Seismic Tomography Using Variational Inference Methods

Citation for published version:
Zhang, X & Curtis, A 2020, 'Seismic Tomography Using Variational Inference Methods', *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 4. https://doi.org/10.1029/2019JB018589

Digital Object Identifier (DOI):
10.1029/2019JB018589

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Geophysical Research: Solid Earth

Seismic tomography using variational inference methods

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

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

Key Points:

- We introduce two variational inference methods: automatic differential variational inference and Stein variational gradient descent.
- We applied 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 non-linear Geophysical inverse and inference problems.

Corresponding author: Xin Zhang, `x.zhang2@ed.ac.uk`

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 datasets 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 probabilistic results. In this study, we applied two variational methods, automatic differential variational inference (ADVI) and Stein variational gradient descent (SVGD), to 2D 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 ADVI provides a biased approximation because of its implicit Gaussian approximation, and cannot be used to find multi-modal posteriors; SVGD can produce 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 obtain maps of subsurface properties in order to understand heterogeneity and 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 parameterised 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 using either the full, known physical relationships, or by using a linearised procedure which involves creating approximate, linearised physics that is assumed to be accurate close to a particular chosen reference model. In the linearised 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 linearised physics. The physics is then re-linearised around this new reference model, and the process is iterated until the preturbations are sufficiently small. Since most tomography problems are under-determined, 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 introduces 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 linearised methods. Therefore, partially or fully nonlinear tomographic methods have been introduced to geophysics which require no linearisation and which provide accurate estimates of uncertainty using a Bayesian probabilistic formulation of the parameter estimation problem. These include Monte Carlo methods (Mosegaard & Tarantola, 1995; Sambridge, 1999; Malinverno et al., 2000; Malinverno, 2002; Malinverno & Briggs, 2004; Bodin & Sambridge, 2009; Galetti et al., 2015, 2017; Zhang et al., 2018) and methods based on neural networks (Röth & Tarantola, 1994; Devilee et al., 1999; Meier et al., 2007b, 2007a; Shahraeeni & Curtis, 2011; Shahraeeni et al., 2012; Käufl et al., 2013, 2015; Earp & Curtis, 2019).

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 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.

Monte Carlo 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 trans-dimensional 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 improves results on otherwise high-dimensional problems (Malinverno et al., 2000; Bodin & Sambridge, 2009; Bodin et al., 2012; Ray et al., 2013; Young et al., 2013; Galetti et al., 2015, 2017; Hawkins & Sambridge, 2015; Piana Agostinetti et al., 2015; Burdick & Lekić, 2017; Galetti & Curtis, 2018; Zhang et al., 2018, 2019). Although many applications have been conducted using McMC sampling methods (previous references, Shen et al., 2012, 2013; Zulfakriza et al., 2014; Zheng et al., 2017; Crowder et al., 2019), they mainly address 1D or 2D tomography problems due to the high computational expense of Monte Carlo methods. Some studies used McMC methods for fully 3D 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, 2019), but the methods demand enormous computational resources. Even in the 1D or 2D case, McMC methods cannot easily be applied to large datasets which are generally expensive to forward model given a set of parameter values. Moreover, McMC methods tend to be inefficient at exploring complex, multi-modal probability distributions (Sivia, 1996; Karlin, 2014), 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 Monte Carlo samples is taken from the prior probability distribution over parameter space, and data are computationally forward modelled 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 using Monte Carlo (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. The trained network then 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 by Devilee et al. (1999) and Meier et al. (2007b, 2007a) and have been applied to surface and body wave tomography in 1D and 2D problems (Meier et al., 2007b, 2007a; Earp & Curtis, 2019). Nevertheless, 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 NN 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 (Robbins & Monro, 1951; Kubrusly & Gravier, 1973) and distributed optimization by dividing large datasets into random minibatches – methods which 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 (Saul & Jordan, 1996; Hoffman & Blei, 2015). 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 (M. A. Nawaz & Curtis, 2018; M. Nawaz & Curtis, 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; M. A. Nawaz & Curtis, 2018; M. Nawaz & Curtis, 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 (Rezende & Mohamed, 2015; Tran et al., 2015; Q. Liu & Wang, 2016; Marzouk et al., 2016); these methods optimise 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 – Q. Liu & Wang, 2016) to a 2D 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 2D 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 characterising 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 $\mathcal{Q}$.

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

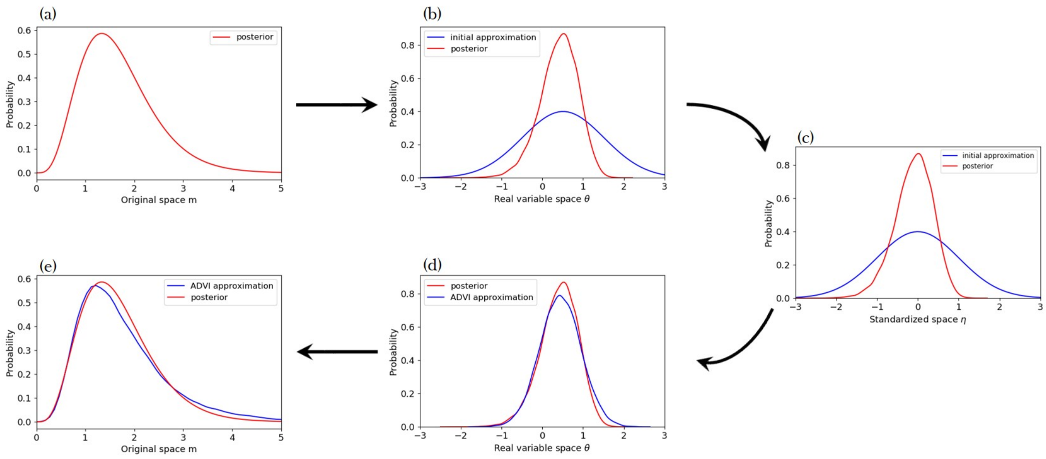

Figure 1. An illustration of the workflow of ADVI. **(a)** An example of a posterior pdf in the original positive half space of parameters $\mathbf{m}$. **(b)** The posterior pdf in the transformed real variable space $\boldsymbol{\theta}$ (red) and an initial Gaussian approximation (blue). **(c)** The posterior pdf (red) and the standard Gaussian distribution (blue) in standardized variable $\boldsymbol{\eta}$; 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.

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., M. A. Nawaz & Curtis, 2018; M. Nawaz & Curtis, 2019). We therefore introduce two more general variational methods: the automatic differential variational inference (ADVI) and the Stein variational gradient descent (SVGD), which can both be applied to general inverse problems.

2.2 Automatic differential variational inference (ADVI)

Kucukelbir et al. (2017) proposed a general variational method called automatic differential variational inference (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 zero) can be transformed to an unconstrained model $\boldsymbol{\theta} = T(\mathbf{m})$, where T is an invertible and differentiable function (Figure 1a and b). 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 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),

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

where $\boldsymbol{\phi}$ 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 re-parameterize 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 $\boldsymbol{\phi}$:

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

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 $q(\boldsymbol{\theta}; \boldsymbol{\phi})$ into a standard Gaussian $\mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})$ (Figure 1c), by $\boldsymbol{\eta} =$

$R_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathbf{L}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})$, thereafter the variational problem becomes:

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

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 one (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 (Çınlar, 2011) and by applying the chain rule – see Appendix B:

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

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

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

where the expectation is computed with respect to a standard Gaussian distribution, which can be estimated by Monte Carlo (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 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 et al., 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 μ and Σ (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 Stein variational gradient descent (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; M. A. Nawaz & Curtis, 2018; M. Nawaz & Curtis, 2019) or other simple structured families (Saul & Jordan, 1996; Hoffman & Blei, 2015). 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 (Rezende & Mohamed, 2015; Tran

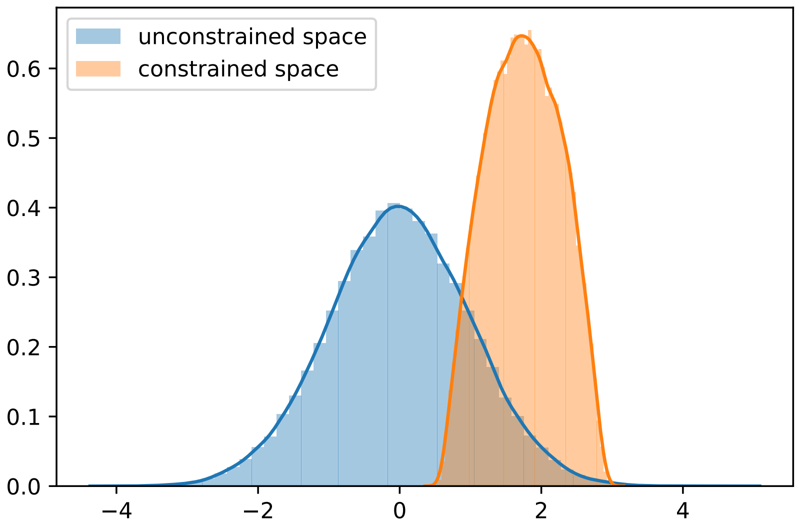

Figure 2. An illustration of the transform in equation (11). The original variable is in a constrained space between 0.5 and 3.0. The blue area shows a standard Gaussian distribution in the transformed unconstrained space and the orange area shows the associated probability distribution in the original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

et al., 2015; Marzouk et al., 2016). 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.

Stein variational gradient descent (SVGD) is one such algorithm based on an incremental transform (Q. Liu & Wang, 2016). In SVGD, a smooth transform $T(\mathbf{m}) = \mathbf{m}+\epsilon\boldsymbol{\phi}(\mathbf{m})$ is used, where $\mathbf{m} = [m_1, ..., m_d]$ and m_i is the i^{th} parameter, and $\boldsymbol{\phi}(\mathbf{m}) = [\phi_1, ..., \phi_d]$ 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 (Q. 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):

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

where $\mathcal{A}_p$ is the Stein operator such that $\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_{\mathbf{m}}\mathrm{log}p(\mathbf{m})\boldsymbol{\phi}(\mathbf{m})^T + \nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})$. 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 (Q. Liu et al., 2016). For two continuous probability densities p and q, the *Stein discrepancy* for a function $\boldsymbol{\phi}$ in a function set $\mathcal{F}$ is defined as:

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

The Stein discrepancy provides another way to quantify the difference between two distribution densities (Stein et al., 1972; Gorham & Mackey, 2015). However the Stein discrepancy is not easy to compute for general $\mathcal{F}$. Therefore, Q. 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 $k(x,y) =<$

$\varphi(x), \varphi(y) >_{\mathcal{H}}$ (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, $f(x) =< f(x'), k(x', x) >_{\mathcal{H}}$, 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 (Q. Liu et al., 2016)

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

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

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

and for which we have $S[q, p] = ||\boldsymbol{\phi}^*_{q,p}(\mathbf{m})||_{\mathcal{H}^d}$. 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:

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

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 (Q. 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 those particles will form better approximations to the posterior as the iteration proceeds. This suggests the following algorithm which is schematically represented in Figure 3:

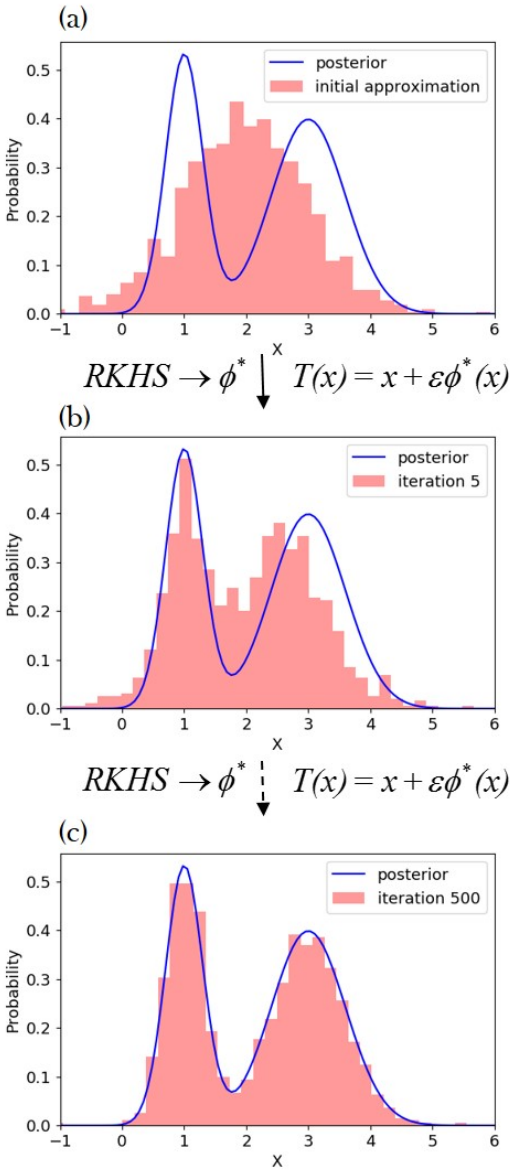

Figure 3. An illustration of the SVGD algorithm. The initial pdf is represented by the density of a set of particles (red histogram) in the top plot. The particles are then updated using a smooth transform $T(x) = x + \epsilon\phi^*(x)$, where ϕ^* is found in a reproducing kernel Hilbert space (RKHS). **(a)** An example of a posterior pdf (blue line) and an initial distribution (red histogram). **(b)** The approximating probability distribution after 5 iterations. **(c)** The approximating probability distribution after 500 iterations.

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

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

   and ϵ_l is the step size at iteration l.
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 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 two gradient terms in equation (19) are balanced (Q. 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 neighbouring 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 towards local high probability areas using the kernel-weighted gradient. If in the kernel $h \to 0$, the algorithm falls into independent gradient ascent that 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 (C. 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, e.g., sequential Monte Carlo methods (Smith, 2013), SVGD requires fewer samples to achieve the same accuracy which makes it a more efficient method (Q. Liu & Wang, 2016). In contrast to sequential Monte Carlo 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 zero and the method reduces to a typical gradient ascent towards the model with the maximum a posterior (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 constrains, 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 constrains 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 2D synthetic example similar to that in Galetti et al. (2015). The true model is a homogeneous background with velocity 2 km/s containing a circular 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 inter-receiver travel time data, each of which is computed using a fast marching method of solving the Eikonal equation over a 100×100 gridded discretisation 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 percent 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 km/s 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 backwards from receiver to (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 trans-

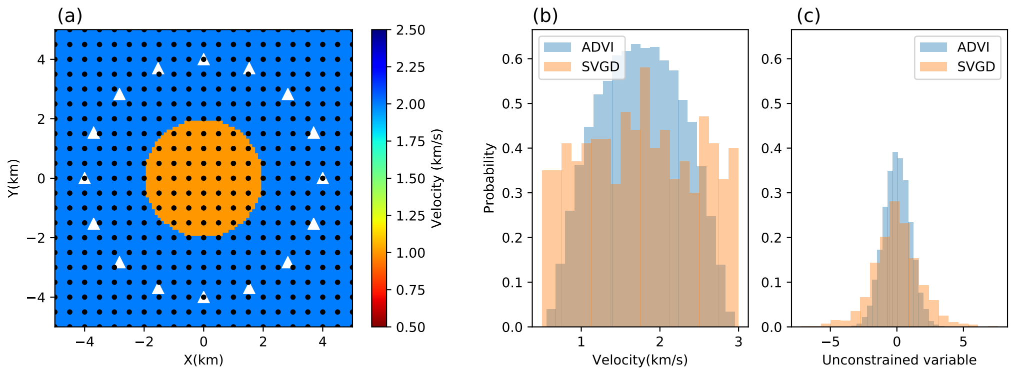

Figure 4. **(a)** The true velocity model and receivers (white triangle) used in the synthetic test. Sources are at the same locations as receivers to simulate a typical ambient noise experiment. Black dots indicate the locations of grid points used in the inversions. The histograms show the initial distributions of a 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 5000 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.

formed 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.

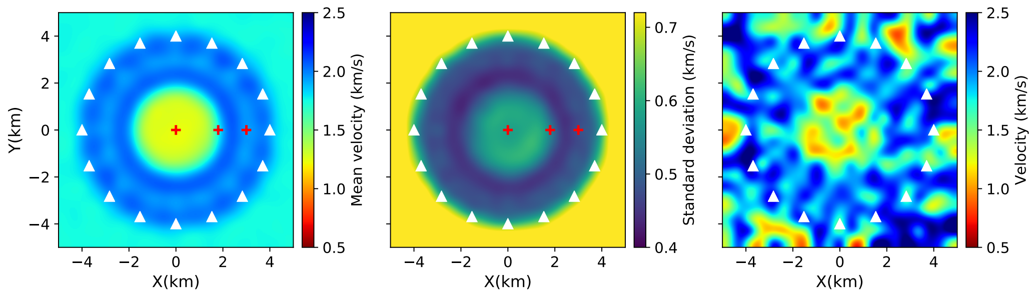

Figure 5. The mean (left), standard deviation (middle) and an individual realization from the approximate posterior distribution (right) obtained using ADVI. The red pluses show locations which are referred to in the main text.

To demonstrate the variational methods we compare the results with the fixed-dimensional Metropolis-Hastings McMC (MH-McMC) method (Metropolis & Ulam, 1949; Hastings, 1970; Mosegaard & Tarantola, 1995; Malinverno et al., 2000) and the rj-McMC method (Green, 1995; Bodin & Sambridge, 2009; 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 used 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 percent. We used a total of 6 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 50^{th} 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 trans-dimensional steps (adding or deleting a cell) the proposal distribution is chosen as the prior pdf (Zhang et al., 2018). We used a total of 6 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.

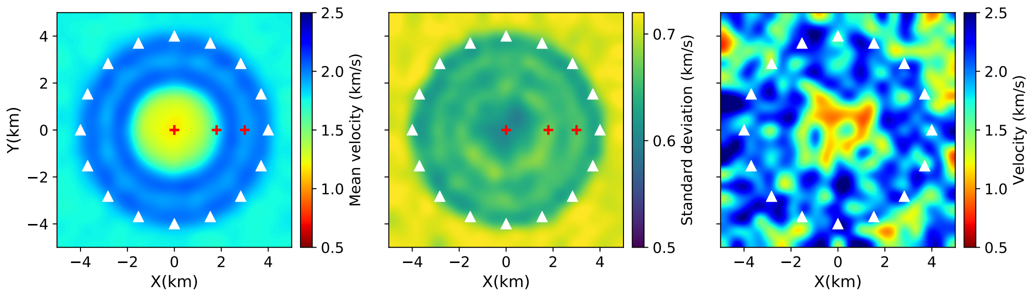

Figure 6. The mean (left), standard deviation (middle) and an individual realization from the approximate posterior distribution (right) obtained using SVGD. The red pluses show locations which are referred to in the main text.

3.1 Results

Figure 5 shows the mean, standard deviation and an individual realization from the approximate posterior distribution calculated using ADVI. The mean model successfully recovers the low velocity anomaly within the receiver array except that the velocity value is slightly higher ($\sim 1.2\,km/s$) than the true value ($1.0\,km/s$). Between the location of the central anomaly and that of the receiver array there is a slightly lower velocity loop. The standard deviation map shows standard deviations similar to that of the prior (0.72 km/s) outside of the array, and clearly higher uncertainties at the location of the central anomaly. The standard deviations around the central anomaly are slightly higher than those at the center. Figure 6 shows the results from SVGD. Similarly, the velocity of the low velocity anomaly ($\sim 1.2\,km/s$) is slightly higher than the true value and a slightly lower velocity loop is also observed between the central anomaly and the receiver array. There is a clear higher uncertainty loop around the central anomaly; this has been observed previously and represent uncertainty due to the trade-off between the velocity of the anomaly and its shape (Galetti et al., 2015; Zhang et al., 2018). There is also another higher uncertainty loop associated with the lower velocity loop between the central anomaly and the receiver array. In contrast to this result, the loop cannot be observed in the results of ADVI.

To validate and better understand these results, Figure 7 shows the results from MH-McMC. The mean velocity model is very similar to the results from ADVI and SVGD. For example, the velocity value of the low velocity anomaly is higher than the true value, which suggests that the mean value of the posterior under the specified parameterization is genuinely biased towards higher values than the true value. A lower velocity loop is also observed between the circular anomaly and the receiver array. The standard deviation map shows similar results to those from SVGD: there is a higher uncertainty loop around the central anomaly and another one associated with the lower velocity loop between the circular anomaly and the receiver array. The latter loop suggests that this area is not well constrained by the data, and therefore the mean velocity tends towards the mean value of the prior which is lower than the true value. We do not observe the clear higher uncertainty loops in the result of ADVI which may be due to the Gaussian approximation which is used to fit a non-Gaussian posterior. In Figure 8 we show the results from rj-McMC. Compared to the results from the fixed-parameterization inversions, the mean velocity is a more accurate estimate of the true model and uncertainty across the model is also lower. For example, the middle low velocity anomaly has almost the same value as the true model and has standard deviation of only $\sim 0.3\,km/s$ compared to values significantly greater than 0.3 km/s for all other methods. Between the middle anomaly and the receivers, the model is determined better than in the fixed-paramterization inversions (with a standard deviation smaller than $0.1\,km/s$). This is because in rj-McMC the model parameterization adapts to the data which usually results in a lower-dimensional parameter space due to the natural parsimony of the method. For example, the average dimensionality of the parameter space in the rj-McMC inversion is around 10; for comparison the fixed-parameterization inversions all have dimensionality fixed to be 441. The standard deviation map from the rj-McMC also shows a clear higher uncertainty loop within the array around the low velocity anomaly, and high uncertainties outside of the array where there is no data coverage.

Note that individual models from fixed-parameterization inversions (ADVI, SVGD and MH-McMC) show complex structures because of their higher dimensionality and the simple Uniform prior distribution that we adopted (right panels in Figure 5, 6 and 7). This might not be appropriate since the real Earth may have a smoother structure (de Pasquale & Linde, 2016; Ray & Myer, 2019). In that case, more informative prior information including some form of regularization might be used to produce smoother individual models (MacKay, 2003).

The results in Figure 8 do not show the double-loop uncertainty structure that is observed in the SVGD and MH-McMC results. The rj-McMC method contains an im-

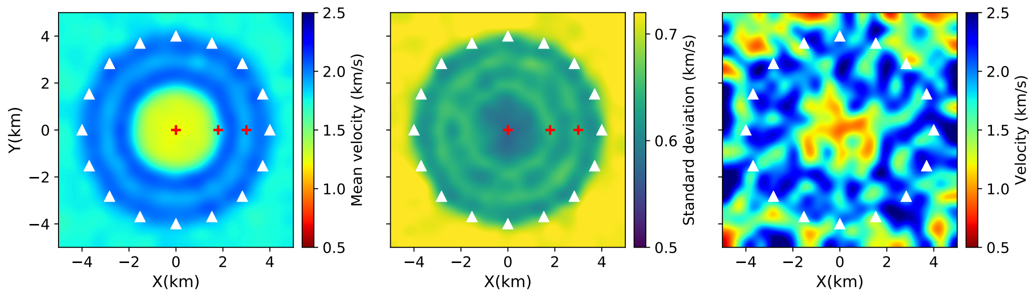

Figure 7. The mean (left), standard deviation (middle) and an individual realization from the approximate posterior distribution (right) obtained using MH-McMC. The red pluses show the point location which are referred to in the text.

plicit natural parsimony – the method tends to use fewer rather than more cells whenever possible. While this may be useful in order to reduce the dimensionality of parameter space, it is also possible that it causes some detailed features of the velocity or uncertainty structure to be omitted, much like a smoothing regularization condition in other tomographic methods. Since the double-loop structure appears to be a robust feature of the image uncertainty, we assume that the parsimony has indeed regularised some of the image structure out of the rj-McMC results.

Note that the result from rj-McMC is fundamentally different from results obtained using the fixed-parameterization inversions (ADVI, SVGD and MH-McMC) because of its entirely different parameterization. While the other inversion results are parameterized over a regular grid and can themselves be regarded as pixelated images, rj-McMC produces a set of models that are vectors containing positions and velocities of Voronoi cells, which can be transformed to an image on a regular grid (right panel in Figure 8). However, the Voronoi parametrization imposes prior restrictions on the pixelated form of models, for example all pixels within each Voronoi cell have idential vleocities. As a result rj-McMC produces very different results to those obtained using the other methods. In fact the choice of parameterizaiton in rj-McMC can impose a variety of restrictions on models, and different parameterizations can produce very different standard deviation structures (Hawkins et al., 2019). Thus the results of rj-McMC must always be interpreted in the light of the specific prior information imposed by the parameterization deployed.

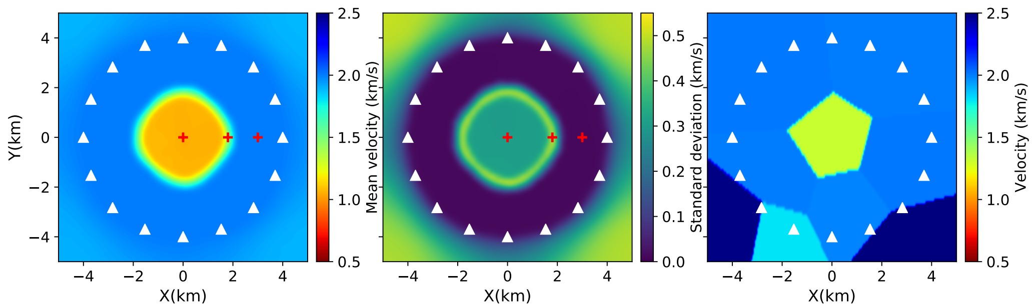

Figure 8. The mean (left), standard deviation (middle) and an individual realization from the approximate posterior distribution (right) obtained using trans-dimensional rj-McMC. The red pluses show the point location which are referred to in the text.

To further analyse the results, in Figure 9 we show marginal probability distributions from the different inversion methods at three points (plus signs in Figure 5, 6, 7, and 8): point (0, 0) at the middle of the model, point (1.8, 0) at the boundary of the low velocity anomaly which has higher uncertainties, and point (3, 0) which also has higher uncertainties in the results from SVGD and MH-McMC. Due to symmetries of the model, marginal distributions at these three points are sufficient to reflect much of the entire set of single-parameter marginal probability distributions. At point (0, 0), the three fixed-parameterization methods produce similar marginal probability distributions. However, the marginal distribution from rj-McMC is narrower and concentrates around the true solution ($1.0\,km/s$). This is likely due to the fact that in rj-McMC we have a much smaller parameter space than in the fixed-parameterization inversions. To assess the convergence we show the marginal distributions obtained by doubling the number of iterations in ADVI and SVGD with an red line in Figure 9a and b. The results show that increasing iterations only slightly improves the marginal distributions, suggesting that they have nearly converged. The black line in Figure 9b shows the marginal distribution obtained using more particles (1,600) with the same number of iterations (500). The result is almost the same as the result obtained using the original set of particles which suggests that 800 particles are sufficient in this case. At point (1.8, 0), the marginal distributions from the three fixed-parameterization inversions become broader which explains the higher uncertainty loops observed in the standard deviation maps. The distribution from ADVI is more centrally focussed than the other two, which is again suggestive of the limitations of that method caused by the Gaussian approximation. The distributions from SVGD and MH-McMC are more similar to each other and are close to the prior – a Uniform distribution – which suggests that the area is not well constrained by the data. By contrast, the result from rj-McMC shows a clearly multimodal distribution with one mode centred around the velocity of the anomaly (1 km/s) and the other around the background velocity (2 km/s) as discussed in Galetti et al. (2015). This multimodal distribution reflects the fact that it is not clear whether this point is inside or outside of the anomaly which produces the higher uncertainty loop in the standard deviation map. This suggests that there are different causes of the higher uncertainty loops in the different models. In the fixed-parameterization inversions (ADVI, SVGD and MH-McMC) the higher uncertainty loops are mainly caused by the low resolution of the data at the boundary of the low velocity anomaly which produces broader marginal distributions. In the rj-McMC inversion, the higher uncertainty loops are mainly caused by multimodality in the posterior pdf. At point (3.0, 0) similarly to the point (0, 0), the marginal distributions from the three fixed-parameterization inversions have similar shape and are much broader than the result from rj-McMC. Compared to the results from SVGD and MH-McMC, the result from ADVI again shows a more centrally-focussed distribution reminiscent of the Gaussian limitation implicit in ADVI. In the result of rj-McMC the marginal distribution concentrates to a very narrow distribution around the true value. Overall the marginal distributions from the fixed-parameterization inversions are broader than the result from rj-McMC due to their far larger parameter space. Note that although the marginal distributions from SVGD and MH-McMC have slightly different shape which causes differences in the magnitudes of their standard deviation maps, the maps are essentially similar from these quite different methods which suggests that the results are (approximately) correct.

3.2 Computational cost

Table 1 summarises the computational cost of the different methods. ADVI involves 10,000 forward simulations which takes 0.45 CPU hours. However, note that in ADVI we used the full-rank covariance matrix which becomes huge in high dimensional parameter spaces which could makes the method inefficient. SVGD involves 400,000 forward simulations which takes 8.53 CPU hours. This appears to make it less efficient than ADVI, however SVGD can produce a more accurate approximation to the posterior pdf than

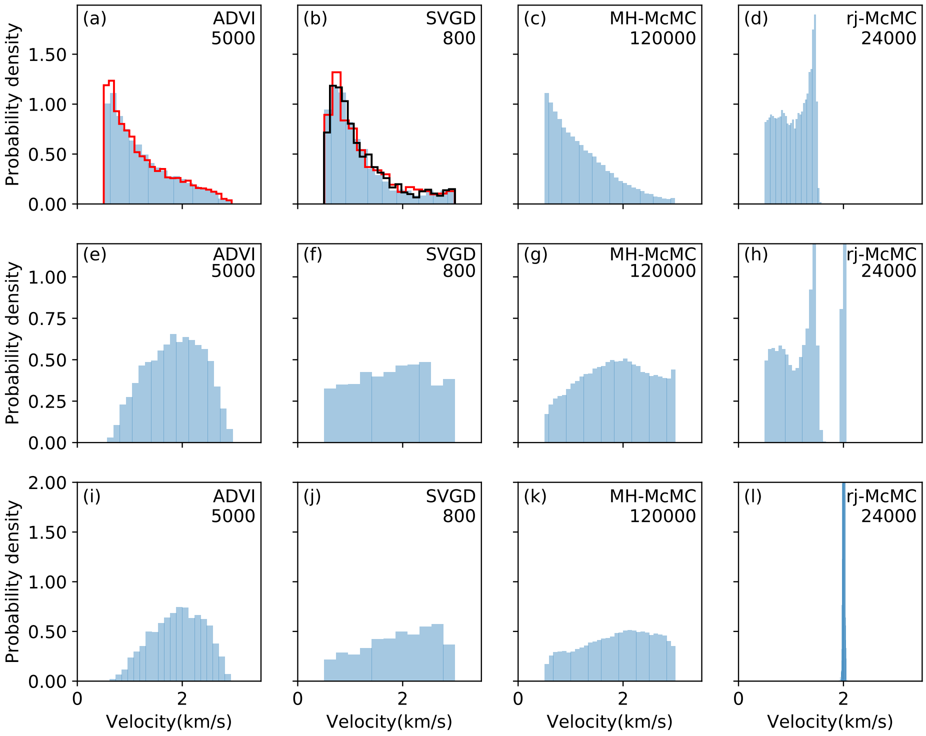

Figure 9. The marginal posterior pdfs of velocity at three points (pluses in Figure 3,4,5,6) derived using different methods. **(a)**, **(b)**, **(c)** and **(d)** show the marginal posterior distributions of velocity at the point (0,0) from ADVI, SVGD, MH-McMC and rj-McMC respectively. **(e)**, **(f)**, **(g)** and **(h)** show the marginal distributions at the point (1.8,0) from the four methods respectively, and **(i)**, **(j)**, **(k)** and **(l)** show the marginal distributions at the point (3,0) from the four methods respectively. The red lines in **(a)** and **(b)** are marginal distributions obtained by doubling the number of iterations and the black line in **(b)** shows the marginal distribution obtained using 1,600 particles. The number at the top-right of each figure shows the number of Monte Carlo samples.

ADVI which is limited by the Gaussian approximation. Note that SVGD can easily be parallelized by computing the gradients in equation (19) in parallel, making the method more time-efficient. For example, the above example takes 0.97 hours when parallelized using 10 cores. In comparison, MH-McMC requires 2,000,000 simulations for one chain which takes about 80.05 CPU hours, so for all 6 chains it requires 480.3 CPU hours in total. The rj-McMC run involved 500,000 simulations for one chain which takes about 17.1 CPU hours, so 102.6 CPU hours in total for 6 chains. The Monte Carlo methods use evaluations of the likelihood and prior distribution at each sample whereas both variational methods also deploy the information in the various gradients in equations 9, 10 and 19. The number of simulations is therefore not a good metric to compare the four methods, since the gradients in this case are calculated by ray tracing which require more calculations per simulation in Table 1 compared to MC. CPU hours is a fairer metric for comparison, but of course this depends on the mechanism by which gradients are obtained: in other forward or inverse problems it is even possible that the variational methods take longer than Monte Carlo if estimating gradients requires extensive computation.

In the comparison in Table 1, rj-McMC is more efficient than MH-McMC due to the fact that rj-McMC explores a much smaller parameter space than the fixed parameterization in MH-McMC. However, note that this might not always be true since trans-dimensional steps in rj-McMC usually have a very low probability of being accepted (Bodin & Sambridge, 2009; Zhang et al., 2018) and the method is generally significantly more difficult to tune (Green & Hastie, 2009). Overall, obtaining solutions from variational methods (ADVI, SVGD) is more efficient than Monte Carlo methods since they turn the Bayesian inference problem into an optimization problem. This also makes variational inference methods applicable to larger-datasets, and offers the advantage that very large datasets can be divided into random minibatches and inverted using stochastic optimization (Robbins & Monro, 1951; Kubrusly & Gravier, 1973) together with distributed computation. Monte Carlo methods are very computationally expensive for large datasets. Of course, the above comparison depends on the methods used to assess convergence for each method, which introduces some subjectivity in the comparison so that the absolute time required by each method may not be entirely accurate. Nevertheless, from all tests that we have conducted it is clear that variational methods produce solutions far more efficiently than Metropolis-Hastings and rj-McMC methods. Note that some other Monte Carlo sampling methods, e.g. Hamiltonian Monte Carlo, also use gradient infor-

Table 1. The comparison of computational cost for all 4 methods

Method	Number of simulations	CPU hours
ADVI	10,000	0.45
SVGD	400,000	8.53
MH-McMC	12,000,000	480.3
rj-McMC	3,000,000	102.6

mation and may be more efficient than Metropolis-Hastings methods (Neal et al., 2011; Sen & Biswas, 2017; Fichtner et al., 2018).

4 Application to Grane field

The Grane field is situated in the North sea, and contains a permanent monitoring system composed of 3458 four-component sensors measuring 3 orthogonal components of particle velocity and water pressure variations due to passing seismic waves. Zhang et al. (2019) used beamforming to show that the noise sources measured in the Grane field are nearly omnidirectional, which allows us to use ambient seismic noise tomography to study the subsurface of the field. To reduce the computational cost, in this study we down-sampled the number of receivers by a factor of 10 which results in 346 receivers, and we only used 35 receivers as virtual sources (Figure 10a). Cross-correlations are computed between vertical component recordings at pairs consisting of a virtual source and a receiver using half-hour time segments, and the set of correlations for each pair were stacked over 6.5 hours. This process produces approximate virtual-source seismograms of Rayleigh-type Scholte waves (Campillo & Paul, 2003; Shapiro et al., 2005; Curtis et al., 2006). Phase velocity dispersion curves for each (virtual) source-receiver pair are then automatically picked using an image transformation technique: for all processing details see Zhang et al. (2019) which presents a complete ambient noise analysis of the field and presents tomographic phase velocity maps at various frequencies as well as estimated shear-velocity structure of the near seabed subsurface. Here we use the recording phase velocity data at 0.9 s period.

We apply the variational inference methods ADVI and SVGD, and rj-McMC to the data to obtain phase velocity maps at 0.9 s and compare the results. For variational methods, the field is parametrized using a regular 26×71 grid with a spacing of 0.2 km at both x and y directions giving a velocity model dimensionality of 1846. Due to its computational cost in high dimensional spaces we do not apply MH-McMC. The data noise level is set to be 0.05 s, which is an average value estimated by the hierarchical Bayesian Monte Carlo inversion of Zhang et al. (2019). The prior pdf of phase velocity in each model cell is set to be a Uniform distribution between $0.35\,km/s$ and $0.55\,km/s$, which is selected to be wider than the minimum ($0.4\,km/s$) and maximum ($0.5\,km/s$) phase velocity picked from cross-correlations. The initial probability distribution for ADVI is chosen similarly to that in the synthetic tests: a standard Gaussian distribution in the unconstrained space (blue histogram in Figure 10c), and its shape in the original space is shown in Figure 10b (blue histogram). For SVGD, the initial distribution is approximated using 1000 particles generated from the prior in the original space (orange histogram in Figure 10b) and transformed to the unconstrained space (orange histogram in Figure 10c). We then applied 10,000 iterations for ADVI and 500 iterations for SVGD. Similarly to the synthetic test above for rj-McMC we use Voronoi cells to parameterize the model. The prior pdf of the number of cells is set to be a discrete Uniform distribution between 30 and 200, and the data noise level is estimated hierarchically during the inversion (Zhang et al., 2018). Proposal distributions are the same as in the synthetic test above. We used a total of 16 chains, each of which contains 800,000 iterations including a burn-in period of 400,000. To reduce the correlation between samples we only retain every 50$^{\text{th}}$ sample post burn-in for our final ensemble.

Figure 11 shows the mean and standard deviation maps from ADVI. The mean phase velocity map shows a clear low velocity anomaly around the centre of the field from Y=6 km to Y=10 km and another at the western edge between Y=8 km and Y=10 km. These were also observed by (Zhang et al., 2019) using Eikonal tomography, who showed that they are correlated with areas of higher density of pockmarks on the seabed, suggesting that they are caused by near surface fluid flow effects. At the western edge between Y=6 km and Y=8 km and at the northwestern edge there are high velocity anomalies which were also observed in the results of Zhang et al. (2019). In the north between Y=11 km and Y=12 km and along the eastern edge between Y=7 km and Y=10 km the model shows some low velocity anomalies. Moreover, there are some small anomalies distributed across the field. For example, to the south of the central low velocity anomaly around Y=6 km there are several other low velocity anomalies. Similarly there is a small low

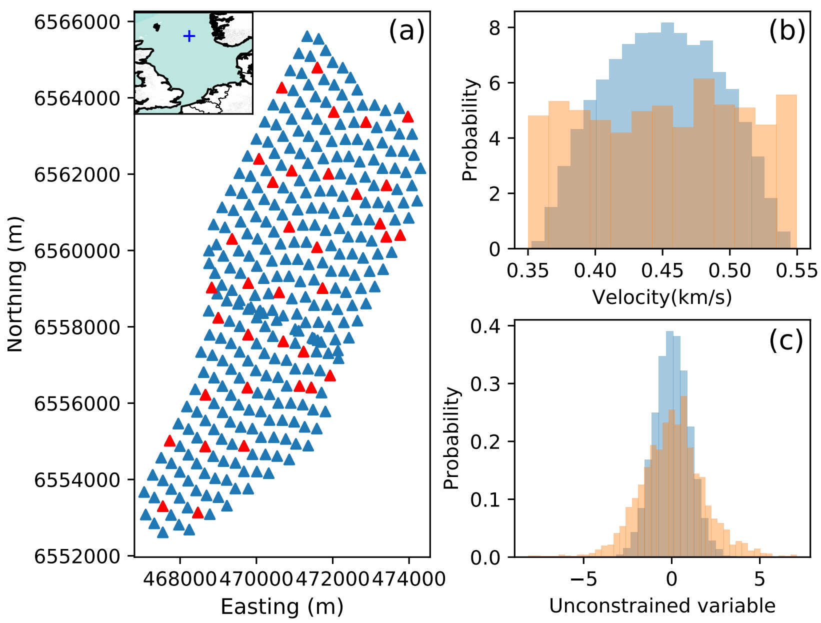

Figure 10. **(a)** The distribution of receiver (blue and red triangles) across the Grane field used in this study. Red triangles show the receivers that were used as virtual sources. The blue plus in the inset map shows the location of Grane field. The histograms show the initial distributions of a parameter in the **(b)** original (velocity) space and **(c)** transformed unconstrained space for ADVI (blue) and for SVGD (orange). Similar to Figure 4, we used 5000 Monte Carlo samples to show probability distributions in both the original and the unconstrained space for ADVI. The initial distribution for SVGD is approximated using 1000 particles generated from the prior (a Uniform distribution) in the original space and transformed to the unconstrained space.

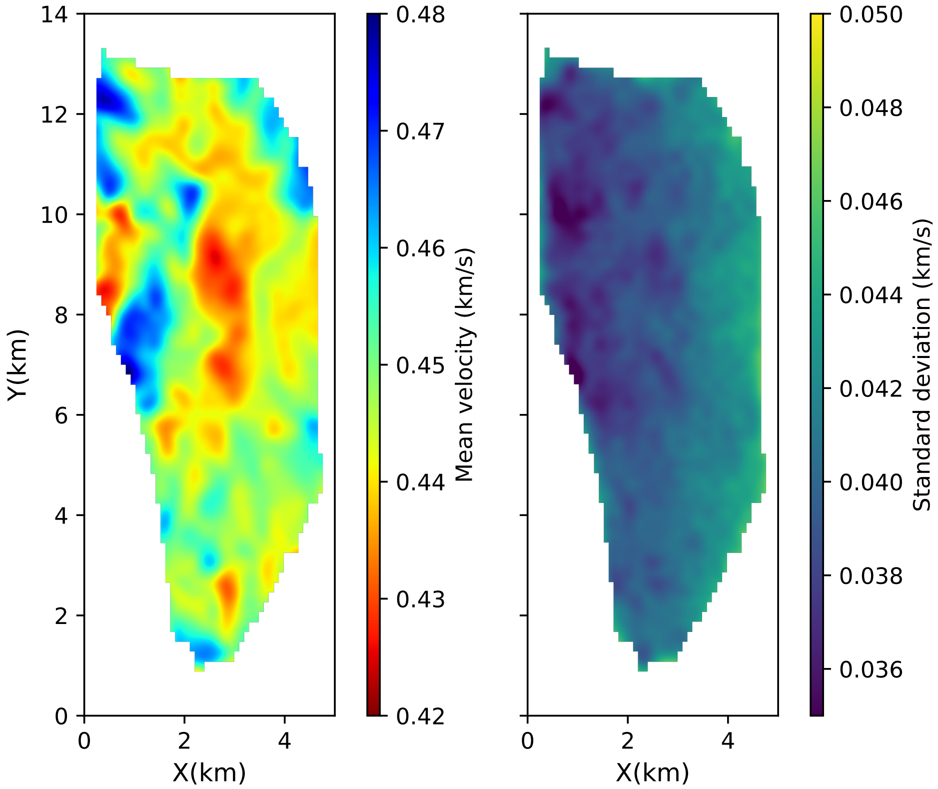

Figure 11. The mean (left) and standard deviation map (right) from ADVI.

velocity anomaly and a small high velocity anomaly in the south of the field around Y=2.5 km, and a small high velocity anomaly in the north around Y=10.5 km.

Overall the standard deviation map shows that uncertainty in the west is lower than in the east. At the western edge there are some low uncertainty areas which are associated with velocity anomalies. For example, the low uncertainty area between Y=6 km and Y=8 km is associated with the high velocity anomaly at the same location. Similarly the high velocity anomaly at the northwestern edge around Y=12 km shows a lower uncertainty, and the middle low velocity anomaly also shows slightly lower uncertainties. This might suggest that these velocity structures are well-constrained by the data. However, in the synthetic tests we noticed that the ADVI can produce biased standard

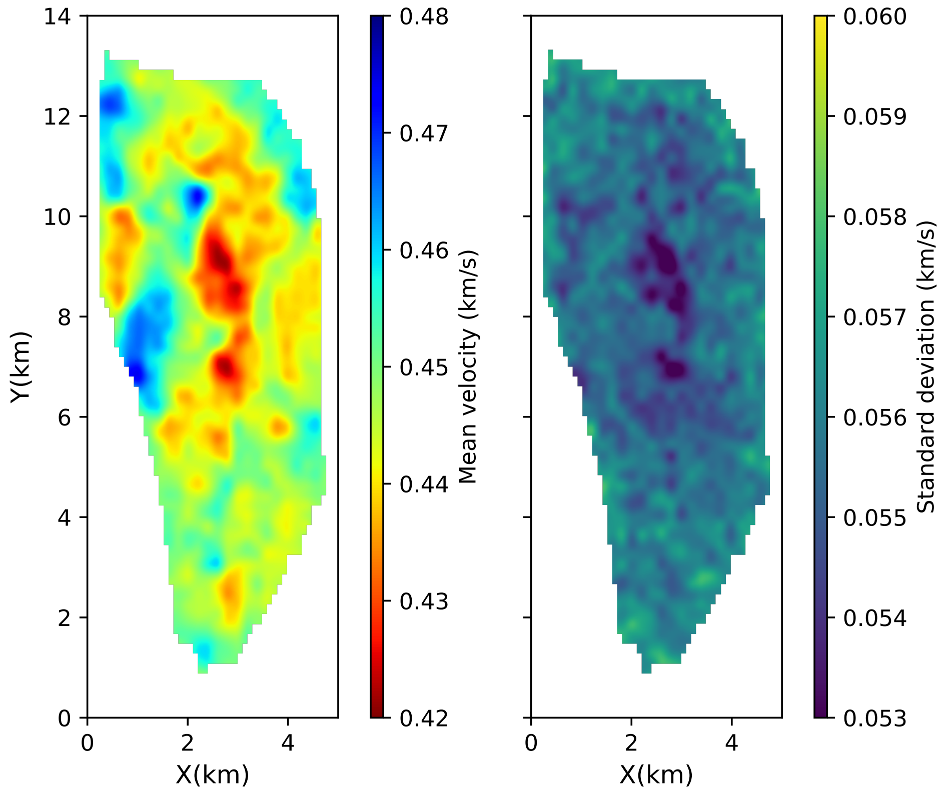

Figure 12. The mean (left) and standard deviation map (right) from SVGD.

deviation maps due to the Gaussian approximation, so these uncertainty properties may not be robust.

We show the mean and standard deviation maps obtained using SVGD in Figure 12. The mean velocity map shows very similar structures to the result from ADVI, except that the velocity magnitudes are slightly different. For example, we observe the central low velocity anomaly and one at the western edge which appeared in the mean velocity map from ADVI and are related to the density distribution of pockmarks. Similarly there are high velocity anomalies at the western edge and a low velocity anomaly at the eastern edge. Even for more detailed structure, e.g., the low velocity anomalies at the north (Y ¿ 10 km), the low velocity anomalies around Y=6 km and the small velocity anomalies around Y=2.5 km, the two results show highly consistent properties be-

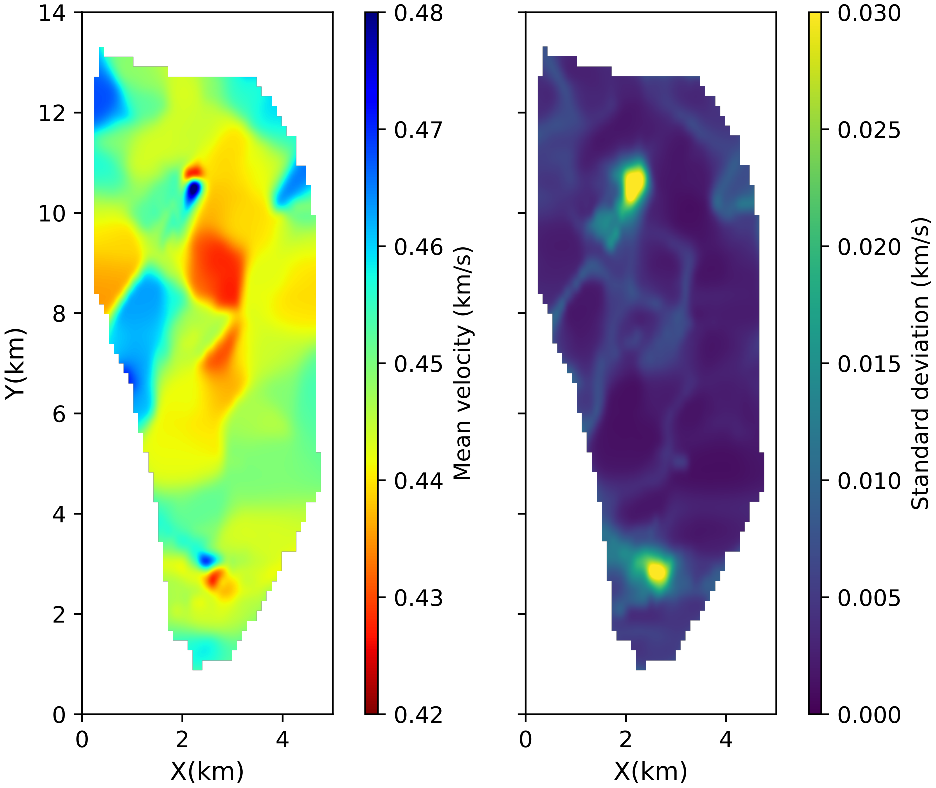

Figure 13. The mean (left) and standard deviation map (right) from rj-McMC.

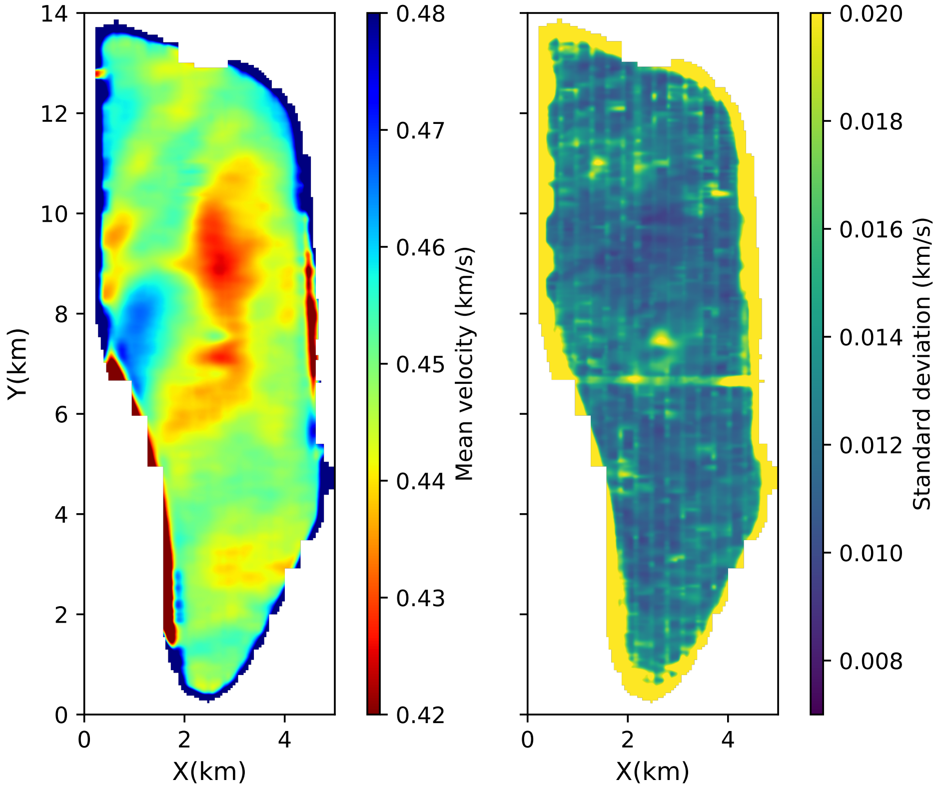

Figure 14. The mean (left) and standard deviation map (right) obtained using Eikonal tomography by Zhang et al. (2019).

tween the two methods. This suggests that we have obtained accurate mean phase velocity maps given the fixed, gridded model parameterization and the observed data.

Despite the similarity in the mean results, the standard deviation map from SVGD is quite different from the results from ADVI, which is consistent with similar variations that we observed in the synthetic tests. For example, there is no clear magnitude difference between the west and the east as appeared in the result from ADVI. There is a clear low uncertainty area associated with the central low velocity anomaly, which is slightly lower in magnitude than the result from ADVI. Similarly there is a slightly lower uncertainty area at the western edge associated with the low velocity anomaly at the same location. The south-central low velocity anomaly around Y=6 km also exhibits relatively lower uncertainties, which suggests that those small low velocity anomalies in this area may reflect true properties of the subsurface. Similarly there are some low uncertainty structures at the north around Y= 11 km which are associated with low velocity anomalies. Note that due to the Gaussian approximation in ADVI, the standard deviation results from SVGD show different magnitudes as we saw in the synthetic tests.

Figure 13 shows the mean and standard deviation maps obtained from rj-McMC. The mean velocity map shows broadly similar structures to the results from ADVI and SVGD. For example, we also observed the middle low velocity anomaly, the low velocity anomalies at the western and eastern edges and the high velocity anomalies at the western edge. However, compared to the previous results these structures are smoother which is probably caused by the natural parsimony that is implicit within the rj-McMC inversion method (Green, 1995; Bodin & Sambridge, 2009) similarly to the synthetic tests above. The small velocity anomalies in the previous results disappear in the result from rj-McMC; this may also be caused by the natural parsimony of rj-McMC, or by overfitting of data in the variational methods due to the fixed parameterization. However, the small high and low velocity anomalies around Y=2.5 km and around Y=10.5 km still exist, which suggests that these detailed velocity structures may represent real properties of the subsurface (or are caused by a consistent bias in the data).

Similarly to the synthetic tests, the standard deviation map from rj-McMC shows significantly smaller uncertainties (< 0.01 km/s) than the results from ADVI (~ 0.04 km/s) and SVGD ($\sim\ 0.055$ km/s), which is probably caused by a lower dimensionality of parameter space used in rj-McMC (around 60 Voronoi cells were used) than in variational methods (1846), resulting in fewer trade-offs between parameters. However, there are higher uncertainties at the location of the small velocity anomalies at Y=2.5 km and at Y=10.5 km, which is probably due to the fact that not all chains found these small structures.

To compare our results with traditional methods, Figure 14 shows the mean and standard deviation maps obtained using Eikonal tomography by Zhang et al. (2019) using all of the available data (3458 virtual sources and 3458 receivers). The mean velocity model shows similar but slightly smoother structures compared to those obtained using ADVI and SVGD. This may be because the larger quantity of data used in Eikonal tomography reduces the noise and stabilizes the results, or because the interpolation used in Eikonal tomography regularizes (smooths) small scale structure. The standard deviation map shows lower uncertainties at the location of the middle low velocity anomaly which is similar to that obtained using SVGD. This again suggests that SVGD can produce a more accurate standard deviation estimate than ADVI. The mean velocity model from rj-McMC shows smoother structures than that from Eikonal tomography, which may suggest that rj-McMC omits small scale structure due to its implicit parsimony. The standard deviation map from rj-McMC also does not show similar structures to those obtained using Eikonal tomography or SVGD due to the completely different parameterizations employed.

In the inversion, ADVI involved 10,000 forward simulations which took 5.1 CPU hours and SVGD involved 500,000 forward simulations which required 141.8 CPU hours. By contrast the rj-McMC involved 12,800,000 forward simulations to obtain an acceptable result which required 1,866.1 CPU hours. In real time, SVGD was in fact parallelised using 12 cores which took 12.1 hours to run, while rj-McMC was parallelised using 16 cores which therefore took about 5 days. We conclude that, although the variational methods produce higher uncertainty estimates, they can produce similar parameter estimates (mean velocity) at hugely reduced computational cost, and indeed our synthetic tests suggest that the variational SVGD image uncertainty results may in fact be more correct.

5 Discussion

We have shown that variational methods (ADVI and SVGD) can be applied to seismic tomography problems and provide efficient alternatives to McMC. ADVI produces biased posterior pdfs because of its implicit Gaussian approximation, and cannot be applied to problems with multi-modal posteriors. However, it still generates an accurate estimate of the mean model. Given that it is very efficient (only requiring 10,000 forward simulations) the method could be useful in scenarios where efficiency is important and a Gaussian approximation is sufficient for uncertainty analysis. Alternatively a mixture of Gaussians approximation might be used to improve the accuracy of the algorithm (Zobay et al., 2014; Arenz et al., 2018). In a very high dimensional case, ADVI could become less efficient because of the increased size of the Gaussian covariance matrix. In that case one could use a mean-field approximation (setting model covariances to zero), or use a sparse covariance matrix to reduce computational cost since seismic velocity in any cell is often most strongly correlated with that in neighbouring cells.

SVGD can produce a good approximation to posterior pdfs. However, since it is based on a number of particles, the method is more computationally costly than ADVI. In this study we parallelized the computation of gradients to improve the efficiency, and for large datasets further improvements can be obtained by using random minibatches to perform the inversion (Q. Liu & Wang, 2016). Such a strategy can be applied to any variational inference method (e.g. also ADVI) since variational methods solve an optimization rather than a stochastic sampling problem. In comparison, this strategy cannot easily be used in McMC based methods since it may break the detailed balance requirement of McMC (Blei et al., 2017). Though it has been shown that SVGD requires fewer particles than particle-based sampling methods (e.g., sequential Monte Carlo) in the sense that it reduces to finding the MAP model if only one particle is used, the optimal choice of the number of particles remains unclear, especially for very high dimensional spaces. In the case of very high dimensionality another possibility is to use normalizing flows – a variational method based on a series of specific invertible transforms (Rezende & Mohamed, 2015).

Monte Carlo and variational inference are different types of methods that solve the same problem. Monte Carlo simulates a set of Markov chains and uses samples of those chains to approximate the posterior pdf, while variational inference solves an optimization problem to find the closest pdf to the posterior within a given family of probability distributions. Monte Carlo methods provide guarantees that samples are asymptotically distributed according to the posterior pdf as the number of samples tends to infinity (Robert & Casella, 2013), while the statistical properties of variational inference algorithms are still unknown (Blei et al., 2017). It is possible to combine the two methods to capitalise on the merits of both. For example, the approximate posterior pdf from an efficient variational method (e.g. ADVI) can be used as a proposal distribution for Metropolis-Hastings (De Freitas et al., 2001) to improve the efficiency of McMC, or McMC steps can be integrated to the variational approximation to improve the accuracy of variational methods (Salimans et al., 2015).

We used a fixed regular grid of cells to parameterize the tomographic model in the variational methods, which might introduce overfitting of the data. For example, the mean velocity models in the synthetic tests show a slightly lower velocity loop between the low velocity anomaly and the receivers, and the uncertainties obtained from fixed-parameterization inversions are significantly higher than the results from rj-McMC. However, although rj-McMC produces lower uncertainty estimates, small scale structures can be omitted in the results of rj-McMC due to their implicitly imposed parsimony. For example, in our real data example, small scale structures in the results of variational inference methods and Eikonal tomography are smoothed out in the results of rj-McMC. Indeed the parameterization used in rj-McMC imposes restrictions on models, and different parameterizations can produce different uncertainties (Hawkins et al., 2019). This makes the interpretation and use of uncertainties from rj-McMC difficult.

It is not easy to determine an optimal grid in variational inference methods since this introduces a trade off between resolution of the model and overfitting of the data. Therefore, it might be necessary to use a more flexible parameterization, e.g., Voronoi cells (Bodin & Sambridge, 2009; Zhang et al., 2018) or wavelet parameterization (Fang & Zhang, 2014; Hawkins & Sambridge, 2015; Zhang & Zhang, 2015). It may also be possible to apply a series of different parameterizations and select the best one using model selection theory (Walter & Pronzato, 1997; Curtis & Snieder, 1997; Arnold & Curtis, 2018). Note that it would make the methods less computationally efficient to find an optimal parameterization because we may need to run a series of optimization problems with different parameterizations. However, in cases with very large datasets which may more suitably be solved by variational inference methods, it might instead be sufficient to use a parameterization with the highest resolution that the frequency of the data could resolve. Instead some more informative prior or regularization may be used to reduce the magnitude of uncertainty estimates and to better constrain the model (MacKay, 2003; Ray & Myer, 2019).

In our experiments the results from rj-McMC are significantly different from the results obtained using variational methods or MH-McMC. This is essentially caused by different parameterizations. In ADVI, SVGD and MH-McMC we invert for a pixelated image, while in rj-McMC we invert for a distribution of parameters that represent locations and shapes of cells and their constant velocities, the pointwise spatial mean of which is visualized as an image. Therefore even though we visualized them in the same way, the results are essentially not directly comparable. Nevertheless, the comparison with rj-McMC is interesting because until now a quite different alternative probabilistic method was never used to estimate the posterior of images from the same realistic tomography problem. The results here demonstrate that the rj-McMC method as applied in most tomography papers gives significantly different solutions than we might previously have thought; specifically, it does not produce the posterior distribution of the pixelated image that is usually shown in scientific papers (e.g., Bodin & Sambridge, 2009; Zulfakriza et al., 2014; Galetti et al., 2015; Crowder et al., 2019). Rather, it samples a probability distribution in a particular irregular and variably parametrized model space and results should be interpreted as such. Note that some other methods, e.g. rj-McMC with Gaussian processes, may provide results that can be compared between all sampling methods, and provide a means of injecting prior information with adaptable complexity into the sampling scheme (Ray & Myer, 2019).

In this study we used a fixed data noise level in the variational methods. It has been shown that an improper noise level can introduce biases in tomographic results (Bodin & Sambridge, 2009; Zhang et al., 2019), so in our example we used the noise level estimated by hierarchical McMC. It can also be estimated by a variety of other methods (Bensen et al., 2009; Yao & Van Der Hilst, 2009; Weaver et al., 2011; Nicolson et al., 2012, 2014), and maximum likelihood methods (Sambridge, 2013; Ray et al., 2016; Ray & Myer, 2019). In future it might also be possible to include the noise parameters in variational methods in a hierarchical way.

In this study we applied variational inference methods to simple 2D tomography problems, but it is straightforward to apply the methods to any geophysical inverse problems whose gradients with respect to the model can be computed efficiently. For example, variational methods can be applied to 3D seismic tomography problems to provide efficient approximation, which generally demands enormous computational resources using McMC methods (Hawkins & Sambridge, 2015; Zhang et al., 2018, 2019). The methods also provide possibilities to perform Bayesian inference for full waveform inversion, which is generally very expensive for McMC (Ray et al., 2017) and suffers from notorious multimodality in the likelihoods. SVGD provides a possible way to approximate these complex distributions given that theoretically it can approximate arbitrary distributions.

6 Conclusion

We introduced two variational inference methods to geophysical tomography – automatic differential variational inference (ADVI) and Stein variational gradient descent (SVGD), and applied them to 2D seismic tomography problems using both synthetic and real data. Compared to the Markov chain Monte Carlo (McMC) method, ADVI provides an efficient but biased approximation to Bayesian posterior probability density functions, and cannot be applied to find multi-modal posteriors because of its implicit Gaussian assumption. In contrast, SVGD is slightly slower than ADVI but produces a more accurate approximation. The real data example shows that ADVI and SVGD produce very similar mean velocity models, even though their uncertainty estimates are different . The mean velocity models are very similar to those produced by reversible jump McMC (rj-McMC), except that the mean model from rj-McMC is smoother because of the much lower dimensionality of its parameter space. Variational methods thus can provide efficient approximate alternatives to McMC methods, and can be applied to many geophysical inverse problems.

Acknowledgments

The authors would like to thank the Grane license partners Equinor ASA, Petoro AS, ExxonMobil E&P Norway AS, and ConocoPhillips Skandinavia AS for allowing us to publish this work. The views and opinions expressed in this paper are those of the authors and are not necessarily shared by the license partners. The authors thank the Edinburgh Interferometry Project sponsors (Schlumberger, Equinor and Total) for supporting this research. This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk). The data used in this study are available at Edinburgh DataShare (https://doi.org/10.7488/ds/2607).

References

Aki, K., & Lee, W. (1976). Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. a homogeneous initial model. *Journal of Geophysical research*, *81*(23), 4381–4399.

Arenz, O., Zhong, M., & Neumann, G. (2018). Efficient gradient-free variational inference using policy search. In *International conference on machine learning* (pp. 234–243).

Arnold, R., & Curtis, A. (2018). Interrogation theory. *Geophysical Journal International*, *214*(3), 1830–1846.

Bensen, G., Ritzwoller, M., & Yang, Y. (2009). A 3-D shear velocity model of the crust and uppermost mantle beneath the United States from ambient seismic noise. *Geophysical Journal International*, *177*(3), 1177–1196.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* springer.

Blatter, D., Key, K., Ray, A., Gustafson, C., & Evans, R. (2019). Bayesian joint inversion of controlled source electromagnetic and magnetotelluric data to image freshwater aquifer offshore new jersey. *Geophysical Journal International*, *218*(3), 1822–1837.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bodin, T., & Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, *178*(3), 1411–1436.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K., & Rawlinson, N. (2012). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth*, *117*(B2).

Burdick, S., & Lekić, V. (2017). Velocity variations and uncertainty from transdimensional P-wave tomography of North America. *Geophysical Journal Interna-*

tional, *209*(2), 1337–1351.

Campillo, M., & Paul, A. (2003). Long-range correlations in the diffuse seismic coda. *Science*, *299*(5606), 547–549.

Çınlar, E. (2011). *Probability and stochastics* (Vol. 261). Springer Science & Business Media.

Crowder, E., Rawlinson, N., Pilia, S., Cornwell, D., & Reading, A. (2019). Trans-dimensional ambient noise tomography of Bass Strait, southeast Australia, reveals the sedimentary basin and deep crustal structure beneath a failed continental rift. *Geophysical Journal International*, *217*(2), 970–987.

Curtis, A., Gerstoft, P., Sato, H., Snieder, R., & Wapenaar, K. (2006). Seismic interferometry – turning noise into signal. *The Leading Edge*, *25*(9), 1082–1092.

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

Curtis, A., & Snieder, R. (1997). Reconditioning inverse problems using the genetic algorithm and revised parameterization. *Geophysics*, *62*(5), 1524–1532.

Curtis, A., & Snieder, R. (2002). Probing the earth's interior with seismic tomography. *International Geophysics Series*, *81*(A), 861–874.

De Freitas, N., Højen-Sørensen, P., Jordan, M. I., & Russell, S. (2001). Variational MCMC. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 120–127).

de Pasquale, G., & Linde, N. (2016). On structure-based priors in bayesian geophysical inversion. *Geophysical Journal International*, *208*(3), 1342–1358.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., & Scheichl, R. (2018). A stein variational newton method. In *Advances in neural information processing systems* (pp. 9169–9179).

Devilee, R., Curtis, A., & Roy-Chowdhury, K. (1999). An efficient, probabilistic neural network approach to solving inverse problems: Inverting surface wave velocities for Eurasian crustal thickness. *Journal of Geophysical Research: Solid Earth*, *104*(B12), 28841–28857.

Dziewonski, A. M., & Woodhouse, J. H. (1987). Global images of the Earth's interior. *Science*, *236*(4797), 37–48.

Earp, S., & Curtis, A. (2019). Probabilistic neural-network based 2D travel time tomography. *arXiv preprint arXiv:1907.00541*.

Fang, H., & Zhang, H. (2014). Wavelet-based double-difference seismic tomography with sparsity regularization. *Geophysical Journal International*, *199*(2), 944–955.

Fichtner, A., Zunino, A., & Gebraad, L. (2018). Hamiltonian monte carlo solution of tomographic inverse problems. *Geophysical Journal International*, *216*(2), 1344–1363.

Galetti, E., & Curtis, A. (2018). Transdimensional electrical resistivity tomography. *Journal of Geophysical Research: Solid Earth*, *123*(8), 6347–6377.

Galetti, E., Curtis, A., Baptie, B., Jenkins, D., & Nicolson, H. (2017). Transdimensional love-wave tomography of the British Isles and shear-velocity structure of the east Irish Sea Basin from ambient-noise interferometry. *Geophysical Journal International*, *208*(1), 36–58.

Galetti, E., Curtis, A., Meles, G. A., & Baptie, B. (2015). Uncertainty loops in travel-time tomography from nonlinear wave physics. *Physical review letters*, *114*(14), 148501.

Gorham, J., & Mackey, L. (2015). Measuring sample quality with Stein's method. In *Advances in neural information processing systems* (pp. 226–234).

Gorham, J., & Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 1292–1301).

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Byesian model determination. *Biometrika*, 711–732.

Green, P. J., & Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, *155*(3), 1391–1403.

Gretton, A. (2013). Introduction to RKHS, and some simple kernel algorithms.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hawkins, R., Bodin, T., Sambridge, M., Choblet, G., & Husson, L. (2019). Trans-dimensional surface reconstruction with different classes of parameterization. *Geochemistry, Geophysics, Geosystems*, *20*(1), 505–529.

Hawkins, R., & Sambridge, M. (2015). Geophysical imaging using trans-dimensional trees. *Geophysical Journal International*, *203*(2), 972–1000.

Hoffman, M. D., & Blei, D. M. (2015). Structured stochastic variational inference.

In *Artificial intelligence and statistics.*

Iyer, H., & Hirahara, K. (1993). *Seismic tomography: Theory and practice.* Springer Science & Business Media.

Karlin, S. (2014). *A first course in stochastic processes.* Academic press.

Käufl, P., Valentine, A., de Wit, R., & Trampert, J. (2015). Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition. *Bulletin of the Seismological Society of America*, *105*(4), 2299–2312.

Käufl, P., Valentine, A. P., O'Toole, T. B., & Trampert, J. (2013). A framework for fast probabilistic centroid-moment-tensor determination – inversion of regional static displacement measurements. *Geophysical Journal International*, *196*(3), 1676–1693.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Byes. *arXiv preprint arXiv:1312.6114*.

Kubrusly, C., & Gravier, J. (1973). Stochastic approximation algorithms and applications. In *1973 ieee conference on decision and control including the 12th symposium on adaptive processes* (pp. 763–766).

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, *18*(1), 430–474.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Liu, C., & Zhu, J. (2018). Riemannian stein variational gradient descent for bayesian inference. In *Thirty-second aaai conference on artificial intelligence.*

Liu, Q. (2017). Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems* (pp. 3115–3123).

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning* (pp. 276–284).

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose Byesian inference algorithm. In *Advances in neural information processing systems* (pp. 2378–2386).

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Malinverno, A. (2002). Parsimonious Byesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, *151*(3), 675–688.

Malinverno, A., & Briggs, V. A. (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Byes and empirical Byes. *Geophysics*, *69*(4), 1005–1016.

Malinverno, A., Leaney, S., et al. (2000). A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data. In *2000 seg annual meeting.*

Marzouk, Y., Moselhy, T., Parno, M., & Spantini, A. (2016). An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*.

Meier, U., Curtis, A., & Trampert, J. (2007a). A global crustal model constrained by nonlinearised inversion of fundamental mode surface waves. *Geophysical Research Letters*, *34*, L16304.

Meier, U., Curtis, A., & Trampert, J. (2007b). Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International*, *169*(2), 706–722.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American statistical association*, *44*(247), 335–341.

Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, *100*(B7), 12431–12447.

Nawaz, M., & Curtis, A. (2019). Rapid discriminative variational Byesian inversion of geophysical data for the spatial distribution of geological properties. *Journal of Geophysical Research: Solid Earth*.

Nawaz, M. A., & Curtis, A. (2018). Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies. *Geophysical Journal International*, *214*(2), 845–875.

Neal, R. M., et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, *2*(11), 2.

Nicolson, H., Curtis, A., & Baptie, B. (2014). Rayleigh wave tomography of the British Isles from ambient seismic noise. *Geophysical Journal International*, *198*(2), 637–655.

Nicolson, H., Curtis, A., Baptie, B., & Galetti, E. (2012). Seismic interferometry and ambient noise tomography in the British Isles. *Proceedings of the Geologists' Association*, *123*(1), 74–86.

Piana Agostinetti, N., Giacomuzzi, G., & Malinverno, A. (2015). Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling. *Geophysical Journal International*, *201*(3), 1598–1617.

Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics* (pp. 814–822).

Ranganath, R., Tran, D., & Blei, D. (2016). Hierarchical variational models. In *International conference on machine learning* (pp. 324–333).

Rawlinson, N., & Sambridge, M. (2004). Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics*, *69*(5), 1338–1350.

Ray, A., Alumbaugh, D. L., Hoversten, G. M., & Key, K. (2013). Robust and accelerated Byesian inversion of marine controlled-source electromagnetic data using parallel tempering. *Geophysics*, *78*(6), E271–E280.

Ray, A., Kaplan, S., Washbourne, J., & Albertin, U. (2017). Low frequency full waveform seismic inversion within a tree based Byesian framework. *Geophysical Journal International*, *212*(1), 522–542.

Ray, A., & Myer, D. (2019). Bayesian geophysical inversion with trans-dimensional gaussian process machine learning. *Geophysical Journal International*, *217*(3), 1706–1726.

Ray, A., Sekar, A., Hoversten, G. M., & Albertin, U. (2016). Frequency domain full waveform elastic inversion of marine seismic data from the alba field using a bayesian trans-dimensional algorithm. *Geophysical Journal International*, *205*(2), 915–937.

Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods.* Springer Science & Business Media.

Röth, G., & Tarantola, A. (1994). Neural networks and inversion of seismic data. *Journal of Geophysical Research: Solid Earth*, *99*(B4), 6753–6768.

Salimans, T., Kingma, D., & Welling, M. (2015). Markov chain Monte Carlo and variational inference: Bridging the gap. In *International conference on machine learning* (pp. 1218–1226).

Sambridge, M. (1999). Geophysical inversion with a neighbourhood algorithm – i. searching a parameter space. *Geophysical journal international*, *138*(2), 479–494.

Sambridge, M. (2013). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, ggt342.

Saul, L. K., & Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In *Advances in neural information processing systems* (pp. 486–492).

Sen, M. K., & Biswas, R. (2017). Transdimensional seismic inversion using the reversible jump hamiltonian monte carlo algorithm. *Geophysics*, *82*(3), R119–R134.

Shahraeeni, M. S., & Curtis, A. (2011). Fast probabilistic nonlinear petrophysical inversion. *Geophysics*, *76*(2), E45–E58.

Shahraeeni, M. S., Curtis, A., & Chao, G. (2012). Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data. *Geophysics*, *77*(3), O1–O19.

Shapiro, N. M., Campillo, M., Stehly, L., & Ritzwoller, M. H. (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science*, *307*(5715), 1615–1618.

Shen, W., Ritzwoller, M. H., & Schulte-Pelkum, V. (2013). A 3-D model of the crust and uppermost mantle beneath the central and western US by joint inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth*, *118*(1), 262–276.

Shen, W., Ritzwoller, M. H., Schulte-Pelkum, V., & Lin, F.-C. (2012). Joint inversion of surface wave dispersion and receiver functions: a Byesian Monte-Carlo approach. *Geophysical Journal International*, *192*(2), 807–836.

Sivia, D. (1996). Data analysis: A Byesian tutorial (oxford science publications).

Smith, A. (2013). *Sequential Monte Carlo methods in practice.* Springer Science & Business Media.

Stein, C., et al. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the* *sixth berkeley symposium on mathematical statistics and probability, volume 2: Probability theory.*

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation* (Vol. 89). SIAM.

Team, S. D., et al. (2016). Stan modeling language users guide and reference manual. *Technical report*.

Tran, D., Ranganath, R., & Blei, D. M. (2015). The variational Gaussian process. *arXiv preprint arXiv:1511.06499*.

Walter, E., & Pronzato, L. (1997). *Identification of parametric models from experimental data.* Springer Verlag.

Weaver, R. L., Hadziioannou, C., Larose, E., & Campillo, M. (2011). On the precision of noise correlation interferometry. *Geophysical Journal International*, *185*(3), 1384–1392.

Yao, H., & Van Der Hilst, R. D. (2009). Analysis of ambient noise energy distribution and phase velocity bias in ambient noise tomography, with application to SE tibet. *Geophysical Journal International*, *179*(2), 1113–1132.

Young, M. K., Rawlinson, N., & Bodin, T. (2013). Transdimensional inversion of ambient seismic noise for 3D shear velocity structure of the Tasmanian crust. *Geophysics*, *78*(3), WB49–WB62.

Zhang, X., Curtis, A., Galetti, E., & de Ridder, S. (2018). 3-D Monte Carlo surface wave tomography. *Geophysical Journal International*, *215*(3), 1644–1658.

Zhang, X., Hansteen, F., & Curtis, A. (2019). Fully 3D Monte Carlo ambient noise tomography over Grane field. In *81st eage conference and exhibition 2019.*

Zhang, X., & Zhang, H. (2015). Wavelet-based time-dependent travel time tomography method and its application in imaging the Etna volcano in Italy. *Journal of Geophysical Research: Solid Earth*, *120*(10), 7068–7084.

Zhdanov, M. S. (2002). *Geophysical inverse theory and regularization problems* (Vol. 36). Elsevier.

Zheng, D., Saygin, E., Cummins, P., Ge, Z., Min, Z., Cipta, A., & Yang, R. (2017). Transdimensional Byesian seismic ambient noise tomography across SE tibet. *Journal of Asian Earth Sciences*, *134*, 86–93.

Zobay, O., et al. (2014). Variational bayesian inference with gaussian-mixture approximations. *Electronic Journal of Statistics*, *8*(1), 355–389.

Zulfakriza, Z., Saygin, E., Cummins, P., Widiyantoro, S., Nugraha, A. D., Lühr, B.-G., & Bodin, T. (2014). Upper crustal structure of central Java, Indonesia, from transdimensional seismic ambient noise tomography. *Geophysical Journal International*, *197*(1), 630–635.

Appendix A The entropy of a Gaussian distribution

The entropy $\mathrm{H}\left[q(\boldsymbol{\theta};\boldsymbol{\phi})\right]$ of a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{L}\mathbf{L}^T)$ is:

$$\begin{aligned}
\mathrm{H}\left[q(\boldsymbol{\theta};\boldsymbol{\phi})\right] &= -\mathrm{E}_q[\log q(\boldsymbol{\theta})] \\
&= -\int \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{L}\mathbf{L}^T)\log\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{L}\mathbf{L}^T)d\boldsymbol{\theta} \\
&= \frac{k}{2}+\frac{k}{2}\log(2\pi)+\frac{1}{2}\log|det(\mathbf{L}\mathbf{L}^T)|
\end{aligned}$$

where k is the dimension of vector $\boldsymbol{\theta}$. The gradients with respect to $\boldsymbol{\mu}$ and $\mathbf{L}$ can be easily calculated (see Appendix B).

Appendix B Gradients of the ELBO in ADVI

We first describe the dominated convergence theorem (DCT) (Çınlar, 2011):

Theorem Assume $X \in \mathcal{X}$ is a random variable and $f : \mathbb{R}\times\mathcal{X} \to \mathbb{R}$ is a function such that $f(t, X)$ is integrable for all t and $\frac{\partial f(t,X)}{\partial t}$ exists for each t. Assume that there is a random variable Z such that $|\frac{\partial f(t,X)}{\partial t}| \leq Z$ for all t and $\mathrm{E}(Z) < \infty$. Then

$$\frac{\partial}{\partial t}\mathrm{E}(f(t,X)) = \mathrm{E}(\frac{\partial}{\partial t}f(t,X))$$

The proof of this theorem is given in Çınlar (2011).

We then calculate the gradients in equation (9) and (10) based on Kucukelbir et al. (2017). The ELBO $\mathcal{L}$ is:

$$\mathcal{L} = \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\log p\big(T^{-1}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\big)+\log|det\mathbf{J}_{T^{-1}}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right)|\right]+\mathrm{H}\left[q(\boldsymbol{\theta};\boldsymbol{\phi})\right]$$

where $\mathrm{H}\left[q(\boldsymbol{\theta};\boldsymbol{\phi})\right] = \mathrm{E}_q\left[\log q(\boldsymbol{\theta}\right]$ is the entropy of distribution q. Assume $\frac{\partial}{\partial\boldsymbol{\phi}}\log p$ is bounded where $\boldsymbol{\phi}$ represents variational parameters $\boldsymbol{\mu}$ and $\mathbf{L}$, then the gradients can be computed by exchanging the derivative and the expectation using the dominated convergence theorem (DCT) and applying the chain rule:

$$\nabla_{\boldsymbol{\mu}}\mathcal{L} = \nabla_{\boldsymbol{\mu}}\left\{\mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\log p\big(T^{-1}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\big)+\log|det\mathbf{J}_{T^{-1}}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right)|\right]+\mathrm{H}\left[q(\boldsymbol{\theta};\boldsymbol{\phi})\right]\right\}$$

Applying the DCT and since H does not depend on $\boldsymbol{\mu}$,

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

Applying the chain rule,

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\mathcal{L} &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{m}}\mathrm{log}p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta})\nabla_{\boldsymbol{\mu}}R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta}) + \nabla_{\boldsymbol{\theta}}\mathrm{log}|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\nabla_{\boldsymbol{\mu}}R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right]\\ &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{m}}\mathrm{log}p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\mathrm{log}|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\right]\end{aligned}$$

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

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L} &= \nabla_{\mathbf{L}}\Big\{\mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\mathrm{log}p\left(T^{-1}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right) + \mathrm{log}|det\mathbf{J}_{T^{-1}}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right)|\right]\\ &\quad +\frac{k}{2} + \frac{k}{2}\mathrm{log}(2\pi) + \frac{1}{2}\mathrm{log}|det(\mathbf{L}\mathbf{L}^T)|\Big\}\end{aligned}$$

Applying the DCT

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L} &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{L}}\left\{\mathrm{log}p\left(T^{-1}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right)\right\} + \nabla_{\mathbf{L}}\left(\mathrm{log}|det\mathbf{J}_{T^{-1}}\left(R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right)|\right)\right]\\ &\quad +\nabla_{\mathbf{L}}\frac{1}{2}\mathrm{log}|det(\mathbf{L}\mathbf{L}^T)|\end{aligned}$$

and applying the chain rule we obtain

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L} &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{m}}\mathrm{log}p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta})\nabla_{\mathbf{L}}R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta}) + \nabla_{\boldsymbol{\theta}}\mathrm{log}|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\nabla_{\mathbf{L}}R_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\right] + (\mathbf{L}^{-1})^T\\ &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\left(\nabla_{\mathbf{m}}\mathrm{log}p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\mathrm{log}|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\right)\boldsymbol{\eta}^T\right] + (\mathbf{L}^{-1})^T\end{aligned}$$

Appendix C Gradients of KL-divergence in SVGD

We calculate the gradient in equation (12) following Q. Liu and Wang (2016). Denote T^{-1} as the inverse transform of T. Then by changing the variable,

$$\mathrm{KL}[q_T||p] = \mathrm{KL}[q||p_{T^{-1}}]$$

and hence

$$\begin{aligned}\nabla_{\epsilon}\mathrm{KL}[q_T||p]\,|_{\epsilon=0} &= \nabla_{\epsilon}\mathrm{KL}[q||p_{T^{-1}}]\,|_{\epsilon=0}\\ &= \nabla_{\epsilon}\left[\mathrm{E}_q\mathrm{log}q(\mathbf{m}) - \mathrm{E}_q\mathrm{log}p_{T^{-1}}(\mathbf{m})\right]\end{aligned}$$

and since $q(\mathbf{m})$ does not depend on ϵ

$$\nabla_{\epsilon}\mathrm{KL}[q_T||p]\,|_{\epsilon=0} = -\mathrm{E}_q\left[\nabla_{\epsilon}\mathrm{log}p_{T^{-1}}(\mathbf{m})\right]$$

where $p_{T^{-1}}(\mathbf{m}) = p(T(\mathbf{m}))\cdot|\det(\nabla_{\mathbf{m}}T(\mathbf{m}))|$. Therefore

$$\nabla_{\epsilon}\mathrm{log}p_{T^{-1}}(\mathbf{m}) = (\nabla_{\mathbf{m}}\log(p(\mathbf{m})))^{\mathrm{T}}\nabla_{\epsilon}T(\mathbf{m}) + trace\left((\nabla_{\mathbf{m}}T(\mathbf{m}))^{-1}\cdot\nabla_{\epsilon}\nabla_{\mathbf{m}}T(\mathbf{m})\right)$$

where $T(\mathbf{m}) = \mathbf{m} + \epsilon\boldsymbol{\phi}(\mathbf{m})$, $\nabla_\epsilon T(\mathbf{m} = \boldsymbol{\phi}(\mathbf{m})$ and $\nabla_\mathbf{m} T(\mathbf{m})|_{\epsilon=0} = \mathbf{I}$, and so

$$\begin{aligned}
\nabla_\epsilon \mathrm{KL}[q_T||p]\,|_{\epsilon=0} &= -\mathrm{E}_q\left[\left(\nabla_\mathbf{m}\log\left(p(\mathbf{m})\right)\right)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{m}) + trace\left(\nabla_\mathbf{m}\boldsymbol{\phi}(\mathbf{m})\right)\right] \\
&= -\mathrm{E}_q\left[trace\left(\nabla_\mathbf{m}\log\left(p(\mathbf{m})\right)\boldsymbol{\phi}(\mathbf{m})^{\mathbf{T}}\right) + trace\left(\nabla_\mathbf{m}\boldsymbol{\phi}(\mathbf{m})\right)\right] \\
&= -\mathrm{E}_q\left[trace\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m})\right)\right]
\end{aligned}$$

where $\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_\mathbf{m}\log p(\mathbf{m})\boldsymbol{\phi}(\mathbf{m})^T + \nabla_\mathbf{m}\boldsymbol{\phi}(\mathbf{m})$ is the Stein operator.