GEOPHYSICS, VOL. 86, NO. 5 (SEPTEMBER-OCTOBER 2021); P. A45–A49, 4 FIGS.
10.1190/GEO2021-0118.1

DOI:10.1190/geo2021-0118.1

Bayesian full-waveform inversion with realistic priors

Xin Zhang[1] and Andrew Curtis[1]

ABSTRACT

Seismic full-waveform inversion (FWI) uses full seismic records to estimate the subsurface velocity structure. This requires a highly nonlinear and nonunique inverse problem to be solved; therefore, Bayesian methods have been used to quantify uncertainties in the solution. Variational Bayesian inference uses optimization to efficiently provide solutions. However, previously the method has only been applied to a transmission FWI problem and with strong prior information imposed on the velocity such as is never available in practice. We have found that the method works well in a seismic reflection setting and with realistically weak prior information, representing the type of problem that occurs in reality. We conclude that the method can produce high-resolution images and reliable uncertainties using data from standard reflection seismic acquisition geometry, realistic nonlinearity, and practically achievable prior information.

INTRODUCTION

Seismic full-waveform inversion (FWI) produces high-resolution subsurface images directly from seismic waveforms (Tarantola, 1984). FWI is traditionally solved by minimizing the difference between predicted and observed seismograms. In such methods, a good starting model is often required because of multimodality of the misfit functions caused by the nonlinearity of the problem. Also, those methods cannot provide accurate estimates of uncertainties, which are required to better understand and interpret the resulting images.

Monte Carlo sampling methods provide a general way to solve nonlinear inverse problems and quantify uncertainties, and they have been applied to solve FWI problems (Ray et al., 2016; Zhao and Sen, 2019; Gebraad et al., 2020; Guo et al., 2020). However, Monte Carlo methods are usually computationally expensive and all Markov chain Monte Carlo (MCMC)-based methods are difficult to fully parallelize. For example, MCMC methods usually require a large number of samples, but they cannot be parallelized across successive samples, which restricts their real-time computational efficiency. In practice, they also can be difficult to tune (Gebraad et al., 2020).

Variational inference provides an efficient, fully parallelizable alternative methodology. This class of methods optimizes an approximation to a probability distribution describing inverted parameter uncertainties (Blei et al., 2017). The method can be parallelized at the sample level, is easy to tune by using adaptive gradient-descent methods (Duchi et al., 2011), and can take advantage of stochastic optimization techniques, which cannot be applied within MCMC methods. Variational inference has been applied to petrophysical inversion (Nawaz and Curtis, 2018), traveltime tomography (Zhang and Curtis, 2020a), and more recently to FWI (Zhang and Curtis, 2020b). In the latter study, strong constraints were imposed on the velocity structure to limit the space of possible models; unfortunately, such strong constraints are almost never available in practice. In addition, the method has only been applied to wavefield transmission problems in which seismic data are recorded on a receiver array that lies above the structure to be imaged, given known double-couple (earthquake-like) sources located underneath the same structure. The transmission FWI problem is less nonlinear than reflection FWI problems, and, in practice, knowledge of such sources is never definitive because it usually depends circularly on the unknown structure itself. Thus, the method has not been demonstrated in a problem with real-world acquisition geometries, nonlinearities, and realistic prior information. In this study, we therefore apply variational inference to solve FWI problems with more practically realistic prior probabilities and using seismic reflection data as would be acquired from active near-surface sources, which represents a realistic problem.

In the next section, we briefly summarize the goals of variational inference and the specific method of the Stein variational gradient descent (SVGD). Then, we demonstrate the method by solving an acoustic reflection FWI problem using the Marmousi model with practically reasonable prior information. To further explore the method, we perform multiple inversions using data from different

Manuscript received by the Editor 18 February 2021; revised manuscript received 21 April 2021; published ahead of production 15 June 2021; published online 30 August 2021.
[1]University of Edinburgh, School of Geosciences, Edinburgh EH9 3FE, UK. E-mail: x.zhang2@ed.ac.uk (corresponding author); andrew.curtis@ed.ac.uk.

frequency ranges, and we demonstrate that the method produces high-resolution velocity models and uncertainties.

METHODS

SVGD

Bayesian inference solves inverse problems by finding the probability distribution function (PDF) of model $\mathbf{m}$ given prior information and observed data $\mathbf{d}_{\text{obs}}$. This is called a posterior PDF written as $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. By Bayes' theorem,

$$p(\mathbf{m}|\mathbf{d}_{\text{obs}}) = \frac{p(\mathbf{d}_{\text{obs}}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{\text{obs}})}, \tag{1}$$

where $p(\mathbf{m})$ is the prior PDF which characterizes the probability distribution of model $\mathbf{m}$ prior to the inversion, $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ is the *likelihood* that represents the probability of observing data $\mathbf{d}_{\text{obs}}$ given model $\mathbf{m}$, and $p(\mathbf{d}_{\text{obs}})$ is a normalization factor called the *evidence*.

Variational inference solves Bayesian inference problems using optimization. The method seeks an optimal approximation to the posterior PDF within a predefined family of PDFs. This is achieved by minimizing the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the approximating PDF and the posterior PDF. Variational inference has been shown to be an efficient alternative to Monte Carlo sampling methods for a range of geophysical applications (Nawaz and Curtis, 2018; Zhang and Curtis, 2020a, 2020b).

SVGD is one such algorithm that iteratively updates a set of models, called particles $\{\mathbf{m}^i\}$ generated from an initial distribution $q(\mathbf{m})$ using a smooth transform:

$$T(\mathbf{m}^i) = \mathbf{m}^i + \epsilon\boldsymbol{\phi}(\mathbf{m}^i), \tag{2}$$

where $\mathbf{m}^i$ is the ith particle, $\boldsymbol{\phi}(\mathbf{m}^i)$ is a smooth vector function representing the perturbation direction, and ϵ is the magnitude of the perturbation. Define q_T as the PDF of q after transforming by T. The gradients of the KL-divergence between q_T and the posterior PDF $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ with respect to ϵ are found to be (Liu and Wang, 2016)

$$\nabla_\epsilon \text{KL}[q_T||p]|_{\epsilon=0} = -\text{E}_q[\text{trace}(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m}))], \tag{3}$$

where $\mathcal{A}_p$ is the Stein operator defined as

$$\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_\mathbf{m} \log p(\mathbf{m}|\mathbf{d}_{\text{obs}})\boldsymbol{\phi}(\mathbf{m})^{\text{T}} + \nabla_\mathbf{m}\boldsymbol{\phi}(\mathbf{m}). \tag{4}$$

The optimal $\boldsymbol{\phi}$ that maximizes the right expectation is found to be

$$\boldsymbol{\phi}^*(\mathbf{m}) \propto \text{E}_{\{\mathbf{m}'\sim q\}}[\mathcal{A}_p k(\mathbf{m}', \mathbf{m})], \tag{5}$$

where $k(\mathbf{m}', \mathbf{m})$ is a kernel function. This is the direction of steepest descent in KL divergence; therefore, the KL divergence can be minimized by iteratively stepping a small distance in that direction. The expectation $\text{E}_{\{\mathbf{m}'\sim q\}}$ is calculated using the set of particles $\{\mathbf{m}^i\}$, and then $\boldsymbol{\phi}^*(\mathbf{m})$ is used to update each particle using equation 2. This process is iterated to equilibrium, at which point the particles are optimally distributed according to the posterior PDF. The kernel function ensures that all pairs of particles interact, which helps the method to jump out of local optima: A particle in a local optimum can be driven out by other particles that are not in that local optimum.

In SVGD, the choice of kernel can affect the efficiency of the method. Instead of the commonly used scalar radial-basis kernel, in this study we apply a matrix-valued kernel to improve efficiency:

$$\mathbf{k}(\mathbf{m}', \mathbf{m}) = \mathbf{Q}^{-1} \exp\left(-\frac{1}{2h}\|\mathbf{m} - \mathbf{m}'\|^2_\mathbf{Q}\right), \tag{6}$$

where $\mathbf{Q}$ is a positive definite matrix, $\|\mathbf{m} - \mathbf{m}'\|^2_\mathbf{Q} = (\mathbf{m} - \mathbf{m}')^{\text{T}}\mathbf{Q}(\mathbf{m} - \mathbf{m}')$, and h is a scaling parameter. Wang et al. (2019) show that by setting $\mathbf{Q}$ to be the Hessian matrix, the method converges faster than with a scalar kernel. However, the Hessian matrix is usually expensive to compute. One alternative is to use the covariance matrix calculated from the particles, but the full covariance matrix may occupy large memory and is difficult to estimate from a relatively small number of samples (Ledoit and Wolf, 2004). We therefore use a diagonal covariance matrix: $\mathbf{Q}^{-1} = \text{diag}(\text{var}(\mathbf{m}))$, where $\text{var}(\mathbf{m})$ is the variance estimated from the particles. For those parameters with higher variance, this choice applies higher weights to the posterior gradients to induce larger perturbations, and it also enables interactions with more distant particles.

Variational FWI

We apply SVGD to solve an acoustic FWI problem. The constant density wave equation is solved using a time-domain finite-difference method. Gradients of the likelihood function with respect to velocity are calculated using the adjoint method (Plessix, 2006). For the likelihood function, we assume Gaussian data errors with a diagonal covariance matrix:

$$p(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto \exp\left[-\frac{1}{2}\sum_i\left(\frac{d_i^{\text{obs}} - d_i(\mathbf{m})}{\sigma_i}\right)^2\right], \tag{7}$$

where i is the index of time samples and σ_i is the standard deviation of each data point.

RESULTS

We apply the preceding method to a 2D acoustic FWI to recover part of the scaled Marmousi model (Martin et al., 2006) from waveform data (Figure 1). The model is discretized in space using a regular 200 × 120 grid. Sources are located at a depth of 20 m in the water layer. In total, 200 equally spaced receivers are located at a depth of 360 m across the horizontal extent of the model. We generated two waveform data sets with a maximum time of 5 s using Ricker wavelets with a dominant frequency of 4 and 10 Hz, respectively. Uncorrelated Gaussian noise with 0.1 standard deviation (1% of the average maximum amplitude of all traces) is added to the data.

Zhang and Curtis (2020b) and Gebraad et al. (2020) impose strong prior information (a uniform distribution over an interval of 0.2 km/s) on the velocity to reduce the complexity of their (identical) inverse problems. In practice, such strong prior information is almost never available. In this study, we use much weaker prior information: a uniform distribution over an interval of 2 km/s at each depth (Figure 1c). We also impose an extra lower velocity bound of 1.5 km/s to ensure that the rock velocity is higher than the acoustic velocity in water. Velocity in the water layer is fixed to be 1.5 km/s

in the inversion. This prior information mimics a practical choice that is applied in real problems.

We perform two independent inversions using the two data sets. For each inversion, we use 600 particles that are initially generated from the prior distribution (an example is shown in Figure 1b) and updated using equation 2 for 600 iterations at which point the average misfit across particles ceases to decrease. Figure 2a shows the mean model obtained using the low-frequency data. In the shallower part (<2 km), the mean model shows features similar to the true model, but it has slightly lower resolution than the true model, which probably reveals the resolution limit imposed by the restricted frequency range. In comparison, the mean model obtained using high-frequency data shows higher resolution (Figure 2d) and is more similar to the true model. In the deeper part (>2 km), mean models are different from the true model: The mean obtained using low-frequency data only shows a large-scale structure, whereas that obtained using high-frequency data shows spatially rapid variations that are different from the true model. This may be because of poor illumination of the deeper part, which causes complex posterior PDFs when using high-frequency data and which cannot be represented properly by a small number of particles. However, we also note that the mean model does not need to reflect the true model in nonlinear problems. For example, Figure 3 compares the data predicted from models with the observed data for the high-frequency inversion. Although the data predicted by a random posterior sample match the observed data (Figure 3a), the data predicted by the mean show differences compared to the observed data (the black box in Figure 3b). This is because in general the mean model does not represent the true structure when the posterior PDF is multimodal (Figure 4).

Both standard deviation models show features that are related to the mean model. For example, in the shallow part (<1 km), the standard deviation is lower at locations of lower velocity anomalies, and, in the deeper part, lower standard deviations are associated with higher velocity anomalies (see the examples denoted by red arrows in Figure 2). This phenomenon has also been found in previous studies (Gebraad et al., 2020; Zhang and Curtis, 2020b). In the shallower part, this is probably due to the fact that the low-velocity layers cause strong changes in traveltimes, to which the L2 misfit is sensitive. Similarly, in the deeper part, those strong high-velocity anomalies can have a large influence on seismic waveforms and hence have lower standard deviations. In most areas, the error between the mean and true models obtained using the two data sets is within two standard deviations (Figure 2c and 2f). In the deeper part (>1.5 km) and close to the sides, errors are higher because of poor illumination. There are also higher errors at the boundaries of anomalies, which suggests that the boundary locations are not well constrained by the data, producing the equivalent of uncertainty loops (Galetti et al., 2015).

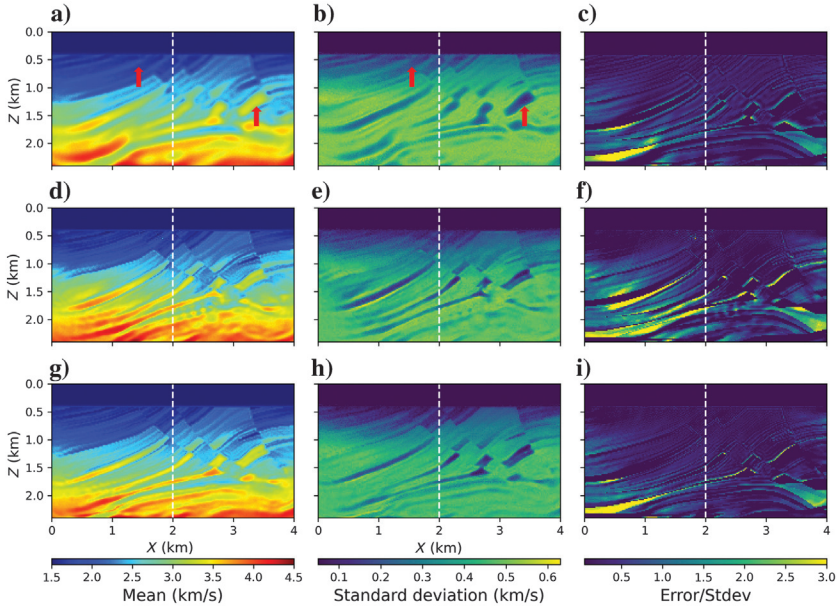

Figure 2. (a, d, and g) Mean, (b, e, and h) standard deviation, and (c, f, and i) the error between the mean and true models divided by the standard deviation obtained, respectively, using low-frequency data, high-frequency data only, and using high-frequency data but starting from the results of low-frequency data. The white lines denote the well location referred to in the main text.

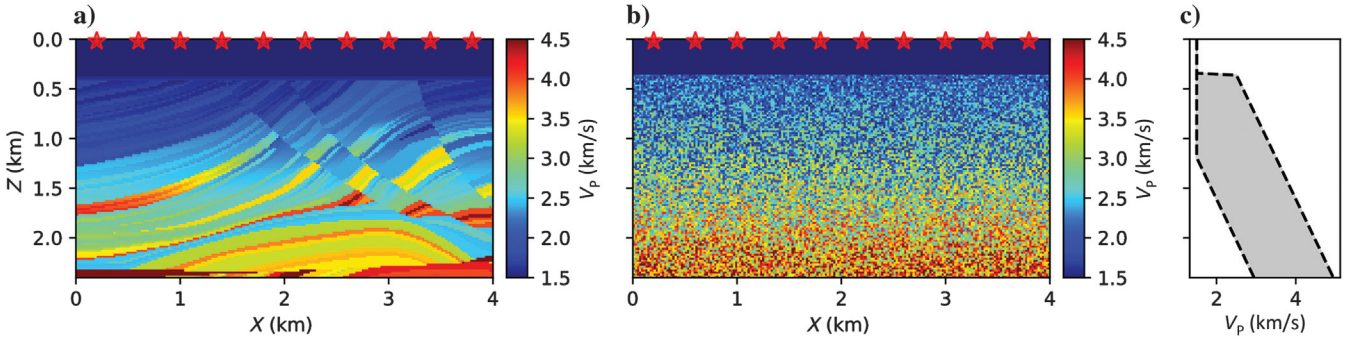

Figure 1. (a) The true velocity model. The red stars denote the locations of 10 sources. The 200 receivers are equally spaced at 0.36 km in depth. (b) A random model generated from the prior distribution. (c) The prior distribution of seismic velocity, which is chosen to be a uniform distribution over an interval of up to 2 km/s at each depth. A lower velocity bound of 1.5 km/s is imposed to ensure that the velocity is higher than the acoustic velocity in water.

To improve the results in the deeper part, we conducted another inversion using high-frequency data but starting the SVGD iterations with the particles generated using the low-frequency data, similarly to the idea of multiscale FWI (Bunks et al., 1995). After running for 300 iterations, the mean model shows features more similar to the true model in the deeper part (Figure 2g). The standard deviation model (Figure 2h) has a smoother structure than in the previous results, and the error between the true model and the mean model is significantly smaller (Figure 2i).

To further understand the results, we show marginal velocity distributions at the horizontal location X = 2 km (the white line in Figure 2) along with 1D histograms at four depths: 0.6, 1.2, 1.8, and 2.4 km. Overall, the marginal distributions obtained using high-frequency data are narrower. In the shallower part (<1.5 km), all marginal distributions show high probabilities around the true velocity (the red lines in Figure 4). In the deeper part, the marginal distributions show complex multimodal distributions, and the high-probability area of the marginal distribution obtained using only high-frequency data deviates from the true values. In comparison, the marginal distributions obtained using the results of low-frequency inversion as starting particles for the high-frequency inversion show high probabilities close to the true values. This clearly indicates that the method can get stuck at local modes in regions of poor illumination when using only high-frequency data; for example, at the depth of 1.8 km, only one incorrect mode is found (Figure 4b). By starting from particles obtained using low-frequency data, this issue can largely be resolved.

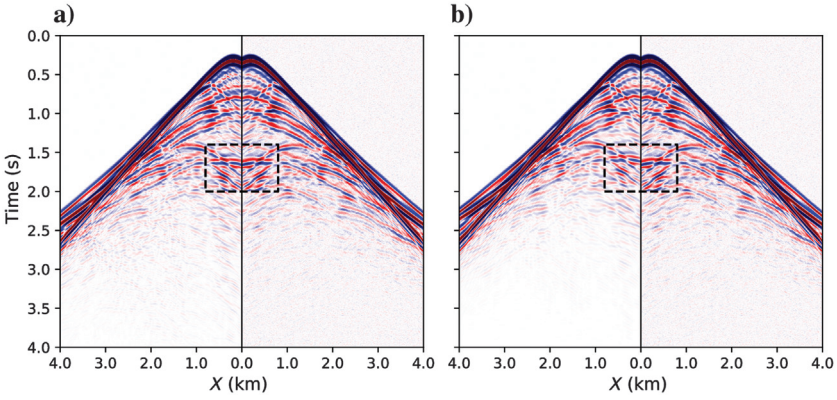

Figure 3. The "butterfly plot" of data comparison for the rightmost shot in Figure 1 obtained using the high-frequency inversion. (a) The data predicted by a random model from the posterior PDF (left) and the observed data (right). (b) The data predicted by the mean model in Figure 2d (left) and the observed data (right). The black box denotes an example location where data predicted by a posterior sample and the mean are different.

DISCUSSION

Because SVGD uses hundreds of particles and updates them iteratively, the method can be computationally expensive (cost similar to running hundreds of standard FWIs). For example, the preceding inversion with 600 iterations took approximately 6703 central processing unit (CPU) hours, which required 74 h to run using 90 Intel Xeon E5-2630 CPU cores. In practice, stochastic minibatch optimization (Robbins and Monro, 1951) can be used to improve computational efficiency for larger data sets and 3D applications. Because the method does not require accurate prior information as shown in our high-frequency example (which is difficult to solve using standard FWI with local optimization methods), we propose that the results obtained using a small data set could be used to provide a reliable starting model for standard FWI with larger data sets to produce higher resolution models.

To improve the method's efficiency, other full matrix kernels might be used, for example, Hessian matrix kernels (Wang et al., 2019) or Stein variational Newton methods (Detommaso et al., 2018). Improved prior information also may be used to improve efficiency, for example, prior velocity information from traveltime tomography. Faster, approximate forward modeling methods also may be used to provide solutions more rapidly, for example, neural network-based forward modeling methods (Meng et al., 2020).

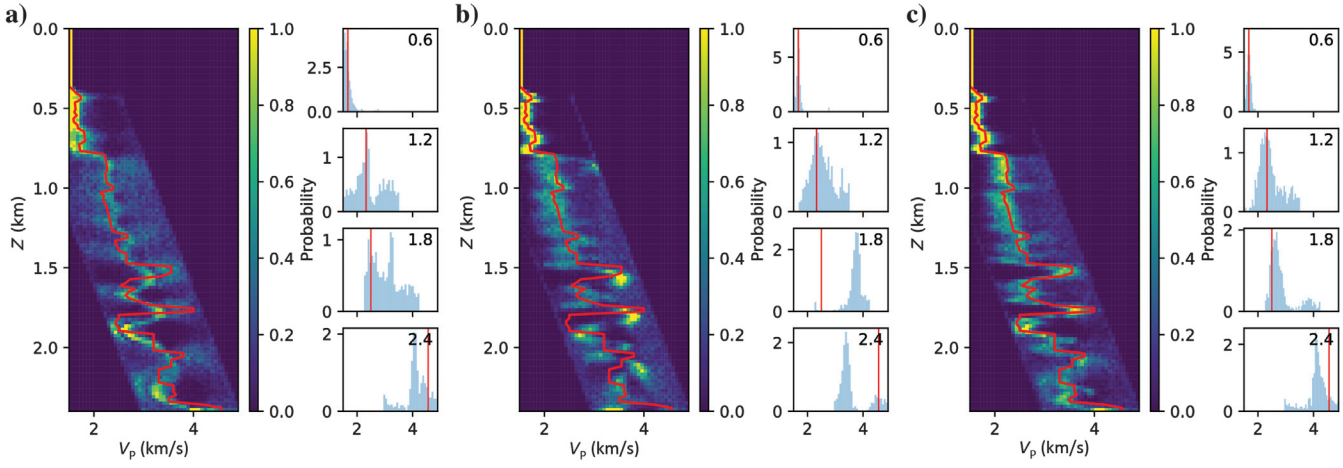

Figure 4. The marginal distributions at horizontal location X = 2 km (the white line in Figure 2) plotted along with four histograms at depths of 0.6, 1.2, 1.8, and 2.4 km. (a-c) Marginal distributions obtained using low-frequency data, high-frequency data only, and using high-frequency data but starting from the results of low-frequency data, respectively. The red lines show true values.

CONCLUSION

In this study, we have presented the first application of variational full-waveform inversion (VFWI) in a seismic reflection setting. To explore the applicability of the method, we imposed realistically weak prior information on seismic velocity: a uniform prior PDF across a 2 km/s interval, and we performed multiple inversions using data from different frequency ranges. The results showed that the method can produce high-resolution mean and uncertainty models using only high-frequency data, but it can get stuck in local modes in areas of poor illumination. This can be resolved by using the results obtained from low-frequency data to initiate high-frequency inversions. We therefore conclude that VFWI may be a useful method to produce high-resolution seismic reflection images with reliable uncertainties.

ACKNOWLEDGMENTS

The authors thank the Edinburgh Imaging Project sponsors (BP, Schlumberger, and Total) for supporting this research. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

DATA AND MATERIALS AVAILABILITY

No real data are used.

REFERENCES

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe, 2017, Variational inference: A review for statisticians: Journal of the American Statistical Association, **112**, 859–877, doi: 10.1080/01621459.2017.1285773.

Bunks, C., F. M. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473, doi: 10.1190/1.1443880.

Detommaso, G., T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl, 2018, A Stein variational Newton method: Advances in Neural Information Processing Systems, 9169–9179.

Duchi, J., E. Hazan, and Y. Singer, 2011, Adaptive subgradient methods for online learning and stochastic optimization: Journal of Machine Learning Research, **12**, 2121–2159.

Galetti, E., A. Curtis, G. A. Meles, and B. Baptie, 2015, Uncertainty loops in travel-time tomography from nonlinear wave physics: Physical Review Letters, **114**, 148501, doi: 10.1103/PhysRevLett.114.148501.

Gebraad, L., C. Boehm, and A. Fichtner, 2020, Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo: Journal of Geophysical Research, Solid Earth, **125**, e2019JB018428, doi: 10.1029/2019JB018428.

Guo, P., G. Visser, and E. Saygin, 2020, Bayesian trans-dimensional full waveform inversion: Synthetic and field data application: Geophysical Journal International, **222**, 610–627, doi: 10.1093/gji/ggaa201.

Kullback, S., and R. A. Leibler, 1951, On information and sufficiency: The Annals of Mathematical Statistics, **22**, 79–86, doi: 10.1214/aoms/1177729694.

Ledoit, O., and M. Wolf, 2004, A well-conditioned estimator for large-dimensional covariance matrices: Journal of Multivariate Analysis, **88**, 365–411, doi: 10.1016/S0047-259X(03)00096-4.

Liu, Q., and D. Wang, 2016, Stein variational gradient descent: A general purpose Byesian inference algorithm: Advances in Neural Information Processing Systems, 2378–2386.

Martin, G. S., R. Wiley, and K. J. Marfurt, 2006, Marmousi2: An elastic upgrade for Marmousi: The Leading Edge, **25**, 156–166, doi: 10.1190/1.2172306.

Meng, X., Z. Li, D. Zhang, and G. E. Karniadakis, 2020, PPINN: Parareal physics-informed neural network for time-dependent PDEs: Computer Methods in Applied Mechanics and Engineering, **370**, 113250, doi: 10.1016/j.cma.2020.113250.

Nawaz, M. A., and A. Curtis, 2018, Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies: Geophysical Journal International, **214**, 845–875, doi: 10.1093/gji/ggy163.

Plessix, R.-E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503, doi: 10.1111/j.1365-246X.2006.02978.x.

Ray, A., A. Sekar, G. M. Hoversten, and U. Albertin, 2016, Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian transdimensional algorithm: Geophysical Journal International, **205**, 915–937, doi: 10.1093/gji/ggw061.

Robbins, H., and S. Monro, 1951, A stochastic approximation method: The Annals of Mathematical Statistics, **22**, 400–407, doi: 10.1214/aoms/1177729586.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266, doi: 10.1190/1.1441754.

Wang, D., Z. Tang, C. Bajaj, and Q. Liu, 2019, Stein variational gradient descent with matrix-valued kernels: Advances in Neural Information Processing Systems, 7836–7846.

Zhang, X., and A. Curtis, 2020a, Seismic tomography using variational inference methods: Journal of Geophysical Research, Solid Earth, **125**, e2019JB018589, doi: 10.1029/2019JB018589.

Zhang, X., and A. Curtis, 2020b, Variational full-waveform inversion: Geophysical Journal International, **222**, 406–411, doi: 10.1093/gji/ggaa170.

Zhao, Z., and M. K. Sen, 2019, A gradient based MCMC method for FWI and uncertainty analysis: 89th Annual International Meeting, SEG, Expanded Abstracts, 1465–1469, doi: 10.1190/segam2019-3216560.1.

Biographies and photographs of the authors are not available.