Bayesian inference, Gibbs sampler and uncertainty estimation in nonlinear geophysical inversion

Conference Paper · June 1994

DOI: 10.3997/2214-4609.201409813

2 authors:

Mrinal K. Sen
University of Texas at Austin
488 PUBLICATIONS **7,500** CITATIONS
SEE PROFILE

P. L. Stoffa
University of Texas at Austin
253 PUBLICATIONS **7,818** CITATIONS
SEE PROFILE

Geophysical Prospecting, 1996, **44**, 313–350

Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion[1]

Mrinal K. Sen[2] and Paul L. Stoffa[3]

Abstract

The posterior probability density function (PPD), $\sigma(\mathbf{m}|\mathbf{d}_{obs})$, of earth model $\mathbf{m}$, where $\mathbf{d}_{obs}$ are the measured data, describes the solution of a geophysical inverse problem, when a Bayesian inference model is used to describe the problem. In many applications, the PPD is neither analytically tractable nor easily approximated and simple analytic expressions for the mean and variance of the PPD are not available. Since the complete description of the PPD is impossible in the highly multi-dimensional model space of many geophysical applications, several measures such as the highest posterior density regions, marginal PPD and several orders of moments are often used to describe the solutions. Calculation of such quantities requires evaluation of multidimensional integrals. A faster alternative to enumeration and blind Monte-Carlo integration is *importance sampling* which may be useful in several applications. Thus how to draw samples of $\mathbf{m}$ from the PPD becomes an important aspect of geophysical inversion such that importance sampling can be used in the evaluation of these multi-dimensional integrals. Importance sampling can be carried out most efficiently by a Gibbs' sampler (GS). We also introduce a method which we called parallel Gibbs' sampler (PGS) based on genetic algorithms (GA) and show numerically that the results from the two samplers are nearly identical.

We first investigate the performance of enumeration and several sampling based techniques such as a GS, PGS and several multiple *maximum a posteriori* (MAP) algorithms for a simple geophysical problem of inversion of resistivity sounding data. Several non-linear optimization methods based on simulated annealing (SA), GA and some of their variants can be devised which can be made to reach very close to the maximum of the PPD. Such MAP estimation algorithms also sample different points in the model space. By repeating these MAP inversions several times, it is possible to sample adequately the most significant portion(s) of the PPD and all these models can

[1] Paper presented at the 56th EAEG meeting, June 1994, Vienna, Austria. Received November 1994, revision accepted August 1995.

[2] Institute for Geophysics, The University of Texas at Austin, 8701 N. Mopac Expressway, Austin, TX 78759-8397, USA.

[3] Institute for Geophysics and Department of Geological Sciences, The University of Texas at Austin, 8701 N. Mopac Expressway, Austin, TX 78759-8397, USA.

be used to construct the marginal PPD, mean, covariance, etc. We observe that the GS and PGS results are identical and indistinguishable from the enumeration scheme. Multiple MAP algorithms slightly underestimate the posterior variances although the correlation values obtained by all the methods agree very well. Multiple MAP estimation required 0.3% of the computational effort of enumeration and 40% of the effort of a GS or PGS for this problem. Next, we apply GS to the inversion of a marine seismic data set to quantify uncertainties in the derived model, given the prior distribution determined from several common midpoint gathers.

Introduction

The geophysical inverse problem of estimating earth model parameters from observations of geophysical data often suffers from the fundamental limitation that several models may fit the observations very well. This phenomenon, which has been called non-uniqueness in the geophysical literature, may be caused by several factors. The most well recognized of these is that in the real earth, the properties vary continuously in all spatial directions (i.e. the model space is truly infinite dimensional) and we are faced with the problem of constructing an earth model from a finite, albeit small, set of measurements. Thus the inverse problem is highly under-determined and will result in many non-unique solutions (e.g. Menke 1984). Since in many cases the earth may be modelled with a discrete (small) set of layers based on independent information, this type of non-uniqueness can be greatly reduced. One other cause of non-uniqueness is related to the problem of identifiability or sensitivity of the model to the data. For example, in seismic problems the information about the seismic wave velocity is contained in the moveout (traveltime as a function of offset) in the seismograms. Therefore, the use of only near-offset traces does not reveal much information on velocity. This means that a suite of velocities will explain the data equally well, resulting in non-unique estimates of the velocities. In seismic tomography problems, the regions of the earth with little or no ray coverage cannot be resolved from the data and the slowness estimates for these regions will be ambiguous. Thus an interpreter is faced with several choices for the earth model and often the number of choices can be reduced based on the prior knowledge of the earth models.

Several attempts have been made to reduce uncertainty by imposing constraints such as regularization or solving for smooth models. Of course, what kind of smoothing is appropriate is highly debatable. In global tomography problems, use of spherical harmonics has been very popular (Dziewonski 1984). Both the spherical harmonics and cubic splines are smoothing operators and may introduce unrealistic features in the model (Shalev 1993) if not used properly. The optimal fitting of splines or any other functional to gridded models is itself an inverse problem. However, they have often been found to generate geologically meaningful solutions. Jackson (1979) correctly pointed out that the aim of inversion is not only to find a best-fitting model but also to characterize the uncertainty in the inversion result.

One other approach to describing the inverse problem is to use a statistical framework (e.g. Jackson 1979; Box, Leonard and Chien-Fu 1983; Tarantola and Valette 1982; Tarantola 1987) and make an attempt to describe or characterize the non-uniqueness of the solution by describing the solution in terms of probability density functions (pdf) in model space. This approach is valid for any model parametrization that we choose. In many situations, we may have prior information to restrict the models to a small set of layers but even then different model parameter values may be altered either independently or dependent on other parameters to explain the observed data. The statistical approach enables us to estimate uncertainty bounds on the resulting model and correlation between different model parameters. Such an approach has also been criticized for the requirement of the prior knowledge of data and theory errors. The advantage of the statistical approach is that it results in the posterior probability density function in a model given the observed data. Several measures of uncertainty in model space can be obtained for a given parametrization. Even though most statistical approaches make simplistic assumptions of Gaussian prior pdfs and uncorrelated data errors, the results obtained from such approaches are physically meaningful.

In our work we take the latter approach to describe the results of inversion by means of a pdf and address issues of estimating the posterior marginal pdf and covariance accurately and rapidly. One problem that makes such an estimation problem computationally intensive is that many geophysical inverse problems are non-linear and the error or misfit function which characterizes the difference between observed and synthetic data is multimodal. We therefore first summarize the Bayesian inference model for the geophysical inversion and then describe several methods of estimating uncertainties using non-linear optimization methods with some numerical examples.

Bayesian formulation

The classical approach to geophysical inversion involves assuming a linear relationship between data and model, and several useful concepts were developed based on linear algebra. The concept of non-uniqueness was very clearly illustrated by examining the length of the data and model vectors to study whether a problem is underdetermined or overdetermined. Concepts of data and model resolutions were also developed and the model resolution essentially describes the uncertainty in the derived result (see Menke 1984 for details). Such a classical approach was modified into the maximum likelihood method (MLM) in which the probabilistic model of error in data was included and then attempts were also made to use prior information on the model in describing the solution of an inverse problem. Tarantola and Valette (1982) and Tarantola (1987) formulated the inverse problem on a strong theoretical and statistical basis and showed how different states of information can be combined to describe an answer to the inverse problem. Tarantola (1987) also showed how gradient based optimization methods can be used in the search for the minimum of an

error function. Bayes's formulation is another approach often used in geophysical problems. This is similar to Tarantola's (1987) formulation that can also be derived from Bayes's rule (Duijndam 1988). Bayes's rule is a very convenient mathematical tool to update our current knowledge when new measurements become available. A detailed description of Bayesian inference models can be found in the text by Box and Tiao (1973). Some of its geophysical applications are given by Jackson and Matsura (1985), Duijndam (1987) and Cary and Chapman (1988). In the following we summarize Bayes's formulation, for completeness.

We will represent the model by a vector $\mathbf{m}$ and data by a vector $\mathbf{d}$ given by

$$\mathbf{m} = [m_1, m_2, m_3, \ldots, m_M]^{\mathrm{T}} \tag{1}$$

and

$$\mathbf{d} = [d_1, d_2, d_3, \ldots, d_{ND}]^{\mathrm{T}},$$

consisting of elements m_i and d_i respectively, each one of which is considered to be a random variable. The quantities M and ND are the number of model parameters and data points respectively and the superscript 'T' represents a matrix transpose. Following Tarantola's (1987) notation, we assume that $p(\mathbf{d}|\mathbf{m})$ is the pdf of $\mathbf{d}$ for given $\mathbf{m}$, $\sigma(\mathbf{m}|\mathbf{d})$ is the conditional pdf of $\mathbf{m}$ for given $\mathbf{d}$, $p(\mathbf{d})$ is the pdf of data $\mathbf{d}$ and $p(\mathbf{m})$ is the pdf of model $\mathbf{m}$ independent of the data. From the definition of the conditional probabilities, we have

$$\sigma(\mathbf{m}|\mathbf{d})p(\mathbf{d}) = p(\mathbf{d}|\mathbf{m})p(\mathbf{m}). \tag{2}$$

From (2) we obtain an equation for the conditional pdf of model $\mathbf{m}$ given the measured data $\mathbf{d}$ as follows

$$\sigma(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \tag{3}$$

which describes the state of information on $\mathbf{m}$ given $\mathbf{d}$. Equation (3) is the so-called Bayes's rule. The denominator $p(\mathbf{d})$ does not depend on $\mathbf{m}$ and can be considered a constant factor in the inverse problem (Duijndam 1988). Replacing the denominator in (3) with a constant, we have

$$\sigma(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m}). \tag{4}$$

Let $\mathbf{d}_{\mathrm{obs}}$ denote the measured data. When we substitute $\mathbf{d} = \mathbf{d}_{\mathrm{obs}}$ in $p(\mathbf{d}|\mathbf{m})$, the result is interpreted as a function of $\mathbf{m}$, denoted by $l(\mathrm{d}_{\mathrm{obs}}|\mathbf{m})$, called the likelihood function (Box and Tiao 1973; Cary and Chapman 1988). Equation (4) can then be written as

$$\sigma(\mathbf{m}|\mathbf{d}_{\mathrm{obs}}) \propto l(\mathbf{d}_{\mathrm{obs}}|\mathbf{m})p(\mathbf{m}). \tag{5}$$

The pdf $p(\mathbf{m})$ is the probability of the model $\mathbf{m}$ independent of the data, i.e. it describes the information we have on the model without the knowledge of the data and is called the *prior* pdf. Similarly the pdf $\sigma(\mathbf{m}|\mathbf{d}_{\mathrm{obs}})$ is the state of information on model $\mathbf{m}$ given the data and is called the *posterior* pdf or the PPD when normalized.

Clearly the PPD is obtained by a product of the likelihood function and the pdf. Thus the prior knowledge in the model is modified by the likelihood function. Box and Tiao (1973) and Duijndam (1988) show different theoretical examples of how the posterior pdf is influenced depending on the relative importance between the prior pdf and the likelihood function. For a uniform prior pdf, the posterior pdf is determined primarily by the likelihood function. Only in very rare circumstances can we have a situation where the prior pdf will dominate the likelihood function in the entire model space in a multiparameter problem. Generally the prior pdf dominates a subspace of the parameter space while the likelihood function dominates other (and usually larger) subspaces.

The choice of the likelihood function depends on the distribution of the noise or error in the data (Box and Tiao 1973; Cary and Chapman 1988). Thus it requires prior knowledge of the error distribution. This is a very important issue since in many situations it is very difficult to obtain an estimate of noise statistics. The error can be due to measurement (e.g. instrument errors) or due to the use of inexact theory in the prediction of the data (Tarantola 1987). Assuming Gaussian errors, the likelihood function takes the form

$$l(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto \exp(-E(\mathbf{m})), \tag{6}$$

where $E(\mathbf{m})$ is the error function given by

$$E(\mathbf{m}) = \left(\tfrac{1}{2}(\mathbf{d}_{\text{obs}} - g(\mathbf{m}))^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}(\mathbf{d}_{\text{obs}} - \gamma(\mathbf{m}))\right), \tag{7}$$

where g is the forward modelling operator and $\mathbf{C}_{\mathrm{D}}$ is called the data covariance matrix. The data covariance matrix consists of two parts, namely the experimental uncertainty or observational error $\mathbf{C}_{\mathrm{d}}$ and the modelling uncertainty or error due to theory $\mathbf{C}_{\mathrm{T}}$ (Tarantola 1987) , i.e.

$$\mathbf{C}_{\mathrm{D}} = \mathbf{C}_{\mathrm{d}} + \mathbf{C}_{\mathrm{T}}. \tag{8}$$

Here it is assumed that the errors due to theory and observation are independent.

In geophysical inversion, a distinction is usually made between theoretical and observational errors. While the observational errors are assumed to be caused by instrumental errors, the theoretical errors are assumed to be due to the use of inexact theory in the prediction of data. Such a distinction is somewhat arbitrary and debatable. In reality, any unmodelled feature in the data can be termed noise. It is justifiable to say that an exact theory should be able to predict every phenomenon including the instrument behaviour. Thus there is no need for a separate observational error term and the entire error in the data can be explained as theoretical error. On the other hand, it can also be argued that no theory is exact and any unmodelled phenomenon not predicted by the theory in use, is observational error. Thus there is no need for a separate theoretical error.

By substituting a Gaussian pdf for the prior pdf $p(\mathbf{m})$, we can derive an equation for the posterior pdf very easily. Note, however, that even under the assumption of

Gaussian prior pdfs, the posterior pdf is non-Gaussian due to the presence of the term $g(\mathbf{m})$ in the likelihood function in (6) (Tarantola 1987). For uniform prior pdfs, the maximum of the PPD coincides with the minimum of the error function $E(\mathbf{m})$. There exist several methods based on local and global search techniques that can be applied to find the minimum of the error function, which will also be the maximum of the PPD for uniform prior pdfs within a predefined search window. Therefore such methods are often referred to as MAP estimation methods. Note that a MAP estimator is a point estimator and does not provide a complete description of the PPD. Often the PPD is assumed to be Gaussian and a Hessian is computed at the MAP point whose inverse approximates the posterior covariance matrix (e.g. Menke 1984).

The expression for the PPD can thus be written as

$$\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}}) \propto \exp(-E(\mathbf{m}))p(\mathbf{m}), \tag{9}$$

or

$$\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}}) = \frac{\exp(-E(\mathbf{m}))p(\mathbf{m})}{\int d\mathbf{m} \exp(-E(\mathbf{m}))p(\mathbf{m})}, \tag{10}$$

where the domain of integration spans the entire model space.

Once the PPD has been identified as given by (10), the description of the solution of the inversion problem is given by the PPD. In reality we are faced with the problem of estimating the PPD in a large multidimensional model space. The most accurate way of evaluating the PPD is to compute the right-hand side of (10) at each point in model space (Tarantola and Valette 1982), i.e. evaluate the forward problem at each point in model space. This is, in general, a very time-consuming task unless the prior information helps us to restrict the model space to a small region.

Even if the PPD was known, there is no way to display it in a multidimensional space. Therefore several measures of dispersion and marginal density functions are often used to describe the answer. The marginal PPD of a particular model parameter, the posterior mean model and the posterior model covariance matrix are given by

$$\sigma(m_i|\mathbf{d}_{\text{obs}}) = \int \mathrm{d}m_1 \int \mathrm{d}m_2 \ldots \int \mathrm{d}m_{i-1} \int \mathrm{d}m_{i+1} \ldots \int \mathrm{d}m_M \sigma(\mathbf{m}|\mathbf{d}_{\text{obs}}), \tag{11}$$

$$\langle \mathbf{m} \rangle = \int \mathrm{d}\mathbf{m}\, \mathbf{m}\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}}) \tag{12}$$

and

$$\mathbf{C}'_{\text{M}} = \int \mathrm{d}\mathbf{m}(\mathbf{m} - \langle \mathbf{m} \rangle)\mathbf{m} - \langle \mathbf{m} \rangle)^{\text{T}} \sigma(\mathbf{m}|\mathbf{d}_{\text{obs}}). \tag{13}$$

All these integrals can be written in the following general form

$$\mathbf{I} = \int \mathrm{d}\mathbf{m} f(\mathbf{m}) \sigma(\mathbf{m}|\mathbf{d}_{\mathrm{obs}}). \tag{14}$$

The primary objective of this paper is to describe methods for evaluating these integrals efficiently for geophysical applications.

Methods of numerical integration

The marginal PPD, posterior covariance and mean model can be calculated by numerical evaluation of integrals of the type given in (14). We review four different approaches to evaluating these integrals: grid search, Monte-Carlo integration, Monte-Carlo importance sampling, and an approximate method based on multiple MAP estimation.

Grid search or enumeration

This approach is the simplest and the most straightforward. The entire region in the multidimensional model space in which the integrals need to be evaluated is discretized and divided into uniform grids. Then, the error function, such as that given by (7), is evaluated at each point in the model space. Thus the marginal posterior PPD, covariance and mean model can be evaluated using numerical integration techniques such as the trapezoidal rule. The integration result depends on the grid spacing and the method of numerical integration employed (Press *et al.* 1990). For most geophysical inversion problems, the model space is usually very large and the forward modelling is very computationally intensive. Since the grid-search technique requires evaluation of the error function (i.e. the forward calculation) at each point in the model space grid, it is impractical for most geophysical applications.

The Monte-Carlo method

Unlike the grid-search method of numerical integration in which the integrand is evaluated at points in a uniform grid over a multidimensional model space, Monte-Carlo integration schemes make use of pseudo random number generators. The integrand is evaluated at points chosen uniformly at random. Although there are some rules (Rubinstein 1981), in practice, it is difficult to estimate how many function evaluations will be required for accurate estimation of the integral.

Importance sampling

The Monte-Carlo method draws samples (random vectors) from a uniform distribution many of which do not contribute significantly to the integral. The

idea of importance sampling is to concentrate the selection of sample points from the regions that are the most important, i.e. which contribute the most to the integral. This requires that, unlike the pure Monte-Carlo method, the random vectors must be drawn from a non-uniform distribution. Such a non-uniform distribution cannot, however, be chosen arbitrarily. Hammersley and Handscomb (1964) and Rubinstein (1981) showed that for an unbiased estimation of the integral (14) with minimum variance, we require that the samples of $\mathbf{m}$ be drawn from the pdf $\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}})$. The estimation of the integral will be biased if the models are drawn from a distribution other than $\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}})$. For many geophysical applications, the prior pdf $p(\mathbf{m})$ is far from being sharply peaked and the PPD is clearly dominated by the likelihood function. Thus generating the models according to the prior pdf alone is not enough and importance sampling by drawing models according to the PPD is essential. This, however, is not a trivial task. The PPD, $\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}})$ as given in (10) contains an integral in the denominator which requires that the error function $E(\mathbf{m})$ be evaluated at each point in model space. Clearly there is no need for importance sampling if we know the value of the error function at each point in model space. We discuss below some ways of drawing models from $\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}})$ without evaluating the error function at each point in the model space. Once the models are drawn accordingly, the frequency distribution of each model parameter represents marginal PPDs for each model parameter.

Multiple MAP estimation

Several MAP estimation algorithms can be developed based on global optimization methods such as SA and GA. In order to reach a near-global minimum of an error function, such methods sample different parts of the model space. Several independent runs of SA and GA can be made; the sampled models must be weighted by their corresponding unnormalized PPD (9) and then integrals of the type given in (14) should be evaluated. The algorithms should be run repeatedly with different starting solutions until the estimates of the marginal PPD, posterior pdf and mean model do not change. This method assumes that the PPD is simple and well behaved and that by repeating several different runs, we are able to sample adequately the most significant portion of the PPD. In general, this approach will result in biased estimates of the integrals. However, the accuracy of the method will depend on the shape of the PPD, the optimization method used and the number of independent runs used in the optimization method.

Optimization methods

For many geophysical applications, the data and the model are related by non-linear functionals. The error function that defines the misfit between the observed and synthetic data have multiple minima. Therefore, most gradient-based methods are not appropriate for searching for the global minimum of an error function. The

objective of this paper is not to promote any particular method of finding the global minimum, but rather to cast the inverse problem into the statistical framework and then devise methods based on global-optimization methods to estimate uncertainties.

Global-optimization methods, such as SA and GA, have been applied to several geophysical inversion problems with the aim of finding the global minimum (e.g. Rothman 1985; Basu and Frazer 1990; Sen and Stoffa 1991; Stoffa and Sen 1991; Sen and Stoffa 1992; Sambridge and Drijkoningen 1992). Here we briefly review SA and some of its variants and GA, and show how they can be used as importance sampling tools.

Simulated annealing: the Gibbs' sampler

The basic concepts of SA are borrowed from problems in statistical mechanics which involve the analysis of properties of a large number of atoms in samples of liquids or solids. It draws an analogy between the parameter (model parameter) of an optimization problem and particles in an idealized physical system. A physical annealing occurs when a solid in a heat-bath is initially heated by increasing the temperature such that all the particles arrange themselves in the low-energy ground state where crystallization occurs. The crystallization process involves simulating the evolution of the physical system as it cools and anneals into a state of minimum energy. There are two basic computer algorithms of SA: Metropolis SA (Metropolis *et al.* 1953; Kirkpatrick, Gelatt and Vecchi 1983), and the heat-bath algorithm (Rothman 1986). Although both the algorithms simulate the same phenomenon, they differ substantially in the details of implementation.

Metropolis SA is a two-step procedure in which a model is drawn at random and then a decision is made whether to accept or reject it. The algorithm starts with a model chosen at random, say $\mathbf{m}_i$. Synthetic data are generated for this model and the error (also called energy) $E(\mathbf{m}_i)$ for this model is computed. Then $\mathbf{m}_i$ is perturbed to obtain a new model $\mathbf{m}_j$ for which the error $E(\mathbf{m}_j)$ is computed. If $\Delta E = E(\mathbf{m}_j) - E(\mathbf{m}_i)$, then if $\Delta E < 0$, $\mathbf{m}_j$ is always accepted. However if $\Delta E > 0$, the new model is accepted with a probability $P = \exp(-\Delta E/T)$, where T is a control parameter called temperature. The generation–acceptance process is repeated several times at a constant temperature. Then the temperature is lowered following a cooling schedule and the process is repeated. The algorithm is stopped when the error does not change after a sufficient number of trials.

Unlike Metropolis SA, heat-bath SA is a one-step procedure. This algorithm computes the relative probability of acceptance for each possible move before any random choice is made, i.e. it produces weighted selections that are always accepted. In this approach, the model parameters are discretized to a desired accuracy. Each model parameter is visited sequentially for all possible values of one model parameter while keeping the values of all other model parameters fixed. Then the following

Gibbs' pdf of a single model parameter, i.e. a conditional pdf, is evaluated

$$p(\mathbf{m}|m_i = m_{ij}) = \frac{\exp\left(-\dfrac{E(\mathbf{m}|m_i = m_{ij})}{T}\right)}{\sum_{j=1}^{N} \exp\left(-\dfrac{E(\mathbf{m}|m_i = m_{ij})}{T}\right)}, \tag{15}$$

where the subscript i (from 1 to M) refers to the model parameters and the subscript j (from 1 to N) refers to the values that the model parameter m_i can take. Thus, there are N^M possible models in the parameter space. E corresponds to the error or energy and T is the control parameter analogous to temperature which has the same dimension as that of the error or energy. A value is drawn from the above distribution and is always retained. The entire procedure is repeated for all M model parameters which involves $N \times M$ forward calculations. This constitutes one iteration. This procedure is repeated for a large number of iterations at a constant temperature, and then the temperature is lowered following a cooling schedule. The entire procedure is repeated until a low-energy state is attained.

Thus the Metropolis and the heat-bath SA algorithms differ substantially. It has been suggested that the heat-bath algorithm may be faster for models with a large number of parameters (Rothman 1986). One important difference between the two methods is that the heat-bath algorithm works with discrete values of one model parameter and Metropolis SA does not require that. Therefore for small increments in model parameter values, the heat-bath algorithm may be slow.

Although the two algorithms differ, they have one thing in common. Each can be modelled as a finite Markov chain that is irreducible and aperiodic. Consequently, it can be shown that at a constant temperature, after a large number of iterations, they attain an equilibrium or a steady-state distribution which is independent of the starting model. They are described in detail by Aarts and Korst (1989), Geman and Geman (1984) and Rothman (1986) and will not be repeated here. The equilibrium distribution at a temperature T attained by both Metropolis and heat-bath SA is given by the following Gibbs' pdf

$$p(\mathbf{m}) = \frac{\exp\left(\dfrac{-E(\mathbf{m})}{T}\right)}{\sum \exp\left(-\dfrac{E(\mathbf{m})}{T}\right)}, \tag{16}$$

where the sum is taken over all the models in the model space.

Thus we have achieved an important goal of the importance sampling discussed above, i.e. the models must be sampled according to a distribution such that we do

not need to evaluate the error function at each point in model space. The Markov chain analysis of SA guarantees that, in equilibrium, the models are sampled according to Gibbs' distribution. The proof, however, is asymptotic, meaning that we require an infinitely large number of iterations at a constant temperature. In practice, it is possible to attain such a distribution in a finitely large number of iterations, which is smaller than that required by enumeration. Below we will show this numerically for a geophysical problem.

Since SA samples models from the Gibbs' distribution, such a sampler is called a GS. The GS, based primarily on the heat-bath algorithm and some of its variants, has recently been applied to several problems outside the field of geophysics (e.g. Gelfand and Smith 1990; Gelfand, Smith and Lee 1992; Gilks and Wild 1992) in computing posterior marginal distributions. Gibbs' sampling based on heat-bath SA will involve sweeping through each model parameter several times at constant temperature. Similarly, a Gibbs' sampling based on Metropolis SA will involve repeating the model generation–acceptance procedure using the Metropolis rule a large number of times at a constant temperature. It may also be useful to make several independent GS runs using different starting models and random number seeds, to avoid the dependence on the computer generated random numbers. The two crucial parameters are the number of iterations for each independent run and the number of independent runs necessary to achieve convergence (Gelfand and Smith 1990). The first parameter is the number of iterations necessary to reach convergence of the error versus iteration curve. The second parameter can be determined from the time it takes to obtain stable estimates of the marginal density functions, mean and posterior covariances. In principle, the heat-bath and the Metropolis-rule-based GS are equivalent; but we prefer the latter over the former, primarily because the heat-bath-based GS works with discrete values for each model parameter and can be quite slow for a large number of model parameters where each model parameter is very finely discretized.

At this stage it is worthwhile re-examining the PPD given by (10). A comparison with (16) reveals that the PPD is essentially the Gibbs' pdf at temperature $T = 1$ (Tarantola 1987) . This means that to evaluate quantities such as marginal PPD, mean and covariance by Monte-Carlo importance sampling, we need to sample the models using a GS at a constant temperature of one. The frequency distribution of the model parameters give the marginal PPDs directly. We also note here that the original Metropolis *et al.* (1953) algorithm was actually invented as a Monte-Carlo importance-sampling technique (Hastings 1970) to study the equilibrium properties, especially ensemble averages, time evolution, etc., of very large systems of interacting components, such as molecules in a gas. Following Metropolis *et al.* (1953), we obtain samples $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{NM}$, such that the mean can be approximated by the following ergodic average

$$\langle \mathbf{m} \rangle \cong \frac{1}{NM} \sum_{j=1}^{NM} \mathbf{m}_i, \tag{17}$$

and the covariance will be given by

$$C_M \cong \frac{1}{NM}\sum_{j=1}^{NM}(\mathbf{m}_i - \langle\mathbf{m}\rangle)(\mathbf{m}_i - \langle\mathbf{m}\rangle)^T, \tag{18}$$

where NM is the number of models.

Genetic algorithm: the parallel Gibbs' sampler

Unlike SA, which is based on analogy with a physical annealing process, GAs are based on analogies with the processes of biological evolution: A GA works with a population of models that are coded in some suitable form and the algorithm seeks to improve the fitness (which is the measure of goodness-of-fit between data and synthetics for the model) of the population generation (iteration) after generation. This is accomplished principally by the genetic processes of selection, crossover and mutation. Excellent reviews of GAs are given by Goldberg (1989). Stoffa and Sen (1991), and Sen and Stoffa (1992) give detailed accounts of a GA for geophysical applications. A simple GA involves three operators, namely, selection, crossover and mutation, acting upon a finite population of binary coded models. The most commonly used selection process, called fitness proportionate selection involves computing the fitness $f(\mathbf{m})$ of each model and reproducing the models based on their probability of selection $p_s^{(i)}$ given by

$$p_s^{(i)} = \frac{f(\mathbf{m}^{(i)})}{\sum_{j=1}^{NM} f(\mathbf{m}^{(i)})}, \tag{19}$$

where NM is the total number of models in the population. Here the fitness may be defined as the negative of an error function.

Next a bit crossover between two models paired or mated via the selection process is carried out with a specified probability, p_x, called the crossover probability. A crossover position in the model is selected uniformly at random and the bits to the right of the crossover position are exchanged. Finally, mutation, which involves a random change of a bit within the model, is applied based on a mutation probability p_m. The three operations result in a new set of models, or a new generation. The models thus generated are evaluated again and the three operations are performed again. The process is repeated until convergence, i.e. when the mean fitness of the entire population approaches the highest fitness value in the population.

The GA described above is called a simple GA. Several modifications of the simple GA have recently been proposed. For example, a tournament selection method (Goldberg and Deb 1991) has been proposed as an alternative to the conventional fitness proportionate selection method. This involves selecting random pairs of models and computing their fitness. This fitness is compared and one of the two

models is accepted based on a probability called the tournament selection number, which is an adjustable parameter. On the other hand, we proposed (Stoffa and Sen 1991) two important modifications of the simple GA. They are as follows:

1. Replace the selection probability as given by (19) with a temperature-dependent selection probability given by

$$p_s^{(i)} = \frac{\exp\left(\dfrac{f(\mathbf{m}^{(i)})}{T_s}\right)}{\displaystyle\sum_{j=1}^{NM} \exp\left(\dfrac{f(\mathbf{m}^{(i)})}{T_s}\right)}, \tag{20}$$

where T_s is a control parameter analogous to temperature in SA, which we call selection temperature. Sen and Stoffa (1992) describe, in detail, how the selection temperature controls the stretching of the fitness function and hence it can be used to control the convergence of the GA.

2. Use an update probability to decide whether a model from a previous generation should be used in place of a current model. Once a new model's fitness has been evaluated, it is compared to the fitness of a model from the previous generation selected uniformly at random and used only once. If the current model's fitness is greater, the current model is always kept. If it is less, the previous generation's model replaces the current model with a specified update probability p_u. This update probability plays the role of acceptance probability as used in Metropolis SA. In fact, Stoffa and Sen (1992) suggested that the Metropolis rule can be used instead of the update probability.

Unlike SA, GAs are not based on rigorous mathematical models and do not have any proof of convergence, asymptotic or otherwise. Recently several attempts have been reported (e.g. Goldberg and Segrest 1987; Eiben, Aarts and Van Hee 1991; Horn 1993; Davis and Principe 1991) to model GAs by means of Markov chains. Most of these are either restrictive or lack mathematical rigour. Vose and Liepins (1991) first proposed a meaningful mathematical model of a simple GA, which was analysed using finite Markov chains (Nix and Vose 1992). The Vose and Liepins (1991) model recognizes the collection of models in a generation to be a state which undergoes the processes of selection, crossover and mutation to generate a new population of models or a new state. The genetic processes of selection, crossover and mutation were explicitly taken into account in the definition of a transition probability matrix that characterizes the transition from one state to another. The Markov chain was assumed ergodic because of the non-zero mutation rate. However, determination of an equilibrium distribution in its explicit form as a function of fitness and algorithm parameters, such as population size, mutation and crossover rate remains an unsolved task. Nix and Vose (1992) further analysed the transition probability in the limit that the population size tends to infinity.

The convergence behaviour of a GA can be modelled by incorporating into a simple GA, some of the elements of SA. At this stage we will re-examine the Stoffa

and Sen (1991) algorithm and consider a simple, albeit less rigorous, model of GA with the aim of comparing it with classical SA.

In order to describe the algorithm we will follow a particular member of the population through different generations. In this algorithm a trial consists of the processes of selection, crossover, mutation and update. Thus this is a two-step procedure. First a model is generated by the processes of selection, crossover and mutation. Then its fitness is compared with that of a model from the previous generation. The new model is always accepted if the fitness of the new model is higher. When the fitness is lower, it is accepted using the well-known Metropolis acceptance criterion as described in the context of SA. Just as in Metropolis SA, we can express the probability of transition from a model i to a model j as a product of two probabilities, i.e.

$$P_{ij} = G_{ij} A_{ij}, \tag{21}$$

where G_{ij} is the generation probability and A_{ij} is the acceptance probability. These are given by the equations

$$G_{ij} = G_j = f_1(p_x) f_2(p_m) p_s^{(j)} = f_1(p_x) f_2(p_m) \frac{\exp\left(\dfrac{E(\mathbf{m}_j)}{T_s}\right)}{\displaystyle\sum_{k=1}^{n} \exp\left(-\frac{E(\mathbf{m}_k)}{T_s}\right)}, \tag{22}$$

and

$$A_{ij} = \exp\left(-\frac{(E(\mathbf{m}_j) - E(\mathbf{m}_i))^+}{T_u}\right), \tag{23}$$

where $a^+ = a$ if $a > 0$ and $a^+ = 0$, otherwise, and where $a = E(\mathbf{m}_j) - E(\mathbf{m}_i)$. f_1 and f_2 are functions of the crossover and mutation probability, respectively. While the operation of crossover combines genetic information from two models, mutation is a true random walk, and we may consider the product of f_1 and f_2 to be a generator of random models for non-zero mutation probability. Note the existence of the control parameter temperature in both the selection (T_s) and acceptance (T_u) probabilities and that they can have different values. Note also that in our analysis in (21)–(23) we use an error function (that we try to minimize) rather than a fitness function, in order to compare the method with SA. That is, we have $f(\mathbf{m}_i) = -E(\mathbf{m}_i)$.

The model described by the above equations is very similar to that used in describing Metropolis SA except that the generation probability in Metropolis SA is uniform while that used in the GA is biased as given by (19). Thus the GA selection process is more global than Metropolis SA. It is intuitively obvious that such an approach is desirable to avoid the high rejection to acceptance ratio in Metropolis SA. This, however, causes problems when we attempt to describe the GA in terms of Markov chains. Although for a finite mutation rate the Markov chain may be considered ergodic (i.e. there is finite probability of reaching every state from every

other state), the selection probability as given by (19) is controlled by the partition function in the denominator which takes on different values for different generations. This can possibly be side-stepped by saying that for a given mutation rate and a very large population size, the sum in the denominator remains nearly constant at different generations. Under such highly restrictive (albeit impractical) conditions, an analysis similar to that used in Metropolis SA can be carried out and an expression for the steady state distribution (which again turns out to be a Gibbs' distribution) can be achieved. However, the main objective of this exercise is to relate GA with SA.

Now consider the case where the selection temperature T_s is very high. Models are then selected uniformly at random from the population. In such a situation, if the crossover rate is zero and the mutation rate is high, the Stoffa and Sen (1991) GA is the same as running several Metropolis SAs in parallel. However, if we consider a high T_s, moderate crossover ($p_x = 0.5$) and moderate mutation ($p_m = 0.5$), then the crossover mixes models between different parallel SAs and mutation allows for a local neighbourhood search. Strictly speaking, such an approach will cause a breakdown of the symmetry property of the generation matrix G_{ij} which is essential to prove the convergence to a stationary distribution. However, due to the presence of high mutation, we expect that when such an algorithm is run for a large number of generations at a constant acceptance temperature, it will attain a Gibbs' distribution at the acceptance temperature. We call this approach a parallel Gibbs' sampler (PGS) and compare the results from PGS and GS with a geophysical example in the following section. Thus our recipe for PGS is as follows. Run several repeat GAs with different initial populations with T_s very large and an acceptance temperature T_u of 1.0 with $p_x = 0.5$ and $p_m = 0.5$. Note that due to this high mutation rate, the population will never become homogeneous, but a single GA can be stopped when the variance of the fitness of the population becomes nearly constant. The number of parallel GAs will be determined by examining when the estimates of marginal PPD, mean and covariance become stable. The frequency distribution of the model parameters sampled by these runs will give the marginal PPDs directly and the posterior mean and covariance will be computed by the formulae given in (17) and (18).

Very fast simulated reannealing

Ingber (1989) proposed several modifications of conventional SA and proposed a new algorithm called very fast simulated reannealing (VFSR). Ingber's (1989) algorithm can be described as follows.

Let us assume that a model parameter m_i at iteration (annealing step or time) k is represented by $m_i^{(k)}$ such that

$$m_i^{\min} \leq m_i^{(k)} \leq m_i^{\max}, \tag{24}$$

where $m_i^{\min}$ and $m_i^{\max}$ are the minimum and maximum values of the model parameter m_i. This model parameter value is perturbed at iteration $(k+1)$ by using the

following relationship:

$$m_i^{(k+1)} = m_i^{(k)} + y_i(m_i^{\max} - m_i^{\min}), \tag{25}$$

such that

$$y_i \in [-1,\ 1] \tag{26}$$

and

$$m_i^{\min} \leq m_i^{(k+1)} \leq m_i^{\max}, \tag{27}$$

where y_i is generated from the following distribution

$$g_T(y) = \prod_{i=1}^{NM} \frac{1}{2(|y_i| + T_i)\ln\left(1 + \frac{1}{T_i}\right)} = \prod_{i=1}^{NM} g_{Ti}(y_i), \tag{28}$$

which has the following cumulative probability

$$G_{Ti}(y_i) = \frac{1}{2} + \frac{\mathrm{sgn}(y_i)}{2} \frac{\ln\left(1 + \frac{|y_i|}{T_i}\right)}{\ln\left(1 + \frac{1}{T_i}\right)}. \tag{29}$$

Ingber (1989) showed that for such a distribution a global minimum can be statistically obtained by using the following cooling schedule:

$$T_i(k) = T_{0i} \exp\left(-c_i k^{1/NM}\right), \tag{30}$$

where T_{0i} is the initial temperature for model parameter i and c_i is a parameter to be used to control the temperature schedule and help tune the algorithm for specific problems.

The acceptance rule of the algorithm is the same as that used in the Metropolis rule. The algorithm described so far can be called very fast simulated annealing (VFSA). Ingber (1989) further showed how the cooling schedule can be changed dynamically based on the sensitivity of the error to model and called it VFSR.

In the following we use VFSA in one of our applications together with Metropolis SA and GA. These are used as MAP estimators and several runs are made with different starting solutions and are used in the evaluation of the integrals (14). We weight each model by the corresponding un-normalized PPD and approximate the integrals (11), (12) and (13) by corresponding discrete sums.

Numerical examples

Inversion of noisy synthetic vertical electric sounding data

Here we investigate the validity of different numerical integration schemes using an example from vertical electric sounding (VES) data. The VES data were chosen for

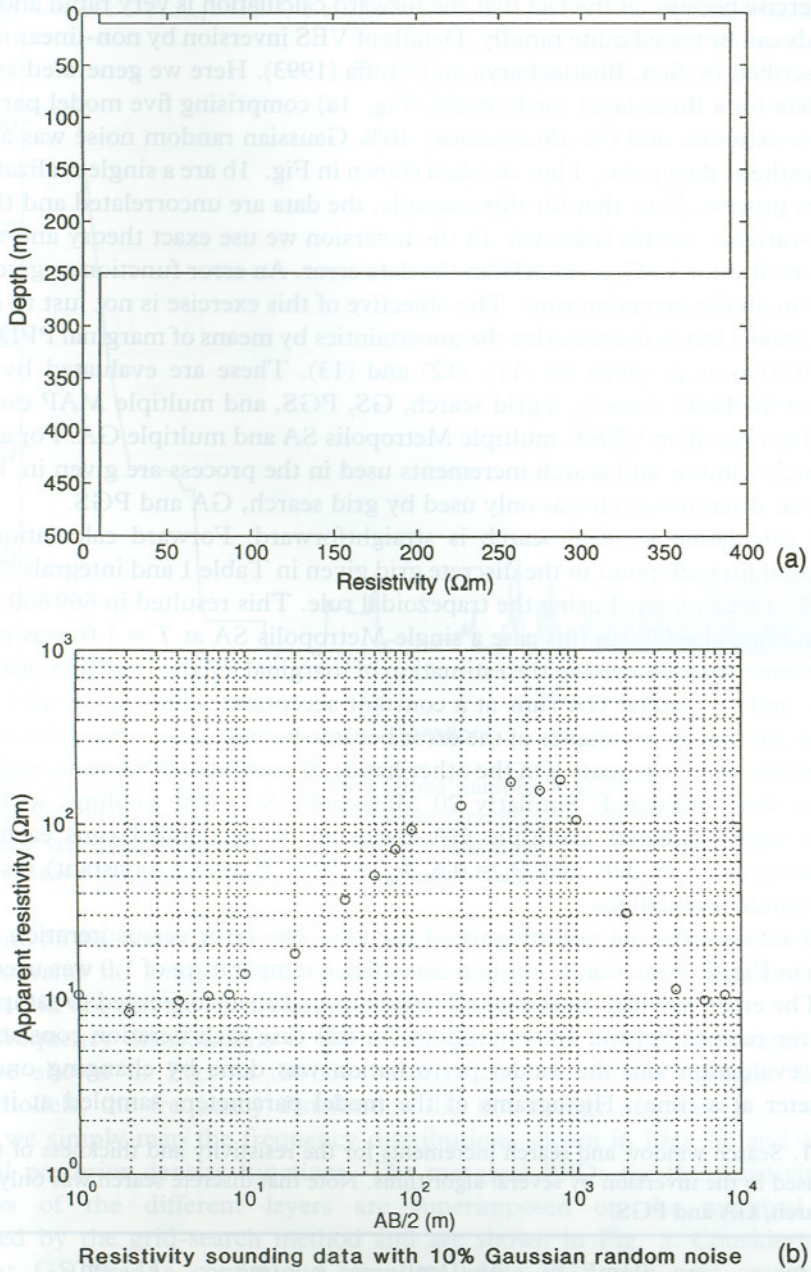

Figure 1. (a) A three layer 1D resistivity model used to generate synthetic Schlumberger vertical electric sounding data to test different sampling based uncertainty estimation algorithms. (b) Resistivity sounding data for the model shown in (a) with 10% Gaussian random noise added to each data sample.

this exercise because of the fact that the forward calculation is very rapid and several methods can be tested quite rapidly. Details of VES inversion by non-linear methods are described by Sen, Bhattacharya and Stoffa (1993). Here we generated synthetic VES data for a three-layer earth model (Fig. 1a) comprising five model parameters (three resistivities and two thicknesses). 10% Gaussian random noise was added to each synthetic data point. Thus the data shown in Fig. 1b are a single realization of a random process. Note that for this example, the data are uncorrelated and the prior data covariance matrix is known. In the inversion we use exact theory and thus the only contribution to $\mathbf{C}_D$ comes from the data error. An error function as given in (7) is used in all the inversion runs. The objective of this exercise is not just to obtain a best fit model but to characterize the uncertainties by means of marginal PPDs, mean and covariances as given by (11), (12) and (13). These are evaluated by several different methods, namely, a grid search, GS, PGS, and multiple MAP estimation algorithms based on VFSA, multiple Metropolis SA and multiple GA. For all these, the search window and search increments used in the process are given in Table 1. Note that discrete search was only used by grid search, GA and PGS.

The calculation by grid search is straightforward. Forward calculations were performed for each point in the discrete grid given in Table 1 and integrals (11), (12) and (13) were evaluated using the trapezoidal rule. This resulted in 669 600 forward calculations. The GS (in this case a single Metropolis SA at $T = 1.0$) was run until convergence was achieved in the estimation of marginal PPDs. For PGS, we used 50 models and 10 parallel GA runs at a constant acceptance temperature of 1.0. Each GA was run until the variance of the error became homogeneous. We will see that 10 parallel runs were adequate. On the other hand, 20 multiple VFSA runs (with 12 000 models) were evaluated. Similarly 20 Metropolis SA (with cooling) with 20 000 models were evaluated. Multiple GA consisted of 10 parallel runs each with a population of 50 models and $P_x = 0.6$, $P_m = 0.01$, $P_u = 0.9$ (constant) resulted in 22 000 model evaluations.

First we describe the performance of the GS. The error versus iteration curve is shown in Fig. 2. Note that a constant acceptance temperature of 1.0 was used in this case. The error was high in the initial iterations and then converged to sampling the low error regions beyond 5000 iterations (in this case each iteration consists of one model evaluation and the model perturbation was done by changing one model parameter at a time). Histograms of the model parameters sampled at iterations

Table 1. Search window and search increments for the resistivity and thickness of the three layers used in the inversion by several algorithms. Note that discrete search was only used by grid search, GA and PGS.

ρ_{min} (Ω m)	ρ_{max} (Ω m)	$\Delta\rho$ (Ω m)	h_{min} (m)	h_{max} (m)	Δh (m)
5.0	15.0	1.0	1.0	20.0	1.0
300.0	450.0	10.0	150.0	400.0	10.0
1.0	20.0	1.0	—	—	—

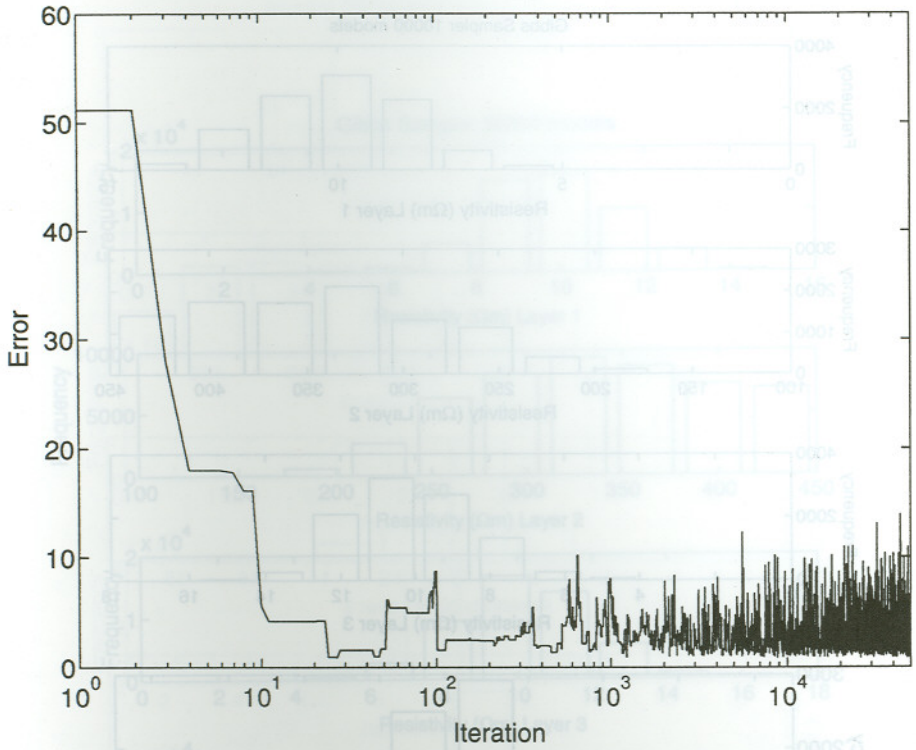

Figure 2. Error versus iteration curve for a Gibbs' sampler at a constant temperature of 1.0 in the inversion of VES data with 10% Gaussian random noise shown in Fig. 1. The error was initially very high but beyond iteration 5000, it sampled low error regions of the model space. Here each iteration corresponds to one model evaluation.

10 000 and 50 000 are shown in Figs 3a and b respectively. Beyond iteration 50 000, the histograms no longer changed and repeating the GS with different starting solutions did not change the frequency distribution either.

The frequency distributions of the model parameters at different stages of the PGS are shown in Figs 4a (24 000 models) and 4b (80 000 models). Beyond 80 000 models, the frequency distribution did not change any more.

Next we simply map the frequency distributions shown in Figs 3b and 4b into marginal posterior density functions. The marginal PPDs for the resistivity and thickness of the different layers are superimposed on the marginal PPD computed by the grid-search method and are shown in Fig. 5. Considering the fact that GS used a continuous sampling while PGS and grid search used discrete sampling between the maximum and minimum values of model parameters given in Table 1, the match between the three estimators is good. Note that the GS results were binned at the same discrete interval as that used by PGS and grid search.

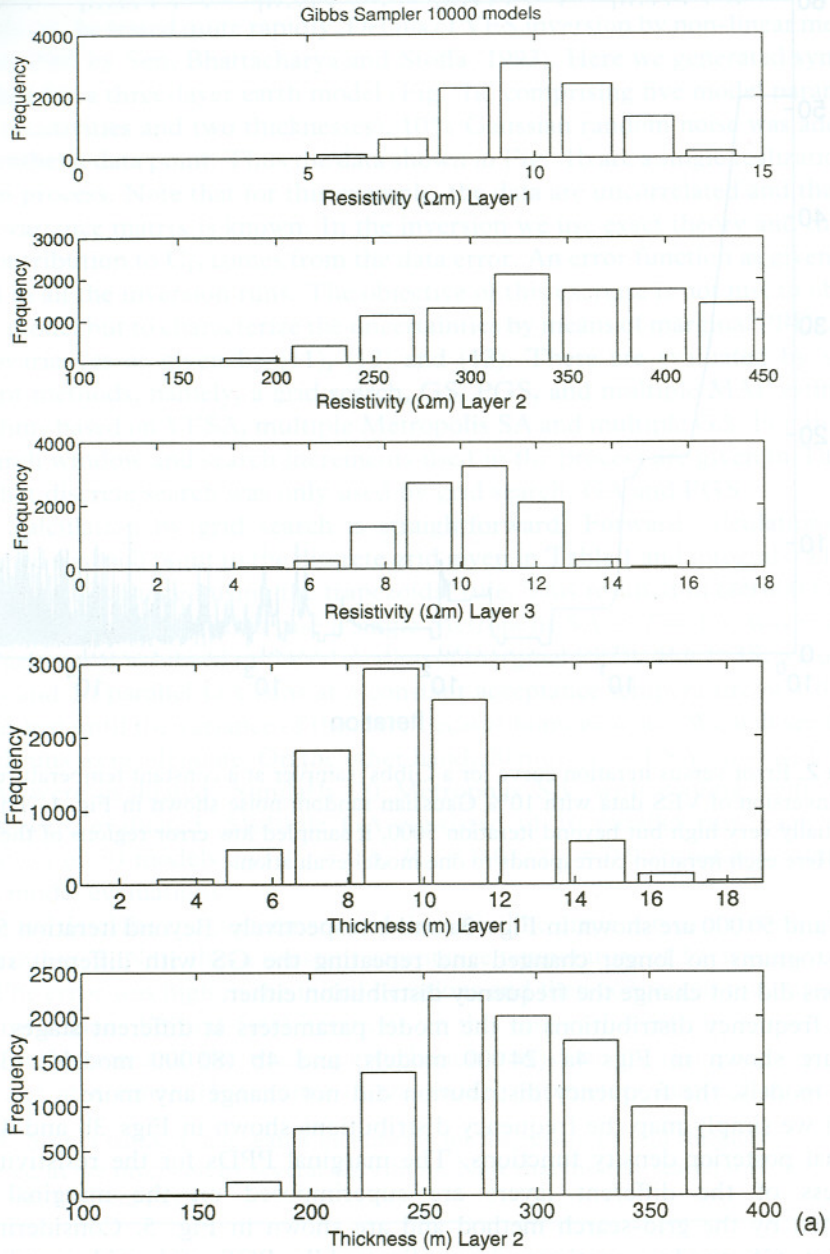

Figure 3. Histograms of the model parameters sampled by a Gibbs' sampler at $T = 1.0$ after (a) 10 000 and (b) 50 000 model evaluations. Note that histograms at the two iterations look very different. However, beyond iteration 50 000, the histograms did not change any more indicating convergence to the stationary distribution.

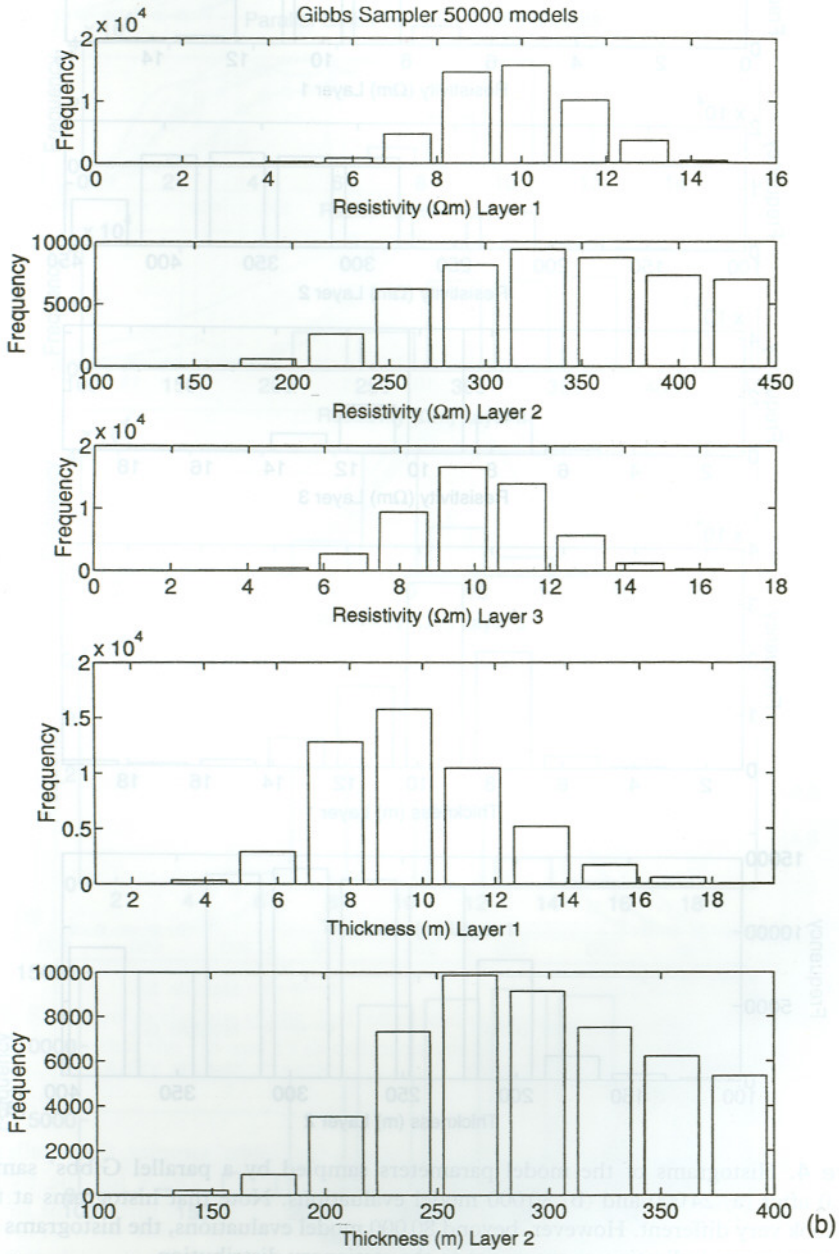

Figure 3. (contd)

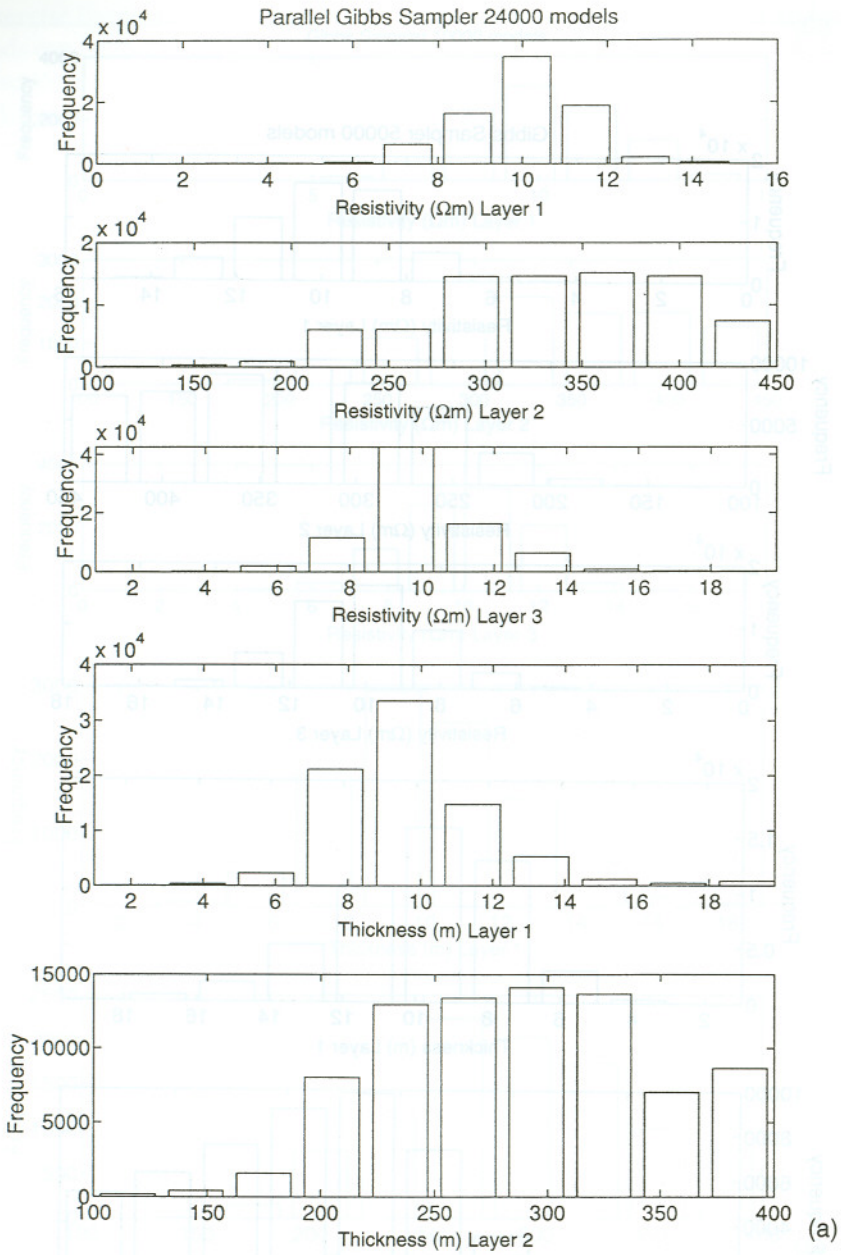

Figure 4. Histograms of the model parameters sampled by a parallel Gibbs' sampler at $T = 1.0$ after (a) 24 000 and (b) 80 000 model evaluations. Note that histograms at the two stages look very different. However, beyond 80 000 model evaluations, the histograms did not change any more, indicating convergence to the stationary distribution.

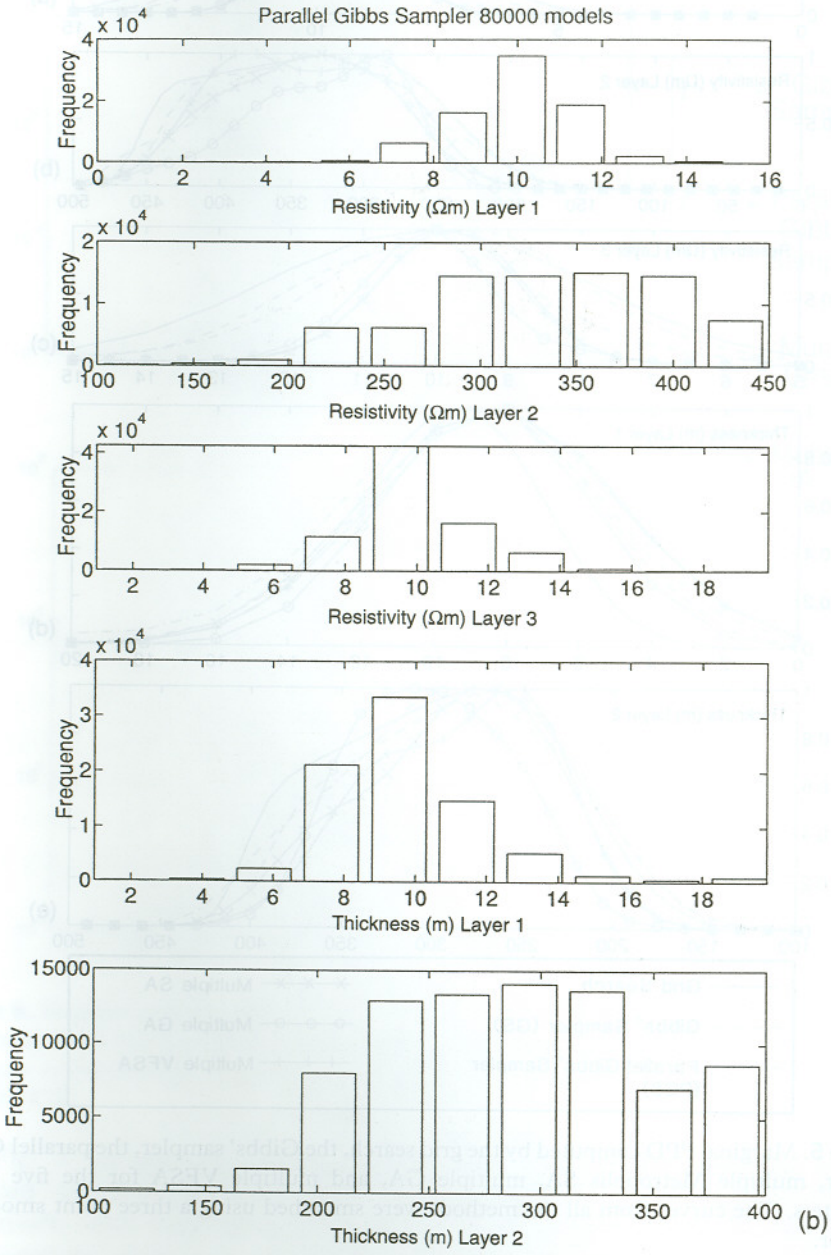

Figure 4. (contd)

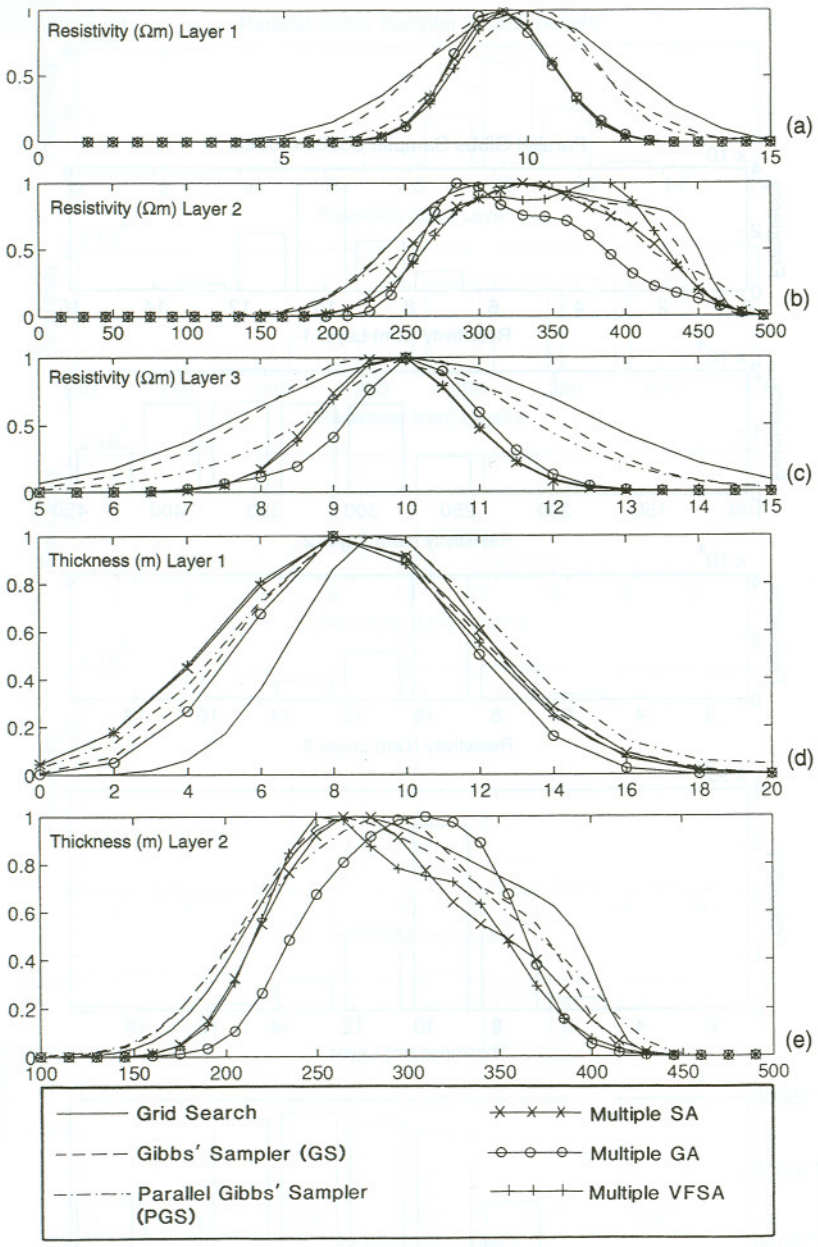

Figure 5. Marginal PPD computed by the grid search, the Gibbs' sampler, the parallel Gibbs' sampler, multiple Metropolis SA, multiple GA, and multiple VFSA for the five model parameters. The curves from all the methods were smoothed using a three point smoothing operator.

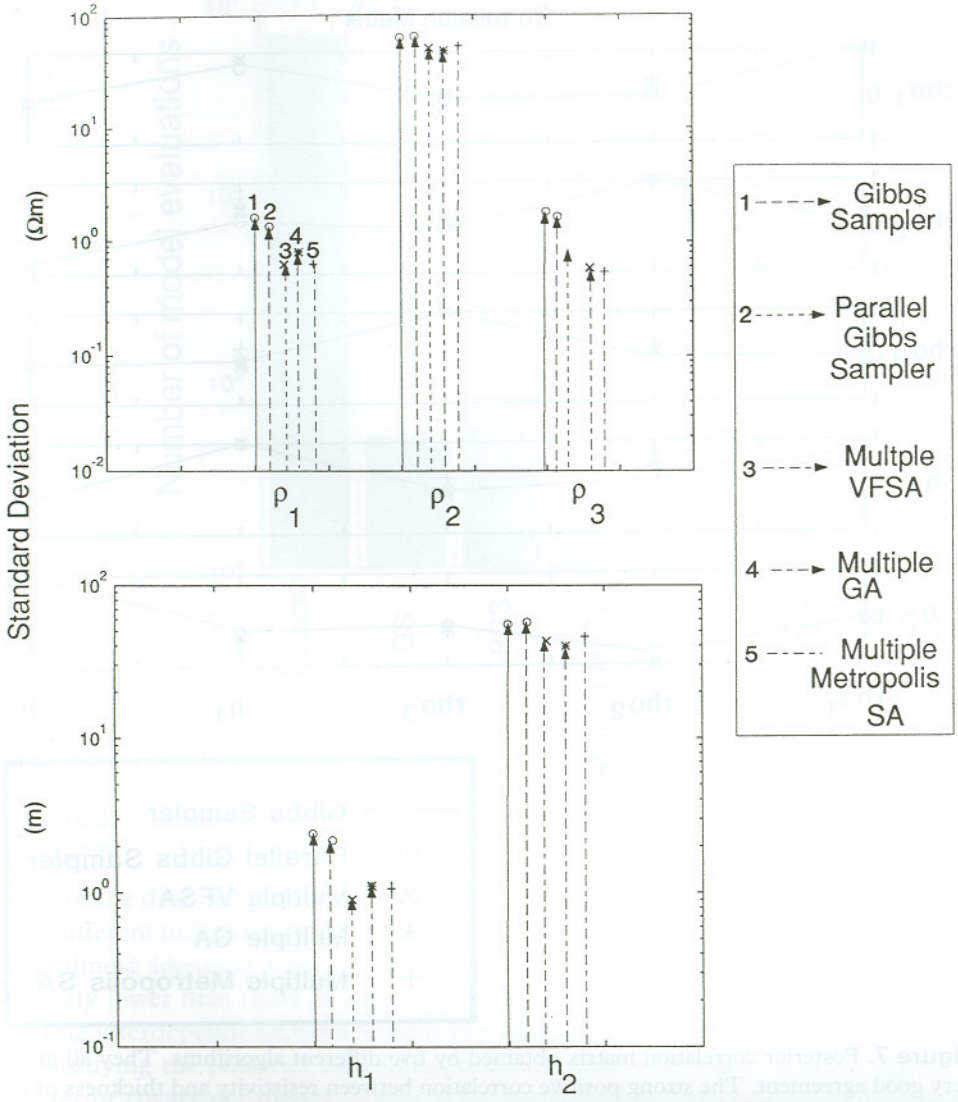

Figure 6. Standard deviations of the derived model parameters as estimated by five different algorithms. Note that the GS and PGS estimates are almost identical. However, the standard deviations estimated by multiple MAP algorithms are slightly lower than those estimated by GS and PGS.

In Fig. 5, we also compare the marginal PPDs for the five model parameters evaluated by multiple Metropolis SA, multiple GA and multiple VFSA, respectively, with those obtained from the GS, PGS and grid search. We notice that overall the width of the distribution obtained by the multiple MAP algorithms, is slightly narrower than the width obtained by the Gibbs' sampler. This means that

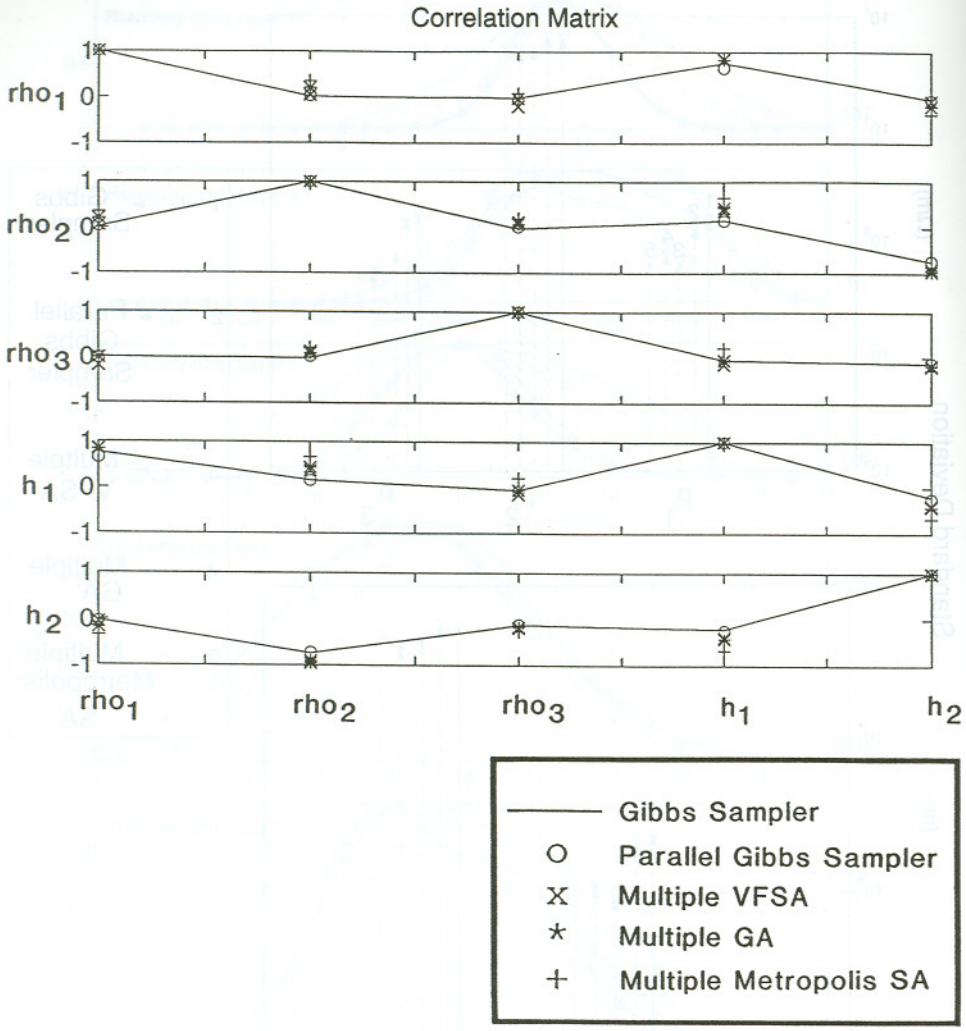

Figure 7. Posterior correlation matrix obtained by five different algorithms. They all are in very good agreement. The strong positive correlation between resistivity and thickness of the first layer, and the strong negative correlation between resistivity and thickness of the second layer are well estimated by all the algorithms.

the posterior variance of these parameters are slightly underestimated by the multiple MAP algorithms. This is to be expected because MAP algorithms are biased towards sampling densely around the maximum of the PPD.

We also note that the plots of the marginal PPDs obtained by all five methods show that the widths of the marginal PPD for the resistivity and thickness of the second layer are generally wide, indicating large uncertainties in the estimated values of these parameters. The standard deviations of different model parameters (square

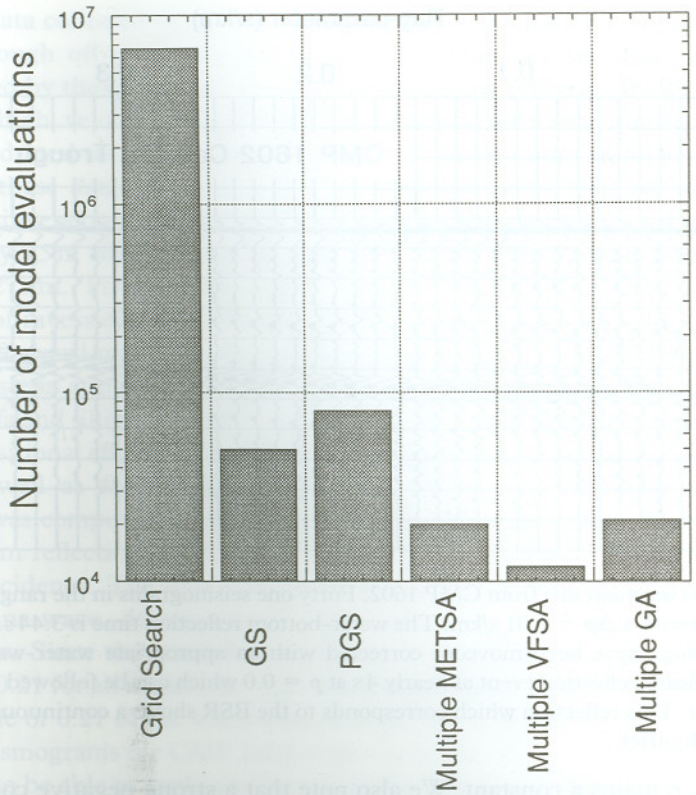

Figure 8. A chart showing the number of forward model evaluations required by the different algorithms.

root of the diagonal element of the posterior model covariance matrix) estimated by the different methods are shown in Fig. 6. The values estimated by the GS and PGS are almost identical while those obtained by the multiple MAP algorithms are slightly lower than those obtained by the GS and PGS.

The interdependence between different model parameters can be best understood by studying the posterior correlation matrix, which is obtained by normalizing the posterior covariance matrix. A plot of the correlation matrix obtained by each of the five methods is shown in Fig. 7. Notice that even though the multiple MAP algorithms slightly underestimated the variances, the correlation matrices obtained by all five methods are nearly identical. In particular, notice the strong positive correlation between the resistivity and thickness of the first layer (> 0.9). This means that if the resistivity and thickness of the first layer are both either increased or decreased, the apparent resistivity curve will not change. This is a classic example of the so-called equivalence problem in vertical electrical sounding interpretation (Koefoed 1979). For this layer, the parameters can be changed (within the variance limits) such that the ratio of the thickness and resistivity, also called the longitudinal

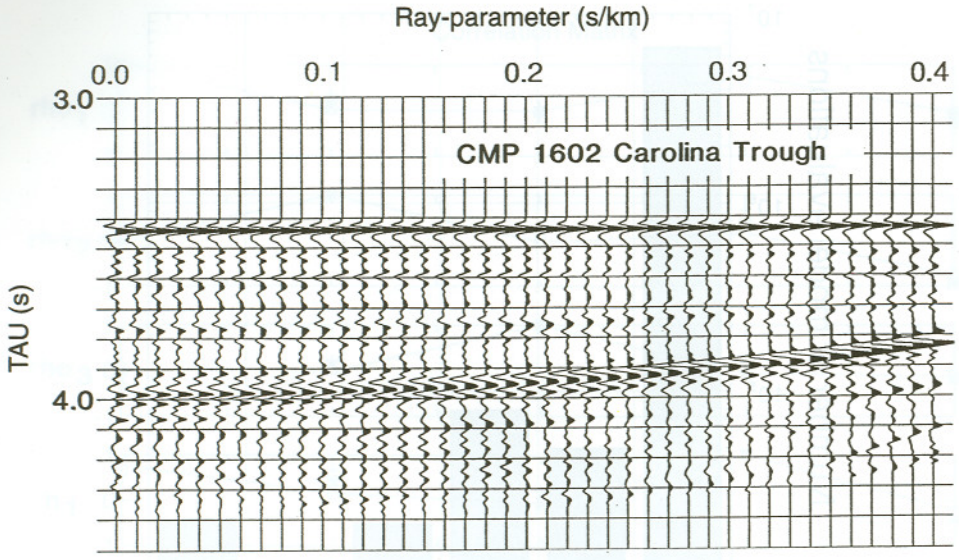

Figure 9. (τ,p) seismograms from CMP 1602. Forty one seismograms in the range (0.0–0.4 s/km) are shown with a $\Delta p = 0.01$ s/km. The water-bottom reflection time is 3.44 s. The water-bottom reflections have been moveout corrected with an appropriate water-wave velocity. Notice a significant reflection event at nearly 4 s at $p = 0.0$ which can be followed along all the ray parameters. This reflection which corresponds to the BSR shows a continuous increase in amplitude with offset .

conductance, remains a constant. We also note that a strong negative correlation is obtained between resistivity and thickness of the second layer. This means that by lowering the resistivity and increasing the thickness or vice versa, the apparent resistivity curve will not be changed. In this case, it is the transverse resistance of the layer, given by the product of resistivity and thickness of the layer, that remains a constant.

Finally we compare the computational requirements of different methods by noting the number of forward modellings, i.e. the number of error function evaluations required, as shown in Fig. 8. The PGS, GA and grid search used a discrete search while the other methods were allowed to sample continuously within the predetermined search window for each model parameter. The grid search required the most computation and was computationally intensive even for a five-parameter problem. The GS and PGS required much less computation time and provided very accurate estimates. The multiple MAP runs were significantly faster than GS and PGS, but the results were not as accurate as those obtained by GS and PGS. However, even though the multiple MAP runs slightly underestimated the variances, the correlation values were quite accurate.

Inversion of seismic waveform data from the Carolina trough

We now describe an application of the GS in the inversion of marine seismic

waveform data collected by the University of Texas Institute for Geophysics in the Carolina trough off the east coast of the U.S. The data sets from the area are characterized by the occurrence of a bottom simulating reflector (BSR). The BSR is caused by high velocity methane hydrate above a reflecting interface and non-methane hydrated sediment below. Wood (1993) carried out careful processing and analysis of these data sets using linear inversion methods. Here we only show an example of inversion for CMP 1602 using a GS. The data consisted of 240 traces separated by 25 m with a sampling interval of 4 ms. The nearest source–receiver offset was 275 m. The original data are collected in the offset–time (x,t) domain. However, for inversions, it is desirable that the data are converted to the (τ,p) domain so that the forward modelling can be done rapidly (Sen and Stoffa 1991). Different processing steps applied to the data are described in detail by Wood (1993) and Wood, Stoffa and Shipley (1994). First the data were static shifted to correct for some of the acquisitional effects such as feathering, varying streamer depths, and electronic delays, as well as from slight sea-floor dip. Next the water-bottom reflection coefficient was computed by the method outlined by Stark (1986). The value of the water-bottom reflection coefficient was found to be 0.21 ± 0.02 (Wood 1993) for all angles of incidence. The data were extrapolated to zero offset and large offset and then a plane-wave decomposition was applied to the data to obtain (τ,p) seismograms. Since the water-bottom reflection coefficient was found to be nearly constant at 0.21 for all angles, the (τ,p) transformed seismograms were scaled to have an amplitude of 0.21 of the water-bottom reflection at all ray-parameters. A plot of the (τ,p) seismograms for CMP 1602 is shown in Fig. 9.

In order to be able to apply any inversion algorithm, we require an estimate of the source function and layer parametrization. We applied a recently developed mean field annealing based deconvolution algorithm (Calderon and Sen, paper presented at 3rd SIAM Conference, San Antonio, TX, USA, 1995) to derive source function and reflectivity spike series from the (τ,p) seismograms shown in Fig. 9. The algorithm requires an initial guess of the source wavelet and iteratively updates the source function and the reflectivity series to arrive at a final estimate of the source function and the reflectivity sequence. Figure 10a shows initial wavelets at each ray-parameter, Fig. 10b shows the final wavelets and Fig. 10c shows the reflectivity sequence. The seismograms obtained from the convolution of wavelets in Fig. 10b with the spikes in Fig. 10c were subtracted from the seismograms shown in Fig. 9 to obtain residual seismograms as shown in Fig. 10d. These residuals show the data not modelled with the convolution model. Note that the residuals mostly show incoherent energy with very small amplitude representing noise present in the data. Thus these residuals can also serve as an estimate of a diagonal data covariance matrix.

We averaged ten wavelets in the range 0–0.09 s/km to obtain a single wavelet for use in the inversion (Fig. 11d). Also ten spike series in the range 0–0.09 s/km were averaged to obtain the spike series shown in Fig. 11c. These spikes were used to identify layer boundaries and their two-way normal-incidence reflection times. We

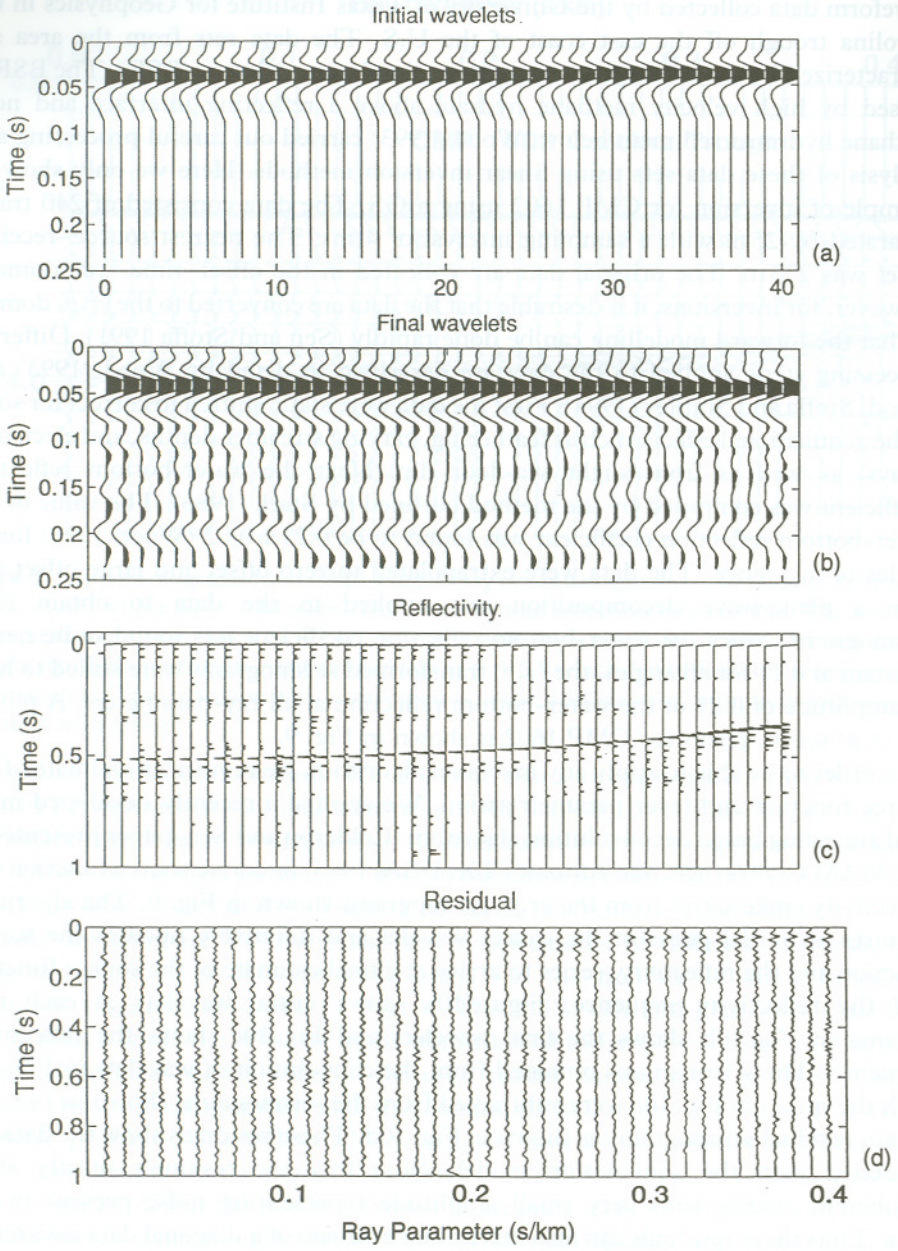

Figure 10. (a) Initial wavelets used in the deconvolution; (b) final wavelets; (c) reflectivity sequence, determined by the deconvolution algorithm; and (d) residual seismograms obtained from the result of the deconvolution.

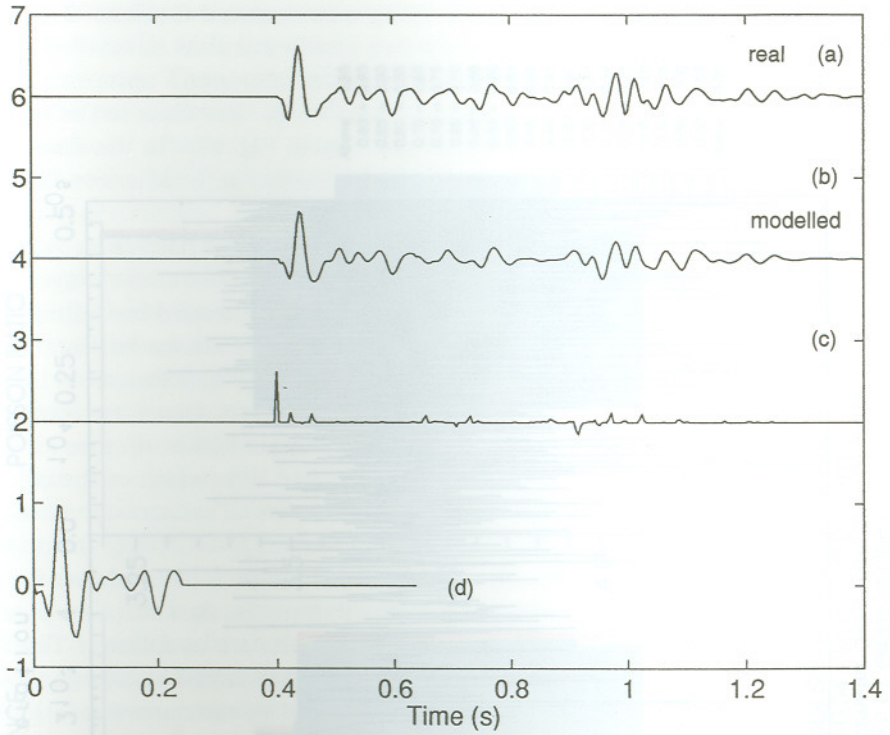

Figure 11. (a) Average of ten seismograms; (b) average of ten modelled seismograms; (c) average of ten reflectivity sequences; and (d) average of ten final wavelets, in the range 0.0–0.09 s/km.

obtained eight layers in the process, which are modelled using four parameters: compressional-wave velocity, impedance, Poisson's ratio and two-way reflection time (layer thickness). The two way times for the reflection events were assigned a ± 2 sample search (± 0.008 s) for each layer. The search window for the compressional wave velocity, impedance and Poisson's ratio were determined by our prior knowledge of the geology of the area, a detailed traveltime analysis of several lines carried out by Wood (1993) and nearby drilling data. The prior model pdf was uniform for each parameter within a predetermined window. The search range for each model parameter was different and was determined from our prior knowledge. For example, the drilling data from a nearby area revealed that the sea-floor is a density discontinuity rather than a velocity discontinuity.

Next, in order to determine the data covariance matrix, we used (τ,p) seismograms from 20 CMPs and aligned the sea-floor reflection. A mean and a variance were computed at each time sample for each ray-parameter seismogram. The variances thus computed were used in the diagonal data covariance matrix (C_D). The assumption made here was that for a 1D earth, the seismograms for each ray parameter should be identical except for random noise in the measurement. We chose

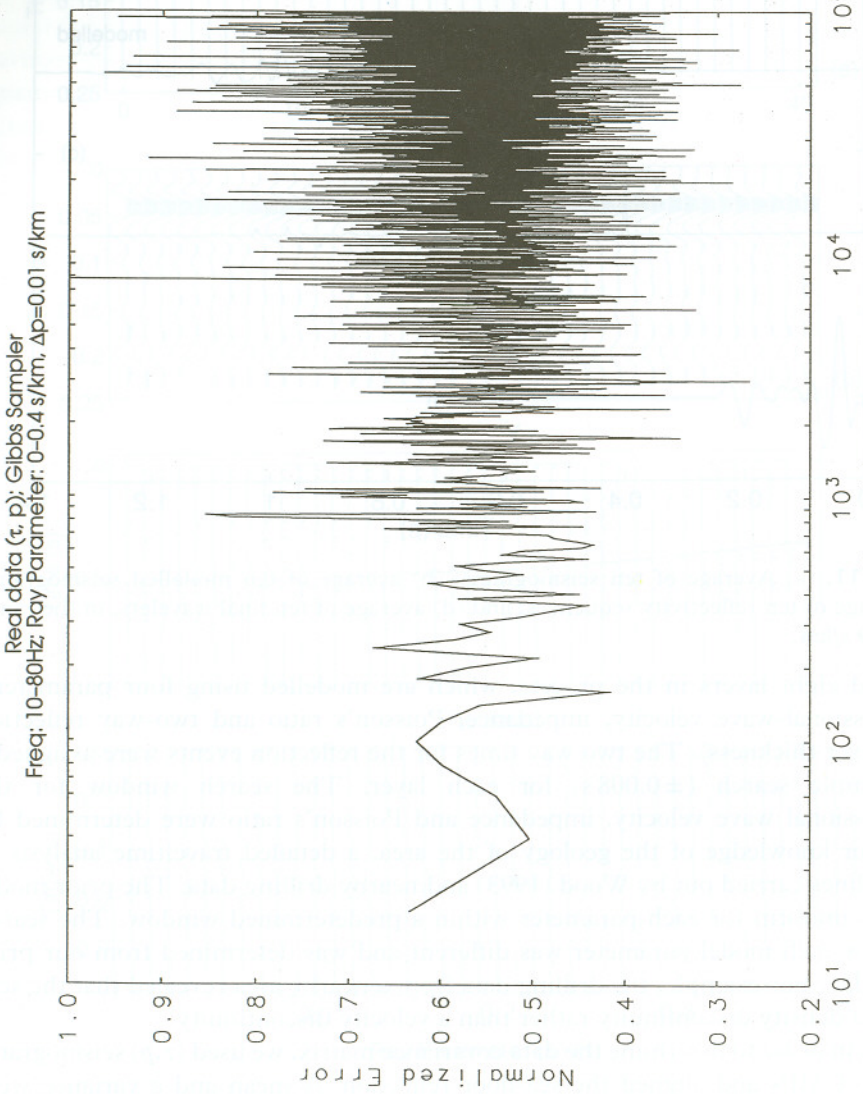

Figure 12. Error versus iteration plot for the Gibbs' sampler run.

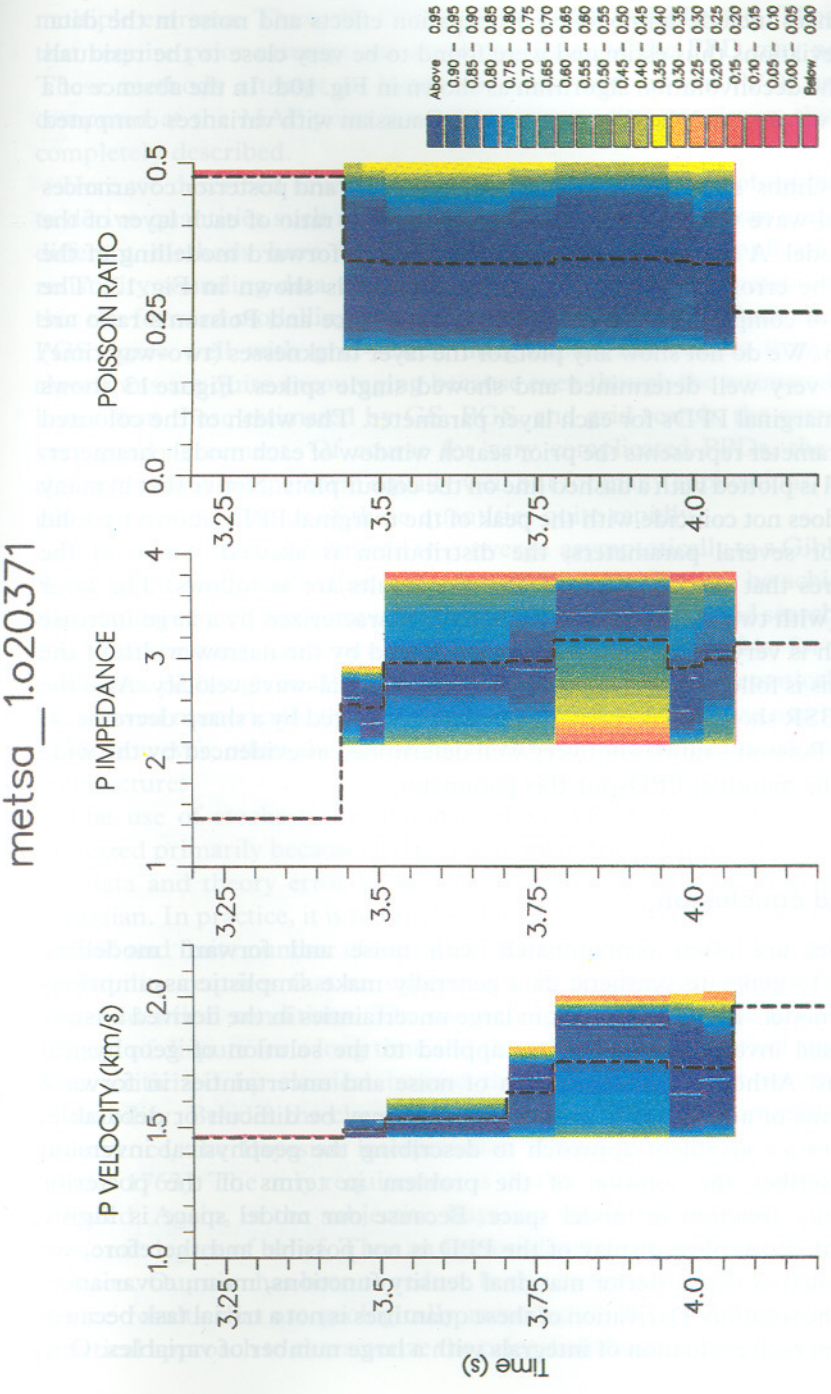

Figure 13. Marginal PPDs for compressional-wave velocity, impedance and Poisson's ratio as determined by the Gibbs' sampler. The dashed line corresponds to the mean model and the solid line corresponds to the peak of the marginal PPD of each model parameter.

20 CMPs that were from a very nearly flat-lying region around CMP 1602. Further, the sea-floor reflections were aligned to remove small variations in sea-floor depth. Thus, the variances include unmodelled propagation effects and noise in the data. The standard deviations thus estimated were found to be very close to the residuals obtained from the deconvolution algorithm as shown in Fig. 10d. In the absence of a better estimate, we assumed the data errors to be Gaussian with variances computed as above.

We now use a Gibbs' sampler to estimate marginal PPDs and posterior covariances of compressional-wave velocity, impedance and Poisson's ratio of each layer of the 8-layer earth model. A reflectivity algorithm was used in forward modelling of the seismograms. The error vs. iteration curve for the GS is shown in Fig. 12. The marginal PPDs of compressional wave velocity, impedance and Poisson's ratio are shown in Fig. 13. We do not show any plot for the layer thicknesses (two-way time) since they were very well determined and showed single spikes. Figure 13 shows colour plots of marginal PPDs for each layer parameter. The width of the coloured area for each parameter represents the prior search window of each model parameter. The mean model is plotted with a dashed line on the colour plots. Notice that in many cases the mean does not coincide with the peak of the marginal PPD (shown by solid line) because for several parameters, the distribution is skewed. Some of the important features that can be observed in these results are as follows: The layer above the BSR (with two-way time 3.8 to 3.95 s) is characterized by a large increase in velocity which is very well determined, as evidenced by the narrow width of the distribution. This is followed by a decrease in compressional-wave velocity. Also the layer above the BSR shows an increase in impedance followed by a sharp decrease. It appears that the Poisson's ratio is not very well determined as evidenced by the wide distribution of the marginal PPDs for this parameter.

Discussion and conclusion

Geophysical data are often contaminated with noise and forward modelling algorithms used to generate synthetic data generally make simplistic assumptions about the earth model. These may result in large uncertainties in the derived answer when model-based inversion methods are applied to the solution of geophysical inverse problems. Although the description of noise and uncertainties in forward modelling in terms of a probability density function may be difficult or debatable, Bayes' rule offers a convenient approach to describing the geophysical inversion problem. It describes the solution of the problem in terms of the posterior probability density function in model space. Because our model space is highly multidimensional, a complete display of the PPD is not possible and therefore, we seek quantities such as the posterior marginal density functions, mean, covariance, etc. to describe the solution. Derivation of these quantities is not a trivial task because they involve numerical evaluation of integrals with a large number of variables. One

additional difficulty is that in most geophysical inverse problems the data have a non-linear relationship with the model parameters and the error functions may have multiple extrema. Thus, the PPD may be highly complicated. Therefore, methods that require prior assumptions about the shape of the PPD may not be adequate. These methods attempt to locate the maximum of the PPD, and the Hessian computed at the MAP point is used to estimate the covariance; thus a Gaussian is completely described.

Here we described several computationally intensive sampling-based approaches to derive quantities such as the posterior marginal PPD, mean and covariance. Six different methods have been tested and compared by application to synthetic resistivity sounding data where the number of model parameters were only five so that the forward modelling was computationally tractable. Methods such as GS and PGS agree well with grid-search integration. The results from multiple MAP algorithms are quite encouraging because even though the estimated variances were lower than those estimated by GS, PGS, and grid search, the correlations were in very good agreement. Of course for very complicated PPDs, the multiple MAP algorithms may not perform equally well but for many applications, they will help in getting rough estimates of these quantities quite rapidly.

The SA method is guaranteed to converge asymptotically to a Gibbs' distribution. We showed numerically that a steady-state distribution can be achieved with a GS that requires much less computation time than a complete grid search. Thus running SA at a constant temperature of 1.0 appears to be feasible for many practical problems. Although our PGS is not based on a rigorous mathematical model as is SA, we showed numerically that the results from GS and PGS are nearly identical. The PGS, however, offers the advantage of running the algorithm on parallel computer architectures.

The use of stochastic inversion methods based on statistical models has been criticized primarily because of the requirement that the prior distribution of noise in the data and theory error be known. Often the prior noise pdf is assumed to be Gaussian. In practice, it is never possible to estimate the true noise distribution with a small and finite number of repeated measurements, since the classical definition of probability requires that an experiment be repeated an infinite number of times under identical conditions. The use and specification of the prior pdf has been a subject of dispute for a long time and thus the use of a statistical approach has often been criticized by classical inverse theorists. There exist several definitions of probability. Besides the classical definition which is based on relative frequency of occurrence, the Bayesian interpretation defines probability as a degree of belief (Bayes 1763). The only requirement is that the axioms of probability theory are not violated. Again, in the subjective Bayesian interpretation, the degree of belief is a personal degree of belief. Thus an a priori distribution may simply reflect an expert's opinion. Our aim has been to reduce both acquisition and computation costs and to derive results that are geologically meaningful. We, therefore, took a simple and practical approach to characterize the prior distribution of noise in the data. Thus for

any practical problem we need to analyse the data carefully and to explore the sources of error in the data in order to arrive at some way of characterizing the uncertainty.

Acknowledgements

We thank Carlos Calderon-Macias for the use of his deconvolution code and Warren T. Wood for providing the marine seismic data and many helpful discussions. We thank an anonymous reviewer for many helpful comments. This work was supported by the National Science Foundation grant EAR 9304417 and Texas Higher Education Coordinating Board grant ARP-264. The University of Texas Institute for Geophysics contribution no. 1066.

References

Aarts E. and Korst J. 1989. *Simulated Annealing and Boltzmann Machines*. John Wiley & Sons, Inc.

Basu A. and Frazer L.N. 1990. Rapid determination of critical temperature in simulated annealing inversion. *Science* **249**, 1409–1412.

Bayes T. 1763. Essay towards solving a problem in the doctrine of chances, republished in *Biometrika*, 1958, **45**, 293–315.

Box G.E.P., Leonard T. and Chien-Fu W. 1983. *Scientific Inferences, Data Analysis, and Robustness*. Academic Press, Inc.

Box G.P. and Tiao G.C. 1973. *Bayesian Inference in Statistical Analysis*. Addison–Wesley Pub. Co.

Cary P.W. and Chapman C.H. 1988. Automatic 1-D waveform inversion of marine seismic reflection data. *Geophysical Journal International* **93**, 527–546.

Davis T.E. and Principe J.C. 1991. A simulated annealing-like convergence theory for the simple genetic algorithm. In: *Proceedings of the Fourth International Conference on Genetic Algorithms* (eds R.K. Belew and L.B. Booker), pp. 174-181.

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

Duijndam A.J.W. 1988. Bayesian estimation in seismic inversion, Part I: Principles. *Geophysical Prospecting* **36**, 878–898.

Dziewonski A.M. 1984. Mapping the lower mantle: Determination of lateral heterogeneity in P velocity up to degree and order 6. *Journal of Geophysical Research*, **89**, 5929–5952.

Eiben A.E., Aarts E. and Van Hee K.M. 1991. Global convergence of genetic algorithms: A Markov chain analysis. In: *Parallel Problem Solving from Nature* (eds H.P. Schwefel and R. Männer), pp. 4–12.

Gelfand A.E. and Smith A.F.M. 1990. Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association* **85**, 398–409.

Gelfand A.E., Smith A.F.M. and Lee T.M. 1992. Bayesian analysis of constrained parameter and truncated data problems using Gibbs' sampling. *Journal of American Statistical Association* **87**, 523–532.

Geman S. and Geman D. 1984. Stochastic relaxation, Gibbs' distribution and Bayesian restoration of images. *IEEE Transactions* **PAMI-6**, 721–741.

Gilks W.R. and Wild P. 1992. Adaptive rejection sampling for Gibbs' sampling. *Applied Statistics* **41**, 337–348.

Goldberg D.E. 1989. *Genetic Algorithm in Search, Optimization and Machine Learning*. Addison–Wesley Pub. Co.

Goldberg D.E. and Deb K. 1991. A comparative analysis of selection schemes used in genetic algorithms. In: *Foundations of Genetic Algorithms* (ed. G.J.E. Rawlins), pp. 69–93.

Goldberg D.E. and Segrest P. 1987. Finite Markov chain analysis of genetic algorithms. In: *Proceedings of the Second International Conference on Genetic Algorithms* (ed. J.J. Grefenstette), pp. 1–8.

Hammersley J.M. and Handscomb D.C. 1964. *Monte-Carlo Methods*. Chapman and Hall.

Hastings W. K. 1970. Monte-Carlo methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Horn J. 1993. *Finite Markov Chain Analysis of Genetic Algorithms with Miching*. Illegal Report No. 93002, University of Illinois at Urbana-Champaign.

Ingber L. 1989. Very fast simulated reannealing. *Mathematical Computer Modeling* **12**, 967–993.

Jackson D.D. 1979. The use of a priori data to resolve non-uniqueness in linear inversion. *Geophysical Journal International* **57**, 121–136.

Jackson D.D. and Matsura M. 1985. A Bayesian approach to nonlinear inversion. *Journal of Geophysical Research* **90**, 581–591.

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

Koefoed O. 1979. *Geosounding Principles, 1, Resistivity Sounding Measurements*. Elsevier Science Publishing Co.

Menke W. 1984. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, Inc.

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

Nix A.E. and Vose M.D. 1992. Modeling genetic algorithms with Markov chains. *Annals of Mathematics and Artificial Intelligence* **5**, 79–88.

Press W.H., Flannerry B.P., Teukolsky S.A. and Vetterling W.T. 1990. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press.

Rothman D.H. 1985. Nonlinear inversion, simulated annealing, and residual statics estimation. *Geophysics* **50**, 2797–2807.

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

Rubinstein R.Y. 1981. *Simulation and the Monte-Carlo Method*. John Wiley & Sons, Inc.

Sambridge M.S. and Drijkoningen G.G. 1992. Genetic algorithms in seismic waveform inversion. *Geophysical Journal International* **109**, 323–342.

Sen M.K., Bhattacharya B.B. and Stoffa P.L. 1993. Nonlinear inversion of resistivity sounding data. *Geophysics* **58**, 4, 496–507.

Sen M.K. and Stoffa P.L. 1991. Nonlinear one-dimensional seismic waveform inversion using simulated annealing. *Geophysics* **56**, 1624–1638.

Sen M.K. and Stoffa P.L. 1992. Rapid sampling of model space using genetic algorithms: Examples from seismic waveform inversion. *Geophysical Journal International* **108**, 281–292.

Shalev E. 1993. Cubic B Splines: Strategies of translating a simple structure to B-spline parameterization. *Bulletin of the Seismological Society of America* **83**, 1617–1627.

Stark T.J. 1986. *Information from deep water reflection data: LASE line* 2. Ph.D. thesis, The University of Texas at Austin.

Stoffa P.L. and Sen M.K. 1991. Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane wave seismograms. *Geophysics* **56**, 1794–1810.

Stoffa P.L. and Sen M.K. 1992. Seismic waveform inversion using global optimization. *Journal of Seismic Exploration* **1**, 9–27.

Tarantola A. 1987. *Inverse Problem Theory, Methods of Data Fitting and Model Parameter Estimation*. Elsevier Science Publishing Co.

Tarantola A. and Valette B. 1982. Inverse problems = quest for information. *Journal of Geophysics* **50**, 159–170.

Vose M.D. and Liepins G.E. 1991. Punctuated equilibria in genetic search. *Complex Systems* 5, 31–44.

Wood W.T. 1993. *Least-squares inversion of field seismic data for an elastic 1-D earth*. Ph.D. thesis, The University of Texas at Austin.

Wood W.T., Stoffa P.L. and Shipley T.H. 1994. Quantitative detection of methane hydrate through high-resolution seismic velocity analysis. *Journal of Geophysical Research* **99**, 9681–9695.