Full-waveform Inversion by Informed-Proposal Monte Carlo

Sarouyeh Khoshkholgh[1], Andrea Zunino[1,2] and Klaus Mosegaard[1]

[1] Department of Physics of Ice, Climate and Earth, Tagensvej 16, 2200 Copenhagen N, Denmark
[2] Now at: Department of Earth Sciences, ETH Zürich, Sonneggstrasse 5, 8092 Zürich, Switzerland

Abstract

Markov Chain Monte Carlo (MCMC) sampling of solutions to large-scale inverse problems is, by many, regarded as being unfeasible due to the large number of model parameters. This statement, however, is only true if arbitrary, local proposal distributions are used. If we instead use a global proposal, informed by the physics of the problem, we can dramatically improve the performance of MCMC and even solve highly nonlinear inverse problems with vast model spaces. This is illustrated by a seismic full-waveform inverse problem in the acoustic approximation, involving close to 10^6 parameters.

Introduction

Full-waveform inversion (FWI) is emerging as a promising method for computing subsurface properties and high-resolution images from seismic data. However, since its introduction in the late 1980s through the theoretical work of Lailly (1983) and Tarantola (1984, 1986, 1988) and subsequent numerical tests [4, 6]. it has been known that inference about the low-wavenumber components of the Earth model (the background velocity field) is a highly nonlinear problem. Early attempts to solve this problem using Markov-Chain Monte Carlo (MCMC) techniques (see, e.g., Koren et al., 1991) ran into serious problems, not only because of the shortcomings of computational resources at the time, but also because large-scale inverse problems have vast parameter spaces.

Later improvements [2, 12, 13, 21], of which the frequency-domain approaches [14, 13, 15] were amongst the most important, gave further momentum to the development of more efficient algorithms, but there were still some way to go before a probabilistic full-waveform inversion could be attempted. Characterizing the full posterior probability distribution, either through a sample of solutions, or in a parameterized form, is a formidable problem, far exceeding the calculation of of best-fitting solutions.

Only recently, improved methods for many-parameter highly nonlinear seismic inversion have come to light, of which inversion based on Hamiltonian Monte Carlo (HMC) [3, 5] and Variational Full-Waveform Inversion (VFWI) [22] are notable examples. HMC is an MCMC with improved sampling efficiency using gradient information of the misfit function - information that can be computed relatively fast for seismic inverse problems through the method of adjoints. VFWI is a non-sampling method seeking a continuous mapping over the parameter space that transforms samples from the prior probability density into approximate samples from the posterior.

A close look at all methods (deterministic and stochastic) used for solution of inverse problems (linear, weakly nonlinear and strongly nonlinear) reveals that efficiency always requires specific properties/assumptions about the problem built into the algorithm: In regular MCMC, good performance depends heavily on smoothness information obtained through initial experimentation with step lenghts of the algorithm. In HMC, efficient computation of misfit gradients through adjoints, which is a specific property of the wave equation, is built into the algorithm. VFWI assumes a predefined family of posterior distributions among which it seeks a solution. The role of information in defining the efficiency of an inversion algorithm is further explored in [7] where it is shown that goal-directed use of specific information, characteristic of the problem, can dramatically improve performance.

In this study we propose a MCMC methodology for large-scale full-waveform inversion, based on the Informed Proposal Monte Carlo (IPMC) technique described by *Khoshkholgh et al.* (2020). Our approach in this paper is pragmatic, in that we use approximate information acquired through classical seismic processing and subsequent interpretation to build a proposal strategy for the MCMC sampler. IMPC ensures that errors and inaccuracies in the proposal information does not pollute the final sampling result. Only the speed by which statistically independent samples are collected is influenced: The closer the proposal is to the posterior, the faster is the sampling.

First we provide a brief overview of the Informed Proposal Monte Carlo method, and then we apply the method to a synthetic 2D test example and compare the results with those obtained by a regular MCMC method.

Methods

Probabilistic Inversion

Probabilistic inverse theory is based on the assumption that any state of information about a parameterized system can be described by a probability density function. In the two most widely used formulations, the Bayesian approach [1, 11] and the approach of Tarantola and Valette (1982), the outcome of the inversion is a so-called posterior probability distribution $\sigma(\mathbf{m})$ over the model parameter space. In both formulations, $\sigma(\mathbf{m})$ is expressed as a product of two distributions

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

where $\rho(\mathbf{m})$ quantifies our (uncertain) prior information about $\mathbf{m}$, and L contains information about $\mathbf{m}$ from uncertain, observed data and from physical law. Since a probabilistic inversion produces, not only one particular solution, but a probability distribution over the entire model space, we are left with the problem of characterizing this distribution. One choice is to look for a *parametric* description of the posterior (as is done in variational methods [22]), but another approach is to create a *non-parametric* description of the posterior. This method, which is essentially equivalent to forming a multi-dimensional histogram of the posterior, is the one used by MCMC- and other sampling methods.

Markov-Chain Monte Carlo

The goal of an MCMC sampler is to produce a collection of models with a sampling density proportional to a given target distribution. Each step of the sampling from a probability density $\sigma(\mathbf{m})$ proceeds from a current $\mathbf{m}$ by first randomly proposing a new model $\mathbf{m}'$ according to a so-called *proposal distribution* $q(\mathbf{m}'|\mathbf{m})$, followed by accepting $\mathbf{m}'$ only with probability

$$P_{\text{acc}}^{\mathbf{m}\to\mathbf{m}'} = \min\left(\frac{\sigma(\mathbf{m}')q(\mathbf{m}|\mathbf{m}')}{\sigma(\mathbf{m})q(\mathbf{m}'|\mathbf{m})}, 1\right). \tag{2}$$

This acceptance probability guarantees that, in the limit where the number of models $N \to \infty$, the distribution $\sigma(\mathbf{m})$ will be correctly sampled. There is great freedom in the choice of the proposal distribution $q(\mathbf{m}'|\mathbf{m})$, as long as equation (2) is well defined for all $\mathbf{m}'$. It is, however, clear from (2) that if we could choose $q(\mathbf{m}'|\mathbf{m}) = \sigma(\mathbf{m}')$ we would have an algorithm with maximum efficiency: It would be allowed to move freely between models of non-zero values of the target probability density, and all moves would be accepted. It is, however, also clear that this choice of q would require that we already had full knowledge of the structure of σ and hence that the solution would be known from the beginning!

In practice, the situation is quite the opposite. We usually have very little information about σ or, to simplify implementation of our algorithm, we ignore this information. A commonly used, minimum information about σ used in an MCMC implementation is about the smoothness of σ, and it is normally found by experimentation. The smoothness tells us how far away from the current point $\mathbf{m}$ in the model space we can go without changing the value of σ significantly, and this allows us to build the proposal q so narrow that any new proposed model have a good chance of being accepted. The advantage of a narrow q is that it is quite similar (near-proportional) to σ in a small neighborhood around $\mathbf{m}$. For this reason it gives a high acceptance probability, but the disadvantage is that the resulting moves away from $\mathbf{m}$ are so small that successive, accepted models become highly correlated. Consequently, it takes many moves to produce new, uncorrelated models.

It is the above-mentioned experience with information-deficient, narrow proposals that leads many to conclude that MCMC is inherently inefficient. We view the situation quite differently: MCMC is an algorithm that allows (possibly slow) sampling of a distribution σ, even in cases where the approximate smoothness is the only thing known about σ. If, on the other hand, we have more comprehensive information about σ, this can be built into the proposal q, and the algorithm can be made much more efficient. The purpose of this study is to demonstrate that the latter approach is very efficient. We use a semi-realistic, synthetic, seismic inversion example where inference about $\approx 10^6$ model parameters is sought.

Using a Proposal, Informed by Approximate Physics

Following the formulation of Tarantola and Valette (1982), a general expression for the joint posterior probability can be written

$$\sigma(\mathbf{d}, \mathbf{m}) = \frac{\rho(\mathbf{d}, \mathbf{m})\theta(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} \tag{3}$$

where $\mathbf{d}$ is data, $\mathbf{m}$ is the model parameters, $\rho(\mathbf{d}, \mathbf{m})$ is the prior and $\mu(\mathbf{d}, \mathbf{m})$ is the homogeneous probability density in the joint $(\mathbf{d}, \mathbf{m})$-space. The density $\theta(\mathbf{d}, \mathbf{m})$ is the distribution of errors/uncertainties of the relation between $\mathbf{m}$ and $\mathbf{d}$, including data uncertainties and possible, physical modelization errors. In this study we will, without loss of generality, assume that μ is constant over the model space, and that $\rho(\mathbf{d}, \mathbf{m}) = \rho(\mathbf{d})\rho(\mathbf{m})$ (prior information about $\mathbf{d}$ and $\mathbf{m}$ are independent). This leads to (ignoring the normalization constant):

$$\sigma(\mathbf{d}, \mathbf{m}) = \rho(\mathbf{m})\rho(\mathbf{d})\theta(\mathbf{d}, \mathbf{m}) \tag{4}$$

Following [7], we now build our proposal distribution as a rough approximation $\tilde{\sigma}_m(\mathbf{m})$ to the marginal posterior in the model space. $\tilde{\sigma}_m$ will be based on sim-

plified physics where the modelization errors overwhelm the observational data uncertainties, hence (ignoring normalizations):

$$\tilde{\sigma}_m(\mathbf{m}) \approx \sigma(\mathbf{d}_{obs}, \mathbf{m}) \approx \rho(\mathbf{m})\theta(\mathbf{d}_{obs}, \mathbf{m}) \; . \tag{5}$$

expressing the approximate posterior as a product of the prior and the approximate likelihood $\theta(\mathbf{d}_{obs}, \mathbf{m})$. In the following we will use $q(\mathbf{m}'|\mathbf{m}) = \tilde{\sigma}_m(\mathbf{m}')$ as a global proposal distribution to speed up the MCMC algorithm. Even if q is only a poor approximation to the likelihood L, it will not bias the final sampling result. Only the efficiency of the algorithm will be influenced.

In this study we obtain $q(\mathbf{m}'|\mathbf{m})$ through traditional processing and interpretation of the seismic data $\mathbf{d}$. This procedure is in the following formally denoted h, a pseudo-inverse operator mapping from data space into the model space: $\tilde{\mathbf{m}} = h(\mathbf{d})$. For real data this may include muting, velocity analysis, normal moveout correction, stacking, deconvolution, migration, and subsequent interpretation. In our illustrative synthetic example we simplify this procedure.

Constructing the Approximate Likelihood

In our study, we take a simple approach and construct $\theta(\mathbf{d}_{obs}, \mathbf{m})$ as follows:

1. Using a simplified approach, we construct an approximate solution $\tilde{\mathbf{m}}$ to the inverse problem.

2. The "true" modelization error is $\delta\mathbf{m}_{true} = \tilde{\mathbf{m}} - \mathbf{m}_{true}$ but since $\mathbf{m}_{true}$ is unknown, we instead compute an approximation to the modelization error:

   $$\delta\mathbf{m}_{approx} = h(g(\tilde{\mathbf{m}})) - \tilde{\mathbf{m}},$$

   where g is the forward function used to compute synthetic data. Since $h(g(\tilde{\mathbf{m}}))$ is an approximate solution to the inverse problem with $g(\tilde{\mathbf{m}})$ as data, the above formula estimates what the modelization error would have been if $\tilde{\mathbf{m}}$ had been the true model. In case $\tilde{\mathbf{m}}$ is close to $\mathbf{m}_{true}$, we expect that $\delta\mathbf{m}_{approx}$ will be close to $\delta\mathbf{m}_{true}$.

3. We now define now $\theta(\mathbf{d}_{obs}, \mathbf{m})$ as a Gaussian with mean $\tilde{\mathbf{m}}$ and a diagonal covariance matrix where the n'th standard deviation is equal to the n'th component of $\delta\mathbf{m}_{approx}$.

Informed Proposal Sampling

Given that $S(\mathbf{m})$ is the (accurately computed) misfit function for the problem, we can now use an MCMC algorithm to sample the posterior, using $\rho(\mathbf{m})$ as the prior, $L(\mathbf{m}) = \exp(-S(\mathbf{m}))$ as the accurately computed likelihood, and $q(\mathbf{m}'|\mathbf{m}) = \rho(\mathbf{m}')\theta(\mathbf{d}_{obs}, \mathbf{m}')$ as the proposal. We assume here that values of the prior can be computed explicitly, so our proposal is now fully defined.

Each step of the algorithm now runs as follows:

1. Perturb the current model $\mathbf{m} \to \mathbf{m}'$ using the proposal $q(\mathbf{m}'|\mathbf{m})$.

2. Accept/reject the perturbed model $\mathbf{m}'$ with probability

$$P_{\text{acc}}^{\rho} = \min\left(\frac{\rho(\mathbf{m}')L(\mathbf{m}')q(\mathbf{m}|\mathbf{m}')}{\rho(\mathbf{m})L(\mathbf{m})q(\mathbf{m}'|\mathbf{m})}, 1\right) = \min\left(\frac{L(\mathbf{m}')\theta(\mathbf{d}_{obs}, \mathbf{m})}{L(\mathbf{m})\theta(\mathbf{d}_{obs}, \mathbf{m}')}, 1\right). \tag{6}$$

   If $\mathbf{m}'$ is rejected, the current model $\mathbf{m}$ will be repeated.

In practice, the rate of accepted models found by the above algorithm can be improved if we introduce a burn-in period (say, within the first N_B iterations) in the sampling procedure where we replace the approximate model $\theta(\mathbf{d}_{obs}, \mathbf{m})$ with better-fitting models discovered in the process. We will see an example of this in the numerical example below.

Results

Model Parameters and Data

The above-mentioned method was applied to acoustic full waveform inversion of synthetic reflection data from a subset of the Marmousi velocity model [20] giving P-wave velocities in a 755 x 1255 Cartesian grid (= 947525 parameters) with a grid size of 1.5 m in vertical and horizontal directions (see Figure (1)). Two shot records were generated by a finite difference algorithm to solve the constant-density, variable-velocity acoustic wave equation in two dimensions (*CREWES Library*: *Youzwishen*, 1999, and *Margrave*, 2000). The algorithm uses second-order finite-difference operators for the time derivative and the Laplacian operator and applies simple, absorbing boundary conditions. Figure (2) shows one of the two shot records used as observed data in our study. The dominant frequency of the Ricker wavelet related to the source term is 40 Hz. The seismic sources were located at $x = 375$ m and $x = 1500$ m, and receivers were located at the surface, equally distributed with distances of 7.5 meters.

The Likelihood Function

Assuming Gaussian white noise on the data, we use the likelihood function

$$L(\mathbf{m}) = \exp\left(-||\mathbf{d} - g(\mathbf{m})||/\sigma^2\right)$$

where g is the forward function calculating synthetic data from a model $\mathbf{m}$, and σ is the standard deviation of the noise. σ was chosen to give a signal-to-noise ratio $S/N \approx 2.0$. Synthetic noise was not added to the data.

The Prior Distribution

In our study, we use a prior probability density assigning non-zero probability only to piecewise constant velocity models that can be derived from:

1. Smooth, continuous 1-1 deformations ('warpings') of the approximate model. A warping is defined by a random, smooth displacement field $\mathbf{u}$ with $|\partial u_i/\partial x_i| < 1$ where x_i $(i = 1, 2)$ are the image location coordinates. Warping takes place in a randomly centered square window with a predefined dimension (here 250×250), a cosine variation of displacement coordinates in both vertical and horizontal direction, and with zero displacement at the boundary. Any such warping will result in a new, piecewise constant velocity model.

2. Choosing new velocities in each layer within given fixed intervals.

All models that can be generated according to the above rules are assumed a priori equally likely. Figure 3 shows four sample models from the prior. It is seen that the prior allows considerable variations in the model, but retains its basic 'topology'.

Building the Proposal Distribution

Our proposal distribution is a rough approximation to the posterior, which in turn is a product of the prior and a Gaussian, approximate likelihood. The approximate likelihood is centered at the approximate solution to the problem and with standard deviations proportional to the modelization errors of each model parameter. Hence, our proposal sampler is based on four components:

1. **A rough approximation to the posterior**. To find this, we assume that we have a 2D seismic reflection profile across the area, from which we will derive a rough subsurface model through classical processing and interpretation. To simulate this situation, we generate zero offset seismic data using the exploding reflector model [10]. The first step in our data processing is

a depth migration using a rough background velocity model derived from the true velocity model through Gaussian smoothing (Figure 4), followed by computing an approximate reflectivity profile through frequency domain spiking deconvolution (Figure 5). Then we simulate seismic interpretation by identifying clearly visible reflectors, and after combining this interpreted image with the long-wavelength migration velocity field, we arrive at an approximate, homogeneous-layer velocity model for the area (Figure 6). This velocity model is a rough solution $\tilde{\mathbf{m}}$ to the full-waveform inverse problem.

2. **Standard deviations of the approximate likelihood**. The difference between the exact likelihood and the approximate likelihood is obtained by generating synthetic reflection data from the approximate model, and then performing a rough inversion of this data through processing and interpretation to obtain a second (and even less accurate) approximate model. Subtracting this from the (first) approximation allows estimation of the modeling error, providing standard deviations of our Gaussian, approximate likelihood. The spatial envelope of the velocity differences at each pixel in the model, computed through a Hilbert transform, is shown in Figure 7 and reveals regions of large and small modelization errors.

3. **A warping scheme that allows deformation of a model into a new model with nonzero prior probability**. Our warping consists of a series of deformations, each in a quadratic window centered at the point chosen for maximum perturbation. The size of the maximum displacement is random and is adjusted such that, in that point, the change will be no larger than the estimated modelization error at that point. The displacement field in each window is smooth and has random directions, and at the boundary of each window the displacement is zero. We chose 100 quadratic windows of size 1/16 of the model size for random warping.

4. **A simple velocity perturbation scheme that allows a random change of velocity within one layer**, within given bounds (here $\pm 5\%$). The size of the velocity perturbation is adjusted such that, in any point, the change will be limited by the estimated modelization error.

This composition of the proposal algorithm ensures that the sampling is guided by (1) information about the approximate solution and its errors, and (2) information about the structure of the solution (piecewise constant).

Producing Samples from the Posterior Distribution:

We use the approximate model derived from processing and interpretation as the starting model from our sampling. In 30 percent of the perturbations the velocity

of a random layer is perturbed and the layer boundaries are kept constant. In the rest of the perturbations we perturb the layer boundaries. During the first 300 iterations we replaced the approximate model $\tilde{\mathbf{m}}$ (which is the center of the proposal distribution) with the best-fitting model found so far by the algorithm, and adjusted the modeling error accordingly. In this way we have deliberately introduced a burn-in sequence of 300 iterations in order to improve the acceptance rate of the sampler.

Figures 8 shows four realizations from the posterior probability density produced after 200, 400, 600 and 800 iterations of the IPMC algorithm. Figure 9 compares the convergence of IPMC with a regular Extended Metropolis Algorithm (EMA) (Mosegaard and Tarantola, 1995). The EMA starts in a random realization from the prior, and its proposal distribution uses the same prior as the IPMC. However, as is customary for the EMA, the proposal establishes a random walk with a limited steplength:

$$q(\mathbf{m}'|\mathbf{m}) = \rho(\mathbf{m}')u(\mathbf{m}'|\mathbf{m}) \tag{7}$$

where $u(\mathbf{m}'|\mathbf{m})$ is a uniformly sampling random walk. Note that, in contrast to the IPMC, the EMA proposal does not use the approximate likelihood $\theta(\mathbf{d}_{obs}, \mathbf{m}')$.

Looking at the log-likelihood curves of the two methods (Figure 9) we see a significant difference between their convergence properties. The IPMC algorithm starts advantageously at a reasonable approximation to the solution with a relatively high value of the log-likelihood. From there it proceeds towards equilibrium, which is attained after around 100 iterations. The log-likelihood curve for the MCMC is, however, showing long-term correlations throughout the 3000 iterations. Even the models sampled after 3000 iterations (not shown) show no similarity to the true model, but rather resembles white noise. The proposed IPMC method shows a remarkable improvement in efficiency as compared to the regular MCMC algorithm.

Discussion

It is often claimed that Markov-Chain Monte Carlo methods are highly inefficient for solution of large-scale inverse problems. This statement is only true if simple, local proposal mechanisms are used. In this study we show that a global proposal, incorporating substantial external information about the problem (from approximate physics, and from knowledge about the character of the solution), dramatically changes the situation and allows MCMC to equilibrate much faster. In our case with $\sim 10^6$ model parameters, the Informed Proposal Monte Carlo converged to equilibrium within ~ 100 iterations whereas the classical local-proposal

MCMC was unable to approach any close-to-equilibrium sampling within this time frame.

As always, when comparing algorithms, or differently tuned versions of the same algorithm, one has to be critical about the conditions under which the comparison was carried out. Here, it is easy to see what the difference between IPMC and MCMC is. The IPMC algorithm suggested in this paper is a specialized MCMC algorithm using much more external information about feasible solutions than the regular MCMC implementation. Knowing in this case that the solution is near-piecewise constant is one important piece of information, and knowing that classical processing and interpretation will produce a reasonable solution is another. Since our approximate likelihood enters only via our proposal and assigns nonzero probability to all models, it does not asymptotically bias our solution and is therefore risk-free to use (it does not exclude any models). It can only, in the best case, speed up the sampling. On the other hand, our prior has the special property that it excludes (assigns zero probability) to all non-piecewise-constant models. For this reason it actually biases the solution. However, this is desirable (and intentional), because we want the prior information to have an imprint on the solutions.

When interpreting our numerical results, one should remember the following:

1. The convergence speeds of both the IPMC and the regular MCMC depend on the noise variance of the data. The lower the variance, the slower the convergence.

2. Our data consist only of two shot records. For this reason the solution is not strongly constrained by the data. This is apparent on the four realizations from the posterior probability density shown in Figure 8. More data would make the solution more well-determined and would probably slow down the sampling process.

3. The proposed IPMC uses a global proposal, and for this reason can generate independent samples quite fast. In our numerical example we could produce uncorrelated samples separated by ~ 200 iterations. This, however, does not mean that a complete description of the posterior can be obtained within short time. More that 10^6 samples may be required to completely describe a posterior defined in a 10^6-dimensional model space, and this is independent of the sampling algorithm used.

Conclusions

In this study we have investigated how external physical information can be used to establish a proposal distribution for efficient MCMC sampling of solutions to

an inverse problem. Our test problem was an acoustic full-waveform inverse problem where earth models consistent with data from two seismic shots we produced. Our aim have been to simulate a practical situation where a preliminary, approximate subsurface model created from processing and interpretation of reflection data was available. The preliminary model was used in two different, independent ways: (1) To define a prior distribution assigning nonzero probability only to models consisting of a stack of homogeneous layers intersected by a major fault, and (2) to quantify the modelization error of combined processing and interpretation and to use this information to build a global proposal distribution centered at the preliminary model and with a dispersion proportional to the modelization error. Our study showed how informed proposal distributions can have significant impact on the computational speed of Monte Carlo sampling of solutions to inverse problems.

Acknowledgements

This work was supported by Innovation Fund Denmark through the OPTION Project (5184-00025B). Klaus Mosegaard would like to thank Dr. Amir Khan and colleagues at the Department of Earth Sciences, ETH, for their hospitality and inspiring discussions during the fall 2017 where this work was initiated.

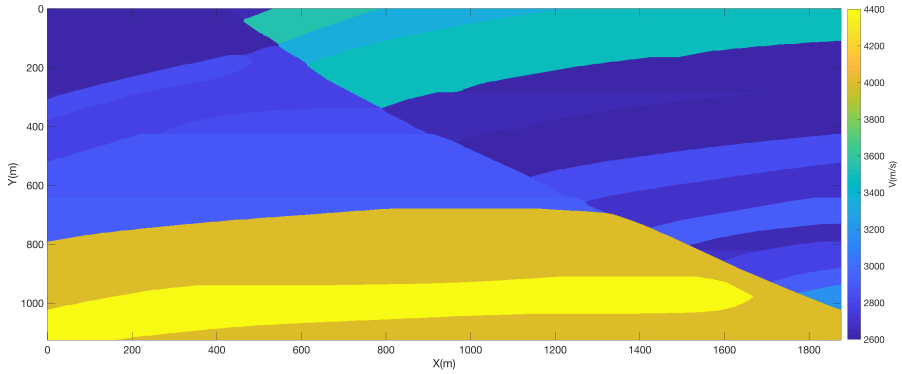

Figure 1: True velocity model

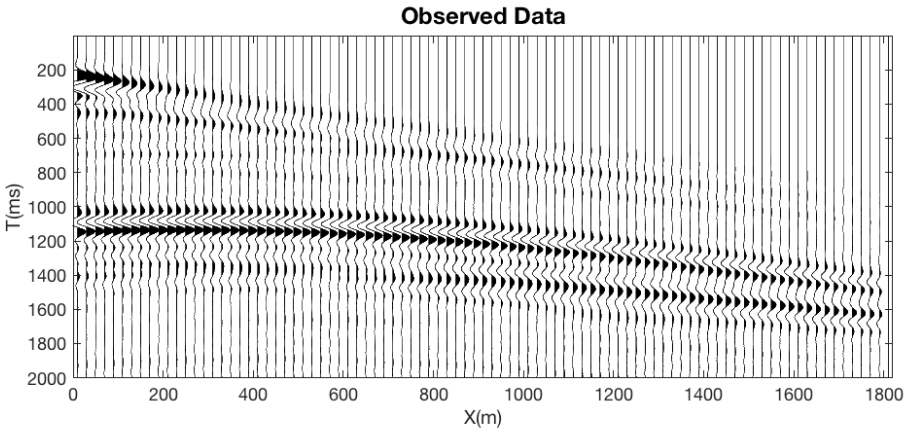

Figure 2: Synthetic shot record generated from true velocity model

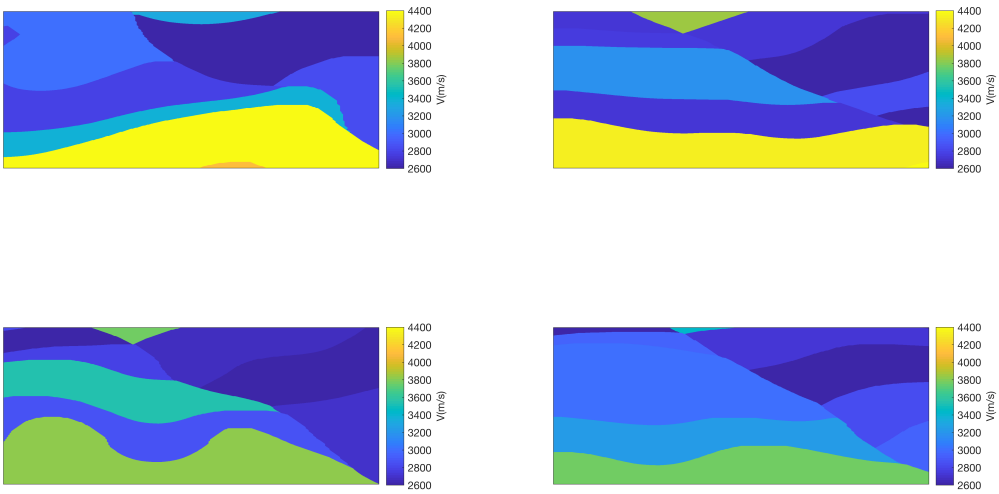

Figure 3: Four realizations from the prior

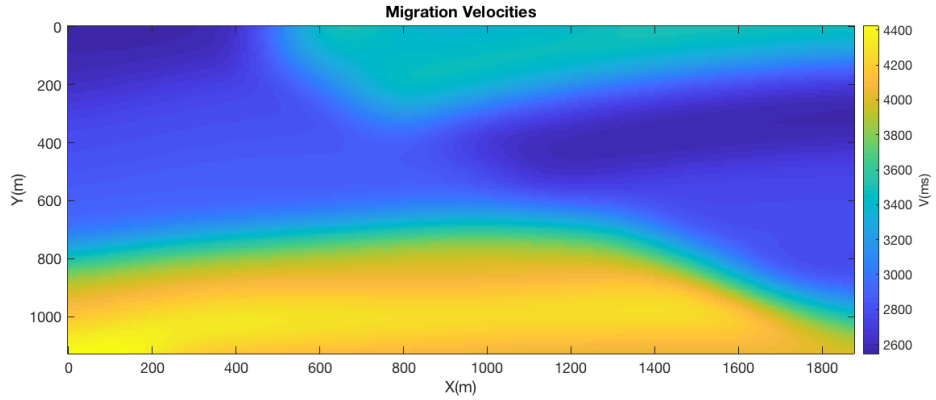

Figure 4: Migration Velocities

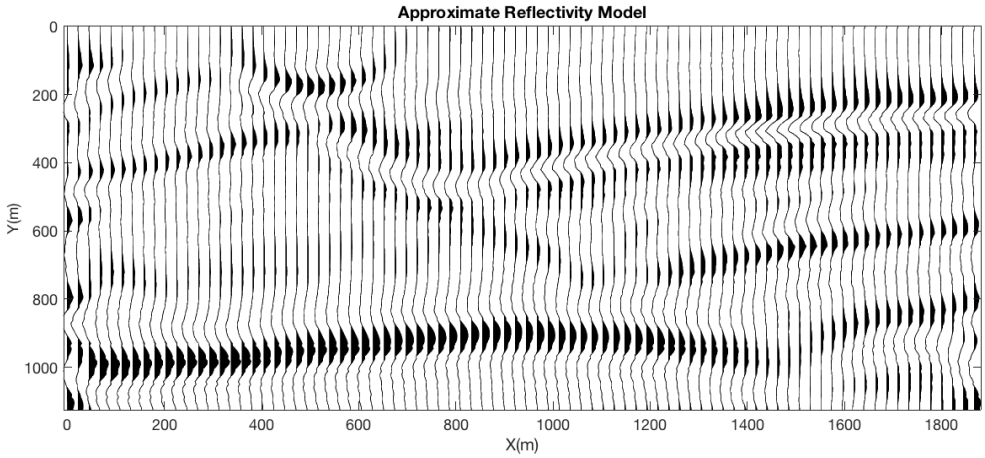

Figure 5: Approximate Reflectivity

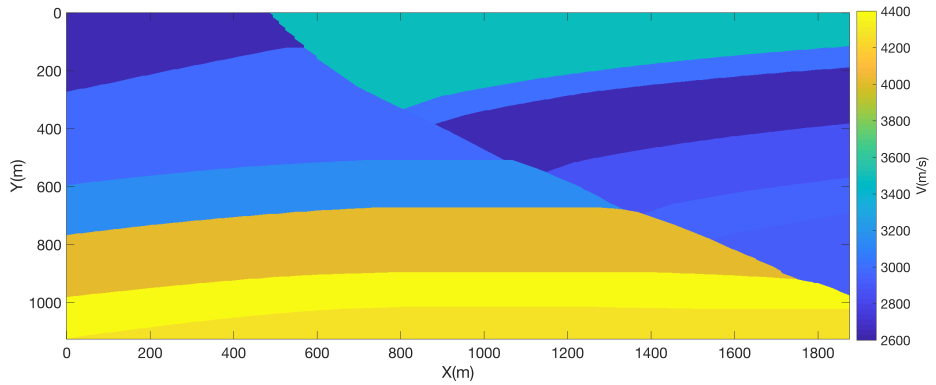

Figure 6: Approximate velocity from interpretation

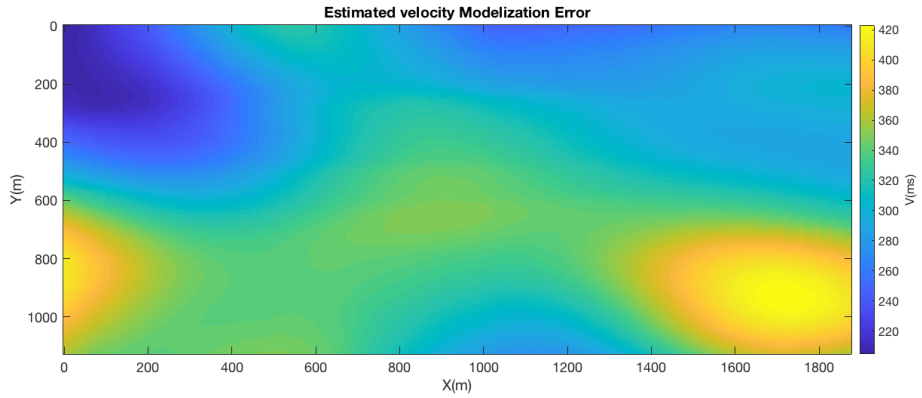

Figure 7: Error envelope obtained from approximation error

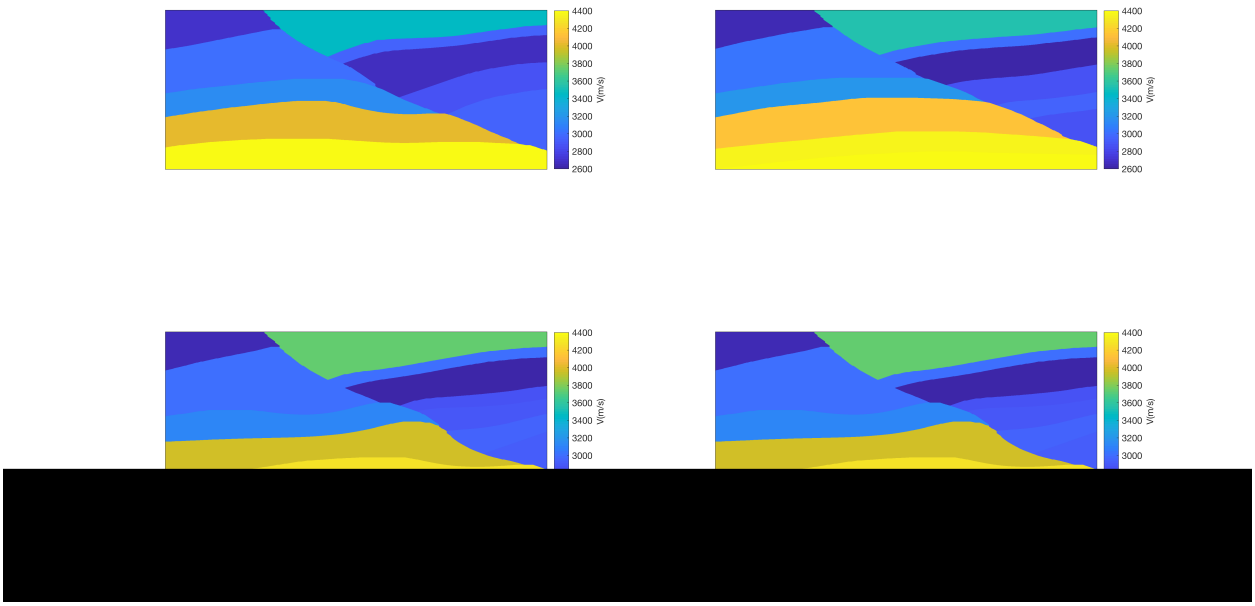

Figure 8: Four realizations from the posterior

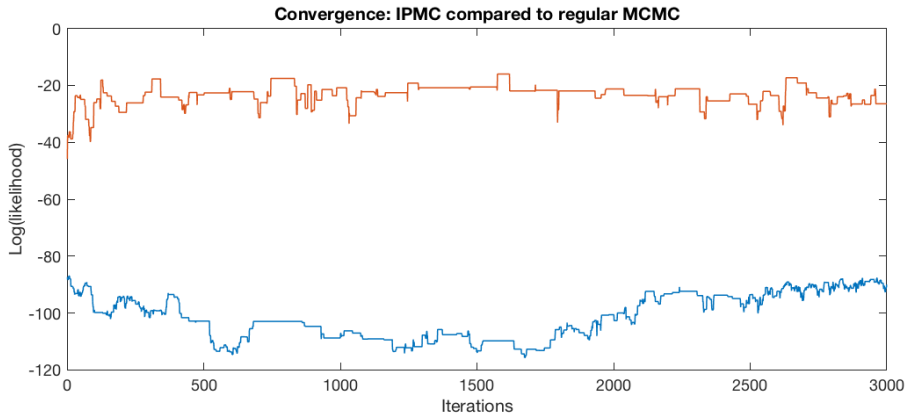

Figure 9: Likelihood vs number of iterations for IPMC compared to a regular MCMC

References

[1] Bayes, T., 1764. An essay towards solving a problem in the doctrine of chances, *Phil. Trans. Roy. Soc.*, **53**, 370–418. Reprinted, with biographical note by G. A. Barnard, 1958, in *Biometrika*, **45**, 293–315.

[2] Bunks, C., Saleck, F. M., Zaleski, S., and Chavent, G., 1995. Multiscale seismic waveform inversion, Geophysics, 60, 1457–1473.

[3] Fichtner, A., Zunino, A., and Gebraad, L., 2018. Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, **216**, 1344-1363.

[4] Gauthier, O., Virieux, J., and Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: numerical results., Geophysics, 51, 1387–1403.

[5] Gebraad, L., Boehm, C., and Fichtner, A., 2020. Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo, J. Geophys. Res., 125.

[6] Igel, H., Djikpesse, H., and Tarantola, A., 1996. Waveform inversion of marine reflection seis- mograms for P impedance and Poisson's ratio., Geophys. J. Int., 124, 363–371.

[7] Khoshkholgh, S., Zunino, A. and Mosegaard, K., 2020. Informed Proposal Monte Carlo. arXiv:2005.14398

[8] Koren, Z., K. Mosegaard, E. Landa, P. Thore, and A. Tarantola, 1991, Monte Carlo Estimation and Resolution Analysis of Seismic Background Velocities: Journal of Geophysical Research, **96**, B12.

[9] Lailly, P., 1983. The seismic inverse problem as a sequence of before stack migrations, in Con- ference on Inverse Scattering: Theory and Application, Soc. Industr. appl. Math., Philadel- phia, PA.

[10] Loewenthal, D., Lu, L., Roberson, R., and Sherwood, J.W.C., 1976, The wave equation applied to migration: *Geophys. Prosp.*, *24*, 380–399.

[11] Jackson, D. D., and Matsu'ura, M., 1985. A Bayesian approach to nonlinear inversion, **90**, 581–591.

[12] Pratt, R., 1990, Frequency-domain elastic modeling by finite differences: a tool for crosshole seismic imaging: Geophysics, 55, 626–632. doi:10.1190/1.1442874

[13] Pratt, R., 1999, Seismic waveform inversion in the frequency domain, part I: theory and verification in a physic scale model: Geophysics, 64, 888–901. doi:10.1190/1.1444597

[14] Pratt, R., Shin, C., and Hicks, G. J., 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, 133, 341–362. doi:10.1046/j.1365-246X.1998. 00498.x

[15] Pratt, R., and Shipp, R., 1999, Seismic waveform inversion in the frequency domain, part 2: fault delineation in sediments using crosshole data: Geophysics, 64, 902–914. doi:10.1190/1.1444598

[16] Tarantola, A., and B. Valette, 1982, Inverse Problems 1?4 Quest for Information: Journal of *Geophysics*, 50, 159?170.?

[17] Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: *Geophysics*, **49**, 1259–1266.

[18] Tarantola, A., 1986, A strategy for nonlinear elastic inversion of seismic reflection data: *Geophysics*, **51**, 1893–1903.

[19] Tarantola, A., 1988, Theoretical Background for the Inversion of Seismic Waveforms, Including Elasticity and Attenuation: *Pure and Applied geophysics* **128**, 1/2, 365–399.

[20] Versteeg, R., 1994. The Marmousi experience: Velocity model determination on a synthetic complex data set. *The Leading Edge* 13 (9): 927–936. doi:10.1190/1.1437051

[21] Virieux, J., and Operto, S., 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, 74, WCC1–WCC26. doi:10.1190/1.3238367

[22] Zhang X., A Curtis, A., 2020. Variational full-waveform inversion. *Geophysical Journal International*, **222,1**, 406-411.