Multi-Step Samplers for Improving Efficiency in Probabilistic Geophysical Inference

Miguel Bosch[1], Christophe Barnes[2], and Klaus Mosegaard[3]

[1] Facultad de Ingeniería, Universidad Central de Venezuela,
Los Chaguaramos, Caracas, VENEZUELA.
mbosch@reacciun.ve
[2] Institut de Physique du Globe de Paris, 4, place Jussieu,
F-75252, Paris, Cedex 05, FRANCE.
barnes@ipgp.jussieu.fr
[3] Niels Bohr Institute, University of Copenhagen, Julianes Maries Vej 30,
DK-2100 Copenhagen Ø, DENMARK.
klaus@gfy.ku.dk

Abstract. Geophysical inference is characterized by non-linear relationships between model and data parameters, large model spaces describing spatial distributions of media properties, and intensive computations related to the numerical resolution of the forward problem. Although sampling approaches are convenient to solve such inverse problems, sometimes the involved computations are demanding. We consider here sampling techniques directed to improve the efficiency of sampling procedures in large real geophysical applications. We propose a sampling algorithm incorporating classical importance sampling within a two-step (or multi-step) Markov chain sampler set-up. The first step of the algorithm is a Metropolis sampler ergodic to an importance density function and the second step is a Metropolis sampler correcting from the bias introduced by the importance density function; the combined algorithm samples the posterior probability density function asymptotically. The importance density is a combination of the prior density and an importance likelihood function, obtained as an approximation of the data likelihood function. Hence, the importance density is an approximation of the posterior density. Although rejection rates of the combined algorithm are larger than the rejection rates of the simple step Metropolis algorithm, the computational sampling efficiency is improved when the calculation of the importance density is much easier than the calculation of the actual posterior density. The algorithm is characterized by high acceptance rates in its second step because the first step serves as a barrier rejecting unlikely samples. Data likelihood approximations can be obtained in several ways : using simplified geophysical simulation, using the likelihood with partial (smaller) sets of the observed data, or using information obtained by preliminary analysis of the data. The technique has been successfully used for the inversion of phase arrival times from an offset vertical seismic profile (OVSP) data.

1 Introduction

Inverse problems in geophysics have been commonly treated using optimization approaches like least squares regression or simulated annealing. The ultimate goal has been to find an optimum (in some sense) model that explains the observations and satisfies some prior information. An alternative approach is sampling models according to a probability density function (*pdf*) combining prior information and the likelihood of the model with the observations. The goal, in this case, is to find a large number of models that explain the data and obey the prior information within uncertainties (Press, [1]; Mosegaard and Tarantola, [2]; Mosegaard, [3]; Bosch, [4]). Sampling solves situations that are difficult to treat by searching an optimal model: multiple solutions located in separate regions of the model space, strongly non-linear relationship between data and model, or presence of categorical parameters in the model.

Nowadays, practical applications demand detailed models, large data sets describing the observations and realistic prior information to include in the inverse problem. Sampling posterior *pdf*s in geophysical applications implies the forward calculation of the geophysical data. In most cases the likelihood function has an open form expression and its evaluation for a particular model represents a significant computation effort. This poses the problem of defining techniques to sample in an efficient way posterior densities defined over very large model spaces (from thousands to millions of parameters).

Extensive work can be found in statistical literature about the Markov-chain samplers (e.g. Smith and Roberts, [5]; Tierney, [6]), commonly used to sample posterior *pdf*s. In particular, the Metropolis sampler is recommended for geophysical inference due to its ability to handle complicated posteriors (Mosegaard and Tarantola, [2]).However, the choice of the sampling technique and its actual performance is highly dependent on the particular type of application.

Several techniques have been used in statistical work to improve the accuracy of the posterior estimates calculated by Monte Carlo methods; they are called *variance reduction techniques* (Wolpert, [7]). One of the better know techniques is *importance sampling* (Hastings, [8]; Kloek and van Dijk, [9]; Evans, [10]). In the present work we describe how this technique can be combined with the Metropolis algorithm, within a multiple step set-up, to improve the sampling efficiency.

Multi-step samplers have been previously used in statistical work to improve sampling efficiency (e.g., Geyer, [11]; Tierney, [6]). Commonly, the first step of the algorithm is a Gibbs sampler or other simpler sampler. This particular formulation is not convenient for geophysical inverse problems because the Gibbs sampler is adequate for simple analytic densities with known and easy to sample conditionals, which is not in general the case in geophysical work. We propose in this work a different technique combining several Metropolis samplers in steps. This algorithm is useful in non-linear geophysical inverse problems, demanding large model spaces and heavy computations.

Sections 2 reviews importance sampling and the Metropolis sampling methods, focusing on their application to statistical inference problems. Section 3

proposes a two-step Metropolis algorithm incorporating an importance density in its first stage. Section 4 discusses different ways to construct importance likelihood functions in geophysical inverse problems, and Section 5 summarizes a practical application of the algorithm to the inversion of real bore-hole travel time seismic data.

2 Statistical Inference

In statistical inference problems, the structure of the posterior *pdf* $\sigma(\mathbf{m})$ is given by

$$\sigma(\mathbf{m}) \propto \rho(\mathbf{m})\, L(\mathbf{m}), \tag{1}$$

with $L(\mathbf{m})$ being the likelihood function, $\rho(\mathbf{m})$ being the prior *pdf*, and $\mathbf{m}$ being a point in the model parameter space $\mathcal{M}$. Inference from the *pdf* involves the calculation of posterior probabilities for model parameters and model structure. The probability that a sample $\mathbf{m}_0$ from the posterior density belongs to a subset $\mathcal{A}$ of the model space is given by the integral,

$$p(\mathbf{m}_0 \in \mathcal{A}) = \int_{\mathcal{A}} \sigma(\mathbf{m}) d\mathbf{m}\,, \tag{2}$$

With more generality, it is required the calculation of the expected value of model functionals $f(\mathbf{m})$ with an integral of the type,

$$E[f] = \int_{\mathcal{M}} f(\mathbf{m})\, \sigma(\mathbf{m})\, d\mathbf{m}. \tag{3}$$

Notice that the above formula reduces to (2) with $p(\mathbf{m}_0 \in \mathcal{A}) = E(f)$, and $f(\mathbf{m})$ being an indicator function ($f(\mathbf{m}) = 1$ if $\mathbf{m} \in \mathcal{A}$; $f(\mathbf{m}) = 0$ if $\mathbf{m} \notin \mathcal{A}$).

The integral (3) is also useful to calculate the moments of an arbitrary integrable functional $y(\mathbf{m})$. The second moment, for example, would be obtained with $f(\mathbf{m}) = y^2(\mathbf{m})$,

$$E[y^2] = \int_{\mathcal{M}} y^2(\mathbf{m})\, \sigma(\mathbf{m})\, d\mathbf{m}. \tag{4}$$

Marginals or conditionals of the posterior density can be also formulated by (3) using Dirac's $\delta(\mathbf{m})$ distributions.

As already discussed, in our inference problems the posterior density is a complicated function, and integrals (2) and (3) can not be solved analytically. Instead, we construct unbiased estimates of $E[f]$ by using Monte Carlo integration methods. A naive estimate can be obtained using N samples $\mathbf{m}_i$ from a uniform distribution in $\mathcal{M}$

$$E^0 = \frac{\sum_{i=1}^{N} f(\mathbf{m}_i)\, \sigma(\mathbf{m}_i)}{\sum_{i=1}^{N} \sigma(\mathbf{m}_i)}. \tag{5}$$

However, the variance of this estimate is very large. The regions of the model space with large values of the integrand are usually poorly represented in the sample.

2.1 Importance Sampling

Instead of using a uniform distribution as in (5) one can draw samples from an arbitrary distribution with density $\tilde{\sigma}(\mathbf{m}) > 0$. The integral (3) can then be estimated by

$$E^1 = \frac{\sum_{i=1}^{N} f(\mathbf{m}_i)\,\sigma(\mathbf{m}_i)/\tilde{\sigma}(\mathbf{m}_i)}{\sum_{i=1}^{N} \sigma(\mathbf{m}_i)/\tilde{\sigma}(\mathbf{m}_i)}. \tag{6}$$

The idea of importance sampling consists of making an adequate selection of the sampling density $\tilde{\sigma}(\mathbf{m})$, currently called the importance density function. Statisticians established (see Kloek and van Dijk [9] or Evans [10] for a review) that the variance of estimate E^1 is significantly reduced when $\tilde{\sigma}(\mathbf{m})$ is taken to be similar to $\sigma(\mathbf{m}_i)$.

The reason behind the variance reduction is that more samples are taken from the regions where the density $\sigma(\mathbf{m}_i)$ has an important contribution to the integral value $E(f)$. Indeed, the variance would be optimal if the samples are taken from a distribution with density proportional to $|f(\mathbf{m})|\sigma(\mathbf{m})$. Nevertheless, as often ones needs to estimate many different functions f in the same problem, it is usually the convenient choice to produce samples from a density as close as possible to $\sigma(\mathbf{m})$. The computational effort needed for the generation of the samples is another practical aspect to take into account in the choice of the importance density function; easy to sample importance densities are naturally preferred.

An important variant of the method is called adaptive importance sampling (e.g., Man-Suk and Berger, [12]). It consists of periodically optimizing the parameters of the importance density, in order to better approximate the posterior density. In this way the data likelihood evaluations that are progressively made along the sampling procedure can be used to improve the importance density.

A straightforward application of importance sampling to inference problems consists in drawing samples from the prior density (i.e. $\tilde{\sigma}(\mathbf{m}) = \rho(\mathbf{m})$) to calculate $E(f)$ by

$$E^2 = \frac{\sum_{i=1}^{N} f(\mathbf{m}_i)\,L(\mathbf{m}_i)}{\sum_{i=1}^{N} L(\mathbf{m}_i)}. \tag{7}$$

Possibly useful for preliminary diagnostics, E^2 is in most cases a very rough estimate of integral $E(f)$ because the likelihood function can be wildly different from the prior density.

A better estimation is obtained by sampling from the importance density $\tilde{\sigma}(\mathbf{m}) \propto \rho(\mathbf{m})I(\mathbf{m})$, with $I(\mathbf{m})$ being an approximation of the likelihood function. In this case, the importance density is closer than the posterior density, than in the precedent case. The integral $E(f)$ is estimated by

$$E^3 = \frac{\sum_{i=1}^{N} f(\mathbf{m}_i)\,L(\mathbf{m}_i)/I(\mathbf{m}_i)}{\sum_{i=1}^{N} L(\mathbf{m}_i)/I(\mathbf{m}_i)}. \tag{8}$$

Of course, the most convenient situation for inference purposes would be established by drawing samples from the posterior density itself (i.e. $\tilde{\sigma}(\mathbf{m}) =$

$\sigma(\mathbf{m})$), yielding to the integral estimate

$$E^4 = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{m}_i). \tag{9}$$

2.2 The Metropolis Sampler

The Metropolis algorithm (Metropolis et al., [13]; Hastings, [8]) is a Markov-chain Monte Carlo method that is used, in long runs, to produce samples from posterior densities, needed in (9). Its convenience to geophysical inference has been shown in the works of Mosegaard and Tarantola [2], Mosegaard [3] and Bosch [4]; these works provide thorough descriptions of the way the Metropolis algorithm is applied to solution of geophysical inverse problems.

Briefly, the Metropolis algorithm operates as follows. Suppose that we have a process, a proposal Markov-chain, that is able to generate samples from an arbitrary proposal density $\phi(\mathbf{m})$ in sufficiently long runs. This property is called ergodicity in statistical work; so consider that we have already constructed such a chain ergodic to the proposal density. The Metropolis algorithm defines a Markov-chain that is ergodic to a target density $\pi(\mathbf{m})$, that we call here the posterior chain. Denoting by $\mathbf{m}_t$ the sample drawn from the posterior chain at the step t, the next sample to be drawn from the posterior chain, $\mathbf{m}_{t+1}$, is obtained by the following rules.

1. Consider $\mathbf{m}_t$ to be the current sample in the proposal chain and generate, conditionally to $\mathbf{m}_t$, a next sample $\mathbf{m}_{t+1}^{can}$ form the proposal chain. This will be the candidate sample for the posterior chain.
2. Accept the candidate sample, making $\mathbf{m}_{t+1} = \mathbf{m}_{t+1}^{can}$ with probability

$$p_1 = \text{Min}\left[1, \frac{\pi(\mathbf{m}_{t+1}^{can})\phi(\mathbf{m}_t)}{\pi(\mathbf{m}_t)\phi(\mathbf{m}_{t+1}^{can})}\right]; \tag{10}$$

   otherwise, make $\mathbf{m}_{t+1} = \mathbf{m}_t$.
3. Return to label (1).

A large amount of work has been centered on practical and theoretical subjects of the Metropolis sampler performance (Tierney, [6]; Robert, [14]; Cowles and Carling [15]) and particularly on the problem of convergence.

Notice that for the case in which the proposal density is the prior density, and the target density is the posterior density in (1), the Metropolis acceptance probability is

$$p_1 = \text{Min}\left[1, \frac{L(\mathbf{m}_{t+1}^{can})}{L(\mathbf{m}_t)}\right], \tag{11}$$

with $L(\mathbf{m})$ being the likelihood function. Hence, the Metropolis rule only depends on the ratio of the likelihood function evaluated at the candidate and at the current model. This is the setting in which we use the Metropolis rule in our

inference problems. We construct the prior chain (i.e. ergodic to the prior density) depending on the particular prior information available. A Gibbs sampler, a Metropolis sampler, sampling from multivariate gaussian densities, or combinations of several sampling methods can be used, depending on the problem.

In geophysical inverse problems the likelihood function is difficult to compute as it involves the solution of a geophysical forward problem. Hence, one important factor related to the sampling efficiency is the acceptance rate p_1. If the Metropolis rule yields too low acceptance rates, samples would be often repeated in consecutive steps of the chain and several likelihood evaluations would be needed in order to produce a transition to a different point of the model space. For this reason, sometimes, the acceptance rate has been used as one criterion of efficiency.

However, the acceptance rate is only a rough criterion of efficiency. For instance, high acceptance rates can be obtained with a proposal chain providing very small changes in model parameters at consecutive chain steps, i.e. small step lengths in the model space. In this case the high acceptance rate would not indicate and efficient sampling of the posterior density, because the variability of the chain (the movement of the chain path in the model space) would still be very slow. A chain is usually said to be fast mixing if it can move reasonably well in the model space in order to explore the important areas of the posterior density support in the given number of steps.

3 The Combined Metropolis Algorithm

In this section we present a two-step Metropolis algorithm that is designed to provide good candidate models for the computer intensive likelihood evaluation, that is, models that have a high probability of being accepted in the second step of the Metropolis algorithm. The algorithm is inspired by the classical idea of importance sampling, but it operates by incorporating the importance density function in the two-step Metropolis setup.

We are going to consider a function $I(\mathbf{m}) \geq 0$, here called the importance likelihood function, to transform expression (1) into the equivalent,

$$\sigma(\mathbf{m}) \propto \rho(\mathbf{m})\, I(\mathbf{m}) \left[\frac{L(\mathbf{m})}{I(\mathbf{m})}\right]. \tag{12}$$

The above equation can be regarded as the product of an importance density function defined as $\tilde{\sigma}(\mathbf{m}) \propto \rho(\mathbf{m})I(\mathbf{m})$ and a likelihood correction term $L(\mathbf{m})/I(\mathbf{m})$. Notice that the data likelihood is divided by the importance likelihood to correct for the bias introduced by the importance density in the first factor.

Applying recursively the conventional Metropolis rule (11) to the structure of the above expression, we can define a posterior Markov-chain (i.e. ergodic to the posterior density) in the following way. As for the single step Metropolis algorithm, assume that a prior Markov-chain (i.e. ergodic to the prior density)

is available. Consider that the model $\mathbf{m}_t$ has been obtained from the posterior Markov-chain. The next model $\mathbf{m}_{t+1}$ in the posterior chain is obtained as follows:

1. Consider $\mathbf{m}_t$ to be the current sample in the prior chain, and draw the next sample, $\mathbf{m}_{t+1}^{\text{precan}}$, from the prior chain. This sample is considered a pre-candidate sample of the posterior.
2. Make $\mathbf{m}_{t+1}^{\text{can}} = \mathbf{m}_{t+1}^{\text{precan}}$ with probability

$$p_2 = \text{Min}[1, I(\mathbf{m}_{t+1}^{\text{precan}})/I(\mathbf{m}_t)];$$

   otherwise, make $\mathbf{m}_{t+1} = \mathbf{m}_t$ and return to label (1).
3. Make $\mathbf{m}_{t+1} = \mathbf{m}_{t+1}^{\text{can}}$ with probability

$$p_3 = \text{Min}\left[1, \frac{L(\mathbf{m}_{t+1}^{\text{can}})\, I(\mathbf{m}_t)}{I(\mathbf{m}_{t+1}^{\text{can}})\, L(\mathbf{m}_t)}\right];$$

   otherwise, make $\mathbf{m}_{t+1} = \mathbf{m}_t$.
4. Return to label (1).

In the above algorithm, steps (1) and (2) alone define a chain ergodic to the importance density. Then, steps (3) and (4) transform this chain to be ergodic to the posterior density. The importance likelihood $I(\mathbf{m})$ can be conceived as a smooth approximation of the likelihood function, with longer tails than the likelihood function (see figure 1). Its role consists of focusing the first sampling stage on some relevant regions of the model space. The evaluation of the importance likelihood does not involve the evaluation of the data likelihood function, which is commonly in geophysics a heavy calculation. Instead, the importance likelihood should be easier to evaluate and may follow some simple multivariate parametric model or a mixture of them. Hence, a high computational cost is only associated with the last step of the sampling.

If $I(\mathbf{m}) \approx a\, L(\mathbf{m})$, with a being a proportionality constant, the acceptance probability of the second sampling step would be $p \approx 1$, and we could expect a high acceptance rate in this last step. Even if we use conveniently large step lengths in the prior walk, the first sampling step of the combined algorithm would serve as a "filter", rejecting unlikely samples before they are presented to the second step. This justifies the use of the importance density function as a way to improve the efficiency of the overall sampling procedure, if the importance density is easier to calculate than the posterior density.

However, the overall rejection rate of the combined algorithm is equal or larger to the rejection rate of the single step Metropolis algorithm. In fact, $p_2 p_3 \leq p_1$ because

$$\text{Min}\left[1, \frac{I(\mathbf{m}_{t+1}^{\text{precan}})}{I(\mathbf{m}_t)}\right] \text{Min}\left[1, \frac{L(\mathbf{m}_{t+1}^{\text{precan}})\, I(\mathbf{m}_t)}{L(\mathbf{m}_t)\, I(\mathbf{m}_{t+1}^{\text{precan}})}\right] \leq \text{Min}\left[1, \frac{L(\mathbf{m}_{t+1}^{\text{precan}})}{L(\mathbf{m}_t)}\right]. \quad (13)$$

Although the two-step algorithm has the same or more overall rejections than the single algorithm, the rejections are concentrated in the first step, which only demands evaluation of the importance likelihood. The overall sampling efficiency

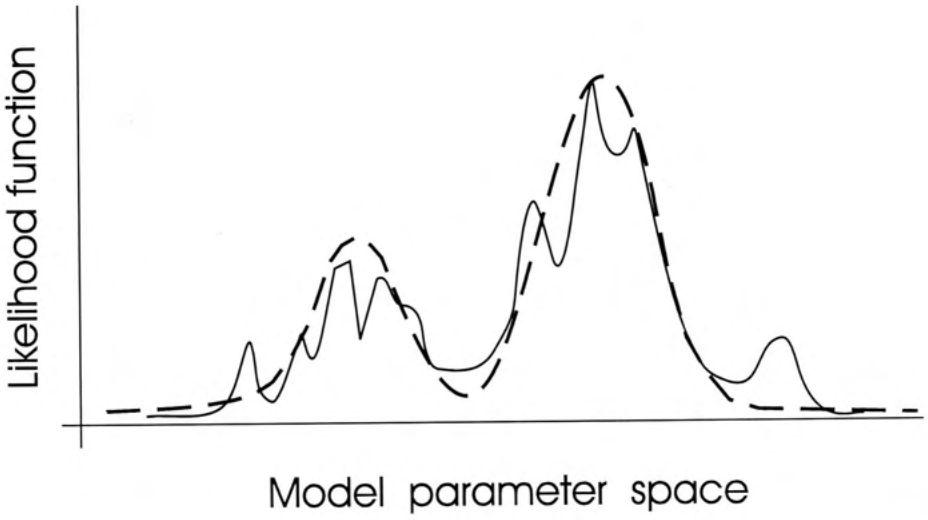

Fig. 1. Simplified illustration of (solid) a data likelihood function defined over a one dimensional parameter space and (dashed) an adequate importance sampling likelihood.

is improved if one defines the function $I(\mathbf{m})$ to be an acceptable approximation to the data likelihood function and one that is inexpensive to calculate. Then, the actual computational cost of pre-candidate rejections is smaller compared to the computational cost of candidate rejections in the single step Metropolis sampler. Consequently, for the same total computational cost, more iterations are produced by the combined algorithm than by the single step Metropolis sampler.

Also, fast mixing prior chains could be incorporated into the algorithm. Because of large rejection rates the simple Metropolis algorithm is not performant with fast mixing prior chains. The combined algorithm here described tolerates large rejection rates, as rejection is concentrated in the first step while keeping a large acceptance rate in the second step. For the same reason, the combined algorithm is able to speed-up the initial *burn-in period*, in which the chain is still influenced by its starting point in the model space.

As for the classical applications of importance sampling, we have here that: (1) The efficiency of the sampling process would be improved when the importance density is approximately similar to the posterior density and easier to calculate. (2) The selection of the importance density preserves the process unbiased affecting only its speed of convergence to the posterior distribution. It is important to remark that the effect of the importance density in the first sampling stage is removed in the second sampling stage by dividing the likelihood function by $I(\mathbf{m})$.

These results on the two-step algorithm are valid for multiple steps by recursivity. Further improvement of the algorithm could incorporate an adaptation of $I(\mathbf{m})$ in order to better approximate the likelihood function as the iterations

progresses. This could be done, as for conventional adaptive importance sampling, by using the likelihood function evaluations to periodically optimize the parameters of the importance likelihood.

4 Importance Likelihoods in Geophysical Problems

Importance likelihood functions can be constructed in geophysical inverse problems in several ways, depending on the particular problem at hand.

4.1 Data Partition

This approach consists in dividing the data set D in a convenient partition of N data subsets D_n and defining a series of consecutively inclusive data subsets $\tilde{D}_n$,

$$\tilde{D}_n = \bigcup_{k=1}^{k=n} D_k \Longrightarrow \tilde{D}_1 \subset \tilde{D}_2 \subset \ldots \subset D. \tag{14}$$

Let us denote by $L(\mathbf{m})$ the likelihood function of a model $\mathbf{m}$ derived from the complete data set D, and similarly denote $\tilde{L}_n(\mathbf{m})$ the likelihood function of the model derived from the data set $\tilde{D}_n$. The likelihood function $\tilde{L}_n(\mathbf{m})$ can be regarded as an approximation of the likelihood function $\tilde{L}_{n+1}(\mathbf{m})$ and can be used as an importance sampling likelihood. In this way we can transform expression (1) for the posterior density into,

$$\sigma(\mathbf{m}) \propto \rho(\mathbf{m})\, \tilde{L}_1(\mathbf{m})\, \frac{\tilde{L}_2(\mathbf{m})}{\tilde{L}_1(\mathbf{m})} \ldots \frac{\tilde{L}_n(\mathbf{m})}{\tilde{L}_{n-1}(\mathbf{m})} \ldots \frac{L(\mathbf{m})}{\tilde{L}_{N-1}(\mathbf{m})}. \tag{15}$$

A multi-step Metropolis sampling algorithm can be adapted to this structure. If one considers a partition of the data set in only two subsets, then a two-step sampler would be implemented as described in the preceding section.

If we consider that uncertainties in data are independent across data subsets, the likelihood function can be factorized and the above equation is simpler,

$$\sigma(\mathbf{m}) \propto \rho(\mathbf{m}) \prod_n L_n(\mathbf{m}), \tag{16}$$

with $L_n(\mathbf{m})$ being the likelihood function of the model derived from the data set D_n. An example of an importance likelihood function derived from the data subset $\tilde{D}_n$ considering a generalized Gaussian model for data uncertainties is,

$$I(\mathbf{m}) = L_n(\mathbf{m}) = \exp\left[-\frac{1}{p} \sum_{d_i^{obs} \in \tilde{D}_n} \frac{|g_i(\mathbf{m}) - d_i^{obs}|^p}{\tilde{\sigma}_i^p} \right], \tag{17}$$

with d_i^{obs} being the ith observed data, $\tilde{\sigma}_i$ the data uncertainty and $g_i(\mathbf{m})$ the simulated data.

Different criteria could be used to setup convenient data partitions. When possible they should be composed by a representative sample of all kinds of data. But differences in efficiency between different kinds of geophysical data simulation should also count, leaving heavy simulations for the last steps of the algorithm.

As in other importance sampling techniques, the efficiency improvement is not warrant in all circumstances. It depends on the choice of data partition and on the problem. We recommend that the partition separates redundant or complementary data. In this way, consecutive likelihood functions in (16) have approximate shapes, yielding to low rejection rates between steps. If there is contradictory data, it should be better grouped into the same data subset. Otherwise, at some level of the algorithm, the proposal models would be contradictory (i.e. located at a different region of the model space) with the likelihood function, producing high rejection rates.

4.2 Fast Geophysical Modeling

The likelihood function based on fast but inaccurate data simulation can be used as importance likelihood function for preliminary rejection of unlikely models before using better quality data simulation. In this way the better quality simulation is made for models that have passed this preliminary selection. An example of importance likelihood function having an approximate forward problem solution $\tilde{g}(\mathbf{m}) \approx g(\mathbf{m})$ and following a generalized Gaussian model is

$$I(\mathbf{m}) = \exp\left[-\frac{1}{p}\sum \frac{|\tilde{g}_i(\mathbf{m}) - d_i^{\text{obs}}|^p}{\tilde{\sigma}_i^p}\right]. \tag{18}$$

The dispersion used in the importance likelihood function should be made greater than the dispersion used in the data likelihood function, $\sigma < \tilde{\sigma}$, to make the first a smoother function than the latter.

4.3 Analysis of Data

Before our inversion, the data may have lead to preliminary characterization of the model parameters through conventional data processing, conventional inversion or qualitative data interpretation. Those results, which may be preliminary images of the study zone (e.g. migrated seismic sections, tomographic images), can well be used to construct an importance sampling likelihood function. As an example, consider that we have a model $\tilde{m}$ which is the result of a preliminary inversion of the data and consider that we associate to each model parameter independent generalized Gaussian uncertainties $\tilde{\sigma}$. Then an importance sampling likelihood function can be

$$I(\mathbf{m}) = \exp\left[-\frac{1}{p}\sum \frac{|m_i - \tilde{m}_i|^p}{\tilde{\sigma}_i^p}\right]. \tag{19}$$

The dispersion $\tilde{\sigma}$ should be sufficiently large to account for the inaccuracy of the preliminary result and to account also for the possible multi-modality of the

data likelihood function. If the preliminary result is obtained by optimization (e.g. least squares) it is important to note that conventional dispersion estimates obtained only measure the local dispersion around the solution and much larger dispersions should be used to define the importance function. If we have preliminary model solutions representing several local maxima of the likelihood function a multi-modal importance function can be well defined as a summation of importance functions for each local maxima. Multi-modal importance density functions are discussed in the work of Man-Suk and Berger [12].

4.4 External Information

In inference problems we often have information that we have not included in the prior density or in the likelihood function for practical implementation reasons or because we do not consider this information reliable enough. This information may come, for instance, from additional data sets or geological hypothesis. This kind of external information, can be represented in $I(\mathbf{m})$ and used to construct the importance likelihood function. If we decide to represent this information by indicating a central model and its parameter variances, a representation of $I(\mathbf{m})$ as in (19) could be used.

It is important to conceptually distinguish importance sampling information represented in $I(\mathbf{m})$ from prior information. Importance sampling information is used as a sampling guide and it is removed in the final step of the sampling algorithm. Prior information, on the contrary, is considered as true information directly contributing to the posterior probability density function. This is another practical advantage of introducing importance sampling. Loose intuitive information about the solution may be used to focus the sampling in some areas of the model space, expecting that this would increase the sampling efficiency. Of course the sampling efficiency would be improved, in fact, only if this external information is nicely consistent with the information represented by the posterior density.

5 An Application to Seismic Data Inversion

The combined Metropolis sampler has been applied with success to the estimation of P-wave and S-wave velocities from bore-hole seismic arrival times provided by an offset vertical seismic profile real experiment (OVSP). To illustrate the method we present a brief description of this application. Further information is presented in the work of Barnes et al. [16].

The seismic data was recorded by 155 three-component geophones (seismographs measuring the velocity vector) located inside a vertical well from 1250 to 4000 m depth. The instruments received the seismic signal from an explosive (air-gun) source located at the surface with 2000 m offset from the top of the well; figure 2 describes the geometry of the offset vertical seismic profile (OVSP) experiment. The obtained seismograms show complicated signals including refracted, reflected and diffracted wave phases due to discontinuities in the medium

physical properties. The traveltimes from the source to geophones for 28 significant phases were picked from the seismogram and used as data in the inverse problem.

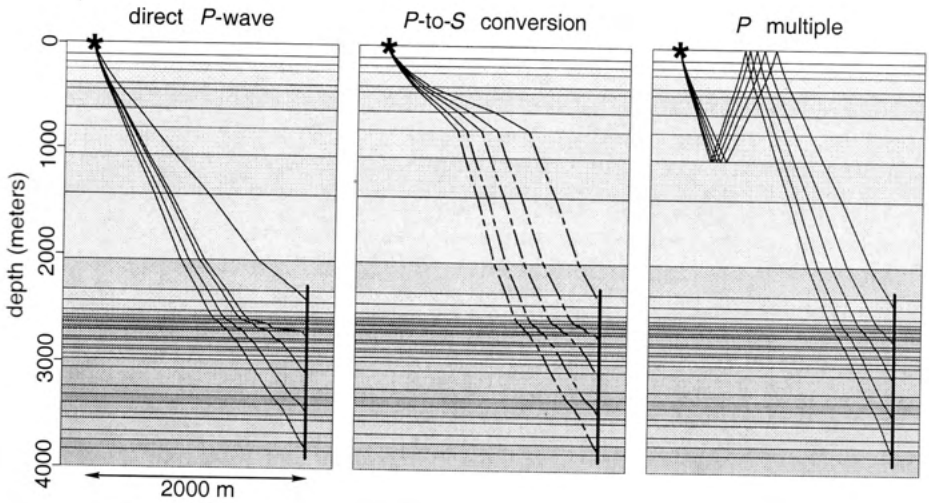

Fig. 2. Layered representation of the media where the geometry of acquisition for the seismic data is shown. The symbol in the surface indicates the source position, and the vertical bold line indicates the position of the array of geophones (inside the well). Raypaths have been plotted to illustrate three different wave phases; solid lines represent P-wave raypaths and dashed lines represent S-wave raypaths.

Prior information was incorporated into the inverse problem:

- structural information from migrated images (low spatial resolution) obtained from seismic data recorded at the surface,
- information on part of the physical parameters from direct physical measurements in the bore-hole (acoustic and density logs), and
- information from regional geological surveys (e.g. information about statistical relationship between P-wave velocity and S-wave velocity).

Structural information showed that the medium was approximately horizontally layered. As regional studies confirmed this observation, we used it in order to justify the horizontally layered assumption made to simplify the parameterization (physical parameters depended only on the depth variable) and thus decrease the number of parameters describing the medium. According to migrated images and the analysis of the log data we considered that the medium presented 34 main layers. Moreover, as the log data showed weak variability of the seismic velocities inside a layer, we considered these parameters homogeneous (constant) for each layer. Hence, we had $34 \times 3 = 102$ model parameters, defining for each layer: (1) the layer bottom depth, (2) the P-wave layer velocity, and (3) the S-wave layer velocity.

Information from log data was used to infer the prior *pdf* for the P-wave velocities and the depths of interfaces. As direct information on S-wave velocity was not available, we used petrophysical studies on samples about P versus S-wave velocity relationship (Castagna et al., [17]), for rocks in the same region, in order to infer statistical information on S-wave velocity over P-wave velocity ratio. The prior information on both velocities distributions was log-normal and dependencies between adjacent layers were taken into account. The prior information about the depths of the layers combined (1) a central value for each layer interface, and (2) statistical constrains about the layer thickness, both obtained from the log analysis and the surface seismic.

The traveltimes were simulated solving the Eikonal equation by a finite difference method (Podvin and Lecomte, [18]). For multiple phases, traveltime computation needs three simulations (down-going from the source then reflected and thus up-going and reflected again thus down-going to the receivers). Errors induced by the space discretization have been estimated and introduced in the inversion as modeling errors. As the traveltime computation is time consuming, we have used a grid size of 30 m which is a trade-off between the computer time cost and accuracy of time calculation (modeling errors should be less than picking errors in order to preserve data information and then resolution in the solution).

To solve the problem it was used a Markov chain Monte Carlo method: the Metropolis sampler. We recall that the Gibbs sampler can not be used here because we do not have the conditionals of the posterior density. The single step Metropolis algorithm was implemented at a first glance, and it turned out to be impossible to achieve equilibration in misfit and posterior parameter statistics within the available computation time. Data misfit reduction was not satisfactorily reduced (large misfit compared with data uncertainties), using several trials for the prior chain step length and step directions. This showed that the algorithm was not completing its initial phase of misfit reduction (burn-in period).

The difficulty for treating this inverse problem with the conventional Metropolis algorithm, suggested the use of the multi-step metropolis sampling scheme. The approximation of the likelihood function with the data partition approach (16-17) was natural here, because the calculation of the arrival times could be separated for the different phases. The arrival time data set was divided into six subsets (see figure 3) and the Metropolis algorithm was decomposed in six steps, correspondingly. Candidate models were confronted with each data subset in sequence, allowing to reject unlikely models using the simulation of a few wave phases. The simulation of additional phases was only performed if the model was accepted in the previous test.

The six subsets were ordered following two criteria: the data precision (energy, picking uncertainty), and the sequence of calculation in the forward problem solution. Group 1 in the sequence, for instance, included the direct P-wave arrival and the first P-wave multiple (water layer multiple). These were the phases with smallest uncertainties and their calculation was necessary for simulating the other phases travel times (saving computing time). Groups 5 and 6

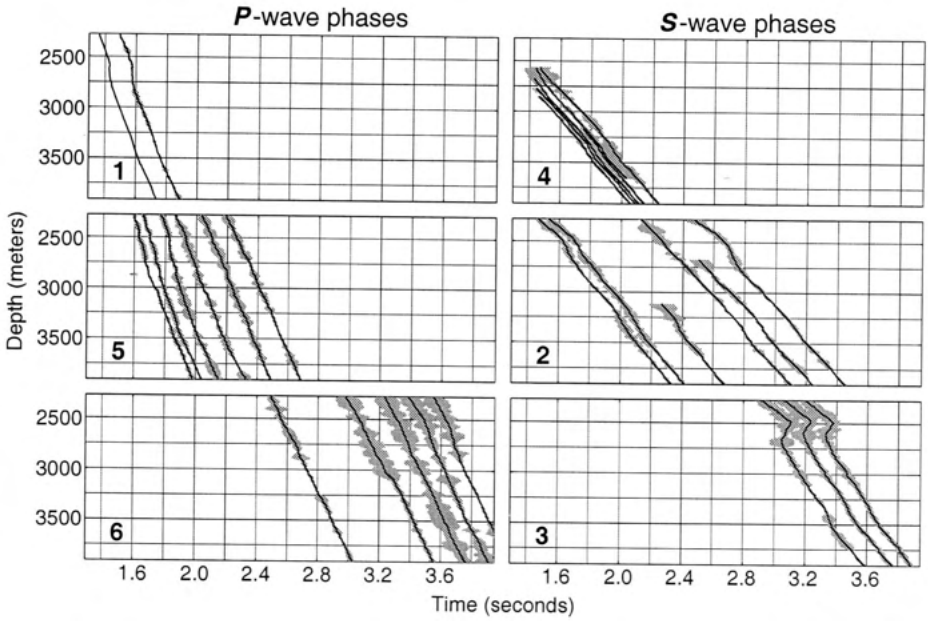

Fig. 3. The arrival times for 28 selected observed wave phases grouped in six data subsets. The solid black lines indicate the arrival times picked from the seismogram and the light gray regions show the uncertainty (mean deviation) associated with the picking operation. The numbers represent the sequence in which the data subsets are used by the multi-step sampling algorithm.

in the sequence were P-multiples and the travel time calculation was more time consuming than for the P-to-S converted phases (groups 2, 3 and 4). Groups 2 and 3 provided accurate information about the S-velocity model (which was not well constrained by the prior information).

The multiple step Metropolis algorithm resulted in a major improvement in sampling efficiency. Choosing an adequate preliminary random walk (found empirically after several trials), it has allowed a fast misfit reduction (burn-in-period) and an acceptable convergence of posterior statistics. To check the convergence of posterior statistics several univariate marginal plots and bivariate marginal cross-plots were monitored, and four independent chains were run. We verified that at least three of the four independent chains converge to similar plots in each monitored variable, or pair of variables. As usual in Monte Carlo, the burn-in period outcome samples were left out from posterior statistics.

Our comparison of the performance of the two algorithms is here qualitative. The multi-step Metropolis algorithm (1) passed the burn-in period by reducing satisfactorily the misfit between calculated and observed data, and (2) fulfilled convergence criteria. Whereas, the single-step algorithm could not pass the prior stage of reducing the data misfit.

The results obtained with the six-step algorithm were satisfactory. Although we did not intend here further division of the data set in smaller subsets, our experience is that breaking the data in too many subsets increases the rejection rate and renders the algorithm inefficient. In this data partition approach there is a compromise between the economy in the likelihood computations and the increment of the rejections.

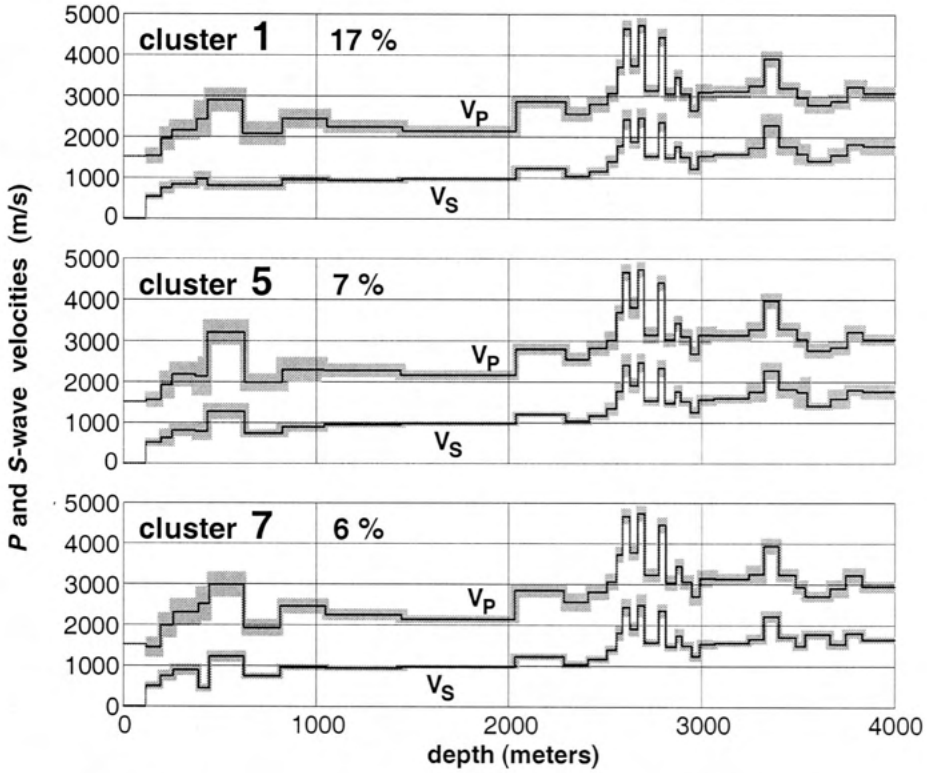

Fig. 4. Three of the most representative model clusters obtained after the analysis of the posterior model set generated by the sampling algorithm. Each cluster being not so far from a multivariate Gaussian statistical model, we represent its centroid (mean model) and the standard deviation of the model parameters. The mean model is indicated by a black bold line, and the standard deviation is given by the gray region around the mean model. The global probability for each mode is given in percent.

The results shown here were obtained using the four independent Markov chains running in parallel (during five days in four work-stations) to produce an overall posterior sample set of 600.000 models (joint P-wave velocity, S-wave velocity models, and depth interface parameters). This posterior model sample set was used to estimate posterior statistics on model parameters. The posterior

pdf being clearly multi-modal, we have used a clustering algorithm which has allowed to identify 15 separated and significant clusters. Figure 4 shows three of these clusters.

The multi-modality of the posterior *pdf* was clearly seen in marginal histograms on model parameters. It was expected because an observed phase could be explained by several different synthetic phases; the set of simulated phases was larger than the actual data phases. The differences among the clusters shown in figure 4 are particularly clear at the uppermost and lower-most zones of the model. This differences are enough to change (shift) the matching between some observed and calculated phases. It was verified, for instance, that the first observed phase of data group 3 was explained in a different way for the clusters one and seven. For the first cluster it was fitted by a conversion on the 5th interface while for the seventh cluster it was fitted by a conversion on the 4th interface.

6 Conclusions

In the first sections of the work we reviewed sampling techniques that are useful in complex inference problems where the evaluation of the likelihood function represents a major computational effort: importance sampling and the Metropolis sampler. Classical importance sampling is a variance reduction technique that allows to focus the sampling in some regions of the model space, without introducing a bias in posterior estimates. If the importance density is suitable selected to mimic the posterior density the technique is able to improve the sampling efficiency.

The Metropolis sampler is an algorithm that provides, asymptotically, a sample of the posterior *pdf*. It is a convenient method in problems with complex posterior information, as it is the case for geophysical inverse problems. However, the actual performance of the Metropolis sampler is known to be affected by the computational limitations in evaluating the likelihood function (and thus a limitation in the total number of collected samples). The speed of convergence, the sample mixing and the accuracy of estimates are highly dependent on the problem at hand and on the particular implementation of the algorithm.

The contribution of the present work is to combine the two techniques portrayed above to design a multi-step Metropolis algorithm been able to significantly improve the sampling performance. The first step of the algorithm samples from an importance density, $\tilde{\sigma}(\mathbf{m}) \propto \rho(\mathbf{m})I(\mathbf{m})$, and the second step samples from the posterior density. The importance density is here defined as the product of the prior density and an importance likelihood function that should be much easier to calculate than the actual data likelihood function. Although the overall acceptance rate of the combined algorithm is smaller than the acceptance rate of the single Metropolis algorithm, the acceptance rate of the second step of the combined algorithm is significantly increased. As the second step is involving the major computational effort, the overall computational efficiency is improved.

Different ways to construct importance likelihood functions for geophysical inverse problems were discussed. This functions can be defined with basis on

(1) the information from the preliminary analysis of the data, (2) the misfit of the observed data with simplified simulations of the geophysical response of the models, (3) the misfit between calculated and observed data for suitably chosen subsets of the total available data set, or (4) additional information that is not incorporated to the posterior density.

The algorithm has been tested on a seismic real inverse problem with success. The design of the importance likelihood function followed here a partition of the travel time data in six subsets according to different groups of wave phases; a six-step Metropolis sampler was implemented correspondingly. The candidate models were confronted progressively with the six subsets of the observed data allowing fast rejection of unlikely models.

Acknowledgements. We thank the Institut de Physique du Globe de Paris, the Central University of Venezuela, the University of Copenhagen and the Geophysic Center of the Ecole des Mines de Paris. Part of this work has been funded by the UMR 7580 of the CNRS (France) and the program GeoFrance-3D. We also gratefully thank ELF company as they provided us the OVSP data. We thank the editor Per Christian Hansen, and the comments of an anonymous referee that helped to improve the manuscript.

References

1. Press, F., Earth models obtained by Monte Carlo inversion, *Journal of Geophysical Research, 73,* 5223-5234, 1968.
2. Mosegaard, K., and A. Tarantola, Monte Carlo sampling of solutions to inverse problems, *Journal of Geophysical Research, 100* 12,431-12,447, 1995.
3. Mosegaard, K., Resolution Analysis of General Inverse Problems through Inverse Monte Carlo Sampling: *Inverse Problems 14,* 405-426, 1998.
4. Bosch, M., Lithologic tomography: from plural geophysical data to lithology estimation, *Journal of Geophysical Research, 104,* 749-766, 1999.
5. Smith, A. F., and G. O. Roberts, Bayesian computation via the Gibbs sampler And related Markov chain Monte Carlo methods, *J. R. Statist. Soc. B, 55,* 3-23, 1993.
6. Tierney, L., Markov-chains for exploring posterior distributions, *Annals of Statistics, 22,* 1702-1762, 1994.
7. Wolpert, R., Monte Carlo integration in Bayesian statistical analysis, *in* Statistical Multiple Integration, Flournoy N. and Tsutakawa R. K. Eds., Contemporary Mathematics Series, American Mathematical Society, v.115, 101-115, 1989.
8. Hastings, W. K., Monte Carlo sampling method using Markov chains and their applications, *Biometrika, 57,* 97-109, 1970.
9. Kloek, T., and H. K. van Dijk, Bayesian estimates of equation system parameters: an application of integration by Monte Carlo, *Econometrica, 46,* 1-19, 1978.
10. Evans, M., Adaptive importance sampling and chaining, *in* Statistical Multiple Integration, Flournoy N. and Tsutakawa R. K. Eds., Contemporary Mathematics Series, American Mathematical Society, v.115, 137-143, 1989.

11. Geyer, Ch. J., Practical Markov Chain Monte Carlo, Statistical Science, 7, 473-511, 1992.
12. Man-Suk, O., and J. O. Berger, Integration of multimodal functions by Monte Carlo importance sampling, *Journal of the American Statistical Association*, 88, 450-456, 1993.
13. Metropolis, N. , A. W. Rosenbluth, M. N Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, *The Journal of Chemical Physics, 21,* 1087-1092, 1953.
14. Robert, Ch. P., Convergence control methods for Markov Chain Monte Carlo Algorithms, *Statistical Science, 10,* 231-253, 1995.
15. Cowles, M. K., and B. P. Carling, Markov Chain Monte Carlo convergence diagnostics: A comparative review, *Journal of the American Statistical Association, 91,* 883-904, 1996.
16. Barnes, Ch., M. Charara, and A. Tarantola, Traveltime inversion of bore-hole seismic data using Monte Carlo Methods, submitted to the Geophysical Journal International, 1999.
17. Castagna, J. .P, M. L Batzle, T. K. and Kan, Rock Physics — The Link Between Rock Properties and AVO Response *Offset-Dependant Reflectivity — Theory and Practice of AVO Analysis*, (Society of Exploration Geophysics: Investigations in Geophysics), p 135, 1993
18. Podvin, P., and I. Lecomte, Finite difference computations of travel times in very contrasted velocity models: massively parallel approach and its associated tools, *Geophysical Journal International,* 105, 271-284, 1991.