflections off a layer with an increase in P wave impedance [*Barton*, 1992]. Alternatively, a zone of increased velocity several kilometers thick with no strong velocity gradients above but a steep negative gradient below can simultaneously explain the observed large amplitude P waves as well as low apparent velocity of these high-amplitude phases and the absence of any head waves [*Faber and Bamford*, 1979]. In either interpretation, the overall change in velocity over the depth range containing the W reflector will be a positive one. As previously noted, our analysis cannot resolve the overall change, but the high probability of a nonpositive contrast toward the top of the zone is closer to the original interpretation of the LISP-B data by *Faber and Bamford* [1979].

Discussion

Small-scale impedance contrasts modeled by our Monte Carlo analysis need not necessarily be consistent with broader velocity gradients sensed by previous refraction surveys. The two do seem to coincide at the Moho at $\sim$ 30-km depth, but not around the W reflector at $\sim$ 50-km depth. The model first proposed in the original interpretation of the LISP-B data [*Faber and Bamford*, 1979] might be appropriate: a gentle positive gradient underlain by a local, steep negative gradient in impedance above another gentle positive gradient. This in effect produces a layer of low impedance, about 1 km thick within a thick layer with overall high-impedance region which cannot be fully resolved by our method at present.

Other recent studies of velocity and impedance structure of the lithospheric layer associated regionally with the Flannan and W reflectors, although also non-definitive, do suggest that a positive gradient occurs at these depths [*Barton*, 1992; *Morgan et al.*, 1994]. Local negative steps in impedance or velocity cannot be excluded by any study and may help explain some features of the record sections. In some locations the analysis may have been made on a different mantle feature from that analyzed in this study. Distinctions have been made between the largely subhorizontal W reflector interpreted to have a Caledonian (400 Ma) or older origin [*Snyder and Flack*, 1990] and the dipping Flannan reflector associated with Permo-Triassic extension and basin formation [*Reston*, 1990].

A limited number of common mineral assemblages are available to produce the appropriate impedance contrasts at 50 km depths that are required by the observation (see *Snyder and Flack*, [1990]) for one recent compilation and references). Eclogite facies metamorphic rocks of gabbroic composition and peridotite are the most probable high-velocity, high-density rocks. Metasomatic deposits provide the most likely source of low-impedance material at these depths. Such deposits typically contain phlogopite mica and other exotic high-pressure minerals deposited by volatile-rich melts, as exemplified by the numerous mineral assemblages ob-

served in Kimberlite pipes in South Africa [*Schulze*, 1989]. Normal faults associated with extension and basin formation typically enhance impedance contrasts by juxtaposing high-impedance rocks below rocks of lower impedance. This provides one possible explanation for the Flannan reflector impedance models [*Morgan et al.*, 1994].

Thermotectonic processes associated with orogenies and stable cratons are more variable but can provide environments suitable for the introduction of negative impedance contrasts with increasing depth by trapping migrating melts or fluids [*Snyder and Flack*, 1990]. Here we give just two examples. In one possible model, the Moho velocity increase results from a phase transition from granulite to eclogite facies that has been steepened by subsequent horizontal shearing localized by this phase transition. The W reflector might then result from a diffuse transition from eclogitic rocks to peridotite, a compositional boundary. Another possible model produces the Moho reflections from a compositional change from a gabbroic granulite to peridotite. Very local negative impedance contrasts associated with the W reflector would then require an anomalous layer imbeded within the peridotite: perhaps a thin, phlogopite-rich metasomatic layer in contact with a mafic intrusive now in eclogite metamorphic facies. The most probable impedance distributions from our Monte Carlo analysis can be interpreted by making reasonable geological assumptions about metasomatic processes and metamorphic grade. That does not necessarily make them correct.

Conclusions

We have provided a strategy for determining polarity of complex reflectors from normal-incidence data. The strategy is based on the Bayesian inversion theory. It consists of two steps: (1) a priori model generator where any type and any number of a priori probability distributions can be used and (2) selection of the model based on fit between synthetic and observe data. We applied this analysis to intriguing features in the mantle beneath northern Scotland that will undoubtedly never be directly sampled to confirm our models and interpretations. Given the uncertainties of the inversion method and the small amount of suitable data that is available with sufficient signal-to-noise ratios to be useful, we feel that the results must first be assessed independently and demonstrated to be reliable and self-consistent. Similar analysis of wide-angle (postcritical) reflections could provide additional new constraints for example but would then only strengthen our present conclusions if the results compared favorably.

As the depth increases, the Moho contains largely positive impedance contrasts, and the W reflector is likely to display a negative or zero overall impedance contrast, at least in its upper levels. This latter conclusion is new and significant information about mantle reflectors. Candidate materials exist which produce both positive and negative impedance contrasts, but to confidently distinguish between the two is an important first step.

Appendix: The Influence of Wavelet Scaling Error on Computed Impedance Functions

For small reflection coefficients (which we have in our case, where $r_i \leq 0.1$), the acoustic impedance $I(T)$ is related to the reflectivity $r(T)$ through [*Peterson et al.*, 1955]

$$I(T) \approx I_0 \exp\left[\frac{2}{\Delta T}\int_0^T r(u)du\right] .$$

where ΔT is the sampling interval for $I(T)$ and $r(T)$ and I_0 is the impedance at the top ($T = 0$) of the considered interval. Let us now look at the acoustic impedance function $\hat{I}(T)$ we obtain if the wavelet is in error with a scale factor k. In that case the computed reflectivity $\hat{r}(T)$ is related to the real reflectivity $r(T)$ through $\hat{r}(T) = r(T)/k$. We get

$$\hat{I}(T) \approx I_0 \exp\left[\frac{2}{\Delta T}\int_0^T \frac{1}{k} r(u)du\right] .$$

The relationship between $\hat{I}(T)$ and $I(T)$ is now

$$\frac{\hat{I}(T)}{I_0} \approx \exp\left[\frac{2}{\Delta T}\int_0^T \frac{1}{k} r(u)du\right] \approx \left[\frac{I(T)}{I_0}\right]^{\frac{1}{k}} .$$

Since

$$x^{\frac{1}{k}} \approx 1 + \frac{1}{k}(x-1)$$

for x close to 1, we get the following simple, approximate relation between scaled and unscaled impedance:

$$\frac{\hat{I}(T)}{I_0} \approx 1 + \frac{1}{k}\left[\frac{\hat{I}(T)}{I_0} - 1\right]$$

or

$$k\Delta\hat{I} \approx \Delta I$$

where $\Delta I = I(T) - I_0$ is the impedance contrast over the considered interval.

Acknowledgments. The authors are indebted to Associate Editor John Scales for his valuable suggestions and criticism. K. Mosegaard has received financial support from BIRPS and from Danish Natural Science Research Council (SNF). H. Wagner has received financial support from SNF. BIRPS is funded by the Natural Environment Research Council and BIRPS Industrial Associates (Amerada-Hess Ldt., BP Exploration Co. Ldt., Chevron UK Ltd., Conoco (UK) Ltd., Lasmo North Sea Plc., Mobil North Sea Ltd., Shell UK Exploration and Production). This is University of Cambridge Department of Earth Sciences contribution 4757.

References

BABEL Working Group, Integrated seismic studies of the Baltic Shield using data in the Gulf of Bothnia region, *Geophys. J. Int.*, *112*, 305–324, 1993.

Barton, P., LISP-B revisited, A new look under the Caledonides of northern Britain, *Geophys. J. Int.*, *110*, 371–391, 1992.

Clowes, R. M., and A. G. Green (Eds.), Seismic reflection probing of the continents and their margins, *Tectonophysics*, *232*, 450 pp., 1994.

Clowes, R. M., F. A. Cook, A. G. Green, C. E. Keen, J. N. Ludden, J. A. Percival, G. M. Quinlan, and G. F. West, LITHOPROBE New perspectives of crustal evolution in Canada, *Can. J. Earth Sci.*, *29*, 1813–1864, 1992.

Faber, S., and D. Bamford, Lithospheric structural contrasts across the Caledonides of northern Britain, *Tectonophysics*, *56*, 17–30, 1979.

Futterman, W.I., Dispersive body waves, *J. Geophys. Res.*, *67*, 5279–5291, 1962.

Ganley, D.C., A method for calculating synthetic seismograms which include the effects of absorption and dispersion, *Geophysics*, *46*, 1100–1107, 1981.

Haskell, N. A., The dispersion of surface waves from point sources in a multilayered media, *Bull. Seismol. Soc. Am.*, *43*, 17–34, 1953.

Hobbs, R., and D. Snyder, Marine seismic sources used for deep seismic reflection profiling, *First Break*, *10*, 417–428, 1993.

Holbrook, W.S., W.D. Mooney, and N. I. Christensen, The seismic velocity structure of the deep continental crust, in *Continental Lower Crust, Dev. Geotectonics*, vol. 23, edited by D.M. Fountain, R. Arculus, and R.W. Kay, pp. 1-34, Elsevier, New York, 1992.

Jannane, M., et al., Wavelengths of structures that can be resolved from seismic reflection data, *Geophysics*, *54*, 906–910, 1989.

Klemperer, S. L., and R. Hobbs, *The BIRPS Atlas: Deep Seismic Reflection Profiles across the British Isles*, 124 pp., Cambridge Univ. Press, New York, 1991.

McGeary, S., and M. R. Warner, Seismic profiling of the continental lithosphere, *Nature*, *317*, 795–797, 1985.

Mooney, W.D., and R. Meissner, Multi-genetic origin of crustal reflectivity: a review of seismic reflection profiling of the continental lower crust and Moho, in *Continental Lower Crust, Dev. Geotectonics*, vol. 23, edited by D.M. Fountain, R. Arculus, and R.W. Kay, pp. 45-79, Elsevier, New York, 1992.

Morgan, J. V., M. Hadwin, M. R. Warner, P. J. Barton, and R. P. L. Morgan, The polarity of deep seismic reflections from the lithospheric mantle: Evidence for a relic subduction zone, *Tectonophysics*, *232*, 319–328, 1994.

Mosegaard, K., and A. Tarantola, Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res.*, *100*, 12,431-12,447, 1995.

Peterson, R.A., W.R. Fillipone, and F.B. Coker, The synthesis of seismograms from well log data, *Geophysics*, *20*, 516–538, 1955.

Reston, T. J., The lower crust and extension of the continental lithosphere: Kinematic analysis of BIRPS deep seismic data, *Tectonics*, *9*, 1235–1248, 1990.

Schulze, D. J., Constraints on the abundance of eclogite in the upper mantle, *J. Geophys. Res.*, *94*, 4205–4212, 1989.

Singh, S. C., and D. P. McKenzie, Layering in the lower crust, *Geophys. J. Int.*, *113*, 622–628, 1993.

Snyder, D., and C. A. Flack, A Caledonian age for reflectors within the mantle lithosphere north and west of Scotland, *Tectonics*, *9*, 903–922, 1990.

Stoffa, P. L., and M. K. Sen, Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane-wave seismograms, *Geophysics*, *56*, 1794–1810, 1991.

Tarantola, A., and B. Valette, Inverse problems = quest for information, *J. Geophys.*, *50*, 159–170, 1982.

Weibe, R. A., The Pleasant Bay layered gabbro-diorite, coastal Maine: Ponding and crystallization of basaltic injections into a silicic magma chamber, *J. Petrol.*, *34*, 461–489, 1993.

Zhao, W., K. D. Nelson, and Project INDEPTH Team, Deep seismic reflection evidence for continental underthrusting beneath southern Tibet, *Nature*, *366*, 557–559, 1993.

Klaus Mosegaard and Helle Wagner, Dept. of Geophysics, Niels Bohr Institute for Astronomy, Physics and Geophysics, Juliane Maries Vej 30, DK-2200 Copenhagen , Denmark. (e-mail: klaus@gfy.ku.dk; hw@miraculix.gfy.ku.dk)

Satish Singh and David Snyder, British Institutions Reflection Profiling Syndicate, Bullard Laboratories, University of Cambridge, Cambridge, England. (e-mail: singh@esc.cam.ac.uk; snyder@esc.cam.ac.uk)

(Received July 16, 1996; accepted August 13, 1996.)

Inverse Problems **14** (1998) 405–426. Printed in the UK PII: S0266-5611(98)92487-X

Resolution analysis of general inverse problems through inverse Monte Carlo sampling

Klaus Mosegaard

Department of Geophysics, Niels Bohr Institute for Astronomy, Physics and Geophysics, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Oe, Denmark

Received 13 March 1998

Abstract. The general inverse problem is characterized by at least one of the following two complications: (1) data can only be computed from the model by means of a numerical algorithm, and (2) the *a priori* model constraints can only be expressed via numerical algorithms. For linear problems and the so-called 'weakly nonlinear problems', which can be locally approximated by a linear problem, analytical methods can provide estimates of the best fitting model and measures of resolution (nonuniqueness and uncertainty of solutions). This is, however, not possible for general problems. The only way to proceed is to use sampling methods that collect information on the posterior probability density in the model space. One such method is the inverse Monte Carlo strategy for resolution analysis suggested by Mosegaard and Tarantola. This method allows sampling of the posterior probability density even in cases where prior information is only available as an algorithm that samples the prior probability density. Once a collection of models sampled according to the posterior is available, it is possible to estimate, not only posterior model parameter covariances, but also resolution measures that are more useful in many applications. For example, posterior probabilities of the existence of interesting Earth structures like discontinuities and flow patterns can be estimated. These extended possibilities for resolution analysis may also provide new insight into problems that are usually treated by means of analytical methods.

1. Introduction

1.1. Resolution analysis for the general inverse problem

In recent years, analysis of nonuniqueness and uncertainty of solutions to inverse problems (in the following called *resolution analysis*) has played an increasing role in astronomy, physics and geophysics. Resolution analysis of linear problems has been well established for many years (Backus 1970a, b, c), whereas methods for resolution analysis of nonlinearizable problems have been linked to recent developments in Monte Carlo techniques (see e.g. Cary and Chapman 1988, Pedersen and Knudsen 1990, Koren *et al* 1991, Mosegaard and Tarantola 1995, Gouveia and Scales 1997, Mosegaard *et al* 1997). Early examples of analysis of nonlinearizable inverse problems were mainly focused on the construction of best fitting models, but today it is widely acknowledged that uncertainty and nonuniqueness analysis is very important for the assessment of scientific conclusions based on inverse calculations.

Resolution analysis can be performed analytically for linear inverse problems and for problems that are well approximated by linearization. Linear relations between model parameters and data parameters and, for example, assumptions about Gaussian noise and a

0266-5611/98/030405+22$19.50

Gaussian *a priori* probability density result in analytical expressions for resolution indicators like, for instance, resolution kernels and *a posteriori* covariance matrices.

The situation is quite different for *the general inverse problem*, which we shall investigate in this paper. In the following we shall define a general inverse problem as an inverse problem that is characterized by at least one of the following two complications.

(1) Data can only be computed from the model by means of a numerical algorithm. In other words, the data parameters cannot be expressed by standard mathematical functions of the model parameters. This is a typical situation when data are calculated by 'simulation', that is, by numerically solving a set of partial differential equations with initial and boundary conditions. In these cases the coefficients of the equations are functions of the model parameters.

(2) The *a priori* model constraints can only be (or are most conveniently) expressed via a numerical algorithm that generates samples from an *a priori* probability density or evaluates this density in given models.

Notable examples of general inverse problems are the so-called *highly nonlinear inverse problems*, that is, problems for which a linearized analysis will fail, and (possibly linear) inverse problems with complicated *a priori* constraints derived from physical or geological information.

The fact that the model/data relationships and/or *a priori* model constraints are only defined through numerical procedures for a general inverse problem means that we are prevented from applying analytical techniques directly to such problems. This situation forces us to apply methods that rely on sampling of the model space. The safest method is *exhaustive sampling* where all points in a dense grid, covering the model space, are visited. This method can be recommended for small inverse problems, but for problems with many model parameters, or in cases where calculation of data from given model parameters is computer intensive, we need sampling methods that give useful results even with a modest number of samples from the model space. The methods we shall investigate in this paper belong to a family of Monte Carlo methods called *importance sampling methods*. These methods are based on a random sampling of the model space, where only models that are consistent with prior information and give a good data fit are sampled frequently. Compared with exhaustive sampling, a dramatic reduction in the number of necessary model samples is obtained, simply because importance sampling algorithms avoid sampling models that are inconsistent with data and prior information.

In this paper we shall adopt a Bayesian point of view on inverse problems. All statements about model parameters will be probabilistic, and the key to these statements is the *a posteriori* probability density in the model space summarizing all information about the model supplied by data, *a priori* information, and the model/data relationship. For the general inverse problem the shape of the posterior probability is unknown, and we may in general expect that it is poorly described by simple mathematical objects such as a mean model vector and a covariance matrix. A large collection of samples from the posterior probability density will be a more safe representation, and the Monte Carlo algorithm we shall describe is therefore designed to produce such a collection of models.

The important question is now how we extract useful information about model resolution from this collection of samples of the posterior. In view of the fact that we have nothing similar to a resolution kernel (which for a linear problem is derived directly from the known linear operator that connects model and data) and that the posterior covariance matrix may poorly describe uncertainties if the posterior deviates strongly from a Gaussian, we will look for alternative ways of describing the resolution.

Our starting point is that resolution analysis should help us to choose between different

interpretations of a given data set. We would like to answer complicated questions that address the correlations between several model parameters. Examples of such questions are:

- How likely is it (given seismic data and *a priori* information) that the Moon has a discontinuity (defined in a given way) in a certain depth interval?
- How likely is it (given data and *a priori* information) that the Earth model is 'cyclic' (defined by certain properties of the spectrum of the model) in a certain depth interval? This question is of interest in exploration for oil/gas reservoirs if one is looking for cyclic carbonate sequences.
- How likely is it (given data and *a priori* information) that a given 'cyclone' exists in the fluid flow at the core/mantle boundary below the Pacific Ocean? This question is of interest in studies of the secular variation of Earth's magnetic field.

In order to answer the above questions, and to calculate covariances and higher order moments, we need to evaluate integrals of the form

$$R(\mathcal{E}, f) = \int_{\mathcal{E}} f(\boldsymbol{m})\sigma(\boldsymbol{m})\,\mathrm{d}\boldsymbol{m} \tag{1}$$

where $\sigma(\boldsymbol{m})$ is the *a posteriori* probability density in the model space, $f(\boldsymbol{m})$ is a given function of the model parameters $\boldsymbol{m}$ and $\mathcal{E}$ is an event in the model space $\mathcal{M}$, containing the models we are interested in. For instance,

$$\mathcal{E} = \{\boldsymbol{m} | \text{ a given range of parameters in } \boldsymbol{m} \text{ is 'cyclic'}\}.$$

In the special case when $\mathcal{E} = \mathcal{M}$, and $f(\boldsymbol{m}) = m_i$, the resolution measure $R(\mathcal{E}, f)$ in equation (1) equals the mean $\langle m_i \rangle$ of the ith model parameter m_i. If $f(\boldsymbol{m}) = (m_i - \langle m_i \rangle)(m_j - \langle m_j \rangle)$, $R(\mathcal{E}, f)$ becomes the covariance between the ith and jth model parameters.

It is clear that in the general inverse problem we are prevented from analytical evaluation of the integral in (1) because we have no analytical expression for $\sigma(\boldsymbol{m})$. However, even for a perfectly linear inverse problem with known $\sigma(\boldsymbol{m})$ we may not be able to evaluate $R(\mathcal{E}, f)$ owing to the complexity of the subset $\mathcal{E}$. Hence, if we want to take advantage of our extended resolution measure, we need methods that allow us to evaluate $R(\mathcal{E}, f)$ in the general case.

One such method is the Monte Carlo algorithm we shall describe in the following. This algorithm will sample a large number of models $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$, according to $\sigma(\boldsymbol{m})$, after which our resolution measure $R(\mathcal{E}, f)$ can be approximated by the following simple average:

$$R(\mathcal{E}, f) \approx \sum_{\{n | \boldsymbol{m}_n \in \mathcal{E}\}} f(\boldsymbol{m}_n).$$

After a brief introduction to the concepts of 'Monte Carlo simulation' and 'inverse Monte Carlo sampling' and a presentation of the probabilistic (Bayesian) formulation of the general, implicit inverse problem, we shall describe the Monte Carlo strategy suggested by Mosegaard and Tarantola (1995). The paper concludes with an example, taken from Mosegaard *et al* (1997), where the method is applied to analysis of seismic reflection data form the Earth's crust and upper mantle.

2. Monte Carlo sampling

2.1. Forward sampling: Monte Carlo simulation

Given a parametrized system whose parameters fluctuate in time, we are interested in the behaviour of some 'data', given as a function of these model parameters. The model parameters may describe the positions and the momenta of particles in a container, and the function may describe the force (pressure) exerted on one of the container walls. The function is mathematically complicated, so we are unable to calculate the probability density for the force directly from the probability densities of the particle positions and momenta. Instead we can use a Monte Carlo algorithm to randomly generate particle positions and momenta according to their probability densities, and then apply the function to these model parameters to obtain randomly generated values of the force. This simple procedure will simulate the physical fluctuations in the force acting on the container wall, and a histogram of the obtained realizations of force values will approximate the probability density for the force.

2.2. Inverse Monte Carlo sampling: the Metropolis algorithm

Imagine that we are in the opposite situation to the one described above. We face the inverse problem where we observe the fluctuating force acting on the container wall, and we want to simulate the fluctuations in the positions and momenta of the particles. Since these particle parameters are not functions of the force on the container wall (there is no unique set of model parameters that corresponds to a given force) the Monte Carlo simulation method above is not applicable and must be replaced by inverse sampling.

The problem of inverse Monte Carlo sampling was solved by Metropolis *et al* (1953). Each iteration of their algorithm consisted of two interacting random steps. The first step was a Monte Carlo simulation step that simulated fluctuations in data generated by *uniformly* distributed model parameters. The second step was accepting or rejecting the model parameter set proposed by the Monte Carlo simulation step, using an acceptance probability derived from the data noise distribution and the model/data relationship. As we shall see later, the two steps of the Metropolis algorithm will, when combined, sample a probability distribution in the model space whose image in the data space (under the considered forward function) is the desired data distribution. This inverse Monte Carlo algorithm can be described as an adaptive procedure: it adjusts its sampling in the model space such that the corresponding sampling in the data space follows the desired probability distribution.

It should be noticed that there is an inherent nonuniqueness in this 'inverse sampling problem'. There exist in general infinitely many probability densities in the model space that map into the known probability density in the data space via our forward relationship. The Metropolis algorithm chooses the least informative of these densities, namely the one $h(\boldsymbol{m})$ that maximizes the entropy

$$\int_{\mathcal{M}} h(\boldsymbol{m}) \log(h(\boldsymbol{m})) \, \mathrm{d}\boldsymbol{m}$$

over the model space $\mathcal{M}$.

The original aim of the algorithm was to estimate thermodynamic averages of functions $f(\omega)$ over high-dimensional state spaces S:

$$\int_S f(\omega) P_{\mathcal{B}}(\omega) \, \mathrm{d}\omega \tag{2}$$

where $P_{\mathcal{B}}(\omega)$ is the Boltzmann distribution

$$P_{\mathcal{B}}(\omega) = \frac{1}{Z} \exp\left(-\frac{E(\omega)}{kT}\right)$$

describing fluctuations in the states of the statistical mechanical system. Here, $E(\omega)$ is the energy of state ω, T is the temperature, k is Boltzmann's constant, and Z is the normalization factor (the 'partition function'). In high-dimensional spaces (in statistical mechanical systems the dimension may be of the order of 10^{23}!) it would be virtually impossible to evaluate the integral (2) from function values in a regular grid of points or in a set of randomly selected grid points distributed uniformly over the state space. The reason is that at low values of the temperature T the Boltzmann distribution typically has localized, narrow peaks, and any practical grid (regular or uniformly random) would be rather coarse and therefore likely to miss these peaks.

The algorithm of Metropolis *et al* (1953) overcomes this problem by generating points in the state space according to the Boltzmann distribution. Sampling in this way (often referred to as 'importance sampling') has two advantages. First, that areas in the model space where the Boltzmann probability is large (and where the system is most likely to be) will be sampled more often than areas where the Boltzmann probability is small. Second, that evaluation of the integral (2) is reduced to a simple averaging:

$$\int_S f(\omega) P_{\mathcal{B}}(\omega)\, \mathrm{d}\omega \approx \frac{1}{N} \sum_{i=1}^{N} f(\omega_i)$$

where $\omega_1, \ldots, \omega_N$ are the states sampled by the algorithm. Their algorithm is the following.

In the ith iteration:

(1) generate a new state $\omega_{i,\text{trial}}$ with uniform probability in the neighbourhood of the current state $\omega_{i,\text{current}}$;

(2) accept the trial state $\omega_{i,\text{trial}}$ with probability

$$P_{\text{accept}} = \text{Min}\left(1, \frac{P_{\mathcal{B}}(\omega_{i,\text{trial}})}{P_{\mathcal{B}}(\omega_{i,\text{current}})}\right) \tag{3}$$

(3) if $\omega_{i,\text{trial}}$ is accepted, then $\omega_{i+1,\text{current}} = \omega_{i,\text{trial}}$, otherwise $\omega_{i+1,\text{current}} = \omega_{i,\text{current}}$.

The set of current states $\omega_{i,\text{current}}$ $(i = 1, \ldots, N)$ generated by this algorithm will, for increasing N, converge towards a set of samples from the Boltzmann distribution $P_{\mathcal{B}}(\omega)$. Obviously, the above algorithm requires a definition of the concept *neighbourhood*. We shall defer this definition, and a statement of the convergence theorem for the algorithm, to a later section where we discuss a wider class of algorithms containing the original Metropolis algorithm as a special case.

3. Probabilistic formulation of the inverse problem

3.1. The Bayesian formulation of the inverse problem

In a Bayesian formulation of the inverse problem, the state of information concerning the physical system after incorporation of both *a priori* information and data information is summarized by the *a posteriori* probability density $\sigma(\boldsymbol{m})$ over the model space (Tarantola and Valette 1982, Tarantola 1987).

Let $\boldsymbol{x}$ be the complete, ordered set of parameters describing the considered physical system. $\boldsymbol{x}$ consists of data and model parameters $\boldsymbol{x} = (\boldsymbol{d}, \boldsymbol{m})$, and its posterior probability

density is given by

$$\sigma(\boldsymbol{x}) = C\frac{\rho(\boldsymbol{x})\theta(\boldsymbol{x})}{\mu(\boldsymbol{x})} \tag{4}$$

(Tarantola and Valette 1982, Tarantola 1987), where C is a normalization constant, $\rho(\boldsymbol{x})$ is the *a priori* probability density for the system parameters, $\theta(\boldsymbol{x})$ is a probability density that contains information from the physical relations between parameters in $\boldsymbol{x}$, and $\mu(\boldsymbol{x})$ is a nonzero reference probability density, called the null information probability density by Tarantola and Valette (1982). The information gained by going from the reference state of information to a state of information given by the probability density $f(\boldsymbol{x})$ is given by

$$I(\mu \rightarrow f) = \int_{\mathcal{X}} f(\boldsymbol{x}) \log\left(\frac{f(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \tag{5}$$

where $\mathcal{X}$ is the parameter space of the system (Tarantola and Valette 1982).

3.2. The general inverse problem

Let us assume that the inverse problem is given by a relation between data and model parameters

$$\boldsymbol{d} = g(\boldsymbol{m}). \tag{6}$$

Besides (6) we have *a priori* information about data (a probability density in the data space, centred around the observed data), and *a priori* information about model parameters. Our problem will be to obtain information about the model parameters from the data.

The distribution $\theta(\boldsymbol{x})$ becomes:

$$\theta(\boldsymbol{x}) = \theta(\boldsymbol{d}|\boldsymbol{m})\mu_{\mathcal{M}}(\boldsymbol{m}) = \delta(\boldsymbol{d} - g(\boldsymbol{m}))\mu_{\mathcal{M}}(\boldsymbol{m}) \tag{7}$$

where δ is Dirac's delta function, and $\mu_{\mathcal{M}}(\boldsymbol{m})$ is the marginal reference probability density for $\boldsymbol{m}$. The above expression for $\theta(\boldsymbol{x})$ is based on the observation that the marginal distribution for the model vector must be the reference (null information) distribution, since a theoretical relationship contains no information on these parameters. If we further assume *a priori* independence between $\boldsymbol{d}$ and $\boldsymbol{m}$: $\mu(\boldsymbol{x}) = \mu_{\mathcal{D}}(\boldsymbol{d})\mu_{\mathcal{M}}(\boldsymbol{m})$ and $\rho(\boldsymbol{x}) = \rho_{\mathcal{D}}(\boldsymbol{d})\rho_{\mathcal{M}}(\boldsymbol{m})$, where $\mu_{\mathcal{D}}(\boldsymbol{d})$ is the marginal reference probability density for $\boldsymbol{d}$, and $\rho_{\mathcal{D}}(\boldsymbol{d})$ and $\rho_{\mathcal{M}}(\boldsymbol{m})$ are the marginal *a priori* probability densities for $\boldsymbol{d}$ and $\boldsymbol{m}$, respectively, we obtain

$$\sigma(\boldsymbol{x}) = C\frac{\rho_{\mathcal{D}}(\boldsymbol{d})\rho_{\mathcal{M}}(\boldsymbol{m})\delta(\boldsymbol{d} - g(\boldsymbol{m}))}{\mu_{\mathcal{D}}(\boldsymbol{d})}. \tag{8}$$

We obtain the desired marginal *a posteriori* distribution over $\boldsymbol{m}$ by integrating over $\boldsymbol{d}$. If we further describe our data in a reference frame where $\mu_{\mathcal{D}}(\boldsymbol{d}) = 1$, we obtain

$$\sigma(\boldsymbol{m}) = C\rho_{\mathcal{M}}(\boldsymbol{m}) \int_{\mathcal{D}} \rho_{\mathcal{D}}(\boldsymbol{d})\delta(\boldsymbol{d} - g(\boldsymbol{m}))\,\mathrm{d}\boldsymbol{d} = C\rho_{\mathcal{M}}(\boldsymbol{m})\rho_{\mathcal{D}}(g(\boldsymbol{m})) = \rho_{\mathcal{M}}(\boldsymbol{m})L(\boldsymbol{m}) \tag{9}$$

where

$$L(\boldsymbol{m}) = C\rho_{\mathcal{D}}(g(\boldsymbol{m})) \tag{10}$$

is the *likelihood function*.

The *a posteriori* probability density $\sigma(\boldsymbol{m})$ over the model space equals the *a priori* probability density $\rho_{\mathcal{M}}(\boldsymbol{m})$ times the likelihood function $L(\boldsymbol{m})$. The likelihood $L(\boldsymbol{m})$ measures the fit between the observed data and 'synthetic' data, calculated from the model vector $\boldsymbol{m}$.

The likelihood function is typically of the form

$$L(\boldsymbol{m}) = C \exp[-S(\boldsymbol{m})]$$

where C is a constant and $S(\boldsymbol{m})$ is a misfit function. $S(\boldsymbol{m})$ is a norm measuring the distance between observed data and 'synthetic' data, calculated from the model vector $\boldsymbol{m}$.

4. Monte Carlo analysis of general inverse problems

Several authors (Geman and Geman 1984, Rothman 1985, Pedersen and Knudsen 1990, Koren *et al* 1991) have suggested that the original Metropolis algorithm could be used to sample the *a posteriori* distribution for an inverse problem directly. The idea is to replace the state space point ω with the parameter vector $\boldsymbol{m}$ and the Boltzmann distribution $P_{\mathcal{B}}(\omega)$ with $\sigma(\boldsymbol{m})$ in the algorithm. In this way the algorithm will sample $\sigma(\boldsymbol{m})$, and the samples can then be used to evaluate the averages

$$R(\mathcal{E}, f) = \int_{\mathcal{E}} f(\boldsymbol{m})\sigma(\boldsymbol{m})\,\mathrm{d}\boldsymbol{m}$$

over the model space $\mathcal{M}$ that are necessary for a general resolution analysis.

This procedure is indeed possible in some cases, but only for problems where the $\sigma(\boldsymbol{m})$ has an explicit mathematical form. For problems with a complex body of *a priori* information, where no explicit mathematical expression or algorithm is available that allows direct calculation of values of the prior distribution $\rho_{\mathcal{M}}(\boldsymbol{m})$, the procedure cannot be used. The latter situation is more a rule than an exception in, for example, Earth science problems.

Of course, any inverse problem can be reduced to one with simple (or no) *a priori* information if such information is deliberately disregarded when not explicitly expressible by a mathematical formula. However, the consequence of disregarding prior information is often that the result obtained from the inversion turns out to be in conflict with requirements that we clearly know must be satisfied by the system we are studying. Examples are numerous: smooth models of subsurface structures that are clearly known to be 'blocky' or layered; models of fluid motion (in, for example, the ocean or in the Earth's core) that are in conflict with the fundamental laws of hydrodynamics; calculated Earth structure whose statistical properties are in conflict with observations from, for example, well data, etc. The obvious way to resolve these inconsistencies is to incorporate as much *a priori* information as possible into the inverse calculations.

The reader may wonder at this point how it is possible to incorporate *a priori* information into an inverse problem if we have neither an explicit mathematical expression for $\rho_{\mathcal{M}}(\boldsymbol{m})$ nor an algorithm that allows calculation of $\rho_{\mathcal{M}}(\boldsymbol{m})$ for a given $\boldsymbol{m}$. As we shall demonstrate in the following, it is possible to design a Monte Carlo method that incorporates *a priori* information even if the only thing available is a procedure that randomly generates parameter vectors $\boldsymbol{m}$ according to the prior $\rho_{\mathcal{M}}(\boldsymbol{m})$ (and an algorithm that allows computation of the likelihood $L(\boldsymbol{m})$ for a given model vector $\boldsymbol{m}$). Designing an algorithm that generates parameter vectors according to the prior is possible for a wide range of inverse problems where the prior information is a complicated mixture of hard constraints (e.g. inequalities) and soft (probabilistic) constraints derived from, for example, hard physical bounds on model parameters, histograms of previously observed values of similar parameters, physical laws that the model parameters must satisfy, geometrical constraints, etc. It is characteristic for such prior information that it is either impossible or very difficult or inelegant to calculate values of $\rho_{\mathcal{M}}(\boldsymbol{m})$ by an algorithm or a mathematical expression.

4.1. Algorithm design

In order to arrive at an algorithm with the above specified requirements, we shall introduce a few mathematical concepts that are useful for the description of any Monte Carlo sampling algorithm based on random walks. The first concept, the *metagraph*, describes aspects of the discretization of the problem as well as the 'rules' followed by the algorithm when sampling the parameter space. The second important concept is *microscopic reversibility*, a notion borrowed from statistical mechanics, describing the equilibrium attained by the algorithm after a large number of iterations. Our exposition will follow Mosegaard and Tarantola (1995), where the topic is treated in full detail.

4.2. The metagraph

We discretize our inverse problem in two steps. In the first step we choose the number of model parameters we shall use to represent the considered system. Since we want our parameters to be constrained only through the *a priori* probability density (and not through an arbitrary oversimplification of the model), we 'overparametrize' the system. In the second discretization of our problem we choose a finite, but dense, set of accessible values that our parameters can attain. In this way our model space becomes a finite space of points, where each point represents a surrounding region in the original, continuous space, small enough for the probability densities under consideration to be almost constant inside it. 'The probability of the point $\boldsymbol{m}_i$' can now be defined as the probability of the region surrounding $\boldsymbol{m}_i$. In practice our computer system will perform this discretization for us through its representation of real numbers by a finite number of bits. In the following, the symbol $\mathcal{M}$ stands for the discretized model space.

We are now in a position where we can define a graph structure, the *metagraph*, describing the possible paths followed by the algorithm when sampling the parameter space. The points (parameter vectors) $\boldsymbol{m}_i$ in $\mathcal{M}$ are the nodes of the metagraph, and when two points $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ in the metagraph are *directly connected* it is permitted for the algorithm to go from $\boldsymbol{m}_i$ to $\boldsymbol{m}_j$ or from $\boldsymbol{m}_j$ to $\boldsymbol{m}_i$ in one move. In other words, we assume that the algorithm is reversible (moves in both directions are permitted). The *neighbourhood* of $\boldsymbol{m}_i$ is defined as the set of points that are directly connected to $\boldsymbol{m}_i$ (figure 1). We shall assume that it is possible to go from any point $\boldsymbol{m}_j$ to itself in one move, that is, the neighbourhood of $\boldsymbol{m}_j$ contains $\boldsymbol{m}_j$ itself.

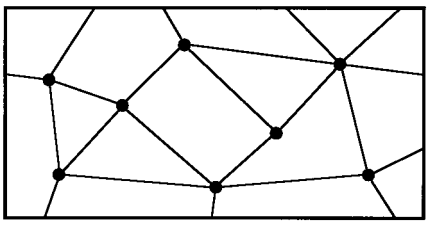


Figure 1. A section of the metagraph in the model space. The black dots indicate models in the discretized model space, and the lines define the possible moves of the random walk. Probabilistic rules allow it to jump from one model to another, if they are directly connected in the metagraph. The random walker will, in the limit of infinitely many iterations, have some probability p_i of being at point i. The neighbourhood of a model is defined as the models to which a random walker can go in one step if it starts at the given model.

4.3. Transition probabilities

To complete the description of the random walk we need only one more concept: the *transition probability* p_{ij}. Given that the random walk has arrived at $\boldsymbol{m}_j$, we define p_{ij} as the conditional probability that it will go to $\boldsymbol{m}_i$ in the next iteration.

4.4. Equilibrium distribution

Given that certain conditions are satisfied by the metagraph and the transition probabilities (see below), a random walk will, as the number of iterations go to infinity, visit the points $\boldsymbol{m}_i$ with a limit probability p_i, the *equilibrium probability*. In other words, if the probability is p_i that the random walker visits $\boldsymbol{m}_i$ after n iterations, it will also be p_i after $n+1$ iterations. If equilibrium for the random walk is attained at a distribution p_i, the random walk *samples* this distribution. In the following, our aim will be to design random walks that sample a particular distribution, namely the discretized version of the *a posteriori* distribution (9)

$$\sigma_i = \rho_i L_i. \tag{11}$$

It can be shown (Feller 1970) that if the metagraph for a random walk is a completely connected graph (it is possible for a random walker to go from any node $\boldsymbol{m}_j$ to any other node $\boldsymbol{m}_i$) of the kind described above, the random walk has a unique equilibrium distribution.

4.5. Choosing the transition probabilities: microscopic reversibility

Given a model space $\mathcal{M}$, there exist infinitely many algorithms (metagraphs, and corresponding sets of transition probabilities) that have σ_i as a unique equilibrium distribution. Our aim is to find one of these algorithms. Let us consider a situation where equilibrium at σ_i is already reached. What shall our algorithm satisfy if we want equilibrium to be maintained?

The probability that a transition takes place from point $\boldsymbol{m}_j$ to $\boldsymbol{m}_i$ is equal to the product of the probability σ_j that the random walker is in $\boldsymbol{m}_j$ before the transition, and the transition probability p_{ij}. Therefore, the probability that point $\boldsymbol{m}_i$ is reached in a given iteration is

$$\sum_j \sigma_j p_{ij} \tag{12}$$

where the summation is over $\boldsymbol{m}_i$'s neighbourhood. Similarly, the probability that the walker at a given time performs a move starting at point $\boldsymbol{m}_i$ in a given iteration is equal to the probability that the walker arrived at $\boldsymbol{m}_i$ after the previous iteration:

$$\sigma_i = \sigma_i \sum_j p_{ji} = \sum_j \sigma_i p_{ji} \tag{13}$$

again summed over $\boldsymbol{m}_i$'s neighbourhood. Equilibrium means that the two probabilities (12) and (13) are equal, and this can be accomplished in infinitely many ways. However, the simplest way is to choose the transition probabilities to impose *microscopic reversibility*, that is, to choose p_{ij} such that the probability that a transition takes place from point $\boldsymbol{m}_j$ to $\boldsymbol{m}_i$ is equal to the probability that a transition takes place from point $\boldsymbol{m}_i$ to $\boldsymbol{m}_j$:

$$\sigma_j p_{ij} = \sigma_i p_{ji} \tag{14}$$

or, equivalently,

$$\rho_j L_j p_{ij} = \rho_i L_i p_{ji}. \tag{15}$$

This can be done by choosing p_{ij} proportional to $\rho_i L_i$. Since we are not able to calculate the *a priori* probabilities ρ_j but are left with an algorithm that is capable of generating points $\boldsymbol{m}_i$ with probability ρ_i through a random walk on the metagraph, we must realize a transition probability p_{ij} proportional to $\rho_i L_i$ by performing iteration n in the following way.

(1) Choose a trial point $\boldsymbol{m}_{\text{trial}}^{(n+1)}$ from the neighbourhood of the current point $\boldsymbol{m}_{\text{current}}^{(n)}$ by running one step of the random walk that samples ρ_i.

(2) Accept $\boldsymbol{m}_{\text{trial}}^{(n+1)}$ with probability

$$p_{\text{accept}} = \min\left(1, \frac{L_{\text{trial}}}{L_{\text{current}}}\right). \tag{16}$$

(3) If $\boldsymbol{m}_{\text{trial}}^{(n+1)}$ is accepted, then $\boldsymbol{m}_{\text{current}}^{(n+1)} = \boldsymbol{m}_{\text{trial}}^{(n+1)}$, otherwise $\boldsymbol{m}_{\text{current}}^{(n+1)} = \boldsymbol{m}_{\text{current}}^{(n)}$.

Since the random walk that samples ρ_i will choose the trial point $\boldsymbol{m}_{\text{trial}}^{(n+1)}$ with a probability proportional to its prior probability, the above algorithm has transition probabilities

$$p_{ij} = C\rho_i \text{Min}\left(1, \frac{L_{\text{trial}}}{L_{\text{current}}}\right) \tag{17}$$

where C is a constant, and therefore satisfies (15). Consequently, it preserves equilibrium sampling at the desired distribution σ_i, once it is established.

The remaining issue is, of course, how we establish equilibrium sampling from a given initial state. Fortunately, a fundamental theorem from the theory of Markov processes states that, for the class of random walks we have chosen here, the sampling distribution will, in the limit of infinitely many iterations, converge (in probability) towards the unique equilibrium distribution for the random walk, independent of the initial state (Feller 1970).

4.6. Speed of convergence

The convergence theorem cited above guarantees only that a convergence to the sampling distribution will take place. Unfortunately, a way of predicting the speed of convergence, given the metagraph and a set of transition probabilities, has not been found. The transition probabilities given in (17) can be shown to be the most efficient in the family of transition probabilities satisfying (15) (Mosegaard and Tarantola 1995), but the most important factor influencing the speed of convergence, the metagraph structure, remains poorly understood.

There exist, however, a few practical ways of testing the performance of a given Monte Carlo algorithm of the above described type.

(1) It is a necessary (but not sufficient) condition for convergence that the algorithm samples a distribution proportional to $\rho_{\mathcal{D}}(\boldsymbol{d})$ in the data space $\mathcal{D}$. One should not rely on the output parameter vectors $\boldsymbol{m}^{(n)}$ before this happens. This can be controlled by observing the likelihood values $L(\boldsymbol{m}^{(n)})$ of the parameter vectors $\boldsymbol{m}^{(n)}$ generated by the algorithm. For instance, in an explicit inverse problem $\boldsymbol{d} = g(\boldsymbol{m})$ with N data, where data uncertainties are independent and Gaussian with covariance matrix $\boldsymbol{C}_d$, the likelihood function is

$$L(\boldsymbol{m}) = \text{constant}\ \exp(-\tfrac{1}{2}(\boldsymbol{d} - g(\boldsymbol{m}))^T \boldsymbol{C}_d^{-1}(\boldsymbol{d} - g(\boldsymbol{m}))). \tag{18}$$

If we have parametrized our problem such that it is purely underdetermined, $\rho_{\mathcal{D}}(\boldsymbol{d})$ is sampled correctly when $\log L(\boldsymbol{m}^{(n)})$ samples the χ^2-distribution with N degrees of freedom. In this case $\log L(\boldsymbol{m}^{(n)})$ will typically be close to $-N/2$.

(2) In order to speed up the convergence of the algorithm, it is often advisable to adjust the size of the neighbourhoods, the 'maximum step length', in the space $\mathcal{M}$. The maximum

step length must be large, so that the algorithm has a chance of exploring $\mathcal{M}$ efficiently. However, the longer the step length, the more moves are rejected. Once an algorithm with large step lengths has reached an area in $\mathcal{M}$ with high values of σ_i, its next move is likely to lead to an area with low values of σ_i, and hence a rejection. The trick is to limit the 'step length' of the algorithm such that most steps away from the current point lead to points whose probability is not too different from that of the current point. In this way many different points will be accepted, and much information about the structure of the probability distribution will be gathered. On the other hand, small steps create another problem, namely that the algorithm may get trapped for a long time in local probability maxima. Practically, it is therefore advisable to study the acceptance rate as a function of step length, and to choose the largest possible step length that maintains a high acceptance rate.

5. Example: inversion of near-vertical reflection data

The following example of inverse Monte Carlo analysis is taken from Mosegaard *et al* (1997), where further details can be found. We consider the inverse problem of generating Earth reflectivity models consistent with geological prior information and near-vertical seismic reflection data from the marine seismic reflection profile Deep Reflections from the Upper Mantle (DRUM), north of Scotland (figure 2(*a*)). The data consist of 10 seismic records from a single explosion, and we will focus on the data recorded in the interval between 8.0 and 10.0 s after the shot. This part of the data (figures 2(*b*)) contains information about the Moho, the transition zone at the base of the Earth's crust.

The main purpose of the following analysis is to answer the following questions, given the seismic data and the *a priori* information.

(1) How likely is it that the strong Moho reflection is due to a cyclic layered sequence?

(2) What is the polarity (sign) of the overall impedance contrast across the Moho?

The answers to these questions are important to the theory of the Earth's lithosphere. There is a general consensus among specialists that the overall impedance contrast across the Moho is positive and that strong Moho reflections are due to cyclic layering, but to what extent are these conclusions backed up by seismic reflection data and quantitative geological information? In the following we shall attempt to give a quantitative answer to this question.

5.1. Model parametrization and calculation of synthetic seismograms

The Earth is approximately horizontally stratified in the considered area, and the model can therefore be parametrized as reflection coefficients specified at discrete, equidistant depths. A model consists of reflection coefficients defined at 128 depths separated 0.008 s, measured as vertical one-way travel time for seismic waves. Since the relation between model and data is only moderately nonlinear in this parametrization, and since the data uncertainties are Gaussian, the shape of $L(\boldsymbol{m})$ is close to a Gaussian. However, the complexity of the *a priori* information means that $\rho(\boldsymbol{m})$ and hence $\sigma(\boldsymbol{m})$ deviates strongly from a Gaussian distribution.

Since inverse Monte Carlo sampling was used for resolution analysis, a fast computation of synthetic seismograms was required in order to be efficient. Synthetic seismograms were computed using a propagator matrix method (Haskell 1953, Ganley 1981), which solves the wave equation exactly in each layer. All multiple reflections, absorption, and dispersion within the target zones were incorporated. Dispersion was derived from Q, using

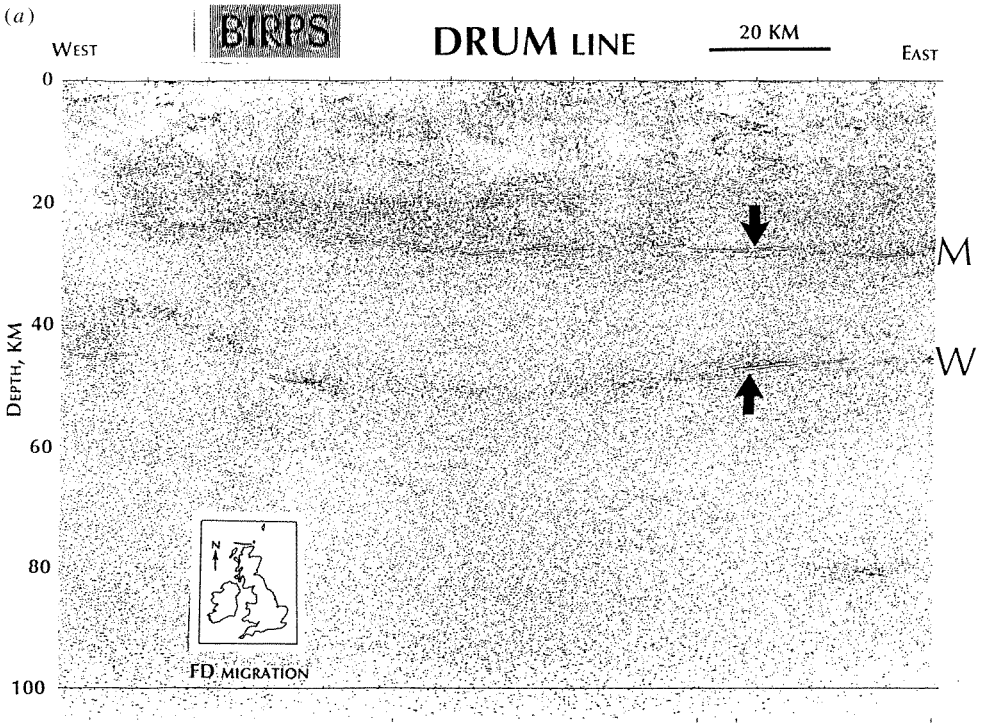


Figure 2. (*a*) The DRUM seismic section showing the Moho (M) and a deep mantle reflector, the W reflector (W). The location of data used in the Monte Carlo study of Mosegaard *et al* (1997) is indicated by arrows. The section was migrated using a two-dimensional velocity field derived from nearby refraction results (Snyder and Flack 1990) and then depth converted using the same velocities. (*b*) The data considered in this paper, covering the time range of 8–10 s that contains the Moho reflection at about 8.9 s. The estimated signal-to-noise (amplitude) ratio of this data set is 1.9. (*c*) The source wavelet used in the inversion. This trace represents a simulated far-field source signature for the entire air gun array provided by the contractor (GECO), convolved with the receiver ghost, recording filter, and reverberations within the 55 m thick water layer. The wavelet was then propagated to the appropriate two-way travel time using a quality factor of 500 to approximate effects of attenuation. From Mosegaard *et al* (1997).

a dispersion model given by Futterman (1962). The seismic source wavelet (source pulse) was estimated from field measurements and a knowledge of water depth and source/receiver depths (figure 2(*c*)).

To facilitate the visual interpretation of the generated *a posteriori* reflectivity models, they were converted to seismic impedance models (the product of mass density and seismic wave propagation velocity) through the recursive relation

$$I_{n+1} = I_n \frac{1 + r_n}{1 - r_n} \tag{19}$$

where I_n is the acoustic impedance in the nth layer, and r_n is the reflection coefficient at the base of the nth layer. I_0 was put equal to 19 000 (m s^{-1}) (g cm^{-3}), a realistic value in the crust near Moho level.

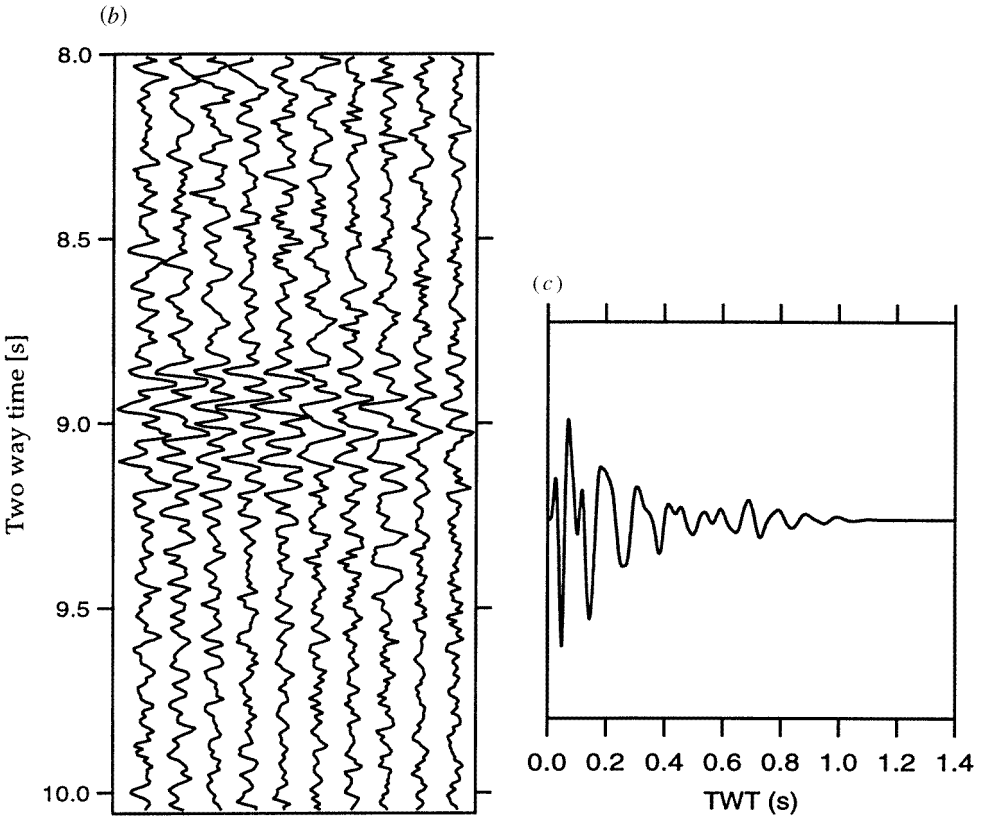


Figure 2. (Continued)

5.2. Data uncertainties and the likelihood function $L(\boldsymbol{m})$

The considered data set consists of 10 seismic records of 2.048 s length (figure 2(*b*)). Because coherent events in the data set are approximately horizontal and laterally invariant, the incoherent noise was estimated by first finding one of the best fitting horizontally stratified models for the data set, by repeating a steepest decent algorithm many times with different starting models. The (10 identical) vertical incidence seismic records computed from this model were then subtracted from the 10 observed data records to form 10 noise records. These noise records were approximately stationary signals, and the distribution of noise values was very close to a Gaussian. Since the data noise has a Gaussian distribution, we define the likelihood function as

$$L(\boldsymbol{m}) = \exp\{-\tfrac{1}{2}[\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}]^T \boldsymbol{C}_d^{-1}[\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{\text{obs}}]\} \tag{20}$$

where $\boldsymbol{C}_d$ is the covariance matrix for the noise. For the considered data we found a signal to noise (amplitude) ratio of 1.9. The estimated marginal distribution of a single noise value, obtained from the histogram of amplitudes from all points in the noise record traces, is shown in figure 3(*a*), and the estimated noise autocorrelation (normalized to one at zero lag) is shown in figure 3(*b*).

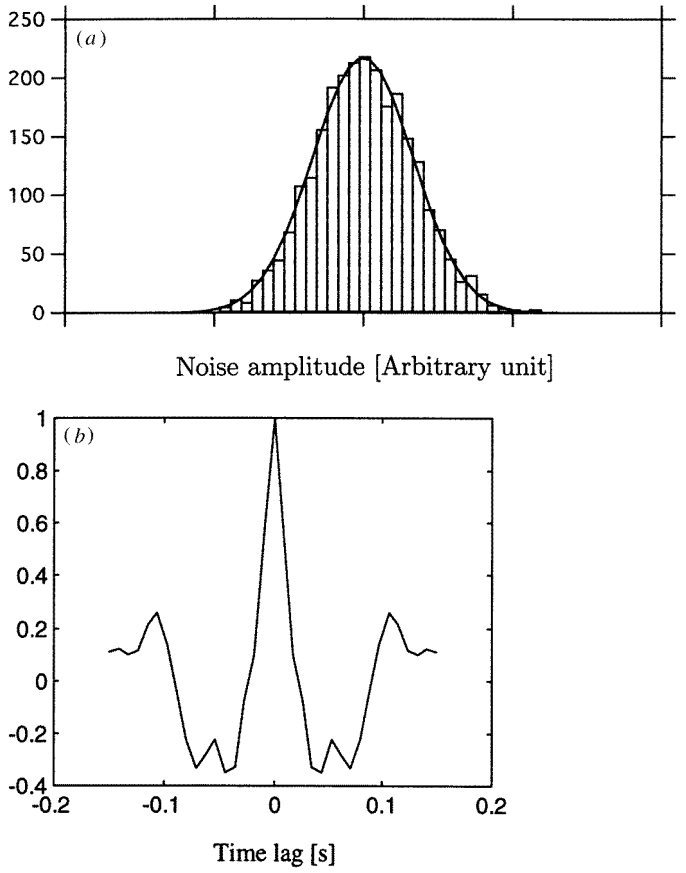


Figure 3. (*a*) Estimated marginal noise distribution for a single data sample. The full curve is a Gaussian distribution. (*b*) Estimated noise autocorrelation. From Mosegaard *et al* (1997).

5.3. A priori information

To obtain appropriate *a priori* information on the structure of deep reflections, possible geological processes which might be responsible for these reflectors were considered. Among a number of possible geological hypotheses the following was chosen as the 'prior hypothesis': layered igneous intrusions in a large magma chamber formed by fractional crystallization of magma derived from the mantle (Weibe 1993). Under this hypothesis, physical properties of rocks found in the igneous intrusive complexes of Rum, an island west of Scotland, and Great Dyke in Zimbabwe formed the basis of the quantitative prior information (Singh and McKenzie 1993). Only that part of the information from the igneous complexes that is expected to be fairly independent of the depth of burial was used, that is, reflection coefficients and layer thicknesses. Any other plausible geological model could have been used as prior information, but the igneous intrusions have been more thoroughly mapped on the surface and relevant statistics are available with greater detail than for other geological models.

Histograms (figure 4) show the distributions of reflection coefficients and layer thicknesses for the intrusive complexes that guided our *a priori* model generator. The thicknesses of the layers have been converted from metres into seconds (seismic wave

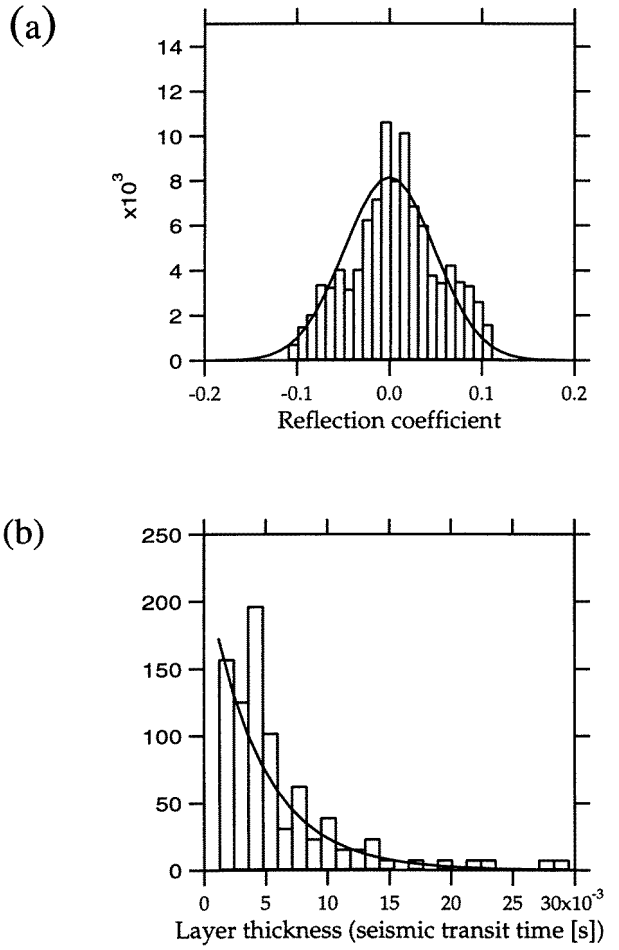


Figure 4. *A priori* information used by the *a priori* model generator. (*a*) Reflection coefficients distribution obtained from studies of the igneous intrusions of Rum, Scotland, and the Great Dyke, Zimbabwe (Singh and McKenzie 1993), that are consistent with other studies such as those of the Pleasant Bay layered gabbro-diorite (Weibe 1993). The full curve has a Gaussian distribution with standard deviation $\sigma = 0.047$. (*b*) Layer thickness distribution as a function of one-way travel time derived from observations of the Rum and Great Dyke intrusions. The full curve has an exponential distribution with $\lambda = 225.0\ \mathrm{s}^{-1}$. From Mosegaard *et al* (1997).

transmission time) using the velocity in each layer (Singh and McKenzie 1993). In order to retain only the general features of the reflectivity and thickness histograms, they were approximated with a Gaussian and an exponential distribution, respectively. In this way we avoided detailed histogram structure that is only characteristic of the specific igneous intrusion complexes considered in this study. All *a priori* models generated by our algorithm are consistent with the approximate distributions; that is, histograms of reflection coefficients and layer thicknesses produced from these models are statistically equivalent to corresponding histograms from the igneous complexes. The combined *a priori* distribution used here allows models with reflection coefficients ranging from about -0.1 to $+0.1$ and thicknesses from about 10 to 2000 m.

An '*a priori* model generator' algorithm was constructed to generate models consistent

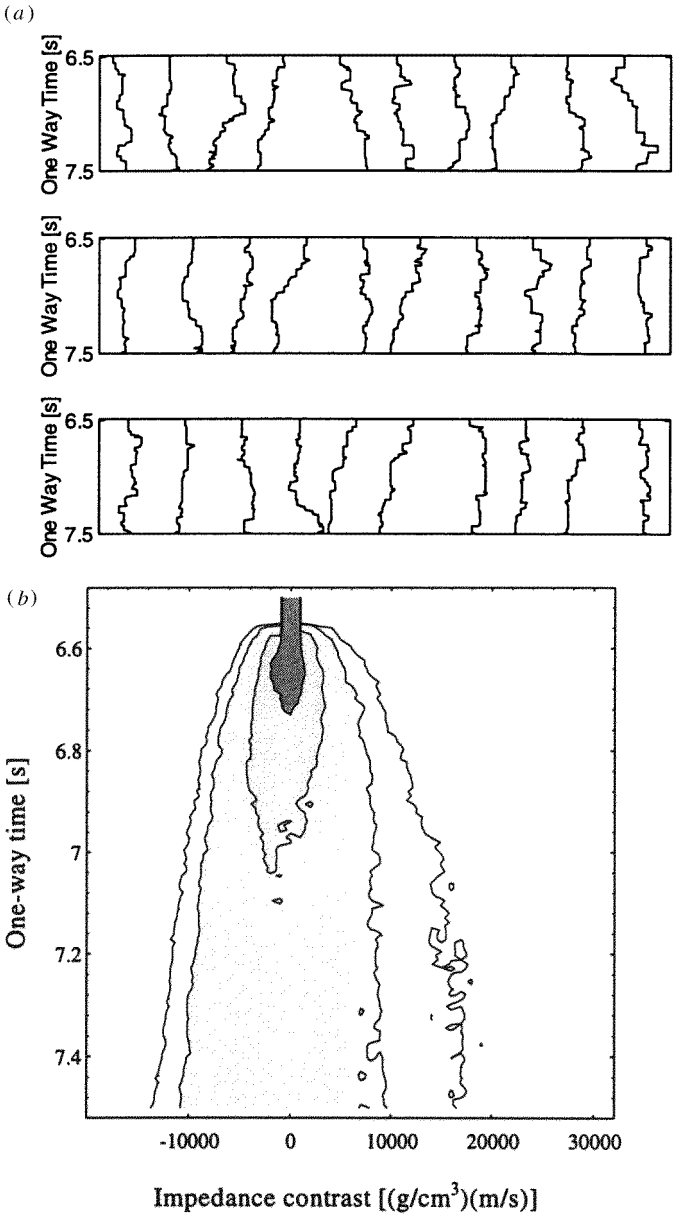


Figure 5. (*a*) A selection of typical *a priori* models generated assuming the distributions shown in figure 4. (*b*) Marginal *a priori* impedance contrast distributions for all one-way times in a 1 s interval. From Mosegaard *et al* (1997).

with the statistics derived from figure 4 for the igneous intrusions. Figure 5 shows a selection of *a priori* impedance models produced by this algorithm. All the *a priori* models are different, but they all have identical statistics. These impedance models have been calculated from their corresponding reflectivity models and, for comparison, all models have the same impedance value, 19 000 m s^{-1} (g cm^{-3}), at the top. One step of the *a priori* model generator perturbed only one reflection coefficient in the model. A perturbation step

consisted of the following two substeps.

For the selected depth, perturb the reflection coefficient r as follows.

(1) Decide if r shall be zero (probability λ) or nonzero (probability $1-\lambda$). In this way, the layer thickness distribution will be exponential

$$\lambda \exp(-\lambda \cdot \Delta t)$$

where Δt is the layer thickness measured in units of the reflectivity sampling distance. The exponential distribution fitted to the layer thickness histogram is shown in figure 4(b). The mean layer thickness is 4.4×10^{-3} s.

(2) If r is chosen to be nonzero, generate its value from a Gaussian with zero mean and standard deviation $\sigma = 0.047$, as suggested by figure 4(a).

The first substep above takes advantage of the fact that the layer thickness distribution is assumed to be exponential. If we had chosen any other distribution $f(t)$ for the layer thicknesses (for instance, we could have obtained $f(t)$ by normalizing the histogram shown in figure 4(b)), we could proceed as follows.

Choose a large positive number N representing the maximum number of layer interfaces in the considered interval. In our case we could put $N = 128$, the number of values representing our discretized reflectivity. The following algorithm considers a stack of layers with N layer interfaces located at one-way times $t_1, t_2, \ldots, t_N$, and in each step it chooses one of the layer interfaces (say, the nth) and perturbs its position within the bounds given by the nearest layer interfaces above and below.

Perturbation of the nth layer interface with initial position t_n consists of the following two substeps.

(1a) If $n = 1$, choose a possible new position t_n' uniformly at random in the interval $[0, t_2]$, but accept it only with probability

$$p_{\text{accept}} = \min\left(1, \frac{f(t_2 - t_1')\, f(t_1')}{f(t_2 - t_1)\, f(t_1)}\right).$$

If the new position is accepted, t_n' replaces t_n as the depth of the nth layer interface. Otherwise t_n is retained.

(1b) If $1 < n < N$, choose a possible new position t_n' uniformly at random in the interval $[t_{n-1}, t_{n+1}]$, but accept it only with probability

$$p_{\text{accept}} = \min\left(1, \frac{f(t_{n+1} - t_n')\, f(t_n' - t_{n-1})}{f(t_{n+1} - t_n)\, f(t_n - t_{n-1})}\right).$$

If the new position is accepted, t_n' replaces t_n as the depth of the nth layer interface. Otherwise t_n is retained.

(1c) If $n = N$, choose a new position t_N' by adding to t_{N-1} a number generated according to the distribution $f(t)$.

(2) Generate a new reflection coefficient for the nth layer interface from a Gaussian with zero mean and standard deviation $\sigma = 0.047$, as suggested by figure 4(a).

The above procedure generates reflectivity models of length t_N, but only the upper one-second interval (one-way time) is used as the perturbed model in later calculations.

Using arguments similar to those presented earlier in the section on algorithm design it can be demonstrated that this algorithm generates reflectivity models with independent, identically distributed layer thicknesses with probability distribution $f(t)$. The distribution of reflection coefficients for the layer interfaces is of course Gaussian.

5.4. Sampling the a posteriori distribution

100 000 iterations of the inverse Monte Carlo sampling were run on the data set. In each iteration, all 128 reflection coefficients were considered by the algorithm, and each of the reflection coefficients was perturbed by running one step of the *a priori* model generator. The second part of the Monte Carlo algorithm, accepting or rejecting models proposed by the *a priori* model generator according to the rule given in (16), used the likelihood function defined by (20) with a noise variance corresponding to a signal-to-noise ratio of 1.9. The series of 'current models' produced by this algorithm are samples from the *a posteriori* distribution $\sigma(\boldsymbol{m})$. The strategy of perturbing only one reflection coefficient at a time preserves most characteristics of the current model, which may have fitted the data well. This strategy is efficient, since we wish to visit many models with a good data fit, but it provides models that are successively correlated. This is a problem since error and resolution analysis requires a collection of statistically independent models from the *a posteriori* distribution. A smaller set of models chosen from among the accepted models in such a way that they are sufficiently separated in time (iterations) constitutes such a set of independent models. Here we chose to save only every 100th model. This waiting time of 100 iterations between accepted models was found by analysing the fluctuations of $L(\boldsymbol{m})$ as the iterations proceeded. Inspection of the autocorrelation function for these fluctuations showed that accepted models separated by 100 iterations were unlikely to be correlated. Thus we had 1000 independent samples of the *a posteriori* distribution from 100 000 sampled models.

The analysis resulted in the *a posteriori* models shown in figures 6(*a*) and (*b*). Figure 6(*a*) shows a selection of *a posteriori* models, randomly chosen from the 1000 *a posteriori* models generated by the algorithm. Figure 6(*b*) shows estimated marginal *a posteriori* distributions for impedance contrasts (impedance functions generated from (19) under the assumption that the impedance at the top of the model is zero) at all depths in the considered interval.

In figure 7, the distribution of synthetic data corresponding to the posterior models is shown. All the synthetic traces are statistically indistinguishable from the observed seismic records, in that their deviations from these records are within the noise level.

The variations in the model that are obtained by the Monte Carlo inversion, and hence permitted by the *a priori* information and the data, are evident. To interpret the output models correctly, it should be remembered that the method is designed to produce particular model features with a frequency proportional to their *a posteriori* probability. This means that the probability of a feature that exists in the impedance model is roughly proportional to the number of times the feature occurs in the collection of output models (and in figure 6(*a*)). If it appears on almost all the *a posteriori* models, it is well resolved.

We are now ready to answer the two questions about the Moho, posed in the introduction to this numerical example. Concerning the first question, it is clear from figure 6(*a*) that the cyclic, layered sequence (the Moho) having a thickness of about 0.3 ± 0.04 s one-way time, near the middle of the considered depth interval, is well resolved. The second question concerns the overall impedance contrast, that is, the increase in impedance from the top to the bottom of the zone. In figure 6(*b*), the marginal *a posteriori* distributions for impedance contrasts at all the considered depths are shown. It is clear from this figure that the magnitude of the overall change in impedance is poorly resolved. However, the polarity (sign) of the overall impedance contrast is well

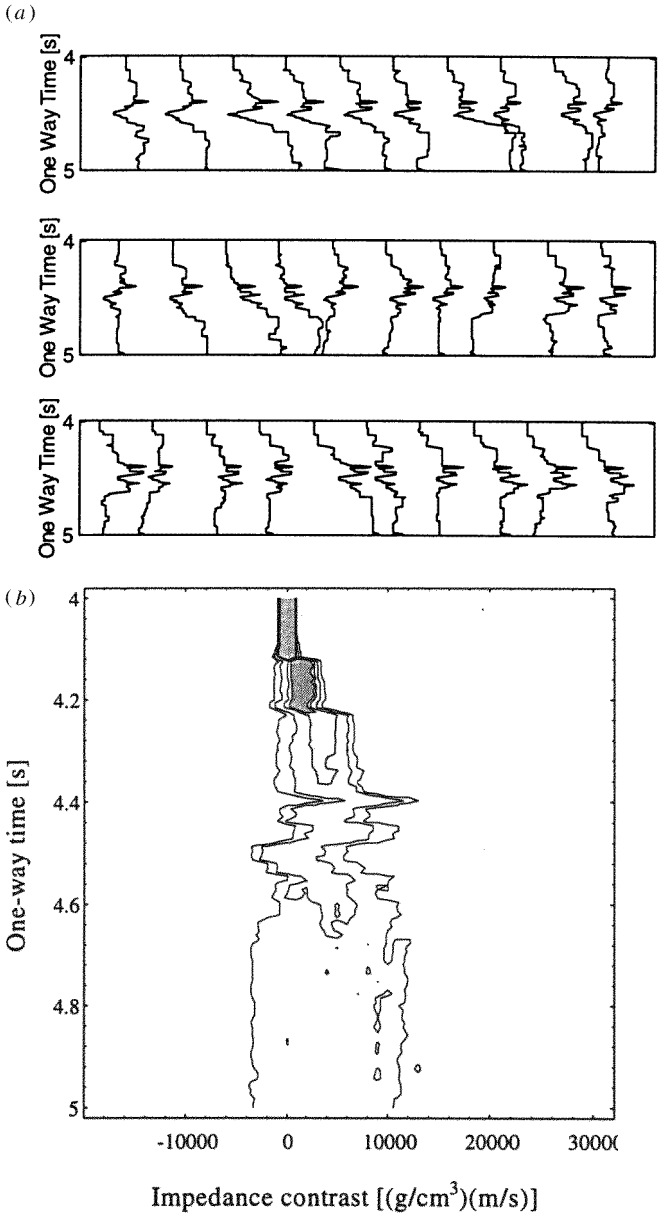


Figure 6. The results of the inversion on the data set. (*a*) A selection of *a posteriori* models. (*b*) Marginal *a posteriori* impedance contrast distributions for all one-way times between 4.0 and 5.0 s. From Mosegaard *et al* (1997).

resolved. As seen from the impedance contrast distribution at 5.0 s one-way time, it has a very high probability of being positive, and this is consistent with previous refraction results that used diving rays and wide-angle reflections to model observed phases (Barton 1992).

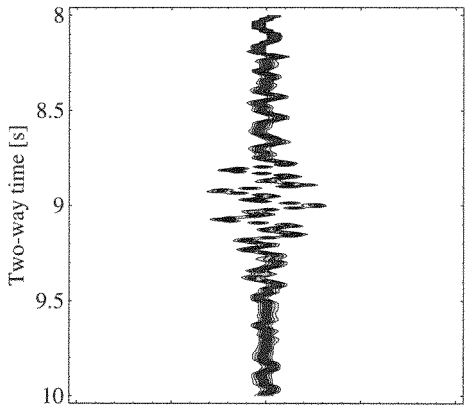


Figure 7. Marginal *a posteriori* data distributions for all two-way times between 8.0 and 10.0 s.

6. Discussion

Inverse Monte Carlo sampling permits resolution analysis of general inverse problems, since it is independent of explicit mathematical formulations of prior information and data/model relationship. Usually, however, there is a price to pay: despite the fact that importance sampling techniques are designed to give information about complex, multidimensional probability densities in relatively few iterations, the required computational work load may be large. It is known that the structure of the metagraph is critical for the speed at which such algorithms converge to their equilibrium (where they sample correctly), but so far there exists no theory that can guide us when designing the algorithms. All we have is a set of design rules to follow that guarantee correct convergence in the limit of infinitely many iterations. Fortunately, practical experience shows that perturbation of one model parameter at a time usually gives reasonable computational efficiency.

An entirely different class of Monte Carlo algorithms, the *genetic algorithms*, have in many cases demonstrated their efficiency for finding a best fitting model for general inverse problems (highly nonlinear problems). These algorithms are, unfortunately, not well adapted for resolution analysis. Owing to the presence of irreversible steps ('selection') in this type of algorithm, no known genetic algorithms used in inverse problem analysis sample the model space according to the posterior probability density. However, they are probably the most efficient to find the maximum *a posteriori* model.

For linear and weakly nonlinear inverse problems the extended resolution analysis offered by inverse Monte Carlo sampling can be a useful supplement to known analytical methods. An interesting comparison between Monte Carlo analysis and Occam's inversion, applied to the same seismic inverse problem, is provided by Gouveia and Scales (1997).

7. Conclusion

The general inverse problem is not susceptible to analytical treatment as the relationship between model and data and/or the *a priori* model constraints are only defined through numerical procedures. Sampling strategies are the only known methods that allow an approximately correct evaluation of model resolution for such problems. The possible complexity of the posterior probability density for general inverse problems forces us to consider a wider class of resolution indicators than is usually necessary for linear problems.

The extended class of resolution indicators provides answers to questions of importance for data interpretation: How likely is it that a certain kind of structure exists in the Earth, given the observed data and certain *a priori* hypotheses? The answer to this type of question is equivalent, in a Bayesian formulation, to the determination of the *a posteriori* probability of the considered structure. Such a probability can be estimated, even when the model space is high-dimensional and the *a posteriori* distribution is highly complex, by inverse sampling in the model space. A rather efficient class of inverse sampling algorithms can be designed through a generalization of the Metropolis algorithm (Mosegaard and Tarantola 1995). Where the original Metropolis algorithm was able to sample a probability density $L(\boldsymbol{m})$, given an algorithm that evaluated $L(\boldsymbol{m})$ for given $\boldsymbol{m}$, the generalized algorithm is able to sample $\sigma(\boldsymbol{m}) = \rho(\boldsymbol{m})L(\boldsymbol{m})$, where $\rho(\boldsymbol{m})$ is any *a priori* probability distribution. The algorithm works even if we do not have a way of evaluating $\rho(\boldsymbol{m})$ at given points $\boldsymbol{m}$, but only an algorithm that samples $\rho(\boldsymbol{m})$. The general algorithm opens the possibility of introducing complex prior information into the solution of inverse problems, a possibility that brings quantitative data analysis a step closer to scientific data interpretation.

Acknowledgments

KM would like to thank the following people for inspiration through cooperation and numerous discussions: Albert Tarantola, Peter Salamon, Jacob Møerch Pedersen, Bjarne Andresen, David Snyder, Satish Singh, Helle Wagner and Camilla Rygaard-Hjalsted. KM would also like to thank Anthony Lomax and an anonymous referee for careful and useful reviews.

References

Backus G 1970a Inference from inadequate and inaccurate data I *Proc. Natl Acad. Sci., USA* **65**, 1–105

——1970b Inference from inadequate and inaccurate data II *Proc. Natl Acad. Sci., USA* **65** 281–7

——1970c Inference from inadequate and inaccurate data III *Proc. Natl Acad. Sci., USA* **67** 282–9

Barton P 1992 LISP-B revisited, a new look under the Caledonides of northern Britain *Geophys. J. Int.* **110** 371–91

Cary P W and Chapman C H 1988 Automatic 1-D waveform inversion of marine seismic refraction data *Geophys. J. R. Astron. Soc.* **93** 527–46

Feller W 1970 *An Introduction to Probability Theory and its Applications* (New York: Wiley)

Futterman W I 1962 Dispersive body waves *J. Geophys. Res.* **67** 5279–91

Ganley D C 1981 A method for calculating synthetic seismograms which include the effects of absorption and dispersion *Geophysics* **46** 1100–7

Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images *IEEE Trans. Pattern Anal. Mach. Intel.* **PAMI-6** 721–41

Gouveia W and Scales J 1997 Resolution of seismic waveform inversion: Bayes versus Occam *Inverse Problems* **13** 323–49

Haskell N A 1953 The dispersion of surface waves from point sources in a multilayered media *Bull. Seismol. Soc. Am.* **43** 17–34

Koren Z, Mosegaard K, Landa E, Thore P and Tarantola A 1991 Monte Carlo estimation and resolution analysis of seismic background velocities *J. Geophys. Res.* **96** 20 289–99

Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **1** 1087–92

Mosegaard K, Singh S C, Snyder D and Wagner H 1997 Monte Carlo analysis of seismic reflections from Moho and the W-reflector *J. Geophys. Res.* B **102** 2969–81

Mosegaard K and Tarantola A 1995 Monte Carlo sampling of solutions to inverse problems *J. Geophys. Res.* B **100** 12 431–47

Pedersen J B and Knudsen O 1990 Variability of estimated binding parameters *Biophys. Chem.* **36** 167–76

Rothman D H 1985 Nonlinear inversion, statistical mechanics, and residual statics estimation *Geophysics* **50** 2784–97

Singh S C and McKenzie D P 1993 Layering in the lower crust *Geophys. J. Int.* **113** 622–8

Tarantola A 1987 *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation* (New York: Elsevier)

Tarantola A and Valette B 1982 Inverse problems = Quest for information *J. Geophys.* **50** 159–70

Weibe R A 1993 The Pleasant Bay layered gabbro-diorite, coastal Maine: Ponding and crystallization of basaltic injections into a silicic magma chamber *J. Petrol.* **34** 461–89

nated to a surface Al, providing a mechanism for the interchange of O_s and O_{ads}. These results also provide evidence for incipient $Al(OH)_3$ formation on the surface. The ultimate structure of the heavily hydrated surface is clearly very complicated and may depend strongly on sample history. It is possible that the $Al(OH)_3$ species can be removed completely (perhaps starting near steps or other defects), leaving a less reactive surface that is completely O_sH-terminated, which is similar to the known surfaces of aluminum hydroxides (*1*).

An idealized model for fully hydroxylated α-Al_2O_3(0001) (*28*) replaces each surface Al with three H atoms (Fig. 5A), yielding a coverage >15 OH per square nanometer. Room-temperature MD simulations of this model revealed a complex dynamic structure (Fig. 5B), with one out of every three OH groups, on average, lying parallel to the surface because of in-plane hydrogen bonding. Calculated O–H vibrational spectra (*25*) yielded two broad peaks at ~3470 and 3650 cm^{-1}, with the peak at ~3470 cm^{-1} corresponding to in-plane OH groups. The peak at ~3650 cm^{-1} is close to the single peak (3720 to 3733 cm^{-1}) that is observed in most measurements on hydroxylated α-Al_2O_3(0001) (*29*) and to the range that is generally assigned to bridging OH groups (*2*, *26*). The peak at ~3470 cm^{-1} is red-shifted by hydrogen bonding and is generally not seen in single-crystal experiments, perhaps because of selection rules or because it is too broad. Our finding of two peaks split by 200 cm^{-1} contradicts all previous classifications of OH vibrations (and subsequent cluster modeling) (*2*) on aluminas, which assume that all OH groups with the same coordination of O and neighboring Al have the same frequency. By this criterion, all of the surface OH groups in Fig. 5 are equivalent; however, their stretch frequencies clearly depend also on longer range environmental effects.

The present investigation of α-Al_2O_3(0001) has elucidated several aspects of the complex interactions of H_2O with an alumina surface, especially the dynamics of dissociation reactions at low and high coverages. On the basis of these results, a consistent interpretation of a diverse set of experimental data on hydroxylated alumina surfaces begins to emerge.

References and Notes

1. K. Wefers and C. Misra, *Alcoa Tech. Pap. 19 (revised)* (Alcoa Laboratories, St. Louis, MO, 1987).
2. H. Knözinger and P. Ratnasamy, *Catal. Rev. Sci. Eng.* **17**, 31 (1978).
3. M. Gautier *et al.*, *J. Am. Ceram. Soc.* **77**, 323 (1994).
4. P. de Sainte Claire, K. C. Hass, W. F. Schneider, W. L. Hase, *J. Chem. Phys.* **106**, 7331 (1997).
5. G. N. Robinson, Q. Dai, A. Freedman, *J. Phys. Chem. B* **101**, 4940 (1997).
6. R. Car and M. Parrinello, *Phys. Rev. Lett.* **55**, 2471 (1985).
7. The gradient-corrected exchange-correlation [Bernstein, Lee, Yang, and Primakoff (BLYP)] functional used here is from A. D. Becke [*Phys. Rev. A* **38**, 3098 (1988)] and C. Lee, W. Yang, and R. Parr [*Phys. Rev. B* **37**, 785 (1988)]. Norm-conserving numerical pseudopotentials were generated for Al and O with the procedure of N. Troullier and J. L. Martins [*ibid.* **43**, 1993 (1991)], and a local analytic pseudopotential was derived for H. This is essentially a softened Coulomb potential with a core radius of 0.25 atomic units. Electron wave functions are expanded in a plane-wave basis set with an energy cutoff of 70 rydbergs (Ry). We used the Car-Parrinello Molecular Dynamics code in the parallelized 2.5 version (developed by J. Hutter and copyrighted by IBM, Armonk, NY). All calculations were performed on a 32-node IBM RS6000 SP at the IBM Watson Research Laboratory (Yorktown Heights, NY).
8. In the MD runs, a value of 400 au was used for the fictitious electron mass of the Car-Parrinello Lagrangian multipliers (*6*), and each hydrogen molecule was replaced by deuterium to improve the separation between electronic and ionic degrees of freedom. The time step in the Verlet algorithm for the integration of the equations of motions was ~0.1 fs.
9. The importance of chemical reaction dynamics in general has recently been highlighted in a special issue of *Science* [Reaction Dynamics, *Science* **279**, 1875–1895 (1998)].
10. A. Curioni *et al.*, *J. Am. Chem. Soc.* **119**, 7218 (1997).
11. V. E. Puchin *et al.*, *Surf. Sci.* **370**, 190 (1997); J. Ahn and J. W. Rabalais, *ibid.* **388**, 121 (1997).
12. See, for example, S. Blonski and S. H. Garofalini, *ibid.* **295**, 263 (1993).
13. See, for example, M. Causa, R. Dovesi, C. Pisani, C. Roetti, *ibid.* **215**, 259 (1989); I. Manassidis, A. De Vita, M. J. Gillan, *Surf. Sci. Lett.* **285**, L517 (1993); I. Frank, D. Marx, M. Parrinello, *J. Chem. Phys.* **104**, 8143 (1996).
14. J. M. McHale, A. Auroux, A. J. Perrotta, A. Navrotsky, *Science* **277**, 788 (1997). For earlier work, see P. A. Thiel and T. E. Madey, *Surf. Sci. Rep.* **7**, 211 (1987) and references therein.
15. J. M. Wittbrodt, W. L. Hase, H. B. Schlegel, *J. Phys. Chem. B* **102**, 6539 (1998).
16. K. C. Hass, W. F. Schneider, A. Curioni, W. Andreoni, in preparation.
17. Earlier calculations used much smaller supercells than the present work. Such studies were therefore limited in their ability to provide accurate adsorbate structures and energies and to study the H_2O coverage dependence and phenomena such as collective effects and surface diffusion.
18. J. Goniakowski and M. J. Gillan, *Surf. Sci.* **350**, 145 (1996); P. J. D. Lindan, N. M. Harrison, J. M. Holdender, M. J. Gillan, *Chem. Phys. Lett.* **261**, 246 (1996); P. J. D. Lindan, N. M. Harrison, M. J. Gillan, *Phys. Rev. Lett.* **80**, 762 (1998).
19. W. Langel and M. Parrinello, *J. Chem. Phys.* **103**, 3240 (1995).
20. Lagrange multipliers were introduced to constrain the relevant H–O_s distance, and the average constraint forces were determined from constant temperature simulations [S. Nosé, *J. Chem. Phys.* **81**, 511 (1984); W. G. Hoover, *Phys. Rev. A* **31**, 1695 (1985)] of at least 0.2 ps.
21. S. Scheiner, in *Proton Transfer in Hydrogen-Bonded Systems*, T. Bountis, Ed. (Plenum, New York, 1992), p. 29.
22. S. Blonski and S. H. Garofalini, *J. Phys. Chem.* **100**, 2201 (1996).
23. The temperature was not controlled but was increased slowly from ~100 to ~350 K. The system was then allowed to evolve for a time interval of >1 ps. The average temperature was 250 K.
24. D. E. Brown, D. J. Moffatt, R. A. Wolkow, *Science* **279**, 542 (1998).
25. Vibrational frequencies were estimated from the power spectra of the (partial) velocity-velocity autocorrelation functions and were rescaled to account for the fictitious electronic mass and the different mass used for the proton.
26. V. I. Lygin and I. S. Muzyka, *Russ. J. Phys. Chem.* **69**, 1829 (1995); A. Tsyganenko and P. Mardilovich, *J. Chem. Soc. Faraday Trans.* **92**, 4843 (1996).
27. B. A. Huggins and P. D. Ellis, *J. Am. Chem. Soc.* **114**, 2098 (1992).
28. M. A. Nygren, D. H. Gay, C. R. A. Catlow, *Surf. Sci.* **380**, 113 (1997).
29. C. Morterra, G. Ghiotti, E. Garrone, F. Boccuzzi, *J. Chem. Soc. Faraday Trans. 1* **72**, 2722 (1976); J. G. Chen, J. E. Crowell, J. T. Yates, *J. Chem. Phys.* **84**, 5906 (1986); V. Coustet and J. Jupille, *Surf. Sci.* **307**, 1161 (1994).

11 August 1998; accepted 3 September 1998

Past Temperatures Directly from the Greenland Ice Sheet

D. Dahl-Jensen,* K. Mosegaard, N. Gundestrup, G. D. Clow, S. J. Johnsen, A. W. Hansen, N. Balling

A Monte Carlo inverse method has been used on the temperature profiles measured down through the Greenland Ice Core Project (GRIP) borehole, at the summit of the Greenland Ice Sheet, and the Dye 3 borehole 865 kilometers farther south. The result is a 50,000-year-long temperature history at GRIP and a 7000-year history at Dye 3. The Last Glacial Maximum, the Climatic Optimum, the Medieval Warmth, the Little Ice Age, and a warm period at 1930 A.D. are resolved from the GRIP reconstruction with the amplitudes –23 kelvin, +2.5 kelvin, +1 kelvin, –1 kelvin, and +0.5 kelvin, respectively. The Dye 3 temperature is similar to the GRIP history but has an amplitude 1.5 times larger, indicating higher climatic variability there. The calculated terrestrial heat flow density from the GRIP inversion is 51.3 milliwatts per square meter.

Measured temperatures down through an ice sheet relate directly to past surface temperature changes. Here, we use the measurements from two deep boreholes on the Greenland Ice Sheet to reconstruct past temperatures. The GRIP ice core (72.6°N, 37.6°W) was successfully recovered in 1992 (*1, 2*), and the 3028.6-m-deep liquid-filled borehole with a diameter of 13 cm was left undisturbed. Temperatures were then measured down through the borehole in 1993, 1994, and 1995 (*3, 4*). We used the measurements from 1995 (Fig. 1) (*4*), because there was no remaining evidence of disturbances from the drilling and

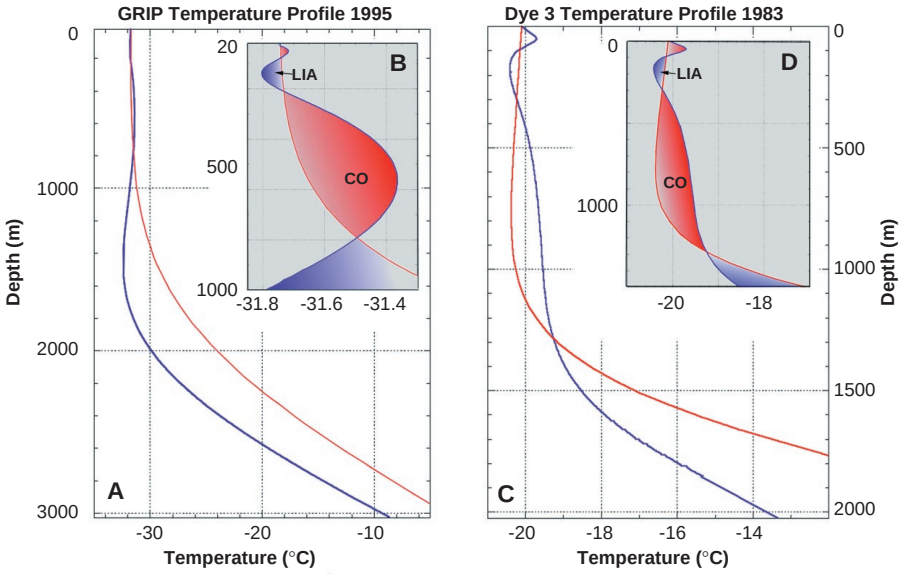


Fig. 1. The GRIP and Dye 3 temperature profiles [blue trace in (A) and (C)] are compared to temperature profiles [red trace in (A) and (C)] calculated under the condition that the present surface temperatures and accumulation rates have been unchanged back in time. (**A**) The GRIP temperature profile measured in 1995. The cold temperatures from the Glacial Period (115 to 11 ka) are seen as cold temperatures between 1200- to 2000-m depth. (**B**) The top 1000 m of the GRIP temperature profiles are enlarged so the Climatic Optimum (CO, 8 to 5 ka), the Little Ice Age (LIA, 1550 to 1850 A.D.), and the warmth around 1930 A.D. are indicated at the depths around 600, 140, and 60 m, respectively. (**C**) The Dye 3 temperature profile measured in 1983. Note the different shape of the temperature profiles when compared to GRIP and the different depth locations of the climate events. (**D**) The top 1500 m of the Dye 3 temperature profiles are enlarged so the CO, the LIA, and the warmth around 1930 A.D. are indicated at the depths around 800, 200, and 70 m, respectively.

the measurements were the most precise (±5 mK). Temperatures measured in a thermally equilibrated shallow borehole near the drill site are used for the top 40 m, because they are more reliable than the GRIP profile over this depth (*5*). The present mean annual surface temperature at the site is –31.70°C. The 2037-m-deep ice core from Dye 3 (65.2°N, 43.8°W) was recovered in 1981. We used temperature data from 1983 measurements with a precession of 30 mK (*6, 7*). The temperatures at the bedrock are –8.58°C at GRIP and –13.22°C at Dye 3. Calculations show that the basal temperatures have been well below the melting point throughout the past 100,000 years (*8*). Because there are still climate-induced temperature changes near the bedrock, we included 3 km of bedrock in the heat flow calculatin.

Past surface temperature changes are indicated from the shape of the temperature profiles (Fig. 1). We used a coupled heat- and ice-flow model to extract the climatic information from the measured temperature profiles. The temperatures down through the ice depend on the geothermal heat flow density (heat flux), the ice-flow pattern, and the past surface temperatures and accumulation rates. The past surface temperatures and the geothermal heat flow density are unknowns, whereas the past accumulation rates and ice-flow pattern are assumed to be coupled to the temperature history through relations found from ice-core studies (*9–11*). The total ice thickness is assumed to vary 200 m as described in (*9*). The coupled heat- and ice-flow equation is (*7, 9, 12*)

$$\rho c \frac{\partial T}{\partial t} = \nabla \cdot (K \nabla T) - \rho c \vec{v} \cdot \nabla T + f$$

where $T(x,z,t)$ is temperature, t is time, z is depth, x is horizontal distance along the flow line, $\rho(z)$ is ice density, $K(T,\rho)$ the thermal conductivity, $c(T)$ is the specific heat capacity, and $f(z)$ is the heat production term. The ice velocities, $\vec{v}$ (x,z,t), are calculated by an ice-flow model (*9, 13*). Model calculations to reproduce a present-day temperature profile through the ice sheet are started 450,000 years ago (ka) at GRIP (100 ka at Dye 3),

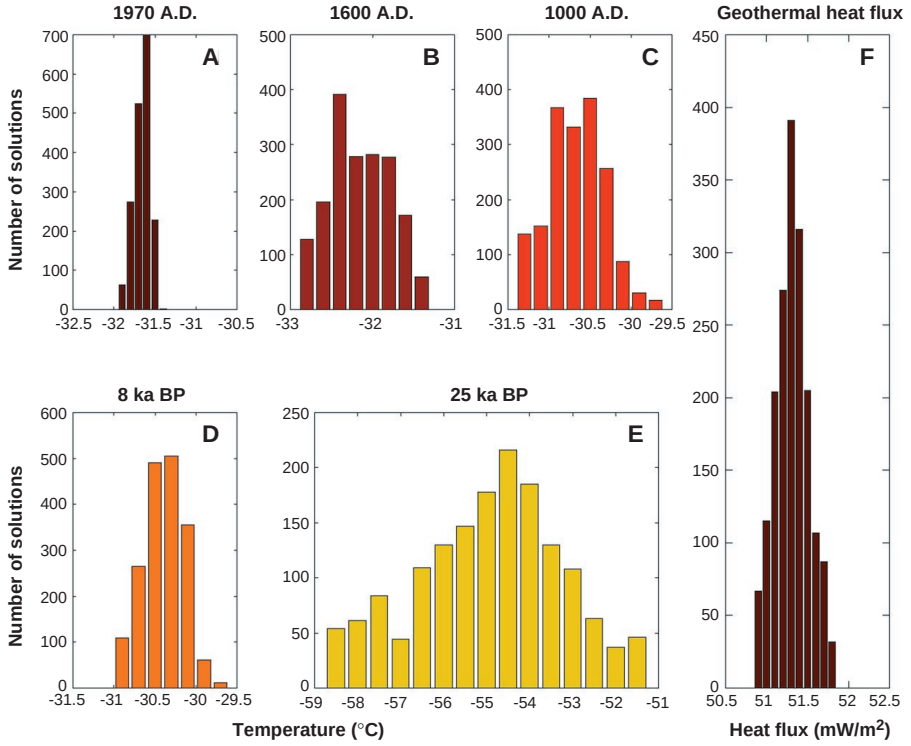


Fig. 2. (**A** through **E**) The probability distributions of the past surface temperatures at the Greenland Ice Sheet summit at selected times before present. They are constructed as histograms of the 2000 Monte Carlo sampled and accepted temperature histories (*17*). All temperature distributions are seen to have a zone with maximum values, the most likely values, which are assumed to be the reconstructed surface temperature at these times (*18*). (**F**) The probability distribution of the sampled geothermal heat flow densities. The most likely value is 51.3 mW/m^2.

D. Dahl-Jensen, K. Mosegaard, N. Gundestrup, S. J. Johnsen, A. W. Hansen, Niels Bohr Institute for Astronomy, Physics and Geophysics, Department of Geophysics, Juliane Maries Vej 30, DK-2100 Copenhagen OE, Denmark. G. D. Clow, USGS-Climate Program, Box 25046, MS 980, Denver Federal Center, Denver, CO 80225, USA. N. Balling, Department of Earth Sciences, Geophysical Laboratory, University of Aarhus, Finlandsgade 8, DK-8200 Aarhus N, Denmark.

*To whom correspondence should be addressed. E-mail: ddj@gfy.ku.dk

more than twice the time scale for thermal equilibrium of the ice-bedrock, so the unknown initial conditions are forgotten when generating the most recent 50,000-year temperature history (7000 years for Dye 3).

We developed a Monte Carlo method to fit the data and infer past climate. The Monte Carlo method tests randomly selected combinations of surface temperature histories and geothermal heat flow densities by using them as input to the coupled heat- and ice-flow model and considering the resulting degrees of fit between the reproduced and measured temperature profiles (*14–16*). Our results for each site are based on tests of 3.3×10^6 combinations of temperature histories and heat flow densities, of which 2000 solutions have been selected (*17*). The 2000 temperature histories and heat flow densities are sampled with a frequency proportional to their likelihood (*14, 15*), and all accepted solutions fit the observations within their limits of uncertainty.

Histograms of the sampled geothermal heat flow densities and of the temperature histories at each time before present can be made (for example, Fig. 2). The distributions in general show that there is a most likely value, a maximum, at all times, which we refer to as the temperature history (*18*). The distribution of accepted geothermal heat flow densities (Fig. 2F) has a median of 51.3 ± 0.2 mW/m^2, which is slightly higher than the heat flow density from Archean continental crust across the Baffin Bay in Canada. A few heat flow measurements have been made from the coast of Greenland (36 and 43 mW/m^2), but these are not corrected for long-term climate variations and are minimum values (*19*). The homogeneous thermal structure of ice is an advantage when the heat flow density and the temperature history are to be reconstructed (*20*).

Histograms from the GRIP reconstruction (Fig. 3) show that temperatures at the Last Glacial Maximum (LGM) were 23 ± 2 K colder than at present (*21*). The temperatures at this time, 25 ka, reflect the cold temperatures seen on the measured temperature profile at a depth of 1200 to 2000 m. Alternative reconstructions of the ice thickness and accumulation rates all reproduce LGM temperatures within 2 K (*9, 10, 22, 23*). The cold Younger Dryas and the warm Bølling/Allerød periods (*24*) are not resolved in the inverse reconstruction. The temperature signals of these periods have been obliterated by thermal diffusion because of their short duration (*25*). After the termination of the glacial period, temperatures in our record increase steadily, reaching a period 2.5 K warmer than present during what is referred to as the Climatic Optimum (CO), at 8 to 5 ka. Following the CO, temperatures cool to a minimum of 0.5 K colder than the present at around 2 ka. The record implies that the medieval period around 1000 A.D. was 1 K warmer than present in Greenland. Two cold periods, at 1550 and 1850 A.D., are observed during the Little Ice Age (LIA) with temperatures 0.5 and 0.7 K below the present. After the LIA, temperatures reach a maximum around 1930 A.D.; temperatures have decreased during the last decades (*26*). The climate history for the most recent times is in agreement with direct measurements in the Arctic regions (*27*). The climate history for the last 500 years agrees with the general understanding of the climate in the Arctic region (*28*) and can be used to verify the temperature amplitudes. The results show that the temperatures in general have decreased since the CO and that no warming in Greenland is observed in the most recent decades.

As seen in Fig. 3, resolution decreases back

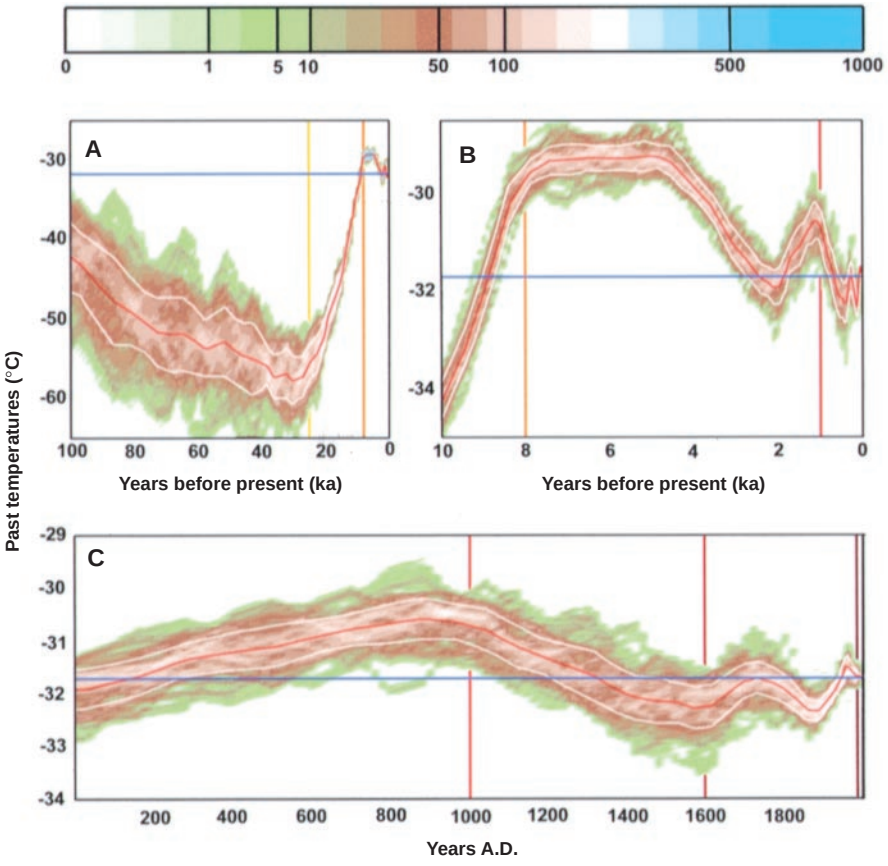


Fig. 3. The contour plots of all the GRIP temperature histograms as a function of time describes the reconstructed temperature history (red curve) and its uncertainty. The temperature history is the history at the present elevation (3240 m) of the summit of the Greenland Ice Sheet (*21*). The white curves are the standard deviations of the reconstruction (*18*). The present temperature is shown as a horizontal blue curve. The vertical colored bars mark the selected times for which the temperature histograms are shown in Fig. 2. (**A**) The last 100 ky BP. The LGM (25 ka) is seen to have been 23 K colder than the present temperature, and the temperatures are seen to rise directly into the warm CO 8 to 5 ka. (**B**) The last 10 ky BP. The CO is 2.5 K warmer than the present temperature, and at 5 ka the temperature slowly cools toward the cold temperatures found around 2 ka. (**C**) The last 2000 years. The medieval warming (1000 A.D.) is 1 K warmer than the present temperature, and the LIA is seen to have two minimums at 1500 and 1850 A.D. The LIA is followed by a temperature rise culminating around 1930 A.D. Temperature cools between 1940 and 1995.

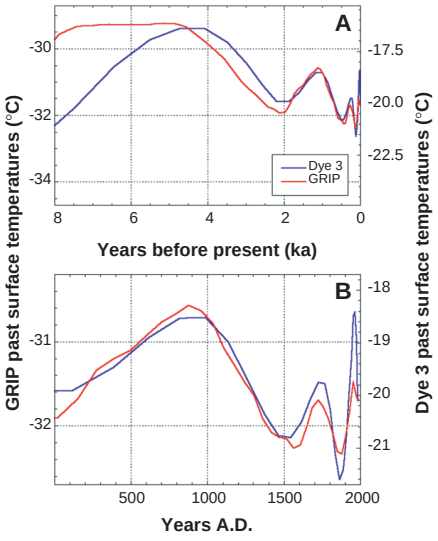


Fig. 4. The reconstructed temperature histories for GRIP (red curves) and Dye 3 (blue curves) are shown for the last 8 ky BP (**A**) and the last 2 ky BP (**B**). The two histories are nearly identical, with 50% larger amplitudes at Dye 3 than found at GRIP. The reconstructed climate must represent events that occur over Greenland, probably the high-latitude North Atlantic region.

in time (*25, 29*). For the GRIP reconstruction, an event with a duration of 50 years and an amplitude of 1 K can be resolved 150 years back in time with a measurement accuracy of 5 mK; an event with a similar amplitude but a duration of 1000 years can be detected back to 5 ka. An event that occurred 50 ka will now be observed in the temperature profile at the bedrock. Climate events for times older than 50,000 years before present (ky BP) are not well resolved (*30*). At Dye 3, the reconstructed climate history extends only to 7 ka, because the ice is 1000 m thinner than at the summit and surface accumulation rate is 50% higher. The LGM is not well resolved in the Dye 3 record, and consequently the geothermal heat flow density is not uniquely determined (*31*). On the other hand, the recent climate history has a higher resolution because of the increased accumulation (Fig. 4).

The Dye 3 record is nearly identical with the GRIP record back to 7 ka, but the amplitudes are 50% higher. Thus, the resolved climate changes have taken place on a regional scale; many are seen throughout the Northern Hemisphere (*27, 28, 32*). GRIP is located 865 km north of Dye 3 and is 730 m higher in elevation. Surface temperatures at the summit are influenced by maritime air coming in from the North Atlantic and air masses arriving from over northeastern Canada (associated with the Baffin trough) (*28, 32, 33*). Temperatures at Dye 3 will be influenced to a greater degree by the North Atlantic maritime air masses. Dye 3 is located closer to the center of the highest atmospheric variability, which is associated with large interseasonal, interannual, and decadal temperature changes (*32, 34*). It is therefore believed that the observed difference in amplitudes between the two sites is a result of their different geographic location in relation to variability of atmospheric circulation, even on the time scale of a millennium.

References and Notes

1. Greenland Ice-Core Project (GRIP) members, *Nature* **364**, 203 (1993).
2. W. Dansgaard *et al.*, *ibid.*, p. 218.
3. N. S. Gundestup, D. Dahl-Jensen, S. J. Johnsen, A. Rossi, *Cold Reg. Sci. Technol.* **21**, 399 (1993).
4. G. D. Clow, R. W. Saltus, E. D. Waddington, *J. Glaciol.* **42**, 576 (1996).
5. The deep borehole is located in a building, and the liquid surface in the borehole is found at a depth of 40 m. The temperatures measured in the top 40 m are very disturbed, so we used measurements from an air-filled shallow borehole (100 m) near the borehole.
6. N. S. Gundestrup and B. L. Hansen, *J. Glaciol.* **30**, 282 (1984).
7. D. Dahl-Jensen and S. J. Johnsen, *Nature* **320**, 250 (1986).
8. D. Dahl-Jensen *et al.*, *J. Glaciol.* **43**, 300 (1997).
9. Between 50 and 20 ka, the ice thickness was 50 m less than at present, even though the ice sheet covered a larger area. The maximum ice thickness of 3230 m is found at 10 ka, after which the ice thickness gradually has decreased to the present 3028.6 m. The depression and uplift of the bedrock influences the elevation of the surface [S. J. Johnsen, D. Dahl-Jensen, W. Dansgaard, N. S. Gundestrup, *Tellus B* **47**, 624 (1995)].
10. K. M. Cuffey and G. D. Clow, *J. Geophys. Res.* **102**, 26383 (1997).
11. The past accumulation rates are determined by coupling them to the past (unknown) temperature through the relation $\lambda(T) = \lambda_O \exp[0.0467(T - T_O) - 0.000227(T - T_O)^2]$, where $\lambda(T)$ is the accumulation rate at the surface temperature T, λ_O is the present ice accumulation rate, which is 0.23 m/year at GRIP and 0.49 m/year at Dye 3, and T_O is the present surface temperatures at the sites: –31.7°C at GRIP and –20.1°C at Dye 3, respectively (*9*).
12. S. J. Johnsen, *IAHS-AISH Publ.* **118**, 388 (1977).
13. S. J. Johnsen and W. Dansgaard, *NATO ASI Ser. I Global Environ. Change* **2**, 13 (1992).
14. K. Mosegaard and A. Tarantola, *J. Geophys. Res.* **100**, 12431 (1995).
15. K. Mosegaard, *Inverse Problems* **14**, 405 (1998).
16. Our Monte Carlo scheme is a random walk in the high dimensional space of all possible models, m (temperature histories and geothermal heat flow densities). The temperature history has been divided in 125 intervals (interval length is 25 ky at 450 ka and 10 years at present). Including the geothermal heat flow density as an unknown the model space is 126-dimensional. In each step of the random walk, a perturbed model, m^i_{pert} of the current model vector m^i is proposed. The next model becomes equal to m^i_{pert} with an acceptance probability $P_{accept} = \min\{1,\exp(-[S(m^i_{pert}) - S(m^i)])\}$, where $S(m) = \sum_j (g^j(m) - d^j_{obs})^2$, which is the misfit function measuring the difference between $g(m)$, the calculated borehole temperatures, and d_{obs}, the observed temperatures. If the perturbated model is rejected, the next model becomes equal to m^i and a new perturbed model is proposed. To ensure an efficient sampling of all possible models, we developed ways of choosing the temperature histories and geothermal heat flow densities to be tested. The main scheme to perturb the models is to randomly select one of the 126 temperature/heat flow density parameters and change its value to a new value chosen uniformly at random within a given interval. A singular value decomposition (SVD) of the matrix $G = \{\partial g_j/\partial m_i\}$, evaluated in a near-optimal model, yields a set of eigenvectors in the model space whose orientations reveal efficient directions of perturbation for the random walk. The SVD method is included as a possible method of perturbing models especially in the start of the process as it speeds the Monte Carlo scheme significantly.
17. Of the 3.3×10^6 models tested during the random walk 30% have been accepted by the Monte Carlo scheme (*16*). Every 500 is chosen of those where the misfit function S (*16*) is less than the variance of the observations. The waiting time of 500 has been chosen to exceed the maximum correlation length of the output model parameters. This is a necessary condition for the 2000 models to be uncorrelated. To further ensure that the output models were uncorrelated, the random walk was frequently restarted at several random selected points in the model space.
18. The probabilistic formulation of the inverse problem leads to definition of a probability distribution in the model space, describing the likelihood of possible temperature histories and geothermal heat flow densities. The Monte Carlo scheme is constructed to sample according to this probability distribution. The histograms in Fig. 2 describe the probability distribution of the geothermal heat flow density and temperatures at times before present. The maxima in the histograms thus describe the most likely values. The method does not constrain the distributions to have a single maximum, indeed there could be histograms with several maxima, reflecting that more than one value of the temperature at this time would give a good fit to the observed temperature in the borehole. The histograms however, are all seen to have a well-defined zone with most likely past temperatures. A soft curve is fitted to the histograms and the maximum value is taken as the most likely value. The standard deviations shown in Fig. 3 are derived as deviations from the maximum value.
19. J. H. Sass, B. L. Nielsen, H. A. Wollenberg, R. J. Munroe, *J. Geophys. Res.* **77**, 6435 (1972).
20. C. Clauser *et al.*, *ibid.* **102**, 18417 (1997); L. Guillou-Frottier, J.-C. Marescal, J. Musset, *ibid.* **103**, 7385 (1998); H. N. Pollack, S. J. Hurter, J. R. Johnson, *Rev. Geophys.* **31**, 267 (1993); W. G. Powell, D. S. Chapman, N. Balling, A. E. Beck, in *Handbook of Terrestrial Heat-Flow Density Determination* (Kluwer Academic, New York, 1988), pp. 167–222.
21. In order to produce a past temperature record from the calculated past surface temperatures, the temperatures have been corrected to the present elevation of the GRIP site (and Dye 3 site respectively) using the surface elevation changes described in (*9*) and a lapse rate of 0.006 K/m.
22. K. M. Cuffey *et al.*, *Science* **270**, 455 (1995).
23. D. Dahl-Jensen, in *Proceedings of the Interdisciplinary Inversion Workshop 2*, Copenhagen, 19 May 1993, K. Mosegaard, Ed. (The Niels Bohr Institute for Astronomy, Physics and Geophysics, University of Copenhagen, Copenhagen, 1993), pp. 11–14.
24. C. U. Hammer *et al.*, *Report on the stratigraphic dating of the GRIP Ice Core. Special Report of the Geophysical Department* (Niels Bohrs Institute for Astronomy, Physics and Geophysics, University of Copenhagen, Copenhagen, in press).
25. J. Firestone, *J. Glaciol.* **41**, 39 (1995).
26. The amplitude of the warming at 1930 A.D. must be considered to be more uncertain. The information leading to this result are the measured temperatures in an open shallow borehole, where air movements could influence the measurements.
27. D. Fisher *et al.*, *NATO ASI Ser. I Global Environ. Change* **41**, 297 (1996); J. W. C. White *et al.*, *J. Geophys. Res.* **102**, 26425 (1997); J. W. Hurrel, *Science* **269**, 676 (1995); P. Frich *et al.*, in *DMI Scientific Report 96-1* (Danish Meteorological Institute, Copenhagen, 1996).
28. R. G. Barry and R. J. Charley, *Atmosphere, Weather & Climate* (Routledge, London, ed. 6, 1992); J. Overpeck *et al.*, *Science* **278**, 1251 (1997); H. H. Lamb, *Climate History and the Modern World* (Routledge, London, ed. 2, 1995); N. W. T. Brink and A. Weidick, *Quat. Res.* **4**, 429 (1974).
29. G. D. Clow, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **98**, 81 (1992).
30. To comply with this resolution the time steps have been chosen with increasing length back in time. The increasing length of the time steps can be considered as an efficient way of calculating the mean temperatures in the intervals so full available resolution is kept but the calculations are rationalized.
31. In (*7*), it is argued that parameter combinations of mean glacial temperature, mean glacial accumulation, and geothermal heat flow density can be found that fit the Dye 3 measurements due to the reduced resolution of the climate history reaching further back than 7 ka. A combination with a geothermal heat flow density of 38.7 mW/m^2 was chosen corresponding to a mean glacial temperature 12 K colder than the present temperatures. If a value of 51 mW/m^2 is chosen as that found for our inversion, the mean glacial temperature is 19 K colder than the present, which is well in agreement with the results found for the GRIP reconstruction. Comparison of the Dye 3 temperature history presented in (*7*) and that presented here shows a general good agreement for the last 7 ky. The history presented in (*7*) is more intuitive and less detailed, and the history has not been corrected for elevation changes. The ice thickness was assumed constant in this reconstruction.
32. L. K. Barlov, J. C. Rogers, M. C. Serreze, R. C. Barry, *J. Geophys. Res.* **102**, 26333 (1997).
33. R. A. Keen, *Occas. Pap. 34* (Institute of Arctic and Alpine Research, University of Colorado, Boulder, 1980).
34. S. Shubert, W. Higgins, C. K. Park, S. Moorthi, M. Svarez, *An Atlas of ECMWF Analyses (1980–87)*. Part II: *Second Moment Quantities*, *NASA Tech. Memorandum 100762* (1990); F. Rex, *World Surv. Climatol.* **4**, 1 (1969).

35. This is a contribution to the Greenland Ice Core Project (GRIP), a European Science Foundation program with eight nations and the European Economic Commission collaborating to drill through the central part of the Greenland Ice Sheet. G.D.C. thanks the USGS Climate History Program and NSF for support.

16 June 1998; accepted 1 September 1998

Inverse Problems **15** (1999) 573–583. Printed in the UK PII: S0266-5611(99)93884-4

Probabilistic analysis of implicit inverse problems

Klaus Mosegaard and Camilla Rygaard-Hjalsted

Department of Geophysics, Niels Bohr Institute for Astronomy, Physics and Geophysics, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark

Received 6 May 1998, in final form 20 October 1998

Abstract. Many inverse problems are most naturally formulated as *implicit* problems where data cannot be expressed in closed form as a function of the unknown model parameters. Notable examples are cases where the data satisfy a partial differential equation with model parameters as constants. When the problem can be recast as an ordinary *explicit* inverse problem the price to be paid is often that the functional relationship between data and model cannot be formulated in closed form. The consequence is typically that a computer time intensive numerical algorithm is needed to perform the forward calculation, thereby making an iterative solution of the problem unreasonably time consuming or impossible. In this paper we present a probabilistic procedure to solve rather general, implicit inverse problems through Monte Carlo sampling of feasible solutions. Our starting point is a probabilistic formulation of inverse problems, and our goal is to produce near-independent samples from the posterior distribution in model space. From these samples important information on the model and its resolution can be obtained. The proposed algorithm is applied to an implicit real-data problem involving analysis of the fluid motions at the Earth's core–mantle boundary from geomagnetic data.

1. Introduction

The first formulation of an inverse problem is often implicit. An important example of this is the estimation of the coefficients of a certain partial differential equation from knowledge of their solutions in given domains. In seismology we face the problem of reconstructing the density $\rho(\boldsymbol{x})$ and components of the elastic tensor $c_{ijkl}(\boldsymbol{x})$ as a function of the space coordinate $\boldsymbol{x}$ from information on solutions $\boldsymbol{u}(\boldsymbol{x}, t)$ to the equation

$$\rho(\boldsymbol{x})\frac{\partial^2 u_i(\boldsymbol{x}, t)}{\partial t^2} - \sum_{jkl}\frac{\partial}{\partial x_j}\left(c_{ijkl}(\boldsymbol{x})\frac{\partial u_k(\boldsymbol{x}, t)}{\partial x_l}\right) = 0 \tag{1}$$

where $u_i(\boldsymbol{x}, t)$ is the ith component of the vector field $\boldsymbol{u}(\boldsymbol{x}, t)$. Since equation (1) cannot be solved directly for the model $m = (c_{ijkl}(\boldsymbol{x}), \rho(\boldsymbol{x}))$ it gives an implicit relation between the data $d = \psi(\boldsymbol{x}, t)$ (the known solutions to the equation) and the model, and has the form

$$L(d, m) = 0 \tag{2}$$

where L is a function (operator) given in closed form.

Another example, which we shall investigate in some detail in the numerical example given in this paper, is geomagnetic studies of fluid motions immediately below the Earth's core–mantle boundary (CMB). In this case we are interested in analysing the velocity field $\boldsymbol{u}$ from the geomagnetic field $\boldsymbol{B}$ and its time derivative $\partial\boldsymbol{B}/\partial t$ at the CMB (estimated from

0266-5611/99/020573+11$19.50

downward continuation of the observed fields at or above the Earth's surface). The key to solving this problem is the diffusionless magnetic induction equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{\nabla} \times (\boldsymbol{u} \times \boldsymbol{B}). \tag{3}$$

Again, equation (3) is an implicit relation $L(d, m) = 0$ between the model $m = \boldsymbol{u}(\boldsymbol{x}, t)$ and data $d = \boldsymbol{B}(\boldsymbol{x}, t)$.

Technical difficulties with solving implicit problems of this kind are usually overcome by recasting them into an explicitly formulated inverse problem where the data d and the model m are related through the forward relation:

$$d = F(m). \tag{4}$$

The price paid through this reformulation is often high: in the original implicit problem the function L is given by a *closed form expression*, whereas in the explicit formulation, the function F has no closed form expression. In practice, F is realized as a numerical algorithm that allows 'simulation' of data d, given the unknown model m. Numerical simulations are often computer time intensive and an iterative search for solutions m to (4) may be very time consuming.

In this paper we shall attack the implicit problem directly. We formulate the inverse problem in a probabilistic framework, and demonstrate how a Markov-chain Monte Carlo method can be used to sample solutions to the problem. To simplify the exposition we shall modify the problem slightly. First, we assume that the problem is discrete and finite-dimensional: the data and model have the form $\boldsymbol{d} = (d_1, \ldots, d_N)$ and $\boldsymbol{m} = (m_1, \ldots, m_M)$, respectively. Second, with little loss of generality we shall assume that the implicit problem

$$g(\boldsymbol{d}, \boldsymbol{m}) = 0 \tag{5}$$

can be formulated as

$$\boldsymbol{d}_1 = f(\boldsymbol{d}_2, \boldsymbol{m}) \tag{6}$$

where $\boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2)$.

Given probabilistic prior information on $\boldsymbol{m}$, and given a statistical description of the observational uncertainties of $\boldsymbol{d}$, it will be our task to describe a random walk in the model space $\mathcal{M}$ that asymtotically samples the posterior probability density, that is, samples models that are consistent with data as well as prior information.

After we have described the algorithm we give an example of its use in connection with analysis of the fluid motions at the CMB from geomagnetic data.

2. Probabilistic formulation of the implicit, inverse problem

Our probabilistic formulation of the inverse problem will follow ideas laid out in Tarantola and Valette (1982) and Tarantola (1987) in which the solution is given as a probability density over the space $\mathcal{X}$ of parameters describing the physical system. This density $\sigma(\boldsymbol{x})$, referred to as the *posterior probability density*, summarizes all information about the physical system after incorporation of both *a priori* information and data information. The components of a vector $\boldsymbol{x} \in \mathcal{X}$ are the unknown 'model parameters' as well as observed 'data parameters'. We can therefore write $\mathcal{X} = \mathcal{D} \times \mathcal{M}$, where $\mathcal{D}$ is the data space, and $\boldsymbol{x} = (\boldsymbol{d}, \boldsymbol{m})$. The prior information on the system parameters is described by the prior probability density $\rho(\boldsymbol{x})$.

2.1. The implicit inverse problem

Let us assume that the inverse problem is given by a (possibly implicit) relation between the system parameters

$$g(\boldsymbol{d}, \boldsymbol{m}) = 0. \tag{7}$$

Without loss of practical applicability, we shall assume that the data parameter vector $\boldsymbol{d}$ can be reorganized (possibly by changing the order of its parameters) into a new parameter vector $(\boldsymbol{d}_1, \boldsymbol{d}_2)$ such that the implicit inverse problem (7) can be formulated as

$$\boldsymbol{d}_1 = f(\boldsymbol{d}_2, \boldsymbol{m}) \tag{8}$$

where f is a function given in closed form. We will use the term *dependent* parameters about the parameters on the left-hand side of (8) and *independent* parameters for those on the right-hand side. The subspaces spanned by possible data vectors $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ will in the following be denoted $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Besides (8) we have prior information about data described by a probability density

$$\rho_{\tilde{\boldsymbol{d}}_1,\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}) = \rho_{\tilde{\boldsymbol{d}}_1,\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}_1, \boldsymbol{d}_2) \tag{9}$$

over the data space $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2$ with the observed data $\tilde{\boldsymbol{d}} = (\widetilde{\boldsymbol{d}}_1, \widetilde{\boldsymbol{d}}_2)$ as expectation, and describing the observational uncertainties. In the following, we will assume that the uncertainties of $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ are independent and hence that $\rho_{\tilde{\boldsymbol{d}}_1,\tilde{\boldsymbol{d}}_2}$ can be factorized as

$$\rho_{\tilde{\boldsymbol{d}}_1,\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}) = \rho_{\tilde{\boldsymbol{d}}_1}(\boldsymbol{d}_1)\rho_{\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}_2). \tag{10}$$

We also have prior information about model parameters, described by a probability density $\rho(\boldsymbol{m})$ over the model space $\mathcal{M}$. We will assume that this information is independent of the prior information over the data space so that

$$\rho(\boldsymbol{x}) = \rho_{\tilde{\boldsymbol{d}}_1}(\boldsymbol{d}_1)\rho_{\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}_2)\rho(\boldsymbol{m}).$$

Our problem will be to obtain information about the model parameters contained in $\boldsymbol{m}$ from the observed data described by $\rho_{\tilde{\boldsymbol{d}}}(\boldsymbol{d})$, taking into account the prior information on models, given by $\rho(\boldsymbol{m})$, and the physical correlation between model parameters and data, given by (8). The total available information on $\boldsymbol{m}$ will be summarized by the *posterior probability density* $\sigma_{\tilde{\boldsymbol{d}}}(\boldsymbol{m})$ over the model space from which model resolution (non-uniqueness and uncertainties) can be analysed. Note that, in the present exposition, $\boldsymbol{m}$, $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ are random variables, whereas $\widetilde{\boldsymbol{d}}_1$ and $\widetilde{\boldsymbol{d}}_2$ are fixed (non-random) vectors used in a parametric description of probability densities. These densities will carry $\widetilde{\boldsymbol{d}}_1$ and/or $\widetilde{\boldsymbol{d}}_2$ as subscripts.

The lack of an explicit, closed form expression for the direct problem forces us to regard the second data subvector $\boldsymbol{d}_2$ as a pseudo model vector with prior probability density $\rho_{\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}_2)$. This allows us to sample the joint posterior probability density $\sigma_{\tilde{\boldsymbol{d}}}(\boldsymbol{d}_2, \boldsymbol{m})$, and thereby the marginal density $\sigma_{\tilde{\boldsymbol{d}}}(\boldsymbol{m})$, by means of a Markov chain Monte Carlo algorithm.

The joint posterior probability density $\sigma_{\tilde{\boldsymbol{d}}}(\boldsymbol{d}_2, \boldsymbol{m})$ is given as the restriction of the prior probability density to the manifold in $\mathcal{X}$ defined by $\boldsymbol{d}_1 = f(\boldsymbol{d}_2, \boldsymbol{m})$. In other words,

$$\sigma(\boldsymbol{d}_2, \boldsymbol{m}) = C\rho_{\tilde{\boldsymbol{d}}_1}(f(\boldsymbol{d}_2, \boldsymbol{m}))\rho_{\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}_2)\rho(\boldsymbol{m}) \tag{11}$$

$$= L_{\tilde{\boldsymbol{d}}_1}(\boldsymbol{d}_2, \boldsymbol{m})\rho_{\tilde{\boldsymbol{d}}_2}(\boldsymbol{d}_2)\rho(\boldsymbol{m}) \tag{12}$$

where C is a normalization constant and

$$L_{\tilde{\boldsymbol{d}}_1}(\boldsymbol{d}_2, \boldsymbol{m}) = C\rho(f(\boldsymbol{d}_2, \boldsymbol{m})) \tag{13}$$

is the *joint likelihood function* for the independent parameter vector $(\boldsymbol{d}_2, \boldsymbol{m})$. It measures the fit between dependent data and 'synthetic' values of these parameters, calculated from the independent parameter vector $(\boldsymbol{d}_2, \boldsymbol{m})$, and is typically of the form

$$L(\boldsymbol{d}_2, \boldsymbol{m}) = k_1 \exp[-S(\boldsymbol{d}_2, \boldsymbol{m})] \tag{14}$$

where k is a constant and $S(\boldsymbol{d}_2, \boldsymbol{m})$ is a misfit function. $S(\boldsymbol{d}_2, \boldsymbol{m})$ is a norm measuring the distance between the dependent parameters and 'synthetic' values of these parameters.

Once we sample the joint *a posteriori* probability density $\sigma_{\tilde{d}}(\boldsymbol{d}_2, \boldsymbol{m})$ over the space $\mathcal{D}_2 \times \mathcal{M}$ of independent variables, the subvectors $\boldsymbol{m}_i$ of these samples $(\boldsymbol{d}_2, \boldsymbol{m})_i$ are samples of the marginal posterior probability density $\sigma_{\tilde{d}}(\boldsymbol{m})$ for the model parameters.

2.2. The general, explicit inverse problem

For the explicit inverse problem, the model alone constitutes the independent parameters, and the data are the dependent parameters, so equation (8) becomes

$$\boldsymbol{d} = f(\boldsymbol{m}) \tag{15}$$

and (11) becomes

$$\sigma(\boldsymbol{m}) = C_2 L(\boldsymbol{m}) \rho_M(\boldsymbol{m}) \tag{16}$$

where C_2 is a normalization constant.

3. Monte Carlo sampling of the *a posteriori* distribution

Given an algorithm that is capable of generating points $\boldsymbol{x}_i$ in the parameter space with prior probability ρ_i, and an algorithm that allows evaluation of the likelihood of any parameter vector, a Markov chain Monte Carlo (MCMC) algorithm can be used to sample the posterior probability density in $\mathcal{X}$ (further details on the application of MCMC sampling to analysis of inverse problems can be found in Mosegaard and Tarantola (1995)).

Algorithm. *Given a random function $V(\boldsymbol{x}^{(n)})$ which samples the prior probability density $\rho(\boldsymbol{x})$ if applied iteratively*

$$\boldsymbol{x}^{(n+1)} = V(\boldsymbol{x}^{(n)}) \tag{17}$$

and a random function $U(0, 1)$ generating a uniformly distributed random number from the interval $[0, 1]$. *The random function W, which iteratively operates on the current parameter vector $\boldsymbol{x}^{(n)}$ and produces the next parameter vector $\boldsymbol{x}^{(n+1)}$*

$$\boldsymbol{x}^{(n+1)} = W(\boldsymbol{x}^{(n)}) = \begin{cases} V(\boldsymbol{x}^{(n)}) & \text{if } U(0,1) \leqslant \min\left[1, \dfrac{L(V(\boldsymbol{x}^{(n)}))}{L(\boldsymbol{x}^{(n)})}\right] \\ \boldsymbol{x}^{(n)} & \text{else} \end{cases} \tag{18}$$

samples the probability density $\sigma(\boldsymbol{x}) = C L(\boldsymbol{x}) \rho(\boldsymbol{x})$, where C is a normalization constant.

The algorithm (18) works if V satisfies the following criteria:

- the iterative procedure (17) must allow access to all points $\boldsymbol{x}$ in the parameter space, given enough iterations;
- it is possible to visit the same point twice (or more) in succession: $\boldsymbol{x}^{(n)} = V(\boldsymbol{x}^{(n)})$.

The algorithm (18) samples the posterior probability density in $\mathcal{X}$ asymptotically. The word 'asymptotically' means in this case that the statistical correlation between samples taken at times separated by, say, n iterations will converge towards zero as n goes to infinity.

The main advantage of sampling the posterior is that sampling is concentrated in areas of the model space where parameter vectors consistent with data as well as prior information exist. Moreover, approximate posterior covariances, higher-order moments, and posterior probabilities of given events are easily evaluated. They can be expressed as

$$R(\mathcal{E}, f) = \int_{\mathcal{E}} f(\boldsymbol{x})\sigma(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{19}$$

where $f(\boldsymbol{x})$ is a given function of the parameters and $\mathcal{E}$ is an event (subset) in the parameter space $\mathcal{M}$. The integral $R(\mathcal{E}, f)$ can be evaluated as the following simple average

$$R(\mathcal{E}, f) \approx \frac{1}{N} \sum_{\{n|\boldsymbol{x}_n \in \mathcal{E}\}} f(\boldsymbol{x}_n) \tag{20}$$

where N is the total number of samples taken in $\mathcal{E}$.

In the following we present a real-data example of analysis of an implicit, nonlinear inverse problem from geomagnetism. We demonstrate how the above algorithm is applied to this problem, and we will also investigate to what extent the solution changes if the problem is modified to an explicit problem, in this case through linearization.

4. Example: inversion of geomagnetic data

4.1. Formulation of the inverse problem

Studies of the fluid flow deep down in the Earth's outer core may provide us with insight into the poorly understood dynamo action responsible for the generation and maintenance of the Earth's magnetic field. The outer core is predominantly made of molten iron, its masses move, and it is an electrical conductor. The magnetic field is believed to have been present throughout most of Earth's history, and the widely accepted hypothesis today is that fluid flow interacts with the magnetic field and thereby generates electrical currents. These currents themselves set up a magnetic field and thus reinforce the Earth's field self-perpetually (Roberts 1987).

We have little direct evidence of the constitution of the fluid flow and the dynamics driving the motions, but we can use magnetic field observations with their time variations to generate an approximate picture of flow at the CMB. The magnetic field varies on many different time scales, but we are concerned with the longest of these, the *secular variation*, since slow changes result from motion within the core. For the sake of mathematical convenience, our 'data' will be spherical harmonic Gauss coefficients $\boldsymbol{g}$ relating to the geomagnetic field $\boldsymbol{B}$, and spherical harmonic coefficients $\dot{\boldsymbol{g}}$ relating to the secular variation, the time derivative of the magnetic field, $\partial \boldsymbol{B}/\partial t$, at the CMB.

The spherical harmonic functions are widely used to generate magnetic field models as they are eigenfunctions of the Laplace equation. The magnetic field is divergence-free and if we can also take it to be curl-free the geomagnetic potential satisfies the Laplacian. Thus, assuming a source-free region at the Earth's surface and downwards to the CMB we can represent the vector field $\boldsymbol{B}$ at the CMB as the negative gradient of a scalar potential V,

$$\boldsymbol{B} = -\boldsymbol{\nabla} V. \tag{21}$$

Here

$$V = a \sum_{l=1}^{\infty} \sum_{m=0}^{l} \left(\frac{a}{r}\right)^{l+1} (g_l^m \cos m\theta + h_l^m \sin m\theta) P_l^m(\cos\theta) \tag{22}$$

where $a = 6371.2$ km is the nominal Earth radius, g_l^m and h_l^m are the Gauss coefficients of degree l and order m, and $P_l^m(\cos\theta)$ are the Schmidt quasi-normalized associated Legendre functions (e.g. Langel 1987, Backus *et al* 1996). For shorthand we contract g_l^m and h_l^m into the vector $\boldsymbol{g}$. A similar expansion is made of the secular variation $\partial \boldsymbol{B}/\partial t$ and similarly we contract the corresponding coefficients $\dot{g}_l^m$ and $\dot{h}_l^m$ into the vector $\dot{\boldsymbol{g}}$.

In this example we consider the problem of inverting these data with their uncertainties with the purpose of modelling the structure of the horizontal fluid flow on a surface immediately below the CMB. We employ the inverse Monte Carlo sampling algorithm to solve this highly non-unique inverse problem, which is further complicated by its implicit relationship between data and model parameters. In this framework we solve for a family of solutions, each of which is consistent with data within its uncertainty and certain *a priori* information. With this subset of solutions we wish to study the resolution given by data and prior information and we do so by comparing fluid flow structures in the individual family members (Mosegaard 1998). By structures we mean, for example, correlated components motions like westward flow (known to govern the flow in a band near the equator in the western hemisphere) or flow circulations in the northern and southern hemispheres. In particular, we look at the horizontal divergence of the fluid flow known as upwelling and downwelling of flow below the CMB. We study the existence and geographical location of these structures (a projection of the world to the CMB is used for reference) in every model and conclude which structures are well resolved and which are poorly resolved by data and prior information. The existence of many solutions that are consistent with the data, their uncertainty and prior information reflects the facts that this inverse problem is fundamentally non-unique (Backus 1968) and that we have a finite amount of inaccurate observables.

4.2. Parametrization

The parametrization of the fluid flow is a set of 390 coefficients† representing fluid velocity. The relationship between the magnetic field $\boldsymbol{B}$ at the CMB, its first-order time derivative, the secular variation, $\partial \boldsymbol{B}/\partial t$ and the velocity vector field $\boldsymbol{u}$ near the CMB is expressed in the magnetic induction equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{\nabla} \times (\boldsymbol{u} \times \boldsymbol{B}) + \eta \boldsymbol{\nabla}^2 \boldsymbol{B} \tag{23}$$

where $\eta = 1/\mu_0 \sigma_{\mathrm{E}}$ is the magnetic diffusivity in the fluid outer core near the CMB, with μ_0 being magnetic permeability ($4\pi \times 10^{-7}$ N A^{-2}) and σ_{E} being electrical conductivity. The two terms on the right-hand side of (23) are respectively magnetic advection and diffusion. Determination of the flow can be made based on just the radial field at the CMB. Throughout this work we neglect the diffusion term for scaling arguments, and thus work under the hypothesis that the field lines are frozen into the fluid and reduce the radial part of (23) to

$$\frac{\partial B_r}{\partial t} = -\boldsymbol{\nabla}_{\mathrm{H}} \cdot (\boldsymbol{u} B_r) \tag{24}$$

where $\boldsymbol{\nabla}_{\mathrm{H}}$ is the horizontal gradient. This relationship is implicit since it cannot be solved for the magnetic field B_r and its time derivative $\partial B_r/\partial t$. As with the parameters of the fluid flow velocity we choose to represent the magnetic field and its secular variation through sets of coefficients (as a result of a spherical harmonic expansion of their corresponding potential

† The spherical harmonic expansion in (22) is an infinite series, but for all practical usages we truncate the series at some finite value. We have chosen $l = 13$. The error introduced by this truncation is not dealt with in this example. The modelled flows presented in this paper are filtered *a posteriori* but not regularized *a priori*. It is likely that a suitable *a priori* regularization of the models would compensate for the unavoidable truncation error. A discussion of truncation errors in geomagnetic modelling is found in Hulot *et al* (1991).

fields, see later) and after some rearrangement (see Whaler 1986 or Jackson and Bloxham 1991) we express (24) in terms of

$$\dot{\boldsymbol{g}} = \mathbf{A}(\boldsymbol{g})\boldsymbol{v} \tag{25}$$

where vectors $\dot{\boldsymbol{g}}$ and $\boldsymbol{g}$ each represent 195 coefficients with $\dot{\boldsymbol{g}}$ being the secular variation and $\boldsymbol{g}$ the magnetic field. $\boldsymbol{v} = (\boldsymbol{t}, \boldsymbol{s})$ is the unknown velocity vector consisting of the so-called toroidal ($\boldsymbol{t}$) and poloidal ($\boldsymbol{s}$) parts of the velocity field, each of 195 coefficients. $\mathbf{A}(\boldsymbol{g})$ is a matrix which depends on $\boldsymbol{g}$. Each component of $\dot{\boldsymbol{g}}$ in equation (25) is calculated from

$$\dot{g}_i = \sum_{j,k} A_{i,j,k} g_j v_k \tag{26}$$

and

$$A_{i,k} = \sum_{j} A_{i,j,k} g_j. \tag{27}$$

The nonlinear, implicit inverse problem posed in (25) has $\boldsymbol{g}$ and $\boldsymbol{v}$ as independent variables and $\dot{\boldsymbol{g}}$ as dependent variables.

4.3. Data, their uncertainty and the likelihood function

The data sets, $\dot{\boldsymbol{g}}$ and $\boldsymbol{g}$, are downward continuations to the CMB of a combination of geomagnetic satellite data from the Magsat mission (1980) and magnetic observatory and survey data from that epoch (ufm1 model, epoch 1980, Bloxham and Jackson 1992). Their uncertainties are estimated from Langel (1987) which we represent in two (Gaussian) probability distributions with variance $N(5 \cdot l^{-1}[\mathrm{nT}])^2$, and $N(1 \cdot l^{-1}[\mathrm{nT/year}])^2$ for $\boldsymbol{g}$ and $\dot{\boldsymbol{g}}$ respectively where $l = 1, 2, 3, \ldots, 13$ is the spherical harmonic degree. Spherical harmonics are solutions of Laplace's equation, which are obeyed by the geomagnetic potential (if the magnetic field is taken to be curl-free). We have expanded the potential fields corresponding to the magnetic field and the secular variation and the horizontal velocity field to spherical harmonic degree $l = 13$. Low degrees represent the long wavelength part of the energy spectrum and hence the monotonic decreasing functional form of the data uncertainty implies highest uncertainty here, where also most of the power of the signal is. However, the relative error (noise to signal) increases throughout the spectrum of l indicating that the gross features of the magnetic field are rather well determined whereas the fine details are poorly determined.

Equation (25) gives us the means to compute the secular variation given a set of toroidal and poloidal coefficients. Since the data uncertainty has a Gaussian distribution the joint likelihood function for $(\boldsymbol{g}, \boldsymbol{v})$ is

$$L_{\tilde{\dot{g}},\mathbf{C}_{\dot{g}}}(\boldsymbol{g}, \boldsymbol{v}) = C_3 \exp\{-\tfrac{1}{2}[\tilde{\dot{\boldsymbol{g}}} - \mathbf{A}(\boldsymbol{g})\boldsymbol{v}]^{\mathrm{T}} \mathbf{C}_{\dot{g}}^{-1} [\tilde{\dot{\boldsymbol{g}}} - \mathbf{A}(\boldsymbol{g})\boldsymbol{v}]\} \tag{28}$$

where C_3 is a normalization constant, $\tilde{\dot{\boldsymbol{g}}}$ is the vector of coefficients of the observed secular variation, and $\mathbf{C}_{\dot{g}}$ is the (diagonal) covariance matrix for the uncertainty in the secular variation.

4.4. A priori information

As *a priori* information we appeal to the knowledge of the dominant force balance in the core, the geostrophic balance. The starting point for describing geostrophic motion is the set of equations governing the dynamics of the fluid flow in the core; the Navier–Stokes equation in the Boussinesq approximation, see for example Bloxham and Jackson (1991),

$$\rho_0 \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{u} + 2\boldsymbol{\Omega} \times \boldsymbol{u} \right) = -\boldsymbol{\nabla} p + \rho' \boldsymbol{f}_g + \boldsymbol{J} \times \boldsymbol{B} - \rho_0 \nu \boldsymbol{\nabla}^2 \boldsymbol{u} \tag{29}$$

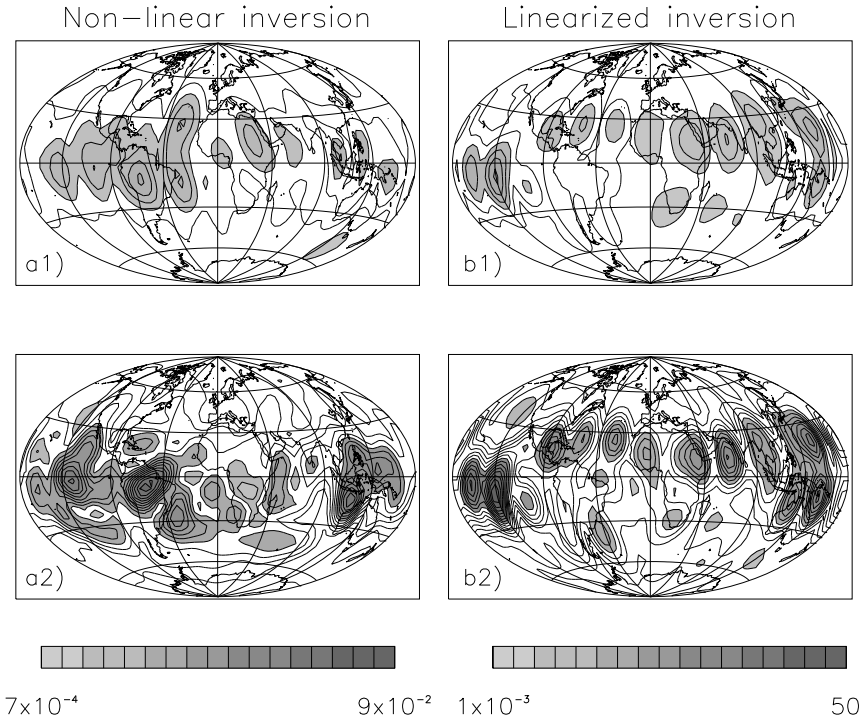


Figure 1. Standard deviation of the horizontal divergence of fluid flow from (a) nonlinear inversion for $\varepsilon = 10^{-2}$; (b) linearized inversion for $\varepsilon = 10^{-2}$; (c) nonlinear inversion for $\varepsilon = 10^{-1}$; (d) linearized inversion for $\varepsilon = 10^{-1}$. Note that the horizontal scale is different for the nonlinear inversions than for the linearized inversions.

where ρ_0 and ρ' are the hydrostatic density and departure from hydrostatic density, respectively, $\boldsymbol{u}$ is the vector of unknown core fluid velocity components, $\boldsymbol{\Omega}$ is the Earth's rotation vector, p is the non-hydrostatic part of the pressure, $\boldsymbol{f_g}$ is the gravitational acceleration vector, ν is kinematic viscosity, and $\boldsymbol{J}$ and $\boldsymbol{B}$ are the current density and magnetic field vectors, respectively.

In the geostrophic momentum equation the Coriolis force on the left-hand side of (29) balances the buoyancy forces on its right-hand side, see for example Le Mouël (1984):

$$2\rho_0(\boldsymbol{\Omega} \times \boldsymbol{u}) = -\boldsymbol{\nabla} p + \rho' \boldsymbol{f_g}. \tag{30}$$

By curling (30) and taking its radial component we get the geostrophic constraint

$$\boldsymbol{\nabla}_{\mathrm{H}} \cdot (\boldsymbol{u}_{\mathrm{H}} \cos\vartheta) = 0 \tag{31}$$

where subscript H reminds us that we only consider the horizontal part of the flow. We undertake solving (25) subject to (31) which implies constraining our numerical system serving to reduce its non-uniqueness.

4.4.1. Relaxing the geostrophic constraint. We *relax* the geostrophic constraint in order to include more of the terms in the Navier–Stokes force balance. In particular, the Lorenz term $\boldsymbol{J} \times \boldsymbol{B}$ is subject to much inquiry in the geomagnetic literature. We first write the left-hand

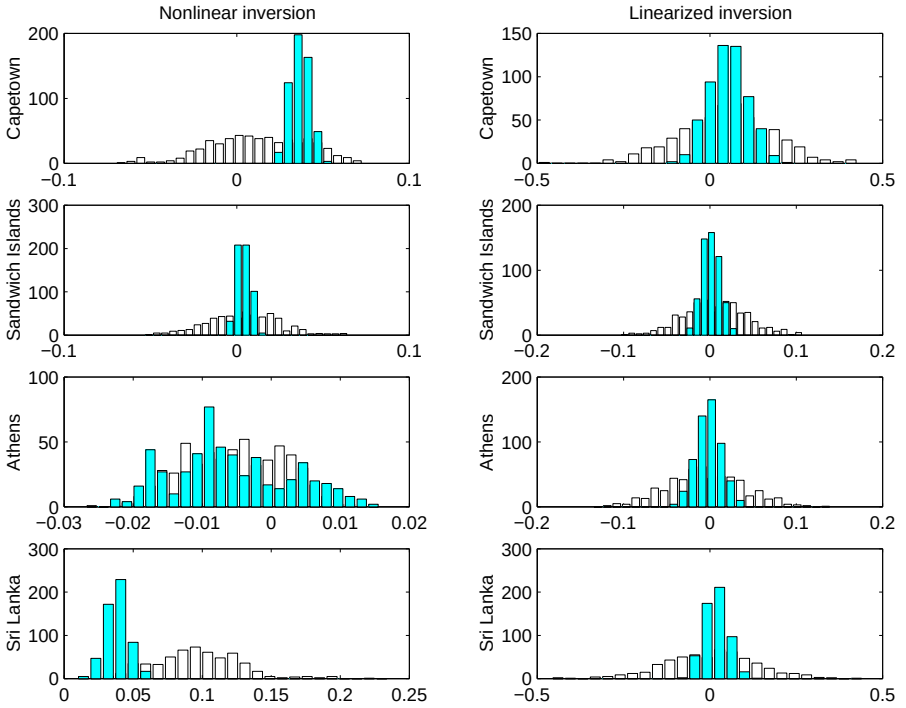


Figure 2. Histograms of the horizontal divergence from 554 models for $\varepsilon = 10^{-1}$ (white bars) and $\varepsilon = 10^{-2}$ (grey bars) from nonlinear inversion (left column) and linear inversion (right column) under Cape Town, Sandwich Islands, Athens and Sri Lanka. Note that the horizontal scale is different in most plots in this figure.

side of (31) as a function of colatitude ϑ and longitude λ, which we in turn separate into a matrix $\mathbf{F}$ and the solution vector $\boldsymbol{v} = (\boldsymbol{t}, \boldsymbol{s})$

$$\boldsymbol{\nabla}_{\mathrm{H}} \cdot (\boldsymbol{u}_{\mathrm{H}} \cos\vartheta) = F(\vartheta, \lambda) = \mathbf{F} \cdot \boldsymbol{v} \tag{32}$$

where the matrix $\boldsymbol{\nabla}_{\mathrm{H}} \cdot (\boldsymbol{u}_{\mathrm{H}} \cos\vartheta)$ is evaluated on a grid of 5 by 5 degrees over the globe.

We now express our *a priori* assumption of near geostrophy (relaxed geostrophy) through a probability density

$$\rho_{\mathbf{F},\varepsilon}(\boldsymbol{v}) = C_4 \cdot \exp\left(-\frac{1}{2}\boldsymbol{v}^{\mathrm{T}} \cdot \frac{\mathbf{F}^{\mathrm{T}}\mathbf{F}}{\varepsilon^2} \cdot \boldsymbol{v}\right) \tag{33}$$

where C_4 is a normalization constant. We use $\rho(\boldsymbol{v})$ as our *a priori* information on the velocity coefficients. As can be seen in (33), ε^2 is treated as a 'variance'. In this paper we present flows with two values of ε, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-1}$.

Owing to the independence of the *a priori* information on the magnetic field coefficients and the velocity field coefficients we can multiply their probability distributions, and the effective *a priori* probability distribution $\rho(\boldsymbol{g}, \boldsymbol{v})$ becomes

$$\rho_{\tilde{g},\mathbf{C}_g,\mathbf{F},\varepsilon}(\boldsymbol{g}, \boldsymbol{v}) = C_5[\exp(-\tfrac{1}{2}(\boldsymbol{g} - \tilde{\boldsymbol{g}})^{\mathrm{T}}\mathbf{C}_g^{-1}(\boldsymbol{g} - \tilde{\boldsymbol{g}}))\exp(-\tfrac{1}{2}\boldsymbol{v}^{\mathrm{T}}\mathbf{C}_v^{-1}\boldsymbol{v})] \tag{34}$$

where C_5 is a normalization constant, $\tilde{\boldsymbol{g}}$ is the vector of coefficients of the observed magnetic field, $\mathbf{C}_g$ is the (diagonal) covariance matrix for the uncertainty in the magnetic field, and the (full) covariance $\mathbf{C}_v$ is the inverse of $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}/\varepsilon^2$ (see equation (33)).

4.5. Sampling the a posteriori distribution

The joint *a posteriori* probability density $\sigma(\boldsymbol{g}, \boldsymbol{v})$ for the considered inverse problem is

$$\sigma_{\tilde{g},\mathbf{C}_g,\tilde{\dot{g}},\mathbf{C}_{\dot{g}},\mathbf{F},\varepsilon}(\boldsymbol{g}, \boldsymbol{v}) = C_6 L_{\tilde{\dot{g}},\mathbf{C}_{\dot{g}}}(\boldsymbol{g}, \boldsymbol{v}) \rho_{\tilde{g},\mathbf{C}_g,\mathbf{F},\varepsilon}(\boldsymbol{g}, \boldsymbol{v}) \tag{35}$$

where C_6 is a normalization constant. We sampled σ for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-1}$ by running the algorithm (18) with 400 000 iterations in each case. A separation of 2000 iterations between saved models was chosen to ensure that they were near-independent samples from the joint posterior. Subvector $\boldsymbol{v}$, extracted from the samples of $(\boldsymbol{g}, \boldsymbol{v})$, are then near-independent samples from the marginal posterior $\sigma_{\tilde{g},\mathbf{C}_g,\tilde{\dot{g}},\mathbf{C}_{\dot{g}},\mathbf{F},\varepsilon}(\boldsymbol{v})$.

4.6. Recasting our problem as an explicit inverse problem

To appreciate the significance of solving the original, unmodified implicit problem, instead of recasting it as an explicit problem, we shall compare results from the two approaches. For this particular problem, a slight modification of the algorithm enables us to analyse the corresponding *linearized*, *explicit* inverse problem where the magnetic field $\boldsymbol{g}$ at the CMB is treated as constant. In this case our model parameters are only the velocity coefficients $\boldsymbol{v} = (\boldsymbol{t}, \boldsymbol{s})$, and the form of the posterior changes as both the likelihood function and the *a priori* probability function are modified

$$L^{\mathrm{lin,ex}}_{\tilde{\dot{g}},\mathbf{C}_{\dot{g}}}(\boldsymbol{v}) = C_7 \exp\{-\tfrac{1}{2}[\tilde{\dot{\boldsymbol{g}}} - \mathbf{A}\boldsymbol{v}]^{\mathrm{T}} \mathbf{C}_{\dot{g}}^{-1} [\tilde{\dot{\boldsymbol{g}}} - \mathbf{A}\boldsymbol{v}]\} \tag{36}$$

$$\rho^{\mathrm{lin,ex}}_{\mathbf{F},\varepsilon}(\boldsymbol{v}) = C_8 \exp(-\tfrac{1}{2}\boldsymbol{v}^{\mathrm{T}} \mathbf{C}_{\boldsymbol{v}}^{-1} \boldsymbol{v}) \tag{37}$$

where C_7 and C_8 are normalization constants. Now the coefficients of the magnetic field are taken to be error free and consequently the design matrix $\mathbf{A}$ becomes constant and no longer dependent on $\boldsymbol{g}$. The form of the linearized likelihood is now Gaussian whereas the real likelihood (28) is of a non-Gaussian form.

4.7. Results and discussion

We present results from solving both the nonlinear, implicit problem and the linearized, explicit problem. The main effect of making the problem linear and explicit is that the overall resolving power decreases. Figure 1 shows the distribution and size of the standard deviations of horizontal divergence of core flow from 554 models. In the results from the nonlinear,

Table 1. Means and standard deviations for horizontal divergence of fluid flow (upwelling and downwelling) for 554 models at four geographical point locations before ($\varepsilon = 10^{-2}$) and after ($\varepsilon = 10^{-1}$) the transition for both nonlinear and linear inversion.

		Mean		Standard deviation	
Location	ε	Nonlinear	Linear	Nonlinear	Linear
Cape Town	10^{-2}	0.0368	0.0535	0.0055	0.0553
Cape Town	10^{-1}	0.0127	0.0451	0.0275	0.1329
Sandwich	10^{-2}	0.0042	0.0012	0.0038	0.0104
Sandwich	10^{-1}	0.0038	0.0058	0.0208	0.0366
Athens	10^{-2}	−0.0058	−0.0015	0.0085	0.0143
Athens	10^{-1}	−0.0054	−0.0054	0.0071	0.0439
Sri Lanka	10^{-2}	0.0383	0.0181	0.0082	0.0336
Sri Lanka	10^{-1}	0.0964	0.0020	0.0307	0.1325

implicit formulation of the problem we see an increase in standard deviation from $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-1}$, corresponding to an increase in uncertainty which in turn reflects the increase in non-uniqueness. In the corresponding solutions to the linearized, explicit problem for respectively $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-1}$ we again see an increase in uncertainty with increasing ε, and furthermore an increased uncertainty over nonlinear, implicit inversion†. Modifying the inverse problem to an linearized, explicit problem in this case results in an increase in the non-uniqueness of the solution. Figure 2 shows four examples of this underneath Cape Town, Sandwich Islands, Athens, and Sri Lanka. Table 1 summarizes the statistics of the distributions shown in figure 2.

References

Backus G E 1968 Kinematics of geomagnetic secular variation in a perfectly conducting core *Phil. Trans. R. Soc. Lond.* A **263** 239–66

Backus G E, Parker R L and Constable C G 1996 *Foundations of Geomagnetism* (New York: Cambridge University Press)

Bloxham J and Jackson A 1991 Fluid flow near the surface of Earth's outer core *Rev. Geophys. Space Phys.* **29** 97–120

——1992 Time-dependent mapping of the magnetic field at the core–mantle boundary *J. Geophys. Res.* **97** 19 537–63

Hulot G, Le Mouël J-L and Wahr J 1991 Taking into account truncation problems and geomagnetic model accuracy in assessing computed flows at the core–mantle boundary *Geophys. J. Int.* **108** 224–46

Jackson A and Bloxham J 1991 Mapping the fluid flow and shear near the core surface using the radial and horizontal components of the magnetic field *Geophys. J. Int.* **105** 199–212

Langel R A 1987 The main field *Geomagnetism* vol 4, ed J A Jacobs (Orlando, FL: Academic)

Le Mouël J L 1984 Outer-core geostrophic flow and secular variation of Earth's magnetic field *Nature* **311** 734–5

Mosegaard K 1998 Resolution analysis of general inverse problems through inverse Monte Carlo sampling *Inverse Problems* **14** 405–26

Mosegaard K and Tarantola A 1995 Monte Carlo sampling of solutions to inverse problems *J. Geophys. Res.* B **100** 12 431–47

Roberts P H 1987 *Origin of the Main Field: Dynamics* (New York: Academic) ch 3

Tarantola A 1987 *Inverse Problem Theory* (New York: Elsevier)

Tarantola A and Valette B 1982 Inverse problems = quest for information *J. Geophys.* **50** 159–70

Whaler K 1986 Geomagnetic evidence for fluid upwelling at the core–mantle boundary *Geophys. J. R. Astron. Soc.* **86** 563–88

† This apparently counter-intuitive result can be understood by considering a simplified version of our problem. A positive datum $\dot{g}$ ('secular variation') is related to another positive datum g ('magnetic field') and a model parameter v ('velocity field') through

$$\dot{g} = gv.$$

For simplicity, let us assume that the data $\dot{g}$ and g are log-normally distributed, so that $\log(\dot{g})$ and $\log(g)$ are normally distributed with variances $\sigma_{\dot{g}}^2$ and σ_g^2, respectively. In this case the distribution of the model parameter v is log-normal, and the variance σ_v^2 of $\log(v)$ satisfies

$$\sigma_{\dot{g}}^2 = \sigma_g^2 + \sigma_v^2.$$

Linearizing this inverse problem corresponds to treating g as a fixed constant, that is, with zero variance. However, since $\sigma_{\dot{g}}^2$ remains fixed, it is seen that in the linearized problem the uncertainty σ_v^2 of v is higher than in the nonlinear case.

16

Probabilistic Approach to Inverse Problems

Klaus Mosegaard
Niels Bohr Institute, Copenhagen, Denmark
Albert Tarantola
Institut de Physique du Globe, Paris, France

1. Introduction

In 'inverse problems' data from indirect measurements are used to estimate unknown parameters of physical systems. Uncertain data (possibly vague) prior information on model parameters, and a physical theory relating the model parameters to the observations are the fundamental elements of any inverse problem. Using concepts from probability theory, a consistent formulation of inverse problems can be made, and, while the most general solution of the inverse problem requires extensive use of Monte Carlo methods, special hypotheses (e.g., Gaussian uncertainties) allow, in some cases, an analytical solution to part of the problem (e.g., using the method of least squares).

1.1 General Comments

Given a physical system, the 'forward' or 'direct' problem consists, by definition, in using a physical theory to predict the outcome of possible experiments. In classical physics this problem has a unique solution. For instance, given a seismic model of the whole Earth (elastic constants, attenuation, etc. at every point inside the Earth) and given a model of a seismic source, we can use current seismological theories to predict which seismograms should be observed at given locations at the Earth's surface.

The 'inverse problem' arises when we do not have a good model of the Earth, or a good model of the seismic source, but we have a set of seismograms, and we wish to use these observations to infer the internal Earth structure or a model of the source (typically we try to infer both).

There are many reasons that make the inverse problem underdetermined (nonunique). In the seismic example, two different Earth models may predict the same seismograms,[1] the finite bandwidth of our data will never allow us to resolve very small features of the Earth model, and there are always experimental uncertainties that allow different models to be 'acceptable.'

The name 'inverse problem' is widely used. The authors of this chapter only like this name moderately, as we see the problem more as a problem of 'conjunction of states of information' (theoretical, experimental, and prior information). In fact, the equations used below have a range of applicability well beyond 'inverse problems': they can be used, for instance, to predict the values of observations in a realistic situation where the parameters describing the Earth model are not 'given' but only known approximately.

We take here a probabilistic point of view. The axioms of probability theory apply to different situations. One is the traditional statistical analysis of random phenomena, another one is the description of (more or less) subjective states of information on a system. For instance, estimation of the uncertainties attached to any measurement usually involves both uses of probability theory: Some uncertainties contributing to the total uncertainty are estimated using statistics, while some other uncertainties are estimated using informed scientific judgment about the quality of an instrument, about effects not explicitly taken into account, etc. The International Organization for Standardization (ISO) in *Guide to the Expression of Uncertainty in Measurement* (1993), recommends that the uncertainties evaluated by statistical methods are named 'type A' uncertainties, and those evaluated by other means (for instance, using Bayesian arguments) be named 'type B' uncertainties. It also recommends that former classifications, for instance into 'random' and 'systematic uncertainties,' should be avoided. In the present text, we accept ISO's basic point of view, and extend it by downplaying the role assigned by ISO to the particular Gaussian model for uncertainties (see Section 4.3) and by not assuming that the uncertainties are 'small.'

INTERNATIONAL HANDBOOK OF EARTHQUAKE AND ENGINEERING SEISMOLOGY, VOLUME 81A ISBN: 0-12-440652-1

In fact, we like to think of an 'inverse' problem as merely a 'measurement.' A measurement that can be quite complex, but the basic principles and the basic equations to be used are the same for a relatively complex 'inverse problem' as for a relatively simple 'measurement.'

We do not normally use, in this text, the term 'random variable,' as we assume that we have probability distributions over 'physical quantities.' This is a small shift in terminology that we hope will not disorient the reader.

An important theme of this paper is *invariant formulation* of inverse problems, in the sense that solutions obtained using different, equivalent, sets of parameters should be consistent, i.e., probability densities obtained as the solution of an inverse problem, using two different set of parameters, should be related through the well-known rule of multiplication by the Jacobian of the transformation.

This chapter is organized as follows. After a brief historical review of inverse problem theory, with special emphasis on seismology, we give a short introduction to probability theory. In addition to being a tutorial, this introduction also aims at fixing a serious problem of classical probability, namely the noninvariant definition of conditional probability. This problem, which materializes in the so-called Borel paradox, has profound consequences for inverse problem theory.

A probabilistic formulation of inverse theory for general inverse problems (usually called 'nonlinear inverse problems') is not complete without the use of Monte Carlo methods. Section 3 is an introduction to the most versatile of these methods, the Metropolis sampler. Apart from being versatile, it also turns out to be the most natural method for implementing our probabilistic approach.

In Sections 4, 5, and 6 time has come for applying probability theory and Monte Carlo methods to inverse problems. All the steps of a careful probabilistic formulations are described, including parametrization, prior information over the parameters, and experimental uncertainties. The hitherto overlooked problem of uncertain physical laws ('forward relations') is given special attention in this text, and it is shown how this problem is profoundly linked to the resolution of the Borel paradox.

Section 7 treats the special case of the mildly nonlinear inverse problems, where deterministic (non-Monte Carlo) methods can be employed. In this section, invariant forms of classical inversion formulas are given.

1.2 Brief Historical Review

For a long time scientists have estimated parameters using optimization techniques. Laplace explicitly stated the least absolute values criterion. This, and the least-squares criterion were later popularized by Gauss (1809). While Laplace and Gauss were mainly interested in overdetermined problems, Hadamard (1902, 1932) introduced the notion of an 'ill-posed problem,' which can be viewed in many cases as an underdetermined problem.

The late 1960s and early 1970s were a golden age for the theory of inverse problems. In this period the first uses of Monte Carlo theory to obtain Earth models were made by Keilis-Borok and Yanovskaya (1967) and by Press (1968). At about the same time, Backus and Gilbert, and Backus alone, in the years 1967–1970, made original contributions to the theory of inverse problems, focusing on the problem of obtaining an unknown *function* from discrete data. Although the resulting mathematical theory is elegant, its initial predominance over the more 'brute force' (but more powerful) Monte Carlo theory was only possible due to the quite limited capacities of the computers at that time. It is our feeling that Monte Carlo methods will play a more important role in the future (and this is the reason why we put emphasis on these methods in this chapter). An investigation of the connection between analog models, discrete models, and Monte Carlo models can be found in a paper by Kennett and Nolet (1978).

Important developments of inverse theory in the fertile period around 1970 were also made by Wiggins (1969), with his method of suppressing 'small eigenvalues,' and by Franklin (1970) by introducing the right mathematical setting for the Gaussian, functional (i.e., infinite dimensional) inverse problem (see also Lehtinen *et al.*, 1989). Other important papers from the period are those of Gilbert (1971) and Wiggins (1972).

A reference that may interest some readers is Parzen *et al.* (1998), where the probabilistic approach of Akaike is described.

To the 'regularizing techniques' of Tikhonov (1963), Levenberg (1944), and Marquardt (1970), we prefer, in this chapter, the approach where the a priori information is used explicitly.

For seismologists, the first bona fide solution of an inverse problem was the estimation of the hypocenter coordinates of an earthquake using the 'Geiger method' (Geiger, 1910), which present-day computers have made practical. In fact, seismologists have been the originators of the theory of inverse problems (for data interpretation), and this is because the problem of understanding the structure of the Earth's interior using only surface data is a difficult one.

3-D tomography of the Earth, using travel times of seismic waves, was developed by Keiiti Aki and his coworkers in a couple of well known papers (Aki and Lee, 1976; Aki, Christofferson and Husebye 1977). Minster and Jordan (1978) applied the theory of inverse problems to the reconstruction of the tectonic plate motions, introducing the concept of 'data importance.' Later, tomographic studies have provided spectacular images of the Earth's interior. Interesting papers on these inversions are by van der Hilst *et al.* (1997) and Su *et al.* (1992).

One of the major current challenges in seismic inversion is the nonlinearity of wave field inversions. This is accentuated by the fact that major experiments in the future most likely will allow us to sample the whole seismic wave field. For low frequencies, wave field inversion is linear. Dahlen (1976)

investigated the influence of lateral heterogeneity on the free oscillations. He showed that the inverse problem of estimating lateral heterogeneity of even degree from multiplet variance and skewance is linear. At the time this was published, data accuracy and unknown ellipticity splitting parameters hindered its application to real data, but later developments, including the works of Woodhouse and Dahlen (1978) on discontinuous Earth models, led to present-day successful inversions of low-frequency seismograms. In this connection the works of Woodhouse, Dziewonski, and others spring to mind.[2] Later, the first attempts to go to higher frequencies and nonlinear inversion were made by Nolet *et al.* (1986), and Nolet (1990).

Purely probabilistic formulations of inverse theory saw the light around 1970 (see, for instance, Kimeldorf and Wahba, 1970). In an interesting paper, Rietsch (1977) made non-trivial use of the notion of a 'noninformative' prior distribution for positive parameters. Jackson (1979) explicitly introduced prior information in the context of linear inverse problems, an approach that was generalized by Tarantola and Valette (1982a,b) to nonlinear problems.

There are three monographs in the area of inverse problems (from the viewpoint of data interpretation). In Tarantola (1987), the general, probabilistic formulation for nonlinear inverse problems is proposed. The small book by Menke (1984) covers several viewpoints on discrete, linear, and nonlinear inverse problems, and is easy to read. Finally, Parker (1994) exposes his view of the general theory of linear problems.

Recently, the interest in Monte Carlo methods, for the solution of inverse problems, has been increasing. Mosegaard and Tarantola (1995) proposed a generalization of the Metropolis algorithm (Metropolis *et al.*, 1953) for analysis of general inverse problems, introducing explicitly prior probability distributions, and they applied the theory to a synthetic numerical example. Monte Carlo analysis was recently applied to real data inverse problems by Mosegaard *et al.* (1997), Dahl-Jensen *et al.* (1998), Mosegaard and Rygaard-Hjalsted (1999), and Khan *et al.* (2000).

2. Elements of Probability

Probability theory is essential to our formulation of inverse theory. This chapter therefore contains a review of important elements of probability theory, with special emphasis on results that are important for the analysis of inverse problems. Of particular importance is our explicit introduction of *distance* and *volume* in data and model spaces. This has profound consequences for the notion of *conditional probability density*, which plays an important role in probabilistic inverse theory.

Also, we replace the concept of conditional probability by the more general notion of 'conjunction' of probabilities, this allowing us to address the more general problem where not only the data, but also the physical laws, are uncertain.

2.1 Volume

Let us consider an abstract space $\mathcal{S}$, where a point $\mathbf{x}$ is represented by some coordinates $\{x^1, x^2, \ldots\}$, and let $\mathcal{A}$ be some region (subspace) of $\mathcal{S}$. The measure associating a volume $V(\mathcal{A})$ to any region $\mathcal{A}$ of $\mathcal{S}$ will be denoted the *volume measure*

$$V(\mathcal{A}) = \int_{\mathcal{A}} d\mathbf{x}\, v(\mathbf{x}) \ , \tag{1}$$

where the function $v(\mathbf{x})$ is the *volume density*, and where we write $d\mathbf{x} = dx^1\, dx^2 \ldots$ The *volume element* is then[3]

$$dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x} \ , \tag{2}$$

and we may write $V(\mathcal{A}) = \int_{\mathcal{A}} dV(\mathbf{x})$. A manifold is called a *metric manifold* if there is a definition of distance between points, such that the distance ds between the point of coordinates $\{x^i\}$ and the point of coordinates $\{x^i + dx^i\}$ can be expressed as[4]

$$ds^2 = g_{ij}(\mathbf{x})\, dx^i\, dx^j \ , \tag{3}$$

i.e., if the notion of distance is 'of the L_2 type.'[5] The matrix whose entries are g_{ij} is the *metric matrix*, and an important result of differential geometry and integration theory is that the volume density of the space, $v(\mathbf{x})$, equals the square root of the determinant of the metric:

$$v(\mathbf{x}) = \sqrt{\det \mathbf{g}(\mathbf{x})} \ . \tag{4}$$

Example 1. *In the Euclidean 3D space, using spherical coordinates, the distance element is* $ds^2 = dr^2 + r^2\, d\theta^2 + r^2 \sin^2\theta\, d\varphi^2$, *from which it follows that the metric matrix is*

$$\begin{pmatrix} g_{rr} & g_{r\theta} & g_{r\varphi} \\ g_{\theta r} & g_{\theta\theta} & g_{\theta\varphi} \\ g_{\varphi r} & g_{\varphi\theta} & g_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} . \tag{5}$$

The volume density equals the metric determinant $v(r, \theta, \varphi) = \sqrt{\det \mathbf{g}(r, \theta, \varphi)} = r^2 \sin\theta$ *and therefore the volume element is* $dV(r, \vartheta, \varphi) = v(r, \vartheta, \varphi)\, dr\, d\vartheta\, d\varphi = r^2 \sin\theta\, dr\, d\vartheta\, d\varphi$.

2.2 Probability

Assume that we have defined over the space, not only the volume $V(\mathcal{A})$ of a region $\mathcal{A}$ of the space, but also its *probability* $P(\mathcal{A})$, which is assumed to satisfy the Kolmogorov axioms (Kolmogorov, 1933). This probability is assumed to be descriptible in terms of a probability density $f(x)$ through the expression

$$P(\mathcal{A}) = \int_{\mathcal{A}} d\mathbf{x}\ f(\mathbf{x}) \ . \tag{6}$$

It is well known that, in a change of coordinates over the space, a probability density changes its value: it is multiplied

by the Jacobian of the transformation (this is the *Jacobian rule*). Normally, the probability of the whole space is normalized to one. If it is not normalizable, we do not say that we have a probability, but a 'measure.' We can state here the following postulate.

Postulate 1. *Given a space $\mathcal{X}$ over which a volume measure $V(\cdot)$ is defined. Any other measure (normalizable or not) $M(\cdot)$ considered over $\mathcal{X}$ is* absolutely continuous *with respect to $V(\cdot)$, i.e., the measure $M(\mathcal{A})$ of any region $\mathcal{A} \subset \mathcal{X}$ with vanishing volume must be zero: $V(\mathcal{A}) = 0 \Rightarrow M(\mathcal{A}) = 0$.*

2.3 Homogeneous Probability Distributions

In some parameter spaces, there is an obvious definition of distance between points, and therefore of volume. For instance, in the 3D Euclidean space the distance between two points is just the Euclidean distance (which is invariant under translations and rotations). Should we choose to parametrize the position of a point by its Cartesian coordinates $\{x, y, z\}$, the volume element in the space would be $dV(x, y, z) = dx\, dy\, dz$, while if we choose to use geographical coordinates, the volume element would be $dV(r, \theta, \varphi) = r^2 \sin\theta\, dr\, d\vartheta\, d\varphi$.

Definition. *The* homogeneous probability distribution *is the probability distribution that assigns to each region of the space a probability proportional to the volume of the region.*

Then, which probability density represents such a homogeneous probability distribution? Let us give the answer in three steps.

- If we use Cartesian coordinates $\{x, y, z\}$, as we have $dV(x, y, z) = dx\, dy\, dz$, the probability density representing the homogeneous probability distribution is constant: $f(x, y, z) = k$.
- If we use geographical coordinates $\{r, \theta, \varphi\}$, as we have $dV(r, \theta, \varphi) = r^2 \sin\theta\, dr\, d\theta\, d\varphi$, the probability density representing the homogeneous probability distribution is $g(r, \theta, \varphi) = kr^2 \sin\theta$.
- Finally, if we use an arbitrary system of coordinates $\{u, v, w\}$, in which the volume element of the space is $dV(u, v, w) = v(u, v, w)\, du\, dv\, dw$, the homogeneous probability distribution is represented by the probability density $h(u, v, w) = kv(u, v, w)$.

This is obviously true, since if we calculate the probability of a region $\mathcal{A}$ of the space, with volume $V(\mathcal{A})$, we get a number proportional to $V(\mathcal{A})$.

From these observations we can arrive at conclusions that are of general validity. First, the homogeneous probability distribution over some space is represented by a constant probability density **only** if the space is flat (in which case rectilinear systems of coordinates exist) and if we use Cartesian (or rectilinear) coordinates. The other conclusions can be stated as rules:

Rule 1. *The probability density representing the homogeneous probability distribution is easily obtained if the expression of the volume element $dV(u_1, u_2, \dots) = v(u_1, u_2, \dots)\, du_1\, du_2 \dots$ of the space is known, as it is then given by $h(u_1, u_2, \dots) = kv(u_1, u_2, \dots)$, where k is a proportionality constant (that may have physical dimensions).*

Rule 2. *If there is a metric $g_{ij}(u_1, u_2, \dots)$ in the space, then the volume element is given by $dV(u_1, u_2, \dots) = \sqrt{\det \mathbf{g}(u_1, u_2, \dots)}\, du_1\, du_2 \cdots$, i.e., we have $v(u_1, u_2, \dots) = \sqrt{\det \mathbf{g}(u_1, u_2, \dots)}$. The probability density representing the homogeneous probability distribution is, then, $h(u_1, u_2, \dots) = k\sqrt{\det \mathbf{g}(u_1, u_2, \dots)}$.*

Rule 3. *If the expression of the probability density representing the homogeneous probability distribution is known in one system of coordinates, then it is known in any other system of coordinates, through the Jacobian rule.*

Indeed, in the expression above, $g(r, \theta, \varphi) = kr^2 \sin\theta$, we recognize the Jacobian between the geographical and the Cartesian coordinates (where the probability density is constant).

For short, when we say *the homogeneous probability density* we mean *the probability density representing the homogeneous probability distribution*. **One should remember that, in general, the homogeneous probability density is *not* constant.**

Let us now examine 'positive parameters,' like a temperature, a period, or a seismic wave propagation velocity. One of the properties of the parameters we have in mind is that they occur in pairs of mutually reciprocal parameters:

Period	$T = 1/\nu$	;	*Frequency*	$\nu = 1/T$
Resistivity	$\rho = 1/\sigma$	;	Conductivity	$\sigma = 1/\rho$
Temperature	$T = 1/(k\beta)$	;	Thermodynamic parameter	$\beta = 1/(kT)$
Mass density	$\rho = 1/\ell$	;	Lightness	$\ell = 1/\rho$
Compressibility	$\gamma = 1/\kappa$	;	Bulk modulus (uncompressibility)	$\kappa = 1/\gamma$
Wave velocity	$c = 1/n$	;	Wave slowness	$n = 1/c$.

When working with physical theories, one may freely choose one of these parameters or its reciprocal.

Sometimes these pairs of equivalent parameters come from a definition, like when we define frequency ν as a function of the period T, by $\nu = 1/T$. Sometimes these parameters arise when analyzing an idealized physical system. For instance, Hooke's law, relating stress $\sigma_{\ell j}$ to strain $\varepsilon_{\ell j}$ can be expressed as $\sigma_{ij} = c_{ij}{}^{k\ell}\varepsilon_{k\ell}$, thus introducing the stiffness tensor $c_{ijk\ell}$, or as $\varepsilon_{ij} = d_{ij}{}^{k\ell}\sigma_{k\ell}$, thus introducing the compliance tensor

$d_{ijk\ell}$, the inverse of the stiffness tensor. Then the respective eigenvalues of these two tensors belong to the class of scalars analyzed here.

Let us take, as an example, the pair conductivity–resistivity (which may be thermal, electric, etc.). Assume we have two samples in the laboratory S_1 and S_2 whose resistivities are respectively ρ_1 and ρ_2. Correspondingly, their conductivities are $\sigma_1 = 1/\rho_1$ and $\sigma_2 = 1/\rho_2$. How should we define the 'distance' between the 'electrical properties' of the two samples? As we have $|\rho_2 - \rho_1| \neq |\sigma_2 - \sigma_1|$, choosing one of the two expressions as the 'distance' would be arbitrary. Consider the following definition of 'distance' between the two samples:

$$D(S_1, S_2) = \left| \log \frac{\rho_2}{\rho_1} \right| = \left| \log \frac{\sigma_2}{\sigma_1} \right| . \tag{7}$$

This definition (i) treats symmetrically the two equivalent parameters ρ and σ and, more importantly, (ii) has an *invariance of scale* (what matters is how many 'octaves' we have between the two values, not the plain difference between the values). In fact, it is the only definition of distance between the two samples S_1 and S_2 that has an invariance of scale and is additive (i.e., $D(S_1, S_2) + D(S_2, S_3) = D(S_1, S_3)$).

Associated to the distance $D(x_1, x_2) = |\log(x_2/x_1)|$ is the distance element (differential form of the distance)

$$dL(x) = \frac{dx}{x} . \tag{8}$$

This being a 'one-dimensional volume,' we can now apply Rule 1 above to get the expression of the homogeneous probability density for such a positive parameter:

$$f(x) = \frac{k}{x} . \tag{9}$$

Defining the reciprocal parameter $y = 1/x$ and using the Jacobian rule, we arrive at the homogeneous probability density for y:

$$g(y) = \frac{k}{y} . \tag{10}$$

These two probability densities have the same form: the two reciprocal parameters are treated symmetrically. Introducing the logarithmic parameters

$$x^* = \log \frac{x}{x_0} ; \qquad y^* = \log \frac{y}{y_0} , \tag{11}$$

where x_0 and y_0 are arbitrary positive constants, and using the Jacobian rule, we arrive at the homogeneous probability densities:

$$f'(x^*) = k ; \qquad g'(y^*) = k . \tag{12}$$

This shows that the logarithm of a positive parameter (of the type considered above) is a 'Cartesian' parameter. In fact, it is the consideration of Eqs. (12), together with the Jacobian rule, that allows full understanding of the (homogeneous) probability densities (9) and (10).

The association of the probability density $f(u) = k/u$ with positive parameters was first made by Jeffreys (1939). To honor him, we propose to use the term *Jeffreys parameters* for all the parameters of the type considered above. The $1/u$ probability density was advocated by Jaynes (1968), and a nontrivial use of it was made by Rietsch (1977) in the context of inverse problems.

Rule 4. *The homogeneous probability density for a Jeffreys quantity u is $f(u) = k/u$.*

Rule 5. *The homogeneous probability density for a 'Cartesian parameter' u (like the logarithm of a Jeffreys parameter, an actual Cartesian coordinate in an Euclidean space, or the Newtonian time coordinate) is $f(u) = k$. The homogeneous probability density for an angle describing the position of a point in a circle is also constant.*

If a parameter u is a Jeffreys parameter with the homogeneous probability density $f(u) = k/u$, then its inverse, its square, and, in general, any power of the parameter is also a Jeffreys parameter, as it can easily be seen using the Jacobian rule.

Rule 6. *Any power of a Jeffreys quantity (including its inverse) is a Jeffreys quantity.*

It is important to recognize when we do *not* face a Jeffreys parameter. Among the many parameters used in the literature to describe an isotropic linear elastic medium we find parameters like the Lamé's coefficients λ and μ, the bulk modulus κ, the Poisson ratio σ, etc. A simple inspection of the theoretical range of variation of these parameters shows that the first Lamé parameter λ and the Poisson ratio σ may take negative values, so they are certainly not Jeffreys parameters. In contrast, Hooke's law $\sigma_{ij} = c_{ijk\ell}\, \varepsilon^{k\ell}$, defining a linearity between stress σ_{ij} and strain ε_{ij}, defines the positive definite stiffness tensor $c_{ijk\ell}$ or, if we write $\varepsilon_{ij} = d_{ijk\ell}\, \sigma^{k\ell}$, defines its inverse, the compliance tensor $d_{ijk\ell}$. The two reciprocal tensors $c_{ijk\ell}$ and $d_{ijk\ell}$ are 'Jeffreys tensors.' This is a notion whose development is beyond the scope of this paper, but we can give the following rule.

Rule 7. *The eigenvalues of a Jeffreys tensor are Jeffreys quantities.*[6]

As the two (different) eigenvalues of the stiffness tensor $c_{ijk\ell}$ are $\lambda_\kappa = 3\kappa$ (with multiplicity 1) and $\lambda_\mu = 2\mu$ (with multiplicity 5), we see that the incompressibility modulus κ and the shear modulus μ are Jeffreys parameters[7] (as are any parameters proportional to them, or any powers of them, including the inverses). If, for some reason, instead of working with κ and μ, we wish to work with other elastic parameters, for instance, the Young modulus Y and the

Poisson ratio σ, or the two elastic wave velocities, then the homogeneous probability distribution must be found using the Jacobian of the transformation (see Appendix H).

Some probability densities have conspicuous 'dispersion parameters,' like the σ's in the normal probability density $f(x) = k \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right)$, in the log-normal probability $g(X) = \frac{k}{X} \exp\left(-\frac{(\log X/X_0)^2}{2\sigma^2}\right)$ or in the Fisher probability density (Fisher, 1953) $h(\vartheta, \varphi) = k \sin\theta \exp(\cos\theta/\sigma^2)$. A consistent probability model requires that when the dispersion parameter σ tends to infinity, the probability density tends to the homogeneous probability distribution. For instance, in the three examples just given, $f(x) \to k$, $g(X) \to k/X$, and $h(\theta, \varphi) \to k \sin\theta$, which are the respective homogeneous probability densities for a Cartesian quantity, a Jeffreys quantity, and the geographical coordinates on the surface of the sphere. We can state the following rule.

Rule 8. *If a probability density has some 'dispersion parameters,' then, in the limit where the dispersion parameters tend to infinity, the probability density must tend to the homogeneous one.*

As an example, using the normal probability density $f(x) = k \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right)$, for a Jeffreys parameter is not consistent. Note that it would assign a finite probability to negative values of a positive parameter that, by definition, is positive. More technically, this would violate our Postulate 1. Using the log-normal probability density for a Jeffreys parameter is consistent.

There is a problem of terminology in the Bayesian literature. The homogeneous probability distribution is a very special distribution. When the problem of selecting a 'prior' probability distribution arises in the absence of any information, except the fundamental symmetries of the problem, one may select as prior probability distribution the homogeneous distribution. But enthusiastic Bayesians do not call it 'homogeneous,' but 'noninformative.' We cannot recommend using this terminology. The homogeneous probability distribution is as informative as any other distribution, it is just the homogeneous one (see Appendix D).

In general, each time we consider an abstract parameter space, each point being represented by some parameters $\mathbf{x} = \{x^1, x^2 \dots x^n\}$, we will start by solving the (sometimes nontrivial) problem of defining a distance between points that respects the necessary symmetries of the problem. Only exceptionally this distance will be a quadratic expression of the parameters (coordinates) being used (i.e., only exceptionally our parameters will correspond to 'Cartesian coordinates' in the space). From this distance, a volume element $dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x}$ will be deduced, from where the expression $f(\mathbf{x}) = k\, v(\mathbf{x})$ of the homogeneous probability density will follow. Sometimes, we can directly define the volume element, without the need of a distance. We emphasize the need of defining a distance—or a volume element—in the parameter space, from which the notion of homogeneity will follow. With this point of view, we slightly depart from the original work by Jeffreys and Jaynes.

2.4 Conjunction of Probabilities

We shall here consider two probability distributions P and Q. We say that a probability R is a product of the two given probabilities, and is denoted $(P \wedge Q)$ if

- $P \wedge Q = Q \wedge P$;
- for any subset $\mathcal{A}$, $(P \wedge Q)(\mathcal{A}) \neq 0 \Rightarrow P(\mathcal{A}) \neq 0$ and $Q(\mathcal{A}) \neq 0$;
- if M denotes the homogeneous probability distribution, then $P \wedge M = P$.

The realization of these conditions leading to the simplest results can easily be expressed using probability densities (see Appendix G for details). If the two probabilities P and Q are represented by the two probability densities $p(\mathbf{x})$ and $q(\mathbf{x})$, respectively, and if the homogeneous probability density is represented by $\mu(\mathbf{x})$, then the probability $P \wedge Q$ is represented by a probability density, denoted $(p \wedge q)(\mathbf{x})$, that is given by

$$(p \wedge q)(\mathbf{x}) = k \frac{p(\mathbf{x})\, q(\mathbf{x})}{\mu(\mathbf{x})} \ , \tag{13}$$

where k is a normalization constant.[8]

The two left columns of Figure 1 represent these probability densities.

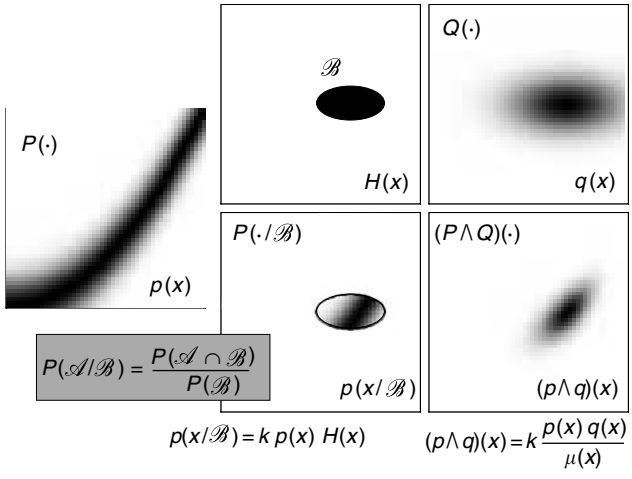


FIGURE 1 The two left columns of the figure illustrate the definition of conditional probability (see text for details). The right of the figure explains that the definition of the AND operation is a generalization of the notion of conditional probability. While a conditional probability combines a probability distribution $P(\cdot)$ with an 'event' $\mathscr{B}$, the AND operation combines two probability distributions $P(\cdot)$ and $Q(\cdot)$ defined over the same space. See text for a detailed explanation.

Example 2. *On the surface of the Earth, using geographical coordinates (latitude ϑ and longitude φ), the homogeneous probability distribution is represented by the probability density $\mu(\vartheta, \varphi) = \frac{1}{4\pi} \cos\vartheta$. An estimation of the position of a floating object at the surface of the sea by an airplane navigator gives a probability distribution for the position of the object corresponding to the probability density $p(\vartheta, \varphi)$, and an independent, simultaneous estimation of the position by another airplane navigator gives a probability distribution corresponding to the probability density $q(\vartheta, \varphi)$. How do we 'combine' the two probability densities $p(\vartheta, \varphi)$ and $q(\vartheta, \varphi)$ to obtain a 'resulting' probability density? The answer is given by the conjunction of the two probability densities:*

$$(p \wedge q)(\vartheta, \varphi) = k \, \frac{p(\vartheta, \varphi)\, q(\vartheta, \varphi)}{\mu(\vartheta, \varphi)} \ . \tag{14}$$

We emphasize here the following:

Example 2 is at the basis of the paradigm that we use below to solve inverse problems.

More generally, the conjunction of the probability densities $f_1(\mathbf{x})$, $f_2(\mathbf{x}) \dots$ is

$$\begin{aligned} h(\mathbf{x}) &= (\, f_1 \wedge f_2 \wedge f_3 \cdots)(\mathbf{x}) \cdots \\ &= k \, \mu(\mathbf{x}) \, \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \, \frac{f_3(\mathbf{x})}{\mu(\mathbf{x})} \cdots \ . \end{aligned} \tag{15}$$

For a formalization of the notion of conjunction of probabilities, the reader is invited to read Appendix G.

2.5 Conditional Probability Density

Given a probability distribution over a space $\mathcal{X}$, represented by the probability density $f(\mathbf{x})$, and given a subspace $\mathcal{B}$ of $\mathcal{X}$ of lower dimension, can we, in a consistent way, infer a probability distribution over $\mathcal{B}$, represented by a probability density $f(\mathbf{x}|\mathcal{B})$ (to be named the conditional probability density 'given $\mathcal{B}$')? The answer is: Using only the elements given, NO, THIS IS NOT POSSIBLE.

The usual way to induce a probability distribution on a subspace of lower dimension is to assign a 'thickness' to the subspace $\mathcal{B}$, to apply the general definition of conditional probability (this time to a region of $\mathcal{X}$, not to a subspace of it) and to take the limit when the 'thickness' tends to zero. But, as suggested in Figure 2, there are infinitely many ways to take this limit, each defining a different 'conditional probability density' on $\mathcal{B}$. Among the infinitely many ways to define a conditional probability density, there is one that is based on the notion of distance between points in the space, and therefore corresponds to an intrinsic definition (see Fig. 2).

Assume that the space $\mathcal{U}$ has p dimensions, the space $\mathcal{V}$ has q dimensions, and define in the $(p+q)$-dimensional space $\mathcal{X} = (\mathcal{U}, \mathcal{V})$ a p-dimensional subspace by the p relations

$$\begin{aligned} v_1 &= v_1(u_1, u_2, \dots, u_p) \\ v_2 &= v_2(u_1, u_2, \cdots, u_p) \\ \cdots &= \cdots \\ v_q &= v_q(u_1, u_2, \dots, u_p) \ . \end{aligned} \tag{16}$$

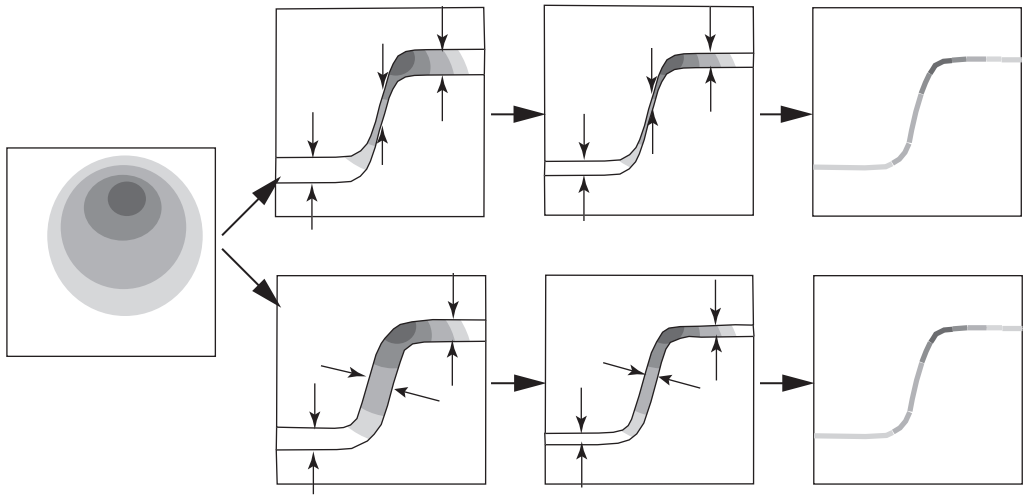


FIGURE 2 An original 2D probability density, and two possible ways (among many) of defining a region of the space whose limit is a given curve. At the top is the 'vertical' limit, while at the bottom is the normal (or orthogonal) limit. Each possible limit defines a different 'induced' or 'conditional' probability density. Only the orthogonal limit gives an intrinsic definition (i.e., a definition invariant under any change of variables). It is, therefore, the only one examined in this work.

The restriction of a probability distribution represented by the probability density $f(\mathbf{x}) = f(\mathbf{u}, \mathbf{v})$ into the subspace defined by the constraint $\mathbf{v} = \mathbf{v}(\mathbf{u})$, can be defined with all generality when it is assumed that we have a metric defined over the $(p+q)$-dimensional space $\mathcal{X} = (\mathcal{U}, \mathcal{V})$. Let us limit here to the special circumstance (useful for a vast majority of inverse problems[9]) where there the $(p+q)$-dimensional space $\mathcal{X}$ is built as the Cartesian product of $\mathcal{U}$ and $\mathcal{V}$ (then we write, as usual, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$). In this case, there is a metric $\mathbf{g}_u$ over $\mathcal{U}$, with associated volume element $dV_u(\mathbf{u}) = \sqrt{\det \mathbf{g}_u}\, d\mathbf{u}$, there is a metric $\mathbf{g}_v$ over $\mathcal{V}$, with associated volume element $dV_v(\mathbf{v}) = \sqrt{\det \mathbf{g}_v}\, d\mathbf{v}$, and the global volume element is simply $dV(\mathbf{u}, \mathbf{v}) = dV_u(\mathbf{u})\, dV_v(\mathbf{v})$.

The restriction of the probability distribution represented by the probability density $f(\mathbf{u}, \mathbf{v})$ on the subspace $\mathbf{v} = \mathbf{v}(\mathbf{u})$ (i.e., the conditional probability density given $\mathbf{v} = \mathbf{v}(\mathbf{u})$) is a probability distribution *on* the submanifold $\mathbf{v} = \mathbf{v}(\mathbf{u})$. We could choose ad-hoc coordinates over this manifold, but as there is a one-to-one correspondence between the coordinates $\mathbf{u}$ and the points on the manifold, the conditional probability density can be expressed using the coordinates $\mathbf{u}$. The restriction of $f(\mathbf{u}, \mathbf{v})$ over the submanifold $\mathbf{v} = \mathbf{v}(\mathbf{u})$ defines the probability density (see Appendix B for the more general case)

$$f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u})) = k\, f(\mathbf{u}, \mathbf{v}(\mathbf{u}))\, \frac{\sqrt{\det\left(\mathbf{g}_u + \mathbf{V}^T \mathbf{g}_v \mathbf{V}\right)}}{\sqrt{\det \mathbf{g}_u}\, \sqrt{\det \mathbf{g}_v}} \Bigg|_{\mathbf{v} = \mathbf{v}(\mathbf{u})} , \tag{17}$$

where k is a normalizing constant, and where $\mathbf{V} = \mathbf{V}(\mathbf{u})$ is the matrix of partial derivatives (see Appendix M for a simple explicit calculation of such partial derivatives)

$$\begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & \cdots & V_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ V_{q1} & V_{q2} & \cdots & V_{qp} \end{pmatrix} = \begin{pmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_2}{\partial u_2} & \cdots & \frac{\partial v_2}{\partial u_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_q}{\partial u_1} & \frac{\partial v_q}{\partial u_2} & \cdots & \frac{\partial v_q}{\partial u_p} \end{pmatrix} . \tag{18}$$

Example 3. *If the hypersurface* $\mathbf{v} = \mathbf{v}(\mathbf{u})$ *is defined by a constant value of* $\mathbf{v}$, *say* $\mathbf{v} = \mathbf{v}_0$, *then Eq. (17) reduces to*

$$f_{u|v}(\mathbf{u}|\mathbf{v} = \mathbf{v}_0) = k\, f(\mathbf{u}, \mathbf{v}_0) = \frac{f(\mathbf{u}, \mathbf{v}_0)}{\int_{\mathcal{U}} d\mathbf{u}\; f(\mathbf{u}, \mathbf{v}_0)} . \tag{19}$$

Elementary definitions of conditional probability density are not based on this notion of distance-based uniform convergence, but use other, ill-defined limits. This is a mistake that, unfortunately, pollutes many scientific works. See Appendix P, in particular, for a discussion on the 'Borel paradox.'

Equation (17) defines the conditional $f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u}))$. Should the relation $\mathbf{v} = \mathbf{v}(\mathbf{u})$ be invertible, it would correspond to a change of variables. It is then possible to show that the alternative conditional $f_{v|u(v)}(\mathbf{v}|\mathbf{u} = \mathbf{u}(\mathbf{v}))$ is related to $f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u}))$ through the Jacobian rule. This is a property that elementary definitions of conditional probability do not share.

2.6 Marginal Probability Density

In the special circumstance described above, where we have a Cartesian product of two spaces, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$, given a 'joint' probability density $f(\mathbf{u}, \mathbf{v})$, it is possible to give an intrinsic sense to the definitions

$$f_u(\mathbf{u}) = \int_{\mathcal{V}} d\mathbf{v}\; f(\mathbf{u}, \mathbf{v}) ; \qquad f_v(\mathbf{v}) = \int_{\mathcal{U}} d\mathbf{u}\; f(\mathbf{u}, \mathbf{v}) . \tag{20}$$

These two densities are called *marginal probability densities*. Their intuitive interpretation is clear, as the 'projection' of the joint probability density respectively over $\mathcal{U}$ and over $\mathcal{V}$.

2.7 Independence and Bayes Theorem

Dropping the index 0 in Eq. (19) and using the second of Eqs. (20) gives

$$f_{u|v}(\mathbf{u}|\mathbf{v}) = \frac{f(\mathbf{u}, \mathbf{v})}{f_v(\mathbf{v})} , \tag{21}$$

or, equivalently, $f(\mathbf{u}, \mathbf{v}) = f_{u|v}(\mathbf{u}|\mathbf{v})\, f_v(\mathbf{v})$. As we can also define $f_{v|u}(\mathbf{v}|\mathbf{u})$, we have the two equations

$$\begin{aligned} f(\mathbf{u}, \mathbf{v}) &= f_{u|v}(\mathbf{u}|\mathbf{v})\; f_v(\mathbf{v}) \\ f(\mathbf{u}, \mathbf{v}) &= f_{v|u}(\mathbf{v}|\mathbf{u})\; f_u(\mathbf{u}) , \end{aligned} \tag{22}$$

that can be read as follows: 'When we work in a space that is the Cartesian product $\mathcal{U} \times \mathcal{V}$ of two subspaces, a joint probability density can always be expressed as the product of a conditional times a marginal.'

From these last equations there follows the expression

$$f_{u|v}(\mathbf{u}|\mathbf{v}) = \frac{f_{v|u}(\mathbf{v}|\mathbf{u})\; f_u(\mathbf{u})}{f_v(\mathbf{v})} , \tag{23}$$

known as the *Bayes theorem*, and generally used as the starting point for solving inverse problems. We do not think this is a useful setting, and we prefer here *not* to use the Bayes theorem (or, more precisely, not to use the intuitive paradigm usually associated with it).

It also follows from Eqs. (22) that the two conditions

$$f_{u|v}(\mathbf{u}|\mathbf{v}) = f_u(\mathbf{u}) ; \qquad f_{v|u}(\mathbf{v}|\mathbf{u}) = f_v(\mathbf{v}) \tag{24}$$

are equivalent. It is then said that **u** and **v** are *independent parameters* (with respect to the probability density $f(\mathbf{u}, \mathbf{v})$). The term 'independent' is easy to understand, as the conditional of any of the two (vector) variables, given the other variable equals the (unconditional) marginal of the variable. Then, one clearly has

$$f(\mathbf{u}, \mathbf{v}) = f_u(\mathbf{u})\ f_v(\mathbf{v}) \tag{25}$$

i.e., for independent variables, the joint probability density can be simply expressed as the product of the two marginals.

3. Monte Carlo Methods

When a probability distribution has been defined, we face the problem of how to 'use' it. The definition of 'central estimators' (such as the mean or the median) and 'estimators of dispersion' (such as the covariance matrix) lacks generality as it is quite easy to find examples (such as multimodal distributions in highly-dimensional spaces) where these estimators fail to have any interesting meaning.

When a probability distribution has been defined over a space of low dimension (say, from one to four dimensions) we can directly represent the associated probability density. This is trivial in one or two dimensions. It is easy in three dimensions, and some tricks may allow us to represent a four-dimensional probability distribution, but clearly this approach cannot be generalized to the high dimensional case.

Let us explain the only approach that seems practical, with help of Figure 3. At the left of the figure, there is an explicit representation of a 2D probability distribution (by means of the associated probability density or the associated (2D) volumetric probability). In the middle, some random points have been generated (using the Monte Carlo method about to be described). It is clear that, if we make a histogram with these points, in the limit of a sufficiently large number of points we recover the representation at the left. Disregarding the histogram possibility, we can concentrate on the individual points. In the 2D example of the figure we have actual points in a plane. If the problem is multidimensional, each 'point' may correspond to some abstract notion. For instance, for a geophysicist a 'point' may be a given model of the Earth. This model may be represented in some way, for instance, by a color plot. Then a collection of 'points' is a collection of such pictures. Our experience shows that, given a collection of randomly generated 'models,' the human eye–brain system is extremely good at apprehending the basic characteristics of the underlying probability distribution, including possible multimodalities, correlations, etc.

When such a (hopefully large) collection of random models is available, we can also answer quite interesting questions. For instance, a geologist might ask: *at which depth is that subsurface structure?* To answer this, we can make a histogram of the depth of the given geological structure over the collection of random models, and the histogram *is* the answer to the question. *What is the probability of having a low-velocity zone around a given depth?* The ratio of the number of models presenting such a low-velocity zone over the total number of models in the collection gives the answer (if the collection of models is large enough).

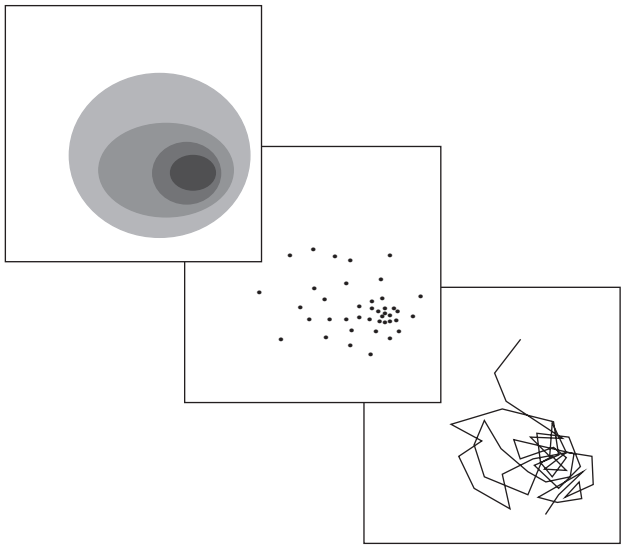


FIGURE 3 An explicit representation of a 2D probability distribution and the sampling of it, using Monte Carlo methods. While the representation at the top left cannot be generalized to high dimensions, the examination of a collection of points can be done in arbitrary dimensions. Practically, Monte Carlo generation of points is done through a 'random walk' where a 'new point' is generated in the vicinity of the previous point.

This is essentially what we propose: looking at a large number of randomly generated models in order to intuitively apprehend the basic properties of the probability distribution, followed by calculation of the probabilities of all interesting 'events.'

Practically, as we will see, the random sampling is not made by generating points independently of each other. Rather, as suggested in the last image of Figure 3, it is done through a 'random walk' where a 'new point' is generated in the vicinity of the previous point.

Monte Carlo methods have a random generator at their core. At present, Monte Carlo methods are typically implemented on digital computers, and are based on pseudo random generation of numbers.[10] As we shall see, any conceivable operation on probability densities (e.g., computing marginals and conditionals, integration, conjunction (the AND operation), etc.) has its counterpart in an operation on/by their corresponding Monte Carlo algorithms.

Inverse problems are often formulated in high-dimensional spaces. In this case a certain class of Monte Carlo algorithms,

the so-called *importance sampling algorithms*, come to the rescue, allowing us to sample the space with a sampling density proportional to the given probability density. In this case excessive (and useless) sampling of low-probability areas of the space is avoided. This is not only important, but in fact vital in high-dimensional spaces.

Another advantage of the importance sampling Monte Carlo algorithms is that we need not have a closed-form mathematical expression for the probability density that we want to sample. Only an algorithm that allows us to evaluate it at a given point in the space is needed. This has considerable practical advantage in analysis of inverse problems where computer-intensive evaluation of, for example, misfit functions plays an important role in calculation of certain probability densities.

Given a probability density that we wish to sample, and a class of Monte Carlo algorithms that samples this density, which one of the algorithms should we choose? Practically, the problem here is to find the most efficient of these algorithms. This is an interesting and difficult problem for which we will not go into detail here. We will, later in this chapter, limit ourselves to only two general methods that are recommendable in many practical situations.

3.1 Random Walks

To escape the dimensionality problem, *any* sampling of a probability density for which point values are available only upon request has to be based on a *random walk*, i.e., in a generation of successive points with the constraint that point $\mathbf{x}_{i+1}$ sampled in iteration $(i+1)$ is in the vicinity of the point $\mathbf{x}_i$ sampled in iteration i . The simplest of the random walks are generated by the so-called Markov Chain Monte Carlo (MCMC) algorithms, where the point $\mathbf{x}_{i+1}$ depends on the point $\mathbf{x}_i$, but not on previous points. We will concentrate on these algorithms here.

If random rules have been defined to select points such that the probability of selecting a point in the infinitesimal 'box' $dx_1 \cdots dx_N$ is $p(\mathbf{x})\, dx_1 \cdots dx_N$, then the points selected in this way are called *samples* of the probability density $p(\mathbf{x})$. Depending on the rules defined, successive samples $i, j, k, \ldots$ may be dependent or independent.

3.2 The Metropolis Rule

The most common Monte Carlo sampling methods are the Metropolis sampler (described below) and the Gibbs sampler (Geman and Geman, 1984). As we believe that the Gibbs sampler is only superior to the Metropolis sampler in low-dimensional problems, we restrict ourselves here to the presentation of the latter.

Consider the following situation. Some random rules define a random walk that samples the probability density $f(\mathbf{x})$. At a given step, the random walker is at point $\mathbf{x}_j$, and the application of the rules would lead to a transition to point $\mathbf{x}_i$. By construction, when all such 'proposed transitions' $\mathbf{x}_i \leftarrow \mathbf{x}_j$ are always accepted, the random walker will sample the probability density $f(\mathbf{x})$. Instead of always accepting the proposed transition $\mathbf{x}_i \leftarrow \mathbf{x}_j$, we reject it sometimes by using the following rule to decide if it is allowed to move to $\mathbf{x}_i$ or if it must stay at $\mathbf{x}_j$:

- If $g(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq g(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then accept the proposed transition to $\mathbf{x}_i$.
- If $g(\mathbf{x}_i)/\mu(\mathbf{x}_i) < g(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to move to $\mathbf{x}_i$, or to stay at $\mathbf{x}_j$, with the following probability of accepting the move to $\mathbf{x}_i$:

$$P = \frac{g(\mathbf{x}_i)/\mu(\mathbf{x}_i)}{g(\mathbf{x}_j)/\mu(\mathbf{x}_j)} \ . \tag{26}$$

Then we have the following theorem.

Theorem 1. *The random walker samples the conjunction* $h(\mathbf{x})$ *of the probability densities* $f(\mathbf{x})$ *and* $g(\mathbf{x})$

$$h(\mathbf{x}) = k\, f(\mathbf{x})\, \frac{g(\mathbf{x})}{\mu(\mathbf{x})} = k\, \frac{f(\mathbf{x})\, g(\mathbf{x})}{\mu(\mathbf{x})} \tag{27}$$

(*see Appendix O for a demonstration*).

It should be noted here that this algorithm nowhere requires the probability densities to be normalized. This is of vital importance in practice, since it allows sampling of probability densities whose values are known only in points already sampled by the algorithm. Obviously, such probability densities cannot be normalized. Also , the fact that our theory also allows unnormalizable probability densities will not cause any trouble in the application of the above algorithm.

The algorithm above is reminiscent (see Appendix O) of the Metropolis algorithm (Metropolis *et al.*, 1953), originally designed to sample the Gibbs–Boltzmann distribution.[11] Accordingly, we will refer to the above acceptance rule as the *Metropolis rule*.

3.3 The Cascaded Metropolis Rule

As above, assume that some random rules define a random walk that samples the probability density $f_1(\mathbf{x})$. At a given step, the random walker is at point $\mathbf{x}_j$.

(1) Apply the rules that unthwarted would generate samples distributed according to $f_1(\mathbf{x})$, to propose a new point $\mathbf{x}_i$.
(2) If $f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, go to point 3; if $f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to go to point 3 or to go back to point 1, with the following probability of going to point 3: $P = (f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j))$.

(3) If $f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, go to point 4; if $f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to go to point 4 or to go back to point 1, with the following probability of going to point 4: $P = (f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j))$.

... ...

(n) If $f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then accept the proposed transition to $\mathbf{x}_i$; if $f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j)$, then decide randomly to move to $\mathbf{x}_i$, or to stay at $\mathbf{x}_j$, with the following probability of accepting the move to $\mathbf{x}_i$: $P = (f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j))$.

Then we have the following theorem.

Theorem 2. *The random walker samples the conjunction* $h(\mathbf{x})$ *of the probability densities* $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})$:

$$h(\mathbf{x}) = k\, f_1(\mathbf{x})\, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \cdots \frac{f_n(\mathbf{x})}{\mu(\mathbf{x})} \quad . \tag{28}$$

(*see the supplementary materials to this chapter on the attached Handbook CD for a demonstration*).

3.4 Initiating a Random Walk

Consider the problem of obtaining samples of a probability density $h(\mathbf{x})$ defined as the conjunction of some probability densitites $f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}) \ldots$,

$$h(\mathbf{x}) = k\, f_1(\mathbf{x})\, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})}\, \frac{f_3(\mathbf{x})}{\mu(\mathbf{x})} \cdots \quad , \tag{29}$$

and let us examine three common situations.

We start with a random walk that samples $f_1(\mathbf{x})$ (optimal situation): This corresponds to the basic algorithm where we know how to produce a random walk that samples $f_1(\mathbf{x})$, and we only need to modify it, taking into account the values $f_2(\mathbf{x})/\mu(\mathbf{x})$, $f_3(\mathbf{x})/\mu(\mathbf{x}) \ldots$, using the cascaded Metropolis rule, to obtain a random walk that samples $h(\mathbf{x})$.

We start with a random walk that samples the homogeneous probability density $\mu(\mathbf{x})$: We can write Eq. (29) as

$$h(\mathbf{x}) = k \left(\left(\left(\mu(\mathbf{x})\, \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \cdots \right) . \tag{30}$$

The expression corresponds to the case where we are not able to start with a random walk that samples $f_1(\mathbf{x})$, but we have a random walk that samples the homogeneous probability density $\mu(\mathbf{x})$. Then, with respect to the example just mentioned, there is one extra step to be added, taking into account the values of $f_1(\mathbf{x})/\mu(\mathbf{x})$.

We start with an arbitrary random walk (worst situation): In the situation where we are not able to directly define a random walk that samples the homogeneous probability distribution, but only one that samples some arbitrary (but known) probability distribution $\psi(\mathbf{x})$, we can write Eq. (29) in the form

$$h(\mathbf{x}) = k \left(\left(\left(\left(\psi(\mathbf{x})\, \frac{\mu(\mathbf{x})}{\psi(\mathbf{x})} \right) \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \cdots \right) . \tag{31}$$

Then, with respect to the example just mentioned, there is one more extra step to be added, taking into account the values of $\mu(\mathbf{x})/\psi(\mathbf{x})$. Note that the closer $\psi(\mathbf{x})$ is be to $\mu(\mathbf{x})$, the more efficient will be the first modification of the random walk.

3.5 Convergence Issues

When has a random walk visited enough points in the space so that a probability density has been sufficiently sampled? This is a complex issue, and it is easy to overlook its importance. There is no general rule: Each problem has its own 'physics,' and the experience of the 'implementer' is crucial here.

Many methods that work for low dimension completely fail when the number of dimensions is high. Typically, a random walk select a random direction and, then, a random step along that direction. The notion of 'direction' in a high-dimensional space is far from the intuitive one we get in the familar three-dimensional space. Any serious discussion on this issue must be problem-dependent, so we do not even attempt one here.

Obviously, a necessary condition for adequate sampling is that any 'output' from the algorithm must 'look stationary.'

4. Probabilistic Formulation of Inverse Problems

A so-called 'inverse problem' arises when a usually complex measurement is made, and information on unknown parameters of the physical system is sought. Any measurement is indirect (we may weigh a mass by observing the displacement of the cursor of a balance), and therefore a possibly nontrivial analysis of uncertainties must be done. Any guide describing good experimental practice (see, for instance the ISO's *Guide to the Expression of Uncertainty in Measurement* (ISO, 1993) or the shorter description by Taylor and Kuyatt, 1994) acknowledges that a measurement involves, at least, two different sources of uncertainties: those estimated using statistical methods, and those estimated using subjective, common-sense estimations. Both are described using the axioms of probability theory, and this chapter clearly takes the probabilistic point of view for developing inverse theory.

4.1 Model Parameters and Observable Parameters

Although the separation of all the variables of a problem into two groups, 'directly observable parameters' (or 'data') and 'model parameters', may sometimes be artificial, we take this point of view here, since it allows us to propose a simple setting for a wide class of problems.

We may have in mind a given physical system, like the whole Earth or a small crystal under our microscope. The system (or a given state of the system) may be described by assigning values to a given set of parameters $\mathbf{m} = \{ m^1, m^2, \ldots, m^{NM} \}$, which we will name the *model parameters*.

Let us assume that we make observations on this system. Although we are interested in the parameters $\mathbf{m}$, they may not be directly observable, so we make indirect measurements such as obtaining seismograms at the Earth's surface for analyzing the Earth's interior, or making spectroscopic measurements for analyzing the chemical properties of a crystal. The set of (*directly*) *observable parameters* (or, by abuse of language, the set of *data parameters*) will be represented by $\mathbf{d} = \{ d^1, d^2, \ldots, d^{ND} \}$.

We assume that we have a physical theory that can be used to solve the *forward problem*, i.e., that given an arbitrary model $\mathbf{m}$, it allows us to predict the theoretical data values $\mathbf{d}$ that an ideal measurement should produce (if $\mathbf{m}$ were the actual system). The generally nonlinear function that associates with any model $\mathbf{m}$ the theoretical data values $\mathbf{d}$ may be represented by a notation such as

$$d^i = f^i(m^1, m^2, \ldots, m^{NM}) ; \qquad i = 1, 2, \ldots, ND , \quad (32)$$

or, for short,

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) . \quad (33)$$

It is in fact this expression that separates the whole set of our parameters into the subsets $\mathbf{d}$ and $\mathbf{m}$, although sometimes there is no difference in nature between the parameters in $\mathbf{d}$ and the parameters in $\mathbf{m}$. For instance, in the classical inverse problem of estimating the hypocenter coordinates of an earthquake, we may put in $\mathbf{d}$ the arrival times of the seismic waves at seismic observatories, and we need to put in $\mathbf{m}$, besides the hypocentral coordinates, the coordinates defining the location of the seismometers—as these are parameters that are needed to compute the travel times—although we estimate arrival times of waves and coordinates of the seismic observatories using similar types of measurements.

4.2 Prior Information on Model Parameters

In a typical geophysical problem, the model parameters contain geometrical parameters (positions and sizes of geological bodies) and physical parameters (values of the mass density, of the elastic parameters, the temperature, the porosity, etc.).

The *prior information* on these parameters is all the information we possess independently of the particular measurements that will be considered as 'data' (to be described below). This prior probability distribution is generally quite complex, as the model space may be high-dimensional, and the parameters may have nonstandard probability densities.

To this generally complex probability distribution over the model space corresponds a probability density that we denote $\rho_m(\mathbf{m})$.

If an explicit expression for the probability density $\rho_m(\mathbf{m})$ is known, it can be used in analytical developments. But such an explicit expression is, by no means, necessary. Using Monte Carlo methods, all that is needed is a set of probabilistic rules that allows us to generate samples distributed according to $\rho_m(\mathbf{m})$ in the model space (Mosegaard and Tarantola, 1995).

Example 4. *Appendix E presents an example of prior information for the case of an Earth model consisting of a stack of horizontal layers with variable thickness and uniform mass density.*

4.3 Measurements and Experimental Uncertainties

Observation of geophysical phenomena is represented by a set of parameters $\mathbf{d}$ that we usually call data. These parameters result from prior measurement operations, and they are typically seismic vibrations on the instrument site, arrival times of seismic phases, gravity or electromagnetic fields. As in any measurement, the data are determined with an associated uncertainty, described by a probability density over the data parameter space, that we denote here $\rho_d(\mathbf{d})$. This density describes not only marginals on individual datum values, but also possible cross-relations in data uncertainties.

Although the instrumental errors are an important source of data uncertainties, in geophysical measurements there are other sources of uncertainty. The errors associated with the positioning of the instruments, the environmental noise, and the human factor (like for picking arrival times) are also relevant sources of uncertainty.

Example 5. Nonanalytic Probability Density. *Assume that we wish to measure the time t of occurrence of some physical event. It is often assumed that the result of a measurement corresponds to something like*

$$t = t_0 \pm \sigma . \quad (34)$$

An obvious question is the exact meaning of the $\pm\sigma$. Has the experimenter in mind that she or he is absolutely certain that the actual arrival time satisfies the strict conditions

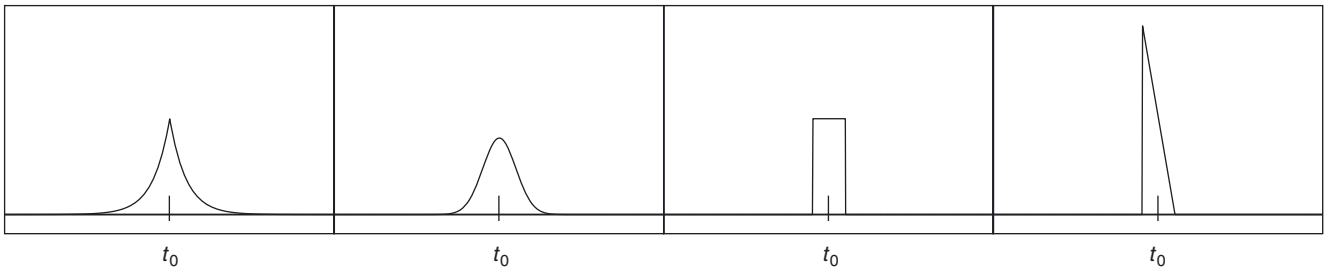


FIGURE 4 What has an experimenter in mind when she or he describes the result of a measurement by something like $t = t_0 \pm \sigma$?

$t_0 - \sigma \le t \le t_0 + \sigma$, or has she or he in mind something like a Gaussian probability, or some other probability distribution (see Fig. 4)? We accept, following ISO's recommendations (1993) that the result of any measurement has a probabilistic interpretation, with some sources of uncertainty being analyzed using statistical methods ('type A' uncertainties), and other sources of uncertainty being evaluated by other means (for instance, using Bayesian arguments) ('type B' uncertainties). But, contrary to ISO suggestions, we do not assume that the Gaussian model of uncertainties should play any central role. In an extreme example, we may well have measurements whose probabilistic description may correspond to a multimodal probability density. Figure 5 shows a typical example for a seismologist: the measurement on a seismogram of the arrival time of a certain seismic wave, in the case one hesitates in the phase identification or in the identification of noise and signal. In this case the probability density for the arrival of the seismic phase does not have an explicit expression like $f(t) = k \exp(-(t - t_0)^2/(2\sigma^2))$, but is a numerically defined function.

Example 6. *The Gaussian model for uncertainties. The simplest probabilistic model that can be used to describe experimental uncertainties is the Gaussian model*

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{d}_{\text{obs}})^T \, \mathbf{C}_D^{-1} \, (\mathbf{d} - \mathbf{d}_{\text{obs}})\right) . \qquad (35)$$

It is here assumed that we have some 'observed data values' $\mathbf{d}_{\text{obs}}$ with uncertainties described by the covariance matrix $\mathbf{C}_D$. If the uncertainties are uncorrelated,

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2}\sum_i \left(\frac{d^i - d^i_{\text{obs}}}{\sigma^i}\right)^2\right) , \qquad (36)$$

where the σ^i are the 'standard deviations.'

Example 7. *The generalized Gaussian model for uncertainties. An alternative to the Gaussian model is to use the*

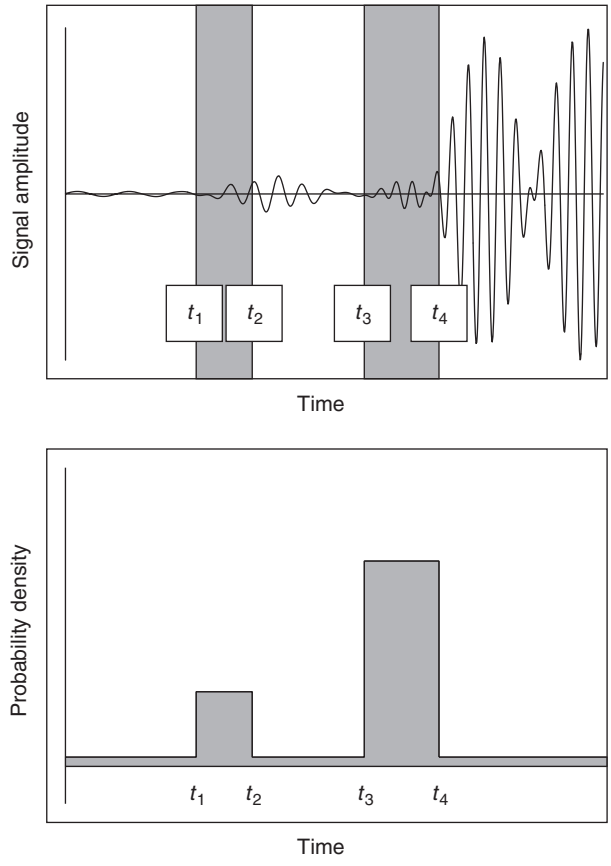


FIGURE 5 A seismologist tries to measure the arrival time of a seismic wave at a seismic station, by 'reading' the seismogram at the top of the figure. The seismologist may find quite likely that the arrival time of the wave is between times t_3 and t_4, and believe that what is before t_3 is just noise. But if there is a significant probability that the signal between t_1 and t_2 is not noise but the actual arrival of the wave, then the seismologist should define a bimodal probability density, as the one suggested at the bottom of the figure. Typically, the actual form of each peak of the probability density is not crucial (here, box-car functions are chosen), but the position of the peaks is important. Rather than assigning a zero probability density to the zones outside the two intervals, it is safer (more 'robust') to attribute some small 'background' value, as we may never exclude some unexpected source of error.

Laplacian (double exponential) model for uncertainties,

$$\rho_d(\mathbf{d}) = k \exp\left(-\sum_i \frac{|d^i - d^i_{\text{obs}}|}{\sigma^i} \right) . \qquad (37)$$

While the Gaussian model leads to least-squares-related methods, this Laplacian model leads to absolute-values methods (see Section 4.5.2), well known for producing robust[12] *results. More generally, there is the L_p model of uncertainties*

$$\rho_p(\mathbf{d}) = k \exp\left(-\frac{1}{p} \sum_i \frac{|d^i - d^i_{\text{obs}}|^p}{(\sigma^i)^p} \right) \qquad (38)$$

(see Fig. 6).

4.4 Joint "Prior" Probability Distribution in the $(\mathcal{M}, \mathcal{D})$ Space

We have just seen that the prior information on model parameters can be described by a probability density in the model space, $\rho_m(\mathbf{m})$, and that the result of measurements can be described by a probability density in the data space $\rho_d(\mathbf{d})$. As by 'prior' information on model parameters we mean information obtained *independently* from the measurements (it often represents information we had before the measurements were made), we can use the notion of independency of variables of Section 2.6 to define a joint probability density in the $\mathcal{X} = (\mathcal{M}, \mathcal{D})$ space as the product of the two 'marginals'

$$\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\, \rho_d(\mathbf{d}) . \qquad (39)$$

Although we have introduced $\rho_m(\mathbf{m})$ and $\rho_d(\mathbf{d})$ separately, and we have suggested building a probability distribution in the $(\mathcal{M}, \mathcal{D})$ space by the multiplication (39), we may have a more general situation where the information we have on $\mathbf{m}$ and on $\mathbf{d}$ is not independent. So, in what follows, let us assume that we have some information in the $\mathcal{X} = (\mathcal{M}, \mathcal{D})$ space, represented by the 'joint' probability density

$$\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d}) , \qquad (40)$$

and let us consider Eq. (39) as just a special case.

Let us in the rest of this chapter denote by $\mu(\mathbf{x})$ the probability density representing the homogeneous probability distribution, as introduced in Section 2.2. We may remember here the Rule 8, stating that the limit of a consistent probability density must be the homogeneous one, so we may formally write

$$\mu(\mathbf{x}) = \lim_{\text{infinite dispersions}} \rho(\mathbf{x}) . \qquad (41)$$

When the partition (39) holds, then, typically (see Rule 8),

$$\mu(\mathbf{x}) = \mu(\mathbf{m}, \mathbf{d}) = \mu_m(\mathbf{m})\, \mu_d(\mathbf{d}) . \qquad (42)$$

4.5 Physical Laws as Mathematical Functions

4.5.1 Physical Laws

Physics analyzes the correlations existing between physical parameters. In standard mathematical physics, these correlations are represented by 'equalities' between physical parameters (as when we write $\mathbf{F} = m\,\mathbf{a}$ to relate the force

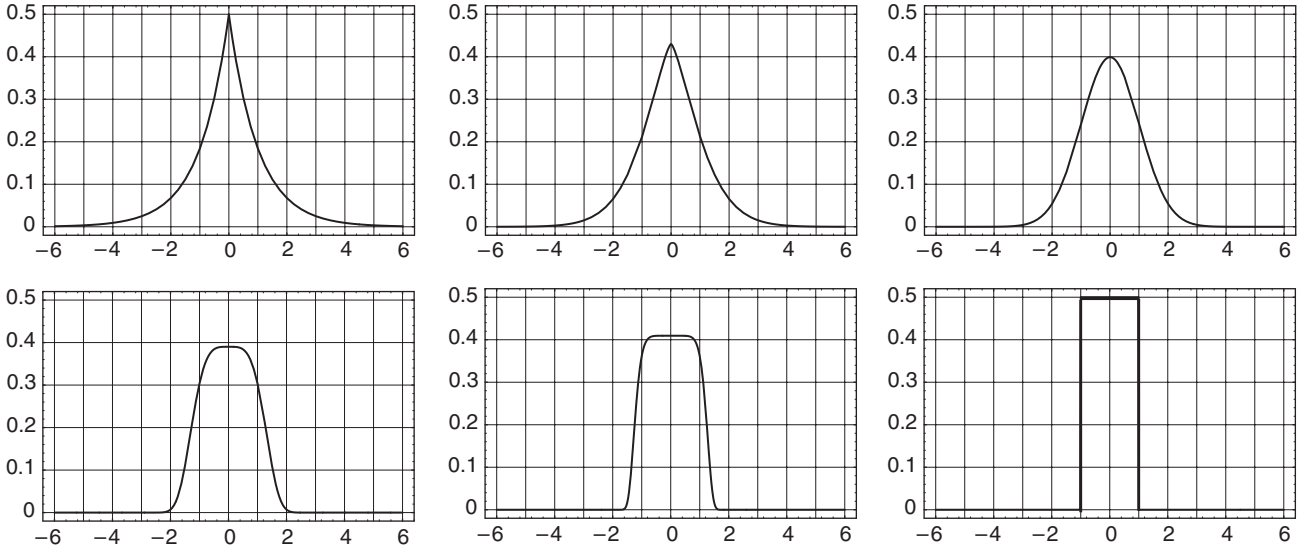


FIGURE 6 Generalized Gaussian for values of the parameter $p = 1$, $\sqrt{2}$, 2, 4, 8, and ∞ .

F applied to a particle, the mass m of the particle and the acceleration **a**). In the context of inverse problems, this corresponds to assuming that we have a function from the 'parameter space' to the 'data space' that we may represent as

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) \ . \tag{43}$$

We do not mean that the relation is necessarily explicit. Given **m** we may need to solve a complex system of equations in order to get **d**, but this nevertheless defines a function $\mathbf{m} \to \mathbf{d} = \mathbf{f}(\mathbf{m})$.

At this point, given the probability density $\rho(\mathbf{m}, \mathbf{d})$ and given the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$, we can define the associated conditional probability density $\rho_{m\,|\,d\,(m)}(\mathbf{m} \mid \mathbf{d} = \mathbf{f}(\mathbf{m}))$. We could here use the more general definition of conditional probability density of Appendix B, but let us simplify the text by using a simplifying assumption: that the total parameter space $(\mathcal{M}, \mathcal{D})$ is just the Cartesian product $\mathcal{M} \times \mathcal{D}$ of the model parameter space $\mathcal{M}$ times the space of directly observable parameters (or 'data space') $\mathcal{D}$. Then, rather than a general metric in the total space, we have a metric $\mathbf{g}_m$ over the model parameter space $\mathcal{M}$ and a metric $\mathbf{g}_d$ over the data space, and the total metric is just the Cartesian product of the two metrics. In particular, then, the total volume element in the space, $dV(\mathbf{m}, \mathbf{d})$ is just the product of the two volume elements in the model parameter space and the data space: $dV(\mathbf{m}, \mathbf{d}) = dV_m(\mathbf{m})\, dV_d(\mathbf{D})$. Most inverse problems satisfy this assumption.[13] In this setting, the formulas of Section 2.4 are valid.

4.5.2 Inverse Problems

In the $(\mathcal{M}, \mathcal{D}) = \mathcal{M} \times \mathcal{D}$ space, we have the probability density $\rho(\mathbf{m}, \mathbf{d})$ and we have the hypersurface defined by the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$. The natural way to 'compose' these two kinds of information is by defining the conditional probability density induced by $\rho(\mathbf{m}, \mathbf{d})$ on the hypersurface $\mathbf{d} = \mathbf{f}(\mathbf{m})$,

$$\sigma_m(\mathbf{m}) \equiv \rho_{m|d(m)}(\mathbf{m}|\mathbf{d} = \mathbf{f}(\mathbf{m})) \ , \tag{44}$$

this gives (see Eq. (17))

$$\sigma_m(\mathbf{m}) = k\, \rho(\mathbf{m}, \mathbf{f}(\mathbf{m}))\, \frac{\sqrt{\det\left(\mathbf{g}_m + \mathbf{F}^T \mathbf{g}_d \mathbf{F}\right)}}{\sqrt{\det \mathbf{g}_m}\sqrt{\det \mathbf{g}_d}}\bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \ , \tag{45}$$

where $\mathbf{F} = \mathbf{F}(\mathbf{m})$ is the matrix of partial derivatives, with components $\mathbf{F}_{i\alpha} = \partial f_i / \partial m_\alpha$, where $\mathbf{g}_m$ is the metric in the model parameter space $\mathcal{M}$ and where $\mathbf{g}_d$ is the metric in the data space $\mathcal{D}$.

Example 8. *Quite often*, $\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\, \rho_d(\mathbf{d})$. *Then*, *Eq.* (45) *can be written*

$$\sigma_m(\mathbf{m}) = k\, \rho_m(\mathbf{m}) \left(\frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_m}} \frac{\sqrt{\det(\mathbf{g}_m + \mathbf{F}^T \mathbf{g}_d \mathbf{F})}}{\sqrt{\det \mathbf{g}_m}} \right)\bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \ . \tag{46}$$

Example 9. *If* $g_{\mathbf{m}}(\mathbf{m}) = constant$ *and* $g_{\mathbf{d}}(\mathbf{d}) = constant$, *and the nonlinearities are weak* $(\mathbf{F}(\mathbf{m}) = constant)$, *then Eq.* (46) *reduces to*

$$\sigma_m(\mathbf{m}) = k\ \rho_m(\mathbf{m})\, \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})}\bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \ , \tag{47}$$

where we have used $\mu_d(\mathbf{d}) = k\ \sqrt{\det \mathbf{g}_d(\mathbf{d})}$ (*see Rule 2*).

Example 10. *We examine here the simplification that we arrive at when assuming that the 'input' probability densities are Gaussian*:

$$\rho_m(\mathbf{m}) = k\ \exp\left(-\frac{1}{2} (\mathbf{m} - \mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}\ (\mathbf{m} - \mathbf{m}_{\text{prior}}) \right) \tag{48}$$

$$\rho_d(\mathbf{d}) = k\ \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1}\ (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right) \ . \tag{49}$$

In this circumstance, *quite often*, *it is the covariance operators* $\mathbf{C}_M$ and $\mathbf{C}_D$ *that are used to define the metrics over the spaces* $\mathcal{M}$ *and* $\mathcal{D}$. *Then*, $\mathbf{g}_m = \mathbf{C}_M^{-1}$ *and* $\mathbf{g}_d = \mathbf{C}_D^{-1}$. *Grouping some of the constant factors in the factor* k , *Eq.* (45) *becomes here*

$$\begin{aligned} \sigma_m(\mathbf{m}) = k\ \exp\Bigg[-\frac{1}{2} \Big(& (\mathbf{m} - \mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}\ (\mathbf{m} - \mathbf{m}_{\text{prior}}) \\ & + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1}\ (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \Big) \Bigg] \\ & \times \frac{\sqrt{\det\left(\mathbf{C}_M^{-1} + \mathbf{F}^T(\mathbf{m})\ \mathbf{C}_D^{-1} \mathbf{F}(\mathbf{m})\right)}}{\sqrt{\det \mathbf{C}_M^{-1}}} \end{aligned} \tag{50}$$

(*the constant factor* $\sqrt{\det \mathbf{C}_M^{-1}}$ *has been left for subsequent simplifications*). *Defining the misfit*

$$S(\mathbf{m}) = -2 \log \frac{\sigma_m(\mathbf{m})}{\sigma_0} \ , \tag{51}$$

where σ_0 *is an arbitrary value of* $\sigma_m(\mathbf{m})$, *gives*, *up to an additive constant*,

$$S(\mathbf{m}) = S_1(\mathbf{m}) - S_2(\mathbf{m}) \ , \tag{52}$$

where $S_1(\mathbf{m})$ is the usual least-squares misfit function

$$S_1(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_{\text{prior}})^t \, \mathbf{C}_M^{-1} \, (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \tag{53}$$

and where [14]

$$S_2(\mathbf{m}) = \log \det \left(\mathbf{I} + \mathbf{C}_M \mathbf{F}^t(\mathbf{m}) \, \mathbf{C}_D^{-1} \, \mathbf{F}(\mathbf{m}) \right) . \tag{54}$$

Example 11. *If, in the context of Example 10, we have* [15] $\mathbf{C}_M \mathbf{F}^t \mathbf{C}_D^{-1} \mathbf{F} \ll \mathbf{I}$ *, we can use the low order approximation for* $S_2(\mathbf{m})$ *, that is* [16]

$$S_2(\mathbf{m}) \approx \text{trace} \, \mathbf{C}_M \mathbf{F}^t(\mathbf{m}) \, \mathbf{C}_D^{-1} \, \mathbf{F}(\mathbf{m}) . \tag{55}$$

Example 12. *If in the context of Example 10 we assume that the nonlinearities are weak, then the matrix of partial derivatives* $\mathbf{F}$ *is approximately constant, and Eq. (50) simplifies to*

$$\sigma_m(\mathbf{m}) = k \exp \left[-\frac{1}{2} \left((\mathbf{m} - \mathbf{m}_{\text{prior}})^t \, \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right) \right] , \tag{56}$$

and the function $S_2(\mathbf{m})$ *is just a constant.*

Example 13. *If the 'relation solving the forward problem'* $\mathbf{d} = \mathbf{f}(\mathbf{m})$ *happens to be a linear relation,* $\mathbf{d} = \mathbf{F}\mathbf{m}$*, then one gets the standard equations for linear problems (see Appendix F).*

Example 14. *We examine here the simplifications that we arrive at when assuming that the 'input' probability densities are Laplacian:*

$$\rho_m(\mathbf{m}) = k \exp \left(-\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} \right) \tag{57}$$

$$\rho_d(\mathbf{d}) = k \exp \left(-\sum_i \frac{|d^i - d^i_{\text{obs}}|}{\sigma_i} \right) . \tag{58}$$

Equation (45) becomes here

$$\sigma_m(\mathbf{m}) = k \exp \left[-\left(\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} + \sum_i \frac{|f^i(\mathbf{m}) - d^i_{\text{obs}}|}{\sigma_i} \right) \right] \Psi(\mathbf{m}) , \tag{59}$$

where $\Psi(\mathbf{m})$ *is a complex term containing, in particular, the matrix of partial derivatives* $\mathbf{F}$ *. If this term is approximately constant (weak nonlinearities, constant metrics), then*

$$\sigma_m(\mathbf{m}) = k \exp \left[-\left(\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} + \sum_i \frac{|f^i(\mathbf{m}) - d^i_{\text{obs}}|}{\sigma_i} \right) \right] . \tag{60}$$

The formulas in the examples above give expressions that contain analytic parts (like the square roots containing the matrix of partial derivatives $\mathbf{F}$) . What we write as $\mathbf{d} = \mathbf{f}(\mathbf{m})$ may sometimes correspond to an explicit expression; sometimes it may correspond to the solution of an implicit equation.[17] Should $\mathbf{d} = \mathbf{f}(\mathbf{m})$ be an explicit expression, and should the 'prior probability densities' $\rho_m(\mathbf{m})$ and $\rho_d(\mathbf{d})$ (or the joint $\rho(\mathbf{m}, \mathbf{d})$) also be given by explicit expressions (as when we have Gaussian probability densities), then the formulas of this section would give explicit expressions for the posterior probability density $\sigma_m(\mathbf{m})$.

If the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is a linear relation, then the expression giving $\sigma_m(\mathbf{m})$ can sometimes be simplified easily (as with the linear Gaussian case to be examined below). More often than not the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is a complex nonlinear relation, and the expression we are left with for $\sigma_m(\mathbf{m})$ is explicit, but complex.

Once the probability density $\sigma_m(\mathbf{m})$ has been defined, there are different ways of 'using' it. If the 'model space' $\mathcal{M}$ has a small number of dimensions (say between one and four) the values of $\sigma_m(\mathbf{m})$ can be computed at every point of a grid and a graphical representation of $\sigma_m(\mathbf{m})$ can be attempted. A visual inspection of such a representation is usually worth a thousand 'estimators' (central estimators or estimators of dispersion). But, of course, if the values of $\sigma_m(\mathbf{m})$ are known at all points where $\sigma_m(\mathbf{m})$ has a significant value, these estimators can also be computed.

If the 'model space' $\mathcal{M}$ has a large number of dimensions (say from five to many millions or billions), then an exhaustive exploration of the space is not possible, and we must turn to Monte Carlo sampling methods to extract information from $\sigma_m(\mathbf{m})$. We discuss the application of Monte Carlo methods to inverse problems, and optimization techniques, in Section 6 and 7, respectively.

4.6 Physical Laws as Probabilistic Correlations

4.6.1 Physical Laws

We return here to the general case where it is not assumed that the total space $(\mathcal{M}, \mathcal{D})$ is the Cartesian product of two spaces.

In Section 4.5 we have examined the situation where the physical correlation between the parameters of the

problem are expressed using an exact, analytic expression $\mathbf{d} = \mathbf{g}(\mathbf{m})$. In this case, the notion of conditional probability density has been used to combine the 'physical theory' with the 'data' and the 'a priori information' on model parameters.

But, we have seen that in order to properly define the notion of conditional probability density, it has been necessary to introduce a metric over the space, and to take a limit using the metric of the space. This is equivalent to put some 'thickness' around the theoretical relation $\mathbf{d} = \mathbf{g}(\mathbf{m})$, and to take the limit when the thickness tends to zero.

But actual theories have uncertainties, and, for more generality, it is better to explicitly introduce these uncertainties. Assume, then, that the physical correlations between the model parameters $\mathbf{m}$ and the data parameters $\mathbf{d}$ are not represented by an analytical expression like $\mathbf{d} = \mathbf{f}(\mathbf{m})$ but by a probability density

$$\vartheta(\mathbf{m}, \mathbf{d}) \ . \tag{61}$$

Example: Realistic 'Uncertainty Bars' around a Functional Relation. In the approximation of a constant gravity field, with acceleration $\mathbf{g}$, the position at time t of an apple in free fall is $\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}_0 t + \frac{1}{2}\mathbf{g}t^2$, where $\mathbf{r}_0$ and $\mathbf{v}_0$ are, respectively, the position and velocity of the object at time $t = 0$. More simply, if the movement is 1D,

$$x(t) = x_0 + v_0 t + \frac{1}{2} g t^2 \ . \tag{62}$$

Of course, for many reasons this equation can never be exact: air friction, wind effects, inhomogeneity of the gravity field, effects of the Earth rotation, forces from the Sun and the Moon (not to mention Pluto), relativity (special and general), and so on.

It is not a trivial task, given very careful experimental conditions, to estimate the size of the leading uncertainty. Although one might think of an equation $x = x(t)$ as a line, infinitely thin, there will always be sources of uncertainty (at least due to the unknown limits of validity of general relativity): looking at the line with a magnifying glass should reveal a fuzzy object of finite thickness. As a simple example, let us examine here the mathematical object we arrive at when assuming that the leading sources of uncertainty in the relation $x = x(t)$ are the uncertainties in the initial position and velocity of the falling apple. Let us assume that:

- the initial position of the apple is random, with a Gaussian distribution centered at x_0, and with standard deviation σ_x;
- the initial velocity of the apple is random, with a Gaussian distribution centered at v_0, and with standard deviation σ_v.

Then it can be shown that at a given time t, the possible positions of the apple are random, with probability density

$$\vartheta(x|t) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2 + \sigma_v^2 t^2}} \times \exp\left(-\frac{1}{2} \frac{\left(x - (x_0 + v_0 t + \frac{1}{2} g t^2)\right)^2}{\sigma_x^2 + \sigma_v^2 t^2} \right) \ . \tag{63}$$

This is obviously a conditional probability density for x, given t. Should we have any reason to choose some marginal probability density $\vartheta_t(t)$, then, the 'law' for the fall of the apple would be

$$\vartheta(x, t) = \vartheta(x|t)\, \vartheta_t(t) \ . \tag{64}$$

See Appendix C for more details.

4.6.2 Inverse Problems

We have seen that the result of measurements can be represented by a probability density $\rho_d(\mathbf{d})$ in the data space. We have also seen that the a priori information on the model parameters can be represented by another probability density $\rho_m(\mathbf{m})$ in the model space. When we talk about 'measurements' and about 'a priori information on model parameters,' we usually mean that we have a joint probability density in the $(\mathcal{M}, \mathcal{D})$ space, that is $\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\rho_d(\mathbf{d})$. Let us consider the more general situation where for the whole set of parameters $(\mathcal{M}, \mathcal{D})$ we have some information that can be represented by a joint probability density $\rho(\mathbf{m}, \mathbf{d})$. Having well in mind the interpretation of this information, let us use the simple term 'experimental information' for it:

$$\rho(\mathbf{m}, \mathbf{d}) \qquad \text{(experimental information)} \ . \tag{65}$$

We have also seen that we have information coming from physical theories, that predict correlations between the parameters, and it has been argued that a probabilistic description of these correlations is well adapted to the resolution of inverse problems.[18] Let $\vartheta(\mathbf{m}, \mathbf{d})$ be the probability density representing this 'theoretical information':

$$\vartheta(\mathbf{m}, \mathbf{d}) \qquad \text{(theoretical information)} \ . \tag{66}$$

A quite fundamental assumption is that in all the spaces we consider, there is a notion of volume that allows us to give meaning to the notion of a 'homogeneous probability distribution' over the space. The corresponding probability density is not constant, but is proportional to the volume element of the space (see Section 2.2):

$$\mu(\mathbf{m}, \mathbf{d}) \qquad \text{(homogeneous probability distribution)} \ . \tag{67}$$

Finally, we have seen examples suggesting that the conjunction of the experimental information with the theoretical information corresponds exactly to the AND operation defined

over the probability densities, to obtain the 'conjunction of information,' as represented by the probability density

$$\sigma(\mathbf{m}, \mathbf{d}) = k \, \frac{\rho(\mathbf{m}, \mathbf{d}) \, \vartheta(\mathbf{m}, \mathbf{d})}{\mu(\mathbf{m}, \mathbf{d})} \qquad \text{(conjunction of information)} \, , \tag{68}$$

with marginal probability densities

$$\sigma_m(\mathbf{m}) = \int_{\mathcal{D}} d\mathbf{d} \, \sigma(\mathbf{m}, \mathbf{d}) \, ; \qquad \sigma_d(\mathbf{d}) = \int_{\mathcal{M}} d\mathbf{m} \, \sigma(\mathbf{m}, \mathbf{d}) \, . \tag{69}$$

Example 15. *We may assume that the physical correlations between the parameters* $\mathbf{m}$ *and* $\mathbf{d}$ *are of the form*

$$\vartheta(\mathbf{m}, \mathbf{d}) = \vartheta_{D|M}(\mathbf{d}|\mathbf{m}) \, \vartheta_M(\mathbf{m}) \, , \tag{70}$$

this expressing that a 'physical theory' gives, on the one hand, the conditional probability for $\mathbf{d}$ *, given* $\mathbf{m}$ *, and, on the other hand, the marginal probability density for* $\mathbf{m}$ *. See Appendix C for more details.*

Example 16. *Many applications concern the special situation where we have*

$$\mu(\mathbf{m}, \mathbf{d}) = \mu_m(\mathbf{m}) \, \mu_d(\mathbf{d}) \, ; \qquad \rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m}) \, \rho_d(\mathbf{d}) \, . \tag{71}$$

In this case, Eqs. (68) and (69) give

$$\sigma_m(\mathbf{m}) = k \, \frac{\rho_m(\mathbf{m})}{\mu_m(\mathbf{m})} \int_{\mathcal{D}} d\mathbf{d} \, \frac{\rho_d(\mathbf{d}) \, \vartheta(\mathbf{m}, \mathbf{d})}{\mu_d(\mathbf{d})} \, . \tag{72}$$

If Eq. (70) holds, then

$$\sigma_m(\mathbf{m}) = k \, \rho_m(\mathbf{m}) \, \frac{\vartheta_m(\mathbf{m})}{\mu_m(\mathbf{m})} \int_{\mathcal{D}} d\mathbf{d} \, \frac{\rho_d(\mathbf{d}) \, \vartheta_{D|M}(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} \, . \tag{73}$$

Finally, if the simplification $\vartheta_M(\mathbf{m}) = \mu_m(\mathbf{m})$ *arises (see Appendix C for an illustration), then*

$$\sigma_m(\mathbf{m}) = k \, \rho_m(\mathbf{m}) \int_{\mathcal{D}} d\mathbf{d} \, \frac{\rho_d(\mathbf{d}) \, \vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} \, . \tag{74}$$

Example 17. *In the context of the previous example, assume that observational uncertainties are Gaussian:*

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t \, \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right) . \tag{75}$$

Note that the limit for infinite variances gives the homogeneous probability density $\mu_d(\mathbf{d}) = k$*. Furthermore, assume that uncertainties in the physical law are also Gaussian:*

$$\vartheta(\mathbf{d}|\mathbf{m}) = k \exp\left(-\frac{1}{2} (\mathbf{d} - \mathbf{f}(\mathbf{m}))^t \, \mathbf{C}_T^{-1} (\mathbf{d} - \mathbf{f}(\mathbf{m})) \right) . \tag{76}$$

Here 'the physical theory says' that the data values must be 'close' to the 'computed values' $\mathbf{f}(\mathbf{m})$*, with a notion of closeness defined by the 'theoretical covariance matrix'* $\mathbf{C}_T$*. As demonstrated in Tarantola (1987, p. 158), the integral in Eq. (74) can be analytically evaluated, and gives*

$$\int_{\mathcal{D}} d\mathbf{d} \, \frac{\rho_d(\mathbf{d}) \, \vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} = k \exp\left(-\frac{1}{2} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t (\mathbf{C}_D + \mathbf{C}_T)^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right) . \tag{77}$$

This shows that when using the Gaussian probabilistic model, observational and theoretical uncertainties combine through addition of the respective covariance operators (a nontrivial result).

Example 18. *In the 'Galilean law' example developed in Section 4.61, we described the correlation between the position* x *and the time* t *of a free falling object through a probability density* $\vartheta(x, t)$ *. This law says that falling objects describe, approximately, a space–time parabola. Assume that in a particular experiment the falling object explodes at some point of its space–time trajectory. A plain measurement of the coordinates* (x, t) *of the event gives the probability density* $\rho(x, t)$ *. By 'plain measurement' we mean here that we have used a measurement technique that is not taking into account the particular parabolic character of the fall (i.e., the measurement is designed to work identically for any sort of trajectory). The conjunction of the physical law* $\vartheta(x, t)$ *and the experimental result* $\rho(x, t)$ *, using expression (68), gives*

$$\sigma(x, t) = k \, \frac{\rho(x, t) \, \vartheta(x, t)}{\mu(x, t)} \, , \tag{78}$$

where, as the coordinates (x, t) *are 'Cartesian,'* $\mu(x, t) = k$ *. Taking the explicit expression given for* $\vartheta(x, t)$ *in Eqs. (63) and (64), with* $\vartheta_t(t) = k$ *,*

$$\vartheta(x, t) = \frac{1}{\sqrt{2\pi} \sqrt{\sigma_x^2 + \sigma_v^2 \, t^2}} \times \exp\left(-\frac{1}{2} \frac{\left(x - (x_0 + v_0 \, t + \frac{1}{2} \, g \, t^2) \right)^2}{\sigma_x^2 + \sigma_v^2 \, t^2} \right) , \tag{79}$$

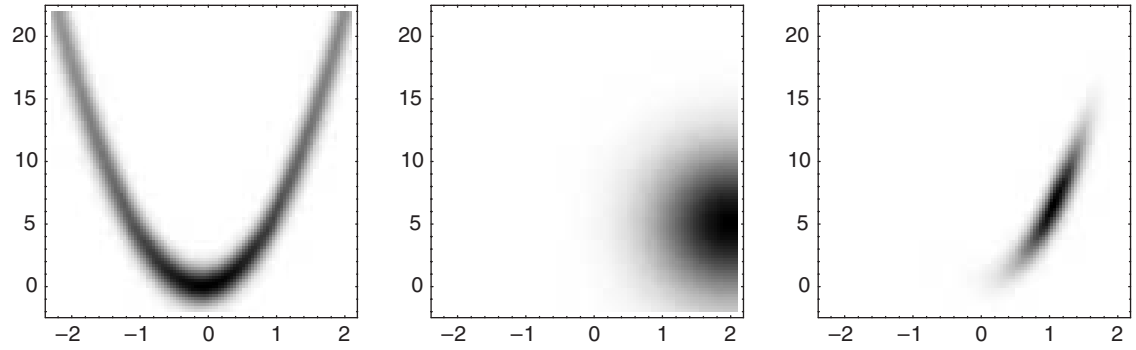


FIGURE 7 This figure has been made with the numerical values mentioned in Figure 17 (see Appendix C) with, in addition, $x_{\text{obs}} = 5.0\,\text{m}$, $\Sigma_x = 4.0\,\text{m}$, $t_{\text{obs}} = 2.0\,\text{sec}$ and $\Sigma_t = 0.75\,\text{sec}$.

and assuming the Gaussian form[19] *for* $\rho(x,t)$,

$$\rho(x,t) = \rho_x(x)\,\rho_t(t) = k \exp\left(-\frac{1}{2}\frac{(x-x_{\text{obs}})^2}{\Sigma_x^2}\right) \exp\left(-\frac{1}{2}\frac{(t-t_{\text{obs}})^2}{\Sigma_t^2}\right) \tag{80}$$

we obtain the combined probability density

$$\sigma(x,t) = \frac{k}{\sqrt{\sigma_x^2+\sigma_v^2 t^2}} \exp\left(-\frac{1}{2}\left(\frac{(x-x_{\text{obs}})^2}{\Sigma_x^2}+\frac{(t-t_{\text{obs}})^2}{\Sigma_t^2} + \frac{\left(x-(x_0+v_0 t+\frac{1}{2}\,g t^2)\right)^2}{\sigma_x^2+\sigma_v^2 t^2}\right)\right) . \tag{81}$$

Figure 7 illustrates the three probability densities $\vartheta(x,t)$, $\rho(x,t)$, *and* $\sigma(x,t)$. *See Appendix C for a more detailed examination of this problem.*

5. Solving Inverse Problems (I): Examination of the Probability Density

The next two sections deal with Monte Carlo and optimization methods. The implementation of these methods takes some programming effort that is not required when we face problems with fewer degrees of freedom (say, between one and five).

When we have a small number of parameters we should directly 'plot' the probability density.

In Appendix K the problem of estimation of a seismic hypocenter is treated, and it is shown there that the examination of the probability density for the location of the hypocenter offers a much better possibility for analysis than any other method.

6. Solving Inverse Problems (II): Monte Carlo Methods

6.1 Basic Equations

The starting point could be the explicit expression (Eq. (46)) for $\sigma_m(\mathbf{m})$ given in Section 4.5.2:

$$\sigma_m(\mathbf{m}) = k\,\rho_m(\mathbf{m})\,L(\mathbf{m}) . \tag{82}$$

where

$$L(\mathbf{m}) = \left(\frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_d(\mathbf{d})}} \frac{\sqrt{\det\left(\mathbf{g}_m(\mathbf{m}) + \mathbf{F}^T(\mathbf{m})\,\mathbf{g}_d(\mathbf{d})\,\mathbf{F}(\mathbf{m})\right)}}{\sqrt{\det \mathbf{g}_m(\mathbf{m})}}\right)\Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} . \tag{83}$$

In this expression the matrix of partial derivatives $\mathbf{F} = \mathbf{F}(\mathbf{m})$, with components $D_{i\alpha} = \partial f_i / \partial m_\alpha$, appears. The 'slope' $\mathbf{F}$ enters here because the steeper the slope for a given $\mathbf{m}$, the greater the accumulation of points we will have with this particular $\mathbf{m}$. This is because we use explicitly the analytic expression $\mathbf{d} = \mathbf{f}(\mathbf{m})$. One should realize that using the more general approach based on Eq. (68) of Section 4.6.2, the effect is automatically accounted for, and there is no need to explicitly consider the partial derivatives.

Equation (82) has the standard form of a conjunction of two probability densities, and is therefore ready to be integrated in a Metropolis algorithm. But one should note that, contrary to many 'nonlinear' formulations of inverse problems, the partial derivatives $\mathbf{F}$ are needed even if we use a Monte Carlo method.

In some weakly nonlinear problems, we have $\mathbf{F}^T(\mathbf{m})\,\mathbf{g}_d(\mathbf{d})\,\mathbf{F}(\mathbf{m}) \ll \mathbf{g}_m(\mathbf{m})$, and then Eq. (83) becomes

$$L(\mathbf{m}) = \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})}\bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} , \tag{84}$$

where we have used $\mu_d(\mathbf{d}) = k\sqrt{\det \mathbf{g}_d(\mathbf{d})}$ (see Rule 2).

This expression is also ready for use in the Metropolis algorithm. In this way, sampling of the prior $\rho_m(\mathbf{m})$ is modified into a sampling of the posterior $\sigma_m(\mathbf{m})$, and the Metropolis Rule uses the 'likelihood function' $L(\mathbf{m})$ to calculate acceptance probabilities.

6.2 Sampling the Homogeneous Probability Distribution

If we do not have an algorithm that samples the prior probability density directly, the first step in a Monte Carlo analysis of an inverse problem is to design a random walk that samples the model space according to the homogeneous probability distribution $\mu_m(\mathbf{m})$. In some cases this is easy, but in other cases only an algorithm (a *primeval random walk*) that samples an arbitrary (possibly constant) probability density $\psi(\mathbf{m}) \neq \mu_m(\mathbf{m})$ is available. Then the Metropolis rule can be used to modify $\psi(\mathbf{m})$ into $\mu_m(\mathbf{m})$ (see Section 3.4). This way of generating samples from $\mu_m(\mathbf{m})$ is efficient if $\psi(\mathbf{m})$ is close to $\mu_m(\mathbf{m})$, otherwise it may be very inefficient.

Once $\mu(\mathbf{m})$ can be sampled, the Metropolis Rule allows us to modify this sampling into an algorithm that samples the prior.

6.3 Sampling the Prior Probability Distribution

The first step in the Monte Carlo analysis is to temporarily 'switch off' the comparison between computed and observed data, thereby generating samples of the prior probability density. This allows us to verify statistically that the algorithm is working correctly, and it allows us to understand the prior information we are using. We will refer to a large collection of models representing the prior probability distribution as the 'prior movie' (in a computer screen, when the models are displayed one after the other, we have a 'movie'). The more models present in this movie, the more accurate the representation of the prior probability density.

6.4 Sampling the Posterior Probability Distribution

If we now switch on the comparison between computed and observed data using, e.g., the Metropolis rule for the actual Eq. (82), the random walk sampling the prior distribution is modified into a walk sampling the posterior distribution.

Since data rarely put strong constraints on the Earth, the 'posterior movie' typically shows that many different models are possible. But even though the models in the posterior movie may be quite different, all of them predict data that, within experimental uncertainties, are models with high likelihood. In other words, we must accept that data alone cannot have a preferred model.

The posterior movie allows us to perform a proper resolution analysis that helps us to choose between different interpretations of a given data set. Using the movie we can answer complicated questions about the correlations between several model parameters. To answer such questions, we can view the posterior movie and try to discover structure that is well resolved by data. Such structure will appear as 'persistent' in the posterior movie.

The 'movie' can be used to answer quite complicated questions. For instance, to answer the question '*Which is the probability that the Earth has this special characteristic, but not having this other special characteristic?*' we can just count the number n of models (samples) satisfying the criterion, and the probability is $P = n/m$, where m is the total number of samples.

Once this 'movie' is generated, it is, of course, possible to represent the 1D or 2D marginal probability densities for all or for some selected parameters: it is enough to concentrate one's attention on those selected parameters in each of the samples generated. Those marginal probability densities may have some pathologies (like being multimodal, or having infinite dispersions), but those are the general characteristics of the joint probability density. Our numerical experience shows that these marginals are, quite often, 'stable' objects, in the sense that they can be accurately determined with only a small number of samples.

If the marginals are, essentially, beautiful bell-shaped distributions, then, one may proceed to merely computing mean values and standard deviations (or median values and mean deviations), using each of the samples and the elementary statistical formulas.

Another, more traditional, way of investigating resolution is to calculate covariances and higher-order moments. For this we need to evaluate integrals of the form

$$R_f = \int_{\mathcal{A}} d\mathbf{m}\ f(\mathbf{m})\ \sigma_m(\mathbf{m}) \tag{85}$$

where $f(\mathbf{m})$ is a given function of the model parameters and $\mathcal{A}$ is an event in the model space $\mathcal{M}$ containing the models we are interested in. For instance,

$$\mathcal{A} = \{\mathbf{m}\,|\text{a given range of parameters in } \mathbf{m} \text{ is } cyclic\}\ . \tag{86}$$

In the special case when $\mathcal{A} = \mathcal{M}$ is the entire model space, and $f(\mathbf{m}) = m_i$, the R_f in Eq. (85) equals the mean $\langle m_i \rangle$ of the ith model parameter m_i . If $f(\mathbf{m}) = (m_i - \langle m_i \rangle)(m_j - \langle m_j \rangle)$, R_f becomes the covariance between the ith and jth model parameters. Typically, in the general inverse problem we cannot evaluate the integral in Eq. (85) analytically because we have no analytical expression for $\sigma(\mathbf{m})$. However, from the samples of the posterior

movie $\mathbf{m}_1, \ldots, \mathbf{m}_n$ we can approximate R_f by the simple average

$$R_f \approx \frac{1}{\text{total number of models}} \sum_{\{i \,|\, \mathbf{m}_i \in \mathcal{A}\}} f(\mathbf{m}_i) \ . \qquad (87)$$

7. Solving Inverse Problems (III): Deterministic Methods

As we have seen, the solution of an inverse problem essentially consists of a probability distribution over the space of all possible models of the physical system under study. In general, this 'model space' is high-dimensional, and the only general way to explore it is by using the Monte Carlo methods developed in Section 3.

If the probability distributions are 'bell-shaped' (i.e., if they look like a Gaussian or like a generalized Gaussian), then one may simplify the problem by calculating only the point around which the probability is maximum, with an approximate estimation of the variances and covariances. This is the problem addressed in this section. Among the many methods available to obtain the point at which a scalar function reaches its maximum value (relaxation methods, linear programming techniques, etc.) we limit our scope here to the methods using the gradient of the function, which we assume can be computed analytically, or at least, numerically. For more general methods, the reader may have a look at Fletcher (1980, 1981), Powell (1981), Scales (1985), Tarantola (1987) or Scales *et al.* (1992).

7.1 Maximum Likelihood Point

Let us consider a space $\mathcal{X}$, with a volume element dV defined. If the coordinates $\mathbf{x} \equiv \{x^1, x^2, \ldots, x^n\}$ are chosen over the space, the volume element has an expression $dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x}$, and each probability distribution over $\mathcal{X}$ can be represented by a probability density $f(\mathbf{x})$. For any fixed small volume ΔV we can search for the point $\mathbf{x}_{\text{ML}}$ such that the probability dP of the small volume, when centered around $\mathbf{x}_{\text{ML}}$, attains a maximum. In the limit $\Delta V \to 0$ this defines the *maximum likelihood point*. The maximum likelihood point may be unique (if the probability distribution is unimodal), may be degenerate (if the probability distribution is 'chevron-shaped'), or may be multiple (as when we have the sum of a few bell-shaped functions).

The maximum likelihood point is *not* the point at which the probability density is maximum. Our definition implies that a maximum must be attained by the ratio between the probability density and the function $v(\mathbf{x})$ defining the volume element:[20]

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff F(\mathbf{x}) = \frac{f(\mathbf{x})}{v(\mathbf{x})} \quad \text{(maximum)} \ . \qquad (88)$$

As the homogeneous probability density is $\mu(\mathbf{x}) = k\, v(\mathbf{x})$ (see Rule 2), we can equivalently define the maximum likelihood point by the condition

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad \text{(maximum)} \ . \qquad (89)$$

The point at which a probability density has its maximum is, in general, not $\mathbf{x}_{\text{ML}}$. In fact, the maximum of a probability density does not correspond to an intrinsic definition of a point: a change of coordinates $\mathbf{x} \mapsto \mathbf{y} = \psi(\mathbf{x})$ would change the probability density $f(\mathbf{x})$ into the probability density $g(\mathbf{y})$ (obtained using the Jacobian rule), but the point of the space at which $f(\mathbf{x})$ is maximum is not the same as the point of the space where $g(\mathbf{y})$ is maximum (unless the change of variables is linear). This contrasts with the maximum likelihood point, as defined by Eq. (89), which is an intrinsically defined point: no matter which coordinates we use in the computation we always obtain the same point of the space.

7.2 Misfit

One of the goals here is to develop gradient-based methods for obtaining the maximum of $F(\mathbf{x}) = f(\mathbf{x})/\mu(\mathbf{x})$. As a quite general rule, gradient-based methods perform quite poorly for (bell-shaped) probability distributions, as when one is far from the maximum the probability densities tend to be quite flat, and it is difficult to get, reliably, the direction of steepest ascent. Taking a logarithm transforms a bell-shaped distribution into a paraboloid-shaped distribution on which gradient methods work well.

The logarithmic volumetric probability, or *misfit*, is defined as $S(\mathbf{x}) = -\log(F(\mathbf{x})/F_0)$, where p' and F_0 are two constants, and is given by

$$S(\mathbf{x}) = -\log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \ . \qquad (90)$$

The problem of maximization of the (typically) bell-shaped function $f(\mathbf{x})/\mu(\mathbf{x})$ has been transformed into the problem of minimization of the (typically) paraboloid-shaped function $S(\mathbf{x})$:

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff S(\mathbf{x}) \quad \text{(minimum)} \ . \qquad (91)$$

Example 19. *The conjunction $\sigma(\mathbf{x})$ of two probability densities $\rho(\mathbf{x})$ and $\vartheta(\mathbf{x})$ was defined (Eq. (13)) as*

$$\sigma(\mathbf{x}) = p\, \frac{\rho(\mathbf{x})\ \vartheta(\mathbf{x})}{\mu(\mathbf{x})} \ . \qquad (92)$$

Then,

$$S(\mathbf{x}) = S_\rho(\mathbf{x}) + S_\vartheta(\mathbf{x}) \ , \qquad (93)$$

where

$$S_\rho(\mathbf{x}) = -\log \frac{\rho(\mathbf{x})}{\mu(\mathbf{x})}\,; \qquad S_\vartheta(\mathbf{x}) = -\log \frac{\vartheta(\mathbf{x})}{\mu(\mathbf{x})}\,. \tag{94}$$

Example 20. *In the context of Gaussian distributions we have found the probability density* (*see Example 12*)

$$\sigma_m(\mathbf{m}) = k \exp\left[-\frac{1}{2}\big((\mathbf{m}-\mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1}(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\big)\right]. \tag{95}$$

The limit of this distribution for infinite variances is a constant, so in this case $\mu_m(\mathbf{m}) = k$ *. The misfit function* $S(\mathbf{m}) = -\log(\sigma_m(\mathbf{m})/\mu_m(\mathbf{m}))$ *is then given by*

$$2\,S(\mathbf{m}) = (\mathbf{m}-\mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\,. \tag{96}$$

The reader should remember that this misfit function is valid only for weakly nonlinear problems (*see Examples 10 and 12*)*. The maximum likelihood model here is the one that minimizes the sum of squares* (*96*)*. This corresponds to the* least-squares criterion.

7.3 Gradient and Direction of Steepest Ascent

One must not consider as synonymous the notions of 'gradient' and 'direction of steepest ascent.' Consider, for instance, an *adimensional* misfit function[21] $S(P,T)$ over a pressure P and a temperature T. Any sensible definition of the gradient of S will lead to an expression like

$$\text{grad}\ S = \begin{pmatrix} \frac{\partial S}{\partial P} \\ \\ \frac{\partial S}{\partial T} \end{pmatrix} \tag{97}$$

and this by no means can be regarded as a 'direction' in the (P,T) space (for instance, the components of this 'vector' does not have the dimensions of pressure and temperature, but of inverse pressure and inverse temperature).

Mathematically speaking, *the gradient of a function* $S(\mathbf{x})$ *at a point* $\mathbf{x}_0$ *is the linear function that is tangent to* $S(\mathbf{x})$ *at* $\mathbf{x}_0$. This definition of gradient is consistent with the more elementary one, based on the use of the first-order expansion

$$S(\mathbf{x}_0+\delta\mathbf{x}) = S(\mathbf{x}_0) + \widehat{\boldsymbol{\gamma}}_0^{\,T}\,\delta\mathbf{x} + \cdots \tag{98}$$

Here $\widehat{\boldsymbol{\gamma}}_0$ is called the gradient of $S(\mathbf{x})$ at point $\mathbf{x}_0$. It is clear that $S(\mathbf{x}_0) + \widehat{\boldsymbol{\gamma}}_0^{\,T}\,\delta\mathbf{x}$ is a linear function, and that it is tangent to $S(\mathbf{x})$ at $\mathbf{x}_0$, so the two definitions are in fact equivalent. Explicitly, the components of the gradient at point $\mathbf{x}_0$ are

$$(\widehat{\boldsymbol{\gamma}}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0)\,. \tag{99}$$

Everybody is well trained in computing the gradient of a function (event if the interpretation of the result as a direction in the original space is wrong). How can we pass from the gradient to the direction of steepest ascent (a bona fide direction in the original space)? In fact, the gradient (at a given point) of a function defined over a given space $\mathcal{E}$) is an element of the dual of the space. To obtain a direction in $\mathcal{E}$ we must pass from the dual to the primal space. As usual, it is the metric of the space that maps the dual of the space into the space itself. So if $\mathbf{g}$ is the metric of the space where $S(\mathbf{x})$ is defined, and if $\widehat{\boldsymbol{\gamma}}$ is the gradient of S at a given point, the *direction of steepest ascent* is

$$\boldsymbol{\gamma} = \mathbf{g}^{-1}\widehat{\boldsymbol{\gamma}}\,. \tag{100}$$

The direction of steepest ascent must be interpreted as follows: if we are at a point $\mathbf{x}$ of the space, we can consider a very small hypersphere around $\mathbf{x}_0$. The direction of steepest ascent points toward the point of the sphere at which $S(\mathbf{x})$ attains its maximum value.

Example 21. *In the context of least squares, we consider a misfit function* $S(\mathbf{m})$ *and a covariance matrix* $\mathbf{C}_M$ *. If* $\hat{\boldsymbol{\gamma}}_0$ *is the gradient of* S *, at a point* $\mathbf{x}_0$ *, and if we use* $\mathbf{C}_M$ *to define distances in the space, the direction of steepest ascent is*

$$\boldsymbol{\gamma}_0 = \mathbf{C}_M\,\widehat{\boldsymbol{\gamma}}_0\,. \tag{101}$$

7.4 The Steepest Descent Method

Consider that we have a probability distribution defined over an n-dimensional space $\mathcal{X}$. Having chosen the coordinates $\mathbf{x} \equiv \{x^1,\ x^2, \ldots, x^n\}$ over the space, the probability distribution is represented by the probability density $f(\mathbf{x})$ whose homogeneous limit (in the sense developed in Section 2.2) is $\mu(\mathbf{x})$. We wish to calculate the coordinates $\mathbf{x}_{\text{ML}}$ of the maximum likelihood point. By definition (Eq. (89)),

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad \text{(maximum)}\,, \tag{102}$$

that is,

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff S(\mathbf{x}) \quad \text{(minimum)}\,, \tag{103}$$

where $S(\mathbf{x})$ is the misfit [Eq. (90)]

$$S(\mathbf{x}) = -k\,\log\frac{f(\mathbf{x})}{\mu(\mathbf{x})}\,. \tag{104}$$

Let us denote by $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ the gradient of $S(\mathbf{x})$ at point $\mathbf{x}_k$, i.e. (Eq. (99)),

$$(\widehat{\boldsymbol{\gamma}}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0) \ . \tag{105}$$

We have seen above that $\widehat{\boldsymbol{\gamma}}(\mathbf{x})$ should not be interpreted as a direction in the space $\mathcal{X}$ but as a direction in the dual space. The gradient can be converted into a direction using a metric $\mathbf{g}(\mathbf{x})$ over $\mathcal{X}$. In simple situations the metric $\mathbf{g}$ will be the one used to define the volume element of the space, i.e., we will have $\mu(\mathbf{x}) = k\, v(\mathbf{x}) = k\sqrt{\det \mathbf{g}(\mathbf{x})}$, but this is not a necessity, and iterative algorithms may be accelerated by astute introduction of ad-hoc metrics.

Given, then, the gradient $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ (at some particular point $\mathbf{x}_k$) to any possible choice of metric $\mathbf{g}(\mathbf{x})$ we can define the direction of steepest ascent associated with the metric $\mathbf{g}$, by (Eq. (101))

$$\boldsymbol{\gamma}(\mathbf{x}_k) = \mathbf{g}^{-1}(\mathbf{x}_k)\, \widehat{\boldsymbol{\gamma}}(\mathbf{x}_k) \ . \tag{106}$$

The algorithm of steepest descent is an iterative algorithm passing from point $\mathbf{x}_k$ to point $\mathbf{x}_{k+1}$ by making a 'small jump' along the local direction of steepest descent,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \varepsilon_k \mathbf{g}_k^{-1}\, \widehat{\boldsymbol{\gamma}}_k \ , \tag{107}$$

where ε_k is an ad-hoc (real, positive) value adjusted to force the algorithm to converge rapidly (if ε_k is chosen too small, the convergence may be too slow; it is chosen too large, the algorithm may even diverge).

Many elementary presentations of the steepest descent algorithm just forget to include the metric $\mathbf{g}_k$ in expression (107). These algorithms are not consistent. Even the physical dimensionality of the equation is not assured. 'Numerical' problems in computer implementations of steepest descent algorithms can often be traced to the fact that the metric has been neglected.

Example 22. *In the context of Example 20, where the misfit function $S(\mathbf{m})$ is given by*

$$\begin{aligned} 2\, S(\mathbf{m}) = &(\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t\ \mathbf{C}_D^{-1}\, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \\ &+ (\mathbf{m} - \mathbf{m}_{\text{prior}})^t\ \mathbf{C}_M^{-1}\, (\mathbf{m} - \mathbf{m}_{\text{prior}}) \ , \end{aligned} \tag{108}$$

the gradient $\widehat{\boldsymbol{\gamma}}$, whose components are $\widehat{\boldsymbol{\gamma}}_\alpha = \partial S/\partial m^\alpha$, is given by the expression

$$\widehat{\boldsymbol{\gamma}}(\mathbf{m}) = \mathbf{F}^t(\mathbf{m})\, \mathbf{C}_D^{-1}\, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1}\, (\mathbf{m} - \mathbf{m}_{\text{prior}}) \ , \tag{109}$$

where $\mathbf{F}$ is the matrix of partial derivatives

$$F^{i\alpha} = \frac{\partial f^i}{\partial m^\alpha} \ . \tag{110}$$

An example of computation of partial derivatives is given in Appendix M.

Example 23. *In the context of Example 22 the model space $\mathcal{M}$ has an obvious metric, namely, that defined by the inverse of the 'a priori' covariance operator $\mathbf{g} = \mathbf{C}_M^{-1}$. Using this metric and the gradient given by Eq. (109), the steepest descent algorithm (107) becomes*

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k \left(\mathbf{C}_M\, \mathbf{F}_k^t\, \mathbf{C}_D^{-1}\, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \ , \tag{111}$$

where $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ and $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$. The real positive quantities ε_k can be fixed after some trial and error by accurate linear search, or by using a linearized approximation.

Example 24. *In the context of Example 22 the model space $\mathcal{M}$ has a less obvious metric, namely, that defined by the inverse of the 'posterior' covariance operator, $\mathbf{g} = \widetilde{\mathbf{C}}_M^{-1}$.*[23] *Using this metric and the gradient given by Eq. (109), the steepest descent algorithm (107) becomes*

$$\begin{aligned} \mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k &\left(\mathbf{F}_k^t\, \mathbf{C}_D^{-1}\, \mathbf{F}_k + \mathbf{C}_M^{-1} \right)^{-1} \left(\mathbf{F}_k^t\, \mathbf{C}_D^{-1}\, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) \right. \\ &\left. + \mathbf{C}_M^{-1}\, (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \ , \end{aligned} \tag{112}$$

where $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ and $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$. The real positive quantities ε_k can be fixed, after some trial and error, by accurate linear search, or by using a linearized approximation that simply gives[24] *$\varepsilon_k \approx 1$.*

The algorithm (112) is usually called a 'quasi-Newton algorithm.' This name is not well chosen: a Newton method applied to minimization of a misfit function $S(\mathbf{m})$ would be a method using the second derivatives of $S(\mathbf{m})$, and thus the derivatives $H^i_{\alpha\beta} = \frac{\partial^2 f^i}{\partial m^\alpha\, \partial m^\beta}$, that are not computed (or not estimated) when using this algorithm. It is just a steepest descent algorithm with a nontrivial definition of the metric in the working space. In this sense it belongs to the wider class of 'variable metric methods,' not discussed in this article.

7.5 Estimating Posterior Uncertainties

In the Gaussian context, the Gaussian probability density that is tangent to $\sigma_m(\mathbf{m})$ has its center at the point given by the iterative algorithm

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k \left(\mathbf{C}_M\, \mathbf{F}_k^t\, \mathbf{C}_D^{-1} (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \ , \tag{113}$$

(Eq. (111)) or, equivalently, by the iterative algorithm

$$\begin{aligned} \mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k &\left(\mathbf{F}_k^t\, \mathbf{C}_D^{-1}\, \mathbf{F}_k + \mathbf{C}_M^{-1} \right)^{-1} \left(\mathbf{F}_k^t\, \mathbf{C}_D^{-1} (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) \right. \\ &\left. + \mathbf{C}_M^{-1}\, (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \end{aligned} \tag{114}$$

(Eq. (112)). The covariance of the tangent Gaussian is

$$\widetilde{\mathbf{C}}_M \approx \left(\mathbf{F}_\infty^t\, \mathbf{C}_D^{-1}\, \mathbf{F}_\infty + \mathbf{C}_M^{-1}\right)^{-1}, \qquad (115)$$

where $\mathbf{F}_\infty$ refers to the value of the matrix of partial derivatives at the convergence point.

7.6 Some Comments on the Use of Deterministic Methods

7.6.1 Linear , Weakly Nonlinear and Nonlinear Problems

There are different degrees of nonlinearity. Figure 8 illustrates four domains of nonlinearity, calling for different

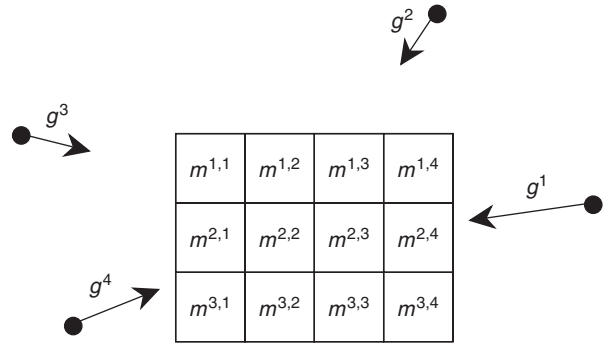


FIGURE 8 A simple example where we are interested in predicting the gravitational field **g** generated by a 2D distribution of mass.

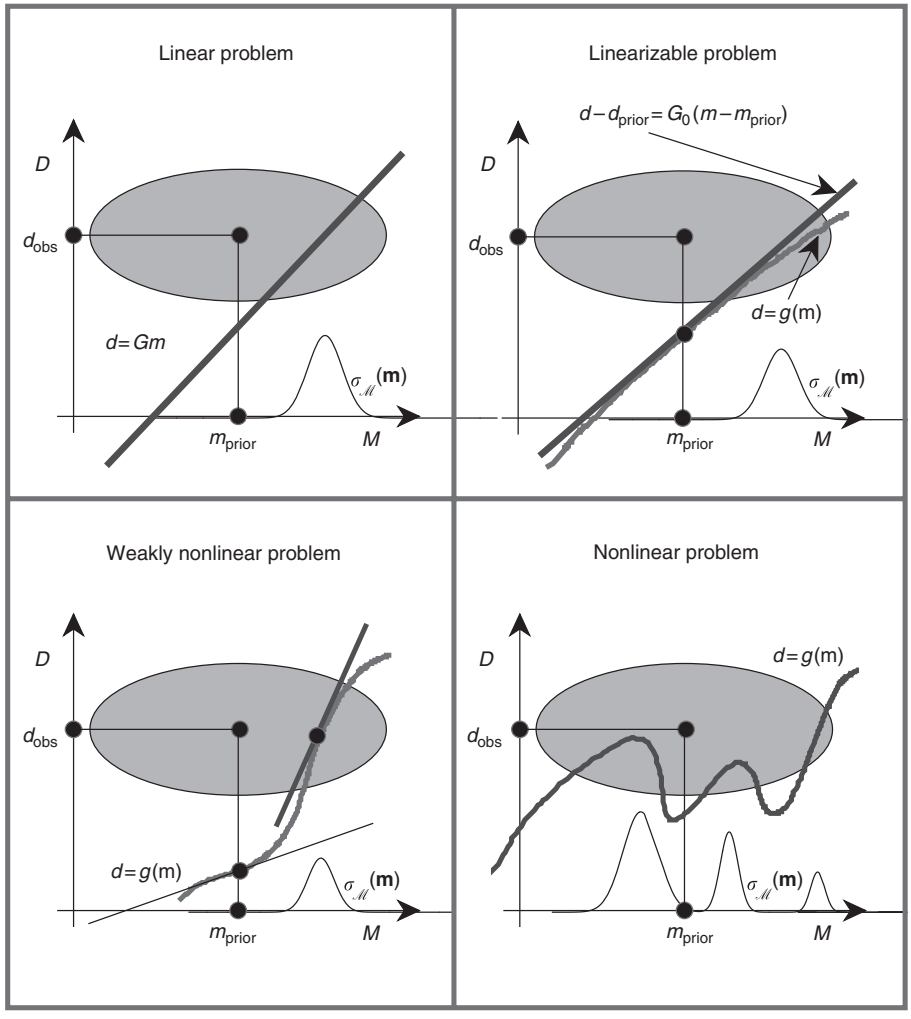


FIGURE 9 Illustration of the four domains of nonlinearity, calling for different optimization algorithms. The model space is symbolically represented by the abscissa, and the data space is represented by the ordinate. The gray oval represents the combination of prior information on the model parameters and information from the observed data. What is important is not an intrinsic nonlinearity of the function relating model parameters to data, but how linear the function is *inside the domain of significant probability*.

optimization algorithms. In this figure the abscissa symbolically represents the model space, and the ordinate represents the data space. The gray oval represents the combination of prior information on the model parameters, and information from the observed data.[25] It is the probability density $\rho(\mathbf{d}, \mathbf{m}) = \rho_d(\mathbf{d})\rho_m(\mathbf{m})$ seen elsewhere.

To fix ideas, the oval suggests here a Gaussian probability, but our distinction between problems according to their nonlinearity will not depend fundamentally on this.

First, there are strictly linear problems. For instance, in the example illustrated by Figure 8 the gravitational field $\mathbf{g}$ depends linearly on the masses inside the blocks.[26]

Strictly linear problems are illustrated at the top left of Figure 9. The linear relationship between data and model parameters, $\mathbf{d} = \mathbf{G}\,\mathbf{m}$, is represented by a straight line. The prior probability density $\rho(\mathbf{d}, \mathbf{m})$ 'induces' on this straight line the posterior probability density[27] $\sigma(\mathbf{d}, \mathbf{m})$ whose 'projection' over the model space gives the posterior probability density over the model parameter space, $\sigma_m(\mathbf{m})$. Should the prior probability densities be Gaussian, then the posterior probability distribution would also be Gaussian: this is the simplest situation.

Quasi-linear problems are illustrated at the bottom left of Figure 9. If the relationship linking the observable data $\mathbf{d}$ to the model parameters $\mathbf{m}$,

$$\mathbf{d} = \mathbf{g}(\mathbf{m}) \ , \tag{116}$$

is approximately linear *inside the domain of significant prior probability* (i.e., inside the gray oval of the figure), then the posterior distribution is just as simple as the prior distribution. For instance, if the prior is Gaussian the posterior is also Gaussian.

In this case also, the problem can be reduced to the computation of the mean and the covariance of the Gaussian. Typically, one begins at some 'starting model' $\mathbf{m}_0$ (typically, one takes for $\mathbf{m}_0$ the 'a priori model' $\mathbf{m}_{\text{prior}}$),[28] linearizing the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around $\mathbf{m}_0$, and one looks for a model $\mathbf{m}_1$ 'better than $\mathbf{m}_0$.'

Iterating such an algorithm, one tends to the model $\mathbf{m}_\infty$ at which the 'quasi-Gaussian' $\sigma_m(\mathbf{m})$ is maximum. The linearizations made in order to arrive to $\mathbf{m}_\infty$ are so far not an approximation: the point $\mathbf{m}_\infty$ is perfectly defined, independently of any linearization and any method used to find it. But once the convergence to this point has been obtained, a linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around this point,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_\infty) = \mathbf{G}_\infty(\mathbf{m} - \mathbf{m}_\infty) \ , \tag{117}$$

allows to obtain a good approximation to the posterior uncertainties. For instance, if the prior distribution is Gaussian this will give the covariance of the 'tangent Gaussian.'

Between linear and quasi-linear problems there are the 'linearizable problems.' The scheme at the top right of Figure 9 shows the case where the linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around the prior model,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_{\text{prior}}) = \mathbf{G}_{\text{prior}}(\mathbf{m} - \mathbf{m}_{\text{prior}}) \ , \tag{118}$$

gives a function that, inside the domain of significant probability, is very similar to the true (nonlinear) function.

In this case, there is no practical difference between this problem and the strictly linear problem, and the iterative procedure necessary for quasi-linear problems is here superfluous.

It remains to analyze the true nonlinear problems that, using a pleonasm, are sometimes called *strongly nonlinear problems*. They are illustrated at the bottom right of Figure 9.

In this case, even if the prior distribution is simple, the posterior distribution can be quite complicated. For instance, it can be multimodal. These problems are in general quite complex to solve, and only a Monte Carlo analysis, as described in the previous chapter, is feasible.

If full Monte Carlo methods cannot be used, because they are too expensive, then one can mix a random part (for instance, to choose the starting point) and a deterministic part. The optimization methods applicable to quasi-linear problems can, for instance, allow us to go from the randomly chosen starting point to the 'nearest' optimal point. Repeating these computations for different starting points, one can arrive at a good idea of the posterior distribution in the model space.

7.6.2 The Maximum Likelihood Model

The *most likely model* is, by definition, that at which the volumetric probability (see Appendix A) $\sigma_\beta(\mathbf{m})$ attains its maximum. As $\sigma_\beta(\mathbf{m})$ is maximum when $S(\mathbf{m})$ is minimum, we see that the most likely model is also the 'best model' obtained when using a 'least-squares criterion.' Should we have used the double exponential model for all the

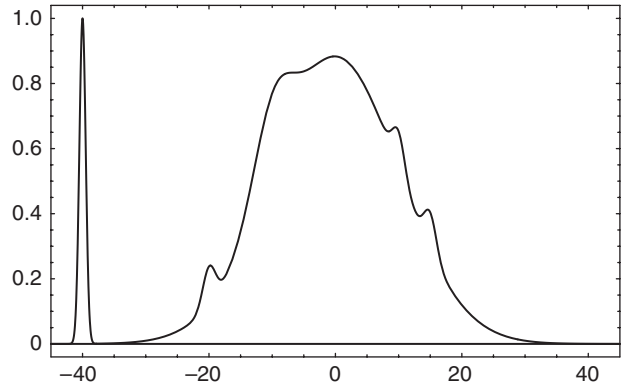


FIGURE 10 One of the circumstances where the 'maximum likelihood model' may not be very interesting is when it corresponds to a narrow maximum with small total probability, as the peak in the left part of this probability distribution.

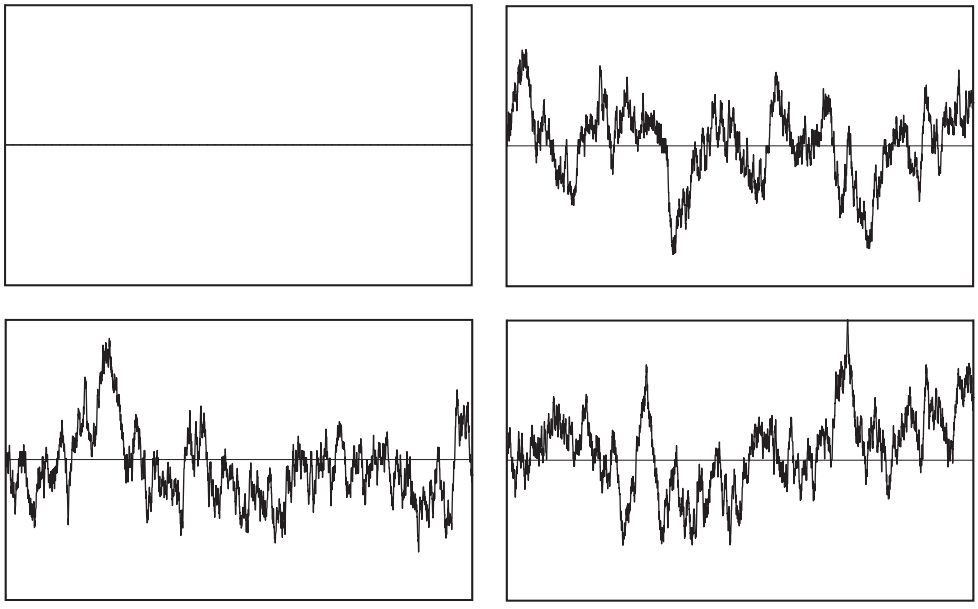


FIGURE 11 At the right, three random realizations of a Gaussian random function with zero mean and (approximately) exponential correlation function. The most likely function, i.e., the center of the Gaussian, is shown at the top left. We see that the most likely function is not a representative of the probability distribution.

uncertainties, then the most likely model would be defined by a 'least absolute values' criterion.

There are many circumstances where the most likely model is not an interesting model. One trivial example is when the volumetric probability has a 'narrow maximum,' with small total probability (see Fig. 10). A much less trivial situation arises when the number of parameters is very large, as for instance when we deal with a random function (that, strictly speaking, corresponds to an infinite number of random variables). Figure 11, for instance, shows a few realizations of a Gaussian function with zero mean and an (approximately) exponential correlation. The most likely function is the center of the Gaussian, i.e., the null function shown at the top left. But this is not a representative sample of the probability distribution, as any realization of the probability distribution will have, with a probability very close to one, the 'oscillating' characteristics of the three samples shown at the right.

8. Conclusions

Probability theory is well adapted to the formulation of inverse problems, although its formulation must be rendered intrinsic (introducing explicitly the definition of distances in the working spaces, by redefining the notion of conditional probability density, and by introducing the notion of conjunction of states of information). The Metropolis algorithm is well adapted to the solution of inverse problems, as its inherent structure allows us to sequentially combine prior information, theoretical information, etc., and allows us to take advantage of the 'movie philosophy.' When a general Monte Carlo approach cannot be afforded, one can use simplified optimization techniques (like least squares). However, this usually requires strong simplifications that can only be made at the cost of realism.

Acknowledgements

We are very indebted to our colleagues (Bartolomé Coll, Miguel Bosch, Guillaume Évrard, John Scales, Christophe Barnes, Frédéric Parrenin, and Bernard Valette) for illuminating discussions. We are also grateful to the students of the Geophysical Tomography Group, and the students at our respective institutes (in Paris and Copenhagen).

References

Aki, K. and W.H.K. Lee (1976). Determination of three-dimensional velocity anomalies under a seismic array using first *P* arrival times from local earthquakes. *J. Geophys. Res.* **81**, 4381–4399.

Aki, K., A. Christofferson, and E.S. Husebye (1977). Determination of the three-dimensional seismic structure of the lithosphere. *J. Geophys. Res.* **82**, 277–296.

Backus, G. (1970a). Inference from inadequate and inaccurate data: I. *Proc. Natl. Acad. Sci. USA* **65**(1), 1–105.

Backus, G. (1970b). Inference from inadequate and inaccurate data: II. *Proc. Natl. Acad. Sci. USA* **65**(2), 281–287.

Backus, G. (1970c). Inference from inadequate and inaccurate data: III. *Proc. Natl. Acad. Sci. USA* **67**(1), 282–289.

Backus, G. (1971). Inference from inadequate and inaccurate data. 'Mathematical Problems in the Geophysical Sciences,' Lectures in Applied Mathematics **14**. American Mathematical Society, Providence, Rhode Island.

Backus, G. and F. Gilbert (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astron. Soc.* **13**, 247–276.

Backus, G. and F. Gilbert (1968). The resolving power of gross Earth data. *Geophys. J. R. Astron. Soc.* **16**, 169–205.

Backus, G. and F. Gilbert (1970). Uniqueness in the inversion of inaccurate gross Earth data. *Philos. Trans. R. Soc. London* **266**, 123–192.

Dahlen, F.A. (1976). Models of the lateral heterogeneity of the Earth consistent with eigenfrequency splitting data. *Geophys. J. R. Astron. Soc.* **44**, 77–105.

Dahl-Jensen, D., K. Mosegaard, N. Gundestrup, G.D. Clow, S.J. Johnsen, A.W. Hansen, and N. Balling (1998). Past temperatures directly from the Greenland Ice Sheet. *Science* (Oct. 9), 268–271.

Fisher, R.A. (1953). Dispersion on a sphere. *Proc. R. Soc. London, A* **217**, 295–305.

Fletcher, R. (1980). 'Practical Methods of Optimization,' Vol. 1: Unconstrained Optimization. Wiley, New York.

Fletcher, R. (1981). 'Practical Methods of Optimization,' Vol. 2: Constrained Optimization. Wiley, New York.

Franklin, J.N. (1970). Well posed stochastic extensions of ill posed linear problems. *J. Math. Anal. Appl.* **31**, 682–716.

Gauss, C.F. (1809). 'Theoria Motus Corporum Cœlestium in Sectionis Conicis Solem Ambientum,' Hamburg. (Also in 'Werke,' Vol. 7. Olmc-Verlag, 1981.)

Geiger, L. (1910). Herdbestimmung bei Erdbeben aus den Ankunftszeiten. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen* **4**, 331–349.

Geman, S. and D. Geman (1984). *IEEE Trans. Pattern Anal. Mach. Int.* **PAMI-6**(6), 721.

Gilbert, F. (1971). Ranking and winnowing gross Earth data for inversion and resolution. *Geophys. J. R. Astron. Soc.* **23**, 215–128.

Hadamard, J. (1902). Sur les problémes aux dérivées partielles et leur signification physique. *Bull. Univ. Princeton* **13**.

Hadamard, J. (1932). 'Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques.' Hermann, Paris.

ISO (1993). 'Guide to the Expression of Uncertainty in Measurement.' International Organization for Standardization, Switzerland.

Jackson, D.D. (1979). The use of a priori data to resolve non-uniqueness in linear inversion. *Geophys. J. R. Astron. Soc.* **57**, 137–157.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* **SSC-4**(3), 227–241.

Jeffreys, H. (1939). 'Theory of Probability.' Clarendon Press, Oxford. (Reprinted in 1961 by Oxford University Press.)

Kandel, A. (1986). 'Fuzzy Mathematical Techniques with Applications.' Addison-Wesley, Reading, MA.

Keilis-Borok, V.J. and T.B. Yanovskaya (1967). Inverse problems in seismology (structural review). *Geophys. J. R. Astron. Soc.* **13**, 223–234.

Kennett, B.L.N. and G. Nolet (1978). Resolution analysis for discrete systems. *Geophys. J. R. Astron. Soc.* **53**, 413–425.

Khan, A., K. Mosegaard, and K.L. Rasmussen (2000). A new seismic velocity model for the Moon from a Monte Carlo inversion of the Apollo lunar seismic data. *Geophys. Res. Lett.* **37**(11), 1591–1594.

Kimeldorf, G. and G. Wahba (1970). A correspondence between Bayesian estimation of stochastic processes and smooting by splines. *Ann. Math. Stat.* **41**, 495–502.

Kullback, S. (1967). The two concepts of information. *J. Am. Stat. Assoc.* **62**, 685–686.

Lehtinen, M.S., L. Päivärinta, and E. Somersalo (1989). Linear inverse problems for generalized random variables. *Inverse Prob.* **5**, 599–612.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least-squares. *Q. Appl. Math.* **2**, 164–168.

Marquardt, D.W. (1963). An algorithm for least squares estimation of nonlinear parameters, SIAM J., **11**, 431–441.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and non-linear estimation. *Technometrics* **12**, 591–612.

Menke, W. (1984). 'Geophysical Data Analysis: Discrete Inverse Theory.' Academic Press, New York.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1**, 1087–1092.

Minster, J.B. and T.M. Jordan (1978). Present-day plate motions. *J. Geophys. Res.* **83**, 5331–5354.

Mosegaard, K. and C. Rygaard-Hjalsted (1999). Bayesian analysis of implicit inverse problems. *Inverse Probl.* **15**, 573–583.

Mosegaard, K. and A. Tarantola (1995). Monte Carlo sampling of solutions to inverse problems. *J. Geophys. Res.* **100**, 12,431–12,447.

Mosegaard, K., S.C. Singh, D. Snyder, and H. Wagner (1997). Monte Carlo analysis of seismic reflections from Moho and the W-reflector. *J. Geophys. Res. B* **102**, 2969–2981.

Nolet, G. (1990). Partitioned wave-form inversion and 2D structure under the NARS array. *J. Geophys. Res.* **95**, 8499–8512.

Nolet, G., J. van Trier, and R. Huisman (1986). A formalism for nonlinear inversion of seismic surface waves. *Geophys. Res. Lett.* **13**, 26–29.

Parker, R.L. (1994). 'Geophysical Inverse Theory.' Princeton University Press, Princeton, NJ.

Parzen, E., K. Tanabe, and G. Kitagawa (Eds.) (1998). 'Selected Papers of Hirotugu Akaike,' Springer Series in Statistics. Springer-Verlag, New York.

Powell, M.J.D. (1981). 'Approximation Theory and Methods.' Cambridge University Press, Cambridge.

Press, F. (1968). Earth models obtained by Monte Carlo inversion. *J. Geophys. Res.* **73**, 5223–5234.

Rietsch, E. (1977). The maximum entropy approach to inverse problems. *J. Geophys.* **42**, 489–506.

Scales, L.E. (1985). 'Introduction to Non-Linear Optimization.' Macmillan, London.

Scales, J.A., M.L. Smith, and T.L. Fischer (1992). Global optimization methods for multimodal inverse problems. *J. Comput. Phys.* **102**, 258–268.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.

Su, W.-J., R.L. Woodward, and A.M. Dziewonski (1992). Deep origin of mid-oceanic ridge velocity anomalies. *Nature* **360**, 149−152.

Tarantola, A. and B. Valette (1982a). Inverse problems = quest for information. *J. Geophys.* **50**, 159−170.

Tarantola, A. and B. Valette (1982b). Generalized nonlinear inverse problems solved using the least-squares criterion. *Rev. Geophys. Space Phys.* **20**, 219−232.

Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics* **49**, 1259−1266.

Tarantola, A. (1986). A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics* **51**, 1893−1903.

Tarantola, A. (1987). 'Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.' Elsevier, Amsterdam.

Taylor, A.E. and Lay, D.C. (1980). 'Introduction to Functional Analysis', John Wiley and Sons, New York.

Taylor, B.N. and C.E. Kuyatt (1994). Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST Technical note 1297.

Tikhonov, A.N. (1963). Resolution of ill-posed problems and the regularization method (in Russian). *Dokl. Akad. Nauk SSSR* **151**, 501−504.

van der Hilst, R.D., S. Widiyantoro, and E.R. Engdahl (1997). Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578−584.

Wiggins, R.A. (1969). Monte Carlo inversion of body-wave observations. *J. Geophys. Res.* **74**, 3171−3181.

Wiggins, R.A. (1972). The general linear inverse problem: implication of surface waves and free oscillations for Earth structure. *Rev. Geophys. Space Phys.* **10**, 251−285.

Woodhouse, J.H. and F.A. Dahlen (1978). The effect of general aspheric perturbation on the free oscillations of the Earth. *Geophys. J. R. Astron. Soc.* **53**, 335−354.

Notes

1. For instance, we could fit our observations with a heterogeneous but isotropic Earth model or, alternatively, with a homogeneous but anisotropic Earth.
2. Preliminary Earth Reference Model (PREM), Dziewonski and Anderson, PEPI, 1981. Inversion for Centroid Moment Tensor (CMT), Dziewonski, Chou and Woodhouse, JGR, 1982. First global tomographic model, Dziewonski, JGR, 1984.
3. The capacity element associated to the vector elements $d\mathbf{r}_1$, $d\mathbf{r}_2, \dots d\mathbf{r}_n$ is defined as $d\tau = \varepsilon_{ij\dots k}\, dr_1^i\, dr_2^j \dots dr_n^k$, where $\varepsilon_{ij\dots k}$ is the Levi-Civita capacity (whose components take the values $\{0, \pm 1\}$). If the metric tensor of the space is $\mathbf{g}(\mathbf{x})$, then $\eta_{ij\dots k} = \sqrt{\det \mathbf{g}}\, \varepsilon_{ij\dots k}$ is a true tensor, as it is the product of a density $\sqrt{\det \mathbf{g}}$ by a capacity $\varepsilon_{ij\dots k}$. Then, the volume element, defined as $dV = \eta_{ij\dots k}\, dr_1^i dr_2^j \dots dr_n^k = \sqrt{\det \mathbf{g}}\, d\tau$, is a (true) scalar.
4. This is a property that is valid for any coordinate system that can be chosen over the space.
5. As a counterexample, the distance defined as $ds = |dx| + |dy|$ is not of the L_2 type (it is L_1).
6. This solves the complete problem for isotropic tensors only. It is beyond the scope of this text to propose rules valid for general anisotropic tensors: the necessary mathematics have not yet been developed.
7. The definition of the elastic constants was made before the tensorial structure of the theory was understood. Seismologists today should not use, at a theoretical level, parameters like the first Lamé coefficient λ or the Poisson ratio. Instead they should use κ and μ (and their inverses). In fact, our suggestion in this IASPEI volume is to use the true eigenvalues of the stiffness tensor, $\lambda_\kappa = 3\kappa$, and $\lambda_\mu = 2\mu$, which we propose to call the *eigen-bulk-modulus* and the *eigen-shear-modulus*, respectively.
8. Assume that $p(\mathbf{x})$ and $q(\mathbf{x})$ are normalized by $\int_{\mathcal{X}} d\mathbf{x}\, p(\mathbf{x}) = \mathbf{1}$ and $\int_{\mathcal{X}} d\mathbf{x}\, \mathbf{q}(\mathbf{x}) = \mathbf{1}$. Then, irrespective of the normalizability of $\mu(\mathbf{x})$ (as explained above, $p(\mathbf{x})$ and $q(\mathbf{x})$ are assumed to be absolutely continuous with respect to the homogeneous distribution), $(p \wedge q)(\mathbf{x})$ is normalizable, and its normalized expression is
$$(p \wedge q)(\mathbf{x}) = \frac{p(\mathbf{x})\, q(\mathbf{x})/\mu(\mathbf{x})}{\int_{\mathcal{X}} d\mathbf{x}\, p(\mathbf{x})\, q(\mathbf{x})/\mu(\mathbf{x})} \,.$$
9. As a counter example, working at the surface of the sphere with geographical coordinates $(\mathbf{u}, \mathbf{v}) = (u, v) = (\vartheta, \varphi)$ this condition is **not** fulfilled, as $g_\varphi = \sin\theta$ is a function of ϑ: the surface of the sphere is not the Cartesian product of two 1D spaces.
10. That is, series of numbers that appear random if tested with any reasonable statistical test.
11. To see this, put $f(\mathbf{x}) = \mathbf{1}$, $\mu(\mathbf{x}) = \mathbf{1}$, and $g(\mathbf{x}) = \dfrac{\exp(-E(\mathbf{x})/T)}{\int \exp(-E(\mathbf{x})/T) d\mathbf{x}}$, where $E(\mathbf{x})$ is an 'energy' associated to the point $\mathbf{x}$, and T is a 'temperature'. The summation in the denominator is over the entire space. In this way, our acceptance rule becomes the classical Metropolis rule: point $\mathbf{x}_i$ is always accepted if $E(\mathbf{x}_i) \le E(\mathbf{x}_j)$, but if $E(\mathbf{x}_i) > E(\mathbf{x}_j)$, it is only accepted with probability $p_{ij}^{\text{acc}} = \exp\big(-\big(E(\mathbf{x}_i) - E(\mathbf{x}_j)\big)/T\big)$.
12. A numerical method is called robust if it is not sensitive to a small number of large errors.
13. It would be violated, for instance, if we use the pair of elastic parameters longitudinal wave velocity − shear wave velocity, as the volume element in the space of elastic wave velocities does not factorize (see Appendix H).
14. We use here the properties $\log \sqrt{\mathbf{A}} = \frac{1}{2} \log \mathbf{A}$, and $\det \mathbf{AB} = \det \mathbf{BA}$
15. Typically, this may happen because the derivatives $\mathbf{F}$ are small or because the variances in $\mathbf{C}_M$ are large.
16. We first use $\log \det \mathbf{A} = \operatorname{trace} \log \mathbf{A}$, and then the series expansion of the logarithm of an operator, $\log(\mathbf{I} + \mathbf{A}) = \mathbf{A} - \frac{1}{2}\mathbf{A}^2 + \cdots$
17. Practically, it may correspond to the output of some 'black box' solving the 'forward problem'.
18. Remember that, even if we wish to use a simple method based on the notion of conditional probability density, an analytic expression like $\mathbf{d} = \mathbf{f}(\mathbf{m})$ needs some 'thickness' before going to the limit defining the conditional probability density. This limit crucially depends on the 'thickness', i.e., on the type of uncertainties the theory contains.
19. Note that taking the limit of $\vartheta(x,t)$ or of $\rho(x,t)$ for infinite variances we obtain $\mu(x,t)$, as we should.
20. The ratio $F(\mathbf{x}) = f(\mathbf{x})\, v(\mathbf{x})$ is what we refer to as *the volumetric probability* associated to the probability density $f(\mathbf{x})$. See Appendix A.
21. We take this example because typical misfit functions are adimensional (have no physical dimensions) but the argument has general validity.

22. As shown in Tarantola (1987), if $\boldsymbol{\gamma}_k$ is the direction of steepest ascent at point $\mathbf{m}_k$, i.e., $\gamma_k = \mathbf{C}_M \, \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}})$, then, a local linearized approximation for the optimal ε_k gives

$$\varepsilon_k = \frac{\boldsymbol{\gamma}_k^t \mathbf{C}_M^{-1} \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^t (\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1}) \, \boldsymbol{\gamma}_k} .$$

23. The 'best estimator' of $\tilde{\mathbf{C}}_M$ is

$$\tilde{\mathbf{C}}_M \approx \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1} \right)^{-1} . \qquad (119)$$

See, e.g., Tarantola (1987).

24. While a sensible estimation of the optimal values of the real positive quantities ε_k is crucial for the algorithm 111, they can in many usual circumstances be dropped from the algorithm 113.
25. The gray oval is the product of the probability density over the model space, representing the prior information, and the probability density over the data space representing the experimental results.
26. The gravitational field at point $\mathbf{x}_0$ generated by a distribution of volumetric mass $\rho(\mathbf{x})$ is given by

$$\mathbf{g}(\mathbf{x}_0) = \int dV(\mathbf{y}) \frac{\mathbf{x}_0 - \mathbf{y}}{\|\mathbf{x}_0 - \mathbf{x}\|^3} \, \rho(\mathbf{x}) .$$

When the volumetric mass is constant inside some predefined (2D) volumes, as suggested in Figure 8, this gives

$$\mathbf{g}(\mathbf{x}_0) = \sum_A \sum_B \mathbf{G}^{\text{A,B}}(\mathbf{x}_0) \, m^{\text{A,B}} .$$

This is a strictly linear equation between data (the gravitational field at a given observation point) and the model parameters (the masses inside the volumes). Note that if instead of choosing as model parameters the total masses inside some predefined volumes one chooses the geometrical parameters defining the sizes of the volumes, then the gravity field is not a linear function of the parameters. More details can be found in Tarantola and Valette (1982b, page 229).

27. Using the 'orthogonal-limit' method described in Section 2.4.
28. The term 'a priori model' is an abuse of language. The correct term is 'mean a priori model'.

Editor's Note

Appendixes A–P are placed on the attached Handbook CD, under the directory \16Mosegaard. An introduction to probability concepts is given in Chapter 82, Statistical Principles for Seismologists, by Vere-Jones and Ogata. See also Chapter 52, Probing the Earth's Interior with Seismic Tomography, by Curtis and Snieder.

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 107, NO. E6, 10.1029/2001JE001658, 2002

An inquiry into the lunar interior: A nonlinear inversion of the Apollo lunar seismic data

A. Khan

Department of Geophysics, Niels Bohr Institute, University of Copenhagen, Denmark

Département de Géophysique Spatiale et Planétaire, Institut de Physique du Globe de Paris, France

K. Mosegaard

Department of Geophysics, Niels Bohr Institute, University of Copenhagen, Denmark

Received 22 September 2001; revised 6 December 2001; accepted 6 December 2001; published 11 June 2002.

[1] This study discusses in detail the inversion of the Apollo lunar seismic data and the question of how to analyze the results. The well-known problem of estimating structural parameters (seismic velocities) and other parameters crucial to an understanding of a planetary body from a set of arrival times is strongly nonlinear. Here we consider this problem from the point of view of Bayesian statistics using a Markov chain Monte Carlo method. Generally, the results seem to indicate a somewhat thinner crust with a thickness around 45 km as well as a more detailed lunar velocity structure, especially in the middle mantle, than obtained in earlier studies. Concerning the moonquake locations, the shallow moonquakes are found in the depth range 50–220 km, and the majority of deep moonquakes are concentrated in the depth range 850–1000 km, with what seems to be an apparently rather sharp lower boundary. In wanting to further analyze the outcome of the inversion for specific features in a statistical fashion, we have used credible intervals, two-dimensional marginals, and Bayesian hypothesis testing. Using this form of hypothesis testing, we are able to decide between the relative importance of any two hypotheses given data, prior information, and the physical laws that govern the relationship between model and data, such as having to decide between a thin crust of 45 km and a thick crust as implied by the generally assumed value of 60 km. We obtain a Bayes factor of 4.2, implying that a thinner crust is strongly favored. *INDEX TERMS:* 6250 Planetology: Solar System Objects: Moon (1221); 5430 Planetology: Solid Surface Planets: Interiors (8147); 3260 Mathematical Geophysics: Inverse theory; 5455 Planetology: Solid Surface Planets: Origin and evolution; *KEYWORDS:* Moon, inverse theory, seismology, interior structure, terrestrial planets

1. Introduction

[2] Using seismology to obtain information about the interior of the Moon saw its advent with the U.S. Apollo missions which were undertaken from July 1969 to December 1972. Seismic stations were deployed at five of the six locations (Apollo 17 did not carry a seismometer) as part of the integrated set of geophysical experiments called the Apollo Lunar Surface Experiment Package (ALSEP). Only four of these five stations (12, 14, 15, and 16), powered by radioactive thermal generators, operated concurrently as a four-station seismic array, which was operative from April 1972 until 30 September 1977, when transmission of seismic data was suspended. Each seismic package consisted of three long-period (LP) seismometers aligned orthogonally to measure one vertical (Z) and two horizontal (X and Y) components of surface motion. The sensor unit also included a short-period seismometer which was sensitive to vertical motion at higher frequencies (for more instrumental details, see *Latham et al.* [1969]). Digital data were transmitted continuously from the lunar surface to receiving stations on Earth and were stored on magnetic tapes for subsequent analysis. All the seismic data were then displayed on a compressed timescale format from which lunar seismic signals were identified [*Nakamura et al.*, 1980].

[3] The lunar seismic network spans the near face of the Moon in an approximate equilateral triangle with 1100 km spacing between stations with two seismometers placed 180 km apart at one corner (see Figure 1) and covers most geological settings on the front side of the Moon. Since the first mission, more than 12,000 events have been recorded and catalogued over a period of 8 years, and it has taken several more to evaluate the data [*Nakamura et al.*, 1981]. Since the compilation of the recorded seismograms, it has been shown that the Moon is very aseismic compared to the Earth. In comparison to the Earth the energy released in seismic activity is 8 orders of magnitude less, being about 10^{10} J yr^{-1}, compared to 10^{18} J yr^{-1} by earthquakes [*Goins et al.*, 1981a], although *Nakamura* [1980] has pointed out that the actual average lunar seismic energy release could be as high as 10^{14} J yr^{-1}. Most of the moonquakes are very small

0148-0227/02/2001JE001658$09.00

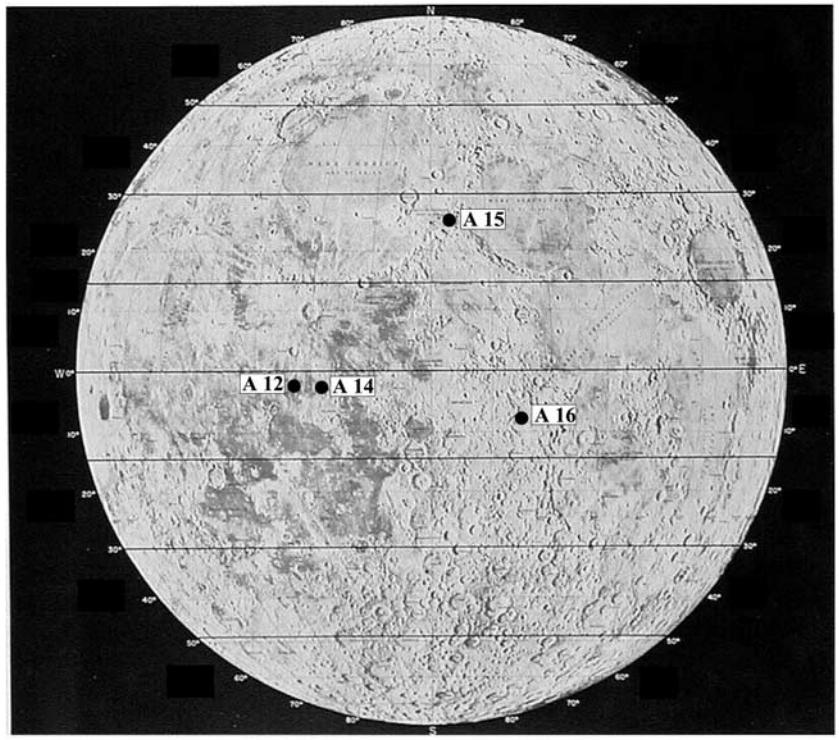


Figure 1. Location map showing the four seismic stations that operated simultaneously.

on the Richter scale with magnitudes ranging up to 4 for shallow events [*Nakamura et al.*, 1974a], whereas the deep events were generally of magnitude <2 [*Goins et al.*, 1981a]. Because of the high sensitivity of the seismometers and the low level of microseismic background noise, however, each station detected on the average between 650 and 1250 moonquakes per year.

[4] Upon examination of the first lunar seismic data returned to Earth in 1969, it became evident that their interpretation would be a somewhat more intricate process, owing to an apparent complexity inherent in lunar seismograms. It turned out that the lunar seismic signals were characterized by being very long, of high frequency, and of reverberating nature with small first arrivals and slowly building amplitudes [*Latham et al.*, 1972]. Unlike the Earth, where seismic pulses in general are of rather short duration, of the order of minutes, the most prominent feature of lunar signals is their anomalous long continuance. Strong signals, as those from the impacts of the upper stage of the Saturn rocket, last several hours. Moonquake and meteoroid impact signals typically continue for 30 min to 2 hours. The general picture that appears upon deciphering lunar seismograms from a meteoroid impact is the following [*Lammlein et al.*, 1974]: lunar signals have emerging beginnings, increase gradually to a maximum, and then slowly decay. Following the first one or two cycles of the *P* wave, ground motion is very complex, with little or no correlation between any two components. The onset of a shear wave from an impact signal is indistinct where it can be identified at all. Coherent surface wave trains displaying dispersion have not been recognized in any recordings to date, although it is believed that scattered surface waves undoubtedly contribute to the signals [*Toksöz et al.*, 1974]. It has been suggested that these signals are caused by intense scattering of the waves in the uppermost layers of the lunar crust [e.g., *Latham et al.*, 1970]. Topographic features, lunar regolith, compositional boundaries, and especially joints and cracks in the crust become very efficient scatterers in the absence of water and the absence of damping. The seismic velocity increases markedly after the first 10–15 km, and deeper material is believed to be sufficiently homogeneous to transmit seismic waves with little scattering. In this lunar environment, seismic waves generated by an impact are intensively scattered near the impact point. Scattered energy gradually diffuses into the lunar interior, in which it propagates normally and undergoes further scattering where it reenters the lunar surface layer. As a result of this, a prolonged wave train with gradual rise and decay is observed at a distant seismic station [*Lammlein et al.*, 1974].

[5] Four distinct types of events have been identified. They are deep moonquakes, shallow moonquakes, and thermal moonquakes, all of which reflect the present dynamic state of the lunar interior, and of course meteroid impacts. For a summary of the catalogued events detected on the LP seismograms during the operation of the network, see *Nakamura et al.* [1981]. The deep moonquakes, by far the most numerous of events, are usually $\sim$1 on the Richter scale [*Goins et al.*, 1981a; *Lammlein*, 1977] and were found to be located halfway toward the center of the Moon in the depth range 700–1200 km [*Nakamura et al.*, 1982]. They consist of repetitive moonquakes that emanate from specific source regions, and many nearly identical wave trains have been observed [*Lammlein et al.*, 1974]. This important observation meant that the locations are fixed, and moreover, it allowed the summing of a large number of moonquake signals, improving the signal-to-noise ratio. At most hypocenters, one moonquake occurred for a period of a few days during a

fixed time in the monthly lunar tidal cycle, giving rise to peaks at 27-day intervals of the observed lunar seismic activity. In addition, a 206-day variation and a 6-year variation in the activity, also due to tidal effects, such as the solar perturbation of the lunar orbit, have been observed [*Lammlein et al.*, 1974; *Lammlein*, 1977]. These periodicities have been taken as evidence that the deep focus moonquakes are related to the tidal forces acting on the Moon and appear to represent merely a process of storage and release of tidal energy without a significant release of tectonic energy [*Nakamura*, 1978; *Koyama and Nakamura*, 1980].

[6] The shallow moonquakes are a manifestation of another type of natural lunar seismicity. These are the most energetic seismic sources observed on the Moon, although they are less abundant than the other types of seismic events, with an average of 4 events per year [*Nakamura*, 1977]. They are also known as high-frequency-teleseismic (HFT) events owing to their unusually high frequency content and the great distances at which they are observed [*Nakamura et al.*, 1974a]. The estimation of the source depth at which HFT events occur has been inconclusive, although several lines of evidence, such as the variation of the observed amplitude of HFT signals with distance, suggested that they originated no deeper than a few hundred kilometers [*Nakamura et al.*, 1979]. It has been pointed out that there is no clear correlation between them and the tides, as is obvious for deep moonquakes [*Nakamura*, 1977]. This led to the conclusion that their origin is likely to be tectonic, given their similarity to intraplate earthquakes [*Nakamura et al.*, 1979, 1982]. Their mechanism, though, could not be explained by plate motion, because of the lack of concentration of events into narrow belts as is observed on the Earth.

[7] A large percentage of the events observed on the short period components are very small moonquakes, occurring with great regularity. It is believed that these events, also termed thermal moonquakes, are triggered by diurnal thermal variations [*Duennebier and Sutton*, 1974].

[8] With the completion of processing of all the lunar seismic data collected during the Apollo seismic network operation, a set of arrival times was obtained, constituting a primary data set from which the interior velocity structure of the Moon could be inferred. The most recent and concise summary of the seismic velocity profile is based on the complete 5-year data set acquired when the four Apollo seismometers were simultaneously operative [*Nakamura et al.*, 1982; *Nakamura*, 1983]. Nakamura and coworkers have used arrival times from various seismic events for the elucidation of the seismic velocity profile of the lunar interior. Analysis of the man-made impacts led to a model of the shallow lunar structure, a model for the upper mantle was constructed on the basis of the shallow moonquakes and the meteroid impacts, and modeling of the deep lunar interior was subject to deep moonquake data. A linearized least squares inversion technique using arrival time data from 41 deep moonquakes, 7 artificial impacts, 18 meteroid impacts, and 14 shallow moonquakes as compared to only 24 deep moonquake sources used by *Goins et al.* [1981b] was applied in steps. These models established the Moon as a highly differentiated body, with a crust and a mantle whose lower parts were thought to be partially molten [*Nakamura et al.*, 1973]. However, velocity variations and the depths of possible discontinuities in the mantle could not be addressed, although they were believed to be present [*Nakamura*, 1983]. The central part of the Moon could also not be ascertained from the seismic data owing to the distribution of seismic sources. All confirmed deep moonquakes occurred on the nearside, except for one source, A_{33}, which is located on the farside beyond the eastern limb. No big meteoroid impacts antipodal to the stations giving rise to unequivocal arrivals were detected, leaving the important question of the existence of a lunar core unanswered, although it has to be noted for the sake of completeness that a farside impact, almost diametrically opposite to station 15, gave tentative evidence for a low-velocity core with a radius around 400 km [*Nakamura et al.*, 1974b; *Sellers*, 1992].

[9] In the earlier studies mentioned above, the analysis of the nonlinear inverse problem was primarily centered on the construction of best fitting models, thereby obviating the analysis of uncertainty and nonuniqueness, which are important items when inferring scientific conclusions from inverse calculations. We [*Khan et al.*, 2000] presented a new *P* and *S* wave velocity structure for the Moon from a Monte Carlo inversion of the Apollo lunar seismic data, where we adopted a Bayesian viewpoint on inverse problems [*Tarantola and Valette*, 1982; *Mosegaard and Tarantola*, 1995] whose main feature is the use of probabilities to describe the model parameters (which are the ones we invert for) and what lends itself to their description is the a posteriori probability density in the model space which summarizes all information about the model we are studying supplied by data, a priori information, and the physical laws relating model and data. Given that for general inverse problems the shape of the posterior probability distribution is not known, it cannot simply be described by mathematical means and covariances. The Markov Chain Monte Carlo (MCMC) algorithm, on the other hand, samples a large suite of models from this probability distribution, thereby rendering us with a better representation. The main purpose of the present study is, on the one hand, to detail the method of analysis underlying that investigation and moreover to extend the Bayesian analysis carried out by *Khan et al.* [2000] (a detailed discussion of the results has already been given in that study and will not be reiterated here) and as an application of this to investigate the suite of models sampled via the MCMC algorithm as to lunar crustal thickness using Bayesian hypothesis testing [*Bernardo and Smith*, 1994]. In regard to the data set, no attempt was made to identify new events or arrivals, and the data used in this study are the same events as those considered in the study by *Nakamura* [1983], comprising first arrivals of *P* and *S* waves. However, seismograms from these events have been reviewed in order to assess the uncertainty and consistency on these arrivals. Finally, as regards the outline of the present manuscript, we have chosen first to present general ideas concerning (1) the use of MCMC algorithms to solve the general inverse problem and (2) the analysis of the posterior distribution. Upon this follows a detailed description of the application of these methods to the Apollo lunar seismic data set.

2. Theory: General Ideas

2.1. Solving the General Inverse Problem Using a MCMC Algorithm

[10] It is customary to commence an investigation such as this one by delineating our physical system by a set of

model parameters, $\mathbf{m} = (m_1, m_2, \ldots, m_s)$, which completely define the system. These parameters are not directly measurable. What we usually are in possession of, though, are certain observable data, $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, obtained through physical measurements. Now, what we could do next is to try to prognosticate the observable parameters by using any theory applicable to our particular predicament. This results in another set of parameters, termed calculated data $\mathbf{d}_{\text{cal}}$, which are dependent on the model parameters. This dependency can be depicted by a relation of the form

$$\mathbf{d} = g(\mathbf{m}), \tag{1}$$

g being a functional relation governing the physical laws that correlate model and data. What is meant by solving the forward problem, then, is the prediction of observable data given a set of model parameters, and conversely, solving the inverse problem is understood to be the inference of values of the model parameters, given observable data. Central to our method is the notion of a state of information over the parameter set. In concordance with the most general description of states of information over a given parameter set, as presented by *Tarantola and Valette* [1982] and *Tarantola* [1987], we shall be employing probability densities over the corresponding parameter space, describing the various states of information inherent to our system. This knowledge embodies the results of measurements of the observable parameters and the a priori information on model parameters as well as the information on the physical correlations between observable and model parameters. Solving the inverse problem, then, shall be formulated as a problem of combining all this information into an alterior state, termed the posterior probability density. The extensive use of probability densities for delineating any information has the advantage of presenting the solution to the inverse problem in the most generic way, thereby implicitly incorporating any nonlinearities [*Tarantola and Valette*, 1982].

[11] It is clear, then, that the most general method for solving nonlinear inverse problems needs an extensive exploration of the model space, since the posterior probability density in the model space contains all the information about the system being studied. Therefore, given probabilistic prior information on $\mathbf{m}$ and a statistical description of the observational uncertainties of $\mathbf{d}$, the main idea is to design a random walk in the model space which samples the posterior probability distribution, that is, samples models which are consistent with data as well as prior information. To this end, we shall use a Markov Chain Monte Carlo algorithm of the following form (the basic premises underlying the use of MC algorithms to solve general inverse problems are reviewed by *Mosegaard* [1998]):

1. Propose a new model, $\mathbf{m}_{\text{pert}}$, by taking a step of a random walk to some current model $\mathbf{m}_{\text{cur}}$, with a probability proportional to $\rho(\mathbf{m})$, the prior probability density on the model parameters.

2. Calculate the likelihood function for the new model using $L(\mathbf{m}) = k \cdot \exp(-S(\mathbf{m}))$, where k is a normalization constant, $S(\mathbf{m})$ is the misfit function, and $L(\mathbf{m})$ is a measure of the degree of data fit.

3. Accept the new model with a probability $P_{\text{acc}} = \min(1, L(\mathbf{m}_{\text{pert}})/L(\mathbf{m}_{\text{cur}}))$.

4. If $\mathbf{m}_{\text{pert}}$ is accepted, then $\mathbf{m}_{\text{cur}} = \mathbf{m}_{\text{pert}}$. If not, then reapply $\mathbf{m}_{\text{cur}}$ and repeat the above steps.

[12] This algorithm will sample the posterior probability density

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

η being a normalization constant, asymptotically. "Asymptotically," in this case, implies that the statistical correlation between samples taken at times separated by n iterations will converge toward zero as n goes to infinity. The advantage of using the above scheme to sample the posterior probability density $\sigma(\mathbf{m})$ is that sampling is conferred to those parts of the model space where model parameters consistent with data and prior information exist.

2.2. Analysis of the Posterior Distribution

[13] Having designated the posterior probability distribution $\sigma(\mathbf{m})$ as the solution to our inverse problem, our main concern is now the analysis of this distribution. However, owing to its complex shape (it might be multimodal, contain infinite variances, etc.), it is not possible to directly access information from it. Instead, the information of most use is obtained by investigating the probability $\mathcal{P}$ that a certain feature resides at a given depth or depth range (corresponding to the model parameters being contained within a given subset of the model space), which can be calculated from

$$\mathcal{P}(\mathbf{m} \in \Lambda) = \frac{\int_\Lambda \sigma(\mathbf{m}) d\mathbf{m}}{\int_\Omega \sigma(\mathbf{m}) d\mathbf{m}}, \tag{3}$$

where Λ denotes a subset of the model space Ω and the denominator is clearly identified as a normalization factor. This can easily be extended to the calculation of means and covariances.

[14] We could also adopt the point of view taken by *Mosegaard* [1998], who states that the main aim of resolution analysis is to aid in having to choose between differing interpretations of a given data set. In line herewith our principal purpose in analyzing the sampled models will basically be to conjure up queries addressing correlations between several model parameters. The one question that we intend to investigate using the sampled models concerns the lunar crust and could be formulated as follows:

- How likely is it, on the basis of the seismic data and their uncertainties as well as prior information, that the Moon has a discontinuity (suitably defined) in a certain depth range, marking for example the crust-mantle interface?

However, answering questions like these involves having to evaluate resolution measures of the form [*Mosegaard*, 1998]

$$\mathcal{R}(\Lambda, f) = \int_\Lambda f(\mathbf{m})\sigma(\mathbf{m}) d\mathbf{m}, \tag{4}$$

where $f(\mathbf{m})$ is a given function of the model parameters $\mathbf{m}$ and Λ is, as in (3), an event or subset of the model space Ω containing the models of current interest. The similarity between (3) and (4) is obvious.

[15] It is clear from the above discussion that in the case of the general inverse problem, (3) and (4) are inaccessible to analytical evaluation, since we do not have an analytical expression for $\sigma(\mathbf{m})$. However, the problem can be solved using the MCMC algorithm, as mentioned before, which samples a large collection of models, $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n$, from $\sigma(\mathbf{m})$, whereby the resolution measure, given by (4), in turn can be approximated by the following simple average [*Mosegaard*, 1998]:

$$\mathcal{R}(\Lambda, f) \approx \frac{1}{N} \sum_{\{n|m_n \in \Lambda\}} f(\mathbf{m}_n), \tag{5}$$

where N normalizes the resolution measure. For the number of samples $N \rightarrow \infty$ the equality of (4) and (5) corresponds to the important ergodicity property used in Monte Carlo integration.

[16] However, we shall avail ourselves of a slightly different approach which is known as hypothesis testing. While the two analyses basically amount to the same, since a hypothesis corresponds to a question, the advantage in hypothesis testing lies in the fact that we are able to compare any two hypotheses against each other.

[17] The classical problem of hypothesis testing concerns itself with making a choice among different hypotheses, $\mathcal{H}_1$, $\mathcal{H}_2, \ldots, \mathcal{H}_n$, on the basis of some observed data $\mathbf{d}$. Or formulated alternatively, we might be in the position of having to decide how the prior probability $\mathcal{P}(\mathcal{H})$, concerning any hypothesis $\mathcal{H}$, has been amended by taking into account observed data $\mathbf{d}$ (henceforth abbreviated data); that is, we are interested in evaluating $\mathcal{P}(\mathcal{H}|\mathbf{d})$, which is usually termed the posterior probability distribution. The mathematical connection linking the prior and the posterior is given by the Bayes theorem

$$\mathcal{P}(\mathcal{H}|\mathbf{d}) = \eta \mathcal{P}(\mathbf{d}|\mathcal{H}) \mathcal{P}(\mathcal{H}), \tag{6}$$

where η is a normalization constant. Extending the analysis in order to compare any two hypotheses, we arrive at the Bayes factor, whose definition has been ascribed to Turing [e.g., *Good*, 1988].

2.2.1. Definition (Bayes factor)

[18] Given two hypotheses $\mathcal{H}_i$, $\mathcal{H}_j$ corresponding to different areas of the model space Ω, for data $\mathbf{d}$, the Bayes factor $\mathcal{B}_{ij}$ in favor of $\mathcal{H}_i$ (and against $\mathcal{H}_j$) is given by the posterior to prior odds ratio.

$$\mathcal{B}_{ij}(\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}|\mathcal{H}_i)}{\mathcal{P}(\mathbf{d}|\mathcal{H}_j)} = \frac{\mathcal{P}(\mathcal{H}_i|\mathbf{d})/\mathcal{P}(\mathcal{H}_j|\mathbf{d})}{P(\mathcal{H}_i)/P(\mathcal{H}_j)}, \tag{7}$$

$\mathcal{P}(\mathbf{d}|\mathcal{H}_i)$ being the probability distribution, usually termed the likelihood (see below). Thus the Bayes factor provides a measure of whether the data $\mathbf{d}$ have increased or decreased the odds on $\mathcal{H}_i$ relative to $\mathcal{H}_j$. Accordingly, if $\mathcal{B}_{ij}(\mathbf{d}) > 1$, $\mathcal{H}_i$ is now more relatively plausible than $\mathcal{H}_j$ in the light of $\mathbf{d}$; on the other hand, if $\mathcal{B}_{ij}(\mathbf{d}) < 1$, it signifies that $\mathcal{H}_j$ has now increased in relative plausibility [*Bernardo and Smith*, 1994]. It is clear that apart from how one defines the crust-mantle transition or any other hypothesis for that matter, this form of analysis is straightforward as seen from the Bayesian viewpoint. This can be appreciated by rewriting the above equation as

$$\frac{\mathcal{P}(\mathcal{H}_i|\mathbf{d})}{\mathcal{P}(\mathcal{H}_j|\mathbf{d})} = \frac{\mathcal{P}(\mathbf{d}|\mathcal{H}_i)}{\mathcal{P}(\mathbf{d}|\mathcal{H}_j)} \times \frac{\mathcal{P}(\mathcal{H}_i)}{\mathcal{P}(\mathcal{H}_j)}, \tag{8}$$

that is, the ratio of posterior odds equals the integrated likelihood ratio times the ratio of prior odds. This equation shows explicitly how the ratio of integrated likelihoods plays the key part in providing the mechanism by which data transform relative prior beliefs into relative posterior beliefs. If we compare this to the form of the posterior probability distribution, equation (2), which we used in sampling solutions to our inverse problem, the connection is complete, in that (2) states in an analogous manner the role of the likelihood function $L(\mathbf{m})$, which contains information on data $\mathbf{d}$ and on the physical theories linking data and model parameters, in changing the prior probability distribution on model parameters $\rho(\mathbf{m})$ into the posterior $\sigma(\mathbf{m})$ [*Mosegaard and Tarantola*, 1995; *Mosegaard*, 1998].

3. Analysis: Application of the MCMC Algorithm to the Apollo Lunar Seismic Data

3.1. Forward Problem

[19] As mentioned, the forward problem consists of calculating data, that is, travel times, given a set of model parameters, that is, a model of the subsurface velocity structure, among other things. In our model of the Moon the assumption of radial symmetry is made. The Moon is partitioned into 56 shells of variable size, with each layer being characterized by its physical extent and material parameters in the form of the velocity. To each shell is assigned a piecewise linearly varying P and S wave velocity of the form $v(r) = v_o + k \cdot r$, which is continuous at layer boundaries. In order to accommodate ray theory, certain limits are placed on the velocity gradients, resulting in a smoothness constraint limiting vertical resolution to roughly 5 km.

[20] Since the Moon is neither spherically symmetric nor geologically homogeneous, an additional asset was introduced to facilitate a more realistic modeling of the lunar interior. Our model of the Moon was stripped of a surficial layer, 1 km thick, which is known to be of very low velocity [*Kovach and Watkins*, 1973]. Because of the extreme velocity gradients encountered by the rays when impinging this layer, they will be almost vertically incident upon reaching the surface. The travel time for a ray in this layer will therefore to a good approximation be given by $T = 1\ \text{km}/v_{\text{surface}}$. So, instead of ray tracing in a sphere with a radius of 1738 km, we traced in one with a radius of 1737 km and added for every station and surface source a time correction to the travel time. Starting off with a surface velocity of, say, 0.5 km s^{-1} [*Nakamura et al.*, 1982] means that we have to add a total time correction of 4 s to the travel time of a ray emanating from a surface source and 2 s for one originating within the Moon. Leaving the time corrections variable, this method has the added advantage of taking localized properties of the surficial material beneath each station into account. For consistency, all time corrections have been correlated in such a way that if a particular station has been assigned a given initial correction, this

same correction will be added to the travel times for the rays emitted from all sources and traveling to this particular receiver. In the same way, the corrections are correlated for the impacts; that is, all rays emanating from a particular impact have the same correction added to their total travel time.

3.2. Inverse Problem

[21] Three parameters as discussed in the previous section are used to describe our physical model of the Moon which included the position of the layer boundaries, r_i, their velocities, v_i, and the time corrections, t_k, in the surficial layer. In choosing how to parameterize our physical system, however, it has to be noted that the choice of which parameterization to employ is an ambiguous matter, in the sense that it is not unique [*Tarantola and Valette*, 1982]. So, instead of examining a number of different parameterizations, such as the velocity or the slowness, we shall avail ourselves of another approach, this being the invariance argument. As the name suggests, invariance implies the procurement of commensurate distributions upon transformation from any one invariant measure. The "log-velocity" possesses exactly this property [*Tarantola*, 1987]; that is, given a uniform distribution in $\log(v/v_o)$, for example, we will achieve a similar distribution whether it be the velocity or the slowness we are transforming to. We shall therefore adopt $\log(v/v_o)$ as the parameterization throughout this study. (In the following we shall continue to use v_i to describe the velocity parameters, but it is tacitly assumed that we actually mean the log-velocity. Throughout this study, $v_o = 1.0$ km s^{-1}.)

[22] To fully characterize our physical system, we also need to incorporate parameters which describe the meteoroid impact as well as the shallow and deep moonquake locations, thus resulting in the addition of three other parameters in the delineation of our model. These parameters are the depth coordinate of every moonquake and the selenographical position of the epicenter, that is, longitude and latitude, which we shall label s_j and the latter two θ_j and ϕ_j, respectively. Our model is thus ultimately given by $\mathbf{m} = \{r_i, v_i, t_k, s_j, \theta_j, \phi_j\}$.

[23] In commencing the MCMC algorithm we started out by assuming some initial model $\mathbf{m}$, henceforth $\mathbf{m}_{\text{cur}}$, comprising a set of values. Now, by perturbing one of these parameters, that is, by either changing the position of a boundary layer, changing the value of the velocity at a given layer boundary, assigning a new time correction for either a source or a station, or amending the hypocentral or epicentral coordinates depending on the particular event, we obtain a new model. The random walker is then set out to sample the parameter space according to the prior information and using a set of random rules whose efficiency has been optimized through several numerical experiments. Let us outline these random rules and the prior information. For every iteration it is first decided which parameter is to be perturbed next. The decision is made whether to change a parameter pertaining to the lunar interior or one regarding the location of a seismic event, which are equally probable. Performing a velocity perturbation has the same probability (0.5) as performing a boundary layer perturbation. When it comes to perturbing the time corrections, these are set to be amended every fifth iteration. When perturbing either the velocity or the position of a boundary, the layer is selected uniformly at random and so is the value of the parameter to be changed. It should be kept in mind that perturbations concerning the placement of boundaries have in a certain sense been restricted in order to accommodate ray theory. Every layer to be perturbed can be assigned any value, except in a range of 5 km within the layer immediately above or below; that is, any given layer i can assume a value within $r_{i-1} + 5 \text{ km} \leq r_i \leq r_{i+1} - 5$ km. Concerning the time amendments, whether it is a source location or the location of a station, which is to be perturbed will be chosen uniformly at random, as are their values. Let us note that we have made the assumption of a uniform distribution in the "log-time" domain as in the case of the velocities, the reason being that the time is directly related to the velocities in the top layer. Concerning the hypocenter coordinates, the source depths are uniformly distributed from the surface to the center of the Moon, and the epicentral coordinates are likewise assumed to be uniformly distributed, in this case across the lunar surface. As regards the sampling of S wave velocities, it has to be remarked that in order to make the inversion tractable the problem was actually divided into two stages. The first entailed the inversion of the P wave arrival times, while the second considered the inversion of the S wave arrivals. Now, it is clear that the two parameters, v_p and v_s, are not independent if described in terms of the elastic moduli, ρ, κ, and μ, these being the density, bulk, and shear modulus, respectively. It could therefore be argued that this division of the problem might result in physically unrealizable models (for further discussion, see section 8). However, being aware of the connection between the two parameters, we did impose certain constraints on the sampling of S wave velocities, corresponding to the addition of prior information, by introducing a Gaussian distribution centered on the average $v_p/\sqrt{3}$ as obtained from the first inversion and with a standard deviation given by $\sigma_{v_p}/\sqrt{3}$. This prior information thus signifies that S wave velocities lying far from this distribution are less likely sampled.

[24] This body of information, then, serves as prior knowledge, in the sense that the random walker will be sampling the model space with a probability density describing exactly this information.

[25] At this point the reader might wonder about what happened to the estimation of parameters pertaining to the origin time of events, since these are also unknowns when it comes to event localization. These are also determined in this study; however, instead of including these in the model parameter vector $\mathbf{m}$ above as another set of parameters to be determined, we adopted a slightly different approach.

[26] Generally, we have the following relation concerning the arrival times:

$$\mathbf{d}_{\text{a}} = \mathbf{m}_o + g(\mathbf{m}),$$

where $\mathbf{d}_{\text{a}}$ is the vector of arrival times, $\mathbf{m}_{\text{o}}$ is a model parameter vector of origin times corresponding to individual arrivals, and $g(\mathbf{m})$ is our set of calculated travel times. Now, for the present purposes we made the assumption that the a

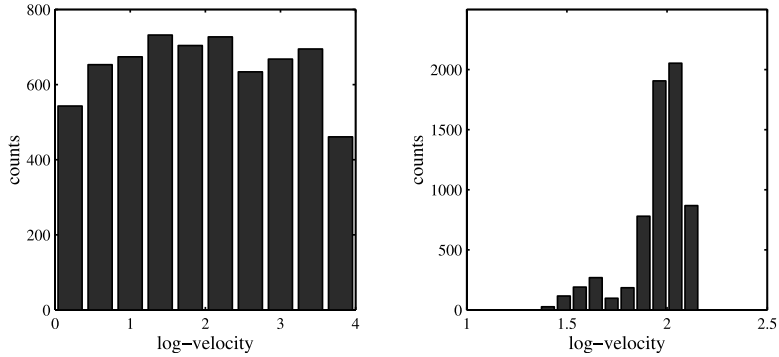


Figure 2. Prior and posterior marginal $\log(v/v_o)$ parameter distribution. Note how the prior is changed from sampling a uniform distribution by taking data into account so as to sample the posterior.

priori information concerning the origin time model parameters was normally distributed with mean values as determined by the Nakamura model [*Nakamura et al.*, 1976; *Nakamura*, 1983] and an a priori uncertainty, typically of a few seconds. This sets the stage for the following formal definition of new data:

$$\mathbf{d} = \mathbf{d}_a - \mathbf{m}_o.$$

To these were assigned a new standard deviation as the sum of the standard deviation of arrival times and the a priori dispersion on origin time model parameters, thus defining a composite uncertainty (to be dealt with later on).

[27] The reason that we ended up choosing what might be labeled rather loose prior knowledge on model parameters, by assuming that they were all uniformly distributed within a large range, was our main aim of wanting to investigate the full model variability inherent in the lunar seismic data.

[28] Having designated prior information, let us continue the MCMC algorithm and assume that the random walker wishes to sample the velocity at some depth. He currently resides at v_i and from there takes a step in some direction to a point v_i^{new}. Now, we do not restrict the random walker to wander within some specified range, but shall rather let him drift around the model space from zero to infinity, although every jump is in a sense constrained so as to assure that the random walker can take only small steps. Sampling the a priori distribution by this method basically corresponds to the process of diffusion, since the random walker can essentially, if given steps enough, explore the whole model space. Mathematically, we would write this condition as $v_i^{\text{new}} = v_i + \xi \cdot (2 \cdot \alpha - 1)$, where α denotes a uniformly distributed number in the interval [0,1] and ξ is a constant, typically of the order of 0.2 km s^{-1} in this study.

[29] After having obtained a new model, $\mathbf{m}_{\text{pert}}$, our next step is to gauge the posterior distribution. This we do by first calculating a new set of arrival times using $g(\mathbf{m}_{\text{pert}})$ and then comparing these to the observed data in order to obtain the misfit function. This misfit is of importance, since we shall by the Metropolis rule use it to either accept or reject the new model, with probability $P_{\text{acc}} = min\{1, L(\mathbf{m}_{\text{pert}})/L(\mathbf{m}_{\text{cur}})\}$. In words, we could state this by saying that the random walker essentially updates his knowledge of the current state of affairs, as prescribed by Bayes' theorem, by merging together prior information with data and theory. The revision of the prior into the posterior is illustrated in Figure 2, which shows how a prior distribution for a given velocity parameter is being changed as data are taken into account. Moreover, we are led to the conclusion that the level of confinement is a direct measure of the knowledge we possess about the system; that is, the more the random walker is constrained when sampling the posterior distribution in the model space, the greater is our degree of knowledge.

[30] Now, the strategy set forth here of perturbing only one parameter at a time is sure to preserve most of the characteristics of the current model which may have resulted in a good data fit. While being efficient, in the sense that we are interested in sampling models with a good data fit, this strategy will result in a sequence of samples, that is, models, that tend to be correlated. This is somewhat unfortunate, since analyses of error and resolution require a collection of statistically independent models from the posterior distribution. The way to proceed is to choose fewer samples from the set of accepted models in such a way that they constitute a set of independent models. This can be done by introducing an elapse time (number of iterations) between retention of samples which was found by analyzing the fluctuations of the likelihood function as the algorithm proceeded to be 100. Inspection of the autocorrelation function for these fluctuations showed that accepted models separated by 100 iterations were unlikely to be correlated. On the other hand, the limited number of iterations performed may not be sufficient to allow the algorithm to visit enough extrema in the model space. To circumvent this problem, we chose to commence at different places in the model space for every 10,000 iterations performed by the algorithm. This was done by restarting the MCMC algorithm with different initial values at those intervals. This procedure should guarantee that the samples are less correlated as well as leading to a better coverage of the probability distribution. (It should be noted that although this method is more efficient in detecting most of the extrema, its downside is the fact that it is biased toward an approximately equal coverage of them all. The global extremum, corresponding to the most likely solution, might end up being sampled on a par with secondary extrema, thereby conferring them with equal weight.) To ensure that the posterior probability density was adequately sampled in our analysis, we monitored the time series of all output parameters from the algorithm to verify that these were indeed stationary over the many iterations performed.

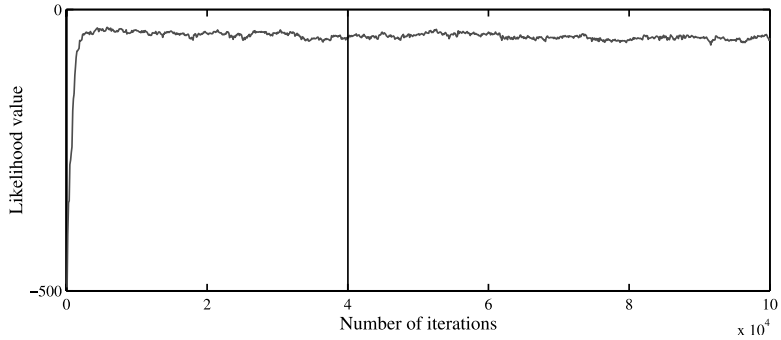


Figure 3. Convergence of the MCMC algorithm. The value of the likelihood function is a rough measure of how well calculated data fit the observables. As convergence has been reached, denoted by the vertical line, sampling of the posterior distribution is initiated.

In addition to showing the convergence (The issue of convergence presents a so far unresolved difficulty in practice, since it is not easy to decide how many samples are actually enough to constitute a good representation of the posterior probability distribution.) of the MCMC algorithm, Figure 3 also depicts an example of how it was decided when to initiate sampling from the distribution, by observing the values of the likelihood function from the time of commencement of the algorithm and as it proceeds. It is customary to start keeping samples only after the likelihood function has stabilized around some value. (It is clear that if our starting model is far removed from any extrema in the model space the longer time it may take, in terms of the sequence of updatings, before the algorithm can actually start to sample the distribution near the extrema where most of the contribution to the posterior distribution ought to come from, rendering the issue of when to start sampling rather important.)

[31] Returning to the sampling algorithm, the next step would be to perturb another model parameter, as explained above, and accept this model with the probability $P_{\text{acc}} = min\{1, L(\mathbf{m}_{\text{pert}})/L(\mathbf{m}_{\text{cur}})\}$. Continuing along this line, we would assemble a suite of models which are distributed in accordance with the posterior distribution. The models thus gathered constitute our main output.

[32] Let us now turn to a point which is of importance when dealing with data suspected of containing outliers. The introduction of an outlier(s) is most probably the result of an erroneously read arrival time, and given the ingrained complexity of the lunar seismograms, as commented earlier, it only seems too natural to suspect inconsistencies in the readings of at least a couple of seismic phases. Now, the central question is, how do we detect these outliers, if there are any present, and furthermore, does their presence in a data set ensue in any form of distortion of the posterior probability density? According to *Tarantola* [1987], an outlier will have a proclivity to "translate" the posterior probability density if we are employing the l_2 norm, that is, if we are assuming independent, identically distributed Gaussian uncertainties, which is what we have done until now, the amount of "translation" necessarily depending on how much the outlier(s) is displaced from the rest of the data points. If a given data set is suspected of harboring an outlier(s), *Tarantola* [1987] advocates the use of the l_1 norm instead, since a contortion of the posterior probability distribution is less prevalent in this case. The l_1 norm assumes that the errors can be modeled using an exponential probability density; that is, instead of having the misfit function given by $S(\mathbf{m}) = \frac{1}{2}\Sigma_{i=1}^{N}\ g^i(\mathbf{m}) - d^i_{obs})^2/\sigma_i^2$, we shall assume the following form for the misfit function: $S(\mathbf{m}) = \Sigma_{i=1}^{N}\ |g^i(\mathbf{m}) - d^i_{\text{obs}}|/\sigma_i$, where σ_i denotes the uncertainty on the ith arrival time reading. In order to investigate as to whether there are any outliers present in our data set and additionally to check the relative robustness of the Gaussian and exponential hypotheses, we undertook an inversion of a more uncertain data set, namely, the meteoroid impacts in combination with the shallow moonquakes. We included in the inversion the artificial impacts, this being mainly for technical reasons. The reason that we have been relying on the l_2 norm hitherto is the fact that the artificial impacts constitute a more reliable data set, since origin times and impact locations are known parameters (except for the 16 SIV-B impact, for which origin time and location are not known parameters because of the loss of tracking of this spacecraft prior to impact [*Latham et al.*, 1972]). Figure 3 also shows how it can be used as an indicator of the presence of outliers in the data set. An estimate of the likelihood value when convergence has been reached using the l_1 norm is roughly given by "minus the number of values in the data vector." If the MCMC algorithm converges on a value differing from this, it is attributed to the presence of conflicting data points in the observed data set. In order to identify outliers, we generated a set of samples of the order of 10^4. The data, that is, the arrival times as calculated by each iterated model, were checked minutely, and 20 of the lunar seismic arrivals (not to be confused with events) were identified as outliers. The rejected outliers deviated significantly from the expected values by more than 4 and up to 15 standard deviations of data noise. Figures 4a and 4b depict actual histograms of calculated arrival time differences for a meteoroid impact and a shallow moonquake as registered at stations 15 and 16, respectively, which deviated by more than 4 standard deviations. The outliers were subsequently removed and the process was initiated all over again until the misfit attained the appropriate value. It should be noted that this method has the added benefit of leading to a reevaluation of the uncertainty on the individual arrival times, which amount to 1 s for the artificial impacts, 4–26 s for the

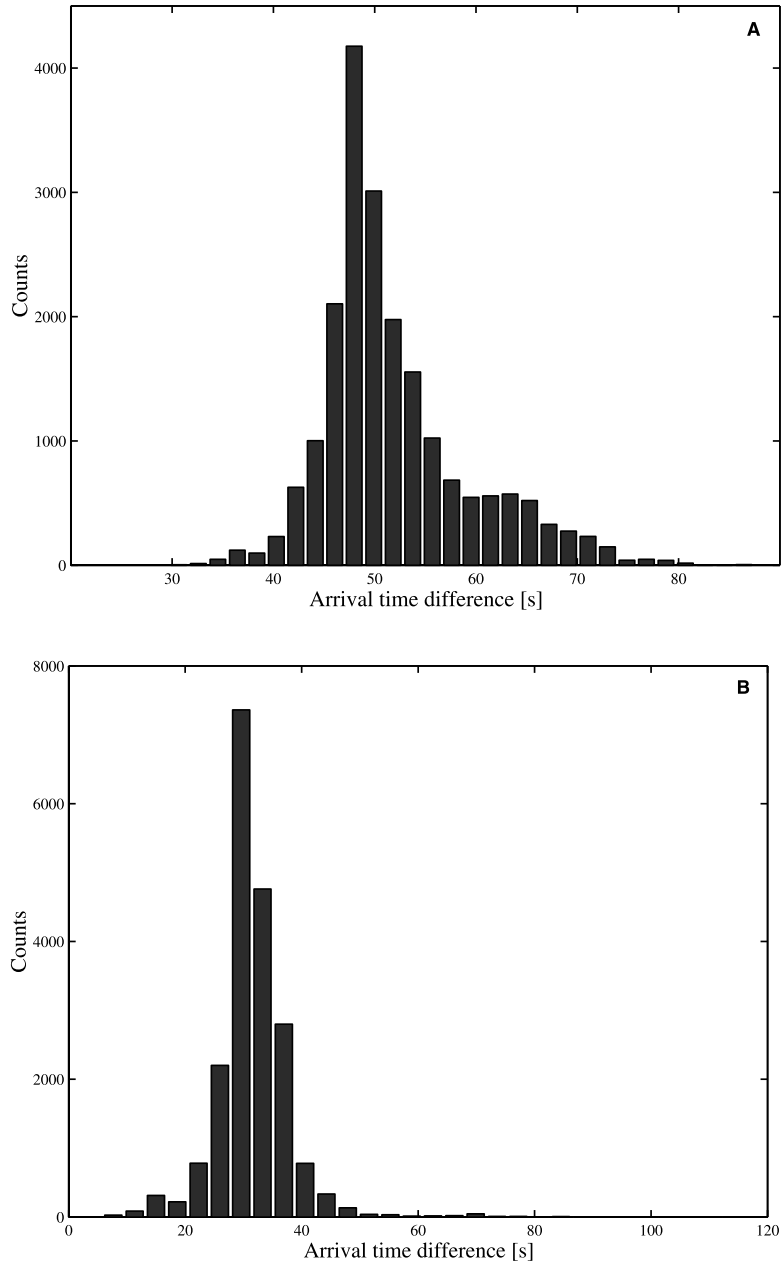


Figure 4. Identification of outliers: (a) the arrival time difference between calculated and observed *P* wave arrival for meteoroid impact on day 349 1974 as recorded by station 15 and (b) the case of a shallow moonquake recorded at station 16 on day 4 1976. The uncertainties on these two readings were assessed to be 7 and 5 s, respectively.

shallow moonquakes and the meteoroid impacts, and 4–7 s for the deep moonquakes. In order to safeguard the results from further inconsistencies in the data set, we simply chose to use the l_1 norm throughout this study. This method of elucidating outliers in a data set corresponds to the statistical technique known from robust M type estimation [*Barnett and Lewis*, 1984; *Hampel et al.*, 1986].

[33] To recapitulate, we started off by randomly generating models which were distributed according to the posterior. In order to enhance the coverage of the model space by the random walker, we perturbed him randomly after a fixed number of iterations. This corresponds to the case in which the random walker, after having sampled some local or global extremum for a preset amount of steps, suddenly finds himself displaced to another region of the model space around which he would steadily peregrinate in the usual manner. This subsequent wandering about leads to the equilibration at yet another extremum, thereby sampling models belonging to a specific class. From Figure 5 we clearly see that the models generated by sampling this particular extremum in the model space are different from the ones produced at other points of the model space, in that they accentuate contrastive properties. On display in Figure 5 are exactly 10 distinct models, and these are each

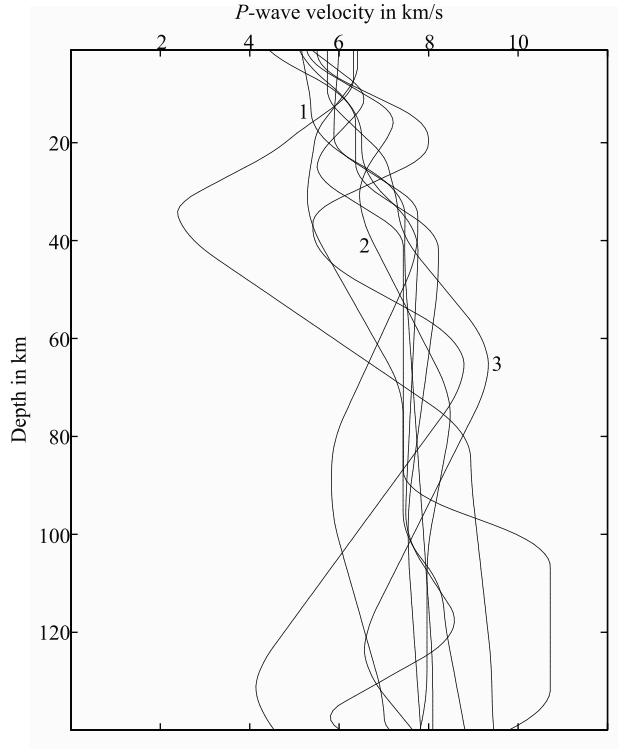


Figure 5. A smoothed version of 10 samples from the posterior distribution, each one obtained by sampling a different extremum in the model space. The models shown here emphasize only the crustal and upper mantle part. The three models numbered 1–3 have been taken as examples of models which individually highlight seemingly different structures but nonetheless are models that all produce a fairly large likelihood value, that is, produce a good fit to the observed data (see Table 1).

representative of a class of models obtained by gathering samples from 10 distinct extrema. Whether 10 reshufflings of the random walker actually result in 10 distinctive families of models is something we clearly cannot be aware of from the outset, and indeed, had we ended up with only three different classes of models, this would have been a revelation in itself, disclosing information about the degree of nonlinearity of the problem. The models juxtaposed in Figure 5 are also displayed to highlight the ingrained ambiguity contained in the data. Although the classes of models presented here are characterized by a fairly large likelihood value, that is, they produce arrival times which are in close agreement with the observed ones, being well within the specified error bounds, they clearly show divergencies among themselves. Figure 5 simply emphasizes that models with differing attributes are accountable for the same data, which is clearly evidenced by Table 1. All three models are seen to be capable of fitting the observational data within error bounds (±1 s). This excellent agreement between observed and calculated data reinforces the method employed here of elucidating the lunar interior from a great number of models rather than just adopting the maximum likelihood model and inferring the velocity structure straight from this.

4. Analysis: Estimating the Crustal Thickness Using Bayesian Hypothesis Testing

[34] Having presented the general theory, we are now in the position to employ the Bayes factor in order to specifically analyze the outcome as to whatever feature we might be interested in. From (5) it is apparent that the number of times a model has been sampled is proportional to its posterior probability. The Bayes factor is thus easily evaluated as the ratio of the number of samples from the posterior distribution to the ratio of the number of samples from the prior distribution for any two hypotheses.

[35] As noted in section 1, our primary concern is the depth to the lunar Moho, and we shall here want to distinguish between the following two hypotheses given the seismic data and prior information:

Hypothesis 1 ($\mathcal{H}_1$): The lunar crustal thickness lies in the range 35–45 km.

Hypothesis 2 ($\mathcal{H}_2$): The lunar crustal thickness lies in the range 50–70 km.

[36] Of a more technical character is the definition of the crust-mantle interface. A model subject to the following two conditions is defined as accommodating a discontinuity:

1. Velocity gradients should reach 0.1 km s^{-1} km^{-1}.
2. The velocity beneath the designated thickness of the crust should attain a value of at least 7.6 km s^{-1}.

[37] This interpretation of the lunar Moho has been constructed so as to agree with the characteristics of all three models discussed in section 1, $\mathcal{H}_1$ and $\mathcal{H}_2$ corresponding to the crustal depths as proposed by (1) *Khan et al.* [2000] and (2) *Toksöz et al.* [1974] and *Nakamura* [1983], respectively. This implies searching through the entire set of sampled posterior and prior models and by (5) counting the number of prior and posterior models, respectively, satisfying the above two definitions of a crust-mantle boundary and simply taking their ratio. As a further technical note, it should be mentioned that the MCMC algorithm is designed to sample models with the highest possibly achievable resolution. However, given the well-known trade-off that exists between resolution and uncertainty, high resolution implies a rather large dispersion in model parameters and vice versa. Since we are presently concerned about searching for velocity discontinuities, small uncertainties on model parameters are paramount. This is easily achieved by smoothing the individual models, although of course at the expense of resolution. We have used the median in filtering the models over a range of eight values around a given velocity point,

Table 1. A Comparison of Observations With Calculations[a]

Impactor	Station	Observed Travel Times, s	Calculated Travel Times, s		
			Model 1	Model 2	Model 3
13 SIV-B	12	27.3	27.2	27.2	27.4
14 SIV-B	12	34.3	34.5	34.2	34.6
14 LM	12	23.7	23.0	23.2	24.0
14 LM	14	16.5	16.1	16.3	16.0
15 SIV-B	12	53.7	53.9	54.1	53.3
15 SIV-B	14	35.3	36.3	34.9	36.2
15 LM	15	20.7	21.3	19.8	20.1
17 SIV-B	12	54.7	55.4	55.3	55.5
17 SIV-B	14	30.7	30.0	31.2	30.5
17 SIV-B	15	149.7	149.1	150.0	150.3
17 SIV-B	16	121.8	122.4	121.1	122.0

[a] The three models refer to Figure 5.

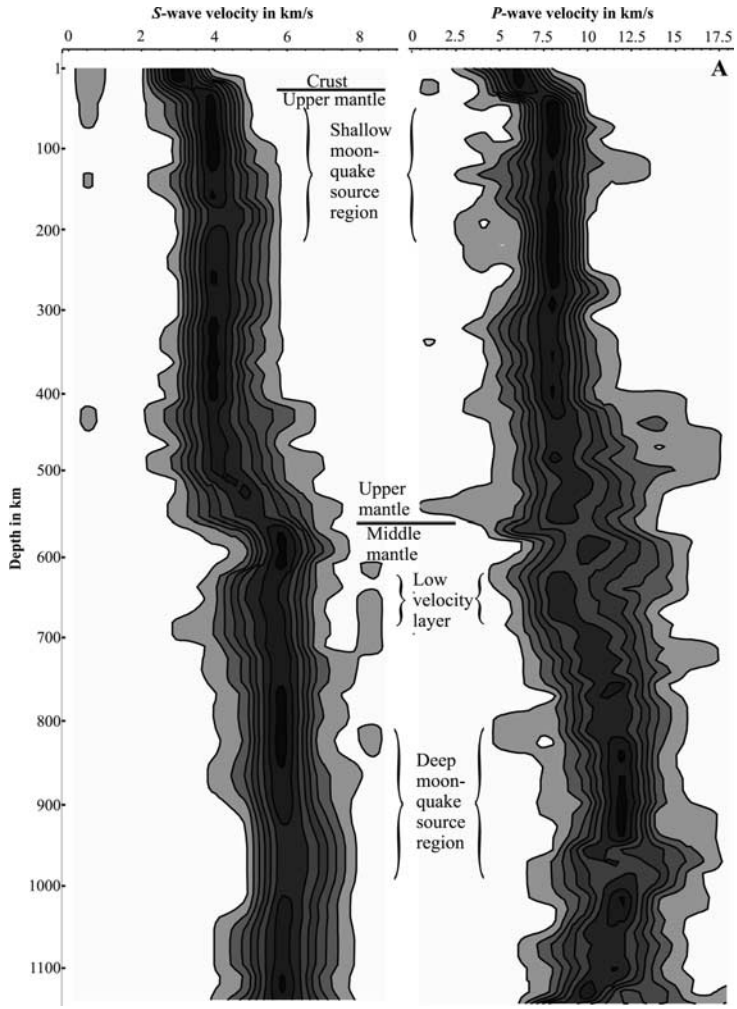


Figure 6. The marginal posterior velocity distributions depicting the velocity structure of the Moon. A total of 50,000 models have been used in constructing the two results. For each kilometer a histogram reflecting the marginal probability distribution of sampled velocities has been set up. By lining up these marginals, the velocity as a function of depth is envisioned as contours directly relating their probability of occurrence. The contour lines define nine equal-sized probability density intervals for the distributions. The uncertainties on the results are in part due to the large uncertainty in arrival time readings. It should be kept in mind that the velocity models are depicted using marginal probability distributions, and as such a model incorporating velocities of maximum probability does not necessarily correspond to the most likely model. See color version of this figure at back of this issue.

since the median preserves any discontinuities rather than smoothing them.

[38] The posterior and prior odds ratios were found to be $\mathcal{P}(\mathcal{H}_1|\mathbf{d})/\mathcal{P}(\mathcal{H}_2|\mathbf{d}) = 2.5$ and $\mathcal{P}(\mathcal{H}_1)/\mathcal{P}(\mathcal{H}_2) = 0.6$, respectively, resulting in a Bayes factor of

$$\mathcal{B}_{12} = \frac{\mathcal{P}(\mathcal{H}_1|\mathbf{d})/\mathcal{P}(\mathcal{H}_2|\mathbf{d})}{\mathcal{P}(\mathcal{H}_1)/\mathcal{P}(\mathcal{H}_2)} = \frac{2.5}{0.6} = 4.2,$$

whereby $\mathcal{H}_1$ is favored to $\mathcal{H}_2$ given data and prior information.

5. Interpretation of the Results

[39] As emphasized previously, the solution to the Bayesian formulation of general inverse problems results in a posterior probability distribution and not single estimates, although these can be obtained from the posterior distribution by numerical integration. What this signifies is that we are solely working with solutions in terms of probability distributions, and moreover, given the multidimensional nature of the posterior distribution, it is not accessible for direct display and only marginals can be exhibited. Therefore care must be exercised in interpreting the outcome, since false conclusions and overinterpretations are easily reached.

[40] Now, from Figure 6, depicting the final velocity model, it would seem very natural to assume the model belonging to the region of highest probability as the prevalent one and therefore pick it as the main lunar velocity structure. However, this particular model will usually not correspond to the most likely model (which is typical of strongly nonlinear problems and is clarified further below), since it has to be remembered that the

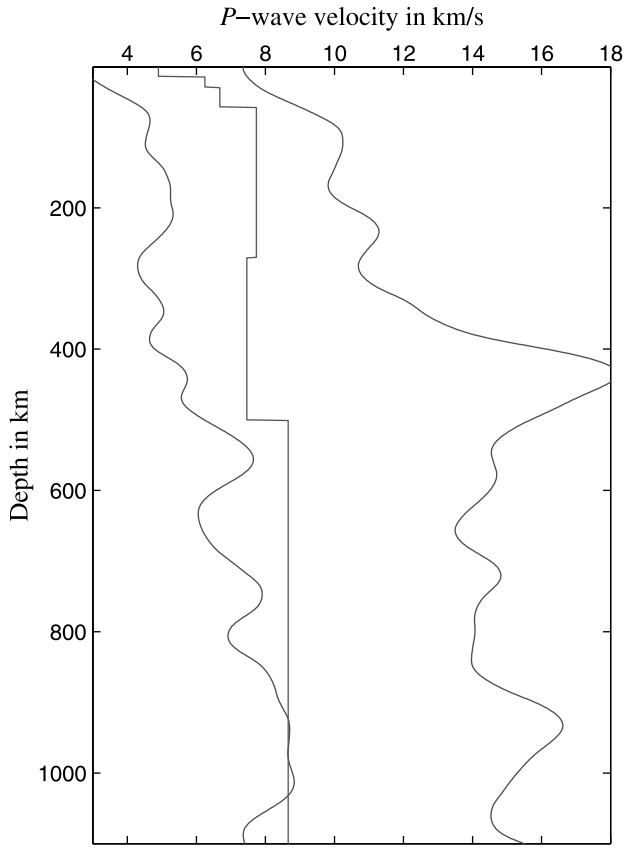


Figure 7. The lunar *P* wave velocity structure shown using 99% credible intervals. The broad velocity ranges evince the uncertainties involved. The structure in the middle shows the Nakamura velocity model [*Nakamura et al.*, 1982] so as to highlight consistency with earlier models.

velocity models shown here have been displayed using marginal posterior probability distributions, one-dimensional (1-D) marginals, actually. What this means is that if we were to choose the most probable velocity at a given depth, this would correspond to the most likely velocity at this particular depth only, while disregarding the information on all other parameters.

[41] Given that $\sigma(\mathbf{m})$ is a rather complicated entity, it is more convenient and probably also sufficient for general orientation regarding the uncertainty about **m** simply to describe regions of given probability under $\sigma(\mathbf{m})$. From this point of view the identification of intervals containing 50, 90, or 99% of the probability under $\sigma(\mathbf{m})$ might actually suffice to give a good idea of the general quantitative messages implicit in $\sigma(\mathbf{m})$, leading us to the calculation of credible intervals. The credible interval is defined as the shortest possible interval containing a given probability, and in Figure 7 we have depicted credible intervals containing 99% of the marginal probability. At first glance these credible intervals appear to be quite broad, which essentially indicates that the absolute value of the velocity in a given layer is not very well determined. As a consequence, the interpretation of the marginals shown in Figure 6 is a subtle issue where the use of qualities such as prudence and restraint are to be advocated. For example, if it happens that at a certain depth or within a certain depth range there might be a large probability for a somewhat high velocity, this is, because of the aforementioned ill-determined nature of the absolute velocities, by no means to be interpreted as the results stating this particular fact.

[42] This conundrum inherent in nonlinear problems is further illustrated with the following simple example. Figure 8a shows the nonlinear function investigated for the purpose (a third-degree polynomial). Also included is the true model, m_{true}, giving rise to noise-free data d_{true}. Let us imagine that we have performed some sort of physical measurement which results in an observed datum, d_{obs}, which is also shown. Let us also assume, mainly for simplicity, that noise concerning this data point is normally distributed with uncertainty σ. In this particular example, d_{obs} turns out to be displaced one σ from d_{true} (experimental details are included in the caption of Figure 8). Now, the situation is this: we have measured a datum, d_{obs}, and given this observation we would like to infer information about the model m (here one model parameter) by computing its posterior probability density (the prior probability density is assumed constant). What has been said so far is exactly what we have been describing hitherto in dealing with inverse problems. Next, the posterior probability distribution for a set of model parameter values is easily calculated, and Figure 8b shows this probability distribution. Here m_{true} has also been depicted, and it is obvious that the true model is situated far from the model of maximum posterior probability.

[43] This example more than anything else reveals the intricacies hidden in nonlinear problems and explains why the most probable model is not necessarily the one of utmost interest, since it might be far displaced, as indeed it is in the example outlined here (by more than 100%), from the true model, and this in spite of the fact that data

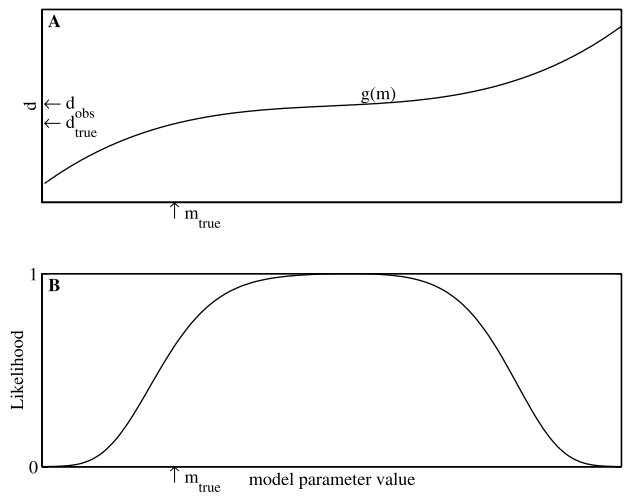


Figure 8. A simple illustration of a nonlinear problem: (a) the nonlinear function and (b) the posterior probability distribution for the model parameter m. Here the true model is displaced by more than 100% from the model with the greatest posterior probability, ultimately showing that care has to be exercised in interpreting the outcome of nonlinear problems. Details of the experiment are as follows: $g(m) = (\alpha m)^3 + \beta m$, where $\alpha = 0.7$ and $\beta = 700$; d_{obs} is assumed to be Gaussian distributed with uncertainty $\sigma = 10^5$ and whose actual value is given by $d_{\text{true}} + \sigma$. The posterior probability distribution is calculated from $L(m) = \exp(-\|d_{obs} - g(m)\|^2/2\sigma^2)$.

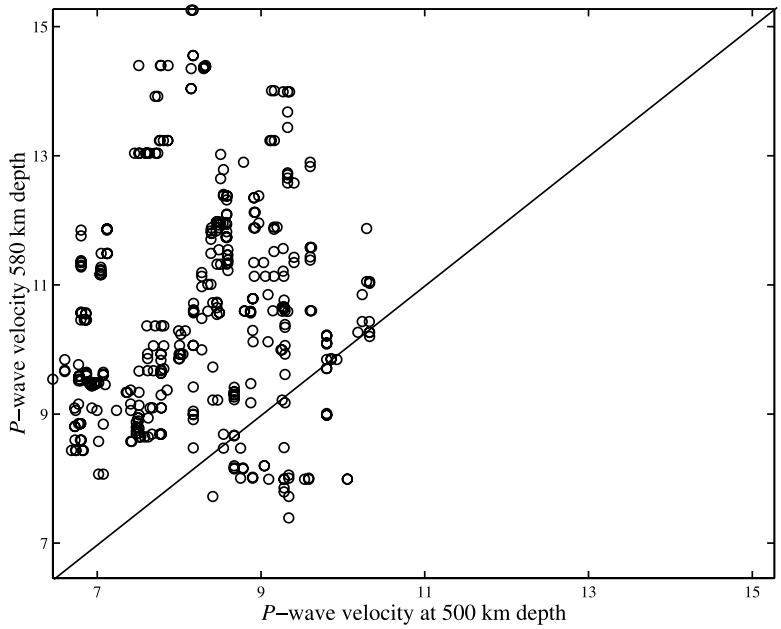


Figure 9. Two-dimensional marginal *P* wave velocity distribution depicting the correlation between sampled velocity parameters at 500 and 580 km depth. The distribution is clearly seen to be translated toward higher velocities at a depth of 580 km, indicating the existence of a velocity jump. Eighty-eight percent of the probability distribution lies above the equality line.

uncertainty has been modeled using a Gaussian distribution. With this example in mind it is also clear why, in the case of nonlinear problems, the model with the highest posterior probability does not necessarily correspond to the right solution and thus why interpretations using the most probable model can be erroneous.

[44] Returning to the credible intervals, we further superposed the Nakamura model on top of our credible intervals to highlight consistency. Although the Nakamura model at certain ranges lies just barely within ours, it is nonetheless contained within the 99% credible region, which shows that there is no inconsistency between our results and his model.

[45] On the other hand, an item which turns out to be well determined is the existence of a velocity discontinuity. This is easily evidenced using two-dimensional marginals. As an example, we plotted in Figure 9 the marginal probability distribution among two model parameters, the *P* wave velocity at 500 km and at 580 km depth, respectively, which depicts the correlation that exists between the two model parameters. From this it is apparent that by far the largest probability is displaced toward the model parameter at 580 km depth, clearly indicating that higher velocities at this depth are favored in comparison to those above. This translation of the probability distribution then is evidence of a velocity discontinuity. Whether the purported discontinuity is sharp or gradual is more difficult to assess with this method. It should be noted that this inference was made without any reference as to absolute velocity values.

[46] Of course, we could also display 3-D marginals, depicting the probability distribution among three model parameters at a time. This is exemplified in Figure 10, where we have shown examples of 3-D marginals depicting the sampled hypocentral coordinates of all deep moonquakes, giving a good representation of the their distribution in space.

6. Results and Discussion

[47] In interpreting the results, we have to be aware of the limitations which have been imposed in part on us by way of the nature of the data and in part by us in our choice of our particular model of the Moon and the inversion. Chief limitations imposed by data are, of course, due to the fact that only four stations were operative, covering only a small area of the front side of the Moon with two stations placed close to each other, raising the question whether they really represent independent degrees of freedom to determine the unknowns. And of course the complex lunar signal characteristics have right from the beginning been a major obstacle, limiting modeling to essentially first arrivals, although attempts at modeling the low-frequency part of the signals have also been undertaken [*Loudin*, 1979; *Khan and Mosegaard*, 2001].

[48] In terms of our model of the Moon the necessary assumption of spherical symmetry has to be cited as a limiting factor, and the obtained velocity structure is thus to be viewed as an average over the front side of the Moon, covered by the four seismic stations, with near-surface lateral heterogeneity modeled using local corrections. Although these local corrections are a simple way of dealing with lateral heterogeneity, by being a vertical average over the 1 km surficial low-velocity layer which has been stripped off, they nonetheless contain a clue to the differences among the four stations. However, whether this difference is due to actual differences in velocity, thereby reflecting different geological settings, or simply due to differences in topography among the four stations cannot be assessed by the local corrections. The importance of the local corrections,

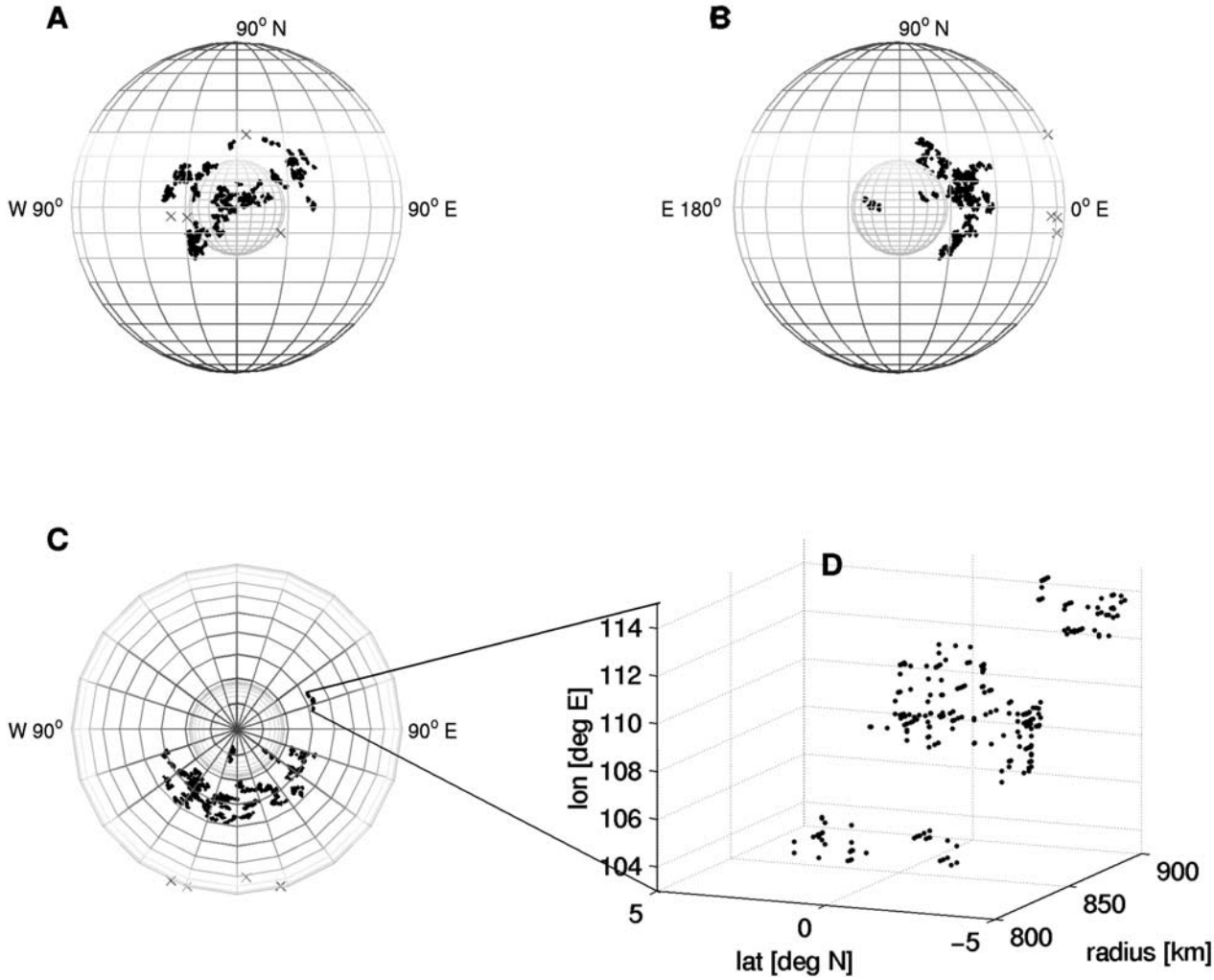


Figure 10. Sampled deep moonquake hypocenter coordinates showing the spatial distribution of quakes in the lunar interior. Figures 10a–10c display all sampled hypocenter coordinates for each deep moonquake source region from a number of different viewpoints. Since the individual source regions contain a cluster of many thousand samples, which are indvidually not distinguishable, Figure 10d depicts the distribution of sampled hypocenter coordinates for the lone farside moonquake, A_{33}, for enhancement. Figures 10a–10c also contain a central sphere with a radius of 500 km, which was included so as to enhance illustration of the spatial distribution of the moonquakes and is not meant to signify a lunar core. In Figure 10c we are viewing down on the lunar north pole, and the only farside moonquake observed to date, A_{33}, lying just over the eastern limb, is clearly visible. Crosses denote seismic stations. See color version of this figure at back of this issue.

however, becomes obvious when one considers the question as to whether or not stations 12 and 14 can be regarded as being independent sources of information. It is clear that with the large number of parameters that have to be determined the arrival times from these two stations should carry full weight. However, if the arrival time is delayed at one station relative to the other and this happens to be due to local structural differences beneath the two stations, such differences directly affect large systematic errors in structural parameters thus determined, as noted by *Nakamura* [1983]. The significance of the local corrections is thus obvious, since structural differences if extant are dealt with by these local parameters. Figure 11 depicts the sampled local corrections in terms of velocities, showing differences to be present beneath each of the four stations.

[49] From the point of view of the inversion there exists the usual trade-off between resolution and uncertainty. In this study it has to be emphasized that we have chosen the highest possible resolution given the constraint we had to satisfy to accommodate ray theory. The uncertainties are therefore to be viewed in the light of this. Had we, for example, chosen to increase the minimum layer thickness, this would have reduced the uncertainties on the velocities. This is partly the reason why the Nakamura model has smaller uncertainties. Accordingly, we have chosen the best possible resolution given ray theory and thus the most conservative estimate of model parameter uncertainties. As regards the MCMC algorithm, it has to be said that in the end we are able to pick only a limited number of samples from the posterior distribution. If the problem is

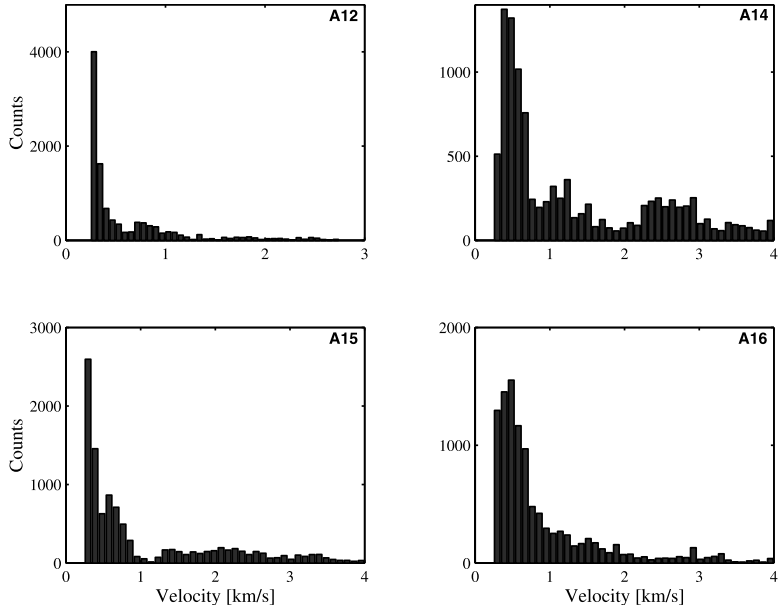


Figure 11. The *P* wave velocity structure of the uppermost 1 km of the lunar crust, comprising the very low velocity layer beneath each of the four Apollo landing sites. Mean values are $v_{12} = 0.6$ km s^{-1}, $v_{14} = 1.3$ km s^{-1}, $v_{15} = 1.0$ km s^{-1}, and $v_{16} = 0.9$ km s^{-1}.

too complex, one might tend to bypass solutions in the model space. This would result in an underestimation of the width of the uncertainty, because of the inadequate coverage of the model space. Stationarity in the likelihood function and model parameters, while sampling, is a necessary condition, but not a sufficient one to ensure that we have actually sampled the posterior distribution in a satisfactory manner. Despite these shortcomings in mathematical details, this method is still a much better tool than deterministic ones (like linearized methods), in that the deterministic methods tend to sample a model which lies close to its start, introducing a bias, since the uncertainty on the obtained model essentially depends on where one starts off.

[50] Another issue concerning this study which merits a comment is the very limited data set, giving in part rise to the error bounds on the results. Earlier studies assumed all arrival time readings equally uncertain. This has the unfortunate consequence of conferring equal weight to all data points. Good as well as bad arrival time readings are thus equally probable, which might lead to inconsistencies. Now, the point of view adopted here of assigning to a given arrival time an a priori uncertainty and the subsequent process of looking for and discarding outliers, resulting in a revision of the posterior uncertainty, should protect the results from these inconsistencies. On the other hand, it must be said that the procedure of eliminating certain data points on the grounds that they deviate from the assigned arrival time reading ± the uncertainty is a somewhat crude approach, since the given data point is strictly an event, which, however, owing to the complex nature of the seismogram, is extremely difficult to pick.

[51] In spite of this, inversion of the artificial impacts led to the crustal velocity structure (to a depth of roughly 100 km), inversion of meteoroid impacts in combination with the shallow moonquakes provided information on the lunar upper mantle (depth range roughly 100 to 500 km), and finally, inversion of the deep moonquakes provided information about the middle mantle velocity structure (depth range 500–1100 km). About 5×10^6 models were sampled in all, from which 5×10^4 were used for analysis in this study. The results are shown in Figure 6. Figure 12 shows the marginal distributions of sampled hypocenter coordinates for deep moonquake source A_7, depicting another way of presenting the full posterior distribution (Figure 10). The hypocentral coordinates for all deep and shallow moonquakes and the epicentral coordinates for the meteoroid impacts are compiled in Tables 2, 3, and 4, respectively.

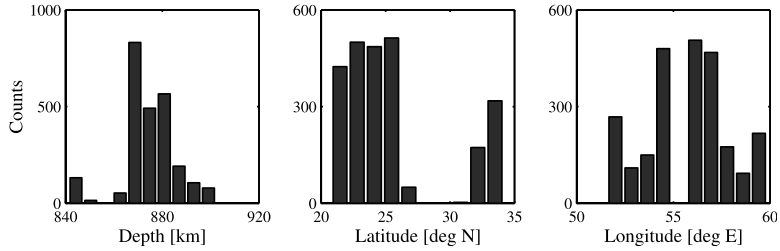


Figure 12. Marginal distributions depicting the sampled source depths and epicentral coordinates in terms of latitude and longitude for deep moonquake source A_7.

Table 2. Shallow Moonquake Hypocenter Estimates[a]

Event		Hypocenter Coordinates		
Year	Day	Depth, km	Latitude, °N	Longitude, °E
1971	107	74.5 ± 21.2	52.5 ± 2.8	26.9 ± 3.3
1971	140	47.5 ± 46.4	37.4 ± 2.1	−19.4 ± 3.7
1971	192	220.8 ± 48.8	49.4 ± 2.5	−27.2 ± 8.7
1972	002	172.5 ± 54.7	56.9 ± 4.2	111.4 ± 5.5
1973	072	48.9 ± 34.9	−81.9 ± 2.9	−130.7 ± 3.4
1973	171	114.8 ± 22.7	−5.4 ± 2.7	−68.1 ± 2.8
1974	192	132.6 ± 40.7	26.1 ± 4.9	85.1 ± 2.3
1975	003	74.8 ± 16.2	27.3 ± 3.0	−102.2 ± 3.5
1975	012	101.9 ± 25.6	63.1 ± 5.5	46.9 ± 3.5
1975	044	129.9 ± 70.3	−16.4 ± 2.3	−24.6 ± 3.1
1975	314	73.2 ± 18.7	−10.6 ± 2.7	58.9 ± 3.2
1976	004	88.0 ± 38.8	44.5 ± 3.2	30.5 ± 3.7
1976	066	74.4 ± 26.9	45.6 ± 3.9	−26.3 ± 4.8
1976	068	119.8 ± 23.9	−18.2 ± 2.3	−10.8 ± 1.6

[a] All estimates are mean values, and the values indicated by ± signify one standard deviation.

[52] A detailed discussion of how to interpret the data has been advanced, and there it was made rather obvious that, because of the large credible region encompassed by the models, the absolute velocities were ill-determined. However, velocity discontinuities were found by employing 2-D marginals. In Figure 9 we depicted the correlation between two model parameters at 500 and 580 km depth, respectively, and it was clearly seen that the probability distribution was displaced toward higher velocities at 580 km depth, indicating a transition in this general region. This estimate can be made quantitatively, since the actual probability for there being such a discontinuity can easily be calculated using (5), which states that all we have to do is to count the number of models purporting whatever feature we might be interested in. In the above case the probability is 88%, which speaks for itself. Figure 13 shows the outcome of a similar analysis conducted for the *S* wave models which gave a 99% probability for there being a velocity increase in the same region. While our results (Figure 6) appear to suggest constant velocity zones for the upper mantle, *Nakamura et al.* [1976] inferred a decrease of shear wave velocity starting at roughly 300 km depth associated with a lower Q for shear waves, using the decay of shear wave amplitude with distance and the relative arrival times of *P* and *S* waves. This decrease in v_s is easily investigated using 2-D marginals. The result is depicted in Figure 14, which clearly shows the probability distribution displaced toward higher velocities at 280 km than at 320 km depth with a value of 95%, although the involved velocity decreases are minor. No significant velocity changes were found for the *P* wave models.

Table 3. Meteoroid Impact Epicenter Estimates[a]

Event		Epicenter Coordinates	
Year	Day	Latitude, °N	Longitude, °E
1971	143	−6.5 ± 4.4	−18.3 ± 3.2
1971	163	27.8 ± 2.5	−36.7 ± 4.4
1971	293	32.1 ± 2.8	−36.4 ± 2.1
1972	134	−0.2 ± 2.1	−14.6 ± 1.3
1972	199	15.2 ± 12.6	133.6 ± 2.5
1972	213	29.0 ± 1.5	−6.3 ± 1.5
1972	324	65.4 ± 3.3	−21.5 ± 6.2
1974	325	−6.3 ± 2.1	22.8 ± 0.9
1974	349	−10.2 ± 7.7	−12.4 ± 2.1
1975	102	5.4 ± 1.9	33.8 ± 1.7
1975	124	−57.4 ± 11.2	−117.4 ± 3.7
1976	025	−2.5 ± 5.0	−71.4 ± 1.7
1976	319	4.1 ± 6.7	−86.9 ± 5.5
1977	107	−37.9 ± 5.9	−60.3 ± 4.6

[a] All estimates are mean values. The values indicated by ± signify one standard deviation.

Table 4. Deep Moonquake Hypocenter Estimates[a]

	Hypocenter Coordinates		
Source	Depth, km	Latitude, °N	Longitude, °E
A_1	921 ± 6	−19.9 ± 1.3	−40.9 ± 1.5
A_5	702 ± 10	17.4 ± 1.5	−39.1 ± 3.4
A_6	847 ± 8	36.2 ± 2.9	54.2 ± 1.4
A_7	876 ± 6	22.9 ± 1.4	55.6 ± 2.7
A_8	942 ± 14	−33.7 ± 2.3	−37.3 ± 2.3
A_9	992 ± 8	−7.9 ± 1.1	−14.4 ± 3.9
A_{11}	822 ± 11	6.6 ± 1.2	6.8 ± 2.0
A_{14}	928 ± 13	−25.2 ± 0.9	−37.6 ± 1.2
A_{15}	820 ± 24	−0.2 ± 2.5	−3.8 ± 3.4
A_{16}	1161 ± 29	7.1 ± 2.1	3.7 ± 1.5
A_{17}	820 ± 37	26.7 ± 1.4	−21.5 ± 1.0
A_{18}	918 ± 7	22.3 ± 1.6	32.1 ± 1.2
A_{19}	799 ± 4	17.0 ± 3.8	30.8 ± 3.3
A_{20}	960 ± 4	25.1 ± 1.3	−34.0 ± 1.6
A_{24}	985 ± 22	−37.8 ± 1.8	−42.7 ± 2.1
A_{25}	959 ± 12	37.0 ± 1.0	67.7 ± 2.4
A_{27}	1056 ± 13	20.9 ± 2.6	21.1 ± 2.0
A_{28}	1048 ± 8	8.3 ± 1.8	28.3 ± 2.1
A_{30}	915 ± 17	13.1 ± 1.5	−33.2 ± 1.6
A_{33}	902 ± 11	3.7 ± 2.1	112.3 ± 2.8
A_{34}	993 ± 13	6.2 ± 2.4	−7.6 ± 1.3
A_{36}	1016 ± 9	64.4 ± 2.5	−12.5 ± 1.5
A_{39}	933 ± 6	−16.9 ± 4.5	−14.6 ± 1.8
A_{40}	905 ± 23	−0.3 ± 1.4	−10.6 ± 2.9
A_{41}	801 ± 4	19.7 ± 1.2	−34.6 ± 0.9
A_{42}	915 ± 9	27.9 ± 3.1	−51.1 ± 3.0
A_{44}	941 ± 8	61.2 ± 2.4	52.9 ± 0.9
A_{50}	836 ± 16	11.4 ± 1.3	−52.9 ± 1.7
A_{53}	932 ± 12	−27.9 ± 3.6	−31.0 ± 2.1
A_{54}	968 ± 18	8.4 ± 1.8	−64.0 ± 4.7
A_{61}	801 ± 5	21.8 ± 1.1	43.4 ± 1.1
A_{71}	945 ± 26	14.3 ± 1.0	−12.4 ± 2.0
A_{84}	879 ± 18	−13.6 ± 2.8	−31.0 ± 1.6
A_{85}	821 ± 7	35.9 ± 2.7	68.1 ± 1.7
A_{97}	948 ± 10	4.5 ± 1.8	12.4 ± 3.6

[a] All estimates are mean values and the values indicated by ± signify one standard deviation. The source numbering follows that of *Nakamura et al.*, [1982].

[53] Turning to a different method of analysis of the posterior distribution leads us to our investigation of the much debated lunar crustal thickness using Bayesian hypothesis testing. Our analysis resulted in a Bayes factor of 4.2, which undeniably leads to the conclusion that hypothesis 1 is more plausible than hypothesis 2, thereby favoring the depth to the lunar Moho in the range 35–45 km. As a further aid in arguing for a shallow lunar crust we have displayed the models, satisfying the two hypotheses, where Figure 15 depicts the posterior velocity models satisfying $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Of prime interest in Figure 15a is the obviously discernible discontinuity at ∼38 ± 3 km depth, where the results indicate an increase in velocity from 6.8 ± 0.7 to 7.8 ± 0.6 km s^{-1}. This is slightly thinner than the value of 45±5 km inferred so far for the crustal thickness [*Khan et al.*, 2000], although it should be stressed that these two values do not mutually disagree, since the former has been obtained from a subset of all

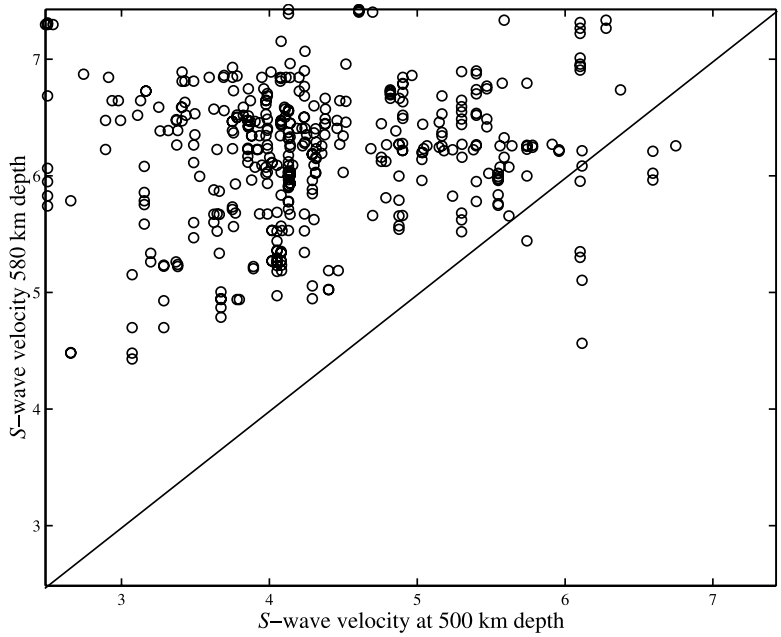


Figure 13. Two-dimensional marginal *S* wave velocity distribution depicting the correlation between sampled velocity parameters at 500 and 580 km depth. The distribution is clearly seen to be translated toward higher velocities at a depth of 580 km, indicating the existence of a velocity jump. Ninety-nine percent of the probability distribution is situated above the equality line.

sampled models, namely, those models that additionally concur with the constraints imposed by $\mathcal{H}_1$. The models which were found to have discontinuities in the depth range 50−70 km are displayed in Figure 15b. One might be tempted to label these models unphysical, but what is actually apparent here is the nonuniqueness inherent in the data we are dealing with, which leads to a multitude of possible models with clearly quite different implications for the lunar crustal velocity structure. The interpretation of Figure 15b is such that it shows that in order for the MCMC algorithm to generate models incorporating velocity discontinuities at depths around 60 km it seems at the same

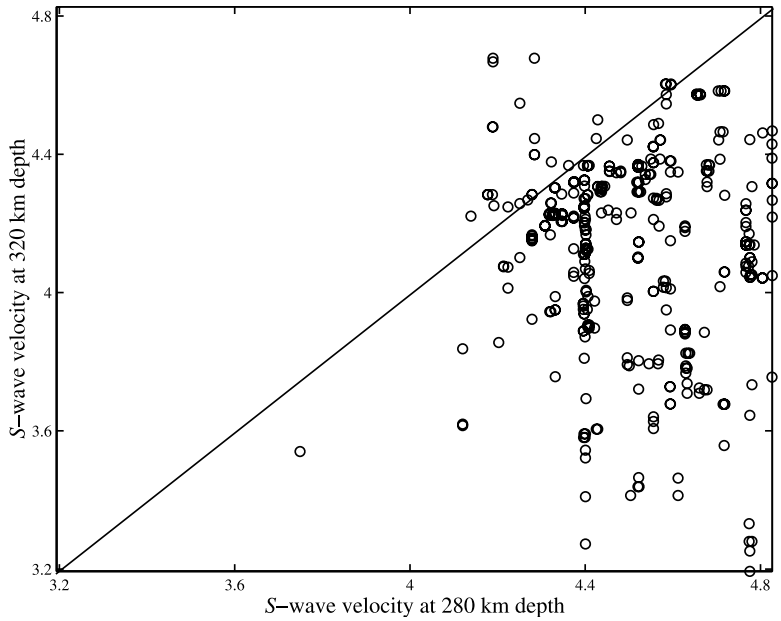


Figure 14. Two-dimensional marginal *S* wave velocity distribution depicting the correlation between sampled velocity parameters at 280 and 320 km depth. The distribution is clearly seen to be translated toward higher velocities at a depth of 280 km, indicating the existence of a velocity decrease in this region. Ninety-five percent of the probability distribution lies below the equality line.

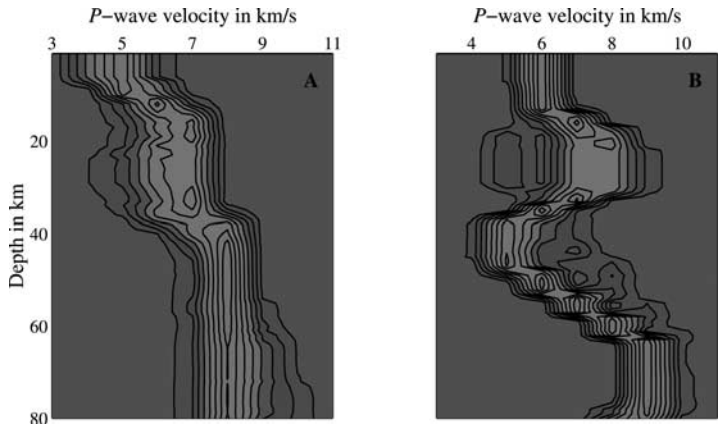
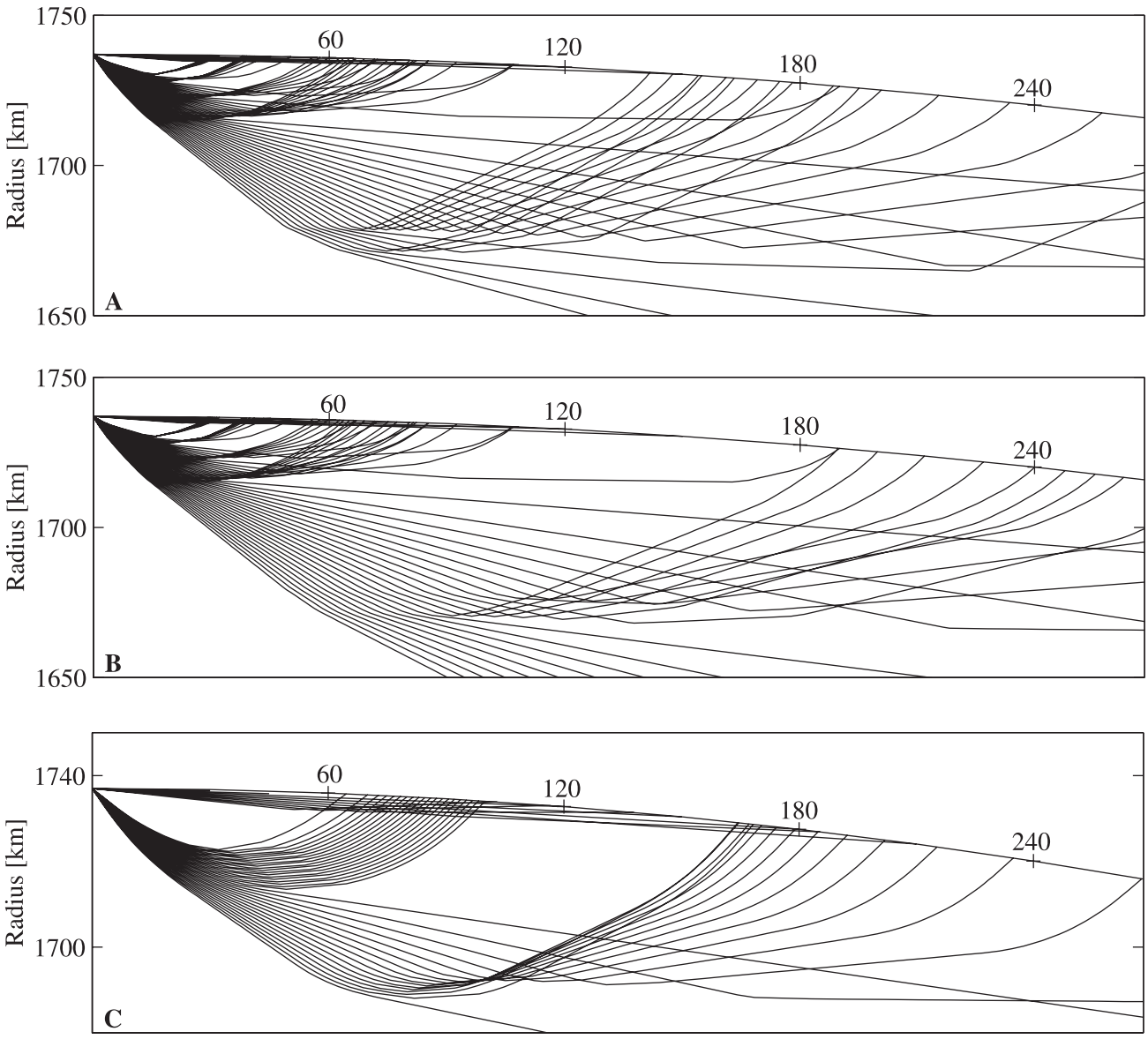


Figure 15. (a) Marginal posterior P wave velocity distributions depicting the lunar crustal velocity structure obtained from inversion of artificial impacts. The figure has been constructed from a total of 2709 models having satisfied $\mathcal{H}_1$. The contour lines define eight equal-sized probability density intervals for the distributions. (b) Marginal posterior P wave velocity distributions for those models that satisfied $\mathcal{H}_2$. The figure has been constructed from a total of 905 models. See color version of this figure at back of this issue.



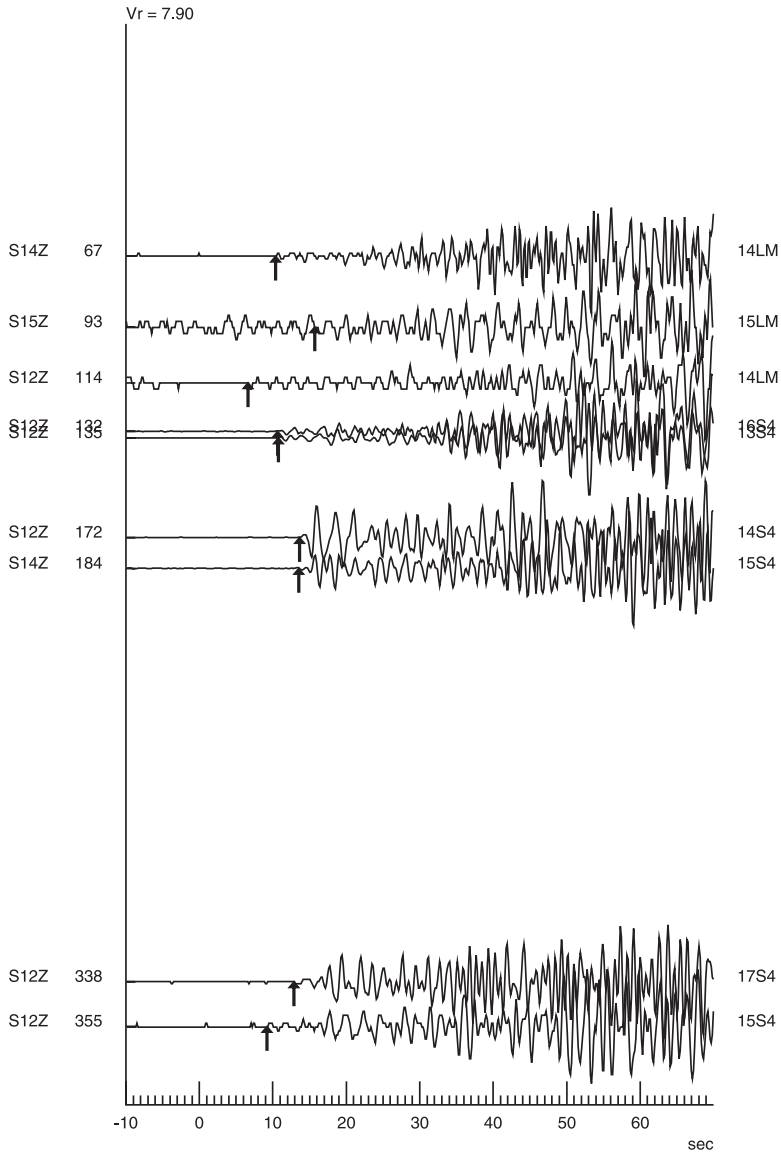


Figure 17. Composite plot of seismograms recorded from all artificial impacts on vertical components as a function of epicentral distance. Notation "S14Z 67" on the left of the first trace means station 14 vertical component recorded at an epicentral distance of 67 km, whereas "14LM" on the right of the trace signifies the mission and the impactor, in this case the lunar module of mission 14. Note the large-amplitude arrival at a distance of 172 km. Arrows indicate the position of first *P* wave arrivals as read by J. Gagnepain-Beyneix (manuscript in preparation, 2002). Seismograms are plotted to the same scale.

time to have a preference for generating substantial velocity decreases in the lower crust. Now, the sampling of such geologically unsound models could easily have been restricted. However, this would have required the introduction of prior information into the inverse problem, which implies constraining the numerical system, thereby reducing its nonuniqueness and not exposing the true model variability, which, as noted, was one of the primary motivations for conducting our reanalysis of the Apollo lunar seismic data. Moreover, what actually constitutes ''geologically unsound'' features in a model is often debatable given its subjectivity.

[54] Of relevance in assessing Bayesian hypothesis testing as a tool for inferring conclusions is to inquire whether the conclusions reached are contingent upon the hypotheses; that is, is the Bayes factor dependent upon the specific

Figure 16. (opposite) Seismic ray paths in the Moon for a surface source. Figures 16a and 16b correspond to the Toksöz model with velocity discontinuities of 2 km/s and 1 km/s placed at 60 km depth. Figure 16c shows ray paths using the velocity model obtained here, with a velocity discontinuity of 1 km/s at a depth of 42 km. Where the density of rays is high, focusing of energy has occurred, leading to large amplitudes. Numbers on the lunar surface denote epicentral distance in kilometers.

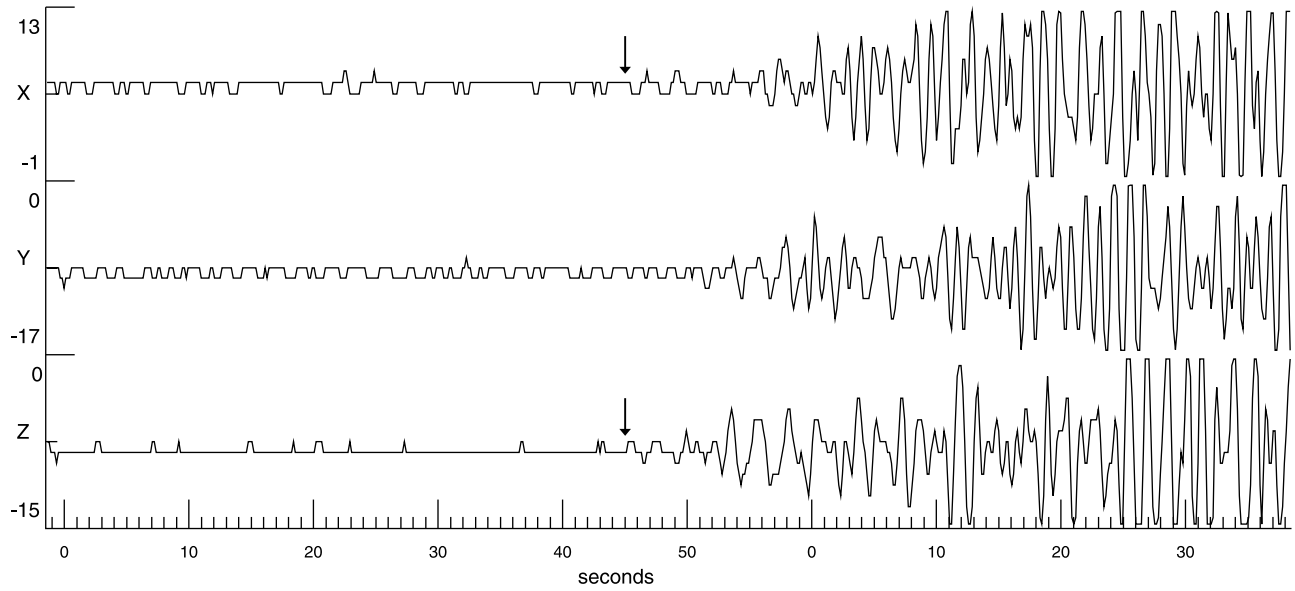


Figure 18. The SIVB 15 impact, 29 July 1971, as recorded at ALSEP 12 (LP) at a distance of 356 km. Traces start at 20 h 58 min 50.5 s. Ground motion is given in du. Here X and Y represent the horizontal components of ground motion, and Z corresponds to vertical motion. The two peaks immediately after the arrow are clearly visible on the Z component and were identified as noise by *Toksöz et al.* [1972]. Ahead of these two pulses the signal itself contains very little noise, and the present authors are therefore more prone to picking the first arrival at the arrow, rather than 5 s later. While the Y component does not reveal anything, the X component contains a signal (indicated by the arrow) that corresponds to the one on the vertical component.

constraints employed in describing the physical features of interest? This is easily verified by simply altering the constraints. As an example, let us look for discontinuities in our models specified according to the following definitions:

1. Velocity gradients should reach 0.3 km s^{-1} km^{-1}.
2. The velocity beneath the crust should reach a value of 8.0 km s^{-1}.

[55] We also extended the depth range in which we seek the crust-mantle boundary of hypothesis 1, from 35–45 km to 30–50 km, so as to have the same size as the range considered in hypothesis 2, but intuitionally searching a greater depth range should just result in an augmented number of models satisfying $\mathcal{H}_1$. The Bayes factor for this analysis is $\mathcal{B}_{12} = 7.5$, leading to a further increase in the plausibility of $\mathcal{H}_1$ relative to $\mathcal{H}_2$. This should confirm the reliability of hypothesis testing.

[56] In view of the discrepancy between the crustal thickness estimate presented in this study and the ones obtained earlier, with values ranging from 65 km [*Toksöz et al.*, 1972] over 55 km [*Toksöz et al.*, 1974] to 58 km [*Nakamura et al.*, 1982] for the region below the Apollo 12 and 14 landing sites, a few comments are merited. (For the sake of completeness, it should be noted that another recent reanalysis of the Apollo lunar seismic data set, although using a different inversion technique, has also resulted in a thinner estimate for the crustal thickness [*Chenet et al.*, 2002].) In the study of *Toksöz et al.* [1972, 1974], amplitude data and synthetic seismograms were employed in addition to the travel times from the man-made impacts.

[57] Now, amplitude data do provide additional information on velocity discontinuities, since phenomena such as geometric focusing, defocusing, and reflection of seismic rays are controlled by velocity gradients. These effects are illustrated in Figures 16a, 16b, and 16c for three different velocity models. Figures 16a and 16b depict ray paths in the crust using the Toksöz model, with velocity discontinuities of 2 and 1 km/s situated at 60 km depth, respectively. It is clear that in order to focus rays at the appropriate epicentral distance with the Toksöz model so as to obtain the observed large first P wave arrival at a distance of 172 km (see Figure 17), a discontinuity of 2 km/s is needed, which is probably difficult to realize. Figure 16c, on the other hand, shows that it is indeed possible to fit the same data by a model corresponding to the one presented in this study, including a 1 km/s discontinuity at a depth of 42 km.

[58] The positive velocity discontinuity will, of course, produce a multiplicity in the travel time curve corresponding to the different arrivals, like the refracted first arrival, the direct arrival, and, if present, any reflected arrivals. Therefore the presence of any later P wave arrivals will provide additional evidence for velocity gradients or discontinuities inside the Moon. In the studies of *Toksöz et al.* [1972, 1974], later arrivals as well as surface reflected phases were reportedly identified in the seismograms and used to further evince the existence of a high-velocity discontinuity. However, the notorious coda seen on lunar seismograms, produced by intense scattering, obscures secondary and all later arriving phases, thereby severely hampering their identification. This

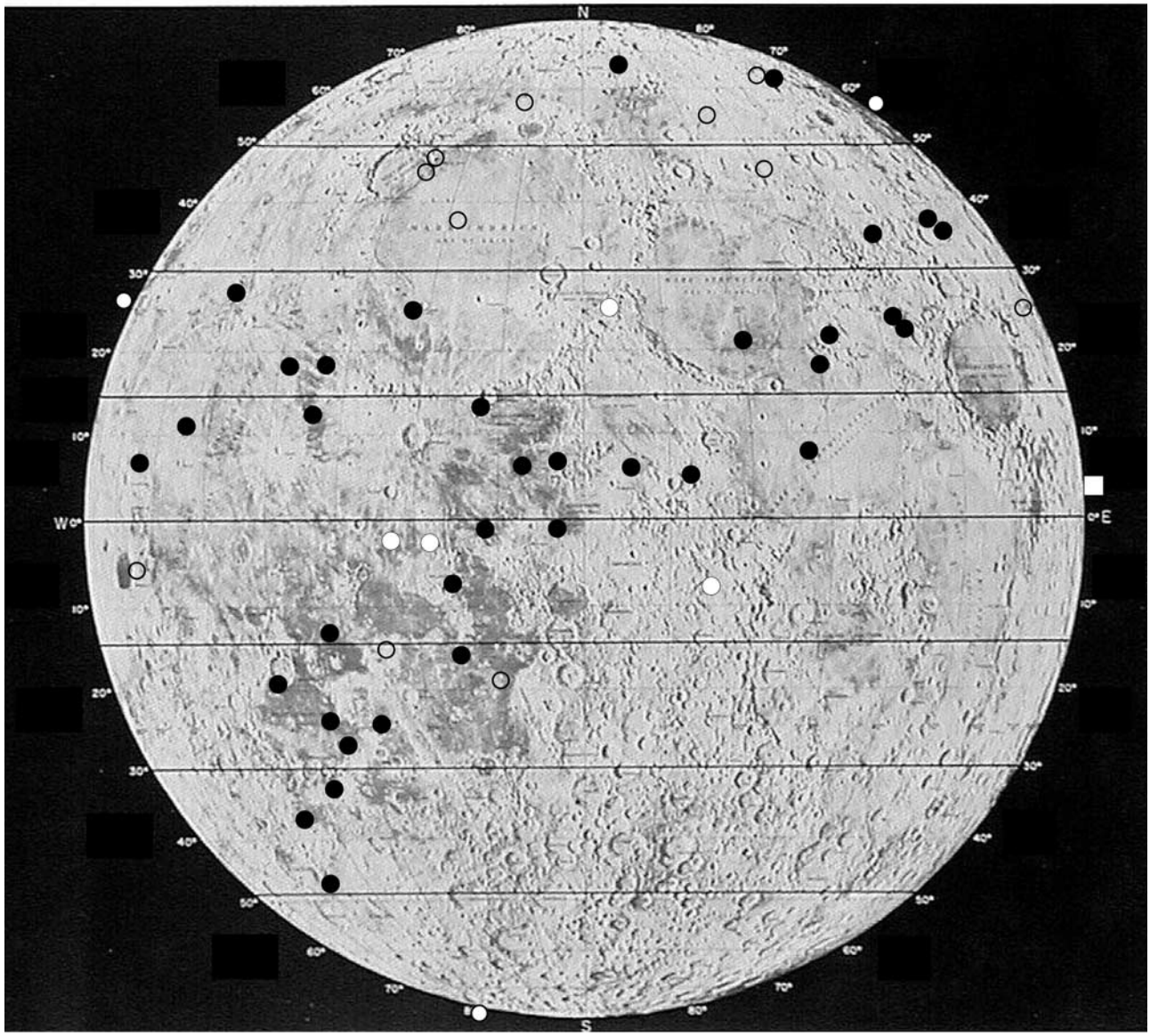


Figure 19. Map of the lunar nearside showing the four Apollo stations as well as deep and shallow moonquake epicenters taken from Tables 2 and 4, respectively. The white dots on the face of the Moon denote the four stations. Black dots indicate deep moonquakes, and open circles denote shallow moonquakes. On the rim of the Moon are additionally located three white dots indicating the location of the shallow moonquakes on the lunar farside. The white square on the right near the lunar equator shows the location of the lone farside deep moonquake.

does not necessarily mean that none of the purported secondary arrivals are real. Some of them may truly represent reflected and converted phases from some interface, but others may quite possibly be due to side scattering (Y. Nakamura, personal communication, 2001).

[59] The calculation of theoretical seismograms defines a more definitive approach, since these can be directly compared to their observed equivalents. The computation of synthetic seismograms is mostly dependent on the velocity structure but, of course, also depends on the seismic source and the response of the seismometer. For details of how seismograms were computed the reader is referred to *Toksöz et al.* [1972]. Figure 6 of that manuscript, in particular, shows a comparison of observed and synthetic seismograms, and it is apparent even from a first look that discrepancies are present, especially when considering the impact recorded at an epicentral distance of 356 km. Concerning this particular seismogram, the authors state that (p. 497): "...the first two peaks of the observed seismograms are noise pulses and can be clearly identified as such in unfiltered seismograms." However, an independent review of this event, by the present authors (see Figure 18) does not qualify the two pulses as noise, but rather as the start of the signal, which, of course, leaves the synthetics and thereby the velocity model used in deriving these questionable.

[60] Having said this, one should be aware that the act of using uncertain data, in the sense of those discussed here when solving inverse problems, should be cautioned, since the probable inconsistency with which, as an example, secondary later arrivals have been read can lead to biasing

of the final result, as it is succinctly stated by *Tarantola* [1987]:

> ....With such kinds of data, it is clear that the subjectivity of the scientist plays a major role. It is indeed the case, whichever inverse method is used, that results obtained by different scientists (for instance for the location of a hypocenter) from such data sets are different. Objectivity can only be attained if the data redundancy is great enough that differences in data interpretation among different observers do not significantly alter the models obtained.

[61] In having performed a relocalization of all events, except the artificial impacts, we will briefly discuss the spatial distribution of the natural seismic events. The current depth distribution (50–220 km depth) of shallow moonquakes confirms earlier evidence suggesting the upper mantle to be tectonically active, implying that this region of the Moon is the only zone in the lunar interior where tectonic stresses thought to arise from thermal contraction and expansion are high enough to cause mechanical failure [*Nakamura et al.*, 1979]. When the epicenters are drawn on the lunar surface, some shallow moonquakes appear to occur, now as then, on the edges of impact basins, which had led to the suggestion that they were correlated [*Nakamura et al.*, 1979] (see Figure 19). However, because of the limited number of detected events, the statistical significance of this correlation must be considered inconclusive.

[62] Concerning the deep moonquakes, it was observed in earlier studies [*Lammlein et al.*, 1974; *Lammlein*, 1977; *Nakamura et al.*, 1982] that their source regions clustered within seismic belts. Two deep belts were identified by *Lammlein et al.* [1974], one belt extending ENE to WSW and another belt extending south. In addition, *Nakamura et al.* [1982] discerned two more belts in the NW quadrant trending W and WNW. The present distribution seems more to define a main SW to NE trending belt extending from source A_{24} at 37.8°S and 42.7°W to A_{85} at 35.9°N and 68.1°E, comprising 60% of the events investigated here and lying on an arc of a great circle (see Figure 19). This belt is ~250 km in width and 3800 km in length as measured on the lunar surface. The linear trends inferred earlier in the NW quadrant are less conspicuous with the present distribution. The only farside moonquake observed to date, situated at 3.7°N and 112.3°E, most clearly presents an isolated, well-located source of moonquake activity and does not appear to be part of the main NE to SW trending belt as previously noted [*Lammlein et al.*, 1974]. The curious lack of moonquake activity in the SE quadrant was also observed earlier and was ascribed as being due to the presence of a localized high attenuation zone or alternatively may simply be due to the absence of deep moonquakes beneath this mostly highland region [*Nakamura et al.*, 1982].

[63] As concerns the depth distribution of deep moonquakes, we have found that 16 events, corresponding to almost 50%, fall in the narrow range from 850 to 950 km.

7. Conclusion

[64] In this paper we have presented a detailed analysis of a general inverse problem as is constituted by the inversion of the Apollo lunar seismic data. This was done by using a MCMC sampling algorithm, that is, by performing a random walk in a multidimensional model space that combines prior information with information from measurements and from the theoretical relationship between data and model parameters. Input to the algorithm were random models generated according to the prior distribution and the likelihood function. As output we assimilated random realizations of the posterior distribution, which contains all the information about the parameterized physical system derivable from available sources.

[65] We furthermore gave examples of how to investigate the models comprising the complex posterior distribution in a statistical fashion, using credible intervals, 2-D marginals, and Bayesian hypothesis testing. Common to all of them is the notion of conjuring up questions relating to any particular property of our interest, like the depth of a particular deep moonquake, for instance. All models exhibit an example of this property, and it may turn out that all of them provide the same value for it, in which case we would say that the depth of this moonquake is well-constrained by data. On the other hand, it might also happen that all models end up giving different answers to that question, thus rendering this feature ill-determined. Whatever the particular method, the point to note here is that within the probabilistic formulation of the general inverse problem, doing statistics is an all important asset which provides us with a clear probabilistic answer.

[66] It was also shown that because of the large credible intervals containing 99% of the probability distribution, the absolute velocities were not very well-determined. Two-dimensional marginals, on the other hand, could be used to investigate the correlation between any two parameters, shedding light on the presence of velocity changes, discontinuities, and the like. Moreover, Bayesian hypothesis testing was shown to be a very informative tool in providing evidence for the relative plausibility between any two hypotheses concerning features of interest. The advantage of using this form of hypothesis testing to examine a set of models as to certain features was made obvious. Instead of having to a priori constrain the system in certain ways from the outset, which means having to run the algorithm again with different sets of prior information, we chose to invoke as few prior constraints as possible so as to facilitate the comparison between any two hypotheses in order to gauge which particular feature is most probable given data and prior information. We specifically employed Bayesian hypothesis testing to distinguish between a thin and a thick lunar crust, which resulted in a Bayes factor $\mathcal{B}_{ij}$ of 4.2, clearly favoring a thinner crust with a thickness around 38 km.

8. Future Work

[67] Finally, we would like to touch on the matter alluded to earlier regarding the sampling of *P* and *S* wave velocities. Let us consider the following two relations governing the wave velocities as a function of the material parameters:

$$v_p = \sqrt{\frac{\kappa + 4/3\mu}{\rho}}, \tag{9}$$

$$v_s = \sqrt{\frac{\mu}{\rho}}. \quad (10)$$

From these two equations it is immediately apparent that the parameters v_p and v_s considered in this study are not independent when considered in terms of the elastic constants describing the medium. This raises the interesting problem of inverting for these parameters instead of the wave velocities using the same arrival time data set, since the inherent dependency that is here brought out would probably result in narrower distributions.

[68] **Acknowledgments.** We are grateful to Yosio Nakamura for a constructive review of this paper. An anonymous reviewer is also thanked for comments. Stimulating and highly informative discussions with Albert Tarantola concerning this study and inverse problems in general are much appreciated by the first author. Finally, we would like to extend our gratitude to Jeannine Gagnepain-Beyneix for making available Figures 17 and 18.

References

Barnett, V., and T. Lewis, *Outliers in Statistical Data*, John Wiley, New York, 1984.

Bernardo, J., and A. Smith, *Bayesian Theory*, John Wiley, 586 pp., New York, 1994.

Chenet, H., J. Gagnepain-Beyneix, and P. Lognonne, A new geophysical view of the Moon, *Lunar Planet. Sci.*, *XXXIII*, abstract 1684, 2002.

Duennebier, F., and G. Sutton, Thermal moonquakes, *J. Geophys. Res.*, *79*, 4351, 1974.

Goins, N., A. Dainty, and M. Toksöz, Seismic energy release of the Moon, *J. Geophys. Res.*, *86*, 378, 1981a.

Goins, N., A. Dainty, and M. Toksöz, Lunar seismology: The internal structure of the Moon, *J. Geophys. Res.*, *86*, 5061, 1981b.

Good, I., The interface between statistics and philosophy of science, *Stat. Sci.*, *3*, 386, 1988.

Hampel, F., P. Rousseeuw, E. Ronchetti, and W. Stabel, *Robust Statistics: The Approach Based on Influence Functions*, John Wiley, New York, 1986.

Khan, A., K. Mosegaard, and K. Rasmussen, A new seismic velocity model for the Moon from a Monte Carlo inversion of the Apollo lunar seismic data, *Geophys. Res. Lett.*, *27*, 1591, 2000.

Khan, A., and K. Mosegaard, New information on the deep lunar interior from an inversion of lunar free oscillation periods, *Geophys. Res. Lett.*, *28*, 1791, 2001.

Kovach, R., and J. Watkins, Apollo 17 seismic profiling—Probing the lunar crust, *Science*, *180*, 1063, 1973.

Koyama, J., and Y. Nakamura, Focal mechanism of deep moonquakes, *Proc. Lunar Planet. Sci. Conf. 11th*, 1855, 1980.

Lammlein, D., Lunar seismicity and tectonics, *Phys. Earth Planet. Inter.*, *14*, 224, 1977.

Lammlein, D., G. Latham, J. Dorman, Y. Nakamura, and M. Ewing, Lunar seismicity, structure and tectonics, *Rev. Geophys.*, *12*, 1, 1974.

Latham, G., M. Ewing, F. Press, and G. Sutton, Apollo Passive Seismic Experiment, *Science*, *165*, 241, 1969.

Latham, G., et al., Seismic data from man-made impacts on the Moon, *Science*, *170*, 620, 1970.

Latham, G., M. Ewing, F. Press, G. Sutton, J. Dorman, Y. Nakamura, M. Toksöz, D. Lammlein, and F. Duennebier, Passive Seismic Experiment, in *Apollo 16 Preliminary Science Report*, *NASA Spec. Publ.*, *NASA SP-315*, sec. 9, 29 pp., 1972.

Loudin, M., Structural-compositional models of the lunar interior, M.S. thesis, Pa. State Univ., Univ. Park, 1979.

Mosegaard, K., Resolution analysis of general inverse problems through inverse Monte Carlo sampling, *Inverse Problems*, *14*, 405, 1998.

Mosegaard, K., and A. Tarantola, Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res.*, *100*, 12,431, 1995.

Nakamura, Y., HFT events: Shallow moonquakes?, *Phys. Earth Planet. Inter.*, *14*, 217, 1977.

Nakamura, Y., A_1 moonquakes: Source distribution and mechanism, *Proc. Lunar Planet. Sci. Conf. 9th*, 3589, 1978.

Nakamura, Y., Shallow moonquakes: How they compare with earthquakes, *Proc. Lunar Sci. Conf. 11th.*, 1847, 1980.

Nakamura, Y., Seismic velocity structure of the lunar mantle, *J. Geophys. Res.*, *88*, 677, 1983.

Nakamura, Y., D. Lammlein, G. Latham, M. Ewing, J. Dorman, F. Press, and M. Toksöz, New seismic data on the state of the deep lunar interior, *Science*, *181*, 49, 1973.

Nakamura, Y., J. Dorman, F. Duennebier, M. Ewing, D. Lammlein, and G. Latham, High-frequency lunar teleseismic events, *Proc. Lunar Sci. Conf. 5th*, 2883, 1974a.

Nakamura, Y., G. Latham, D. Lammlein, M. Ewing, F. Duennebier, and J. Dorman, Deep lunar interior inferred from recent seismic data, *Geophys. Res. Lett.*, *1*, 137, 1974b.

Nakamura, Y., F. Duennebier, G. Latham, and H. Dorman, Structure of the lunar mantle, *J. Geophys. Res.*, *81*, 4818, 1976.

Nakamura, Y., G. Latham, H. Dorman, A. Ibrahim, J. Koyama, and P. Horvath, Shallow moonquakes: Depth, distribution and implications as to the present state of the lunar interior, *Proc. Lunar Planet. Sci. Conf. 10th*, 2299, 1979.

Nakamura, Y., G. Latham, and J. Dorman, How we processed Apollo lunar seismic data, *Phys. Earth Planet. Inter.*, *21*, 218, 1980.

Nakamura, Y., G. Latham, J. Dorman, and J. Harris, Passive Seismic Experiment long-period event catalog, Final Version, 1969 day 202–1977 day 273, *Galveston Geophys. Lab. Contrib. 491*, 314 pp., Univ. of Tex. at Austin, 1981.

Nakamura, Y., G. Latham, and J. Dorman, Apollo Lunar Seismic Experiment—Final summary, *J. Geophys. Res.*, *87*, A117, 1982.

Sellers, P., Seismic evidence for a low-velocity lunar core, *J. Geophys. Res.*, *97*, 11,663, 1992.

Tarantola, A., *Inverse Problem Theory*, 613 pp., Elsevier Sci., New York, 1987.

Tarantola, A., and B. Valette, Inverse problems: Quest for information, *J. Geophys.*, *50*, 159, 1982.

Toksöz, M., Geophysical data and the interior of the Moon, *Annu. Rev. Earth Planet. Sci.*, *2*, 151, 1974.

Toksöz, M., et al., Velocity structure and properties of the lunar crust, *Moon*, *4*, 490, 1972.

Toksöz, M., A. Dainty, S. Solomon, and K. Anderson, Structure of the Moon, *Rev. Geophys.*, *12*, 539, 1974.

A. Khan and K. Mosegaard, Department of Geophysics, Niels Bohr Institute for Astronomy, Physics and Geophysics, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen O, Denmark. (amir@gfy.ku.dk; klaus@gfy.ku.dk)

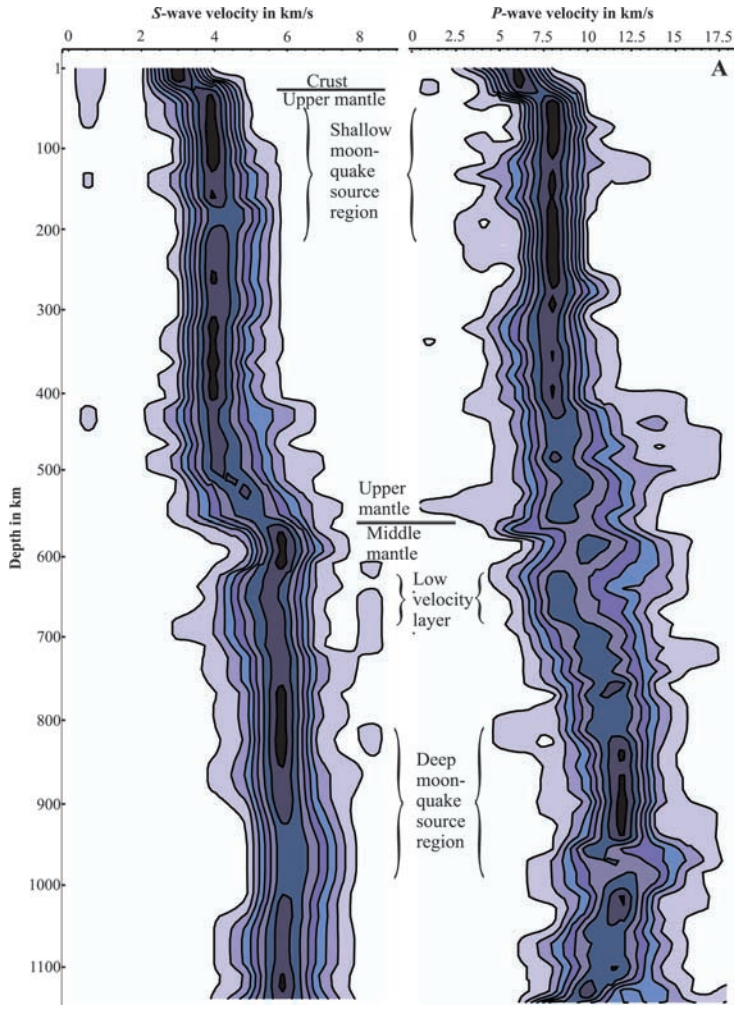


Figure 6. The marginal posterior velocity distributions depicting the velocity structure of the Moon. A total of 50,000 models have been used in constructing the two results. For each kilometer a histogram reflecting the marginal probability distribution of sampled velocities has been set up. By lining up these marginals, the velocity as a function of depth is envisioned as contours directly relating their probability of occurrence. The contour lines define nine equal-sized probability density intervals for the distributions. The uncertainties on the results are in part due to the large uncertainty in arrival time readings. It should be kept in mind that the velocity models are depicted using marginal probability distributions, and as such a model incorporating velocities of maximum probability does not necessarily correspond to the most likely model.

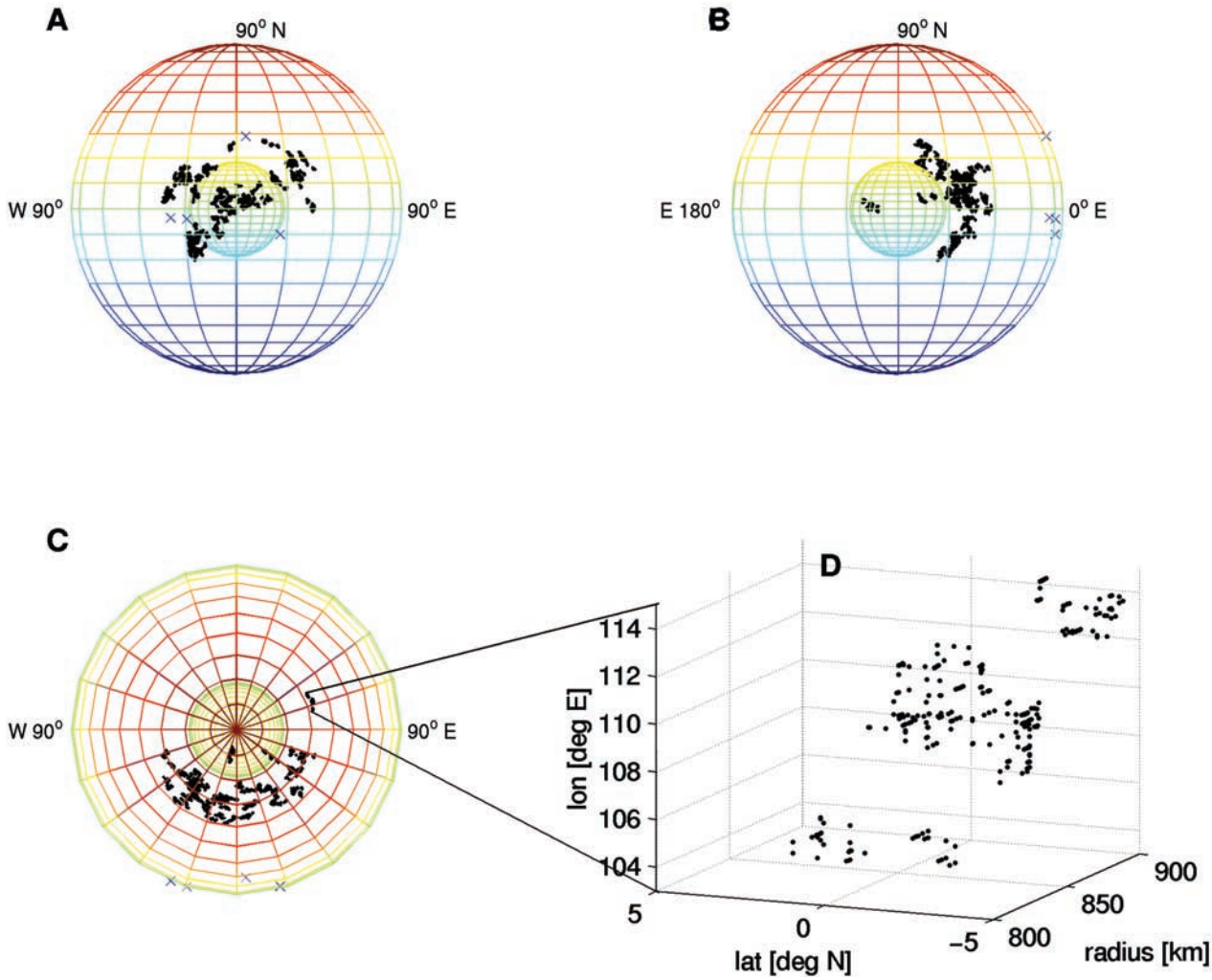


Figure 10. Sampled deep moonquake hypocenter coordinates showing the spatial distribution of quakes in the lunar interior. Figures 10a–10c display all sampled hypocenter coordinates for each deep moonquake source region from a number of different viewpoints. Since the individual source regions contain a cluster of many thousand samples, which are indvidually not distinguishable, Figure 10d depicts the distribution of sampled hypocenter coordinates for the lone farside moonquake, A_{33}, for enhancement. Figures 10a–10c also contain a central sphere with a radius of 500 km, which was included so as to enhance illustration of the spatial distribution of the moonquakes and is not meant to signify a lunar core. In Figure 10c we are viewing down on the lunar north pole, and the only farside moonquake observed to date, A_{33}, lying just over the eastern limb, is clearly visible. Crosses denote seismic stations.

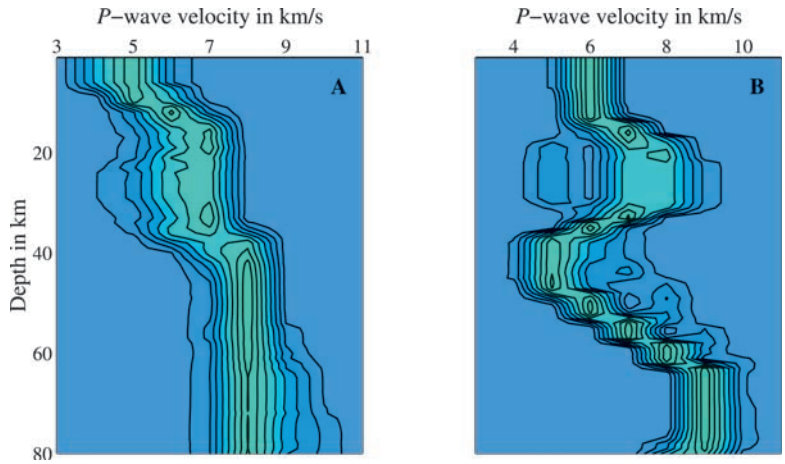


Figure 15. (a) Marginal posterior P wave velocity distributions depicting the lunar crustal velocity structure obtained from inversion of artificial impacts. The figure has been constructed from a total of 2709 models having satisfied $\mathcal{H}_1$. The contour lines define eight equal-sized probability density intervals for the distributions. (b) Marginal posterior P wave velocity distributions for those models that satisfied $\mathcal{H}_2$. The figure has been constructed from a total of 905 models.

INSTITUTE OF PHYSICS PUBLISHING INVERSE PROBLEMS

Inverse Problems **18** (2002) R29–R54 PII: S0266-5611(02)79161-2

TOPICAL REVIEW

Monte Carlo analysis of inverse problems

Klaus Mosegaard[1,3] **and Malcolm Sambridge**[2]

[1] Niels Bohr Institute, Department of Geophysics, Juliane Maries Vej 30, 2100 Copenhagen, Denmark
[2] Research School of Earth Sciences, Australian National University, Canberra, ACT 0200, Australia

E-mail: klaus@gfy.ku.dk

Received 19 July 2001, in final form 9 January 2002
Published 8 April 2002
Online at stacks.iop.org/IP/18/R29

Abstract

Monte Carlo methods have become important in analysis of nonlinear inverse problems where no analytical expression for the forward relation between data and model parameters is available, and where linearization is unsuccessful. In such cases a direct mathematical treatment is impossible, but the forward relation materializes itself as an algorithm allowing data to be calculated for any given model.

Monte Carlo methods can be divided into two categories: the sampling methods and the optimization methods. Monte Carlo sampling is useful when the space of feasible solutions is to be explored, and measures of resolution and uncertainty of solution are needed. The Metropolis algorithm and the Gibbs sampler are the most widely used Monte Carlo samplers for this purpose, but these methods can be refined and supplemented in various ways of which the neighbourhood algorithm is a notable example.

Monte Carlo optimization methods are powerful tools when searching for globally optimal solutions amongst numerous local optima. Simulated annealing and genetic algorithms have shown their strength in this respect, but they suffer from the same fundamental problem as the Monte Carlo sampling methods: no provably optimal strategy for tuning these methods to a given problem has been found, only a number of approximate methods.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

About a hundred years ago, it was recognized that integrals of the form

$$I = \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \tag{1}$$

[3] Author to whom any correspondence should be addressed.

where $h(\boldsymbol{x})$ and $f(\boldsymbol{x})$ are functions for which $h(\boldsymbol{x})f(\boldsymbol{x})$ is integrable over the space $\mathcal{X}$, and $f(\boldsymbol{x})$ is a non-negative valued, integrable function satisfying

$$\int_{\mathcal{X}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1 \tag{2}$$

could, in principle, be evaluated numerically by generating random samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ of $\boldsymbol{x}$ using $f(\boldsymbol{x})$ as a probability distribution (see, e.g., Housholder (1951)). An approximation to the integral could then be calculated as

$$I \approx \frac{1}{N}\sum_{n=1}^{N} h(\boldsymbol{x}_n). \tag{3}$$

Practical use of this method has today become a reality through application of *Monte Carlo* algorithms running on high-speed computers. Monte Carlo methods are numerical processes that produce so-called *pseudo-random numbers*, that is, a series of numbers that appear random if tested with any reasonable statistical test. The basic operation of a Monte Carlo algorithm is generation of pseudo-random numbers uniformly distributed over the interval [0, 1]. Once such a sample x_i is produced, it can be transformed into a pseudo-random sample from any given one-dimensional probability distribution $f(x)$, using simple rules.

As long as we work in one dimension, the Monte Carlo method is inefficient, and hence not a useful alternative to more direct methods for numerical evaluation of (1). If however, $\boldsymbol{x}$ belongs to a high-dimensional space, the Monte Carlo method may become the only feasible method. All other numerical methods, using N points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ in an M-dimensional space $\mathcal{X}$ to produce an approximation to, e.g., (1), have an absolute error that decreases no faster than $N^{-1/M}$, whereas the absolute error of the Monte Carlo method decreases as $N^{-1/2}$, that is, independent of the dimension of the space (see, e.g., Fishman (1996)).

The history of Monte Carlo methods is long, but their application to the solution of scientific problems begins with von Neumann, Ulam and Fermi who used a Monte Carlo method in nuclear reaction studies. The name 'the Monte Carlo method' was first used by Metropolis and Ulam (1949), and 4 years later Metropolis *et al* (1953) published their Markov chain-based algorithm for (asymptotic) sampling of Gibbs–Boltzmann distributions in high-dimensional spaces. This algorithm, now known as the Metropolis algorithm, was in fact the first major scientific algorithm to be run on a digital computer. It was a biased random walk whose individual steps (iterations) were based on very simple probabilistic rules. One important feature of the Metropolis algorithm was that full information on the distribution $p(\boldsymbol{x})$ to be sampled was unnecessary: as long as ratios $p(\boldsymbol{x}_i)/p(\boldsymbol{x}_j)$ between the value of p at any two selected points, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, could be calculated upon request (by some numerical procedure), the algorithm worked.

Monte Carlo methods are becoming increasingly important for the solution of nonlinear inverse problems in two different, but related, situations. In the first situation we need a near-optimal solution (measured in terms of data fit and adherence to given constraints) to the problem. In the second situation, the inverse problem is formulated as a search for solutions fitting the data within a certain tolerance, given by data uncertainties. In a non-probabilistic setting this means that we search for solutions with calculated data whose distance from the observed is less than a fixed, positive number. In a Bayesian context, the tolerance is 'soft': a large number of samples of statistically near-independent models from the a posterior probability distribution are sought. Such solutions are consistent with data and prior information, as they fit the data 'within error bars', and adhere to 'soft' *prior* constraints given by a prior probability distribution.

Early examples of the solution of inverse problems by means of Monte Carlo methods are abundant in geophysics and other disciplines of applied physics. Since Keilis-Borok and

Yanovskaya (1967) and (Press 1968, 1970a) made the first attempts at randomly exploring the space of possible Earth models consistent with seismological data, there has been considerable advances in computer technology, and therefore an increasing interest in these methods. Geman and Geman (1984) applied simulated annealing to Bayesian image restoration, and derived an expression for the posterior distribution from the prior distribution, a model of the image blurring mechanism, and a Gaussian noise model. Through an identification of the posterior distribution with a Gibbs–Boltzmann distribution, they estimated a maximum *a posteriori* model, using a simulated annealing algorithm. In their paper, they suggested using the Metropolis algorithm, not only for maximum *a posteriori* estimation, but also to sample the posterior distribution. This idea was taken up by Rothman (1985, 1986) who was the first to employ *importance sampling* to solve a strongly nonlinear (that is, a multi-modal) optimization problem arising in seismic reflection surveys. Later, Cary and Chapman (1988), Landa *et al* (1989), Mosegaard and Vestergaard (1991), and Koren *et al* (1991) applied Monte Carlo methods to seismic waveform fitting.

Cary and Chapman (1988) used the Monte Carlo method to analyse the seismic waveform inversion problem in a Bayesian formulation. Waveforms and travel times from source to receiver were used as data, and the model parameters defined a horizontally stratified Earth with depth as a function of the seismic wave propagation velocity. They improved efficiency of the sampling through a method described by Wiggins (1969, 1972) in which the model space was sampled according to a prior distribution $\rho(\boldsymbol{m})$.

Marroquin *et al* (1987) adopted an approach similar to that of Geman and Geman. However, they used the Metropolis algorithm to generate samples from the posterior distribution, from which they computed model estimates.

Recent examples of using Bayes theorem and the Metropolis algorithm for calculating approximate *a posteriori* probabilities for an inverse problem are given by Pedersen and Knudsen (1990), Koren *et al* (1991), Gouveia and Scales (1998), Dahl-Jensen *et al* (1998), Khan *et al* (2000), Rygaard-Hjalsted *et al* (2000), and Khan and Mosegaard (2001).

Not all Monte Carlo inversion adopted the Bayesian viewpoint. The initial introduction of Monte Carlo techniques into geophysics by Keilis-Borok and Yanovskaya (1967) and Press (1968, 1970a) was concerned with only uniform sampling of a parameter space, without taking a Bayesian approach. These papers generated considerable interest, and also a debate over how to interpret the resulting ensemble of Earth models, especially when extra constraints were imposed, Haddon and Bullen (1969), Anderssen (1970), Anderssen and Senata (1971, 1972), Anderssen *et al* (1972), Wiggins (1972), Kennett and Nolet (1978). In addition, uniform random search techniques have been used for model optimization in geophysical inversion (again without invoking Bayesian principles). Notable examples include: estimation of seismic attenuation structure in the Earth, Burton and Kennett (1974a, 1974b), Burton (1977); seismic surface wave attenuation studies, Biswas and Knopoff (1974), Mills and Fitch (1977); estimation of seismic and density structure, Worthington *et al* (1972, 1974), Goncz and Cleary (1976), Kennett (1998); magnetotelluric studies, Jones and Hutton (1979); estimation of mantle viscosity, Ricard *et al* (1989); and plate rotation vectors, Jestin *et al* (1994).

Most of the applications of genetic algorithms (Holland 1975) within geophysical inverse problems are also non-Bayesian. That is, they are Monte Carlo techniques that generate an ensemble of samples from a parameter space, which must then be made use of in some way. The introduction of genetic algorithms into geophysics occurred in the early 1990s (Stoffa and Sen 1991, Gallagher *et al* 1991, Wilson and Vasudevan 1991, Sambridge and Drijkoningen 1992, Scales *et al* 1992, Sen and Stoffa 1992, Smith *et al* 1992). They have since been applied to a wide range of geophysical problems. (Many references on geophysical applications can be found in Gallagher and Sambridge (1994) and Sambridge and Mosegaard (2001).)

More recently a new class of ensemble-based Monte Carlo search technique known as a neighbourhood algorithm (Sambridge 1999a) has been developed for geophysical inverse problems. This approach follows a non-Bayesian approach for sampling of a parameter space, but can be used within a Bayesian formulation for analysing the resulting ensemble (Sambridge 1999b).

Monte Carlo methods are essentially practical tools for dealing with (usually complicated) probability distributions, and this is the main reason for their usefulness in inverse problem analysis. Solutions $\boldsymbol{m}$ to inverse problems can usually be described by a function $f(\boldsymbol{m})$ over the model space, measuring the model's ability to fit the data and/or given *a priori* constraints. In this paper, we will call $f(\boldsymbol{m})$ the *fitness function*, a term borrowed from the theory of genetic algorithms.

In Bayesian inversion the fitness function is the so-called posterior probability distribution given by

$$f(\boldsymbol{m}) = C_f L(\boldsymbol{m})\rho(\boldsymbol{m}) \tag{4}$$

where C_f is a normalization constant, $L(\boldsymbol{m})$ is a likelihood function and $\rho(\boldsymbol{m})$ a prior probability distribution. The likelihood function is usually of the form

$$L(\boldsymbol{m}) = C_L \exp(-S(\boldsymbol{m})) \tag{5}$$

where C_L is a constant, and $S(\boldsymbol{m})$ is a *misfit function*, measuring the deviation of the observed data from the data calculated from $\boldsymbol{m}$. The prior distribution assigns a data-independent weight to $\boldsymbol{m}$ depending on how acceptable it is according to other available information.

In a non-probabilistic context, the function may be an indicator function describing which models are acceptable according to data, and which are not. For instance, for a given positive constant S we may define

$$f(\boldsymbol{m}) = \begin{cases} 1 & \text{if } S(\boldsymbol{m}) \leqslant S \\ 0 & \text{if } S(\boldsymbol{m}) > S. \end{cases} \tag{6}$$

As mentioned in the introduction, another way of using Monte Carlo methods for analysis of inverse problems is for *model construction*, that is, finding a set of model parameters that, in some sense, gives a (near-) optimal fit to data and prior information. Model construction is an optimization problem, which in a Bayesian context can be formulated as a search for the model(s) where $f(\boldsymbol{m})$ attains its maximum. In a non-probabilistic context, the problem can be formulated as a search for the minimum of the misfit function $S(\boldsymbol{m})$.

Theoretically, the model construction problem can be viewed as a special case of the more general sampling problem: sampling, e.g., a likelihood function (5), where we have artificially decreased the standard deviation of the data uncertainties, corresponds to minimizing the misfit $S(\boldsymbol{m})$. We shall therefore begin the next section with a review of the essential Markov chain sampling theory. Later we will specialize this to cover the model construction problem. The heuristic methods, simulated annealing and genetic algorithms, designed to improve the speed of model construction will then be covered. This order of presentation is not in accordance with the historic development, but it reveals some of the fundamental strengths and weaknesses of the Monte Carlo methods, and hopefully points in directions where future research efforts should be made.

2. Monte Carlo methodology

Monte Carlo methods are natural when solving inverse problems within a purely probabilistic framework. The reason for this is the following: the fundamental building blocks of the

theory are probability densities, and these objects can be viewed mathematically as limits of histograms (normalized, and with vanishing column widths) generated by a random process, e.g., a Monte Carlo algorithm. For this reason, any probability distribution can be represented by a group of Monte Carlo algorithms, namely the Monte Carlo algorithms that sample it. On the other hand, to every Monte Carlo algorithm there is a probability distribution, namely the one that it samples. Most conceivable operations on probability densities (e.g., computing marginals and conditionals, integration, combining independent information, etc) have their counterparts in operations on/by their corresponding Monte Carlo algorithms. In this way, Monte Carlo algorithms provide a way of manipulating probability densities—even densities that cannot be expressed mathematically in closed form.

Our first task will be to describe which distribution is sampled by a given Monte Carlo algorithm, and how a Monte Carlo algorithm can be made to sample a prescribed probability distribution.

2.1. Sampling known one-dimensional probability distributions

If random rules have been defined to select points such that the probability of selecting a point in the volume element $\mathrm{d}x_1 \ldots \mathrm{d}x_N$ is $p(\boldsymbol{x})\mathrm{d}x_1 \ldots \mathrm{d}x_N$, then the points selected in that way are called *samples* of the distribution $p(\boldsymbol{x})$. Depending on the rules defined, successive samples $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k, \ldots$, may be dependent or independent.

Before going into more complex sampling situations, let us briefly review a few methods for sampling a probability distribution that is 'completely known' in the sense that it can be described by an explicit mathematical expression. Two important methods are given below (formulated for a probability distribution over a one-dimensional space).

Method 1. Let p be an everywhere nonzero probability distribution with distribution function P, given by

$$P(s) = \int_{-\infty}^{x} p(s)\,\mathrm{d}s, \tag{7}$$

and let r be a random number chosen uniformly at random between 0 and 1. Then the random number x generated through the formulae

$$x = P^{-1}(r) \tag{8}$$

has probability distribution p.

More special, yet useful, is the following way of generating Gaussian random numbers.

Method 2. Let r_1 and r_2 be random numbers chosen uniformly at random between 0 and 1. Then the random numbers x_1 and x_1 generated through the formulae

$$x_1 = \sqrt{-2\ln r_2}\cos(2\pi r_1) \tag{9}$$

$$x_2 = \sqrt{-2\ln r_2}\sin(2\pi r_1) \tag{10}$$

are independent and Gaussian distributed with zero mean and unit variance.

These theorems are easily demonstrated, and straightforward to use in practice. They can be extended to higher dimensions.

2.2. Sampling 'unknown' probability distributions

For most practical inverse problems, the model space is so vast, and evaluation of the fitness function $f(\boldsymbol{x})$ is so computer intensive, that only algorithms which evaluate $f(\boldsymbol{x})$ once (or only a few times) in each iteration, are useful. In such cases we will say that, for the algorithm, values of $f(\boldsymbol{x})$ 'are only available on request'.

If $p(\boldsymbol{x})$ is a multi-dimensional probability distribution over $\mathcal{X}$ whose values are only available on request, it is still straightforward to design a random walk that samples it. Imagine, for instance, that we have a number M satisfying

$$M \geqslant \max(p(\boldsymbol{x})). \tag{11}$$

Then, the following algorithm will sample $p(\boldsymbol{x})$.

Algorithm 1 (the primitive algorithm). *In the nth step of the algorithm, choose a candidate sampling point $\boldsymbol{x}_{cand,n}$ randomly (using a uniform distribution over the sampling space) but accept it only with probability*

$$p_{accept} = \frac{p(\boldsymbol{x}_{cand,n})}{M}. \tag{12}$$

The set of thus accepted candidate points are samples from the probability distribution $p(\boldsymbol{x})$.

However, this algorithm would not be useful in practice, because in many problems $p(\boldsymbol{x})$ could have narrow maxima, (which contribute significantly to integrals over $p(\boldsymbol{x})$), but are sampled rarely (or not at all) when only a limited number of samples can be taken. One of the things that slows down the primitive algorithm is that the ratio

$$\frac{p(\boldsymbol{x})}{M} \tag{13}$$

is very small almost everywhere in $\mathcal{X}$, especially when $p(\boldsymbol{x})$ has large values only in a small part of the space $\mathcal{X}$, and hence the likelihood of acceptance is very small.

The *Metropolis algorithm*, an example of an *importance sampling algorithm*, works around this problem by comparing $p(\boldsymbol{x}_{cand,n})$, not with a large number M, but with the smaller number $p(\boldsymbol{x}_{current,n})$, where $\boldsymbol{x}_{current,n}$ is the current point being visited. The probability of accepting the candidate point $\boldsymbol{x}_{cand,n}$ in the Metropolis algorithm is

$$p_{accept} = \begin{cases} \dfrac{p(\boldsymbol{x}_{cand,n})}{p(\boldsymbol{x}_{current,n})} & \text{when } p(\boldsymbol{x}_{cand,n}) \leqslant p(\boldsymbol{x}_{current,n}) \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

Furthermore, in this algorithm $\boldsymbol{x}_{cand,n}$ is often chosen from a relatively small neighbourhood around $\boldsymbol{x}_{current,n}$ so as to further reduce the chance that the ratio $p(\boldsymbol{x}_{cand,n})/p(\boldsymbol{x}_{current,n})$ is small.

Importance sampling allows us to sample the space with a sampling density proportional to the given probability density, without excessive (and useless) sampling of low-probability areas of the space. This is not only important, but in fact vital in high-dimensional spaces, where a very large proportion (approaching 100%) of the volume may have near-zero probability density. Figure 1 illustrates the superiority of the Metropolis algorithm over the primitive algorithm, even in a one-dimensional example.

3. Sampling through random walks

In the following, we shall describe the essential properties of random walks performing *importance sampling*.

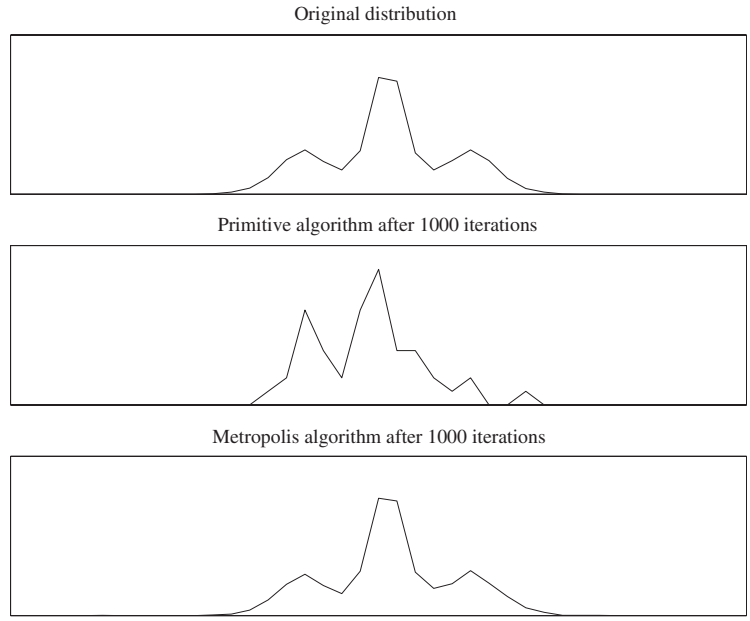


Figure 1. A probability distribution (top) were sampled 1000 times by the primitive algorithm (middle) and the Metropolis algorithm (bottom). The superiority of the Metropolis algorithm over the primitive algorithm is seen even in this one-dimensional example. In the Metropolis case all 1000 iterations yielded a sample contributing to the final result, whereas for the primitive algorithm, only 42 samples were accepted for histogram building.

A random walk is a Markov chain. This means that the probability of moving from a point $\boldsymbol{x}_i$ to a point $\boldsymbol{x}_j$ in the space $\mathcal{X}$ in a given step (iteration) is independent of the path travelled by the 'walker' in the past. Let us define the conditional probability distribution $P_{ij}(\boldsymbol{x}_i|\boldsymbol{x}_j)$ of visiting $\boldsymbol{x}_i$, given that the walker currently is at point $\boldsymbol{x}_j$. We will call $P_{ij}(\boldsymbol{x}_i|\boldsymbol{x}_j)$ the *transition probability distribution*, and for simplicity in the following we will drop the double subscript and write $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$. P is a probability distribution with respect to its first argument, so we have

$$\int_{\mathcal{X}} P(\boldsymbol{x}_i|\boldsymbol{x}_j)\,\mathrm{d}\boldsymbol{x}_i = 1. \tag{15}$$

As a special case, P may be a discrete distribution, in which case we imagine the set of points $\boldsymbol{x}_j$ $(i = 1, \ldots, J)$ to be a discrete, finite subset of $\mathcal{X}$. The random walk we are about to construct here will then only be allowed to visit points in the considered grid, and $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$ will be a discrete distribution over the grid points.

Given a random walk defined by the transition probability distribution $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$. Assume now that we have a distribution $s_n(\boldsymbol{x})$ describing the position of the random walker after n steps. Each step of the random walk will modify this distribution, and if $s_n(\boldsymbol{x}) \to p(\boldsymbol{x})$ for $n \to \infty$ we say that $p(\boldsymbol{x})$ is an *equilibrium (or stable) probability distribution* for the random walk. That $p(\boldsymbol{x})$ is an equilibrium distribution means that if it is the distribution of the position of the random walker at one time, it remains the distribution of its position after one more step (and hence, forever). Technically, this can be expressed by the fact that $p(\boldsymbol{x})$ is an eigenfunction with eigenvalue 1 of the linear integral operator with kernel $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$:

$$\int_{\mathcal{M}} P(\boldsymbol{x}_i|\boldsymbol{x}_j)\,p(\boldsymbol{x}_j)\,\mathrm{d}\boldsymbol{x}_j = p(\boldsymbol{x}_i). \tag{16}$$

If the random walk, for any initial distribution $s_0(\boldsymbol{x})$, equilibrates to the same distribution $p(\boldsymbol{x})$, we say that $p(\boldsymbol{x})$ is *the* (unique) equilibrium distribution for $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$.

The next section describes the basic idea behind importance sampling by the Metropolis algorithm, which is designed to have any chosen function $p(\boldsymbol{x})$ as its unique equilibrium distribution.

3.1. Design of random walks with given equilibrium distributions

A random walk with a given equilibrium distribution $p(\boldsymbol{x})$ must satisfy the condition that once the sampling distribution $s(\boldsymbol{x})$ is equal to $p(\boldsymbol{x})$ it must remain equal to $p(\boldsymbol{x})$. This equilibrium can also be expressed through the following condition.

Condition 1 (microscopic reversibility). *The probability, at any time, that the random walker enters an infinitesimal neighbourhood $\mathcal{N}_j$, surrounding the point $\boldsymbol{x}_j$, equals the probability that it leaves $\mathcal{N}_j$.*

There are infinitely many ways of satisfying the above requirement. As we shall see in a later section, the way a genetic algorithm establishes microscopic reversibility (equilibrating to its unknown stable distribution) is rather complex. In contrast, the Metropolis algorithm and the Gibbs sampler rely on a very simple principle, summarized in the following condition.

Condition 2 (detailed balance). *For any pair of points $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$, the probability, at any time, that the random walker jumps from the infinitesimal neighbourhood $\mathcal{N}_j$, surrounding $\boldsymbol{x}_j$, to the infinitesimal neighbourhood $\mathcal{N}_i$ (of the same volume), surrounding $\boldsymbol{x}_i$, equals the probability that it jumps from $\mathcal{N}_i$ to $\mathcal{N}_j$.*

From this last principle, an algorithm that has a given distribution $p(\boldsymbol{x})$ as an equilibrium distribution can easily be derived.

Consider two points in $\mathcal{X}$, say $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$. Detailed balance means that the transition probability distribution satisfies the following symmetry condition:

$$P(\boldsymbol{x}_i|\boldsymbol{x}_j)\,p(\boldsymbol{x}_j)\,\mathrm{d}\boldsymbol{x}_j\,\mathrm{d}\boldsymbol{x}_i = P(\boldsymbol{x}_j|\boldsymbol{x}_i)\,p(\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x}_i\,\mathrm{d}\boldsymbol{x}_j \tag{17}$$

where $p(\boldsymbol{x})$ is the desired equilibrium distribution, which we will, without loss of generality, assume to be nonzero everywhere in $\mathcal{X}$.

Equation (17) means that $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$ must satisfy

$$\frac{P(\boldsymbol{x}_i|\boldsymbol{x}_j)}{P(\boldsymbol{x}_j|\boldsymbol{x}_i)} = \frac{p(\boldsymbol{x}_i)}{p(\boldsymbol{x}_j)}. \tag{18}$$

There are, of course, infinitely many solutions $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$ to the above equation, but we shall try to design an efficient algorithm by keeping the transition probabilities everywhere as large as possible between pairs of points.

The first question is now: how can we implement an algorithm with transition probability densities satisfying (18)? The answer to this question is interesting, as it reveals two important, and conflicting, characteristics of random-walk-based algorithms.

In the so-called Metropolis algorithm, (18) is implemented through two randomized operations which, together, form an iteration:

(i) The exploration step. The first operation consists of proposing a 'candidate' point $\boldsymbol{x}_i$ using a so-called *proposal distribution* $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$, where $\boldsymbol{x}_j$ is the currently visited point. The proposal distribution is symmetric:

$$U(\boldsymbol{x}_i|\boldsymbol{x}_j) = U(\boldsymbol{x}_j|\boldsymbol{x}_i), \tag{19}$$

but otherwise it is arbitrary, in the sense that its form is chosen before running the algorithm, and is, in principle, independent of any knowledge on $p(\boldsymbol{x})$. $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$ embodies the 'strategy' by which the algorithm explores $\mathcal{X}$, when searching for new samples from the distribution $p(\boldsymbol{x})$.

(ii) The exploitation step. The second operation is to decide if the candidate point should be accepted as the next sample. Any acceptance probability of the form

$$p_{accept} = \psi(\boldsymbol{x}_i, \boldsymbol{x}_j)/p(\boldsymbol{x}_j), \tag{20}$$

where $\psi(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a symmetric function, can be used. The simplest acceptance probability (and the one giving the highest transition probabilities for given $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$) is obtained by putting $\psi(\boldsymbol{x}_i, \boldsymbol{x}_j) = \min(p(\boldsymbol{x}_i), p(\boldsymbol{x}_j))$. This gives the traditional Metropolis acceptance probability

$$p_{accept} = \begin{cases} \dfrac{p(\boldsymbol{x}_i)}{p(\boldsymbol{x}_j)} & \text{for } p(\boldsymbol{x}_i) \leqslant p(\boldsymbol{x}_j) \\ 1 & \text{otherwise.} \end{cases} \tag{21}$$

If the candidate point is rejected, the current point is repeated (thus counting one more time). The acceptance probability describes the 'greediness' of the algorithm. The smaller p_{accept} is for $p(\boldsymbol{x}_i) \leqslant p(\boldsymbol{x}_j)$, the more 'greedy' the algorithm is.

The above procedure means that

$$P(\boldsymbol{x}_i|\boldsymbol{x}_j) = U(\boldsymbol{x}_i|\boldsymbol{x}_j)\, p_{accept} \tag{22}$$

giving

$$\frac{P(\boldsymbol{x}_i|\boldsymbol{x}_j)}{P(\boldsymbol{x}_j|\boldsymbol{x}_i)} = \frac{U(\boldsymbol{x}_i|\boldsymbol{x}_j)}{U(\boldsymbol{x}_j|\boldsymbol{x}_i)} \frac{\psi(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\psi(\boldsymbol{x}_j, \boldsymbol{x}_i)} \frac{p(\boldsymbol{x}_i)}{p(\boldsymbol{x}_j)}. \tag{23}$$

Due to the imposed symmetry (19) on $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$ and $\psi(\boldsymbol{x}_i, \boldsymbol{x}_j)$, equation (23) is the detailed balance condition (18).

It can be shown (see, e.g., Kaipio *et al* 2000) that if our transition probability distribution $P(\boldsymbol{x}_i|\boldsymbol{x}_j)$, satisfies two particular conditions (in addition to microscopic reversibility), then $p(\boldsymbol{x})$ will be the only equilibrium distribution for the algorithm, and so it will converge towards $p(\boldsymbol{x})$ no matter what the starting distribution is. The two conditions are:

(i) Aperiodicity. The probability that an iteration of the algorithm results in the trivial move $\boldsymbol{x}_j \longrightarrow \boldsymbol{x}_j$ is non-zero. That this is clearly satisfied, can easily be seen from the form (21) of the acceptance probability.

(ii) Irreducibility. It is possible to go from any point $\boldsymbol{x}_j$ to any other point $\boldsymbol{x}_i$ in $\mathcal{X}$, given a sufficient number of iterations. It is up to the designer of the proposal distribution to make sure that this requirement is satisfied.

The majority of Monte Carlo methods used in practice satisfy the above two conditions, and therefore converge toward (equilibrate at) a unique distribution. In this paper we describe three algorithms belonging to this category: the *Metropolis algorithm* (described above), the *Gibbs sampler* and the *genetic algorithm*. The first two algorithms equilibrate at *known* distributions (and they are, in fact, designed to equilibrate at these distributions), whereas the equilibrium distribution of the genetic algorithm, defined in terms of the parameters of the algorithm, has not yet been found.

Convergence issues. The second important question is now: How do we maximize the efficiency of our algorithm? Unfortunately, a definitive answer to this question has not yet been found, although it is of utmost importance for the practical applicability of the Metropolis algorithm. The proposal distribution $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$ embodies the complete search strategy of the algorithm, so the problem of efficiency can, e.g., be formulated in the following way.

Given a measure of distance $\mathrm{dist}(p_1, p_2)$ in the space of probability densities over $\mathcal{X}$, and a (usually small) positive number M. Find an irreducible proposal distribution $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$ that minimizes the expected number of iterations needed to obtain

$$\mathrm{dist}(s, p) \leqslant M, \tag{24}$$

where $s(\boldsymbol{x})$ is the sampling distribution and $p(\boldsymbol{x})$ is the equilibrium distribution. A possible distance measure can, e.g., be

$$\mathrm{dist}(p_1, p_2) = \max_{\boldsymbol{x}\in\mathcal{X}} |p_1(\boldsymbol{x}) - p_2(\boldsymbol{x})|. \tag{25}$$

In the special case where $\mathcal{X}$ is a discrete space with a relatively small number of points $\boldsymbol{x}_j$, and the algorithm can be described by a completely known transition probability matrix $\boldsymbol{P} = \{P(\boldsymbol{x}_i|\boldsymbol{x}_j)\}$, the convergence speed may be estimated through an eigenvalue analysis of P_{ij}. However, this situation virtually never occurs in practice. Knowing $\boldsymbol{P}$ completely would require knowing $p(\boldsymbol{x}_j)$ for all j, and in that case the inverse problem is already solved!

Andresen *et al* (1988) showed that an approximate eigenvalue analysis can be made by lumping points of similar values of $p(\boldsymbol{x})$ into a small number of 'states' between which transition probabilities can be estimated empirically. For a given choice of $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$ an initial run of the algorithm is used to monitor the frequency of transitions between the lumped states. The matrix of (normalized) transition frequencies is then used as an approximation to the transition probability matrix for transitions between lumped states. Finally, the second largest eigenvalue of this transition probability matrix is used to estimate the *relaxation time* for the algorithm (the time taken by the algorithm to reduce the distance between its sampling distribution $s(\boldsymbol{x})$ and its equilibrium distribution $p(\boldsymbol{x})$ by a factor $1/e$). The relaxation time is related to the convergence time we are looking for, although the authors do not explicitly explore this relationship.

Lacking theoretical guidance, it has become common practice to tune $U(\boldsymbol{x}_i|\boldsymbol{x}_j)$ empirically (Hastings 1970). An acceptable proposal distribution must keep the so-called *burn-in period* for the algorithm at a minimum. The burn-in period is the time it takes for the algorithm, from its initial state, before it reaches a point where its outputs (probability densities of sampled models, parameters of sampled models, frequency of accepted models, etc) are approximately stationary over the considered number of iterations. Experience has shown that a frequency of accepted models (after the burn-in period) of 25–50% indicates that the algorithm is performing well (Gelman *et al* 1996).

Multistep iterations. Often, it is convenient to split up an iteration into a number of substeps, having their own transition probability densities. A typical example is a random walk in an N-dimensional space where we are interested in dividing an iteration of the random walk into N substeps, where the nth move of the random walker is in the direction parallel to the nth axis.

The question is now: if we want to form an iteration consisting of a series of substeps, can we give a sufficient condition to be satisfied by each substep such that the complete iteration has the desired convergence properties?

It is easy to see that if the individual substeps in an iteration all have the same distribution $p(\boldsymbol{x})$ as their equilibrium distribution (not necessarily unique), then the complete iteration also has $p(\boldsymbol{x})$ as an equilibrium distribution. This follows from the fact that the equilibrium

distribution is an eigenfunction with eigenvalue 1 for the integral operators corresponding to each of the substep transition probability distributions. Then it is also an eigenfunction with eigenvalue 1, and hence an equilibrium distribution, for the integral operator corresponding to the transition probability distribution for the complete iteration.

If this transition probability distribution is to be the unique equilibrium distribution for the complete iteration, then the random walk must be irreducible. That is, it must be possible to go from any point to any other point by performing iterations consisting of the specified substeps.

If the substeps of an iteration satisfy these sufficient conditions, there is also another way of defining an iteration with the desired, unique equilibrium distribution. Instead of performing an iteration as a series of substeps, it is possible to define the iteration as consisting of one of the substeps, chosen randomly (with any distribution having nonzero probabilities) among the possible substeps. In this case, the transition probability distribution for the iteration is equal to a linear combination of the transition probability densities for the individual substeps. The coefficient of the transition probability distribution for a given substep is the probability that this substep is selected. Since the desired distribution is an equilibrium distribution (eigenfunction with eigenvalue 1) for the integral operators corresponding to each of the substep transition probability distribution, and since the sum of all the coefficients in the linear combination is equal to 1, it is also an equilibrium distribution for the integral operator corresponding to the transition probability distribution for the complete iteration. This equilibrium distribution is unique, since it is possible, following the given substeps, to go from any point to any other point in the space.

Of course, a substep of an iteration can, in the same way, be built from sub-substeps, and in this way acquire the same (not necessarily unique) equilibrium distribution as the sub-substeps.

The Gibbs sampler. In the Gibbs sampler, one iteration consists of a number of substeps, each having its own transition probability distribution. In a typical implementation of the Gibbs sampler, operating in a K-dimensional model space, each iteration consists of K substeps, one for each parameter. The kth substep perturbs only the kth parameter, and it has its own transition probability distribution $P_k(\boldsymbol{x}_i|\boldsymbol{x}_j)$. The kth substep runs as follows:

(i) The proposal distribution is defined as

$$U_k(\boldsymbol{x}_i|\boldsymbol{x}_j) = \begin{cases} \dfrac{p(\boldsymbol{x}_i)}{\sum_{\boldsymbol{x}_k \in \mathcal{N}_j^k} p(\boldsymbol{x}_k)} & \boldsymbol{x}_i \in \mathcal{N}_j^k \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

where $\mathcal{N}_j^k$ is the set of points deviating from $\boldsymbol{x}_j$ only in the kth parameter.

(ii) The acceptance probability is

$$p_{accept} = 1 \tag{27}$$

for all the proposed candidate models.

In each substep of the Gibbs sampler, no more than one parameter is perturbed (or is possibly left unchanged), so after completion of one iteration (consisting of all K substeps), all parameters have been perturbed.

That this algorithm samples p can be seen in the following way: the transition probability distribution for each substep is given by

$$P_k(\boldsymbol{x}_i|\boldsymbol{x}_j) = U_k(\boldsymbol{x}_i|\boldsymbol{x}_j), \tag{28}$$

which satisfies detailed balance equation (18), since if $\boldsymbol{x}_i \in \mathcal{N}_j^k$, then

$$P_k(\boldsymbol{x}_i|\boldsymbol{x}_j)\,p(\boldsymbol{x}_j) = U_k(\boldsymbol{x}_i|\boldsymbol{x}_j)\,p(\boldsymbol{x}_j) \tag{29}$$

$$= \frac{p(\boldsymbol{x}_i)}{\sum_{\boldsymbol{x}_k \in \mathcal{N}_j^k} p(\boldsymbol{x}_k)}\, p(\boldsymbol{x}_j) \tag{30}$$

$$= \frac{p(\boldsymbol{x}_j)}{\sum_{\boldsymbol{x}_k \in \mathcal{N}_i^k} p(\boldsymbol{x}_k)}\, p(\boldsymbol{x}_i) \tag{31}$$

$$= U_k(\boldsymbol{x}_j|\boldsymbol{x}_i)\,p(\boldsymbol{x}_i) \tag{32}$$

$$= P_k(\boldsymbol{x}_j|\boldsymbol{x}_i)\,p(\boldsymbol{x}_i) \tag{33}$$

where we have used that $\mathcal{N}_j^k = \mathcal{N}_i^k$. Since each substep of an iteration satisfies microscopic reversibility, so does the entire iteration, and the algorithm samples the target distribution p asymptotically.

The advantage of the Gibbs sampler is that the acceptance probability is always 1, so there are no rejected moves. The disadvantage lies in the construction of the proposal distribution. In contrast to the Metropolis algorithm, $U_k(\boldsymbol{x}_i|\boldsymbol{x}_j)$ is here constructed directly from the desired equilibrium distribution by evaluating $p(\boldsymbol{x})$ 'along a line' in $\mathcal{X}$. This is feasible when calculation of $p(\boldsymbol{x})$ is computationally inexpensive, but this is not the case in many practical inverse problems. When evaluation of $p(\boldsymbol{x})$ is computer intensive, the Metropolis algorithm may be a better choice. An example of the use of a Gibbs sampler in a case where p can be efficiently evaluated for all perturbations of a single model parameter can be found in Rothman (1986).

3.2. Bayesian inference

We mentioned in the introduction that in a Bayesian formulation, the fitness function is the so-called posterior probability distribution over the model space, given by

$$f(\boldsymbol{m}) = C_f L(\boldsymbol{m})\rho(\boldsymbol{m}).$$

This distribution carries all information available on models originating from the data, and from data-independent prior information.

If a Monte Carlo algorithm (typically the Metropolis algorithm or a Gibbs sampler) is used to generate a large number of samples $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$ from $f(\boldsymbol{m})$, we can use these samples to estimate averages over the model space. Any average of a function $h(\boldsymbol{m})$ over the model space $\mathcal{M}$ (e.g., an expectation or a covariance) can be approximated by (3). The mean $\langle m_i \rangle$ of the ith model parameter m_i can be estimated by putting $h(\boldsymbol{m}) = m_i$, and the posterior covariance between the ith and jth model parameters is approximated by putting $h(\boldsymbol{m}) = (m_i - \langle m_i \rangle)(m_j - \langle m_j \rangle)$. If we wish to calculate the probability of an event $\mathcal{E}$ in $\mathcal{M}$, containing all models in model space with a given feature, it is done by putting $h(\boldsymbol{m})$ equal to the indicator function

$$h(\boldsymbol{m}) = \begin{cases} 1 & \text{if } \boldsymbol{m} \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

As an example, $\mathcal{E}$ may be all sampled models of the Earth containing a sharp boundary (appropriately defined) in a certain depth interval. Finally, it should be mentioned that samples from the one-dimensional marginal $f(m_k)$ are obtained simply by collecting values of m_k from samples $\boldsymbol{m} = (m_1, \ldots, m_k, \ldots, m_M)$ of $f(\boldsymbol{m})$.

The above procedure is general and simple, but the practical problem is often to *discover* model features that have a high probability. In classical terminology, these features are *well*

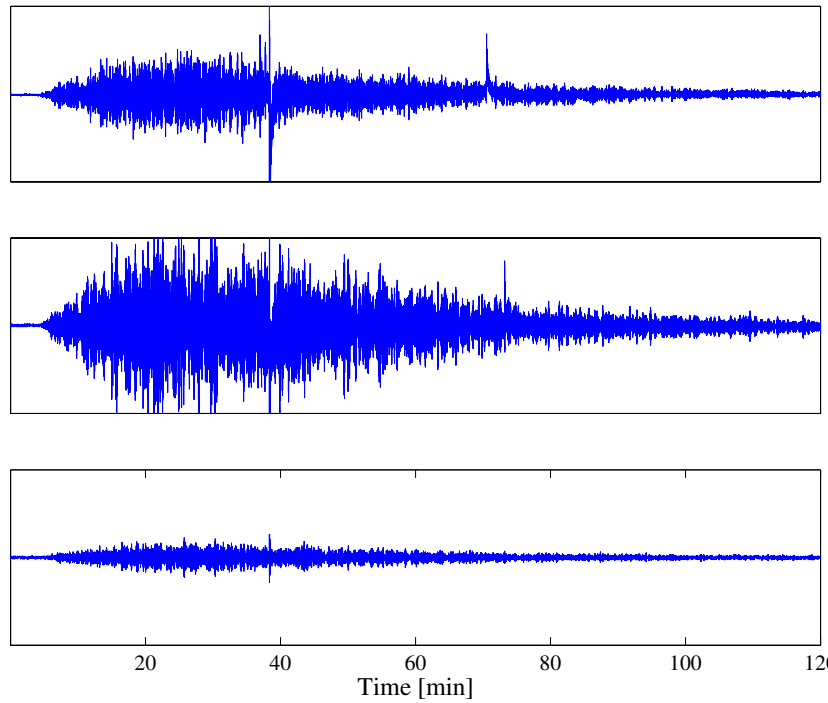


Figure 2. This figure shows a Lunar three-component seismic record from a meteoroid impact recorded at station 16 on day 319, 1976. The seismogram at the top is the N–S component (S positive), the middle is E–W (W positive) and the seismogram at the bottom is the vertical component (up positive). The seismograms commence at 23 h 16 min 50 s.

resolved. Simply looking at the Monte Carlo-generated samples from $f(\boldsymbol{m})$ is often the most efficient way to discover such structure. Well-resolved structure is seen as a structure that occurs frequently in the output, and the simplest way to discover it, is to plot all output models side by side, or on top of each other. Another way is to display all output models sequentially as pictures in a movie, preferably with different degrees of smoothing (Mosegaard and Tarantola 1995).

3.3. Example: inversion of seismic travel-time data from the Moon

Khan *et al* (2000) used the Metropolis algorithm to reanalyse the Lunar seismic data from the Apollo project, 1969–1977. The problem was to infer Lunar P- and S-wave velocity structure from observed arrival times of seismic disturbances, generated by moonquakes, meteorite impacts and artificial impacts (spacecraft modules hitting the Moon surface). Figure 2 shows a sample Lunar three-component seismic record from a deep moonquake, recorded on February 2, 1970. First arrival times for the P- and S-waves were read from 177 seismograms, and a spherically symmetric P- and S-wave velocity model, as well as source locations and epicentre times (time of seismic energy release), were sought to explain this dataset. The P- and S-wave velocity models were assumed to be piecewise linear functions of radius, with 56 break points each (the Lunar radius is 1738 km). Depths and velocities at these break points were used as the unknown model parameters of the problem, together with the source locations and epicentre times for 80 seismic events. All in all, the total number of unknown parameters for this problem was 450.

The *a priori* probability density was defined as follows: each velocity parameter was allowed to vary in the interval between 1 and 50 km s^{-1}, its logarithm assigning a uniform distribution between these values. Ray theory requires a certain smoothness of the model, which was realized by assuming a minimum layer thickness (distance between two consecutive break points) of 5 km. Source coordinates were unconstrained.

This problem has the typical 'pathology' that no analytical relation between data and model parameters exists. Only a rather computer intensive algorithm, tracing rays through any given P- and S-velocity model, is available, so Monte Carlo analysis is the most suitable method for this problem.

The Metropolis algorithm used in this case updated one velocity (or source-) parameter at a time. The likelihood function was given by (5), where

$$S(\boldsymbol{m}) = \sum_n \frac{|d_{obs}^{(n)} - d^{(n)}(\boldsymbol{m})|}{\sigma_n}$$

$d_{obs}^{(n)}$ denoting the observed data (travel times), $d^{(n)}(\boldsymbol{m})$ synthetic data, computed (by ray tracing) from the model $\boldsymbol{m}$, and σ_n is the uncertainty (standard deviation) of the nth datum. The uncertainty was 1 s for artificial impacts, between 4 and 7 s for deep moonquakes, and between 4 and 26 s for shallow moonquakes and meteorite impacts.

The proposal distribution was defined as follows: in each iteration, a new model parameter was chosen at random. It was then perturbed using a uniform distribution centred at the current value, and having a half-width of 1.1 km s^{-1} for velocities, 2 km for the upper 11 layer boundaries, 8 km for the lower 45 layer boundaries, 1° for source longitudes/latitudes and 4 km for source depths. Layer boundary perturbations were, however, modified so as to maintain the ordering of boundaries. This was done by appropriate reduction of the half-width of the distribution of proposed boundary depths.

The chosen proposal distribution gives a burn-in time of approximately 1000 iterations (out of 1 370 000 iterations in total) and a subsequent average frequency of acceptance of about 40–50%. Marginal *a posteriori* frequency distributions of P- and S-wave velocities are shown in figure 3. A certain layering of the Lunar velocity structure is revealed, of which an 'upper mantle' thickness of about 560 km and a crustal thickness of approximately 45 km were the most surprising features (see Khan *et al* (2000) and Hood and Zuber (2000), for a further discussion of these results). However, it is evident from the figure that the marginal uncertainties are large, and that the output should be interpreted with great care.

Another instructive example of Monte Carlo analysis of an inverse problem, this time from electrical impedance tomography, can be found in Kaipio *et al* (2000). This paper also gives further details on the theory of Monte Carlo methods.

3.4. Random walks in non-Bayesian ensemble inference

Although random walks are at the heart of every Monte Carlo algorithm, they are not only useful within a Bayesian formulation of an inverse problem. Many authors have used random walks to sample parameter spaces without defining a posterior probability distribution (4). As mentioned above, the earliest Monte Carlo algorithms used in geophysics were non-Bayesian. An example is the work of Press (1978, 1970a), which consisted of a series of nested uniform random searches in parameter space. A convenient way to view non-Bayesian ensemble inference is in terms of a two-stage approach, consisting of a search stage and an appraisal stage.

In the search stage, an algorithm (based on random walks) is used to collect samples, and the predictions of these models are compared to the data. In many cases the search

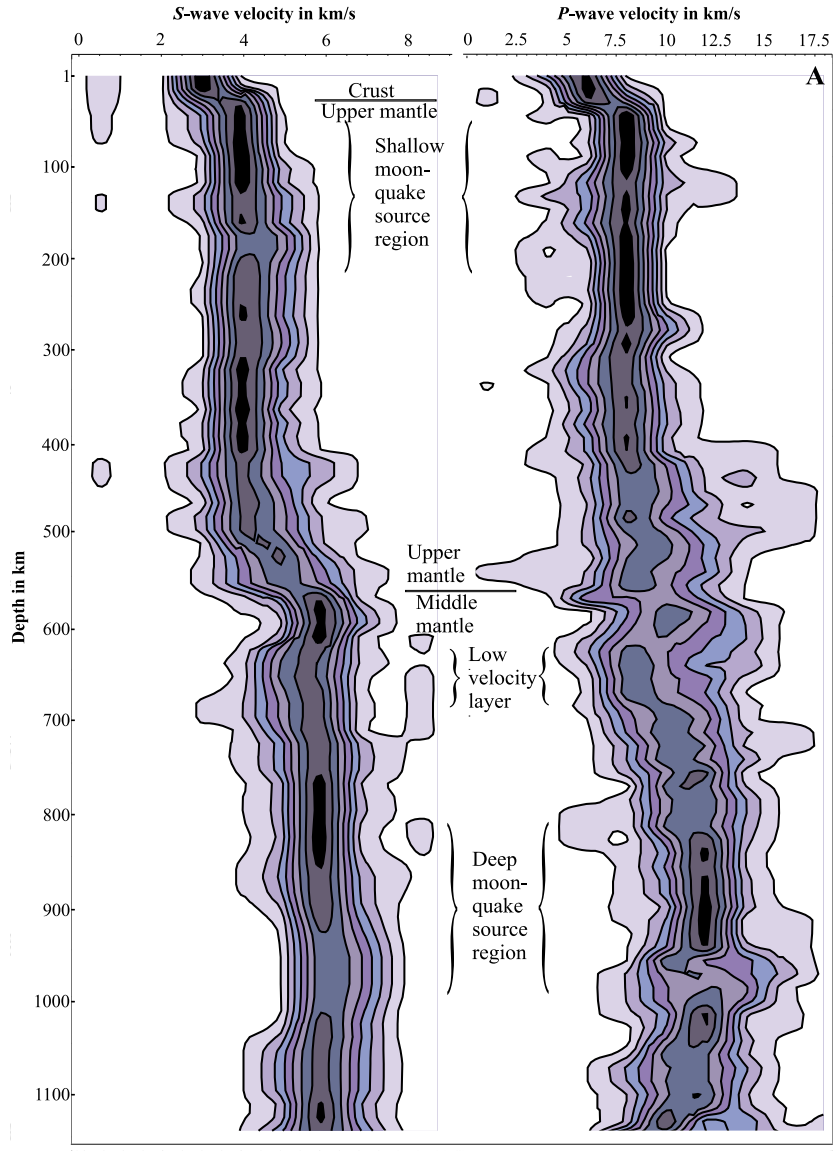


Figure 3. Marginal posterior velocity distributions for the velocity structure of the Moon. 50 000 models were used in constructing the two results. For each kilometre in depth, a histogram reflecting the marginal *a posteriori* probability distribution of sampled velocities has been computed. These marginals are lined up, and contour lines define nine equal-sized probability density intervals for the distributions.

process is adaptive, i.e. it makes use of the samples collected to guide the search for new models. Clearly, there is a strong connection with optimization problems, and indeed many direct search (i.e. non-derivative based) optimization algorithms have been used as search algorithms in inverse problems. Examples include simulated annealing, genetic algorithms and evolutionary programming techniques, and the neighbourhood algorithm (see below). Since many of these overlap with the area of global optimization we defer a discussion of them to the next section.

In the appraisal stage the objective is to make use of the complete ensemble of parameter space samples to draw inferences from the data. If an optimization algorithm has been used in the search stage to maximize a likelihood function (5), then the temptation is often to select the best data fitting model only and examine it in detail. However, this is almost always insufficient because of the noise in the data and the non-uniqueness of the underlying inverse problem. Even within a finite-dimensional parameter space one usually finds that if one model fits the data to an acceptable level (taking into account noise in the data), then an infinite number will, and so the best data fit model may well be mis-leading. (Just as in a Bayesian formulation the maximum of the posterior is seldom representative of the overall ensemble.)

An alternative to taking single 'best fit' models is to try and characterize the subset of data acceptable models in the collected ensemble, which may be useful if the search algorithm has sufficiently explored the parameter space. The earliest efforts in this direction consisted of simply directly comparing the acceptable models in the ensemble. This was the approach taken by Press (1968, 1970a, 1970b), however in that case just six Earth models were found from 5 million which fit all of the available seismic data and prior constraints.

Another way of characterizing an acceptable ensemble is to look for particular properties, or features, which all data acceptable models share, e.g. ones which have the least structure, or detail, as measured by some particular criteria. This extremal model approach was proposed by Parker (1977) in the context of nonlinear inverse problems, and can also be applied to the appraisal problem. Again the degree to which one can sensibly draw inferences from the ensemble depends crucially on the type of sampling performed during the search stage. Therefore, just as in a Bayesian approach, the search and appraisal stages are linked. For example, to generate many data acceptable models one needs the search algorithm to first find and then 'map out' the regions of parameter space where the fit to data is acceptable (which is often a difficult task, see Sambridge (2001)); while for the extremal model approach one needs the search stage to optimize a combination of data fit and model property, (often called a regularizing property).

Other methods have been proposed for characterizing a multi-dimensional ensemble of data-acceptable models. A summary can be found in Sen and Stoffa (1995). Examples include Vasco *et al* (1993) who used cluster analysis techniques, while Douma *et al* (1996) used a projection onto empirical orthogonal functions to determine the common features in their data-acceptable ensemble of Earth models (seismic wave speed as a function of depth). Another popular approach has been through semi-graphical methods (Basu and Frazer 1990, Nolte and Frazer 1994, Lomax and Snieder 1995, Kennett 1998). A draw back of many of these techniques is that they are best suited to the situation where the acceptable ensemble forms a single cluster, and not multiple unconnected clusters in parameter space. Recently, Sambridge (2001) has proposed a technique to map out the acceptable ensemble which is applicable to the multiple cluster case.

3.5. Design of random walks for optimization

3.5.1. Simulated annealing: Metropolis and Gibbs as optimizers.

The simulated annealing algorithm. When crystalline material is slowly cooled through its melting point, highly ordered, low-energy crystals are formed. The slower the cooling, the lower the final lattice energy. This physical process is a natural optimization method where the lattice energy E is the objective function to be minimized. Large numerical systems can be run through a similar optimization process if we identify parameters of the system with state space variables, and the objective function of the optimization problem with the energy E.

Thermal fluctuations in the system are simulated by randomly perturbing its parameters, and the size of the fluctuations are controlled by a temperature parameter T.

The simulated annealing algorithm (Kirkpatrick *et al* 1983) is a specialization and modification of the Metropolis algorithm, in that the desired equilibrium distribution for a constant value of the temperature parameter is the Gibbs–Boltzmann distribution

$$p_B(\boldsymbol{x}) = \exp(-E(\boldsymbol{x})/T)Z(T) \tag{35}$$

where $1/Z(T)$ is a normalization constant. This is the distribution over the state space of a statistical mechanical system in equilibrium with a heat bath of temperature T, and this is the kind of system we wish to simulate with this algorithm. Simulation of chemical annealing is now performed by gradually lowering the temperature T from a high value to near-zero. Close to $T = 0$ the Gibbs–Boltzmann distribution approximates a delta function at the global minimum for $E(\boldsymbol{x})$ (if it is unique).

Simulated annealing can also be realized by generating samples from the Gibbs–Boltzmann distribution by means of a Gibbs sampler. Rothman (1986) solves the so-called residual statics problem of reflection seismology in this way.

The Nulton–Salamon annealing schedule. Strictly speaking, simulated annealing is only guaranteed to work for infinitely slow 'cooling'. The practical problem is therefore: How can we decrease the temperature of the system in a finite number of steps, such that the final value of E, on average, is as close as possible to the global minimum for E?

Nulton and Salamon (1988) proposed an annealing method based on the idea that the numerical system should be kept as close to 'thermal equilibrium' as possible at all times. This was done by keeping the actual mean value $\langle E \rangle$ of the objective function at a constant distance

$$v = \frac{\langle E \rangle - \langle E \rangle_{eq}}{\sigma_E(T)} \tag{36}$$

from the theoretical (but never realized) equilibrium mean value $\langle E \rangle_{eq}$. In (36) $\sigma_E(T)$ is the standard deviation of $E(\boldsymbol{x})$ which, of course, fluctuates from iteration to iteration. v is known as the 'thermodynamic speed' and sometimes also as the 'thermodynamic distance'. The following differential equation for the annealing temperature schedule $T(t)$ can now be derived:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = -\frac{vT}{\epsilon(T)\sqrt{C(T)}} \tag{37}$$

where $C(T)$ is the *heat capacity* of the system, and $\epsilon(T)$ is its *relaxation time*. Andresen *et al* (1988) estimate an approximate, temperature-dependent, transition probability matrix $P_E(T)$ for transitions between 'energy levels' by monitoring transition frequencies during the annealing process. For each temperature, the heat capacity $C(T)$ can be evaluated from the eigenvector of $P_E(T \to \infty)$ with eigenvalue 1, and the relaxation time $\epsilon(T)$ can be calculated from the second largest eigenvalue of $P_E(T)$.

The thermodynamic speed v in equation (37) is calibrated, through some initial experimentation, to the problem at hand, such that the annealing schedule is close to zero after a predetermined number of iterations. The total number of iterations is, of course, limited by the available computer resources (Jakobsen *et al* 1988, Andresen *et al* 1988, Koren *et al* 1991).

In a more direct approach, the so-called *ensemble-based simulated annealing* (EBSA), heat capacities and relaxation times are estimated using the statistics of a set (ensemble, population) of parallel simulated-annealing walkers at the same temperature. In contrast to genetic algorithms, the individual members of an EBSA ensemble/population do not interact directly. Instead, thermodynamic properties of the problem are calculated from ensemble

averages, and optimal annealing schedules can be calculated on-the-fly or after an initial test run (Salamon *et al* 2002, Mosegaard and Vestergaard 1991).

3.5.2. Genetic algorithms. Genetic algorithms were originally devised as a model of adaptation in an artificial system, by Holland (1975). Early reference works are by Davis (1987), Goldberg (1989), and a useful tutorial can be found in Whitley (1994). Genetic algorithms fall into the class of Monte Carlo techniques because they also use random numbers to control the sampling of a parameter space. In contrast to simulated annealing which uses an analogy with a physical optimization process, genetic algorithms are based on analogy with biological evolution.

Within the past decade genetic algorithms have been applied in a wide range of the areas within the physical sciences, and for a range of purposes, only one of which is optimization. Many excellent references and web pages exist describing genetic algorithms and their variants (see the above works and references cited therein). Here we describe a simple genetic algorithm and compare it to other Monte Carlo importance sampling techniques by examining how it might be used to sample a particular probability distribution.

A key feature of genetic algorithms is the representation of physical variables of interest by a simple string data structure, and usually a binary string. It is straightforward to represent many optimization problems involving real-valued variables into a set of binary string representations. For a single variable, x_i, the simplest approach would be to choose upper and lower bounds, u_i and l_i, together with a discretization interval, and assign a binary number to each value the variable could take. For example, one might encode a real variable x_i into a binary number b_i, using

$$b_i = B\left\{\frac{x_i - l_i}{u_i - l_i}(2^l - 1)\right\} \tag{38}$$

where l is the length of the binary string produced, and we have introduced the operator $B\{y\}$ to indicate taking the binary value of y after rounding down to the nearest integer.

The genetic algorithm uses the bit string data structure to manipulate an ensemble of parameter space samples at each iteration. Here we restrict ourselves to a description of the basic three-operator genetic algorithm. Many variants exist, details of which can be found in the references cited.

Selection. From the initial population $\pi = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, of N strings an interim population of N parents is generated by randomly selecting strings with replacement, where the probability of selection, p_s, is a function of the fitness of each string. A common choice is

$$p_s(\boldsymbol{x}_j) = A \exp[-f(\boldsymbol{x}_j)/B] \tag{39}$$

where A is a normalizing constant,

$$A^{-1} = \sum_i \exp[-f(\boldsymbol{x}_i)/B] \tag{40}$$

and B is some measure of the spread of fitness values in the current population, e.g. the standard deviation. Since the population size is unchanged, the selection operator will generate multiple copies of some models in $\boldsymbol{x}_i$ at the expense of others, which may be extinguished completely. It is in this stage that the fitness of each string influences its chances of survival.

Crossover. From the parent population of N strings a new generation of N strings is produced by mixing pairs together. All of the parents are randomly paired to produce $N/2$ couples.

A cross-over probability p_c is assigned and each pair of strings is chosen for crossover with probability, p_c. If crossover is selected then the two strings are cut at a uniformly random point along their paths and two new strings are formed by interchanging the sub-strings. If crossover is not selected then the two parent strings are passed unscathed to the next generation.

Mutation. The final stage involves the process whereby any bit in an individual string is allowed to flip between 0 and 1 with probability p_m. This allows some degree of local diversity to be introduced into the population.

The action of the three types of process is to produce a new generation of N bit strings each of which has a new fitness value and the whole process can be repeated. After many iterations the population has the potential to evolve towards a fitter on average state. In general there appears to be no clear theory on how to choose the values of control parameters p_s, p_m N and B, to obtain an optimal performance for particular problems. Usually some level of empirical tuning is common. More sophisticated versions of genetic algorithms may also introduce extra control parameters, e.g. using different types of mapping from fitness to survival probability in (39), or perhaps different forms of crossover involving multiple string cuts.

The fundamental theorem. A theoretical analysis of the way genetic algorithms process information was performed by Holland (1975), which resulted in the *fundamental theorem*. Goldberg (1989) discusses the theorem in detail and shows how it explains the processing of *Schema* over each iteration. Schema are simply patterns within the binary strings representing each sample in parameter space. For a bit string of length seven ($l = 7$) an example of a schemata would be,

$$H = *11*0*1 \tag{41}$$

where the asterisk represents either a 1 or 0. H is the subset of bit strings differing in only the 1.0, 4.0 and 6.0 bits, and clearly represents many different possible combinations of bit-strings. Two properties of schema can be defined. The first is the *order*, represented by the operator $o(H)$, which gives the number of fixed positions (i.e. the number of 1 s or 0 s in the schemata), and the second is the *defining length*, $\delta(H)$, which is the distance between the first and last specified string position. For example, the schemata in (41) would have,

$$o(H) = 4, \qquad \text{and} \qquad \delta(H) = 5.$$

These quantities appear in the fundamental theorem which predicts how the number of schema changed between iterations of a genetic algorithm due to the three basic operations of selection, crossover and mutation. If we define $m(H, t)$ as the expected number of examples of schema H at iteration t, then the fundamental theorem says,

$$m(H, t+1) \geqslant m(H, t)\frac{f(H)}{\overline{f}}\left[1 - p_c\frac{\delta(H)}{l-1} - o(H)p_m\right] \tag{42}$$

where $f(H)$ represents the average fitness of all strings in the current population that correspond to schema H, and $\overline{f}$ is the average fitness of the population. Since each bit-string in a population belongs to many schema the genetic algorithm manipulates many schema in parallel at each iteration. The main conclusion that follows from (42) is that short, low-order, above-average schemata receive exponentially increasing numbers of copies in subsequent generations. Most theoretical analysis of genetic algorithms performance in optimization has centred on the way in which schema are created and destroyed.

During the 1990s, convergence properties of genetic algorithms were studied extensively. Conditions for guaranteed convergence towards optimal solutions were worked out for

important classes of algorithms, and convergence rates in terms of control parameters were estimated. Here, we shall not elaborate further on this interesting topic, but readers who want further details may, e.g., consult papers by Gao (1998) and Greenhalgh and Marshall (2000).

Genetic algorithms and sampling of probability distributions. Genetic algorithms were not designed to sample given probability distributions, but the efficiency of these algorithms for optimization has led researchers to search for ways of modifying them, so that they could be used for sampling (see, e.g., Sen and Stoffa (1995)). Here we shall take a close look at the problems we are facing when we try to develop GAs for sampling, and explain why the problem has not yet been solved.

A genetic algorithm is not based on a single random walk in $\mathcal{X}$. Instead, it is based on an ensemble of, say, N random walkers, operating simultaneously, and dependent on each other. This process can be understood if we consider a new space $\mathcal{P} = \mathcal{X}^N$, the *population space*, whose points are all possible *populations* consisting of points (here called *individuals*) from the original space $\mathcal{X}$. Note that, traditionally, $\mathcal{X}$ is a discrete space where the points are represented by strings of binary numbers. In $\mathcal{P}$ we define the probability of a population as the product of the probabilities of the individual points in the population.

We now re-examine each of three operators in a simple genetic algorithm, to determine whether they can be designed to sample $\mathcal{P}$ uniformly (that is, with a constant probability distribution). If this is the case then it is a simple matter to modify the GA such that it samples a given distribution $p(\boldsymbol{x})$ over $\mathcal{P}$. The reader can easily verify that the following algorithm will indeed sample $p(\boldsymbol{x})$:

(i) Given a point $\boldsymbol{x}_j \in \mathcal{P}$ currently visited by the algorithm. Use the uniform sampler to propose a point $\boldsymbol{x}_i$ as a candidate to be the next sample.
(ii) Accept only $\boldsymbol{x}_i$ with probability

$$p_{accept} = \begin{cases} \dfrac{p(\boldsymbol{x}_i)}{p(\boldsymbol{x}_j)} & \text{for } p(\boldsymbol{x}_i) \leqslant p(\boldsymbol{x}_j) \\ 1 & \text{otherwise.} \end{cases} \tag{43}$$

As we shall see, it is straightforward to design the mutation and the crossover substeps to give a uniform distribution, but not the selection substep. We will consider them in reverse order.

The mutation substep. The mutation substep is easy to design such that it has the uniform distribution over $\mathcal{P}$ as its equilibrium distribution. In this substep we select one individual $\boldsymbol{x}_j$ from the population space $\pi = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, uniformly at random (that is, giving each individual probability $1/N$ of being chosen). We now generate a new candidate individual $\boldsymbol{x}_i$ using an irreducible proposal distribution over $\mathcal{X}$ that is uniform in a neighbourhood surrounding (and including $\boldsymbol{x}_j$). Finally, we accept it with probability 1. This operation is one step of a random walk in $\mathcal{X}$ that, should it be iterated, would sample $\mathcal{X}$ uniformly. It is also clear that the entire population will, in this way, sample the population space $\mathcal{P}$ uniformly.

The crossover substep. In crossover we select one pair of individuals, $\xi_q = (\boldsymbol{x}_j, \boldsymbol{x}_k)$ from the population, uniformly at random, and generate a new candidate pair $\xi_r = (\boldsymbol{x}_i, \boldsymbol{x}_l)$ using a proposal distribution $U(\xi_r|\xi_q)$ over $\mathcal{X}^2$ (which in this case is *not* irreducible) that assigns equal probability to all possible pairs ξ_r obtained by choosing an integer $1 \leqslant \nu \leqslant \dim(\mathcal{X})$ uniformly at random, and then replacing the first ν components of $\boldsymbol{x}_j$ with the first ν components of $\boldsymbol{x}_k$, and vice versa. Finally, we accept the new pair ξ_r with probability 1.

This operation is one step of a random walk in $\mathcal{X}^2$ with the uniform distribution over $\mathcal{X}^2$ as an equilibrium distribution. It is also clear that the corresponding move of the entire population is one step of a random walk in $\mathcal{P}$ with the uniform distribution over $\mathcal{P}$ as an equilibrium distribution.

The selection substep. The problems arise in the selection substep. It is tempting, from the population π_j of N individuals to generate a new population π_i of N individuals by selecting models from π_j with uniform probability of selection, i.e. set $p_s(\pi_j)$ equal to a constant rather than using (39). As stated above this will generate multiple copies of some models in π_i at the expense of others, which may be removed completely.

The problem is now that the selection substep so defined violates microscopic reversibility and therefore has a non-uniform equilibrium distribution. The point is that selection is irreversible: it can only remove individuals from the population, not create them again. Hence, our results concerning multi-step iterations do not apply. Forming complete iterations by combining mutation, crossover and selection will not lead to a uniform sampling, and no modification in selection probabilities will make it do so.

These observations do not mean that it is impossible to find mutation-, crossover- and selection-substeps that, taken together in each iteration, will sample the population space uniformly. It only means that these three substeps cannot be designed independently. Some mechanism must be built into the mutation- and crossover substeps that compensate for the irreversibility of the selection substep, such that microscopic reversibility is restored. Unfortunately, no solution has yet been found to this intricate problem. If it can be found, it will no doubt have great practical significance.

It is sometimes claimed that a GA who 'has a desired distribution as an approximate equilibrium distribution' is a sufficiently good solution to the above-mentioned problem, because even a Metropolis algorithm designed to sample the distribution will only give approximate solutions (due to limited sampling time). There are two reasons why this argument is incorrect:

(i) A clear definition what is meant by 'an approximate equilibrium distribution' is usually lacking.

(ii) A genetic algorithm with only an 'approximately correct' equilibrium distribution will, asymptotically, deviate more and more from the desired equilibrium distribution (and nobody knows how much). In contrast to this, a Metropolis algorithm having the desired equilibrium distribution will, asymptotically, sample this distribution.

This reasoning points to a clear difference between genetic algorithms and other Monte Carlo sampling methods.

3.5.3. The neighbourhood algorithm. The neighbourhood algorithm is a recently developed Monte Carlo search algorithm that has found a number of applications in seismic inverse problems (Sambridge 1998, 1999a, 1999b, 2001). This is an ensemble based search technique that uses concepts from computational geometry to guide the sampling of a parameter space.

The basic idea is that at each iteration the multi-dimensional parameter space is partitioned into a set of *Voronoi polytopes* constructed (uniquely) about the samples generated from previous iterations, $\boldsymbol{x}_j, (j = 1, \ldots, n)$. Voronoi polytopes (cells) are simply nearest-neighbour regions, as determined by a suitable distance measure. An L_2-norm is a common choice. More formally, any new point $\boldsymbol{x}$, lies within the ith Voronoi cell if,

$$\|\boldsymbol{x} - \boldsymbol{x}_i\| < \|\boldsymbol{x} - \boldsymbol{x}_j\| \qquad (j = 1, \ldots, i-1, i+1, \ldots, n). \tag{44}$$

Voronoi cells have some useful properties in that they are always unique, space filling, and adapt to the density of the samples about which they are constructed (one per cell), i.e. as the

local sampling density increases the size of each Voronoi cell decreases. Since each Voronoi cell forms a neighbourhood about one of the previously generated samples, they can be used to guide further sampling. One choice is to place samples uniformly randomly in selected Voronoi cells. A Gibbs sampler is ideal for this purpose. For example, if the jth Voronoi cell were chosen for sampling then a random walk would begin from model $\boldsymbol{x}_j$. The proposal distribution for the Gibbs sampler at the kth substep, is to be given by (26), with

$$p(\boldsymbol{x}_i) = \begin{cases} 1, & \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leqslant \|\boldsymbol{x}_i - \boldsymbol{x}_l\| \ (l = 1, \ldots, n) \\ 0, & \text{otherwise} \end{cases} \tag{45}$$

and the acceptance probability is unity. Selection of cells to re-sample can be made either deterministically or probabilistically on the basis of the fitness function evaluated at the corresponding samples. A simple choice would be to re-sample only, say, the 10% of cells with the higher fitness values. The number of cells chosen, n_r, would become a control parameter of the algorithm. Since these cells are not necessarily neighbours of each other, sampling can be performed in multiple regions simultaneously.

An alternative to re-sampling chosen Voronoi cells is to again use a standard Gibbs sampler to generate new samples distributed according to the *neighbourhood approximation* of the the fitness/objective function. This is formed by setting the fitness to a constant inside each Voronoi cell. When a Gibbs sampler is applied to the neighbourhood approximation, it is equivalent to setting the fitness function, $f(\boldsymbol{x})$, of any point equal to that of its nearest point among the current set, i.e. we have,

$$f_{NA}(\boldsymbol{x}) = f(\boldsymbol{x}_j) \qquad \{\boldsymbol{x}_j : \min(\|\boldsymbol{x} - \boldsymbol{x}_l\|) \ (l = 1, \ldots, n)\}. \tag{46}$$

In this case the proposal distribution for the Gibbs sampler becomes (26), with

$$p(\boldsymbol{x}_i) = f_{NA}(\boldsymbol{x}_i). \tag{47}$$

Note that the Gibbs sampler is used to sample from the neighbourhood approximation of the fitness function (47), or the Voronoi cells directly (45), and in both cases no further fitness evaluations are required for the random walks. An iteration is complete when the required number of new models, n_s, are selected from the chain of samples produced by the random walks. Only then is the fitness of each new sample calculated and the Voronoi cells (updated to included the latest n_s samples) for the next iteration. (Note that n_s is a second control parameter.) In each of the two scenarios represented by (47) and (45), multiple random walks can be performed in parallel starting from the previous samples $f(\boldsymbol{x}_j)(j = 1, \ldots, n)$.

Like a genetic algorithm the neighbourhood approach generates a new ensemble of samples at each iteration, but does so using the Voronoi cell concept to identify 'promising regions' of parameter space. It also has similarities with other Monte Carlo methods, since it makes use of a Gibbs sampler. We note here that one could equally well make use of the neighbourhood approximation of the fitness function within other optimization procedures like simulated annealing or genetic algorithms, i.e. to intermittently replace expensive evaluations of the fitness function, with cheaper but more approximate evaluations of $f_{NA}(\boldsymbol{x})$. Sambridge (1999a) has shown that the neighbourhood algorithm can result in quite a complex 'self-adaptive' search process in global optimization.

In addition to searching a parameter space for optimization the neighbourhood algorithm may also be used in the appraisal problem, i.e. analysing the resulting ensemble. Sambridge (1999b) has shown how one can use the neighbourhood approximation within a Bayesian formulation to estimate resolution and parameter covariances from an arbitrarily distributed ensemble of samples. Initial results with this technique for both search and appraisal are quite promising but more experience is required. In addition it is not yet known whether the search algorithm could be applied to combinatorial optimization, because of the lack of a general definition of a Voronoi cell (nearest neighbour region) for combinatorial problems.

4. Conclusions

Monte Carlo methods provide a systematic way of dealing with (discrete) inverse problems for which we have incomplete knowledge of the relationship between data and model parameters. This is the case, e.g., for many highly nonlinear problems, where the forward relation is insusceptible to mathematical analysis, and is only given by a complex algorithm. Monte Carlo methods can be divided into two groups, the first of which is devoted to sampling from a probability density, and the second is designed to search for near-optimal solutions to the problem.

The most widely used sampling algorithms are the Metropolis algorithm and the Gibbs sampler, which are used in cases where a thorough resolution and uncertainty analysis is called for. Although the equilibrium theory of these algorithms is simple and well mapped-out, many theoretical and practical problems concerning their 'speed of equilibration' remain to be solved.

The dominating stochastic optimization algorithms are simulated annealing and the genetic algorithms. In many applications, the genetic algorithm has shown its strength as a search method. Attempts at constructing a sampling algorithm from the same principles as the genetic algorithm has failed so far, but if the problem can be solved, it will no doubt be a great step forward in many practical applications.

In this paper we have covered the basic Monte Carlo algorithms currently in use in applications. Several variants of these algorithms exist, many of which are adaptations of the basic methods, or exploit special properties of the problem considered. The readers are referred to the literature for information on these methods. Important examples are *reversible jump MCMC* (Green 1995) and *Simulated Tempering* (Geyer and Thompson 1995, Marini and Parisi 1992). Readers who wish to dive into the basic theory of Markov chains may consult Kemeny and Snell (1976), or Seneta (1981).

Acknowledgments

The authors would like to thank Amir Khan for providing the figures illustrating the inversion of the Apollo seismic data. KM would like to thank Albert Tarantola, Peter Salamon, Bjarne Andresen and many other colleagues for stimulating cooperation on Monte Carlo methods over the years.

References

Anderssen R S 1970 The character of non-uniqueness in the conductivity modelling problem for the Earth *Pure Appl. Geophys.* **80** 238–59

Anderssen R S and Senata E 1971 A simple statistical estimation procedure for Monte Carlo inversion in geophysics *Pure Appl. Geophys.* **91** 5–13

Anderssen R S and Senata E 1972 A simple statistical estimation procedure for Monte Carlo inversion in geophysics: efficiency and Hempel's paradox *Pure Appl. Geophys.* **96** 5–14

Anderssen R S, Worthington M H and Cleary J R 1972 Density modelling by Monte Carlo inversion—I methodology *Geophys. J. R. Astron. Soc.* **29** 433–44

Andresen B, Hoffman K H, Mosegaard K, Nulton J, Pedersen J M and Salamon P 1988 *J. Physique* **49** 1485

Basu A and Frazer L N 1990 Rapid determination of critical temperature in simulated annealing inversion *Science* **249** 1409–12

Biswas N N and Knopoff L 1974 The structure of the upper mantle under the United States from dispersion of Rayleigh waves *Geophys. J. R. Astron. Soc.* **36** 515–39

Burton P W 1977 Inversions of high frequency $Q_\gamma^{-1}(f)$ *Geophys. J. R. Astron. Soc.* **48** 29–51

Burton P W and Kennett B L N 1974a Upper mantle zone of low *Q Nature* **238** 84

Burton P W and Kennett B L N 1974b Upper mantle zone of low *Q Nature* **238** 87–90

Cary P W and Chapman C H 1988 Automatic 1D waveform inversion of marine seismic refraction data *Geophys. J.* **93** 527–46

Dahl-Jensen D, Mosegaard K, Gundestrup N, Clow G D, Johnsen S J, Hansen A W and Balling N 1998 Past temperatures directly from the Greenland ice sheet *Science* **9** 268–71

Davis L 1987 Genetic algorithms and simulated annealing *Research Notes in Artificial Intelligence* (London: Pitman)

Douma H, Snieder R and Lomax A 1996 Ensemble inference in terms of empirical orthogonal functions *Geophys. J. Int.* **127** 363–78

Fishman G S 1996 *Monte Carlo. Concepts, Algorithms, and Applications* (New York: Springer)

Gallagher K and Sambridge M 1994 Genetic algorithms: a powerful tool for large-scale non-linear optimization problems *Comput. Geosci.* **20** 1229–36

Gallagher K, Sambridge M S and Drijkoningen G G 1991 Genetic algorithms: an evolution on Monte Carlo methods in strongly non-linear geophysical optimization problems *Geophys. Res. Lett.* **18** 2177–80

Gao Y 1998 An upper bound on the convergence rates of canonical genetic algorithms *Complexity International* vol 5

Gelman A, Roberts G O and Gilks W R 1996 Efficient Metropolis jumping rules *Bayesian Statistics* vol 5, ed J M Bernardo, J O Berger, A P Dawid and A F M Smith (Oxford: Clarendon) pp 599–608

Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images *IEEE Trans. Patt. Analysis Mach. Int.* **6** 721–41

Geyer C J and Thompson E A 1995 Annealing Markov chain Monte Carlo with applications to ancestral inference *J. Am. Stat. Assoc.* **90** 909–20

Goldberg D E 1989 *Genetic Algorithms in Search, Optimization, and Machine Learning* (Reading, MA: Addison-Wesley)

Goncz J H and Cleary J R 1976 Variations in the structure of the upper mantle beneath Australia, from Rayleigh wave observations *Geophys. J. R. Astron. Soc.* **44** 507–16

Gouveia W P and Scales J A 1998 Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis *J. Geophys. Res.* **103** 2759–79

Green P 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination *Biometrica* **82** 711–32

Greenhalgh D and Marshall S 2000 Convergence criteria for genetic algorithms *SIAM J. Comput.* **30** 269–82

Haddon R A W and Bullen K E 1969 An earth model incorporating free earth oscillation data *Phys. Earth Planet. Inter.* **2** 35–49

Hastings W K 1970 Monte Carlo sampling methods using Markov Chain and their applications *Biometrika* **57** 97–109

Holland J H 1975 *Adaptation in Natural and Artificial Systems* (Ann Arbor, MI: The University of Michigan Press)

Hood L and Zuber M 2000 Recent refinements in geophysical constraints on Lunar origin and evolution *The Origin of the Earth and Moon* ed R Canup and K Righter (Tucson, AZ: University of Arizona Press)

Housholder A S (ed) 1951 *Monte Carlo Method (Mathematics Series 12)* (Washington, DC: National Bureau of Standards)

Jakobsen M O, Mosegaard K and Pedersen J M 1988 Global model optimization in reflection seismology by simulated annealing *Model Optimization in Exploration Geophysics 2* ed E Vogel (Wiesbaden: Braunschweig) p 361

Jestin F, Huchon P and Gaulier J M 1994 The Somalia plate and the East African rift system; present-day kinematics *Geophys. J. Int.* **116** 637–54

Jones A G and Hutton R 1979 A multi-station magnetotelluric study in southern Scotland; II, Monte Carlo inversion of the data and its geophysical and tectonic implications *Geophys. J. R. Astron. Soc.* **56** 351–68

Kaipio J P, Kolehmainen V, Somersalo E and Vauhkonen M 2000 Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography *Inverse Problems* **16** 1083–618

Keilis-Borok V I and Yanovskaya T B 1967 Inverse problems of seismology *Geophys. J.* **13** 223–34

Kemeny J G and Snell J L 1976 Finite Markov chains *Springer Undergraduate Texts in Mathematics* 2nd edn

Kennett B L N 1998 On the density distribution within the earth *Geophys. J. Int.* **132** 374–82

Kennett B L N and Nolet G 1978 Resolution analysis for discrete systems *Geophys. J. R. Astron. Soc.* **53** 413–25

Kennett B L N and Sambridge M 1992 Earthquake location; genetic algorithms for teleseisms *Phys. Earth Planet. Inter.* **75** 103–10

Khan A and Mosegaard K 2001 New information on the deep lunar interior from an inversion of lunar free oscillation periods *Geophys. Res. Lett.* **28** 1791

Khan A, Mosegaard K and Rasmussen K L 2000 A new seismic velocity model for the Moon from a Monte Carlo inversion of the Apollo Lunar seismic data *Geophys. Res. Lett.* **27** 1591–4

Kirkpatrick S C, Gelatt D and Vecchi M P 1983 Optimization by simulated annealing *Science* **220** 671–80

Koren Z, Mosegaard K, Landa E, Thore P and Tarantola A 1991 Monte Carlo estimation and resolution analysis of seismic background velocities *J. Geophys. Res.* **96** 20 289–99

Landa E, Beydoun W and Tarantola A 1989 Reference velocity model estimation from prestack waveforms; coherency optimization by simulated annealing *Geophysics* **54** 984–90

Lomax A and Snieder R 1995 Identifying sets of acceptable solutions to non-linear geophysical inverse problems which have complicated misfit functions *Nonlinear Process. Geophys.* **2** 222–7

Marini E and Parisi G 1992 Simulated tempering: a new Monte Carlo scheme *Europhys. Lett.* **19** 451–8

Marroquin J, Mitter S and Poggio T 1987 Probabilistic solution of ill-posed problems in computational vision *J. Am. Stat. Assoc.* **82** 76–89

Metropolis N, Rosenbluth M N, Rosenbluth A W, Teller A H and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087–92

Metropolis N and Ulam S M 1949 The Monte Carlo method *J. Am. Stat. Assoc.* **44** 335–41

Mills J M and Fitch T J 1977 Thrust faulting and crust-upper mantle structure in East Australia *Geophys. J. R. Astron. Soc.* **48** 351–84

Mosegaard K and Rygaard-Hjalsted C 1999 Bayesian analysis of implicit inverse problems *Inverse Problems* **15** 573–83

Mosegaard K and Tarantola A 1995 Monte Carlo sampling of solutions to inverse problems *J. Geophys. Res.* **100** 12 431–47

Mosegaard K and Vestergaard P 1991 A simulated annealing approach to seismic model optimization with sparse prior information *Geophys. Prospect.* **39** 599–611

Nolte B and Frazer L N 1994 Vertical seismic profile inversion with genetic algorithms *Geophys. J. Int.* **117** 162–79

Nulton J D and Salamon P 1988 Statistical mechanics of combinatorial optimization *Phys. Rev.* A **37** 1351–6

Parker R L 1977 Understanding inverse theory *Ann. Rev. Earth Planet Sci.* **5** 35–64

Parker R L 1994 *Geophysical Inverse Theory* (Princeton, NJ: Princeton University Press)

Pedersen J B and Knudsen O 1990 Variability of estimated binding parameters *Biophys. Chem.* **36** 167–76

Press F 1968 Earth models obtained by Monte Carlo inversion *J. Geophys. Res.* **73** 5223–34

Press F 1970a Earth models consistent with geophysical data *Phys. Earth Planet. Inter.* **3** 3–22

Press F 1970b Regionalized Earth models *J. Geophys. Res.* **75** 6575–81

Ricard Y, Vigny C and Froidevaux C 1989 Mantle heterogeneities, geoid, and plate motion; a Monte Carlo inversion *J. Geophys. Res.* **94** 13 739–54

Rothman D H 1985 Nonlinear inversion statistical mechanics, and residual statics corrections *Geophysics* **50** 2784–96

Rothman D H 1986 Automatic estimation of large residual statics corrections *Geophysics* **51** 332–46

Rygaard-Hjalsted C, Mosegaard K and Olsen N 2000 Resolution studies of fluid flow models near the core–mantle boundary through Bayesian inversion of geomagnetic data *Methods and Applications of Inversion: Proc. IIC98 Conf. (Copenhagen, 1998)* ed P C Hansen, B H Jacobsen and K Mosegaard, pp 255–75

Salamon P, Sibani P and Frost R 2002 *Facts, Conjectures and Improvements for Simulated Annealing (SIAM Monographs on Mathematical Modeling and Computation)* at press

Sambridge M 1998 Exploring multi-dimensional landscapes without a map *Inverse Problems* **14** 427–40

Sambridge M 1999a Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space *Geophys. J. Int.* **138** 479–94

Sambridge M 1999b Geophysical inversion with a neighbourhood algorithm—II. Appraising the ensemble *Geophys. J. Int.* **138** 727–46

Sambridge M 2001 Finding acceptable models in nonlinear inverse problems using a neighbourhood algorithm *Inverse Problems* **17** 387–403

Sambridge M and Drijkoningen G G 1992 Genetic algorithms in seismic waveform inversion *Geophys. J. Int.* **109** 323–42

Sambridge M and Mosegaard K 2001 Monte Carlo methods in geophysical inverse problems *Rev. Geophys.* submitted

Scales J A, Smith M L and Fischer T L 1992 Global optimization methods for multimodel inverse problems *J. Comput. Phys.* **103** 258–68

Sen M K and Stoffa P L 1992 Rapid sampling of model space using genetic algorithms *Geophys. J. Int.* **108** 281–92

Sen M K and Stoffa P L 1995 Global optimization methods in geophysical inversion *Advances in Exploration Geophysics* vol 4 (Amsterdam: Elsevier)

Seneta E 1981 *Non-Negative Matrices and Markov Chains* 2nd edn (Berlin: Springer)

Shibutani T, Sambridge M and Kennett B 1996 Genetic algorithm inversion for receiver functions with application to crust and uppermost mantle structure beneath Eastern Australia *Geophys. Res. Lett.* **23** 1829–32

Smith M L, Scales J A and Fischer T L 1992 Global search and genetic algorithms *Geophysics: The Leading Edge of Exploration* pp 22–6

Stoffa P L and Sen M K 1991 Nonlinear multiparameter optimization using genetic algorithms: inversion of plane wave seismograms *Geophysics* **56** 1794–810

Vasco D W, Johnson L R and Majer E L 1993 Ensemble inference in geophysical inverse problems *Geophys. J. Int.* **117** 711–28

Whitley D L 1994 A genetic algorithm tutorial *Stat. Comput.* **4** 65–85

Wiggins R A 1969 Monte Carlo inversion of body wave observations *J. Geophys. Res.* **74** 3171–81

Wiggins R A 1972 The general inverse problem: implication of surface waves and free oscillations for Earth structure *Rev. Geophys. Space Phys.* **10** 251–85

Wilson W G and Vasudevan K 1991 Application of the genetic algorithm to residual statics estimation *Geophys. Res. Lett.* **18** 2181–4

Worthington M H, Cleary J R and Anderssen R S 1972 Density modelling by Monte Carlo inversion—II comparison of recent earth models *Geophys. J. R. Astron. Soc.* **29** 445–57

Worthington M H, Cleary J R and Anderssen R S 1974 Upper and lower mantle shear velocity modelling by Monte Carlo inversion *Geophys. J. R. Astron. Soc.* **36** 91–103