Geophysical
Prospecting

EAGE

A gradient-based Markov chain Monte Carlo algorithm for elastic pre-stack inversion with data and model space reduction

Journal:	*Geophysical Prospecting*
Manuscript ID	GP-2020-3420.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Aleardi, Mattia; University of Pisa, Earth Sciences
Keyword:	AVA, Seismic inversion, Uncertainties

EAGE Publications B.V., PO Box 59, 3990 DB, Houten, The Netherlands

A gradient-based Markov chain Monte Carlo algorithm for elastic pre-stack inversion with data and model space reduction

Mattia Aleardi

University of Pisa, Earth Sciences Department, via S. Maria 53, 56126, Pisa, Italy

Corresponding author: Mattia Aleardi, mattia.aleardi@unipi.it

ABSTRACT

The main challenge of Markov Chain Monte Carlo sampling is to define a proposal distribution that simultaneously is a good approximation of the posterior probability while being inexpensive to manipulate. We present a gradient-based Markov Chain Monte Carlo inversion for elastic pre-stack inversion in which the posterior sampling is accelerated by defining a proposal that is a local, Gaussian approximation of the posterior model, while a non-parametric prior distribution is assumed for the distribution of the elastic properties. The proposal is constructed from the local Hessian and gradient information of the log posterior, whereas the non-linear, exact Zoeppritz equations constitute the forward modeling engine for the inversion procedure. Hessian and gradient information is made computationally tractable by a reduction of data and model spaces through a Discrete Cosine Transform reparameterization. This reparameterization acts as a regularization operator in the model space, while also preserving the spatial and temporal continuity of the elastic properties in the sampled models. We test the implemented algorithm on synthetic pre-stack inversions under different signal-to-noise ratios in the observed data. We also compare the results provided by the presented method when a computationally expensive (but accurate) finite-difference scheme is used for the Jacobian computation, with those obtained when the Jacobian is derived from a linearization of the exact Zoeppritz equations. The outcomes of the proposed approach are also compared against those yielded by a gradient-free Monte Carlo sampling method and by a deterministic least-squares inversion. Our tests demonstrate that the gradient-based sampling reaches accurate uncertainty estimations with a much lower computational effort than the gradient-free approach.

Keywords: AVA; Seismic inversion; Uncertainties.

INTRODUCTION

The great challenge in solving geophysical inverse problems lies in the fact that they are usually ill-posed: different combinations of model parameters are consistent with the observed data. No uniqueness in the recovered solution arises from noisy measurements, sparse observations, prior uncertainties, and approximation in the forward model that maps the model parameters into the observed data. The deterministic approach to geophysical inversion guarantees a rapid convergence toward a best-fitting model, but is incapable of accounting for the uncertainties affecting the recovered solution. On the contrary, Bayesian inference provides a systematic framework for incorporating and propagating the uncertainties in observed data, prior knowledge, and forward operator into the uncertainties affecting the recovered model (Tarantola, 2005). The final solution of a Bayesian inversion is the so-called posterior probability density (PPD) function in model space that fully quantifies the uncertainties in the recovered solution. However, an analytical uncertainty assesment is only possible for linear forward operators and Gaussian assumptions about model, data, and noise distributions. In all the other cases, Markov Chain Monte Carlo sampling methods can be used to numerically assess the posterior density (Sen and Stoffa, 1996; Sambridge and Moseegard, 2002; Sen and Stoffa, 2013). However, expensive forward model operators and high-dimensional parameter spaces make the application of MCMC algorithms computationally unfeasible. Indeed, it is known that the sampling ability of these methods dramatically decreases in large-dimensional problems due to the so-called curse of dimensionality issue (Curtis and Lomax 2001). In these contexts, traditional sampling methods might require billions of forward evaluations before converging to stable posterior uncertainties.

More in detail, MCMC algorithms generate samples by perturbing the current state of the chain (current model) according to a proposal distribution. Once generated, the Metropolis-Hasting criterion is used to either accept or reject the proposed sample. This process generates a chain of samples whose distribution asymptotically converges to the target PPD. Theoretically, for an infinite number of samples, the estimated distribution does not depend on the choice of the proposal. However, from a more practical perspective, the Monte Carlo sampling is maximally efficient when the proposal is a good approximation of the target density. For this reason, the definition of an appropriate proposal is of crucial importance for an efficient probabilistic sampling. The setting of an optimal proposal is especially of great importance in large-dimensional parameter spaces, where a significant mismatch between the proposal and the target density can drastically affect the performance of the sampling: persistent rejections of models, entrapment in local maxima of the PPD, and a dramatic increase in the number of forward evaluations needed to attain stable uncertainty estimations. When the classical random walk Metropolis algorithm is employed, a good compromise between the exploitation and exploration of the sampling is usually determined by a trial and error procedure in which different hyperparameters defining the proposal are tuned. However, more sophisticated MCMC recipes can be adopted (e.g. self-adaptive MCMC algorithms, preconditioned MCMC, hybrid MCMC approaches; Tierney and Mira 1999; Haario et al. 1999; Haario et al. 2001; Haario et al. 2006; ter Braak and Vrugt 2008; Turner and Sederberg 2012; Sambridge 2013; Vrugt 2016; Holmes et al. 2017). For example, self-adaptive algorithms, iteratively adjust the proposal to the local shape of the posterior. As an alternative, Gradient-Based MCMC (GB-MCMC) (e.g. Hamiltonian Monte Carlo, Langevin Monte Carlo; Sen and Biswaw; 2017; Fichtner and Simutè, 2018; Fichtner and Zunino, 2019; Fichtner et al. 2019; Gebrad et al. 2020; Aleardi and Salusti 2020; Aleardi, 2020a) exploit the gradient information of the misfit function (the negative natural logarithm of the posterior) to efficiently explore the model space and to rapidly converge toward stable posterior uncertainties (MacKay, 2003; Neal 2011). The main computational requirement of these methods is the need for computing derivatives, although this information is highly beneficial to speeding up the convergence of the sampling and to guarantee high independence of the samples while maintaining high acceptance rates.

It is also well known that MCMC algorithms work well in reduced spaces (Lieberman et al. 2010), and hence a popular approach to deal with high-dimensional problems is to use a reparameterization strategy that decreases the number of unknowns. In this case, the full state space is projected onto a limited number of basis functions and the algorithm generates samples in this reduced domain. This technique must be applied taking in mind that part of the information in the original (unreduced) parameter space could be lost in the reduced space and for this reason, the model parameterization must always constitute a compromise between model resolution and model uncertainty (Dejtrakulwong et al. 2012; Lochbühler et al. 2014; Aleardi 2019; Grana et al. 2019; Aleardi 2020b).

Here we propose a sampling strategy in which a gradient-based MCMC algorithm is combined with a compression of data and model space through a Discrete Cosine Transform (DCT). In particular, on the line of Martin et al. (2012), we exploit the geometrical properties of the misfit function to greatly speed up the probabilistic sampling. The approach is derived by analogy with the classical Newton approach to deterministic inversion and it defines a proposal density based on a local Gaussian approximation to the target PPD informed by local Hessian information. We apply this strategy to solve a Bayesian amplitude versus angle (AVA) inversion in which the subsurface elastic properties of P-wave velocity (*Vp*), S-wave velocity (*Vs*), and density are inferred from partially stacked seismic data at different incidence angles, while the exact Zoeppritz equations constitute the forward modeling operator. In our approach the dimensions of the Jacobian matrix are significantly reduced through a compression of both model and data spaces, thereby rendering the Hessian and Gradient manipulations computationally feasible. The DCT expands a signal (e.g. expressing the subsurface *Vp* model) into a series of cosine functions oscillating at different frequencies. The low-order discrete cosine transform coefficients express most of the variability of the original signal, and the model compression is simply accomplished by zeroing the numerical coefficients beyond a certain threshold. Therefore, the compression also helps to reduce the ill-conditioning of the inversion and mitigate the curse of dimensionality issue.

A crucial aspect of AVA inversion is the preservation of both the mutual and spatial/temporal relationships between the elastic parameters as inferred, for example, from available well log data (Aleardi et al. 2015). Usually, to avoid inverting large matrices, the AVA inversion is solved for each seismic gather independently. However, with this strategy, the spatial continuity of the elastic properties in the predicted model could be lost, especially in case of severe noise contamination. In this context, the advantage of the DCT lies in the possibility to apply this transformation to multidimensional signals as well (e.g. 2-D images). In this case, the order of the retained non-zero coefficients determines the wavelength of the recovered, compressed image along different (i.e. vertical, horizontal) directions. In our implementation, the compression is applied both to the elastic parameters (*Vp*, *Vs,* and density) and the seismic data that are treated as 2-D and 3-D images, respectively. This strategy allows for a simultaneous estimation of the elastic parameters over the entire considered area while guaranteeing the preservation of the temporal and spatial continuity of the elastic properties in all the sampled models.

After discussing the theoretical aspects of the proposed inversion scheme, we consider an analytical probability density function to illustrate the benefits of the implemented GB-MCMC algorithm over standard gradient-free MCMC approaches. Then, the method is applied to synthetic seismic data computed on a realistic subsurface elastic model that mimics a clastic geological setting in which a turbiditic sequence host gas saturated sand intervals. The outcomes of the implemented GB-MCMC approach are also validated and compared with those yielded by a gradient-free MCMC sampling (i.e. the Differential Evolution Markov Chain “DEMC”; Vrugt 2016) still running in the reduced data and model spaces and with those provided by a linearized least-squares algorithm that inverts each seismic gather separately working in the full model and data spaces. The proposed approach needs computing the Jacobian matrix associated with each sampled model. Therefore, we also compare the predictions provided by two GB-MCMC implementations: The former uses a computationally intensive, but accurate forward finite-difference scheme to compute the Jacobian matrix around each considered model. The latter replaces the Jacobian with a matrix operator derived from a linear approximation of the exact Zoeppritz equations after projection onto the compressed space (Aleardi and Salusti, 2020).

The main novelty of this paper is the combination of a gradient-based MCMC sampling and a DCT compression of both data and model space to efficiently solve the Bayesian non-linear pre-stack inversion.

METHODS

Gradient-based MCMC sampling

Gradient-based deterministic inversions are aimed at minimizing a previously defined misfit function, which usually is a linear combination of data error and a model regularization term. For Gaussian-distributed noise and model parameters, the error function can be written as follows (Menke 2018; Aster et al. 2018):

$$E(\mathbf{m}) = \left|\left|\mathbf{C}_d^{-\frac{1}{2}}(\mathbf{d} - G(\mathbf{m}))\right|\right|_2^2 + \left|\left|\mathbf{C}_m^{-\frac{1}{2}}(\mathbf{m} - \mathbf{m}_{prior})\right|\right|_2^2, \qquad (1)$$

where the vectors $\mathbf{m}$ and $\mathbf{d}$ identify the model parameters and the observed data, respectively; $\mathbf{C}_d^{-1/2}$ and $\mathbf{C}_m^{-1/2}$ are the data and prior model covariance matrices; $\mathbf{m}_{prior}$ is the prior model vector, and G is the forward modeling operator that maps the model into the corresponding data. The minimum of E $(\mathbf{m})$ can be iteratively approached through a local quadratic approximation of the error function around the current model $\mathbf{m}_k$:

$$E(\mathbf{m}) = E(\mathbf{m}_k + \Delta\mathbf{m}) \approx \tilde{E}(\mathbf{m}) == E(\mathbf{m}_k) + \Delta\mathbf{m}^T\nabla_{\mathbf{m}}\mathbf{E}(\mathbf{m}_k) + \frac{1}{2}\Delta\mathbf{m}^T\nabla^2_{\mathbf{m}}\mathbf{E}(\mathbf{m}_k)\Delta\mathbf{m} + O(||\Delta\mathbf{m}||^3), \quad (2)$$

where $\Delta\mathbf{m} = \mathbf{m} - \mathbf{m}_k$, whereas $\nabla_{\mathbf{m}}\mathbf{E}(\mathbf{m}_k)$ and $\nabla^2_{\mathbf{m}}\mathbf{E}(\mathbf{m}_k)$ represent the first and second derivative of $E(\mathbf{m})$ computed around $\mathbf{m}_k$. In particular, it results that:

$$\nabla_{\boldsymbol{m}}\mathbf{E}(\mathbf{m}_k) = \mathbf{g} = \mathbf{J}^T\mathbf{C}_d^{-1}\Delta\mathbf{d}(\mathbf{m}_k) + \mathbf{C}_m^{-1}(\mathbf{m}_k - \mathbf{m}_{prior}), \quad (3)$$

and

$$\nabla^2_{\mathbf{m}}\mathbf{E}(\mathbf{m}_k) = \mathbf{H} = \left(\mathbf{J}^T\mathbf{C}_d^{-1}\mathbf{J}\right)^{-1} + \frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}\mathbf{C}_d^{-1}\left(\Delta\mathbf{d}(\mathbf{m}_k)...\Delta\mathbf{d}(\mathbf{m}_k)\right) + \mathbf{C}_m^{-1} = \mathbf{H_o} + \mathbf{B} + \mathbf{C}_m^{-1}, \quad (4)$$

where $\Delta\mathbf{d}(\mathbf{m}_k) = G(\mathbf{m}_k) - \mathbf{d}$, $\mathbf{B} = \frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}\mathbf{C}_d^{-1}\left(\Delta\mathbf{d}(\mathbf{m}_k)...\Delta\mathbf{d}(\mathbf{m}_k)\right)$, whereas $\mathbf{J}$ denotes the Jacobian matrix expressing the partial derivative of the data with respect to model parameters. In practical applications and for computational feasibility reason, the Hessian matrix is approximated as $\mathbf{H} \approx \mathbf{H}_{\boldsymbol{a}} = \mathbf{H_o} + \mathbf{C}_m^{-1}$, thus neglecting the partial derivative of the Jacobian with respect to the model. The number of rows and columns of the Hessian is equal to the number of data points and model parameters, respectively. The quadratic approximation of the error function can be compactly written as:

$$\tilde{E}(\mathbf{m}) = \frac{1}{2}\left(\mathbf{m} - \mathbf{m}_k + \mathbf{H}_a^{-1}\mathbf{g}\right)^{\boldsymbol{T}}\mathbf{H}_{\boldsymbol{a}}\left(\mathbf{m} - \mathbf{m}_k + \mathbf{H}_a^{-1}\mathbf{g}\right) + const., \quad (5)$$

Equation 5 shows that the minimizer of $\tilde{E}(\mathbf{m})$ can be computed as

$$\mathbf{m}_{k+\mathbf{1}} = \mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g}, \quad (6)$$

where $\mathbf{H}_a^{-1}\mathbf{g}$ is called the Newton step. In the context of deterministic inversions, an approximated uncertainty quantification can be computed from the inverse of the Hessian matrix at the convergence point. In other terms, a local quadratic approximation of the inverse of the curvature of the error function gives the uncertainties affecting the recovered solution.

Differently, a Bayesian inversion aims to estimate the full posterior distribution in the model space given by:

$$p(\mathbf{m} \,|\, \mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \quad (7)$$

where $p(\mathbf{m} \,|\, \mathbf{d})$ is the posterior probability density (PPD), $p(\mathbf{d}|\mathbf{m})$ is the so-called likelihood function, whereas $p(\mathbf{m})$ and $p(\mathbf{d})$ are the a-priori distributions of model parameters and data, respectively. For problems in which the $p(\mathbf{m} \,|\, \mathbf{d})$ can not be expressed in a closed form, an MCMC algorithm can be used for a numerical assessment of the posterior model. In this context, the probability to move from the current state of the chain $\mathbf{m}_k$ to the next, proposed state $\mathbf{m}_{k+\mathbf{1}}$ is determined according to the Metropolis-Hasting rule:

$$\alpha = p(\mathbf{m}_{k+\mathbf{1}} \,|\, \mathbf{m}_k) = \min\left[1, \frac{p(\mathbf{m}_{k+\mathbf{1}})}{p(\mathbf{m}_k)} \times \frac{p(\mathbf{d}|\mathbf{m}_{k+\mathbf{1}})}{p(\mathbf{d}|\mathbf{m}_k)} \times \frac{q(\mathbf{m}_k|\mathbf{m}_{k+\mathbf{1}})}{q(\mathbf{m}_{k+\mathbf{1}}|\mathbf{m}_k)}\right], \quad (8)$$

where $q(.)$ is the proposal distribution that defines the new state (i.e. model) $\mathbf{m}_{k+\mathbf{1}}$ as a random deviate from a probability distribution $q(\mathbf{m}_{k+\mathbf{1}}|\mathbf{m}_k)$ conditioned only on the current state $\mathbf{m}_k$. The proposal ratio term vanishes if symmetric proposals are used. For example, the most popular proposal strategy uses a Gaussian step centered on the current state $\mathbf{m}_{k+\mathbf{1}} = \mathbf{m}_k + \mathcal{N}(0,\mathbf{C})$, where $\mathbf{C}$ is the selected covariance matrix of the proposal and $\mathcal{N}$ denotes the Gaussian distribution. This method is referred to as the Random Walk Metropolis. If $\mathbf{m}_{k+1}$ is accepted, $\mathbf{m}_k = \mathbf{m}_{k+1}$. Otherwise, $\mathbf{m}_k$ is repeated in the chain and another state is generated as a random deviate from $\mathbf{m}_k$. The ensemble of sampled states after the burn-in period is used to numerically compute the statistical properties (e.g. mean, mode, standard deviations, marginal densities) of the target posterior probability. Now we can formulate the Bayesian inversion framework in terms of $E(\mathbf{m})$, $\mathbf{H}$ and $\mathbf{g}$, under Gaussian assumptions for data, noise, and model parameter distributions; we can write (Tarantola, 2005):

$$p(\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{prior})^{\boldsymbol{T}} \mathbf{C}_{\boldsymbol{m}}^{-\mathbf{1}} (\mathbf{m} - \mathbf{m}_{prior})\right), \quad (9)$$

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{d} - G(\mathbf{m}))^{\boldsymbol{T}} \mathbf{C}_{\boldsymbol{d}}^{-\mathbf{1}} (\mathbf{d} - G(\mathbf{m}))\right), \quad (10)$$

$$p(\mathbf{m} \,|\, \mathbf{d}) \propto \exp(-E(\mathbf{m})), \quad (11)$$

If we substitute equation 5 into equation 11 we obtain the approximation of the posterior around $\mathbf{m}_k$:

$$p(\mathbf{m} \,|\, \mathbf{d}) \approx \tilde{p}(\mathbf{m} \,|\, \mathbf{d}) \propto \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}\left(\mathbf{m} - \left(\mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g}\right)\right)^{\boldsymbol{T}} \mathbf{H}_{\boldsymbol{a}} \left(\mathbf{m} - \left(\mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g}\right)\right)\right), \quad (12)$$

Equation 12 indicates that the approximation of the PPD is Gaussian distributed $\tilde{p}(\mathbf{m}|\mathbf{d}) = \mathcal{N}(\mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g};\mathbf{H}_{\boldsymbol{a}})$ with mean equal to the minimizer of $\tilde{E}(\mathbf{m})$ and covariance equal to the inverse of the Hessian matrix. After constructing a local Gaussian approximation of the posterior density, we can now define a sampling method that uses the following proposal density:

$$q(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{m} - \left(\mathbf{m}_k - \lambda\mathbf{H}_a^{-1}\mathbf{g}\right)\right)^T \frac{\mathbf{H}_{\boldsymbol{a}}}{\mu^2}\left(\mathbf{m} - \left(\mathbf{m}_k - \lambda\mathbf{H}_a^{-1}\mathbf{g}\right)\right)\right). \quad (13)$$

Each proposed model is accepted according to the Metropolis Hasting rule taking in mind that in this case the proposal is not symmetric and for this reason, the proposal ratio should be evaluated. However, since the proposal is Gaussian both $q(\mathbf{m}_{k+\mathbf{1}}|\mathbf{m}_k)$ and $q(\mathbf{m}_k|\mathbf{m}_{k+\mathbf{1}})$ can be analytically computed. λ and μ^2 are tunable parameters that determine the step length along the negative gradient direction and the variance of the random perturbation around the minimizer of $\tilde{E}(\mathbf{m})$. These parameters must be properly set to get the desired acceptance rate or in other words to find a good compromise between exploitation and exploration of the sampling. More in detail, the λ value should be large enough to make the proposal dependent on the gradient information, but small enough so that the model update is not dominated by the deterministic information. On the contrary, the $\mu^{\mathbf{2}}$ value should be large enough to ensure an efficient exploration of the model space, but small enough so that the gradient information is not completely masked by the random update. We will consider the full Hessian and not only its diagonal entries so that possible posterior correlations between the inverted parameters are fully taken into account.

Therefore, we have "tailored" the proposal density $q(\mathbf{m})$ to the underlying local Gaussian approximation of the posterior probability using the derivative information of the error function. From a practical point of view, the proposed model can be straightforwardly generated according to:

$$\mathbf{m}_{k+\mathbf{1}} = \mathbf{m}_k - \lambda\mathbf{H}_a^{-1}\mathbf{g} + \mu\mathbf{H}_{\boldsymbol{a}}^{-\frac{1}{2}}\mathbf{n}, \quad (14)$$

with $\mathbf{H}_{\boldsymbol{a}}^{-\mathbf{1}} = \mathbf{H}_{\boldsymbol{a}}^{-\mathbf{1/2}}(\mathbf{H}_{\boldsymbol{a}}^{-\mathbf{1/2}})^T$, whereas $\mathbf{n}$ is a random column vector with the number of rows equal to the number of model parameters drawn from $\mathcal{N}(0,\mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix.

Note that for a Gaussian PPD and an exact Hessian, the proposed method results in a perfect sampling, in which all the samples are independently drawn from the posterior density with an acceptance probability equal to 1 (Martin et al. 2012). Also, note that for $\mu = 0$ equation 14 gives the standard gradient descent model update. On the contrary, if $\lambda = 0$ we have the standard random walk with some constraints given by the inverse Hessian. It can also be demonstrated that the previous GB-MCMC approach is related to the Hamiltonian Monte Carlo and Langevin Monte Carlo approaches (Martin et al. 2012). Finally, even though the proposal is derived by assuming a local Gaussian assumption, it can be used to sample from whatever type of posterior model and under whatever a-priori assumption (e.g. non-parametric), as it has been done in the following examples.

The major computational requirement of the implemented approach is the need for computing the Jacobian associated with each sampled model. When the forward is expressed by a partial differential equation the adjoint state method can be used to rapidly estimate the Gradient and Hessian with a reduced number of forward evaluations. The Jacobian can be also evaluated using a finite-difference scheme or in the case of weakly non-linear problems, a linearized approximation of the non-linear forward operator can be adopted as well. An extra computational workload also arises in large dimensional spaces due to the manipulation of large Hessian matrices and gradient vectors. In these contexts, a compression strategy would be useful to reduce the number of data points and model parameters and hence the dimensions of $\mathbf{H}_{\boldsymbol{a}}$ and $\mathbf{g}$. If a finite difference scheme is employed, the compression of the model parameter space also reduces the number of forward evaluations needed for the Jacobian computation.

Discrete Cosine Transform

Several variants of discrete cosine transform exist with slightly modified definitions, but in this work, we use the so-called DCT-II formulation that is the most common (Britanak et al. 2010). Hereafter we simply refer to the DCT-II as the DCT. We employ this parameterization because it exhibits superior compression power over other compression methods (Lochbühler et al. 2014).

This compression technique can be applied to multidimensional signals (i.e. 2-D matrices) and such multi-dimensional transform follows straightforwardly from the one-dimensional definition because it is simply a separable product (equivalently, a composition) of DCTs along each dimension. For example, if we assume a 2-D density model $\boldsymbol{\rho}(x,y)$ in which x=[0,1,…,M_x-1] and y=[0,1,…,M_y-1] represent the horizontal and vertical coordinates, respectively, the associated 2-D transform is defined as follows:

$$\begin{cases} \mathbf{R}(k_x,k_y) = \dfrac{1}{\sqrt{M_x}}\dfrac{1}{\sqrt{M_y}}\displaystyle\sum_{x=0}^{M_x+1}\sum_{y=0}^{M_y+1}\boldsymbol{\rho}(x,y),\ if k_x = k_y = 0 \\ \mathbf{R}(k_x,k_y) = \sqrt{\dfrac{2}{M_x}}\sqrt{\dfrac{2}{M_y}}\displaystyle\sum_{x=0}^{M_x+1}\sum_{y=0}^{M_y+1}\boldsymbol{\rho}(x,y)cos\left(\dfrac{(2x+1)\pi k_x}{2M_x}\right)cos\left(\dfrac{(2y+1)\pi k_y}{2M_y}\right), if k_x,k_y \neq 0 \end{cases} ,(15)$$

where $\mathbf{R}(k_x,k_y)$ represent the k_x-th and k_y-th coefficient. The values within the matrix $\mathbf{R}$ represent the unknowns to be estimated in a reparametrized inverse problem. Equation 15 can be compactly rearranged in matrix form:

$$\mathbf{R} = \mathbf{B}_y\boldsymbol{\rho}\mathbf{B}_x^T, \quad (16)$$

where $\mathbf{B}_x$ and $\mathbf{B}_y$ are the matrices with dimensions $M_x \times M_x$ and $M_y \times M_y$, respectively that contain the basis functions spanning the compressed space, whereas the $M_y \times M_x$ matrix $\mathbf{R}$ expresses the DCT coefficients. This approach concentrates most of the information of the original signal into the low-order coefficients, and hence an approximation of the subsurface density model can be obtained as follows:

$$\tilde{\boldsymbol{\rho}} = \left(\mathbf{B}_y^q\right)^T\mathbf{R}_{qp}\mathbf{B}_x^p, \ (17)$$

where $\tilde{\boldsymbol{\rho}}$ is the approximated $[M_y \times M_x]$ density model, $\mathbf{B}_y^q$ is a $[q \times M_y]$ matrix containing only the first q rows of $\mathbf{B}_y$; $\mathbf{B}_x^p$ is a $[p \times M_x]$ matrix containing only the first p rows of $\mathbf{B}_x$, whereas the matrix $\mathbf{R}_{qp}$ represents the first q rows and p columns of $\mathbf{R}$. In other words, the scalar q and p represent the retained number of base functions along the y and x directions used to derive the approximated model. Therefore, the DCT transformation allows for a reduction of the $(M_y \times M_x)$-D full density model space to a $(q \times p)$-D DCT-compressed parameter space with $p < M_x$ and $q < M_y$. In this context the $p \times q$ non-zero numerical coefficients of the $\mathbf{R}_{qp}$ matrix becomes the unknowns to be estimated after a compression of the model space. Estimating the retained coefficients reduces the parameter space dimensionality and can significantly improve the computational efficiency of the inversion procedure. Figure 1 shows some DCT base functions of different orders in a 2-D space. Note that the variability of the solution along each dimension is directly determined by the orders of the retained coefficients.

The implemented AVA inversion scheme

We consider a 2-D subsurface model in which the parameters to be estimated are the *Vp*, *Vs,* and density values. The observed data are partial angle stacks computed by separately applying the Zoeppritz equations to the elastic properties at each spatial location. For a $M_y \times M_x$ subsurface model and for *N* incidence angles (usually *N*=3; near, mid, and far stacks), we have $3 \times M_y \times M_x$ model parameters to be estimated from $N \times (M_y - 1) \times M_x$ data points. The spatial dimensions M_y, M_x are usually large, and hence the simultaneous estimation of the *Vp*, *Vs,* and density over the entire study area becomes computationally impractical for both deterministic and MCMC methods: In the former case, the large dimension of the Hessian and gradient matrices makes their manipulation and/or computation problematic. In the latter, the convergence of the probabilistic sampling is hampered by the curse of dimensionality issue. For this reason, common deterministic and probabilistic inversion approaches separately estimate the elastic properties at each seismic gather location. This means that M_x inversions are run, each one estimating $3 \times M_y$ parameters from $3 \times (M_y - 1)$ observations. Although this method makes the inversion computationally feasible it does not preserve the spatial continuity on the predicted elastic models. To overcome this issue, we compress both data and model space. In more details, the *Vp*, *Vs,* and density models are treated as separate $M_y \times M_x$ images to which the 2-D DCT is applied. Therefore, each $M_y \times M_x$ matrix expressing a given elastic property is approximated with a reduced number of coefficients contained within a $q \times p$ matrix $(p < M_x$ and

$q < M_y$). This reduces the full $(3 \times M_y \times M_x)$-D elastic space to a compressed $(3 \times p \times q)$-D space. The compression is also applied to decrease the dimensionality of the data space. In this case, we apply a 3-D DCT in which the first two coordinates represent the spatial and temporal directions, while the third axis identifies the incidence angles. The application of this transformation allows for a reduction of the original $(N \times (M_y - 1) \times M_x)$-D data space to a $(b \times v \times \mathrm{c})$-D space with $b < N$, $v < M_y - 1$ and $c < M_x$ (Figure 2). The map between the full data and model spaces is constituted by the Zoeppritz equations that are separately applied to the elastic properties at each spatial location and provide the seismic gathers associated with each sampled model.

In this context, the GB-MCMC algorithm samples the compressed $(3 \times p \times q)$-D model space and estimate the DCT coefficients expressing the elastic properties from the retained $b \times v \times \mathrm{c}$ basis in the data space. This means that the computation of the proposal ratio, likelihood ratio, and prior ratio for each sampled model (see equation 8) is performed in the compressed model and data domains. A schematic representation of this strategy is given in Figure 3. We note that multiple forward and inverse transformations are needed in each iteration. However, these transformations can be run with a negligible computational cost. The sampled models after the burn-in phase are projected onto the elastic space through equation 17 to numerically compute the statistical characteristics of the PPD in the *Vp*, *Vs,* and density domain.

We assume a non-parametric prior for the elastic parameters in order to properly model their facies-dependent behavior, while a stationary Gaussian variogram expresses their lateral and temporal variability. Similarly, we assume a Gaussian noise model. The non-parametric prior in the elastic domain impedes an analytical derivation of the prior in the compressed space and for this reason, the prior model in the compressed space is numerically computed by applying the kernel density estimation algorithm to prior elastic realizations projected onto the DCT space. Differently, the assumed Gaussian noise model allows for an analytical derivation of the data covariance matrix in the compressed data space.

The main limitation of any GB-MCMC approach arises from the need for computing the gradient of the posterior model, and hence this strategy is usually applied to problems in which such derivative information can be computed quickly (Neal, 2011). In our case of elastic pre-stack inversion, the Jacobian matrix can be derived, for example, by adopting an accurate but computationally quite expensive forward finite-difference scheme. In this case, $3 \times p \times q$ forward modeling runs are needed to compute the Jacobian associated with the current compressed model. The good news is that each column of the Jacobian can be independently computed and hence the finite difference computation can be easily distributed across different cores.

Another and much less demanding strategy replaces the Jacobian with an analytical operator derived from a linear approximation of the full Zoeppritz equations (for example the linear equation proposed by Aki and Richards, 1980) properly projected onto the compressed model and data spaces (Aleardi, 2020). Note that, in this case, we are inherently assuming that the curvature of the misfit function, and hence the variance of the proposal distribution is constant over the entire model space. This simplification could decrease the convergence speed of the algorithm, but dramatically reduces the computing time of the entire sampling with respect to the finite difference strategy (Aleardi and Salusti, 2020). However, it should be also noted that any linear approximation of the Zoeppritz equations, although widely employed in AVA studies, is theoretically valid in case of weak elastic contrasts at the reflecting interfaces and within a limited angle range (usually not beyond 30-35 degrees). For this reason, the suitability of this approach should be evaluated case-by-case. In the following, we solve the GB-MCMC inversion using both approaches. Their results are also validated against those provided by a gradient-free sampling approach (i.e. the Differential Evolution Markov Chain; DEMC, Vrugt, 2016) running in the reduced model and data spaces, and with the outcomes of a linearized least-squares inversion running in the full elastic and data spaces and inverting each seismic gather separately.

RESULTS

Analytical example

Before applying the GB-MCMC algorithm, we briefly illustrate the benefits provided by the gradient-based sampling over a more standard, gradient-free sampling method. In this section, we aim to draw samples from a posterior model derived from the 2-D Rosenbrock function. This function has challenging features: its minimum is located at the bottom of a narrow parabolic valley where a small change in direction can lead to a steep increase of the gradient. The Rosenbrock can be turned into a probability density that maintains the same basic characteristics of the original function, and hence it has been frequently adopted to test sampling methods (Christen and Fox, 2010; Pagani et al. 2019). In this example the posterior model can be expressed as follows:

$$p(x,y) \propto \exp\left(-\left(100\left(y-x^2\right)^2+(1-x)^2\right)\right). \quad (18)$$

We compare the GB-MCMC approach with a random walk Metropolis (RWM). Both algorithms have been run for 80000 iterations employing 10 parallel chains and under uninformative prior for the x and y variables. The standard deviation of the proposal distribution for the random walk Metropolis has been properly set in order to get an acceptance rate lying in the interval [0.2, 0.4].

Figure 4 illustrates that both MCMC approaches provide similar posterior estimations, in good agreement with the target density. To assess the convergence of the sampling we use the potential scale reduction factor (PSRF), a popular convergence diagnostic tool proposed by Brooks and Gelman (1998) to which we refer the reader for its formal definition. This tool compares within-chain variances to the variance computed from all mixed chains for a given parameter. In practice, one can consider that the convergence to a stable posterior model has been achieved if the potential scale reduction factor is lower than 1.1. By the inspection of the PSRF evolution for the two unknown parameters (Figures 5a-c), we observe that 50000 iterations are needed by the random walk Metropolis to converge toward a stable posterior, while the GB-MCMC converges after only 2000 iterations. This significant difference is related to the different sampling strategies employed by the two approaches. Indeed, Figures 5d-e illustrates that the GB-MCMC not only focus the sampling around the most promising zones of the parameter space (i.e. those characterized by high posterior density values) but also uses a proposal that incorporates information about the local covariance structure of the target density as provided by the inverse of the Hessian matrix. Differently, the random walk proposal is not influenced by the local, geometrical properties of the target posterior and thus the proposed model could also be located far away from the posterior maximum. This difference also indicates that the acceptance rate for the GB-MCMC is usually much higher than that of the random walk Metropolis: in this example, the GB-MCMC acceptance oscillates around 60-80%, whereas only the 30%, on average, of the proposed states, were accepted by the random walk Metropolis. This is another strength of the gradient-based sampling methods because avoid wasting computing time to run forward evaluations for proposed models with a low probability to be accepted.

Synthetic inversion tests

For the lack of available real seismic data, we discuss synthetic experiments in which we applied the implemented approach to invert seismic data generated on a reference model that simulates a realistic geological context in which a turbiditic sequence hosts a gas-saturated reservoir (see Figure 6a). This subsurface model has been derived by integrating the borehole information provided by several wells with an accurate geologic interpretation. The true model represents an in-line section with 61 time samples and 91 cross-lines. The time sampling is 0.004 s, whereas 50 m is the cross-line distance. Figure 6a shows that significant elastic contrasts occur at the interface separating the encasing shales from the reservoir sands.

A forward modeling based on the full Zoeppritz equations computes the observed seismic gathers by convolving the angle-dependent reflectivity series with a 30-Hz, zero-phase Ricker wavelet. We consider three partial angle stacks corresponding to incidence angles of 0 (near stack), 20 (mid stack), and 40 (far stack) degrees. This means that the full model space comprises $61 \times 91 \times 3 = 16653$ parameters to be estimated from $60 \times 91 \times 3 = 16380$ data points.

Two columns extracted from the reference models at the horizontal coordinates of 1000 and 3000 m have been considered as available well log data, used to derive the prior information. We assume a non-parametric distribution for the elastic parameters, which has been computed by applying the kernel density estimation algorithm (Parzen 1962) to the available well log information (Figure 6b). We also assume a stationary 2D Gaussian variogram model in which the vertical and lateral ranges have been inferred from the vertical variability of the available well log data and the lateral variability of the observed seismic data, respectively. The ranges of the variogram are equal to 0.008 s and 160 m along the temporal (vertical) and spatial (lateral) directions, respectively.

The previously defined elastic prior model must be projected onto the DCT space where the MCMC sampling runs. To this end and given the non-parametric prior, we adopt a Monte Carlo simulation approach. The direct-sequential co-simulation method with joint probability distribution (Horta and Soares, 2010) has been used to draw 5000 2-D elastic models in accordance with the prior assumptions. Such models have been projected onto the compressed space, and the kernel density algorithm has been again employed to numerically compute the non-parametric prior in the reduced domain for the *Vp*, *Vs,* and density (Figure 6c). Two examples of *Vp*, *Vs,* and density prior realizations are represented in Figure 7.

The next step is to define the optimal number of coefficients needed to approximate the elastic profiles. To this end, we quantified how the explained variability of the elastic properties changes as the number of basis functions increases. The selection of the optimal number of coefficients is a very delicate step that must guarantee uncertainty estimations, model resolution, and data fitting comparable to those achieved by an inversion running in the full, uncompressed space. A detailed discussion on how the model and data compressions affect the AVA inversion results is far beyond the scope of this work and for this reason, we refer the interested reader to Grana et al. (2019) for more theoretical insights. Figure 8 shows the explained variability for a *Vp*, *Vs,* and density model drawn from the prior as the number of retained coefficients increases. We note that only 25 coefficients per elastic property (q=p=25) along the two DCT spatial dimensions explain almost the total variability of the three elastic parameters. This means that the compression allows for a reduction of the 16653-D model space to $25 \times 25 \times 3 = 1875$-D domain. A similar analysis has been carried out on some seismic gathers derived from prior elastic realizations. An example is shown in Figure 9a where the green rectangle encloses the $40 \times 45 \times 2 = 3600$ retained coefficients in the data space that explain almost the total variability of the uncompressed seismic gather (Figure 9b). Therefore, in this case, the full 16380-D data space has been reduced to a 3600-D domain. These data and model parameter reductions not only guarantees a considerable speed-up in the finite-difference Jacobian computation but also drastically reduces the computational cost of the Hessian and gradient manipulation. For example, the 16653×16653 Hessian in the full domain has been reduced to a 3600×3600 matrix in the compressed space.

In the following inversion tests, we consider two different scenarios: in the former (Test 1) the data computed on the reference model have been contaminated with uncorrelated Gaussian random noise with a standard deviation of 0.03 that corresponds to the 20% of the total standard deviation of the noise-free dataset. However, the popular assumption of uncorrelated noise usually constitutes an oversimplification because in real data applications correlated noise can be ascribed, for example, to residual of multiple reflections or diffractions not successfully removed during the processing phase. For this reason, in the second example (named Test 2 in the following), the observed data have been contaminated with both incoherent and coherent Gaussian noise with the same standard deviation value of 0.06. The temporal and lateral correlation pattern of the coherent noise is the same as the elastic prior model.

In what follows, we discuss the results provided by the GB-MCMC approaches when the Jacobian is computed with a forward finite-difference scheme (GB-MCMC-FD), and when the Jacobian is replaced by the linear operator derived from the Aki and Richards equation (GB-MCMC-L). The outcomes of these inversions are also benchmarked against the predictions of a gradient-free DEMC inversion still running in the reduced model and data spaces and with the results provided by a deterministic linearized least-squares inversion running in the full data and model spaces and inverting each seismic gather independently. All the considered MCMC inversions take advantage of parallel implementations. In the DEMC and GB-MCMC-L each chain is run in parallel, while the

GB-MCMC-FD runs the chains serially but distributes the Jacobian computation across different cores.

We start with the results of Test 1 in which the noise model and the source wavelet are assumed perfectly known during the inversion phase. Figure 10 and Figure 11 show the posterior mean models and posterior standard deviations provided by the GB-MCMC-FD and GB-MCMC-L approaches, respectively. The GB-MCMC-FD and GB-MCMC-L have been run for 10000 iterations and employing 10 independent chains. Both algorithms yield similar and congruent estimates of the posterior mean and the associated uncertainties. We note that the posterior standard deviation increases as the velocities and density values increase. This indicates that the curvature of the error function is expected to change over the model space. Figure 12 compares the elastic properties extracted at two different spatial coordinates (1200 and 2500 m, respectively) with the posterior mean and the 95 % confidence interval estimated by the two GB-MCMC algorithms. We observe that the mean model closely reproduces the vertical variations of the true model, and more importantly, the posterior mean usually lies within the range depicted by the 95% confidence interval, thus ensuring us about the reliability of the final predictions. As an example, Figure 13 shows a comparison between the observed data and the data predicted on the mean model provided by the GB-MCMC-L inversion. The close similarity between the two seismic datasets prove that the predicted mean model can accurately reproduce the observed seismic amplitudes. A similar conclusion would have been drawn for the GB-MCMC-FD algorithm. Figures 10-13 demonstrate that both algorithms provide similar and congruent model and uncertainty estimations, thereby confirming that in both cases a stable posterior model has been reached within the selected number of iterations. For both the GB-MCMC-FD and GB-MCMC-L inversion we set the λ and μ^2 values to 0.2 and 0.95, respectively. This combination resulted in an acceptance rate oscillating around 0.7-0.85.

However, If we analyze the evolution of the negative log-likelihood we observe that the two GB-MCMC implementations are characterized by different convergence speeds toward the stationary regime (Figure 14). The GB-MCMC-FD converges to the steady-state in less than 5 iterations, while

50 iterations are needed by the GB-MCMC-L, although in both cases the same final likelihood value has been reached. This fact is related to the different strategies used to define the Jacobian matrix. In other words, a more accurate Jacobian reflects into a more accurate estimate of the local curvature of the error function thus guaranteeing a faster convergence toward the stationary regime. At a closer inspection, we also observe another difference between the two approaches (see the two close-ups on the right of Figure 14). The GB-MCMC-FD shows strongly variable misfit values with iterations, while the GB-MCMC-L misfit oscillates with a longer period. This proves that the use of an accurate Jacobian guarantees the sampling of maximally decoupled models, while for a linear approximation the successively sampled models are mutually correlated. Therefore, the sampling is expected to attain accurate uncertainty estimations with a lower number of iterations when the finite-difference strategy is adopted (MacKay, 2003). Indeed, Figure 15 shows the evolution of the potential scale reduction factor for all the model parameters in the compressed space and for the two algorithms. In both cases, as expected, the sampling converges faster for the *Vp* coefficients since this is the elastic parameter better constrained by the data, while a longer sampling is needed to attain stable PPDs for the *Vs* and density coefficients (i.e., *Vs* and density are less informed by the seismic data). From the evolution of the PSRF values, we can claim that the GB-MCMC-FD attains convergence for all the parameters with 1000 iterations, while 4000 iterations are needed by the GB-MCMC-L. However, the computational costs of a single GB-MCMC-FD and GB-MCMC-L iteration are very different: 30 s for the former and just 2.5 s, for the latter. This means that, despite the less accurate approximation of the Hessian matrix, the GB-MCMC-L attains convergence in less than 3 hours, while more than 8 hours are needed by the GB-MCMC-FD (see Table 1).

Figure 16 compares the assumed Gaussian correlograms and the average vertical and spatial correlograms computed on the true model and on the posterior solution provided by the GB-MCMC-L inversion. We observe that the assumed correlogram is well reproduced by the estimated model, which also shows a good agreement with the actual lateral and temporal variability patterns. The match between the marginal distributions derived on the true model with those computed on the posterior mean GB-MCMC-L solution also demonstrates that the implemented method guarantees a good reproduction of the actual distribution of the elastic parameters in the investigated area (Figure 17).

We now present the results of the DEMC and the linearized inversion for Test 1. For the DEMC we employ 10 parallel chains evolving for 100000 iterations, with a burn-in period of 70000 samples. In Figure 18a we observe that the linearized approach estimates elastic profiles affected by lateral scattering related to noise propagation from data to model space. This method converges in a few seconds to the final solution but it hampers accurate uncertainty assessments. Figure 18b shows that the DEMC algorithm has not reached accurate model estimations and stable uncertainty appraisals within the selected number of iterations. In particular, we note scattered standard deviation maps completely different from those provided by the two GB-MCMC algorithms. Indeed, the evolution of the negative log-likelihood values (Figure 19) proves that the gradient-free sampling has not even reached the stationary regime within the selected number of iterations. We point out that the acceptance rate of the DEMC oscillated around the optimal values of 0.22-0.33. For this reason, the slow convergence toward the steady-state is not related to an erroneous hyperparameters setting but to the difficulty in sampling the high-dimensional parameter space starting from random prior realizations. In other terms, due to the curse of dimensionality issue, a much higher number of iterations is now needed for accurate uncertainty estimations. To reduce the burn-in stage, the starting model can be generated from the results of a previous inversion step (for example a fast analytical inversion; de Figueiredo et al. 2018). The total computing time for running 100000 DEMC iterations was 11.1 hours, while a single iteration of this approach takes on average 0.4 s (Table 1). However, since the gradient-free sampling does not even reach the stationary regime within the selected number of iterations, we envisage that a much higher computing time is needed to achieve stable posterior assessments. The results of Test 1 indicate that, although the extra time needed for vector/matrix manipulation and Jacobian computation, both gradient-based MCMC algorithms outperform the gradient-free method because they achieve accurate model estimations and uncertainty appraisals with a much lower computational effort.

In the second test, we want to assess the applicability of the proposed approach to a more realistic scenario with a low signal-to-noise ratio and both coherent, and uncorrelated noise affecting the data. For the sake of conciseness, we will only present the results provided by the GB-MCMC-L and linearized approaches. Indeed, on the one hand, the two GB-MCMC strategies still provided very similar predictions, with the GB-MCMC-FD still needing a lower number of iterations to converge, but a higher computing time with respect to the GB-MCMC-L. On the other hand, the DEMC was again severely affected by the curse of dimensionality issue: thus, it would have needed a much higher computing time than the two GB-MCMC implementations to attain stable posterior estimations. In this example, only the uncorrelated Gaussian random noise is taken into account by the data covariance matrix, while the source wavelet is again assumed to be known. The hyperparameter setting for the GB-MCMC-L inversion is the same previously used in Test 1.

Figure 20 compares the results of the deterministic and GB-MCMC-L algorithms. We observe that the inclusion of coherent noise and the overestimation of the signal-to-noise ratio of the data has severely decreased the overall quality of the predictions. The linearized inversion provides final estimates severely affected by lateral scattering. In this case, the lateral formation boundaries of the main gas-saturated reservoir can not be mapped with reasonable accuracy. Differently, in the GB-MCMC-L predictions, we can still appreciate the significant decrease of *Vp*, *Vs,* and density occurring at the interface separating the reservoir sand and the encasing shale. As expected, the posterior uncertainty is increased with respect to the previous example (compare Figures 20c and 11b), such as the sample-by-sample difference between the observed and predicted seismic amplitudes (Figure 21). The direct comparison of the outcomes of the GB-MCMC-L and deterministic approach better highlights the superior predictions achieved by the proposed approach (Figure 22): the mean model estimated by the GB-MCMC is usually closer to the true model than the deterministic results. However, differently from the previous example we now note that the erroneous assumption in the statistical properties of the noise results in estimated confidence intervals that sometimes do not include the true model. Finally, the inspection of the evolution of the potential scale reduction factor (Figure 23) shows that similarly to Test 1, the GB-MCMC-L reaches accurate uncertainty appraisals for all the model parameters in 4000 iterations, approximately.

DISCUSSION

The aim of this work was twofold: implementing an sampling algorithm for accurate and fast uncertainty assessments in non-linear AVA inversion and mitigating the curse-of-dimensionality issues, thus allowing for a simultaneous estimation of the elastic properties along the entire considered 2-D section. To this end, we combined a GB-MCMC sampling with a DCT reparameterization of both data and model spaces.

We compared two different implementations of the GB-MCMC method: The first uses a finite-difference scheme to compute the Jacobian (named GB-MCMC-FD), while the second replaces the Jacobian with a matrix operator derived from a linearization of the Zoepprtiz equation (named GB-MCMC-L). Theoretically, the validity of the linear approximation of the Zoeppritz equations depends on the considered angle range and the magnitude of the elastic contrasts at the reflecting interfaces. However, in our tests, this strategy provided satisfactory model predictions and uncertainty quantifications comparable to those yielded by the GB-MCMC-FD algorithm, although the reference model was characterized by significant elastic contrasts at the interface separating the encasing shale from the reservoir sand. Besides, the GB-MCMC-L approach made it also possible for a significant reduction of the computational cost of a single GB-MCMC inversion. This reduced computational effort occurs at the expense of a slower convergence toward the stationary regime and to an overall decrease in the independence of successively sampled models. The choice of replacing the Jacobian with the linear matrix operator must be considered case-by-case and should constitute a compromise between the sampling efficiency and the computational cost of the GB-MCMC inversion. Another possibility is to employ the finite-difference strategy only for the first iterations when the stationary regime is not yet attained and hence maintaining the same Jacobian during the sampling stage. This recipe should guarantee a faster convergence toward the steady-state and a more efficient sampling, with a limited extra computational cost.

The computing times shown in Table 1 refer to Matlab codes running on a single server equipped with two deca-core intel E5-2630 at 2.2 GHz (128 Gb RAM). So there is still room for a substantial decrease of the computational costs of the GB-MCMC inversion, for example by running the codes on a large computer cluster or utilizing fast computing units and/or adopting a more efficient implementation (e.g., codes written in a lower-level programming language). The computational cost of the GB-MCMC inversion related to the computation of the inverse Hessian can be also reduced by dropping the off-diagonal entries of $\mathbf{H}_{\boldsymbol{a}}$. This strategy results in a proposal distribution that neglects the possible correlation between model parameters, which might have a negative impact on the convergence rate of the sampling.

A proper setting of the hyperparameters λ and μ is important for the efficiency of the sampling. Indeed, a poorly chosen parameter combination would result in a slow convergence toward stable uncertainty estimations. A good parameter combinations would guarantee a good compromise between exploitation and exploration, rendering reasonable acceptance rates. The λ parameter acts as the step length in gradient descent methods. Its value should be similar to the one used in gradient-based local optimization methods so that the linearized Taylor expansion is still locally valid. The μ parameter determines the variance of the proposal distribution: A too small μ would results in poor mixing, while a too-large μ would decrease the acceptance rate. In our experiments, we found the optimal combination using a trial-and-error procedure in which our goal was to get an acceptance rate around 0.7-0.8 during the sampling stage. Another viable strategy could be employing a self-adaptive scheme (Haario et al., 2001; Atchadé, 2006) that automatically adjusts the proposal variance during the sampling process. However, the many inversion tests we carried out showed that the optimal acceptance rate can be achieved by many different parameter combinations and hence a proper selection of λ and μ is not that hard to find; for example in the previous tests a good compromise between exploitation and exploration is guaranteed for λ and μ^2 values lying in the range [0.1, 0.7] and [0.5, 1.5], respectively. From our experiments also emerged that if the approximated Hessian is used, a good λ values should lie in the range]0, 1] because a higher λ puts more emphasis on the exploitation while penalizing the exploration. On the other hand, an optimal μ^2 value is usually around 1. Appendix A uses a didactic example to analyze the effect of the λ and μ^2 values on the sampling efficiency.

The implemented method can be also extended to 3-D models and in this case, a 4-D transformation must be used to compress the data space. Some experiments on a 3D elastic model with 61 time samples, 91 cross-line and 91 in-line have been carried out employing the same Matlab implementation previously considered. In these preliminary tests, the DCT allowed for a compression of the full 1515423-D elastic space into a $25 \times 25 \times 25 \times 3 = 46875$-D domain. However, the current Matlab implementation and the limited available hardware resources make the computation of the Jacobian, the derivation of the inverse Hessian, and also the manipulation of both the Hessian and Gradient, prohibitive. In this context, the GB-MCMC-FD approach is unfeasible, while the GB-MCMC-L works but requires more than a week of computing time to converge. For this reason, a more scalable inversion code and additional hardware resources are needed to invert 3D data. Regarding the performance scaling of the adopted GB-MCMC recipe, Martins et al. (2012) observed similar convergence rates for different model space dimensionalities. They claimed that although this desirable characteristic is not yet proved theoretically, the numerical observations seem to indicate an insensitivity of convergence of the proposed GB-MCMC method to the parameter dimension. We refer the interested reader to Martins et al. (2012) for a more in-depth discussion of this aspect.

CONCLUSIONS

We presented a gradient-based MCMC method for casting the non-linear elastic pre-stack inversion into a solid Bayesian framework that also guarantees fast convergence toward stable PPD assessments. The key idea is to guide the parameter sampling by exploiting the gradient and Hessian information of the PPD, thereby generating proposal densities that are locally a good approximation of the target posterior. This results in a proposal distribution that is easy to construct, and in an increased efficiency of the probabilistic sampling: the gradient guides the sampling toward “better” solutions, whereas the random perturbation term avoids entrapments in local maxima of the PPD. The good compromise between the gradient and random perturbation (that is the optimal compromise between exploitation and exploration) can be found by adjusting two hyperparameters (λ and μ). We reduced the computational effort related to Hessian and gradient manipulation and Jacobian computation by employing a discrete cosine transform reparameterization of data and model spaces.

Our synthetic inversion experiments showed very promising results, in which the posterior mean model well reproduced the ground truth even when coherent noise contaminates the seismic gathers, and for erroneous assumptions about the noise properties. Our results indicated that the exploitation of the Hessian and gradient information always guarantees a much faster convergence toward stable uncertainty estimations than a gradient-free MCMC algorithm. The use of the finite-difference scheme reduced the number of iterations needed to achieve stable PPDs, but it required a significant extra computational cost per iteration for the Jacobian computation. Deriving the Jacobian from a linear approximation of the Zoeppritz equations decreased the sampling efficiency (e.g. slower convergence toward the stationary regime and increase of the correlation value between successively sampled models) but it also greatly reduced the computing time to attain convergence. However, the applicability of this strategy should be evaluated case-by-case because the validity of the linearization of the Zoeppritz equations depends on the considered angle range and the elastic contrasts at the reflective interfaces. The computational cost of the GB-MCMC inversion is orders of magnitude higher than that of the deterministic approaches. However, the main advantage of any MCMC algorithm over deterministic inversions is the possibility to evaluate the posterior uncertainties.

Conflict of interest

The author declares no conflict of interest.

Data Availability

Data available on request from the authors.

APPENDIX A

To investigate in more detail the effects of the λ and μ^2 values on the sampling efficiency of the GB-MCMC inversion we consider a simple example with a 2-D multivariate target density. We run two different tests: in the first, we set $\lambda = 0.05$ and $\mu^2 = 3$, whereas in the second $\lambda = 0.5$ and $\mu^2 = 0.5$. Both tests use 5 independent chains running for just 1000 iterations. Figure 24 demonstrates that in both cases we get a reasonable prediction of the target density, despite the limited maximum number of iterations considered. However, the inspection of the PSRF highlights that in the first test more than 500 iterations are needed to reach the threshold of convergence, while in the second case the convergence is attained in less than 150 iterations. This proves that in the first case we select a too low λ value and a too-high μ^2 thus meaning that we are promoting the exploration at the expense of the exploitation. Instead, in the second case, the hyperparameter setting guarantees an efficient sampling of the parameter space that results in a rapid convergence toward a stable PPD.

FIGURE LEGENDS

Figure 1: 2-D DCT base functions of different orders. Dark and light colors code low and high numerical values, respectively.

Figure 2: Derivation of data and model space vectors in the DCT space from the elastic properties and seismic gathers.

Figure 3: Schematic representation of the GB-MCMC inversion scheme. Green and pink rectangles refer to steps performed in the reduced and full spaces, respectively.

Figure 4: a) True posterior density function. b) Posterior density provided by the random walk Metropolis. c) Posterior density estimated by the GB-MCMC. The colormap codes the normalized probability values.

Figure 5: a) Evolution of the potential scale reduction factor for the random walk Metropolis. b) Evolution of the potential scale reduction factor for the GB-MCMC. c) Close-up of b). In a)-c) the horizontal dotted green lines represent the threshold of convergence, whereas the blue and red lines refer to the *x* and *y* variables, respectively. d) Example of current, proposed model, and proposal distribution for the random walk Metropolis. e) Example of current, proposed model, and proposal distribution for the GB-MCMC. In d) and e) the magenta curves represents the contour lines of the proposal while the colored curves are the contour lines of the Rosenbrock error function.

Figure 6: a) The elastic properties of *Vp*, *Vs,* and density of the reference model. In a) the black arrows point toward the main sand reservoir body, whereas the dotted red lines depict the columns of the model considered as available well log data for defining the a-priori elastic distribution. b) The marginal non-parametric prior distributions for the three elastic properties derived from the two wells shown in a). c) The marginal prior projected onto the compressed space through a Monte Carlo simulation.

Figure 7: a), b) Two examples of *Vp*, *Vs,* and density model drawn from the non-parametric elastic prior.

Figure 8: Examples of explained model variability for an elastic model extracted from the prior and as the number of coefficients along the 1st and 2nd DCT dimension increases. In each plot, the numerical value with coordinate (*x*, *y*) indicates the explained variability if the first *x*, and *y* coefficients along the 1st and 2nd dimensions, respectively, are used for compressing the model. It emerges that 25 coefficients along both the 1st dimension explain almost the 100 % of the variability of the uncompressed *Vp*, *Vs,* and density profiles.

Figure 9: a) DCT decomposition of a seismic gather computed on an elastic model drawn from the prior. Blue and red colors code low and high values, respectively while the green rectangles enclose the retained coefficients in the data space. b) Explained data variability as the number of considered basis functions increases.

Figure 10: Results provided by the GB-MCMC-FD approach for Tests 1. a) A-posteriori mean model. b) Posterior standard deviation. In a), and b) the *Vp*, *Vs,* and density are represented from left to right.

Figure 11: As in Figure 10 but for the GB-MCMC-L approach.

Figure 12: Comparison between the true model, the posterior mean, and 95% confidence interval at two different spatial locations. a) GB-MCMC-FD. b) GB-MCMC-L. The leftmost plot refers to the spatial position of 1200 m, while the plot on the right refers to the spatial position of 2500 m.

Figure 13: Comparison between observed data (left column), predicted data (central column), and their sample-by-sample difference (right column) for Test 1. The predicted data have been computed on the mean posterior model estimated by the GB-MCMC-L algorithm. a), b), and c) refer to near, mid and far stack, respectively.

Figure 14: Evolution of the negative log-likelihood values for the GB-MCMC-FD and GB-MCMC-L inversions (part a) and b), respectively). Each color represents a different chain.

Figure 15: Evolution of the potential scale reduction factor over iterations for the DCT coefficients associated with the three elastic properties. a) GB-MCMC-FD. b) GB-MCMC-L. The red dotted lines depict the threshold of convergence.

Figure 16: Comparison between the lateral (a) and vertical (b) assumed correlogram functions with the average correlograms computed on the true model (blue line) and on the posterior mean estimated by the GB-MCMC-L algorithm (red lines). From left to right we represent *Vp*, *Vs,* and density.

Figure 17: Marginal probabilities for the three elastic parameters computed on the true model and on the posterior mean estimated by the GB-MCMC-L algorithm.

Figure 18: a) Results of the linearized least-squares inversion. b) Estimated mean model by the DEMC algorithm. c) Posterior standard deviation estimated by the DEMC algorithm.

Figure 19: Evolution of the negative log-likelihood value during the DEMC sampling. Each color refers to a different chain.

Figure 20: Results for Test 2: a) *Vp*, *Vs*, and density profiled estimated by the linearized least-squares approach. b) Posterior mean model provided by the GB-MCMC-L approach. c) Posterior standard deviation estimated by the GB-MCMC-L inversion.

Figure 21: Comparison between observed data (left column), predicted data (central column), and their sample-by-sample difference (right column) for Test 2. The predicted data have been computed on the mean posterior model estimated by the GB-MCMC-L algorithm. a), b) and c) refer to near, mid and far stack, respectively.

Figure 22: Comparison between the true model, the deterministic inversion results, the posterior mean, and 95% confidence interval estimated by the GB-MCMC- L approach. a) refers to the spatial position of 1200 m, while b) refers to the spatial position of 2500 m.

Figure 23: Evolution of the potential scale reduction factor over iterations and for the coefficients associated with the three elastic properties. The dotted red lines depict the threshold of convergence.

Figure 24: GB-MCMC sampling of a 2D multivariate density for different hyperparameter settings. a) $\lambda = 0.05$ and $\mu^2 = 3$. b) $\lambda = 0.5$ and $\mu^2 = 0.5$. From left to right we represent the target probability density, the estimated probability density, and the evolution of the PSRF for the two parameters. Blue and yellow colors code low and high probability values, respectively. On the rightmost plot, the horizontal dotted green line represents the threshold of convergence, whereas the blue and red lines refer to the *x* and *y* variable, respectively.

REFERENCES

Aki, K., and Richards, P. G. (1980). Quantative seismology: Theory and methods. New York, 801.

Aleardi, M., and Salusti, A. (2020). Hamiltonian Monte Carlo algorithms for target-and interval-oriented amplitude versus angle inversions. Geophysics, 85(3), R177-R194.

Aleardi, M. (2020a). Combining discrete cosine transform and convolutional neural networks to speed up the Hamiltonian Monte Carlo inversion of pre-stack seismic data. Geophysical Prospecting, 68(9), 2738-2761.

Aleardi, M. (2020b). Discrete cosine transform for parameter space reduction in linear and non-linear AVA inversions. Journal of Applied Geophysics, 104106.

Aleardi, M. (2019). Using orthogonal Legendre polynomials to parameterize global geophysical optimizations: Applications to seismic-petrophysical inversion and 1D elastic full-waveform inversion. Geophysical Prospecting, 67(2), 331-348.

Aleardi, M., Mazzotti, A., Tognarelli, A., Ciuffi, S., and Casini, M. (2015). Seismic and well log characterization of fractures for geothermal exploration in hard rocks. Geophysical Journal International, 203(1), 270-283.

Aster, R. C., Borchers, B., and Thurber, C. H. (2018). Parameter estimation and inverse problems. Elsevier.

Atchadé Y. F., (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift: Methodology and Computing in applied Probability, 8, 2, 235–254.

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 7(4), 434-455.

Britanak, V., Yip, P. C., and Rao, K. R. (2010). Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations. Elsevier.

Christen, J. A., and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t-walk). Bayesian Analysis, 5(2), 263-281.

Curtis, A., and Lomax, A. (2001). Prior information, sampling distributions, and the curse of dimensionality. Geophysics, 66(2), 372-378.

de Figueiredo, L. P., Grana, D., Bordignon, F. L., Santos, M., Roisenberg, M., and Rodrigues, B. B. (2018). Joint Bayesian inversion based on rock-physics prior modeling for the estimation of spatially correlated reservoir properties. Geophysics, 83(5), M49-M61.

Dejtrakulwong, P., Mukerji, T., and Mavko, G. (2012). Using kernel principal component analysis to interpret seismic signatures of thin shaly-sand reservoirs. In SEG Technical Program Expanded Abstracts 2012. Society of Exploration Geophysicists.

Fichtner, A., and Simutė, S. (2018). Hamiltonian Monte Carlo inversion of seismic sources in complex media. Journal of Geophysical Research: Solid Earth, 123(4), 2984-2999.

Fichtner, A., and Zunino, A. (2019). Hamiltonian nullspace shuttles. Geophysical research letters, 46(2), 644-651.

Fichtner, A., Zunino, A., and Gebraad, L. (2019). Hamiltonian Monte Carlo solution of tomographic inverse problems. Geophysical Journal International, 216(2), 1344-1363.

Gebraad, L., Boehm, C., and Fichtner, A. (2020). Bayesian elastic Full-Waveform Inversion using Hamiltonian Monte Carlo. Journal of Geophysical Research: Solid Earth, 125(3).

Grana, D., Passos de Figueiredo, L., and Azevedo, L. (2019). Uncertainty quantification in Bayesian inverse problems with model and data dimension reduction. Geophysics, 84(6), M15-M24.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. Computational Statistics, 14(3), 375-396.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. Bernoulli, 7(2), 223-242.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: efficient adaptive MCMC. Statistics and computing, 16(4), 339-354.

Holmes, C., Krzysztof, L. and Pompe, E. (2017). Adaptive MCMC for multimodal distributions. Technical report. https://pdfs.semanticscholar.org/c75d/f035c23e3c0425409e70d457cd43b174076f.pdf.

Horta, A., and Soares, A. (2010). Direct sequential co-simulation with joint probability distributions. Mathematical Geosciences, 42(3), 269-292.

Lieberman, C., Willcox, K., and Ghattas, O. (2010). Parameter and state model reduction for large-scale statistical inverse problems. SIAM Journal on Scientific Computing, 32(5), 2523-2542.

Lochbühler, T., Breen, S. J., Detwiler, R. L., Vrugt, J. A., and Linde, N. (2014). Probabilistic electrical resistivity tomography of a CO2 sequestration analog. Journal of Applied Geophysics, 107, 80-92.

MacKay, D.J., (2003). Information Theory, Inference and Learning Algorithms. Cambridge University Press.

Martin, J., Wilcox, L. C., Burstedde, C., and Ghattas, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM Journal on Scientific Computing, 34(3), A1460-A1487.

Menke, W. (2018). Geophysical data analysis: Discrete inverse theory. Academic press.

Neal, R.M., (2011). MCMC using Hamiltonian dynamics: in Handbook of Markov Chain Monte Carlo, Brooks, S., Gelman, A., Jones, G. and Meng, X.: Chapman and Hall, 113–162.

Pagani, F., Wiegand, M., and Nadarajah, S. (2019). An n-dimensional Rosenbrock Distribution for MCMC Testing. arXiv preprint arXiv:1903.09556.

Sambridge, M., and Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. Reviews of Geophysics, 40(3), 3-1.

Sambridge, M. (2014). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. Geophysical Journal International, 196(1), 357-374.

Sen, M. K., and Stoffa, P. L. (1996). Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion 1. Geophysical Prospecting, 44(2), 313-350.

Sen, M. K., and Stoffa, P. L. (2013). Global optimization methods in geophysical inversion. Cambridge University Press.

Sen, M. K., and Biswas, R. (2017). Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. Geophysics, 82(3), R119-R134.

Tarantola, A. (2005). Inverse problem theory and methods for model parameter estimation. siam.

Ter Braak, C. J., and Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. Statistics and Computing, 18(4), 435-446.

Tierney, L., and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. Statistics in medicine, 18(17-18), 2507-2515.

Turner, B. M., and Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. Journal of Mathematical Psychology, 56(5), 375-385.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software, 75, 273-316.

A gradient-based Markov chain Monte Carlo algorithm for elastic pre-stack inversion with data and model space reduction

Mattia Aleardi

University of Pisa, Earth Sciences Department, via S. Maria 53, 56126, Pisa, Italy

Corresponding author: Mattia Aleardi, mattia.aleardi@unipi.it

ABSTRACT

The main challenge of Markov Chain Monte Carlo sampling is to define a proposal distribution that simultaneously is a good approximation of the posterior probability while being inexpensive to manipulate. We present a gradient-based Markov Chain Monte Carlo inversion for elastic pre-stack inversion in which the posterior sampling is accelerated by defining a proposal that is a local, Gaussian approximation of the posterior model, while a non-parametric prior distribution is assumed for the distribution of the elastic properties. The proposal is constructed from the local Hessian and gradient information of the log posterior, whereas the non-linear, exact Zoeppritz equations constitute the forward modeling engine for the inversion procedure. Hessian and gradient information is made computationally tractable by a reduction of data and model spaces through a Discrete Cosine Transform reparameterization. This reparameterization acts as a regularization operator in the model space, while also preserving the spatial and temporal continuity of the elastic properties in the sampled models. We test the implemented algorithm on synthetic pre-stack inversions under different signal-to-noise ratios in the observed data. We also compare the results provided by the presented method when a computationally expensive (but accurate) finite-difference scheme is used for the Jacobian computation, with those obtained when the Jacobian is derived from a linearization of the exact Zoeppritz equations. The outcomes of the proposed approach are also compared against those yielded by a gradient-free Monte Carlo sampling method and by a deterministic least-squares inversion. Our tests demonstrate that the gradient-based sampling reaches accurate uncertainty estimations with a much lower computational effort than the gradient-free approach.

Keywords: AVA; Seismic inversion; Uncertainties.

INTRODUCTION

The great challenge in solving geophysical inverse problems lies in the fact that they are usually ill-posed: different combinations of model parameters are consistent with the observed data. No uniqueness in the recovered solution arises from noisy measurements, sparse observations, prior uncertainties, and approximation in the forward model that maps the model parameters into the observed data. The deterministic approach to geophysical inversion guarantees a rapid convergence toward a best-fitting model, but is incapable of accounting for the uncertainties affecting the recovered solution. On the contrary, Bayesian inference provides a systematic framework for incorporating and propagating the uncertainties in observed data, prior knowledge, and forward operator into the uncertainties affecting the recovered model (Tarantola, 2005). The final solution of a Bayesian inversion is the so-called posterior probability density (PPD) function in model space that fully quantifies the uncertainties in the recovered solution. However, an analytical uncertainty assesment is only possible for linear forward operators and Gaussian assumptions about model, data, and noise distributions. In all the other cases, Markov Chain Monte Carlo sampling methods can be used to numerically assess the posterior density (Sen and Stoffa, 1996; Sambridge and Moseegard, 2002; Sen and Stoffa, 2013). However, expensive forward model operators and high-dimensional parameter spaces make the application of MCMC algorithms computationally unfeasible. Indeed, it is known that the sampling ability of these methods dramatically decreases in large-dimensional problems due to the so-called curse of dimensionality issue (Curtis and Lomax 2001). In these contexts, traditional sampling methods might require billions of forward evaluations before converging to stable posterior uncertainties.

More in detail, MCMC algorithms generate samples by perturbing the current state of the chain (current model) according to a proposal distribution. Once generated, the Metropolis-Hasting criterion is used to either accept or reject the proposed sample. This process generates a chain of samples whose distribution asymptotically converges to the target PPD. Theoretically, for an infinite number of samples, the estimated distribution does not depend on the choice of the proposal. However, from a more practical perspective, the Monte Carlo sampling is maximally efficient when the proposal is a good approximation of the target density. For this reason, the definition of an appropriate proposal is of crucial importance for an efficient probabilistic sampling. The setting of an optimal proposal is especially of great importance in large-dimensional parameter spaces, where a significant mismatch between the proposal and the target density can drastically affect the performance of the sampling: persistent rejections of models, entrapment in local maxima of the PPD, and a dramatic increase in the number of forward evaluations needed to attain stable uncertainty estimations. When the classical random walk Metropolis algorithm is employed, a good compromise between the exploitation and exploration of the sampling is usually determined by a trial and error procedure in which different hyperparameters defining the proposal are tuned. However, more sophisticated MCMC recipes can be adopted (e.g. self-adaptive MCMC algorithms, preconditioned MCMC, hybrid MCMC approaches; Tierney and Mira 1999; Haario et al. 1999; Haario et al. 2001; Haario et al. 2006; ter Braak and Vrugt 2008; Turner and Sederberg 2012; Sambridge 2013; Vrugt 2016; Holmes et al. 2017). For example, self-adaptive algorithms, iteratively adjust the proposal to the local shape of the posterior. As an alternative, Gradient-Based MCMC (GB-MCMC) (e.g. Hamiltonian Monte Carlo, Langevin Monte Carlo; Sen and Biswaw; 2017; Fichtner and Simutè, 2018; Fichtner and Zunino, 2019; Fichtner et al. 2019; Gebrad et al. 2020; Aleardi and Salusti 2020; Aleardi, 2020a) exploit the gradient information of the misfit function (the negative natural logarithm of the posterior) to efficiently explore the model space and to rapidly converge toward stable posterior uncertainties (MacKay, 2003; Neal 2011). The main computational requirement of these methods is the need for computing derivatives, although this information is highly beneficial to speeding up the convergence of the sampling and to guarantee high independence of the samples while maintaining high acceptance rates.

It is also well known that MCMC algorithms work well in reduced spaces (Lieberman et al. 2010), and hence a popular approach to deal with high-dimensional problems is to use a reparameterization strategy that decreases the number of unknowns. In this case, the full state space is projected onto a limited number of basis functions and the algorithm generates samples in this reduced domain. This technique must be applied taking in mind that part of the information in the original (unreduced) parameter space could be lost in the reduced space and for this reason, the model parameterization must always constitute a compromise between model resolution and model uncertainty (Dejtrakulwong et al. 2012; Lochbühler et al. 2014; Aleardi 2019; Grana et al. 2019; Aleardi 2020b).

Here we propose a sampling strategy in which a gradient-based MCMC algorithm is combined with a compression of data and model space through a Discrete Cosine Transform (DCT). In particular, on the line of Martin et al. (2012), we exploit the geometrical properties of the misfit function to greatly speed up the probabilistic sampling. The approach is derived by analogy with the classical Newton approach to deterministic inversion and it defines a proposal density based on a local Gaussian approximation to the target PPD informed by local Hessian information. We apply this strategy to solve a Bayesian amplitude versus angle (AVA) inversion in which the subsurface elastic properties of P-wave velocity (*Vp*), S-wave velocity (*Vs*), and density are inferred from partially stacked seismic data at different incidence angles, while the exact Zoeppritz equations constitute the forward modeling operator. In our approach the dimensions of the Jacobian matrix are significantly reduced through a compression of both model and data spaces, thereby rendering the Hessian and Gradient manipulations computationally feasible. The DCT expands a signal (e.g. expressing the subsurface *Vp* model) into a series of cosine functions oscillating at different frequencies. The low-order discrete cosine transform coefficients express most of the variability of the original signal, and the model compression is simply accomplished by zeroing the numerical coefficients beyond a certain threshold. Therefore, the compression also helps to reduce the ill-conditioning of the inversion and mitigate the curse of dimensionality issue.

A crucial aspect of AVA inversion is the preservation of both the mutual and spatial/temporal relationships between the elastic parameters as inferred, for example, from available well log data (Aleardi et al. 2015). Usually, to avoid inverting large matrices, the AVA inversion is solved for each seismic gather independently. However, with this strategy, the spatial continuity of the elastic properties in the predicted model could be lost, especially in case of severe noise contamination. In this context, the advantage of the DCT lies in the possibility to apply this transformation to multidimensional signals as well (e.g. 2-D images). In this case, the order of the retained non-zero coefficients determines the wavelength of the recovered, compressed image along different (i.e. vertical, horizontal) directions. In our implementation, the compression is applied both to the elastic parameters (*Vp*, *Vs,* and density) and the seismic data that are treated as 2-D and 3-D images, respectively. This strategy allows for a simultaneous estimation of the elastic parameters over the entire considered area while guaranteeing the preservation of the temporal and spatial continuity of the elastic properties in all the sampled models.

After discussing the theoretical aspects of the proposed inversion scheme, we consider an analytical probability density function to illustrate the benefits of the implemented GB-MCMC algorithm over standard gradient-free MCMC approaches. Then, the method is applied to synthetic seismic data computed on a realistic subsurface elastic model that mimics a clastic geological setting in which a turbiditic sequence host gas saturated sand intervals. The outcomes of the implemented GB-MCMC approach are also validated and compared with those yielded by a gradient-free MCMC sampling (i.e. the Differential Evolution Markov Chain "DEMC"; Vrugt 2016) still running in the reduced data and model spaces and with those provided by a linearized least-squares algorithm that inverts each seismic gather separately working in the full model and data spaces. The proposed approach needs computing the Jacobian matrix associated with each sampled model. Therefore, we also compare the predictions provided by two GB-MCMC implementations: The former uses a computationally intensive, but accurate forward finite-difference scheme to compute the Jacobian matrix around each considered model. The latter replaces the Jacobian with a matrix operator derived from a linear approximation of the exact Zoeppritz equations after projection onto the compressed space (Aleardi and Salusti, 2020).

The main novelty of this paper is the combination of a gradient-based MCMC sampling and a DCT compression of both data and model space to efficiently solve the Bayesian non-linear pre-stack inversion.

METHODS

Gradient-based MCMC sampling

Gradient-based deterministic inversions are aimed at minimizing a previously defined misfit function, which usually is a linear combination of data error and a model regularization term. For Gaussian-distributed noise and model parameters, the error function can be written as follows (Menke 2018; Aster et al. 2018):

$$E(\mathbf{m}) = \left\|\mathbf{C}_d^{-\frac{1}{2}}(\mathbf{d} - G(\mathbf{m}))\right\|_2^2 + \left\|\mathbf{C}_m^{-\frac{1}{2}}(\mathbf{m} - \mathbf{m}_{prior})\right\|_2^2, \qquad (1)$$

where the vectors $\mathbf{m}$ and $\mathbf{d}$ identify the model parameters and the observed data, respectively; $\mathbf{C}_d^{-1/2}$ and $\mathbf{C}_m^{-1/2}$ are the data and prior model covariance matrices; $\mathbf{m}_{prior}$ is the prior model vector, and G is the forward modeling operator that maps the model into the corresponding data. The minimum of E $(\mathbf{m})$ can be iteratively approached through a local quadratic approximation of the error function around the current model $\mathbf{m}_k$:

$$E(\mathbf{m}) = E(\mathbf{m}_k + \Delta\mathbf{m}) \approx \tilde{E}(\mathbf{m}) == E(\mathbf{m}_k) + \Delta\mathbf{m}^{\boldsymbol{T}}\nabla_{\boldsymbol{m}}\mathbf{E}(\mathbf{m}_k) + \frac{1}{2}\Delta\mathbf{m}^{\boldsymbol{T}}\nabla_{\boldsymbol{m}}^{\mathbf{2}}\mathbf{E}(\mathbf{m}_k)\Delta\mathbf{m} + O(||\Delta\mathbf{m}||^3), \quad (2)$$

where $\Delta\mathbf{m} = \mathbf{m} - \mathbf{m}_{\boldsymbol{k}}$, whereas $\nabla_{\boldsymbol{m}}\mathbf{E}(\mathbf{m}_k)$ and $\nabla_{\boldsymbol{m}}^{\mathbf{2}}\mathbf{E}(\mathbf{m}_k)$ represent the first and second derivative of $E(\mathbf{m})$ computed around $\mathbf{m}_k$. In particular, it results that:

$$\nabla_{\boldsymbol{m}}\mathbf{E}(\mathbf{m}_k) = \mathbf{g} = \mathbf{J}^T\mathbf{C}_d^{-1}\Delta\mathbf{d}(\mathbf{m}_k) + \ \mathbf{C}_m^{-1}(\mathbf{m}_k - \mathbf{m}_{prior}), \quad (3)$$

and

$$\nabla^2_{\mathbf{m}}\mathbf{E}(\mathbf{m}_k) = \mathbf{H} = \left(\mathbf{J}^T\mathbf{C}_d^{-1}\mathbf{J}\right)^{-1} + \frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}\mathbf{C}_d^{-1}\left(\Delta\mathbf{d}(\mathbf{m}_k)...\Delta\mathbf{d}(\mathbf{m}_k)\right) + \mathbf{C}_m^{-1} = \mathbf{H_o} + \mathbf{B} + \mathbf{C}_m^{-1}, \quad (4)$$

where $\Delta\mathbf{d}(\mathbf{m}_k) = G(\mathbf{m}_k) - \mathbf{d}$, $\mathbf{B} = \frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}\mathbf{C}_d^{-1}\left(\Delta\mathbf{d}(\mathbf{m}_k)...\Delta\mathbf{d}(\mathbf{m}_k)\right)$, whereas $\mathbf{J}$ denotes the Jacobian matrix expressing the partial derivative of the data with respect to model parameters. In practical applications and for computational feasibility reason, the Hessian matrix is approximated as $\mathbf{H} \approx \mathbf{H}_{\boldsymbol{a}} = \mathbf{H_o} + \mathbf{C}_m^{-1}$, thus neglecting the partial derivative of the Jacobian with respect to the model. The number of rows and columns of the Hessian is equal to the number of data points and model parameters, respectively. The quadratic approximation of the error function can be compactly written as:

$$\tilde{E}(\mathbf{m}) = \frac{1}{2}\left(\mathbf{m} - \mathbf{m}_k + \mathbf{H}_a^{-1}\mathbf{g}\right)^{\boldsymbol{T}}\mathbf{H}_{\boldsymbol{a}}\left(\mathbf{m} - \ \mathbf{m}_k + \mathbf{H}_a^{-1}\mathbf{g}\right) + const., \quad (5)$$

Equation 5 shows that the minimizer of $\tilde{E}(\mathbf{m})$ can be computed as

$$\mathbf{m}_{k+\mathbf{1}} = \mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g}, \quad (6)$$

where $\mathbf{H}_a^{-1}\mathbf{g}$ is called the Newton step. In the context of deterministic inversions, an approximated uncertainty quantification can be computed from the inverse of the Hessian matrix at the convergence point. In other terms, a local quadratic approximation of the inverse of the curvature of the error function gives the uncertainties affecting the recovered solution.

Differently, a Bayesian inversion aims to estimate the full posterior distribution in the model space given by:

$$p(\mathbf{m}\,|\,\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \quad (7)$$

where $p(\mathbf{m}\,|\,\mathbf{d})$ is the posterior probability density (PPD), $p(\mathbf{d}|\mathbf{m})$ is the so-called likelihood function, whereas $p(\mathbf{m})$ and $p(\mathbf{d})$ are the a-priori distributions of model parameters and data, respectively. For problems in which the $p(\mathbf{m}\,|\,\mathbf{d})$ can not be expressed in a closed form, an MCMC algorithm can be used for a numerical assessment of the posterior model. In this context, the probability to move from the current state of the chain $\mathbf{m}_k$ to the next, proposed state $\mathbf{m}_{k+1}$ is determined according to the Metropolis-Hasting rule:

$$\alpha = p(\mathbf{m}_{k+1} \,|\, \mathbf{m}_k) = \min\left[1, \frac{p(\mathbf{m}_{k+1})}{p(\mathbf{m}_k)} \times \frac{p(\mathbf{d}|\mathbf{m}_{k+1})}{p(\mathbf{d}|\mathbf{m}_k)} \times \frac{q(\mathbf{m}_k|\mathbf{m}_{k+1})}{q(\mathbf{m}_{k+1}|\mathbf{m}_k)}\right], \quad (8)$$

where $q(.)$ is the proposal distribution that defines the new state (i.e. model) $\mathbf{m}_{k+1}$ as a random deviate from a probability distribution $q(\mathbf{m}_{k+1}|\mathbf{m}_k)$ conditioned only on the current state $\mathbf{m}_k$. The proposal ratio term vanishes if symmetric proposals are used. For example, the most popular proposal strategy uses a Gaussian step centered on the current state $\mathbf{m}_{k+1} = \mathbf{m}_k + \mathcal{N}(0,\mathbf{C})$, where $\mathbf{C}$ is the selected covariance matrix of the proposal and $\mathcal{N}$ denotes the Gaussian distribution. This method is referred to as the Random Walk Metropolis. If $\mathbf{m}_{k+1}$ is accepted, $\mathbf{m}_k = \mathbf{m}_{k+1}$. Otherwise, $\mathbf{m}_k$ is repeated in the chain and another state is generated as a random deviate from $\mathbf{m}_k$. The ensemble of sampled states after the burn-in period is used to numerically compute the statistical properties (e.g. mean, mode, standard deviations, marginal densities) of the target posterior probability. Now we can formulate the Bayesian inversion framework in terms of $E(\mathbf{m})$, $\mathbf{H}$ and $\mathbf{g}$, under Gaussian assumptions for data, noise, and model parameter distributions; we can write (Tarantola, 2005):

$$p(\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{prior})^{\boldsymbol{T}}\mathbf{C}_{\boldsymbol{m}}^{-\mathbf{1}}(\mathbf{m} - \mathbf{m}_{prior})\right), \quad (9)$$

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{d} - G(\mathbf{m}))^{\boldsymbol{T}}\mathbf{C}_{\boldsymbol{d}}^{-\mathbf{1}}(\mathbf{d} - G(\mathbf{m}))\right), \quad (10)$$

$$p(\mathbf{m} \,|\, \mathbf{d}) \propto \exp(-E(\mathbf{m})), \quad (11)$$

If we substitute equation 5 into equation 11 we obtain the approximation of the posterior around $\mathbf{m}_k$:

$$p(\mathbf{m} \,|\, \mathbf{d}) \approx \tilde{p}(\mathbf{m} \,|\, \mathbf{d}) \propto \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{m} - (\mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g}))^{\boldsymbol{T}}\mathbf{H}_{\boldsymbol{a}}(\mathbf{m} - (\mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g}))\right), \quad (12)$$

Equation 12 indicates that the approximation of the PPD is Gaussian distributed $\tilde{p}(\mathbf{m}|\mathbf{d}) = \mathcal{N}(\mathbf{m}_k - \mathbf{H}_a^{-1}\mathbf{g};\mathbf{H}_{\boldsymbol{a}})$ with mean equal to the minimizer of $\tilde{E}(\mathbf{m})$ and covariance equal to the inverse of the Hessian matrix. After constructing a local Gaussian approximation of the posterior density, we can now define a sampling method that uses the following proposal density:

$$q(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{m} - \left(\mathbf{m}_k - \lambda\mathbf{H}_a^{-1}\mathbf{g}\right)\right)^T \frac{\mathbf{H}_{\boldsymbol{a}}}{\mu^2}\left(\mathbf{m} - \left(\mathbf{m}_k - \lambda\mathbf{H}_a^{-1}\mathbf{g}\right)\right)\right). \quad (13)$$

Each proposed model is accepted according to the Metropolis Hasting rule taking in mind that in this case the proposal is not symmetric and for this reason, the proposal ratio should be evaluated. However, since the proposal is Gaussian both $q(\mathbf{m}_{k+\mathbf{1}}|\mathbf{m}_k)$ and $q(\mathbf{m}_k|\mathbf{m}_{k+\mathbf{1}})$ can be analytically computed. λ and μ^2 are tunable parameters that determine the step length along the negative gradient direction and the variance of the random perturbation around the minimizer of $\tilde{E}(\mathbf{m})$. These parameters must be properly set to get the desired acceptance rate or in other words to find a good compromise between exploitation and exploration of the sampling. More in detail, the λ value should be large enough to make the proposal dependent on the gradient information, but small enough so that the model update is not dominated by the deterministic information. On the contrary, the $\mu^{\mathbf{2}}$ value should be large enough to ensure an efficient exploration of the model space, but small enough so that the gradient information is not completely masked by the random update. We will consider the full Hessian and not only its diagonal entries so that possible posterior correlations between the inverted parameters are fully taken into account.

Therefore, we have "tailored" the proposal density $q(\mathbf{m})$ to the underlying local Gaussian approximation of the posterior probability using the derivative information of the error function. From a practical point of view, the proposed model can be straightforwardly generated according to:

$$\mathbf{m}_{k+\mathbf{1}} = \mathbf{m}_k - \lambda\mathbf{H}_a^{-1}\mathbf{g} + \mu\mathbf{H}_{\boldsymbol{a}}^{-\frac{1}{2}}\mathbf{n}, \quad (14)$$

with $\mathbf{H}_{\boldsymbol{a}}^{-\mathbf{1}} = \mathbf{H}_{\boldsymbol{a}}^{-\mathbf{1/2}}(\mathbf{H}_{\boldsymbol{a}}^{-\mathbf{1/2}})^T$, whereas $\mathbf{n}$ is a random column vector with the number of rows equal to the number of model parameters drawn from $\mathcal{N}(0,\mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix.

Note that for a Gaussian PPD and an exact Hessian, the proposed method results in a perfect sampling, in which all the samples are independently drawn from the posterior density with an acceptance probability equal to 1 (Martin et al. 2012). Also, note that for $\mu = 0$ equation 14 gives the standard gradient descent model update. On the contrary, if $\lambda = 0$ we have the standard random walk with some constraints given by the inverse Hessian. It can also be demonstrated that the previous GB-MCMC approach is related to the Hamiltonian Monte Carlo and Langevin Monte Carlo approaches (Martin et al. 2012). Finally, even though the proposal is derived by assuming a local Gaussian assumption, it can be used to sample from whatever type of posterior model and under whatever a-priori assumption (e.g. non-parametric), as it has been done in the following examples.

The major computational requirement of the implemented approach is the need for computing the Jacobian associated with each sampled model. When the forward is expressed by a partial differential equation the adjoint state method can be used to rapidly estimate the Gradient and Hessian with a reduced number of forward evaluations. The Jacobian can be also evaluated using a finite-difference scheme or in the case of weakly non-linear problems, a linearized approximation of the non-linear forward operator can be adopted as well. An extra computational workload also arises in large dimensional spaces due to the manipulation of large Hessian matrices and gradient vectors. In these contexts, a compression strategy would be useful to reduce the number of data points and model parameters and hence the dimensions of $\mathbf{H}_{\boldsymbol{a}}$ and $\mathbf{g}$. If a finite difference scheme is employed, the compression of the model parameter space also reduces the number of forward evaluations needed for the Jacobian computation.

Discrete Cosine Transform

Several variants of discrete cosine transform exist with slightly modified definitions, but in this work, we use the so-called DCT-II formulation that is the most common (Britanak et al. 2010). Hereafter we simply refer to the DCT-II as the DCT. We employ this parameterization because it exhibits superior compression power over other compression methods (Lochbühler et al. 2014).

This compression technique can be applied to multidimensional signals (i.e. 2-D matrices) and such multi-dimensional transform follows straightforwardly from the one-dimensional definition because it is simply a separable product (equivalently, a composition) of DCTs along each dimension. For example, if we assume a 2-D density model $\boldsymbol{\rho}(x,y)$ in which x=[0,1,…,M_x-1] and y=[0,1,…,M_y-1] represent the horizontal and vertical coordinates, respectively, the associated 2-D transform is defined as follows:

$$\begin{cases} \mathbf{R}(k_x,k_y) = \frac{1}{\sqrt{M_x}}\frac{1}{\sqrt{M_y}}\sum_{x=0}^{M_x+1}\sum_{y=0}^{M_y+1}\boldsymbol{\rho}(x,y),\ if k_x = k_y = 0 \\ \mathbf{R}(k_x,k_y) = \sqrt{\frac{2}{M_x}}\sqrt{\frac{2}{M_y}}\sum_{x=0}^{M_x+1}\sum_{y=0}^{M_y+1}\boldsymbol{\rho}(x,y)cos\left(\frac{(2x+1)\pi k_x}{2M_x}\right)cos\left(\frac{(2y+1)\pi k_y}{2M_y}\right), if k_x,k_y \neq 0 \end{cases},(15)$$

where $\mathbf{R}(k_x,k_y)$ represent the k_x-th and k_y-th coefficient. The values within the matrix $\mathbf{R}$ represent the unknowns to be estimated in a reparametrized inverse problem. Equation 15 can be compactly rearranged in matrix form:

$$\mathbf{R} = \mathbf{B}_y\boldsymbol{\rho}\mathbf{B}_x^T, \quad (16)$$

where $\mathbf{B}_x$ and $\mathbf{B}_y$ are the matrices with dimensions $M_x \times M_x$ and $M_y \times M_y$, respectively that contain the basis functions spanning the compressed space, whereas the $M_y \times M_x$ matrix $\mathbf{R}$ expresses the DCT coefficients. This approach concentrates most of the information of the original signal into the low-order coefficients, and hence an approximation of the subsurface density model can be obtained as follows:

$$\tilde{\boldsymbol{\rho}} = \left(\mathbf{B}_y^q\right)^T\mathbf{R}_{qp}\mathbf{B}_x^p, \ (17)$$

where $\tilde{\boldsymbol{\rho}}$ is the approximated $[M_y \times M_x]$ density model, $\mathbf{B}_y^q$ is a $[q \times M_y]$ matrix containing only the first q rows of $\mathbf{B}_y$; $\mathbf{B}_x^p$ is a $[p \times M_x]$ matrix containing only the first p rows of $\mathbf{B}_x$, whereas the matrix $\mathbf{R}_{qp}$ represents the first q rows and p columns of $\mathbf{R}$. In other words, the scalar q and p represent the retained number of base functions along the y and x directions used to derive the approximated model. Therefore, the DCT transformation allows for a reduction of the $(M_y \times M_x)$-D full density model space to a $(q \times p)$-D DCT-compressed parameter space with $p < M_x$ and $q < M_y$. In this context the $p \times q$ non-zero numerical coefficients of the $\mathbf{R}_{qp}$ matrix becomes the unknowns to be estimated after a compression of the model space. Estimating the retained coefficients reduces the parameter space dimensionality and can significantly improve the computational efficiency of the inversion procedure. Figure 1 shows some DCT base functions of different orders in a 2-D space. Note that the variability of the solution along each dimension is directly determined by the orders of the retained coefficients.

The implemented AVA inversion scheme

We consider a 2-D subsurface model in which the parameters to be estimated are the *Vp*, *Vs,* and density values. The observed data are partial angle stacks computed by separately applying the Zoeppritz equations to the elastic properties at each spatial location. For a $M_y \times M_x$ subsurface model and for *N* incidence angles (usually *N*=3; near, mid, and far stacks), we have $3 \times M_y \times M_x$ model parameters to be estimated from $N \times (M_y - 1) \times M_x$ data points. The spatial dimensions M_y, M_x are usually large, and hence the simultaneous estimation of the *Vp*, *Vs,* and density over the entire study area becomes computationally impractical for both deterministic and MCMC methods: In the former case, the large dimension of the Hessian and gradient matrices makes their manipulation and/or computation problematic. In the latter, the convergence of the probabilistic sampling is hampered by the curse of dimensionality issue. For this reason, common deterministic and probabilistic inversion approaches separately estimate the elastic properties at each seismic gather location. This means that M_x inversions are run, each one estimating $3 \times M_y$ parameters from $3 \times (M_y - 1)$ observations. Although this method makes the inversion computationally feasible it does not preserve the spatial continuity on the predicted elastic models. To overcome this issue, we compress both data and model space. In more details, the *Vp*, *Vs,* and density models are treated as separate $M_y \times M_x$ images to which the 2-D DCT is applied. Therefore, each $M_y \times M_x$ matrix expressing a given elastic property is approximated with a reduced number of coefficients contained within a $q \times p$ matrix $(p < M_x$ and

$q < M_y$). This reduces the full $(3 \times M_y \times M_x)$-D elastic space to a compressed $(3 \times p \times q)$-D space. The compression is also applied to decrease the dimensionality of the data space. In this case, we apply a 3-D DCT in which the first two coordinates represent the spatial and temporal directions, while the third axis identifies the incidence angles. The application of this transformation allows for a reduction of the original $(N \times (M_y - 1) \times M_x)$-D data space to a $(b \times v \times \mathrm{c})$-D space with $b < N$, $v < M_y - 1$ and $c < M_x$ (Figure 2). The map between the full data and model spaces is constituted by the Zoeppritz equations that are separately applied to the elastic properties at each spatial location and provide the seismic gathers associated with each sampled model.

In this context, the GB-MCMC algorithm samples the compressed $(3 \times p \times q)$-D model space and estimate the DCT coefficients expressing the elastic properties from the retained $b \times v \times \mathrm{c}$ basis in the data space. This means that the computation of the proposal ratio, likelihood ratio, and prior ratio for each sampled model (see equation 8) is performed in the compressed model and data domains. A schematic representation of this strategy is given in Figure 3. We note that multiple forward and inverse transformations are needed in each iteration. However, these transformations can be run with a negligible computational cost. The sampled models after the burn-in phase are projected onto the elastic space through equation 17 to numerically compute the statistical characteristics of the PPD in the *Vp*, *Vs,* and density domain.

We assume a non-parametric prior for the elastic parameters in order to properly model their facies-dependent behavior, while a stationary Gaussian variogram expresses their lateral and temporal variability. Similarly, we assume a Gaussian noise model. The non-parametric prior in the elastic domain impedes an analytical derivation of the prior in the compressed space and for this reason, the prior model in the compressed space is numerically computed by applying the kernel density estimation algorithm to prior elastic realizations projected onto the DCT space. Differently, the assumed Gaussian noise model allows for an analytical derivation of the data covariance matrix in the compressed data space.

The main limitation of any GB-MCMC approach arises from the need for computing the gradient of the posterior model, and hence this strategy is usually applied to problems in which such derivative information can be computed quickly (Neal, 2011). In our case of elastic pre-stack inversion, the Jacobian matrix can be derived, for example, by adopting an accurate but computationally quite expensive forward finite-difference scheme. In this case, $3 \times p \times q$ forward modeling runs are needed to compute the Jacobian associated with the current compressed model. The good news is that each column of the Jacobian can be independently computed and hence the finite difference computation can be easily distributed across different cores.

Another and much less demanding strategy replaces the Jacobian with an analytical operator derived from a linear approximation of the full Zoeppritz equations (for example the linear equation proposed by Aki and Richards, 1980) properly projected onto the compressed model and data spaces (Aleardi, 2020). Note that, in this case, we are inherently assuming that the curvature of the misfit function, and hence the variance of the proposal distribution is constant over the entire model space. This simplification could decrease the convergence speed of the algorithm, but dramatically reduces the computing time of the entire sampling with respect to the finite difference strategy (Aleardi and Salusti, 2020). However, it should be also noted that any linear approximation of the Zoeppritz equations, although widely employed in AVA studies, is theoretically valid in case of weak elastic contrasts at the reflecting interfaces and within a limited angle range (usually not beyond 30-35 degrees). For this reason, the suitability of this approach should be evaluated case-by-case. In the following, we solve the GB-MCMC inversion using both approaches. Their results are also validated against those provided by a gradient-free sampling approach (i.e. the Differential Evolution Markov Chain; DEMC, Vrugt, 2016) running in the reduced model and data spaces, and with the outcomes of a linearized least-squares inversion running in the full elastic and data spaces and inverting each seismic gather separately.

RESULTS

Analytical example

Before applying the GB-MCMC algorithm, we briefly illustrate the benefits provided by the gradient-based sampling over a more standard, gradient-free sampling method. In this section, we aim to draw samples from a posterior model derived from the 2-D Rosenbrock function. This function has challenging features: its minimum is located at the bottom of a narrow parabolic valley where a small change in direction can lead to a steep increase of the gradient. The Rosenbrock can be turned into a probability density that maintains the same basic characteristics of the original function, and hence it has been frequently adopted to test sampling methods (Christen and Fox, 2010; Pagani et al. 2019). In this example the posterior model can be expressed as follows:

$$p(x,y) \propto \exp\left(-\left(100\left(y-x^2\right)^2+(1-x)^2\right)\right). \quad (18)$$

We compare the GB-MCMC approach with a random walk Metropolis (RWM). Both algorithms have been run for 80000 iterations employing 10 parallel chains and under uninformative prior for the *x* and *y* variables. The standard deviation of the proposal distribution for the random walk Metropolis has been properly set in order to get an acceptance rate lying in the interval [0.2, 0.4].

Figure 4 illustrates that both MCMC approaches provide similar posterior estimations, in good agreement with the target density. To assess the convergence of the sampling we use the potential scale reduction factor (PSRF), a popular convergence diagnostic tool proposed by Brooks and Gelman (1998) to which we refer the reader for its formal definition. This tool compares within-chain variances to the variance computed from all mixed chains for a given parameter. In practice, one can consider that the convergence to a stable posterior model has been achieved if the potential scale reduction factor is lower than 1.1. By the inspection of the PSRF evolution for the two unknown parameters (Figures 5a-c), we observe that 50000 iterations are needed by the random walk Metropolis to converge toward a stable posterior, while the GB-MCMC converges after only 2000 iterations. This significant difference is related to the different sampling strategies employed by the two approaches. Indeed, Figures 5d-e illustrates that the GB-MCMC not only focus the sampling around the most promising zones of the parameter space (i.e. those characterized by high posterior density values) but also uses a proposal that incorporates information about the local covariance structure of the target density as provided by the inverse of the Hessian matrix. Differently, the random walk proposal is not influenced by the local, geometrical properties of the target posterior and thus the proposed model could also be located far away from the posterior maximum. This difference also indicates that the acceptance rate for the GB-MCMC is usually much higher than that of the random walk Metropolis: in this example, the GB-MCMC acceptance oscillates around 60-80%, whereas only the 30%, on average, of the proposed states, were accepted by the random walk Metropolis. This is another strength of the gradient-based sampling methods because avoid wasting computing time to run forward evaluations for proposed models with a low probability to be accepted.

Synthetic inversion tests

For the lack of available real seismic data, we discuss synthetic experiments in which we applied the implemented approach to invert seismic data generated on a reference model that simulates a realistic geological context in which a turbiditic sequence hosts a gas-saturated reservoir (see Figure 6a). This subsurface model has been derived by integrating the borehole information provided by several wells with an accurate geologic interpretation. The true model represents an in-line section with 61 time samples and 91 cross-lines. The time sampling is 0.004 s, whereas 50 m is the cross-line distance. Figure 6a shows that significant elastic contrasts occur at the interface separating the encasing shales from the reservoir sands.

A forward modeling based on the full Zoeppritz equations computes the observed seismic gathers by convolving the angle-dependent reflectivity series with a 30-Hz, zero-phase Ricker wavelet. We consider three partial angle stacks corresponding to incidence angles of 0 (near stack), 20 (mid stack), and 40 (far stack) degrees. This means that the full model space comprises $61 \times 91 \times 3 = 16653$ parameters to be estimated from $60 \times 91 \times 3 = 16380$ data points.

Two columns extracted from the reference models at the horizontal coordinates of 1000 and 3000 m have been considered as available well log data, used to derive the prior information. We assume a non-parametric distribution for the elastic parameters, which has been computed by applying the kernel density estimation algorithm (Parzen 1962) to the available well log information (Figure 6b). We also assume a stationary 2D Gaussian variogram model in which the vertical and lateral ranges have been inferred from the vertical variability of the available well log data and the lateral variability of the observed seismic data, respectively. The ranges of the variogram are equal to 0.008 s and 160 m along the temporal (vertical) and spatial (lateral) directions, respectively.

The previously defined elastic prior model must be projected onto the DCT space where the MCMC sampling runs. To this end and given the non-parametric prior, we adopt a Monte Carlo simulation approach. The direct-sequential co-simulation method with joint probability distribution (Horta and Soares, 2010) has been used to draw 5000 2-D elastic models in accordance with the prior assumptions. Such models have been projected onto the compressed space, and the kernel density algorithm has been again employed to numerically compute the non-parametric prior in the reduced domain for the *Vp*, *Vs,* and density (Figure 6c). Two examples of *Vp*, *Vs,* and density prior realizations are represented in Figure 7.

The next step is to define the optimal number of coefficients needed to approximate the elastic profiles. To this end, we quantified how the explained variability of the elastic properties changes as the number of basis functions increases. The selection of the optimal number of coefficients is a very delicate step that must guarantee uncertainty estimations, model resolution, and data fitting comparable to those achieved by an inversion running in the full, uncompressed space. A detailed discussion on how the model and data compressions affect the AVA inversion results is far beyond the scope of this work and for this reason, we refer the interested reader to Grana et al. (2019) for more theoretical insights. Figure 8 shows the explained variability for a *Vp*, *Vs,* and density model drawn from the prior as the number of retained coefficients increases. We note that only 25 coefficients per elastic property (*q*=*p*=25) along the two DCT spatial dimensions explain almost the total variability of the three elastic parameters. This means that the compression allows for a reduction of the 16653-D model space to 25 × 25 × 3 = 1875-D domain. A similar analysis has been carried out on some seismic gathers derived from prior elastic realizations. An example is shown in Figure 9a where the green rectangle encloses the $40 \times 45 \times 2 = 3600$ retained coefficients in the data space that explain almost the total variability of the uncompressed seismic gather (Figure 9b). Therefore, in this case, the full 16380-D data space has been reduced to a 3600-D domain. These data and model parameter reductions not only guarantees a considerable speed-up in the finite-difference Jacobian computation but also drastically reduces the computational cost of the Hessian and gradient manipulation. For example, the 16653×16653 Hessian in the full domain has been reduced to a 3600×3600 matrix in the compressed space.

In the following inversion tests, we consider two different scenarios: in the former (Test 1) the data computed on the reference model have been contaminated with uncorrelated Gaussian random noise with a standard deviation of 0.03 that corresponds to the 20% of the total standard deviation of the noise-free dataset. However, the popular assumption of uncorrelated noise usually constitutes an oversimplification because in real data applications correlated noise can be ascribed, for example, to residual of multiple reflections or diffractions not successfully removed during the processing phase. For this reason, in the second example (named Test 2 in the following), the observed data have been contaminated with both incoherent and coherent Gaussian noise with the same standard deviation value of 0.06. The temporal and lateral correlation pattern of the coherent noise is the same as the elastic prior model.

In what follows, we discuss the results provided by the GB-MCMC approaches when the Jacobian is computed with a forward finite-difference scheme (GB-MCMC-FD), and when the Jacobian is replaced by the linear operator derived from the Aki and Richards equation (GB-MCMC-L). The outcomes of these inversions are also benchmarked against the predictions of a gradient-free DEMC inversion still running in the reduced model and data spaces and with the results provided by a deterministic linearized least-squares inversion running in the full data and model spaces and inverting each seismic gather independently. All the considered MCMC inversions take advantage of parallel implementations. In the DEMC and GB-MCMC-L each chain is run in parallel, while the

GB-MCMC-FD runs the chains serially but distributes the Jacobian computation across different cores.

We start with the results of Test 1 in which the noise model and the source wavelet are assumed perfectly known during the inversion phase. Figure 10 and Figure 11 show the posterior mean models and posterior standard deviations provided by the GB-MCMC-FD and GB-MCMC-L approaches, respectively. The GB-MCMC-FD and GB-MCMC-L have been run for 10000 iterations and employing 10 independent chains. Both algorithms yield similar and congruent estimates of the posterior mean and the associated uncertainties. We note that the posterior standard deviation increases as the velocities and density values increase. This indicates that the curvature of the error function is expected to change over the model space. Figure 12 compares the elastic properties extracted at two different spatial coordinates (1200 and 2500 m, respectively) with the posterior mean and the 95 % confidence interval estimated by the two GB-MCMC algorithms. We observe that the mean model closely reproduces the vertical variations of the true model, and more importantly, the posterior mean usually lies within the range depicted by the 95% confidence interval, thus ensuring us about the reliability of the final predictions. As an example, Figure 13 shows a comparison between the observed data and the data predicted on the mean model provided by the GB-MCMC-L inversion. The close similarity between the two seismic datasets prove that the predicted mean model can accurately reproduce the observed seismic amplitudes. A similar conclusion would have been drawn for the GB-MCMC-FD algorithm. Figures 10-13 demonstrate that both algorithms provide similar and congruent model and uncertainty estimations, thereby confirming that in both cases a stable posterior model has been reached within the selected number of iterations. For both the GB-MCMC-FD and GB-MCMC-L inversion we set the λ and μ^2 values to 0.2 and 0.95, respectively. This combination resulted in an acceptance rate oscillating around 0.7-0.85.

However, If we analyze the evolution of the negative log-likelihood we observe that the two GB-MCMC implementations are characterized by different convergence speeds toward the stationary regime (Figure 14). The GB-MCMC-FD converges to the steady-state in less than 5 iterations, while

50 iterations are needed by the GB-MCMC-L, although in both cases the same final likelihood value has been reached. This fact is related to the different strategies used to define the Jacobian matrix. In other words, a more accurate Jacobian reflects into a more accurate estimate of the local curvature of the error function thus guaranteeing a faster convergence toward the stationary regime. At a closer inspection, we also observe another difference between the two approaches (see the two close-ups on the right of Figure 14). The GB-MCMC-FD shows strongly variable misfit values with iterations, while the GB-MCMC-L misfit oscillates with a longer period. This proves that the use of an accurate Jacobian guarantees the sampling of maximally decoupled models, while for a linear approximation the successively sampled models are mutually correlated. Therefore, the sampling is expected to attain accurate uncertainty estimations with a lower number of iterations when the finite-difference strategy is adopted (MacKay, 2003). Indeed, Figure 15 shows the evolution of the potential scale reduction factor for all the model parameters in the compressed space and for the two algorithms. In both cases, as expected, the sampling converges faster for the *Vp* coefficients since this is the elastic parameter better constrained by the data, while a longer sampling is needed to attain stable PPDs for the *Vs* and density coefficients (i.e., *Vs* and density are less informed by the seismic data). From the evolution of the PSRF values, we can claim that the GB-MCMC-FD attains convergence for all the parameters with 1000 iterations, while 4000 iterations are needed by the GB-MCMC-L. However, the computational costs of a single GB-MCMC-FD and GB-MCMC-L iteration are very different: 30 s for the former and just 2.5 s, for the latter. This means that, despite the less accurate approximation of the Hessian matrix, the GB-MCMC-L attains convergence in less than 3 hours, while more than 8 hours are needed by the GB-MCMC-FD (see Table 1).

Figure 16 compares the assumed Gaussian correlograms and the average vertical and spatial correlograms computed on the true model and on the posterior solution provided by the GB-MCMC-L inversion. We observe that the assumed correlogram is well reproduced by the estimated model, which also shows a good agreement with the actual lateral and temporal variability patterns. The match between the marginal distributions derived on the true model with those computed on the posterior mean GB-MCMC-L solution also demonstrates that the implemented method guarantees a good reproduction of the actual distribution of the elastic parameters in the investigated area (Figure 17).

We now present the results of the DEMC and the linearized inversion for Test 1. For the DEMC we employ 10 parallel chains evolving for 100000 iterations, with a burn-in period of 70000 samples. In Figure 18a we observe that the linearized approach estimates elastic profiles affected by lateral scattering related to noise propagation from data to model space. This method converges in a few seconds to the final solution but it hampers accurate uncertainty assessments. Figure 18b shows that the DEMC algorithm has not reached accurate model estimations and stable uncertainty appraisals within the selected number of iterations. In particular, we note scattered standard deviation maps completely different from those provided by the two GB-MCMC algorithms. Indeed, the evolution of the negative log-likelihood values (Figure 19) proves that the gradient-free sampling has not even reached the stationary regime within the selected number of iterations. We point out that the acceptance rate of the DEMC oscillated around the optimal values of 0.22-0.33. For this reason, the slow convergence toward the steady-state is not related to an erroneous hyperparameters setting but to the difficulty in sampling the high-dimensional parameter space starting from random prior realizations. In other terms, due to the curse of dimensionality issue, a much higher number of iterations is now needed for accurate uncertainty estimations. To reduce the burn-in stage, the starting model can be generated from the results of a previous inversion step (for example a fast analytical inversion; de Figueiredo et al. 2018). The total computing time for running 100000 DEMC iterations was 11.1 hours, while a single iteration of this approach takes on average 0.4 s (Table 1). However, since the gradient-free sampling does not even reach the stationary regime within the selected number of iterations, we envisage that a much higher computing time is needed to achieve stable posterior assessments. The results of Test 1 indicate that, although the extra time needed for vector/matrix manipulation and Jacobian computation, both gradient-based MCMC algorithms outperform the gradient-free method because they achieve accurate model estimations and uncertainty appraisals with a much lower computational effort.

In the second test, we want to assess the applicability of the proposed approach to a more realistic scenario with a low signal-to-noise ratio and both coherent, and uncorrelated noise affecting the data. For the sake of conciseness, we will only present the results provided by the GB-MCMC-L and linearized approaches. Indeed, on the one hand, the two GB-MCMC strategies still provided very similar predictions, with the GB-MCMC-FD still needing a lower number of iterations to converge, but a higher computing time with respect to the GB-MCMC-L. On the other hand, the DEMC was again severely affected by the curse of dimensionality issue: thus, it would have needed a much higher computing time than the two GB-MCMC implementations to attain stable posterior estimations. In this example, only the uncorrelated Gaussian random noise is taken into account by the data covariance matrix, while the source wavelet is again assumed to be known. The hyperparameter setting for the GB-MCMC-L inversion is the same previously used in Test 1.

Figure 20 compares the results of the deterministic and GB-MCMC-L algorithms. We observe that the inclusion of coherent noise and the overestimation of the signal-to-noise ratio of the data has severely decreased the overall quality of the predictions. The linearized inversion provides final estimates severely affected by lateral scattering. In this case, the lateral formation boundaries of the main gas-saturated reservoir can not be mapped with reasonable accuracy. Differently, in the GB-MCMC-L predictions, we can still appreciate the significant decrease of *Vp*, *Vs,* and density occurring at the interface separating the reservoir sand and the encasing shale. As expected, the posterior uncertainty is increased with respect to the previous example (compare Figures 20c and 11b), such as the sample-by-sample difference between the observed and predicted seismic amplitudes (Figure 21). The direct comparison of the outcomes of the GB-MCMC-L and deterministic approach better highlights the superior predictions achieved by the proposed approach (Figure 22): the mean model estimated by the GB-MCMC is usually closer to the true model than the deterministic results. However, differently from the previous example we now note that the erroneous assumption in the statistical properties of the noise results in estimated confidence intervals that sometimes do not include the true model. Finally, the inspection of the evolution of the potential scale reduction factor (Figure 23) shows that similarly to Test 1, the GB-MCMC-L reaches accurate uncertainty appraisals for all the model parameters in 4000 iterations, approximately.

DISCUSSION

The aim of this work was twofold: implementing an sampling algorithm for accurate and fast uncertainty assessments in non-linear AVA inversion and mitigating the curse-of-dimensionality issues, thus allowing for a simultaneous estimation of the elastic properties along the entire considered 2-D section. To this end, we combined a GB-MCMC sampling with a DCT reparameterization of both data and model spaces.

We compared two different implementations of the GB-MCMC method: The first uses a finite-difference scheme to compute the Jacobian (named GB-MCMC-FD), while the second replaces the Jacobian with a matrix operator derived from a linearization of the Zoepprtiz equation (named GB-MCMC-L). Theoretically, the validity of the linear approximation of the Zoeppritz equations depends on the considered angle range and the magnitude of the elastic contrasts at the reflecting interfaces. However, in our tests, this strategy provided satisfactory model predictions and uncertainty quantifications comparable to those yielded by the GB-MCMC-FD algorithm, although the reference model was characterized by significant elastic contrasts at the interface separating the encasing shale from the reservoir sand. Besides, the GB-MCMC-L approach made it also possible for a significant reduction of the computational cost of a single GB-MCMC inversion. This reduced computational effort occurs at the expense of a slower convergence toward the stationary regime and to an overall decrease in the independence of successively sampled models. The choice of replacing the Jacobian with the linear matrix operator must be considered case-by-case and should constitute a compromise between the sampling efficiency and the computational cost of the GB-MCMC inversion. Another possibility is to employ the finite-difference strategy only for the first iterations when the stationary regime is not yet attained and hence maintaining the same Jacobian during the sampling stage. This recipe should guarantee a faster convergence toward the steady-state and a more efficient sampling, with a limited extra computational cost.

The computing times shown in Table 1 refer to Matlab codes running on a single server equipped with two deca-core intel E5-2630 at 2.2 GHz (128 Gb RAM). So there is still room for a substantial decrease of the computational costs of the GB-MCMC inversion, for example by running the codes on a large computer cluster or utilizing fast computing units and/or adopting a more efficient implementation (e.g., codes written in a lower-level programming language). The computational cost of the GB-MCMC inversion related to the computation of the inverse Hessian can be also reduced by dropping the off-diagonal entries of $\mathbf{H}_{\boldsymbol{a}}$. This strategy results in a proposal distribution that neglects the possible correlation between model parameters, which might have a negative impact on the convergence rate of the sampling.

A proper setting of the hyperparameters λ and μ is important for the efficiency of the sampling. Indeed, a poorly chosen parameter combination would result in a slow convergence toward stable uncertainty estimations. A good parameter combinations would guarantee a good compromise between exploitation and exploration, rendering reasonable acceptance rates. The λ parameter acts as the step length in gradient descent methods. Its value should be similar to the one used in gradient-based local optimization methods so that the linearized Taylor expansion is still locally valid. The μ parameter determines the variance of the proposal distribution: A too small μ would results in poor mixing, while a too-large μ would decrease the acceptance rate. In our experiments, we found the optimal combination using a trial-and-error procedure in which our goal was to get an acceptance rate around 0.7-0.8 during the sampling stage. Another viable strategy could be employing a self-adaptive scheme (Haario et al., 2001; Atchadé, 2006) that automatically adjusts the proposal variance during the sampling process. However, the many inversion tests we carried out showed that the optimal acceptance rate can be achieved by many different parameter combinations and hence a proper selection of λ and μ is not that hard to find; for example in the previous tests a good compromise between exploitation and exploration is guaranteed for λ and μ^2 values lying in the range [0.1, 0.7] and [0.5, 1.5], respectively. From our experiments also emerged that if the approximated Hessian is used, a good λ values should lie in the range]0, 1] because a higher λ puts more emphasis on the exploitation while penalizing the exploration. On the other hand, an optimal μ^2 value is usually around 1. Appendix A uses a didactic example to analyze the effect of the λ and μ^2 values on the sampling efficiency.

The implemented method can be also extended to 3-D models and in this case, a 4-D transformation must be used to compress the data space. Some experiments on a 3D elastic model with 61 time samples, 91 cross-line and 91 in-line have been carried out employing the same Matlab implementation previously considered. In these preliminary tests, the DCT allowed for a compression of the full 1515423-D elastic space into a $25 \times 25 \times 25 \times 3 = 46875\text{-D}$ domain. However, the current Matlab implementation and the limited available hardware resources make the computation of the Jacobian, the derivation of the inverse Hessian, and also the manipulation of both the Hessian and Gradient, prohibitive. In this context, the GB-MCMC-FD approach is unfeasible, while the GB-MCMC-L works but requires more than a week of computing time to converge. For this reason, a more scalable inversion code and additional hardware resources are needed to invert 3D data. Regarding the performance scaling of the adopted GB-MCMC recipe, Martins et al. (2012) observed similar convergence rates for different model space dimensionalities. They claimed that although this desirable characteristic is not yet proved theoretically, the numerical observations seem to indicate an insensitivity of convergence of the proposed GB-MCMC method to the parameter dimension. We refer the interested reader to Martins et al. (2012) for a more in-depth discussion of this aspect.

CONCLUSIONS

We presented a gradient-based MCMC method for casting the non-linear elastic pre-stack inversion into a solid Bayesian framework that also guarantees fast convergence toward stable PPD assessments. The key idea is to guide the parameter sampling by exploiting the gradient and Hessian information of the PPD, thereby generating proposal densities that are locally a good approximation of the target posterior. This results in a proposal distribution that is easy to construct, and in an increased efficiency of the probabilistic sampling: the gradient guides the sampling toward “better” solutions, whereas the random perturbation term avoids entrapments in local maxima of the PPD. The good compromise between the gradient and random perturbation (that is the optimal compromise between exploitation and exploration) can be found by adjusting two hyperparameters (λ and μ). We reduced the computational effort related to Hessian and gradient manipulation and Jacobian computation by employing a discrete cosine transform reparameterization of data and model spaces.

Our synthetic inversion experiments showed very promising results, in which the posterior mean model well reproduced the ground truth even when coherent noise contaminates the seismic gathers, and for erroneous assumptions about the noise properties. Our results indicated that the exploitation of the Hessian and gradient information always guarantees a much faster convergence toward stable uncertainty estimations than a gradient-free MCMC algorithm. The use of the finite-difference scheme reduced the number of iterations needed to achieve stable PPDs, but it required a significant extra computational cost per iteration for the Jacobian computation. Deriving the Jacobian from a linear approximation of the Zoeppritz equations decreased the sampling efficiency (e.g. slower convergence toward the stationary regime and increase of the correlation value between successively sampled models) but it also greatly reduced the computing time to attain convergence. However, the applicability of this strategy should be evaluated case-by-case because the validity of the linearization of the Zoeppritz equations depends on the considered angle range and the elastic contrasts at the reflective interfaces. The computational cost of the GB-MCMC inversion is orders of magnitude higher than that of the deterministic approaches. However, the main advantage of any MCMC algorithm over deterministic inversions is the possibility to evaluate the posterior uncertainties.

Conflict of interest

The author declares no conflict of interest.

Data Availability

Data available on request from the authors.

APPENDIX A

To investigate in more detail the effects of the λ and μ^2 values on the sampling efficiency of the GB-MCMC inversion we consider a simple example with a 2-D multivariate target density. We run two different tests: in the first, we set $\lambda = 0.05$ and $\mu^2 = 3$, whereas in the second $\lambda = 0.5$ and $\mu^2 = 0.5$. Both tests use 5 independent chains running for just 1000 iterations. Figure 24 demonstrates that in both cases we get a reasonable prediction of the target density, despite the limited maximum number of iterations considered. However, the inspection of the PSRF highlights that in the first test more than 500 iterations are needed to reach the threshold of convergence, while in the second case the convergence is attained in less than 150 iterations. This proves that in the first case we select a too low λ value and a too-high μ^2 thus meaning that we are promoting the exploration at the expense of the exploitation. Instead, in the second case, the hyperparameter setting guarantees an efficient sampling of the parameter space that results in a rapid convergence toward a stable PPD.

FIGURE LEGENDS

Figure 1: 2-D DCT base functions of different orders. Dark and light colors code low and high numerical values, respectively.

Figure 2: Derivation of data and model space vectors in the DCT space from the elastic properties and seismic gathers.

Figure 3: Schematic representation of the GB-MCMC inversion scheme. Green and pink rectangles refer to steps performed in the reduced and full spaces, respectively.

Figure 4: a) True posterior density function. b) Posterior density provided by the random walk Metropolis. c) Posterior density estimated by the GB-MCMC. The colormap codes the normalized probability values.

Figure 5: a) Evolution of the potential scale reduction factor for the random walk Metropolis. b) Evolution of the potential scale reduction factor for the GB-MCMC. c) Close-up of b). In a)-c) the horizontal dotted green lines represent the threshold of convergence, whereas the blue and red lines refer to the *x* and *y* variables, respectively. d) Example of current, proposed model, and proposal distribution for the random walk Metropolis. e) Example of current, proposed model, and proposal distribution for the GB-MCMC. In d) and e) the magenta curves represents the contour lines of the proposal while the colored curves are the contour lines of the Rosenbrock error function.

Figure 6: a) The elastic properties of *Vp*, *Vs,* and density of the reference model. In a) the black arrows point toward the main sand reservoir body, whereas the dotted red lines depict the columns of the model considered as available well log data for defining the a-priori elastic distribution. b) The marginal non-parametric prior distributions for the three elastic properties derived from the two wells shown in a). c) The marginal prior projected onto the compressed space through a Monte Carlo simulation.

Figure 7: a), b) Two examples of *Vp*, *Vs,* and density model drawn from the non-parametric elastic prior.

Figure 8: Examples of explained model variability for an elastic model extracted from the prior and as the number of coefficients along the 1st and 2nd DCT dimension increases. In each plot, the numerical value with coordinate (*x*, *y*) indicates the explained variability if the first *x*, and *y* coefficients along the 1st and 2nd dimensions, respectively, are used for compressing the model. It emerges that 25 coefficients along both the 1st dimension explain almost the 100 % of the variability of the uncompressed *Vp*, *Vs,* and density profiles.

Figure 9: a) DCT decomposition of a seismic gather computed on an elastic model drawn from the prior. Blue and red colors code low and high values, respectively while the green rectangles enclose the retained coefficients in the data space. b) Explained data variability as the number of considered basis functions increases.

Figure 10: Results provided by the GB-MCMC-FD approach for Tests 1. a) A-posteriori mean model. b) Posterior standard deviation. In a), and b) the *Vp*, *Vs,* and density are represented from left to right.

Figure 11: As in Figure 10 but for the GB-MCMC-L approach.

Figure 12: Comparison between the true model, the posterior mean, and 95% confidence interval at two different spatial locations. a) GB-MCMC-FD. b) GB-MCMC-L. The leftmost plot refers to the spatial position of 1200 m, while the plot on the right refers to the spatial position of 2500 m.

Figure 13: Comparison between observed data (left column), predicted data (central column), and their sample-by-sample difference (right column) for Test 1. The predicted data have been computed on the mean posterior model estimated by the GB-MCMC-L algorithm. a), b), and c) refer to near, mid and far stack, respectively.

Figure 14: Evolution of the negative log-likelihood values for the GB-MCMC-FD and GB-MCMC-L inversions (part a) and b), respectively). Each color represents a different chain.

Figure 15: Evolution of the potential scale reduction factor over iterations for the DCT coefficients associated with the three elastic properties. a) GB-MCMC-FD. b) GB-MCMC-L. The red dotted lines depict the threshold of convergence.

Figure 16: Comparison between the lateral (a) and vertical (b) assumed correlogram functions with the average correlograms computed on the true model (blue line) and on the posterior mean estimated by the GB-MCMC-L algorithm (red lines). From left to right we represent *Vp*, *Vs,* and density.

Figure 17: Marginal probabilities for the three elastic parameters computed on the true model and on the posterior mean estimated by the GB-MCMC-L algorithm.

Figure 18: a) Results of the linearized least-squares inversion. b) Estimated mean model by the DEMC algorithm. c) Posterior standard deviation estimated by the DEMC algorithm.

Figure 19: Evolution of the negative log-likelihood value during the DEMC sampling. Each color refers to a different chain.

Figure 20: Results for Test 2: a) *Vp*, *Vs*, and density profiled estimated by the linearized least-squares approach. b) Posterior mean model provided by the GB-MCMC-L approach. c) Posterior standard deviation estimated by the GB-MCMC-L inversion.

Figure 21: Comparison between observed data (left column), predicted data (central column), and their sample-by-sample difference (right column) for Test 2. The predicted data have been computed on the mean posterior model estimated by the GB-MCMC-L algorithm. a), b) and c) refer to near, mid and far stack, respectively.

Figure 22: Comparison between the true model, the deterministic inversion results, the posterior mean, and 95% confidence interval estimated by the GB-MCMC- L approach. a) refers to the spatial position of 1200 m, while b) refers to the spatial position of 2500 m.

Figure 23: Evolution of the potential scale reduction factor over iterations and for the coefficients associated with the three elastic properties. The dotted red lines depict the threshold of convergence.

Figure 24: GB-MCMC sampling of a 2D multivariate density for different hyperparameter settings. a) $\lambda = 0.05$ and $\mu^2 = 3$. b) $\lambda = 0.5$ and $\mu^2 = 0.5$. From left to right we represent the target probability density, the estimated probability density, and the evolution of the PSRF for the two parameters. Blue and yellow colors code low and high probability values, respectively. On the rightmost plot, the horizontal dotted green line represents the threshold of convergence, whereas the blue and red lines refer to the *x* and *y* variable, respectively.

REFERENCES

Aki, K., and Richards, P. G. (1980). Quantative seismology: Theory and methods. New York, 801.

Aleardi, M., and Salusti, A. (2020). Hamiltonian Monte Carlo algorithms for target-and interval-oriented amplitude versus angle inversions. Geophysics, 85(3), R177-R194.

Aleardi, M. (2020a). Combining discrete cosine transform and convolutional neural networks to speed up the Hamiltonian Monte Carlo inversion of pre-stack seismic data. Geophysical Prospecting, 68(9), 2738-2761.

Aleardi, M. (2020b). Discrete cosine transform for parameter space reduction in linear and non-linear AVA inversions. Journal of Applied Geophysics, 104106.

Aleardi, M. (2019). Using orthogonal Legendre polynomials to parameterize global geophysical optimizations: Applications to seismic-petrophysical inversion and 1D elastic full-waveform inversion. Geophysical Prospecting, 67(2), 331-348.

Aleardi, M., Mazzotti, A., Tognarelli, A., Ciuffi, S., and Casini, M. (2015). Seismic and well log characterization of fractures for geothermal exploration in hard rocks. Geophysical Journal International, 203(1), 270-283.

Aster, R. C., Borchers, B., and Thurber, C. H. (2018). Parameter estimation and inverse problems. Elsevier.

Atchadé Y. F., (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift: Methodology and Computing in applied Probability, 8, 2, 235–254.

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 7(4), 434-455.

Britanak, V., Yip, P. C., and Rao, K. R. (2010). Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations. Elsevier.

Christen, J. A., and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t-walk). Bayesian Analysis, 5(2), 263-281.

Curtis, A., and Lomax, A. (2001). Prior information, sampling distributions, and the curse of dimensionality. Geophysics, 66(2), 372-378.

de Figueiredo, L. P., Grana, D., Bordignon, F. L., Santos, M., Roisenberg, M., and Rodrigues, B. B. (2018). Joint Bayesian inversion based on rock-physics prior modeling for the estimation of spatially correlated reservoir properties. Geophysics, 83(5), M49-M61.

Dejtrakulwong, P., Mukerji, T., and Mavko, G. (2012). Using kernel principal component analysis to interpret seismic signatures of thin shaly-sand reservoirs. In SEG Technical Program Expanded Abstracts 2012. Society of Exploration Geophysicists.

Fichtner, A., and Simutė, S. (2018). Hamiltonian Monte Carlo inversion of seismic sources in complex media. Journal of Geophysical Research: Solid Earth, 123(4), 2984-2999.

Fichtner, A., and Zunino, A. (2019). Hamiltonian nullspace shuttles. Geophysical research letters, 46(2), 644-651.

Fichtner, A., Zunino, A., and Gebraad, L. (2019). Hamiltonian Monte Carlo solution of tomographic inverse problems. Geophysical Journal International, 216(2), 1344-1363.

Gebraad, L., Boehm, C., and Fichtner, A. (2020). Bayesian elastic Full-Waveform Inversion using Hamiltonian Monte Carlo. Journal of Geophysical Research: Solid Earth, 125(3).

Grana, D., Passos de Figueiredo, L., and Azevedo, L. (2019). Uncertainty quantification in Bayesian inverse problems with model and data dimension reduction. Geophysics, 84(6), M15-M24.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. Computational Statistics, 14(3), 375-396.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. Bernoulli, 7(2), 223-242.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: efficient adaptive MCMC. Statistics and computing, 16(4), 339-354.

Holmes, C., Krzysztof, L. and Pompe, E. (2017). Adaptive MCMC for multimodal distributions. Technical report. https://pdfs.semanticscholar.org/c75d/f035c23e3c0425409e70d457cd43b174076f.pdf.

Horta, A., and Soares, A. (2010). Direct sequential co-simulation with joint probability distributions. Mathematical Geosciences, 42(3), 269-292.

Lieberman, C., Willcox, K., and Ghattas, O. (2010). Parameter and state model reduction for large-scale statistical inverse problems. SIAM Journal on Scientific Computing, 32(5), 2523-2542.

Lochbühler, T., Breen, S. J., Detwiler, R. L., Vrugt, J. A., and Linde, N. (2014). Probabilistic electrical resistivity tomography of a CO2 sequestration analog. Journal of Applied Geophysics, 107, 80-92.

MacKay, D.J., (2003). Information Theory, Inference and Learning Algorithms. Cambridge University Press.

Martin, J., Wilcox, L. C., Burstedde, C., and Ghattas, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM Journal on Scientific Computing, 34(3), A1460-A1487.

Menke, W. (2018). Geophysical data analysis: Discrete inverse theory. Academic press.

Neal, R.M., (2011). MCMC using Hamiltonian dynamics: in Handbook of Markov Chain Monte Carlo, Brooks, S., Gelman, A., Jones, G. and Meng, X.: Chapman and Hall, 113–162.

Pagani, F., Wiegand, M., and Nadarajah, S. (2019). An n-dimensional Rosenbrock Distribution for MCMC Testing. arXiv preprint arXiv:1903.09556.

Sambridge, M., and Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. Reviews of Geophysics, 40(3), 3-1.

Sambridge, M. (2014). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. Geophysical Journal International, 196(1), 357-374.

Sen, M. K., and Stoffa, P. L. (1996). Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion 1. Geophysical Prospecting, 44(2), 313-350.

Sen, M. K., and Stoffa, P. L. (2013). Global optimization methods in geophysical inversion. Cambridge University Press.

Sen, M. K., and Biswas, R. (2017). Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. Geophysics, 82(3), R119-R134.

Tarantola, A. (2005). Inverse problem theory and methods for model parameter estimation. siam.

Ter Braak, C. J., and Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. Statistics and Computing, 18(4), 435-446.

Tierney, L., and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. Statistics in medicine, 18(17-18), 2507-2515.

Turner, B. M., and Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. Journal of Mathematical Psychology, 56(5), 375-385.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software, 75, 273-316.

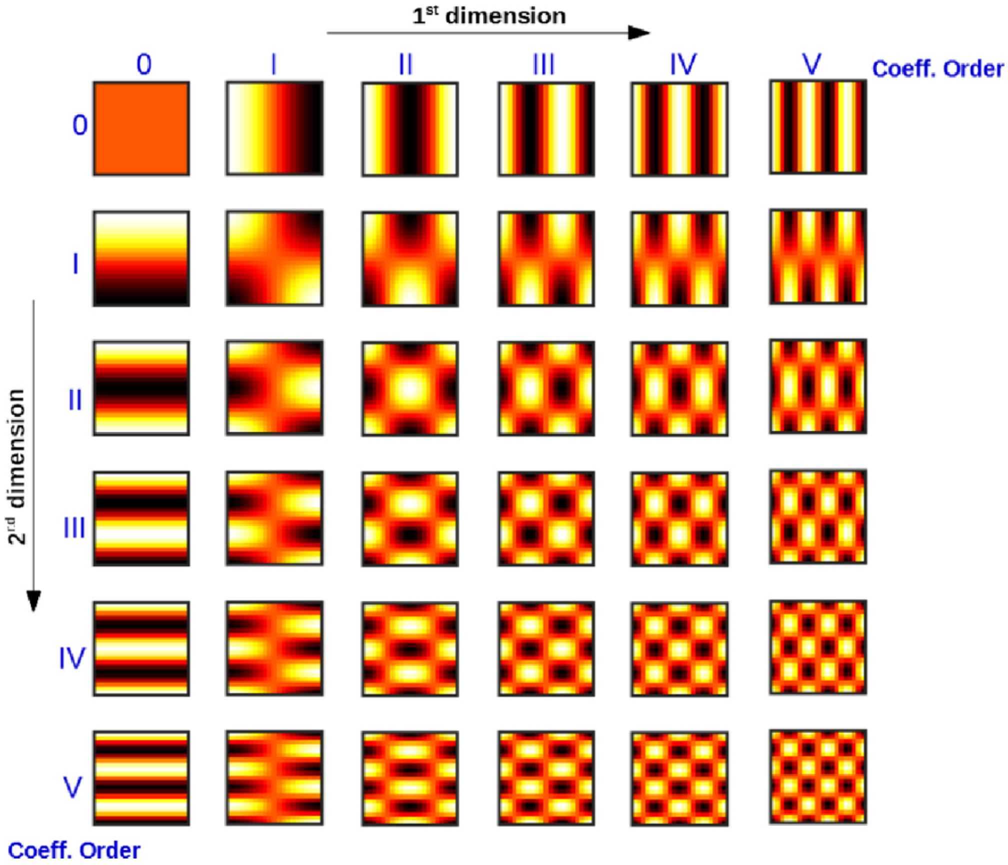

Figure 1: 2-D DCT base functions of different orders. Dark and light colors code low and high numerical values, respectively.

90x76mm (300 x 300 DPI)

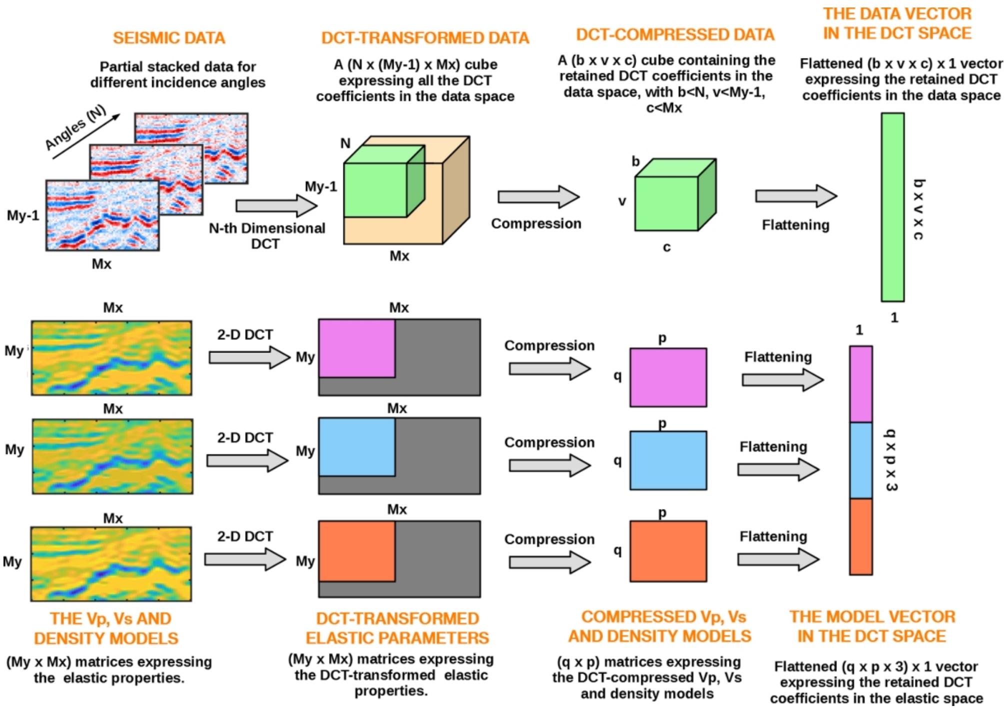
SEISMIC DATA
Partial stacked data for different incidence angles
Angles (N)
My-1
Mx
N-th Dimensional DCT
DCT-TRANSFORMED DATA
A (N x (My-1) x Mx) cube expressing all the DCT coefficients in the data space
N
My-1
Mx
Compression
DCT-COMPRESSED DATA
A (b x v x c) cube containing the retained DCT coefficients in the data space, with b<N, v<My-1, c<Mx
b
v
c
Flattening
THE DATA VECTOR IN THE DCT SPACE
Flattened (b x v x c) x 1 vector expressing the retained DCT coefficients in the data space
b x v x c
1
Mx
My
2-D DCT
Mx
My
Compression
p
q
Flattening
1
Mx
My
2-D DCT
Mx
My
Compression
p
q
Flattening
q x p x 3
Mx
My
2-D DCT
Mx
My
Compression
p
q
Flattening
THE Vp, Vs AND DENSITY MODELS
(My x Mx) matrices expressing the elastic properties.
DCT-TRANSFORMED ELASTIC PARAMETERS
(My x Mx) matrices expressing the DCT-transformed elastic properties.
COMPRESSED Vp, Vs AND DENSITY MODELS
(q x p) matrices expressing the DCT-compressed Vp, Vs and density models
THE MODEL VECTOR IN THE DCT SPACE
Flattened (q x p x 3) x 1 vector expressing the retained DCT coefficients in the elastic space

Figure 2: Derivation of data and model space vectors in the DCT space from the elastic properties and seismic gathers.

160x111mm (300 x 301 DPI)

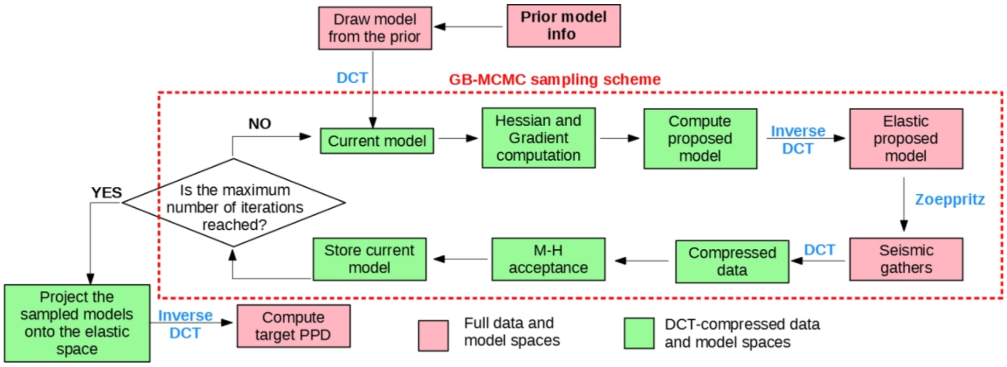

Figure 3: Schematic representation of the GB-MCMC inversion scheme. Green and pink rectangles refer to steps performed in the reduced and full spaces, respectively.

150x53mm (300 x 300 DPI)

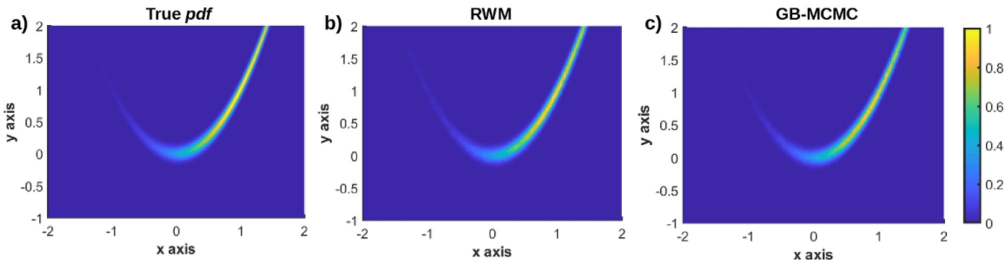

Figure 4: a) True posterior density function. b) Posterior density provided by the random walk Metropolis. c) Posterior density estimated by the GB-MCMC. The colormap codes the normalized probability values.

130x32mm (300 x 300 DPI)

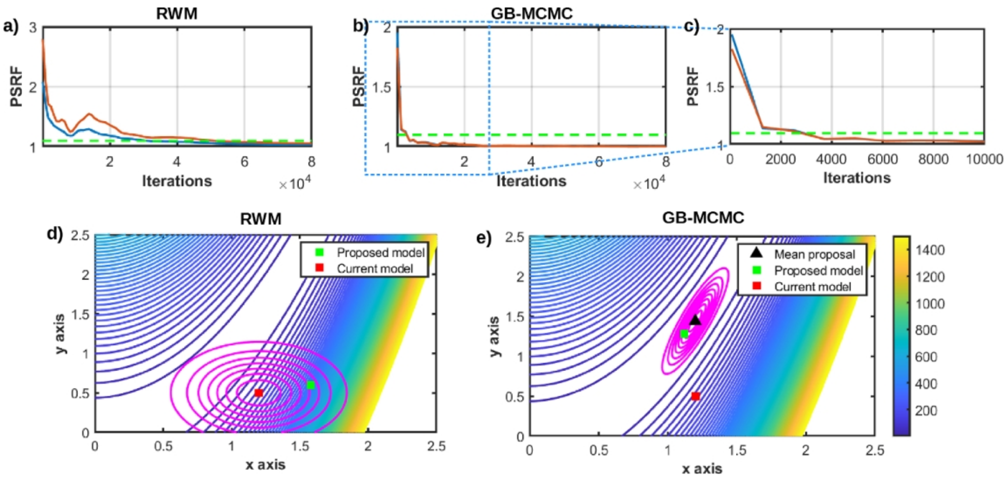

Figure 5: a) Evolution of the potential scale reduction factor for the random walk Metropolis. b) Evolution of the potential scale reduction factor for the GB-MCMC. c) Close-up of b). In a)-c) the horizontal dotted green lines represent the threshold of convergence, whereas the blue and red lines refer to the x and y variables, respectively. d) Example of current, proposed model, and proposal distribution for the random walk Metropolis. e) Example of current, proposed model, and proposal distribution for the GB-MCMC. In d) and e) the magenta curves represents the contour lines of the proposal while the colored curves are the contour lines of the Rosenbrock error function.

130x60mm (300 x 300 DPI)

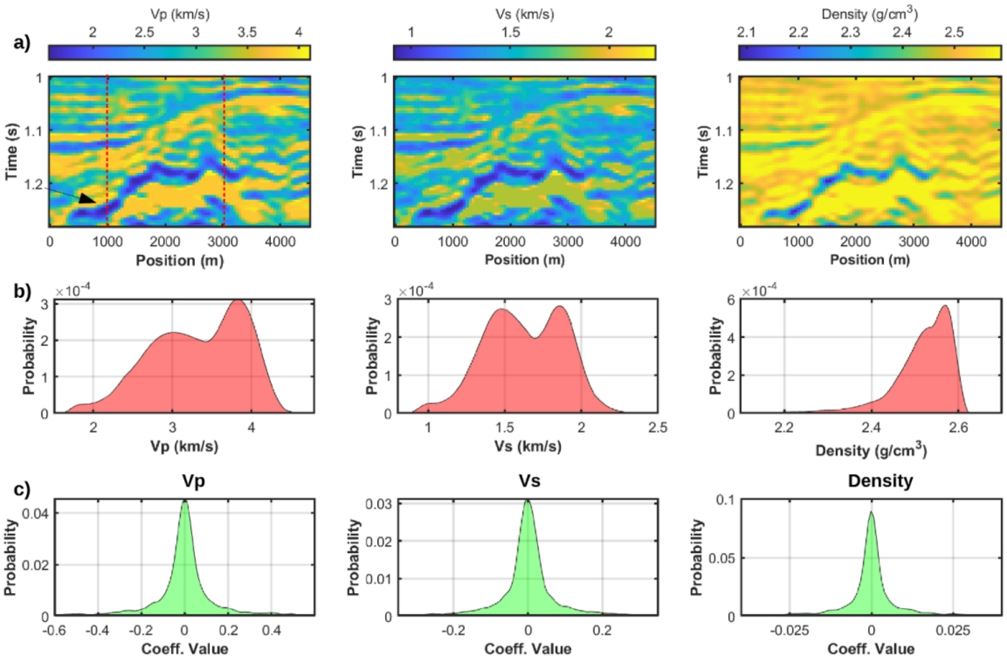

Figure 6: a) The elastic properties of Vp, Vs, and density of the reference model. In a) the black arrows point toward the main sand reservoir body, whereas the dotted red lines depict the columns of the model considered as available well log data for defining the a-priori elastic distribution. b) The marginal non-parametric prior distributions for the three elastic properties derived from the two wells shown in a). c) The marginal prior projected onto the compressed space through a Monte Carlo simulation.

130x84mm (300 x 300 DPI)

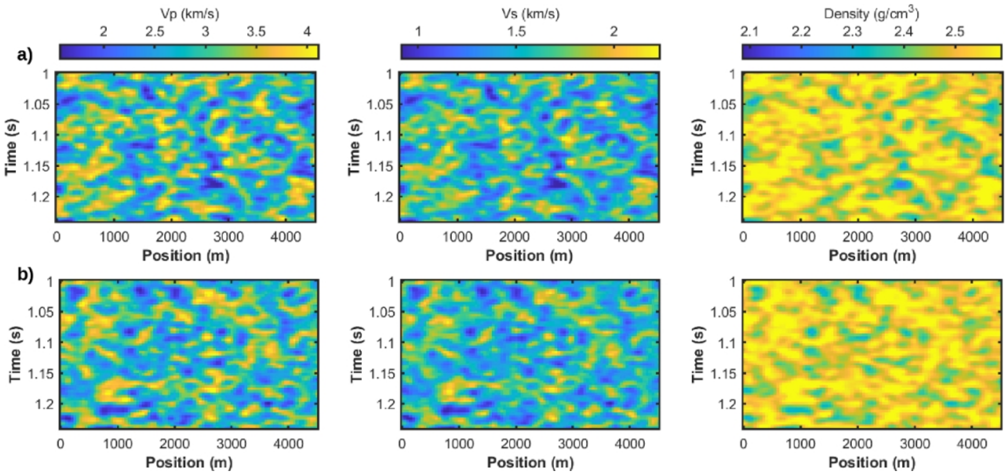

Figure 7: a), b) Two examples of Vp, Vs, and density model drawn from the non-parametric elastic prior.

120x56mm (300 x 300 DPI)

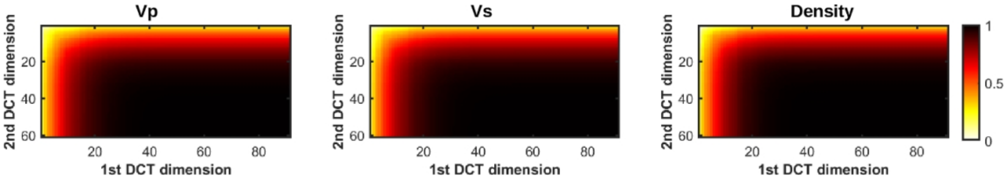

Figure 8: Examples of explained model variability for an elastic model extracted from the prior and as the number of coefficients along the 1st and 2nd DCT dimension increases. In each plot, the numerical value with coordinate (x, y) indicates the explained variability if the first x, and y coefficients along the 1st and 2nd dimensions, respectively, are used for compressing the model. It emerges that 25 coefficients along both the 1st dimension explain almost the 100 % of the variability of the uncompressed Vp, Vs, and density profiles.

120x20mm (300 x 300 DPI)

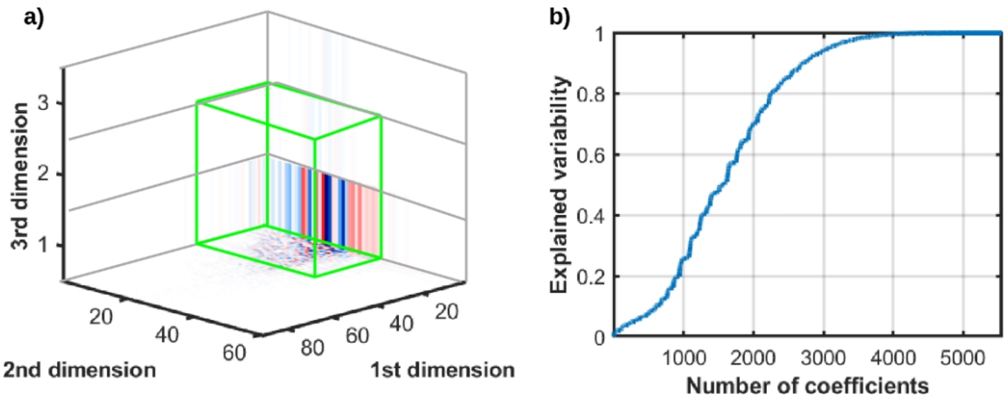

Figure 9: a) DCT decomposition of a seismic gather computed on an elastic model drawn from the prior. Blue and red colors code low and high values, respectively while the green rectangles enclose the retained coefficients in the data space. b) Explained data variability as the number of considered basis functions increases.

100x38mm (300 x 300 DPI)

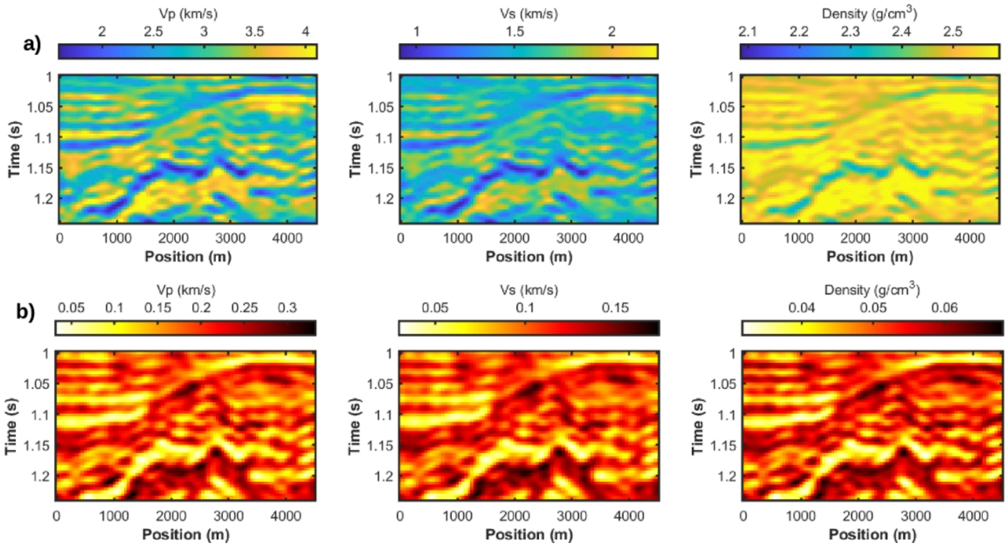

Figure 10: Results provided by the GB-MCMC-FD approach for Tests 1. a) A-posteriori mean model. b) Posterior standard deviation. In a), and b) the Vp, Vs, and density are represented from left to right.

120x64mm (300 x 300 DPI)

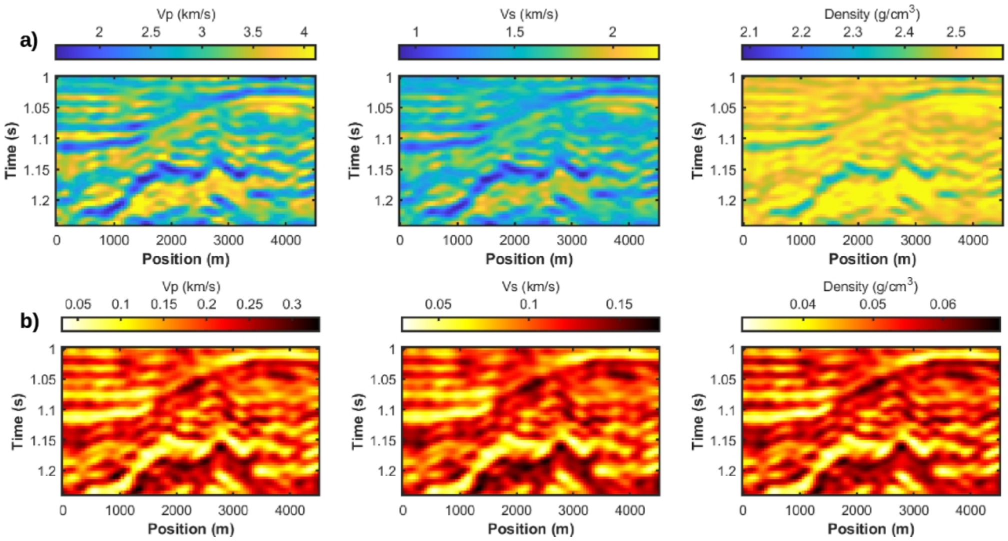

Figure 11: As in Figure 10 but for the GB-MCMC-L approach.

120x64mm (300 x 300 DPI)

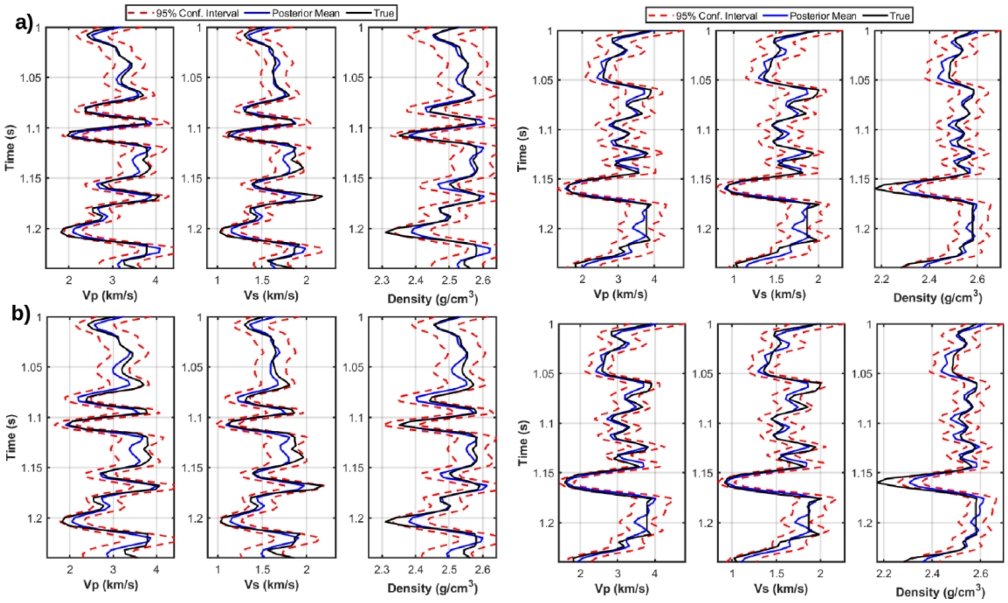

Figure 12: Comparison between the true model, the posterior mean, and 95% confidence interval at two different spatial locations. a) GB-MCMC-FD. b) GB-MCMC-L. The leftmost plot refers to the spatial position of 1200 m, while the plot on the right refers to the spatial position of 2500 m.

140x83mm (300 x 300 DPI)

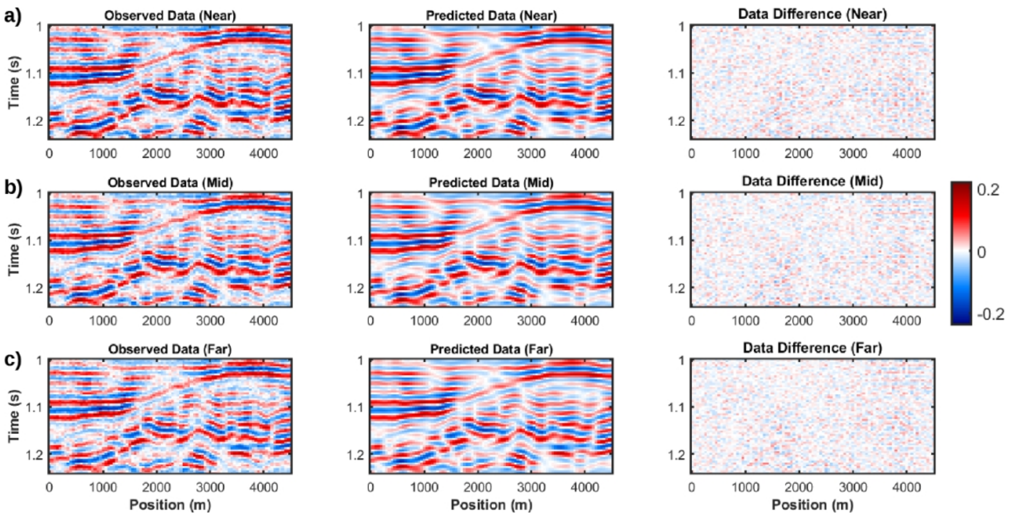

Figure 13: Comparison between observed data (left column), predicted data (central column), and their sample-by-sample difference (right column) for Test 1. The predicted data have been computed on the mean posterior model estimated by the GB-MCMC-L algorithm. a), b), and c) refer to near, mid and far stack, respectively.

120x60mm (300 x 300 DPI)

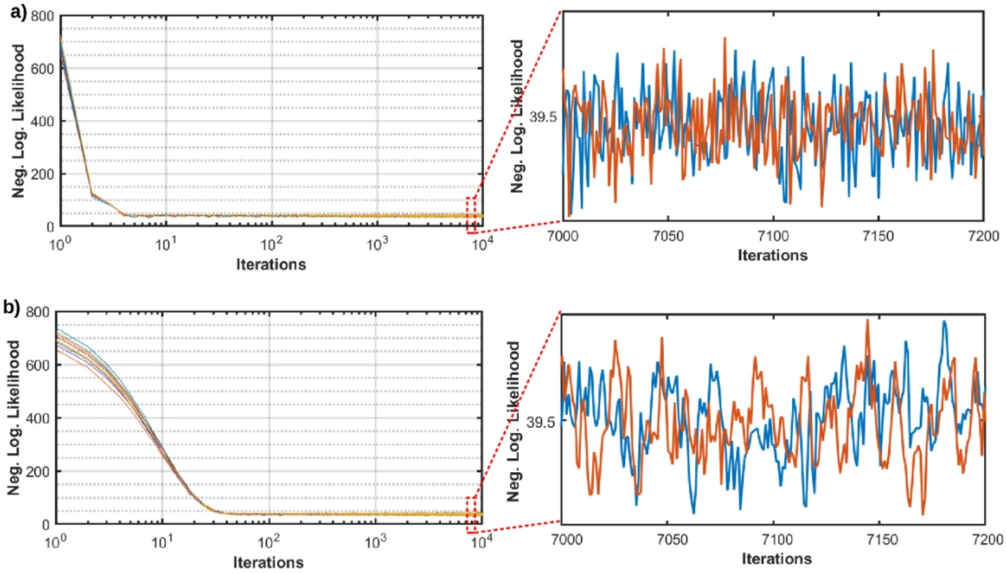

Figure 14: Evolution of the negative log-likelihood values for the GB-MCMC-FD and GB-MCMC-L inversions (part a) and b), respectively). Each color represents a different chain.

110x62mm (300 x 300 DPI)

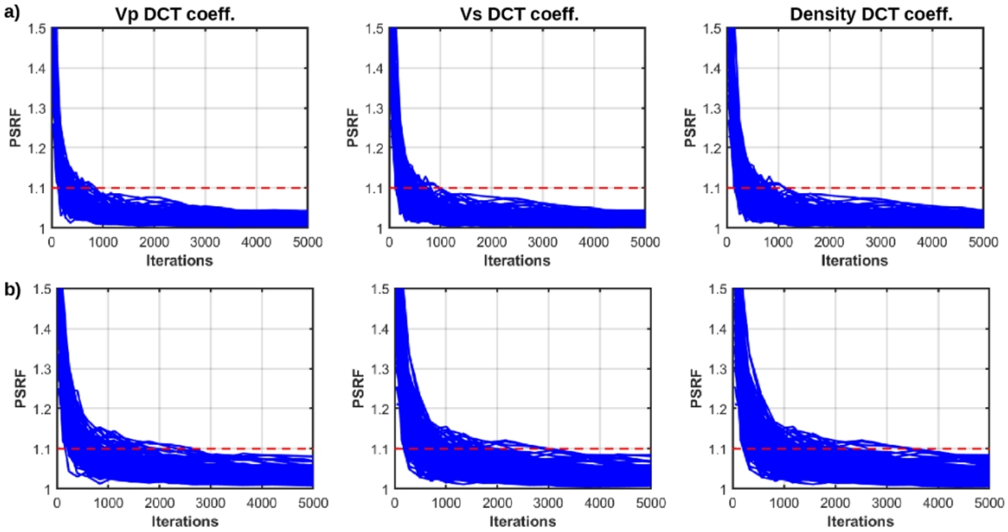

Figure 15: Evolution of the potential scale reduction factor over iterations for the DCT coefficients associated with the three elastic properties. a) GB-MCMC-FD. b) GB-MCMC-L. The red dotted lines depict the threshold of convergence.

110x57mm (300 x 300 DPI)

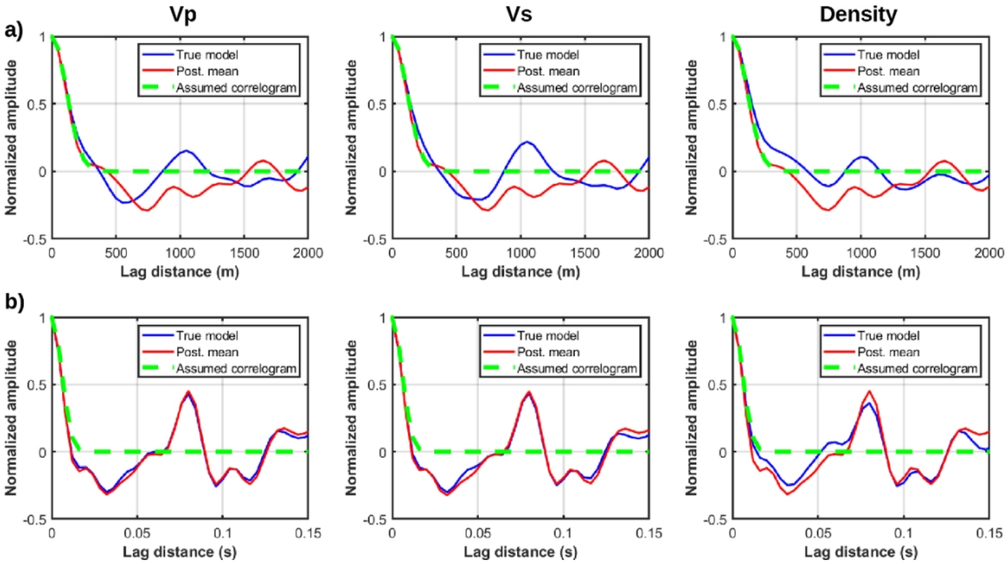

Figure 16: Comparison between the lateral (a) and vertical (b) assumed correlogram functions with the average correlograms computed on the true model (blue line) and on the posterior mean estimated by the GB-MCMC-L algorithm (red lines). From left to right we represent Vp, Vs, and density.

110x61mm (300 x 300 DPI)

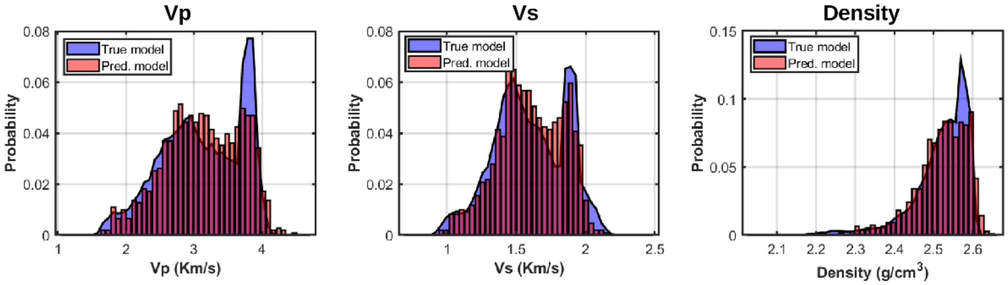

Figure 17: Marginal probabilities for the three elastic parameters computed on the true model and on the posterior mean estimated by the GB-MCMC-L algorithm.

100x27mm (300 x 300 DPI)

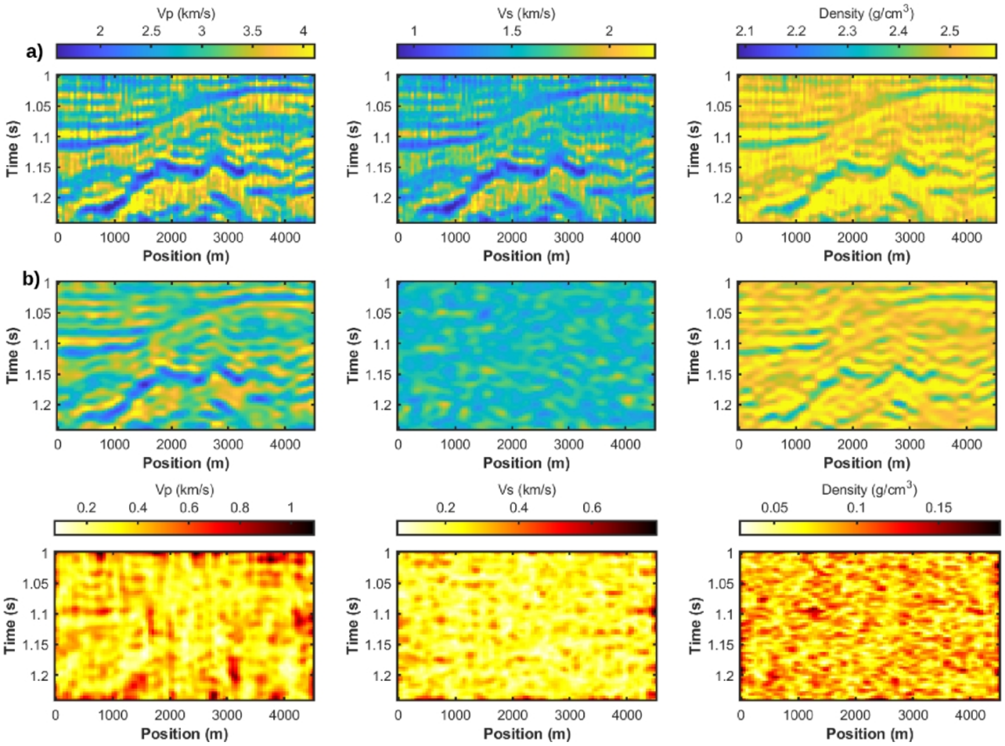

Figure 18: a) Results of the linearized least-squares inversion. b) Estimated mean model by the DEMC algorithm. c) Posterior standard deviation estimated by the DEMC algorithm.

120x88mm (300 x 300 DPI)

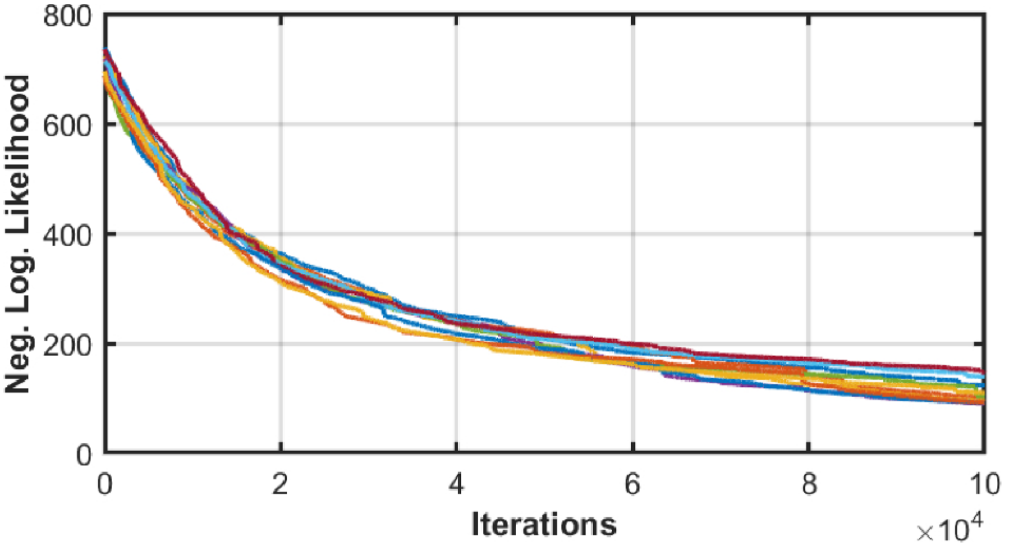

Figure 19: Evolution of the negative log-likelihood value during the DEMC sampling. Each color refers to a different chain.

80x43mm (300 x 300 DPI)

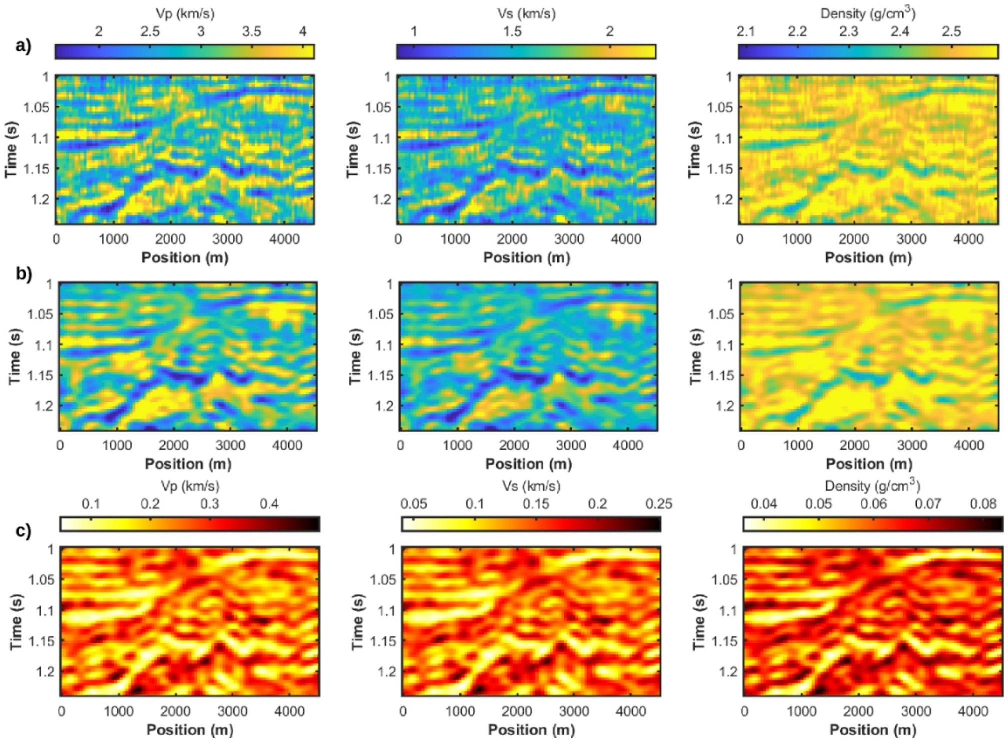

Figure 20: Results for Test 2: a) Vp, Vs, and density profiled estimated by the linearized least-squares approach. b) Posterior mean model provided by the GB-MCMC-L approach. c) Posterior standard deviation estimated by the GB-MCMC-L inversion.

120x88mm (300 x 300 DPI)

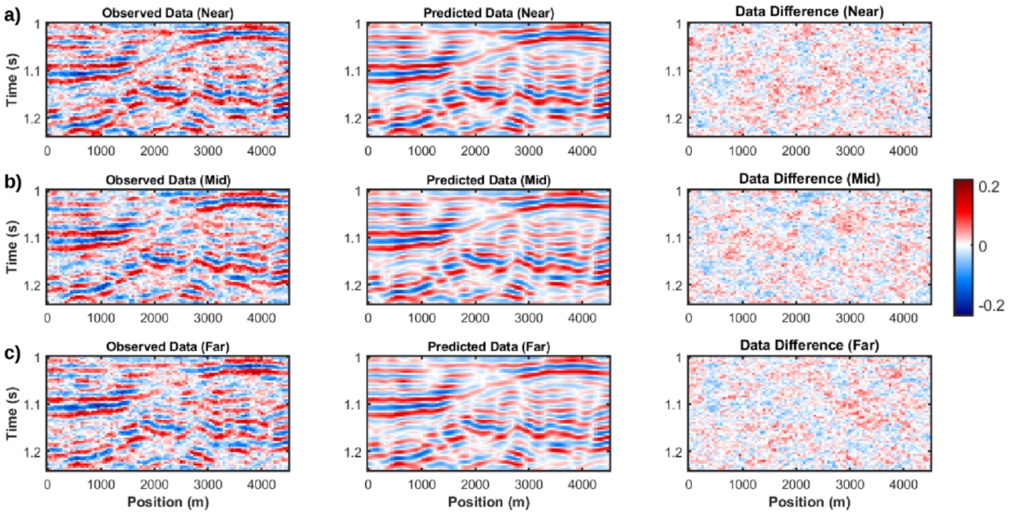

Figure 21: Comparison between observed data (left column), predicted data (central column), and their sample-by-sample difference (right column) for Test 2. The predicted data have been computed on the mean posterior model estimated by the GB-MCMC-L algorithm. a), b) and c) refer to near, mid and far stack, respectively.

120x60mm (300 x 300 DPI)

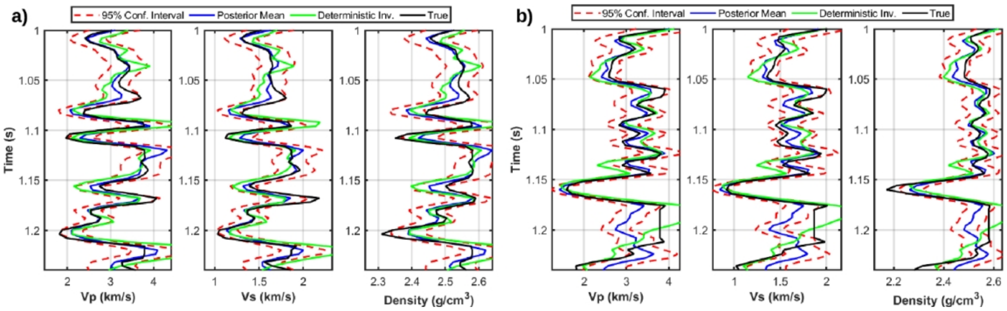

Figure 22: Comparison between the true model, the deterministic inversion results, the posterior mean, and 95% confidence interval estimated by the GB-MCMC- L approach. a) refers to the spatial position of 1200 m, while b) refers to the spatial position of 2500 m.

130x39mm (300 x 300 DPI)

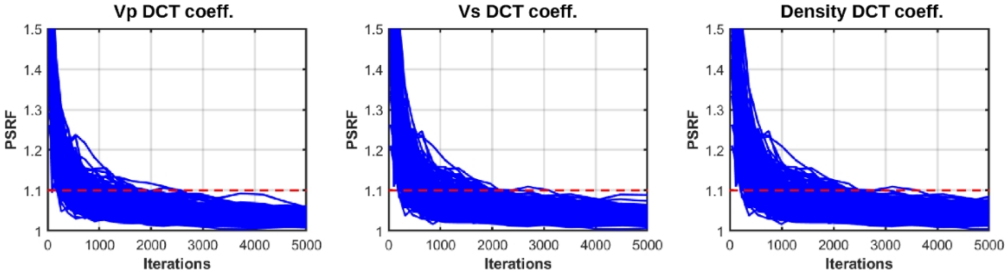

Figure 23: Evolution of the potential scale reduction factor over iterations and for the coefficients associated with the three elastic properties. The dotted red lines depict the threshold of convergence.

110x29mm (300 x 300 DPI)

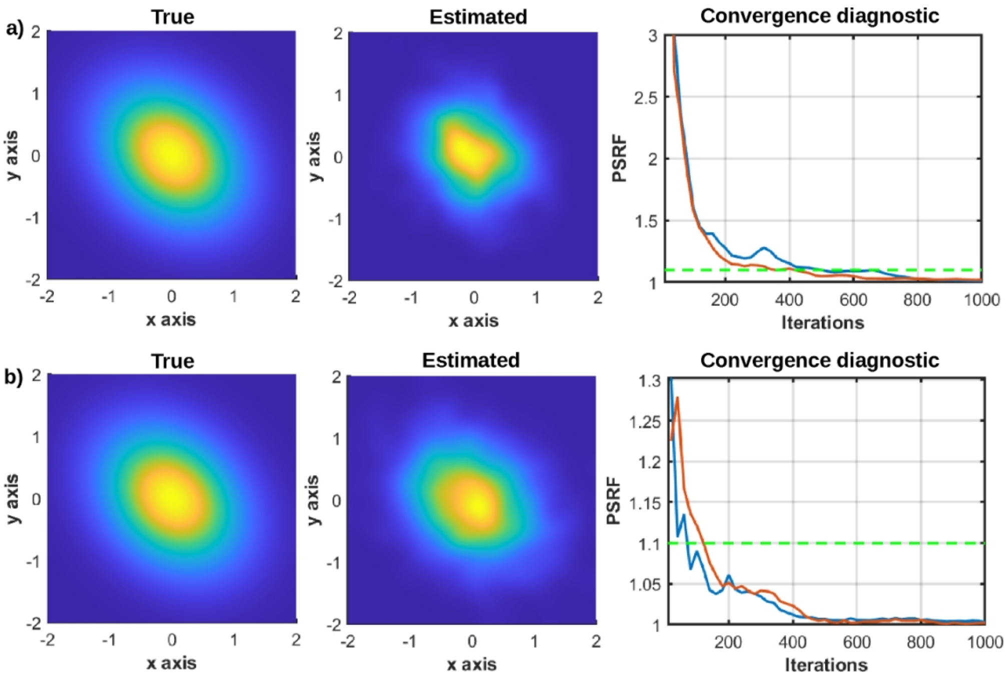

Figure 24: GB-MCMC sampling of a 2D multivariate density for different hyperparameter settings. a) λ=0.05 and μ^2=3. b) λ=0.5 and μ^2=0.5. From left to right we represent the target probability density, the estimated probability density, and the evolution of the PSRF for the two parameters. Blue and yellow colors code low and high probability values, respectively. On the rightmost plot, the horizontal dotted green line represents the threshold of convergence, whereas the blue and red lines refer to the x and y variable, respectively.

120x79mm (300 x 300 DPI)

TABLE

Method	**Time per iteration (s)**	**# iterations to converge**	**Time to converge (hours)**
GB-MCMC-FD	30	1000	8.3
GB-MCMC-L	2.5	4000	2.8
DEMC	0.4	>>100000	>> 11.1

Table 1: Some characteristics of the GB-MCMC-FD, GB-MCMC-L, and DEMC inversions (see the text for details). The computational cost of the deterministic inversion is negligible compared to the MCMC algorithms.