AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR Solid Earth

RESEARCH ARTICLE
10.1029/2019JB018589

Key Points:
- We introduce two variational inference methods: automatic differential variational inference and Stein variational gradient descent
- We apply the methods to solve synthetic and real data seismic tomography, producing similar probabilistic results to Monte Carlo methods
- Variational methods are efficient alternatives to Monte Carlo for generally nonlinear Geophysical inverse and inference problems

Correspondence to:
X. Zhang,
x.zhang2@ed.ac.uk

Citation:
Zhang, X., & Curtis, A. (2020). Seismic tomography using variational inference methods. *Journal of Geophysical Research: Solid Earth*, *125*, e2019JB018589. https://doi.org/10.1029/2019JB018589

Received 27 AUG 2019
Accepted 5 NOV 2019
Accepted article online 12 NOV 2019

Seismic Tomography Using Variational Inference Methods

Xin Zhang[1] **and Andrew Curtis**[1,2]

[1]School of Geosciences, University of Edinburgh, Edinburgh, UK, [2]Department of Earth Sciences, ETH Zürich, Zürich, Switzerland

Abstract Seismic tomography is a methodology to image the interior of solid or fluid media and is often used to map properties in the subsurface of the Earth. In order to better interpret the resulting images, it is important to assess imaging uncertainties. Since tomography is significantly nonlinear, Monte Carlo sampling methods are often used for this purpose, but they are generally computationally intractable for large data sets and high-dimensional parameter spaces. To extend uncertainty analysis to larger systems, we use variational inference methods to conduct seismic tomography. In contrast to Monte Carlo sampling, variational methods solve the Bayesian inference problem as an optimization problem yet still provide fully nonlinear, probabilistic results. In this study, we applied two variational methods, automatic differential variational inference and Stein variational gradient descent, to 2-D seismic tomography problems using both synthetic and real data, and we compare the results to those from two different Monte Carlo sampling methods. The results show that automatic differential variational inference provides a biased approximation because of its implicit transformed-Gaussian approximation, and it cannot be used to find generally multimodal posteriors; Stein variational gradient descent produces more accurate approximations to the results of Monte Carlo sampling methods. Both methods estimate the posterior distribution at significantly lower computational cost, provided that gradients of parameters with respect to data can be calculated efficiently. We expect that the methods can be applied fruitfully to many other types of geophysical inverse problems.

1. Introduction

In a variety of geoscientific applications, scientists need to create maps of subsurface properties in order to understand both the heterogeneity and the processes taking place within the Earth. Seismic tomography is a method that is widely used to generate those maps. The maps of interest are usually parameterized in some way, and data are recorded that can be used to constrain the parameters. Tomography is therefore a parameter estimation problem, given the data and a physical relationship between data and parameters; since the physical relationships usually predict data given parameter values but not the reverse, seismic tomography involves solving an inverse problem (Curtis & Snieder, 2002).

Tomographic problems can be solved either using the full, known physical relationships or through a linearized procedure which involves creating approximate, linearized physics that is assumed to be accurate close to a particular chosen reference model. In the linearized procedure one seeks an optimal solution by perturbing the model so as to minimize the misfit between the observed data and the data predicted by the linearized physics. The physics is then relinearized around this new reference model, and the process is iterated until the perturbations are sufficiently small. Since most tomography problems are underdetermined, some form of regularization must be introduced to solve the system (Aki & Lee, 1976; Dziewonski & Woodhouse, 1987; Iyer & Hirahara, 1993; Tarantola, 2005). However, regularization is usually chosen using ad hoc criteria, which introduce poorly understood biases in the results; thus, valuable information can be concealed by regularization (Zhdanov, 2002). Moreover, in nonlinear problems it is almost always impossible to estimate accurate uncertainties in results using linearized methods. Therefore, partially or fully nonlinear tomographic methods have been introduced to geophysics, which require no linearization and which provide accurate estimates of uncertainty using a Bayesian probabilistic formulation of the parameter estimation problem. These include Monte Carlo (MC) methods (Bodin & Sambridge, 2009; Galetti et al., 2015, 2017; Mosegaard & Tarantola, 1995; Malinverno & Leaney, 2000; Malinverno, 2002; Malinverno & Briggs, 2004; Sambridge, 1999; Zhang et al., 2018) and methods based on neural networks (Devilee et al., 1999; Earp & Curtis, 2019; Käufl et al., 2013, 2015, Meier et al., 2007a, 2007b; Röth & Tarantola, 1994; Shahraeeni & Curtis, 2011; Shahraeeni et al., 2012).

Bayesian methods use Bayes' theorem to update a prior probability distribution function (pdf—either a conditional density function or a discrete set of probabilities) with new information from data. The prior pdf describes information available about the parameters of interest prior to the inversion. Bayes' theorem combines the prior pdf with information derived from the current data to produce the total state of information about the parameters post inversion, described by a so-called posterior pdf—this process is referred to as Bayesian inference. Thus, in our case Bayesian inference is used to solve the tomographic inverse problem.

MC methods generate a set (or chain) of samples from the posterior pdf describing the probability distribution of the model given the observed data; thereafter, these samples can be used to estimate useful information about that pdf (mean, standard deviation, etc.). The methods are quite general from a theoretical point of view so that in principle they can be applied to any tomographic problems. They have been extended to transdimensional inversion using the reversible jump Markov chain Monte Carlo (rj-McMC) algorithm (Green, 1995), in which the number of parameters (hence the dimensionality of parameter space) can vary in the inversion. Consequently, the parameterization itself can be simplified by adapting to the data, which can improve results on otherwise high-dimensional problems (Bodin & Sambridge, 2009; Bodin et al., 2012; Burdick & Lekić, 2017; Galetti et al., 2015, 2017; Galetti & Curtis, 2018; Hawkins & Sambridge, 2015; Malinverno & Leaney, 2000; Ray et al., 2013; Piana Agostinetti et al., 2015; Young et al., 2013; Zhang et al., 2018, 2020). Although many tomographic applications have been conducted using McMC sampling methods (previous references, Crowder et al., 2019; Shen et al., 2012, 2013; Zheng et al., 2017; Zulfakriza et al., 2014), they mainly address 1-D or 2-D tomography problems due to the high computational expense of MC methods. Some studies used McMC methods for fully 3-D tomography using body wave travel time data (Hawkins & Sambridge, 2015; Piana Agostinetti et al., 2015; Burdick & Lekić, 2017) and surface wave dispersion (Zhang et al., 2018, 2020), but the methods demand enormous computational resources. Even in the 1-D or 2-D case, McMC methods cannot easily be applied to large data sets, which are generally expensive to forward model given a set of parameter values. Moreover, McMC methods tend to be inefficient at exploring complex, multimodal probability distributions (Karlin, 2014; Sivia, 1996), which appear to be common in seismic tomography problems.

Neural network-based methods offer an efficient alternative for certain classes of tomography problems that will be solved many times with new data of the same type. An initial set of MC samples is taken from the prior probability distribution over parameter space, and data are computationally forward modeled for each parameter vector. Neural networks are flexible mappings that can be regressed (trained) to emulate the mapping from data to parameter space by fitting the set of examples of that mapping generated by MC (Bishop, 2006). Since for each input data vector the neural network only produces one parameter vector, trade-offs between parameters are not clearly represented in the mapping from data to model parameters. Nevertheless, the trained network interpolates the inverse mapping between the examples and can be applied efficiently to any new, measured data to estimate corresponding parameter values. The first geophysical application of neural network tomography was Röth and Tarantola (1994), but that application did not estimate uncertainties. Forms of networks that estimate tomographic uncertainties were introduced to Geophysics by Devilee et al. (1999) and Meier et al. (2007a, 2007b) and have been applied to surface and body wave tomography in 1-D and 2-D problems (Earp & Curtis, 2019; Meier et al., 2007a, 2007b). Unfortunately neural networks still suffer from the computational cost of generating the initial set of training examples. That set may have to include many more samples than are required for standard Bayesian MC, because the training set must span the prior pdf, whereas standard applications of MC tomography sample the posterior pdf which is usually more tightly constrained. Neural networks have the advantage that the training samples need only be calculated once for any number of data sets, whereas MC inversion must perform sampling for every new data set. However, in high-dimensional problems the cost of sampling may be prohibitive for both MC and neural network-based methods due to the curse of dimensionality (the exponential increase in the hypervolume of parameter space as the number of parameters increases; Curtis & Lomax, 2001).

Variational inference provides a different way to solve a Bayesian inference problem: Within a predefined family of probability distributions, one seeks an optimal approximation to a target distribution, which in this case is the Bayesian posterior pdf. This is achieved by minimizing the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951)—one possible measure of the difference between two given pdfs (Blatter et al., 2019), in our case the difference between approximate and target pdfs (Bishop, 2006; Blei et al., 2017). Since the method casts the inference problem into an optimization problem, it can be computationally more efficient than either MC sampling or neural network methods and provides better scaling to higher-dimensional

problems. Moreover, it can be used to take advantage of methods such as stochastic optimization (Kubrusly & Gravier, 1973; Robbins & Monro, 1951) and distributed optimization by dividing large data sets into random minibatches—methods that are difficult to apply for McMC methods since they may break the reversibility property of Markov chains, which is required by most McMC methods.

In variational inference, the complexity of the approximating family of pdfs determines the complexity of the optimization. A complex variational family is generally more difficult to optimize than a simple family. Therefore, many applications are performed using simple mean-field approximation families (Bishop, 2006; Blei et al., 2017) and structured families (Hoffman & Blei, 2015; Saul & Jordan, 1996). For example, in Geophysics the method has been used to invert for the spatial distribution of geological facies given seismic data using a mean-field approximation (Nawaz & Curtis, 2018, 2019).

Even using those simple families, applications of variational inference methods usually involve tedious derivations and bespoke implementations for each type of problem, which restricts their applicability (Bishop, 2006; Blei et al., 2017; Nawaz & Curtis, 2018, 2019). The simplicity of those families also affects the quality of the approximation to complex distributions. To make variational methods easier to use, "black box" variational inference methods have been proposed (Kingma & Welling, 2013; Ranganath et al., 2014, 2016). Based on these ideas, Kucukelbir et al. (2017) proposed an automatic variational inference method, which can easily be applied to many Bayesian inference problems. Another set of methods has been proposed based on probability transformations (Liu & Wang, 2016; Marzouk et al., 2016; Rezende & Mohamed, 2015; Tran et al., 2015); these methods optimize a series of invertible transforms to approximate the target probability and in this case it is possible to approximate arbitrary probability distributions.

We apply automatic differential variational inference (ADVI; Kucukelbir et al., 2017) and Stein variational gradient descent (SVGD; Liu & Wang, 2016) to a 2-D seismic tomography problem. In the following we first describe the basic idea of variational inference and then the ADVI and SVGD methods. In section 3 we apply the two methods to a simple 2-D synthetic seismic tomography example and compare their results with both fixed-dimensional McMC and rj-McMC. In section 4 we apply the two methods to real data from Grane field, North Sea, to study the phase velocity map at 0.9 s and compare the results to those found using rj-McMC. We thus demonstrate that variation inference methods can provide efficient alternatives to McMC methods while still producing reasonably accurate approximations to Bayesian posterior pdfs. Our aim is to introduce variational inference methods to the geoscientific community and to encourage more research on this topic.

2. Methods

2.1. Variational Inference

Bayesian inference involves calculating or characterizing a posterior probability density function $p(\mathbf{m}|\mathbf{d}_{obs})$ of model parameters $\mathbf{m}$ given the observed data $\mathbf{d}_{obs}$. According to Bayes' theorem,

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

where $p(\mathbf{d}_{obs}|\mathbf{m})$ is called the *likelihood* which is the probability of observing data $\mathbf{d}_{obs}$ conditional on model $\mathbf{m}$, $p(\mathbf{m})$ is the prior which describes known information about the model that is independent of the data, and $p(\mathbf{d}_{obs})$ is a normalization factor called the *evidence*, which is constant for a fixed model parameterization. The likelihood is usually assumed to follow a Gaussian probability density function around the data predicted synthetically from model $\mathbf{m}$ (using the known physical relationships), as this is assumed to be a reasonable approximation to the pdf of uncertainties or errors in the measured data, and because noise reduction is performed by stacking, which through the central limit theorem justifies the use of a Gaussian distribution.

Variational inference approximates the above pdf $p(\mathbf{m}|\mathbf{d}_{obs})$ using optimization. First, a family (set) of known distributions $\mathcal{Q} = \{q(\mathbf{m})\}$ is defined. The method then seeks the best approximation to $p(\mathbf{m}|\mathbf{d}_{obs})$ within that family by minimizing the KL-divergence:

$$\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] = \mathrm{E}_q[\log\ q(\mathbf{m})] - \mathrm{E}_q[\log p(\mathbf{m}|\mathbf{d}_{obs})] \tag{2}$$

where the expectation is taken with respect to distribution $q(\mathbf{m})$. It can be shown that $\mathrm{KL}[q||p] \geq 0$ and has zero value if and only if $q(\mathbf{m})$ equals $p(\mathbf{m}|\mathbf{d}_{obs})$ (Kullback & Leibler, 1951). Distribution $q^*(\mathbf{m})$ that minimizes the KL-divergence is therefore the best approximation to $p(\mathbf{m}|\mathbf{d}_{obs})$ within the family $\mathcal{Q}$.

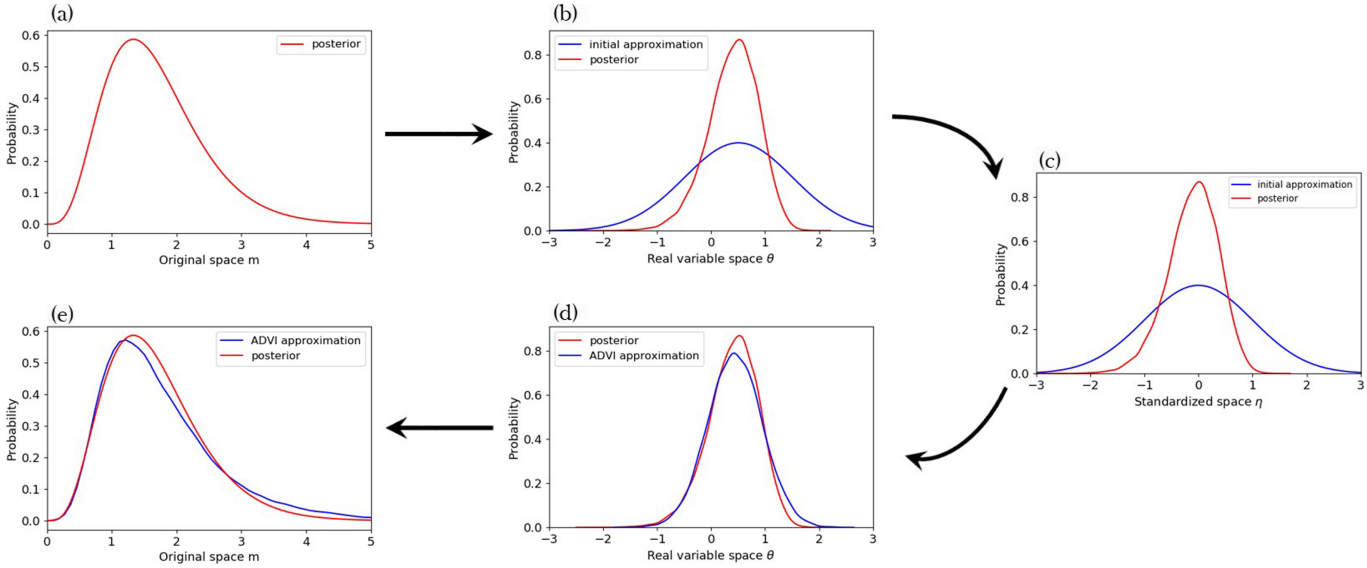

Figure 1. An illustration of the workflow of ADVI. (a) An example of a posterior pdf in the original positive half-space of parameters $\mathbf{m}$. (b) The posterior pdf in the transformed real variable space θ (red) and an initial Gaussian approximation (blue). (c) The posterior pdf (red) and the standard Gaussian distribution (blue) in standardized variable space η; gradients with respect to variational parameters are calculated in this space. (d) and (e) show the posterior pdf (red) and the approximation obtained using ADVI (blue) in the unconstrained real variable space and the original space, respectively.

Combining equations (1) and (2), the KL-divergence becomes

$$\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] = \mathrm{E}_q[\log q(\mathbf{m})] - \mathrm{E}_q[\log p(\mathbf{m}, \mathbf{d}_{obs})] + \log p(\mathbf{d}_{obs}) \tag{3}$$

The evidence term $\log p(\mathbf{d}_{obs})$ generally cannot be calculated since it involves the evaluation of a high-dimensional integral, which takes exponential time. Instead, we calculate the evidence lower bound (ELBO), which is equivalent to the KL-divergence up to an unknown constant and is obtained by rearranging equation (3) and using the fact that $\mathrm{KL}[q||p] \geq 0$:

$$\begin{aligned} \mathrm{ELBO}[q] &= \mathrm{E}_q[\log p(\mathbf{m}, \mathbf{d}_{obs})] - \mathrm{E}_q[\log q(\mathbf{m})] \\ &= \log p(\mathbf{d}_{obs}) - \mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] \end{aligned} \tag{4}$$

Thus, minimizing the KL-divergence is equivalent to maximizing the ELBO.

In variational inference, the choice of the variational family is important because the flexibility of the variational family determines the power of the approximation. However, it is usually more difficult to optimize equation (4) over a complex family than a simple family. Therefore, many applications are performed using the *mean-field* variational family, which means that the parameters $\mathbf{m}$ are treated as being mutually independent (Bishop, 2006; Blei et al., 2017). However, even under that simplifying assumption, traditional variational methods require tedious model-specific derivations and implementations, which restricts their applicability to those problems for which derivations have been performed (e.g., Nawaz & Curtis, 2018, 2019). We therefore introduce two more general variational methods: the ADVI and the SVGD, which can both be applied to general inverse problems.

2.2. ADVI

Kucukelbir et al. (2017) proposed a general variational method called ADVI based on a Gaussian variational family. In ADVI, a model with constrained parameters is first transformed to a model with unconstrained real-valued variables. For example, the velocity model $\mathbf{m}$ that usually has hard bound constraints (such as velocity being greater than 0) can be transformed to an unconstrained model $\boldsymbol{\theta} = T(\mathbf{m})$, where T is an invertible and differentiable function (Figures 1a and 1b). The joint probability $p(\mathbf{m}, \mathbf{d}_{obs})$ then becomes

$$p(\boldsymbol{\theta}, \mathbf{d}_{obs}) = p(\mathbf{m}, \mathbf{d}_{obs})|\det \mathbf{J}_{T^{-1}}(\boldsymbol{\theta})| \tag{5}$$

where $\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})$ is the Jacobian matrix of the inverse of T, which accounts for the volume change of the transform, and $|\cdot|$ represents the absolute value. This transform makes the choice of variational approximations

independent of bounds on the original model since transformed variables lie in the common unconstrained space of real numbers.

In ADVI, we choose a Gaussian variational family (e.g., blue line in Figure 1b):

$$q(\boldsymbol{\theta};\phi) = \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\mu}, \sum\right) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T) \tag{6}$$

where ϕ represents variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix. As in Kucukelbir et al. (2017), for computational purposes we use a Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{L}$ is a lower-triangular matrix, to reparameterize the covariance matrix to ensure that it is positive semidefinite (covariance is positive semidefinite by definition). If $\boldsymbol{\Sigma}$ is a diagonal matrix, q reduces to a mean-field approximation in which the variables are mutually independent; in order to include spatial correlations in the velocity model, we use a full-rank covariance matrix, noting that this incurs a computational cost since it increases the number of variational parameters.

In the transformed space, the variational problem is solved by maximizing the ELBO, written as $\mathcal{L}$, with respect to variational parameters ϕ:

$$\begin{aligned}\phi^* &= \arg\max_{\phi} \mathcal{L}[q(\boldsymbol{\theta};\phi)] \\ &= \arg\max_{\phi} \mathrm{E}_q[\log p(T^{-1}(\boldsymbol{\theta}), \mathbf{d}_{obs}) + \log|det\, \mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|] - \mathrm{E}_q[\log q(\boldsymbol{\theta})]\end{aligned} \tag{7}$$

This is an optimization problem in an unconstrained space and can be solved using gradient ascent methods without worrying about any constrains on the original variables.

However, the gradients of variational parameters are not easy to calculate since the ELBO involves expectations in a high-dimensional space. We therefore transform the Gaussian distribution $q(\boldsymbol{\theta};\phi)$ into a standard Gaussian $\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})$ (Figure 1c), by $\boldsymbol{\eta} = R_\phi(\boldsymbol{\theta}) = \mathbf{L}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})$; thereafter, the variational problem becomes

$$\begin{aligned}\phi^* &= \arg\max_{\phi} \mathcal{L}[q(\boldsymbol{\theta};\phi)] \\ &= \arg\max_{\phi} \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}[\log p(T^{-1}(R_\phi^{-1}(\boldsymbol{\eta})), \mathbf{d}_{obs}) + \log|\det \mathbf{J}_{T^{-1}}(R_\phi^{-1}(\boldsymbol{\eta}))|] - \mathrm{E}_q[\log q(\boldsymbol{\theta})]\end{aligned} \tag{8}$$

where the first expectation is taken with respect to a standard Gaussian distribution $\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})$. There is no Jacobian term related to this transform since the determinant of the Jacobian is equal to 1 (Kucukelbir et al., 2017). The second expectation $-\mathrm{E}_q[\log q(\boldsymbol{\theta})]$ is not transformed since it has a simple analytic form as does its gradient (Kucukelbir et al., 2017)—see Appendix A.

Since the distribution with respect to which the expectation is taken now does not depend on variational parameters, the gradient with respect to variational parameters can be calculated by exchanging the expectation and derivative according to the dominated convergence theorem (DCT; Çınlar, 2011) and by applying the chain rule—see Appendix B:

$$\nabla_{\boldsymbol{\mu}}\mathcal{L} = \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}[\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}} T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log|\det \mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|] \tag{9}$$

The gradient with respect to $\mathbf{L}$ can be obtained similarly:

$$\nabla_{\mathbf{L}}\mathcal{L} = \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}[(\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}} T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log|\det \mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|)\boldsymbol{\eta}^T] + (\mathbf{L}^{-1})^T \tag{10}$$

where the expectation is computed with respect to a standard Gaussian distribution, which can be estimated by MC integration. MC integration provides a noisy, unbiased estimation of the expectation and its accuracy increases with the number of samples. Nevertheless, it has been shown that in practice a low number or even a single sample can be sufficient at each iteration since the mean is taken with respect to the standard Gaussian distribution (see discussions and experiments in Kucukelbir et al., 2017). For distributions $p(\mathbf{m}, \mathbf{d}_{obs})$ for which the gradients have analytic forms, the whole process of computing gradients can be automated (Kucukelbir et al., 2017), hence the name "automatic differential". We can then use a gradient ascent method to update the variational parameters and obtain an approximation to the pdf $p(\mathbf{m}|\mathbf{d}_{obs})$ (e.g., Figure 1d).

Note that although the method is based on Gaussian variational approximations, the actual shape of the approximation to the posterior $p(\mathbf{m}|\mathbf{d}_{obs})$ over the original parameters $\mathbf{m}$ is determined by the transform

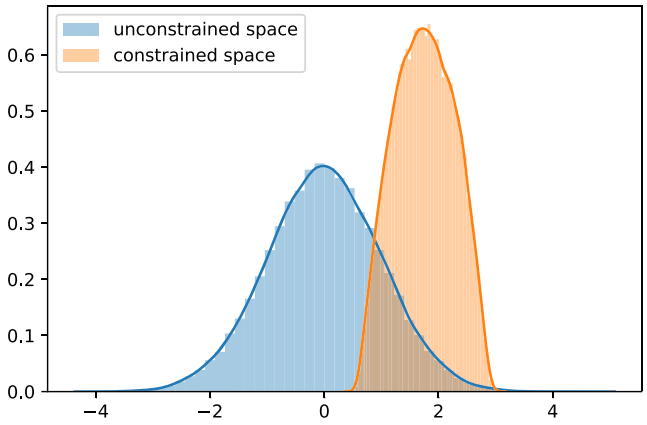

Figure 2. An illustration of the transform in equation (11). The original variable is in a constrained space between 0.5 and 3.0. The blue area shows a standard Gaussian distribution in the transformed unconstrained space, and the orange area shows the associated probability distribution in the original space. The probability distributions are estimated using Monte Carlo samples. The orange curve is the distribution fitted using Gaussian kernels.

T (Figure 1e). It is difficult to determine an optimal transform since that is related to the properties of the unknown posterior (Kucukelbir et al., 2017). In this study we use a commonly used invertible logarithmic transform (Team, 2016):

$$\begin{aligned} \theta_i &= T(m_i) = \log(m_i - a_i) - \log(b_i - m_i) \\ m_i &= T^{-1}(\theta_i) = a_i + \frac{(b_i - a_i)}{1 + exp(-\theta_i)} \end{aligned} \tag{11}$$

where m_i represents each original constrained parameter, θ_i is the transformed unconstrained variable, a_i is the original lower bound, and b_i the upper bound on m_i. Therefore, the quality of the ADVI approximation is limited by the Gaussian approximation in the unconstrained space and by the specific transform T in equation (11).

To illustrate the effects of the transform in equation (11), we show an example in Figure 2. The original variable lies in a constrained space between 0.5 and 3.0 (a typical phase velocity range of seismic surface waves). The space is transformed to an unconstrained space using equation (11). If, as in ADVI, we assume a standard Gaussian distribution in the transformed space (blue area in Figure 2), the associated probability distribution in the original space is shown in orange in Figure 2. The actual shape of the distribution in the original space is not Gaussian but is determined by the transform T in equation (11). However, under this choice of T it is likely that the probability distribution in the original space is still unimodal. We thus see that ADVI provides a unimodal approximation of the target posterior pdf around a local optimal parameter estimate. This suggests that the method will not be effective for multimodal distributions, and the estimated probability distribution depends on the initial value of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Kucukelbir et al., 2017). However, since the maximum a posteriori probability (MAP) estimate has been shown to be effective for parameter estimation in practice, the ADVI method could still be used to provide a good approximation of the distribution around a MAP estimate.

2.3. SVGD

In practice, most applications of variational inference use simple families of posterior approximations such as a Gaussian approximation (Kucukelbir et al., 2017), mean-field approximations (Blei et al., 2017; Nawaz & Curtis, 2018, 2019), or other simple structured families (Hoffman & Blei, 2015; Saul & Jordan, 1996). These simple choices significantly restrict the quality of derived posterior approximations. In order to employ a broader family of variational approximations, variational methods based on invertible transforms have been proposed (Marzouk et al., 2016; Rezende & Mohamed, 2015; Tran et al., 2015). In these methods instead of choosing specific forms for variational approximations, a series of invertible transforms are applied to an initial distribution, and these transforms are optimized by minimizing the KL divergence. This provides a way to approximate arbitrary posterior distributions since a pdf can be transformed to any other pdf as long as the probability measures are absolutely continuous.

SVGD is one such algorithm based on an incremental transform (Liu & Wang, 2016). In SVGD, a smooth transform $T(\mathbf{m}) = \mathbf{m} + \epsilon\boldsymbol{\phi}(\mathbf{m})$ is used, where $\mathbf{m} = [m_1, \ldots, m_d]$ and m_i is the ith parameter, and $\boldsymbol{\phi}(\mathbf{m}) = [\phi_1, \ldots, \phi_d]$ is a smooth vector function that describes the perturbation direction and where ϵ is the magnitude of the perturbation. It can be shown that when ϵ is sufficiently small, the transform is invertible since the Jacobian of the transform is close to an identity matrix (Liu & Wang, 2016). Say $q_T(\mathbf{m})$ is the transformed probability distribution of the initial distribution $q(\mathbf{m})$. Then the gradient of KL-divergence with respect to ϵ can be computed as (see Appendix C):

$$\nabla_{\boldsymbol{\epsilon}} \mathrm{KL}[q_T||p] \mid_{\epsilon=0} = -\mathrm{E}_q \left[trace \left(\mathcal{A}_p \, \boldsymbol{\phi}(\mathbf{m}) \right) \right] \tag{12}$$

where $\mathcal{A}_p$ is the Stein operator such that $\mathcal{A}_p \, \boldsymbol{\phi}(\mathbf{m}) = \nabla_{\mathbf{m}} \log p(\mathbf{m}) \, \boldsymbol{\phi}(\mathbf{m})^T + \nabla_{\mathbf{m}} \boldsymbol{\phi}(\mathbf{m})$. This suggests that maximizing the right-hand expectation with respect to $q(\mathbf{m})$ gives the steepest descent of the KL divergence, and consequently, the KL divergence can be minimized iteratively.

It can be shown that the negative gradient of the KL divergence in equation (12) can be maximized by using the kernelized Stein discrepancy (Liu et al., 2016). For two continuous probability densities p and q, the *Stein*

discrepancy for a function ϕ in a function set $\mathcal{F}$ is defined as follows:

$$S[q,p] = \arg\max_{\boldsymbol{\phi}\in\mathcal{F}} \left\{ \left(\mathrm{E}_q\left[trace\left(\mathcal{A}_p\,\boldsymbol{\phi}(\mathbf{m})\right)\right]\right)^2 \right\} \tag{13}$$

The Stein discrepancy provides another way to quantify the difference between two distribution densities (Gorham & Mackey, 2015; Stein, 1972). However, the Stein discrepancy is not easy to compute for general $\mathcal{F}$. Therefore, Liu et al. (2016) proposed a kernelized Stein discrepancy by maximizing equation (13) in the unit ball of a reproducing kernel Hilbert space (RKHS) as follows.

A Hilbert space is a space $\mathcal{H}$ on which an inner product $<,>_{\mathcal{H}}$ is defined. A function is called a *kernel* if there exists a real Hilbert space and a function φ such that $k(x,y) = <\varphi(x),\varphi(y)>_{\mathscr{H}}$ (Gretton, 2013). A kernel is said to be positive definite if the matrix defined by $K_{ij} = k(x_i,x_j)$ is positive definite. Assuming a positive definite kernel $k(\mathbf{m},\mathbf{m}')$ on $\mathcal{M}\times\mathcal{M}$, its reproducing kernel Hilbert space $\mathcal{H}$ is defined by the closure of the linear span $\{f : f(\mathbf{m}) = \sum_{i=1}^{n} a_i k(\mathbf{m},\mathbf{m}^i), a_i \in \mathcal{R}, n \in \mathcal{N}, \mathbf{m}^i \in \mathcal{M}\}$ with inner products $\langle f,g\rangle_{\mathcal{H}} = \sum_{ij} a_i b_j k(\mathbf{m}^i,\mathbf{m}^j)$ for $g(\mathbf{m}) = \sum_i b_i k(\mathbf{m},\mathbf{m}^i)$. The RKHS has an important reproducing property, that is, $f(x) = \langle f(x'),k(x',x)\rangle_{\mathcal{H}}$, such that the evaluation of a function f at x can be represented as an inner product in the Hilbert space. In a RKHS, the kernelized Stein discrepancy can be defined as (Liu et al., 2016)

$$S[q,p] = \arg\max_{\boldsymbol{\phi}\in\mathcal{H}^d} \left\{ \left(\mathrm{E}_q\left[trace\left(\mathcal{A}_p\,\boldsymbol{\phi}(\mathbf{m})\right)\right]\right)^2, \quad s.t. \quad ||\,\boldsymbol{\phi}\,||_{\mathcal{H}^d} \le 1 \right\} \tag{14}$$

where $\mathcal{H}^d$ is the RKHS of d-dimensional vector functions. The right side of equation (14) is found to be equal to

$$\boldsymbol{\phi}^* = \boldsymbol{\phi}^*_{q,p}(\mathbf{m}) / ||\,\boldsymbol{\phi}^*_{q,p}(\mathbf{m})||_{\mathcal{H}^d} \tag{15}$$

where

$$\boldsymbol{\phi}^*_{q,p}(\mathbf{m}) = \mathrm{E}_{\{\mathbf{m}'\sim q\}}\left[\mathcal{A}_p k(\mathbf{m}',\mathbf{m})\right] \tag{16}$$

and for which we have $S[q,p] = ||\,\boldsymbol{\phi}^*_{q,p}(\mathbf{m})||_{\mathcal{H}^d}{}^2$. Thus, the optimal $\boldsymbol{\phi}$ in equation (12) is $\boldsymbol{\phi}^*$ and $\nabla_\epsilon \mathrm{KL}[q_T||p]\,|_{\epsilon=0} = -\sqrt{S[q,p]}$.

Given the above solution, the SVGD works as follows: We start from an initial distribution q_0 then apply the transform $T_0^*(\mathbf{m}) = \mathbf{m} + \epsilon\,\boldsymbol{\phi}^*_{q_0,p}(\mathbf{m})$ where we absorb the normalization term in equation (15) into ϵ; this updates q_0 to $q_{[T_0]}$ with a decrease in the KL divergence of $\epsilon * \sqrt{S[q,p]}$. This process is iterated to obtain an approximation of the posterior p:

$$q_{l+1} = q_{l\left[T_l^*\right]}, \quad \text{where} \quad T_l^*(\mathbf{m}) = \mathbf{m} + \epsilon_l\,\boldsymbol{\phi}^*_{q_l,p}(\mathbf{m}) \tag{17}$$

and for sufficiently small $\{\epsilon_l\}$ the process eventually converges to the posterior pdf p. Note that a large stepsize may lead the Jacobian matrix of transform T to be singular, which in turn makes the approximation probability fail to converge to the true posterior (Liu, 2017).

To calculate the expectation in equation (16), we start from a set of particles (models) generated using q_0, and at each step the $\boldsymbol{\phi}^*_{q,p}(\mathbf{m})$ can be estimated by computing the mean in equation (16) using those particles. Each particle is then updated using the transform in equation (17), and the resulting particles will form better approximations to the posterior as the iteration proceeds. This suggests the following algorithm, which is schematically represented in Figure 3:

1. Draw a set of particles $\{\mathbf{m}_i^0\}_{i=1}^n$ from an initial pdf estimate (e.g., the prior).
2. At iteration l, update each particle using

$$\mathbf{m}_i^{l+1} = \mathbf{m}_i^l + \epsilon_l\,\boldsymbol{\phi}^*_{q_l,p}(\mathbf{m}_i^l) \tag{18}$$

where

$$\boldsymbol{\phi}^*_{q_l,p}(\mathbf{m}) = \frac{1}{n}\sum_{j=1}^{n}\left[k(\mathbf{m}_j^l,\mathbf{m})\nabla_{\mathbf{m}_j^l}\log p(\mathbf{m}_j^l) + \nabla_{\mathbf{m}_j^l} k(\mathbf{m}_j^l,\mathbf{m})\right] \tag{19}$$

and ϵ_l is the step size at iteration l.

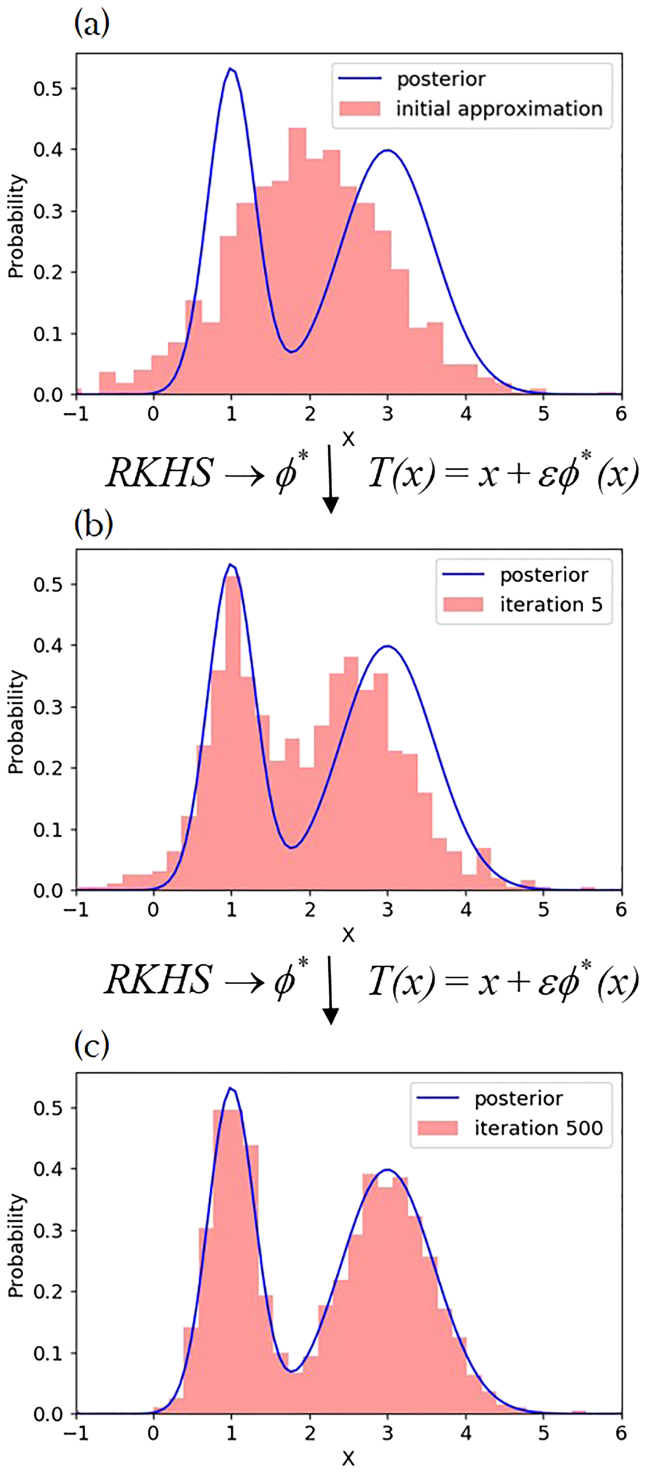

Figure 3. An illustration of the SVGD algorithm. The initial pdf is represented by the density of a set of particles (red histogram) in the top plot. The particles are then updated using a smooth transform $T(x) = x + \epsilon\phi^*(x)$, where ϕ^* is found in a reproducing kernel Hilbert space (RKHS). (a) An example of a posterior pdf (blue line) and an initial distribution (red histogram). (b) The approximating probability distribution after five iterations. (c) The approximating probability distribution after 500 iterations.

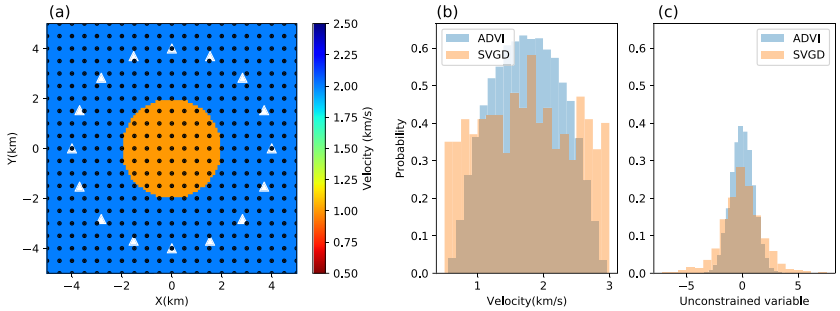

Figure 4. (a) The true velocity model and receivers (white triangles) used in the synthetic test. Sources are at the same locations as receivers to simulate a typical ambient noise interferometry experiment. Black dots indicate the locations of grid points used in the inversions. The histograms show the initial distribution of each parameter in the (b) original space (velocity) and (c) transformed unconstrained space for ADVI (blue) and SVGD (orange). In ADVI, the initial distribution is a standard Gaussian in unconstrained space. For simplicity we generated 5,000 samples from the standard Gaussian and transformed to the original space to show the initial distribution in the original space. In SVGD the initial distribution is approximated using 800 particles generated from a Uniform distribution in the original space and transformed to the unconstrained space.

3. Calculate the density of the final set of particles $\{\mathbf{m}_i^*\}_{i=1}^n$, which approximates the posterior probability density function.

For kernel $k(\mathbf{m}, \mathbf{m}')$ we use the radial basis function $k(\mathbf{m}, \mathbf{m}') = \exp(-\frac{1}{h}||\mathbf{m} - \mathbf{m}'||^2)$, where h can take any positive value. Here h is taken to be $\tilde{d}^2/\log n$ where $\tilde{d}$ is the median of pairwise distances between all particles. This choice of h is based on the intuition that $\sum_j k(\mathbf{m}_i, \mathbf{m}_j) \approx n\exp(-\frac{1}{h}\tilde{d}^2) = 1$, so that for particle $\mathbf{m}_i$ the contribution from its own gradient and the influence from the other particles in equation (19) are balanced (Liu & Wang, 2016). For the radial basis function kernel the second term in equation (19) becomes $\sum_j \frac{2}{h}(\mathbf{m} - \mathbf{m}_j)k(\mathbf{m}_j, \mathbf{m})$, which drives the particle $\mathbf{m}$ away from neighboring particles for which the kernel takes large values. Therefore, the second term in equation (19) acts as a *repulsive force* preventing particles from collapsing to a single mode, while the first term moves particles toward local high probability areas using the kernel-weighted gradient. If in the kernel $h \to 0$, the algorithm falls into independent gradient ascent which maximizes log$p$ for each particle.

Note that since SVGD uses kernelized Stein discrepancy, the choice of kernels may affect the efficiency of the algorithm. In this study we adopted a commonly used kernel: a radial basis function. However, in some cases other kernels may provide a more efficient algorithm, for example, an inverse multiquadric kernel (Gorham & Mackey, 2017), a Hessian kernel (Detommaso et al., 2018), and kernels on a Riemann manifold (Liu & Zhu, 2018).

In SVGD, the accuracy of the approximation increases with the number of particles. It has been shown that compared to other particle-based methods, for example, sequential MC methods (Smith, 2013), SVGD requires fewer samples to achieve the same accuracy, which makes it a more efficient method (Liu & Wang, 2016). In contrast to sequential MC, which is a stochastic process, SVGD acts as a deterministic sampling method. If only one particle is used, the second term in equation (19) becomes 0 and the method reduces to a typical gradient ascent toward the model with the maximum a posteriori (MAP) pdf value. This suggests that even for a small number of particles the method could still produce a good parameter estimate since MAP estimation can be an effective method in practice. Thus, in practice, one could start from a small number of particles and gradually increase the number to find an optimal choice.

In seismic tomography velocities are usually constrained to lie within a given velocity range. In order to ensure that velocities always lie within the constraints, we first apply the same transform used in ADVI (equation (11)) so that the parameters are in an unconstrained space. We can then simply use equation (18) to update particles without explicitly considering the constraints on seismic velocities. The final seismic velocities can be obtained by transforming particles back to the constrained space.

3. Synthetic Tests

We first apply the above methods to a simple 2-D synthetic example similar to that in Galetti et al. (2015) and Zhang et al. (2018). The true model is a homogeneous background with velocity 2 km/s containing a circular

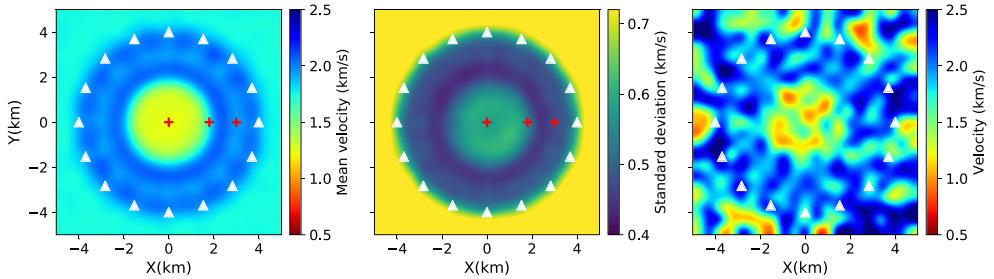

Figure 5. The mean (left), standard deviation (middle), and an individual realization from the approximate posterior distribution (right) obtained using ADVI. The red pluses show locations which are referred to in the main text.

low velocity anomaly with a radius of 2 km with velocity 1 km/s. The 16 receivers are evenly distributed around the anomaly approximating a circular acquisition geometry with radius 4 km (Figure 4a). Each receiver is also treated as a source to simulate a typical ambient noise interferometry experiment (Campillo & Paul, 2003; Curtis et al., 2006; Galetti et al., 2015). This produces a total of 120 interreceiver travel time data, each of which is computed using a fast marching method of solving the Eikonal equation over a 100×100 gridded discretization in space (Rawlinson & Sambridge, 2004).

For variational inversions we use a fixed 21×21 grid of cells to parameterize the velocity model $\mathbf{m}$ (Figure 4a). The noise level is fixed to be 0.05 s (<5% of travel times) for all inversions. The prior pdf of the velocity in each cell is set to be a Uniform distribution between 0.5 and 3.0 km/s to encompass the true model. Travel times are calculated using the same fast marching method as above over a 100×100 grid but using the lower spatial resolution of model properties parameterized in $\mathbf{m}$. The gradients for velocity models are calculated by tracing rays backward from each receiver to each (virtual) source using the gradient of the travel time field for each receiver pair (Rawlinson & Sambridge, 2004). For ADVI, the initial mean of the Gaussian distribution in the transformed space is chosen to be the value, which is the transform of the mean value of the prior in the original space; the initial covariance matrix is simply set to be an identity matrix, which turns out to give a standard Gaussian in our case (see blue histogram in Figure 4c). The shape of the initial distribution in the original space is shown in Figure 4b (blue histogram). We then used 10,000 iterations to update the variational parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$). In order to visualize the results, we generated 5,000 models from the final approximate posterior probability density in the original space and computed their mean and standard deviation. For SVGD, we used 800 particles generated from the prior pdf (orange histogram in Figure 4b) and transformed to an unconstrained space using equation 11 (orange histogram in Figure 4c). Each particle is then updated using equation (17) for 500 iterations, then transformed back to seismic velocity. The mean and standard deviation are then calculated using the values of those particles.

To demonstrate the variational methods, we compare the results with the fixed-dimensional Metropolis-Hastings (MH) McMC method (Hastings, 1970; Malinverno & Leaney, 2000; Metropolis & Ulam, 1949; Mosegaard & Tarantola, 1995) and the rj-McMC method (Bodin & Sambridge, 2009; Green, 1995; Galetti et al., 2015; Zhang et al., 2018). For MH-McMC inversion we used the same parameterization as for the variational methods (a 21×21 grid). A Gaussian perturbation is used as the proposal distribution to generate potential McMC samples, for which the step length is chosen by trial and error to give an acceptance ratio between 20% and 50%. We used a total of six chains, each of which used 2,000,000 iterations with a burn-in period of 1,000,000 iterations. To reduce the correlation between samples, we only retain every fiftieth sample in each chain after the burn-in period. The mean and standard deviation are then calculated using those samples. For rj-McMC inversion we use Voronoi cells to parameterize the model (Bodin & Sambridge, 2009), for which the prior pdf of the number of cells is set to be a Uniform distribution between 4 and 100. The proposal distribution for fixed-dimensional steps (changing the velocity of a cell or moving a cell) is chosen in a similar way as in MH-McMC. For transdimensional steps (adding or deleting a cell) the proposal distribution is chosen as the prior pdf (Zhang et al., 2018). We used a total of six chains, each of which contained 500,000 iterations with a burn-in period of 300,000. Similarly to the fixed-dimensional inversion the chain was thinned by a factor of 50 post burn-in.