Monte Carlo full waveform inversion of tomographic crosshole data using complex geostatistical a priori information

Knud S. Cordua, Thomas M. Hansen, and Klaus Mosegaard, Technical University of Denmark, Department of Informatics and Mathematical Modelling*

Summary

This paper presents a Monte Carlo full waveform inversion strategy based on a Bayesian formulation of the inverse problem. Existing full waveform inversion strategies often relies on a migration based approach, which suffers from lack of uncertainty estimates. Using a Bayesian approach, the solution to the inverse problem is formulated as an a posteriori probability density. We demonstrate that samples from the solution to the full waveform inverse problem can be obtained using the extended Metropolis algorithm in conjunction with complex a priori information. The a priori information is described by a training image using a geostatistical algorithm. A posteriori samples from the solution to the inverse problem provide a means of obtaining resolution analysis of the solution. The suggested strategy is tested on synthetic crosshole full waveform ground penetrating radar (GPR) data, but is equally well applicable to seismic waveform data. The forward problem is solved using finite-difference time-domain calculations of Maxwell's equations. To our knowledge this is the first example of performing full waveform inversion using the extend Metropolis Algorithm and, in this way, provide an uncertainty estimate of a tomographic full waveform inverse problem.

Introduction

Ground penetrating radar (GPR) crosshole tomography is a popular method used to obtain tomographic images of near-surface geological structures and geophysical parameters. The crosshole GPR experiment involves a transmitting radar antenna (20 MHz – 1 GHz, (see Reynolds, 1997)) lowered into a borehole and a receiving antenna placed in an adjacent borehole. The boreholes are typically separated by a distance of 5m – 20m and are 10 m – 20 m deep. The transmitting antenna is kept fixed in one borehole while the signal is recorded at multiple locations in the opposite borehole. This procedure is repeated for multiple transmitter positions until the inter borehole section has been covered with measurements. For a corresponding seismic experiment the source pulse frequency is typically in the order of 500Hz – 1000 Hz (e.g. Paasche et al., 2006; Belina et al., 2009).

Ray-based inversion of first arrival travel times and amplitudes of GPR or seismic data provide estimates of the signal velocity and attenuation distribution of the inter-borehole region. Unfortunately, ray-theory is based on a high-frequency approximation and accounts only for a small part of the information content of the full waveform signal. Therefore, tomographic images obtained using ray-theory are limited in resolution.

Ernst et al. (2007a,b) introduced and applied an algorithm for inversion of tomographic crosshole GPR data. This code was modified by Belina et al. (2009) and applied to seismic waveform data. Ernst et al. (2007a,b) and Belina et al. (2009) demonstrated that sub-wavelength features can be resolved by properly modelling the wave propagation and, hence, utilizing the full information content of the waveform data. The algorithm introduced by Ernst et al. (2007a,b) relies on the migration-based full waveform inversion concept introduced by Tarantola (1984). This method is desirable because it, in relatively few iterations, is able to infer geophysical parameters in order to minimise a misfit function between the observed and modelled waveforms. However, the migration based approach has some limitations: Firstly, the method depends on a good initial guess of the large scale features of the solution in order to ensure convergence. Secondly, when using this method in a tomographic crosshole setup, a dense receiver and transmitter coverage is needed in order to avoid numerical artefacts along the boreholes. Finally, the method is based on an optimization algorithm and does not provide any uncertainty estimate of the solution.

Here, we formulate the GPR crosshole problem as a probabilistic, Bayesian, full waveform inverse problem. We use the extended Metropolis Algorithm (Mosegaard and Tarantola, 1995) to obtain samples of the a posteriori pdf. In this way, not only a single estimate, but multiple realizations of tomographic images which all honour the data uncertainties and a priori information are provided. This, in turn, provides a means of performing resolution analysis of the solution (Mosegaard, 1998). The a priori information about the problem is controlled through a geostatistical algorithm. This approach allow us to incorporate complex a priori information described by both 2-point and multiple-point statistics. In this study we limit the example only to concern a priori information described by a training image (i.e. multiple-point statistics). We demonstrate how geostatistical formulated a priori information serves as a guide in the initial part of the inversion procedure. In this way convergence of the suggested full-waveform inversion algorithm becomes independent of the initial model.

Monte Carlo full waveform inversion

Methodology

Consider that the subsurface can be represented by a discrete set of model parameters, $\mathbf{m}$, and that a data set, $\mathbf{d}$, of indirect observations of the model parameters is provided. The model parameters describe some physical properties of the subsurface that influences the data observations. Hence, the forward relation between the model parameters (i.e. the model) and the data observations can be expressed as (e.g. Tarantola, 2005):

$$\mathbf{d} = g(\mathbf{m}), \quad (1)$$

where g is a linear or non-linear mapping operator which often relies on a physical law. Here, the forward relation in equation (1) is given as a finite-difference time-domain solution of Maxwell's equations. However, any numerical wave propagation modelling strategy for GPR or seismic signals can be applied. The inverse problem is to infer information about the model parameters based on a set of observations, a priori information about the model, and the forward relation between the model and the data observations.

In a Bayesian formulation the solution to the inverse problem is given as an a posteriori probability density, which can be formulated as (e.g. Tarantola, 2005):

$$\sigma_M(\mathbf{m}) = k\rho_M(\mathbf{m})L(\mathbf{m}), \quad (2)$$

where k is a normalization constant, $\rho_M(\mathbf{m})$ is the a priori probability density, and $L(\mathbf{m})$ is the likelihood function. $\rho_M(\mathbf{m})$ describes the probability that the model satisfies the a priori information. $L(\mathbf{m})$ describes how well the modelled data explains the observed data given a data uncertainty. Hence, the a posteriori probability density describes the probability that a certain model is a solution to the inverse problem.

A highly nonlinear inverse problem refers to the case where the a priori probability density is far from being Gaussian or the forward relation between the model and data are far from being linear. In the case of full waveform inversion the forward relation is expected to be highly nonlinear. Moreover, the a priori information described by a training image is highly non-Gaussian.

The extended Metropolis algorithm is a versatile tool which, in particular, is useful to obtain samples from solutions to non-linear inverse problems using arbitrarily complex a priori information. The minimum requirement of the algorithm is; 1) a "black box" algorithm that is able to sample the a priori probability density and, 2) a "black box" algorithm that is able to compute the likelihood for a given set of model parameters. The flowchart of the extended Metropolis algorithm is as follows: 1) The a priori sampler proposes a sample, $\mathbf{m}_{propose}$, from the a priori probability density, which is a perturbation of a previous accepted model, $\mathbf{m}_{accept}$. 2) The proposed sample is accepted with the probability (known as the Metropolis rule):

$$P_{accept} = \min\left(1, \frac{L(\mathbf{m}_{propose})}{L(\mathbf{m}_{accept})}\right) \quad (3)$$

3) If the proposed model is accepted, $\mathbf{m}_{propose}$ is a sample of the a posteriori probability and $\mathbf{m}_{propose}$ becomes $\mathbf{m}_{accept}$. Otherwise $\mathbf{m}_{propose}$ is rejected. 4) The procedure is continued until a desirable number of models have been accepted.

In this study the algorithm that provides the a priori information is the Single Normal Equation SIMulation (snesim) algorithm, which is a fast geostatistical algorithm that produces samples (conditional or unconditional) from an a priori probability density defined by a training image for a relatively low number of categorical values (Strebelle, 2002). Hansen et al. (2008) suggest a strategy termed perturbed simulation, which is capable of producing perturbations of spatial distributions using geostatistical algorithms. Thus, perturbed simulation serves as a "black box" that produces samples of a priori probability densities described by both two-point and multiple-point statistics. The flow of this algorithm is as follows: 1) An initial unconditional sample of the a priori probability density (here defined by a training image) is provided. 2) A subarea of the sample is randomly chosen. 3) The model parameters within this area are set to unknown. 4) The unknown model parameters are resimulated conditional to the rest of the model parameters using a geostatistical algorithm (here snesim) and a perturbation is obtained. 5) This procedure is repeated in order to obtain multiple samples of the a priori probability density.

The size of the perturbation area governs the exploratory nature of the Metropolis algorithm. The size of the perturbation area is chosen subjectively. In the extreme case where the area covers the entire model the outcome of the perturbed simulation algorithm is uncorrelated to the previous model. Contrary, if the area only constitutes a single model parameter, the perturbed model is highly correlated with the initial model. According to the metropolis rule a small perturbation area results in a proposed model that is more probable of being accepted as compared to a proposed model obtained using a larger perturbation area. Therefore, the perturbation area should

be chosen carefully in order to ensure an efficient algorithm. Gelman et al. (1996) found that the acceptance rate should be around 23% for high-dimensional distributions. For large acceptance rates the algorithm is exploring the a posteriori probability density too slowly. On the other hand, for smaller acceptance rates too many computationally expensive trials are performed. Therefore, we suggest to automatically change the size of the perturbation area while running the algorithm such that a certain acceptance rate is maintained. A constant acceptance rate results in a larger perturbation area in the burn-in period than in the subsequent sampling period. This effect is beneficial because the algorithm needs to perform large perturbations in the initial part in order to find models of large probability and, hence, produce representative samples of the a posteriori probability density.

Finally, the likelihood function is defined as a Gaussian distribution:

$$L(\mathbf{m}) = k \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (g(\mathbf{m})^i - d_{obs}^i) / \sigma^2 \right), \qquad (4)$$

where $g(\mathbf{m})^i$ represents the amplitude of the individual sample points of all the simulated waveforms obtained through equation (1) (i.e. the FDTD algorithm) and d_{obs}^i are the sample points of the observed waveform data. σ is the standard deviation of the expected amplitude uncertainty of the waveform data.

Results and discussion

Figure 1 shows a training image that mimic a matrix of clay with embedded channels of unconsolidated sand. Electromagnetic signals in near surface sediments are sensitive to the dielectric permittivity and the electrical conductivity of the materials. In this study we limit ourselves only to consider the influence of the dielectric permittivity, which is primarily governing the phase velocity of the signal. Water saturation of clay is often high compared to sandy deposits. Therefore, the dielectric permittivity of the clay is set to a relative dielectric permittivity of $\varepsilon_r \approx 4{,}57$ (0,14m/ns) and the permittivity of the sand channels is set to $\varepsilon_r \approx 2{,}75$ (0,18m/ns) (e.g. Topp et al., 1980). Figure 2 (left) is the synthetic reference to be considered and is, at the same time, an unconditional sample of the training image obtained using snesim. The electrical conductivity is set to a constant value of 3 mS/m and is, in the following, assumed known.

A full waveform synthetic data set is calculated using the FDTD algorithm. A Ricker wavelet with a central frequency of 100 MHz is used as source pulse. The source pulse is assumed known during the inversion. The transmitter and receiver positions are separated by 2 m and 0.25 m, respectively (see figure 2 left). Data acquired with a transmitter-receiver angle larger than 45 degrees from horizontal are omitted since, in practice, these data are violated by effects of wave guiding in the boreholes (cf. Peterson, 2001). This leads to a total of 248 data observations (i.e. recorded waveforms).

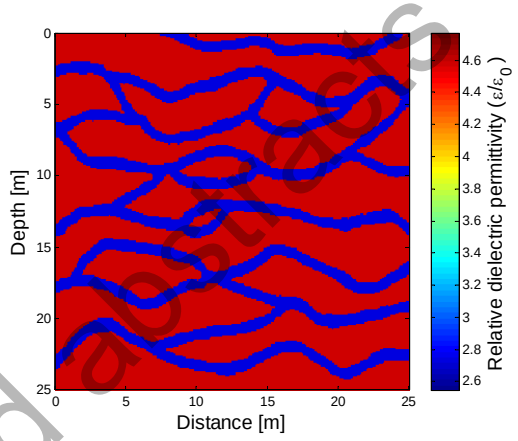

Figure 1. Training image which mimic sandy channel structures embedded in a matrix of clay deposits.

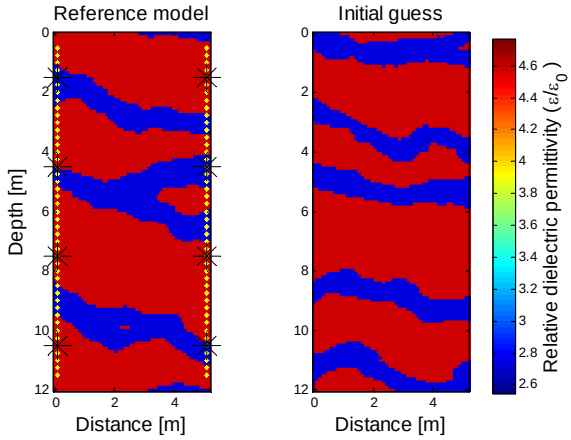

Figure 2. Left) Synthetic reference model. Black asterisks show transmitter positions and the yellow dots show receiver positions. Right) The initial model used as input for the inversion.

Noise is subsequently added to the data by performing a random phase shifting of the synthetic waveforms. The phase shift is normal distributed with zero mean and a standard deviation of 0.4 ns since this is a typical magnitude found in GPR travel time data (e.g. Looms et al, in press). The phase shift results in an amplitude

uncertainty with a standard deviation of 10^{-3}. Accordingly, the standard deviation of the data uncertainties, σ, is set to 10^{-3}. The standard deviation of the amplitude noise is indicated by the red error bar in figure 3 and compared with two waveforms recorded at 0 degrees (short offset / high amplitude) and 45 degrees transmitter-receiver angles (long offset / low amplitude), respectively.

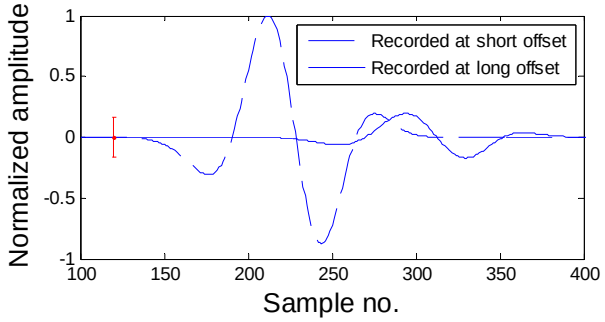

Figure 3. Dashed line is a waveform recorded at short transmitter-receiver offset (0 degrees). Solid line is a waveform recorded at long offset (45 degrees). The height of the red errorbar indicates 2 times the standard deviation of the noise added to the data.

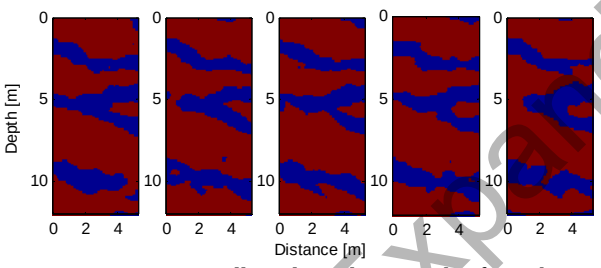

Figure 4. Five statistically independent samples from the a posteriori probability density of the full waveform inverse problem.

The initial model used for the Metropolis algorithm is chosen as an unconditional sample of the training image using a different random seed than for the reference model (see figure 2 right). Burn-in was reached after 2000 accepted models. Hereafter samples accepted by the Metropolis rule are representative samples of the a posteriori probability density. Figure 4 shows the 2000th, 6000th, 1000th, 14000th, and 18000th accepted sample using a priori information defined by the training image (figure 1) and Gaussian data uncertainty (equation 4). Only a slight deviation between the individual samples is seen, which indicates little a posteriori model uncertainty. Figure 5 shows the a posteriori mean and variance based on 18000 samples from the a posteriori probability density after the burn-in period. From the a posteriori variance it is seen that the overall structures of the model is recovered (variance close to or equal to zero) whereas the higher variances along the edges between the clay and sand deposits are subjected to uncertainty. A comparison between the reference model (figure 2 left) and the mean of the samples (figure 4 right) confirms that the waveform inversion is able to recover the sand structures very well. Moreover, it should be noted that the algorithm is able to reach these results even though it is initiated in a model uncorrelated with the reference model (compare figure 2 right and left).

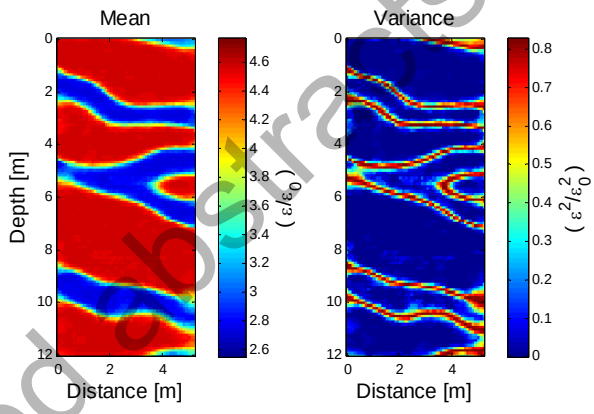

Figure 5. Mean (left) and variance (right) of 18000 samples drawn from the a posteriori probability density of the solution to the full waveform inverse problem.

The suggested Monte Carlo inversion strategy is preferable in that it allows for arbitrary antennae geometry. Furthermore, complex a priori inversion can be included using the perturbed simulation algorithm. Finally, a full data covariance matrix can be included in order to account for correlated data errors, which are often present in tomographic inverse problems (e.g. Maurer and Musil, 2004; Cordua et al., 2008, 2009). However, note that this exhaustive sampling strategy needs substantially more computationally expensive forward calculations compared to the traditional migration based approach.

Conclusions

We have demonstrated the potential of producing samples of the solution to a tomographic full waveform inverse problem using the extended Metropolis algorithm with complex a priori information. The methodology provides a means of evaluating the a posteriori uncertainty, which is not provided using optimisation based strategies for full waveform inversion. Moreover, the present approach is robust with regard to the initial guess of the solution and the transmitter-receiver density. Finally, the extended Metropolis Algorithm is flexible regarding the choice of a priori information and specification of data uncertainties.

References

Belina, F. A., J. R. Ernst, and K. Holliger, 2009, Integration of diverse physical-property models: Subsurface zonation and petrophysical parameter estimation based on fuzzy *c*-means cluster analyses: Journal of Applied Geophysics, **68**, 85 – 94.

Cordua, K. S., M.C. Looms, and L. Nielsen, 2008, Accounting for correlated data errors during inversion of cross-borehole ground penetrating radar data: Vadose Zone Journal, **7**, 263 – 271.

Cordua, K.S., L. Nielsen, M. C. Looms, T. M. Hansen, and A. Binley, 2009, Quantifying the influence of static-like errors in least-squares-based inversion and sequential simulation of cross-borehole ground penetrating radar data: Journal of Applied Geophysics, **68**, 71 – 84.

Ernst, J. R., A. G. Green, H. Maurer, and K. Holliger, 2007a, Application of a new 2D time-domain full-waveform inversion scheme to crosshole radar data: Geophysics, **72**, J53 – J64.

Ernst, J.R., H. Maurer, A. G. Green, and K. Holliger, 2007b, Full-waveform inversion of crosshole radar data based on 2- D finite-difference time-domain solutions of maxwell's equations: IEEE Transactions on Geoscience and Remote Sensing, **45**, 2807 – 2828.

Gelman, A., G. O. Roberts, and W. R. Gilks, 1996, Efficient Metropolis jumping rules: Bayesian statistics, **5**, 599 – 607.

Hansen, T. M., K. Mosegaard, and K. S. Cordua, 2008, Using geostatistics to describe complex a priori information for inverse problems, in J. M. Ortiz and X. Emery, eds., Geostatistics 2008 Chile, vol. 1.

Maurer, H.R. and M. Musil, 2004, Effects and removal of systematic errors in crosshole georadar attenuation tomography: Journal of Applied Geophysics, **55**, 261 -260.

Mosegaard, K., 1998, Resolution analysis of general inverse problems through inverse Monte Carlo sampling: Inverse Problems, **14**, 405 – 426.

Mosegaard, K., and A. Tarantola, 1995, Monte Carlo sampling of solutions to inverse problems: Journal of geophysical research, **100**, 431 – 447.

Paasche, H., J. Tronicke, K. Holliger, A. G. Green, and H. Maurer, 2006, Integration of diverse physical-property models: subsurface zonation and petrophysical parameter estimation based on fuzzy c-means cluster analyses: Geophysics, **71**, H33–H44.

Peterson, J.E., 2001, Pre-inversion correction and analysis of radar tomographic data: Journal of Environmental and Engineering Geophysics, **6**, 1 – 18.

Reynolds, J.M., 1997, An introduction to applied and environmental geophysics: John Wiley and Sons Ltd. Pp.

Strebelle, S., 2002, Conditional simulation of somplex geological structures using multiple-point statistics: Matehmatical geology, **34**, 1 – 21.

Tarantola, A., 2005, Inverse problem theory and methods for model parameter estimation: Society of Industrial and Applies Mathematics, Philadelphia, PA.

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

Topp, G.C., J. L. Davis, and A. P. Annan, 1980, Electromagnetic determination of soil water content: Measurements coaxial transmission lines: Water Resources Research, **16**, 574–582.