REPORT
doi:10.26443/seismica.v3i1.1143

VIP - Variational Inversion Package with example implementations of Bayesian tomographic imaging

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

[1]School of Engineering and Technology, China University of Geosciences, Beijing, China, [2]School of GeoSciences, University of Edinburgh, UK

Author contributions: *Conceptualization*: Xin Zhang, Andrew Curtis. *Writing - Original draft*: Xin Zhang.

Abstract Bayesian inference has become an important methodology to solve inverse problems and to quantify uncertainties in their solutions. Variational inference is a method that provides probabilistic, Bayesian solutions efficiently by using optimisation. In this study we present a Python Variational Inversion Package (VIP), to solve inverse problems using variational inference methods. The package includes automatic differential variational inference (ADVI), Stein variational gradient descent (SVGD) and stochastic SVGD (sSVGD), and provides implementations of 2D travel time tomography and 2D full waveform inversion including test examples and solutions. Users can solve their own problems by supplying an appropriate forward function and a gradient calculation code. In addition, the package provides a scalable implementation which can be deployed easily on a desktop machine or using modern high performance computational facilities. The examples demonstrate that VIP is an efficient, scalable, extensible and user-friendly package, and can be used to solve a wide range of low or high dimensional inverse problems in practice.

Production Editor:
Gareth Funning
Handling Editor:
Paula Koelemeijer
Copy & Layout Editor:
Hannah F. Mark

Received:
October 27, 2023
Accepted:
May 2, 2024
Published:
May 29, 2024

1 Introduction

In a variety of academic and practical applications that concern the Earth's subsurface we wish to find answers to specific scientific questions. In the geosciences this is often achieved by imaging subsurface properties using data recorded on the surface, and by interpreting those images to address questions of interest. The subsurface is usually parameterised in some way, and a physical relationship is defined that predicts data that would be recorded for any particular set of model parameters, while the inverse relationship can not be determined uniquely. Once real data have been observed, the imaging problem is thus established as an inverse problem (Tarantola, 2005).

Because of non-linearity in the physical relationship, insufficient data coverage and noise in the data, inverse problems almost always have non-unique solutions: many sets of parameter values can fit the data to within their uncertainty. It is therefore important to characterize the family of possible solutions (in other words, the solution uncertainty) in order to interpret the results with the correct level of confidence, and to provide well-justified and robust answers to the scientific questions (Arnold and Curtis, 2018).

Solutions to an inverse problem are often found by seeking an optimal set of parameter values that minimizes the difference or misfit between observed data and model-predicted data to within the data noise. Since most inverse problems have non-unique solutions, some form of regularization is often imposed on the parameters in order to make the computational solution unique (Aki and Lee, 1976; Tarantola, 2005; Aster et al., 2018). Many codes have been developed using this class of methods (Rawlinson, 2005; Rücker et al., 2017; Afanasiev et al., 2019; Wathelet et al., 2020; Komatitsch et al., 2023). However, since regularization is often chosen using ad-hoc criteria, these methods produce deliberately biased results, and valuable information can be concealed in the process (Zhdanov, 2002). Moreover, no such optimisation method can provide accurate estimates of uncertainty. To overcome these issues, the SOLA-Backus-Gilbert inversion method has recently been applied to large scale linearised tomographic problems. This method evaluates the weighted average of the true model parameters and provides both resolution and uncertainty estimates (Zaroli, 2016; Zaroli et al., 2017). In addition, the method does not require regularization and can be conducted in a parameter-free way which avoids bias caused by parameterisation (Zaroli, 2019). Unfortunately, the method is only developed for linear problems; since most Geophysical problems are significantly nonlinear, our goal is to provide methods that estimate solutions and uncertainties for that class of problems.

Bayesian inference solves both linear and nonlinear inverse problems by updating a *prior* probability density function (pdf) with new information contained in the data to produce a *posterior* pdf which describes the full state of information about the parameters post inversion (Tarantola, 2005). If we define the prior pdf as $p(\mathbf{m})$, the posterior pdf $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ can be computed using Bayes' theorem:

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

*Corresponding author: xzhang@cugb.edu.cn

where $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ is the *likelihood* function which describes the probability of observing the recorded data $\mathbf{d}_{\text{obs}}$ if model parameters took the values in $\mathbf{m}$, and $p(\mathbf{d}_{\text{obs}})$ is a normalization factor called the *evidence*. This posterior pdf describes the full uncertainty in parameter values by combining the prior information and the uncertainty contained in the data.

Markov chain Monte Carlo (McMC) is one commonly-used method to solve Bayesian inference problems and has been used widely in many fields. The method constructs a set (chain) of successive samples that are distributed according to the posterior pdf by performing a structured random walk through parameter space (Brooks et al., 2011); thereafter, these samples can be used to estimate statistical information about parameters in the posterior pdf (Mosegaard and Tarantola, 1995; Tarantola, 2005) and to find answers to specific scientific questions (Arnold and Curtis, 2018; Siahkoohi et al., 2022b; Zhang and Curtis, 2022; Zhao et al., 2022b; McKean et al., 2023). The Metropolis-Hastings algorithm is one such method that originates from physics (Metropolis and Ulam, 1949; Hastings, 1970), and has been applied to a range of geophysical inverse problems (Mosegaard and Tarantola, 1995; Malinverno et al., 2000; Andersen et al., 2001; Mosegaard and Sambridge, 2002; Sambridge and Mosegaard, 2002; Ramirez et al., 2005; Gallagher et al., 2009). However, the algorithm becomes inefficient in high dimensional space because of poor scaling due to its random walk behaviour.

In order to solve Bayesian inference problems more efficiently, a variety of more advanced methods have been introduced to geophysics, such as reversible-jump McMC (Green, 1995; Malinverno, 2002; Bodin and Sambridge, 2009; Galetti et al., 2015; Zhang et al., 2018b), Hamiltonian Monte Carlo (Duane et al., 1987; Sen and Biswas, 2017; Fichtner et al., 2018; Gebraad et al., 2020), Langevin Monte Carlo (Roberts et al., 1996; Siahkoohi et al., 2020), stochastic Newton McMC (Martin et al., 2012; Zhao and Sen, 2019), and parallel tempering (Hukushima and Nemoto, 1996; Dosso et al., 2012; Sambridge, 2013). Gaussian process models have also been used to solve linearised probabilistic problems (Valentine and Sambridge, 2020). Based on these studies a range of methods and codes have been developed to solve geophysical inverse problems using McMC (Bodin and Sambridge, 2009; Shen et al., 2012; Hawkins and Sambridge, 2015; Zhang et al., 2018b; Zunino et al., 2023). Nevertheless, these papers mainly address 1D, 2D or sparsely-parametrised 3D spatial imaging problems; Bayesian solutions to large scale problems (e.g., those involving thousands of parameters to be estimated) remain intractable because of their unaffordable computational cost due to the curse of dimensionality (Curtis and Lomax, 2001).

In an attempt to improve the efficiency of Bayesian inference for certain types of problems, variational inference has been introduced to geophysics as an alternative to McMC. In variational inference one seeks a best approximation to the posterior pdf within a predefined family of (simplified) probability distributions by minimizing the difference between the approximating pdf and the posterior pdf (Bishop, 2006; Blei et al., 2017). One commonly-used measure of the difference between the pdfs is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) as it is easier to estimate computationally than other measures. Variational inference therefore solves Bayesian inference problems by minimizing the KL divergence, which is an optimisation rather than a stochastic sampling problem. The method has been demonstrated to be computationally more efficient and more scalable to high dimensionality in some classes of problems (Bishop, 2006; Zhang et al., 2018a). The method can also be applied to large datasets by dividing the data set into random minibatches and using stochastic and distributed optimisation (Robbins and Monro, 1951; Kubrusly and Gravier, 1973). By contrast, the same strategy cannot easily be used for McMC because it breaks the detailed balance condition required by most McMC methods (O'Hagan and Forster, 2004). In addition, variational inference methods can usually be parallelized at the individual sample level, whereas in McMC this cannot be achieved because of dependence between successive samples.

Variational inference has been applied to a range of geophysical inverse problems. Nawaz and Curtis (2018) used *mean-field* variational inference to invert for subsurface geological facies distributions and petrophysical properties using seismic data, with further developments by Nawaz and Curtis (2019) and Nawaz et al. (2020). Although these methods are computationally efficient, the mean-field approximation ignores correlations between parameters, and the methods of Nawaz and Curtis involved the development of bespoke mathematical derivations and implementations for each class of problem. While these developments result in exceptional speed of calculation, this approach restricts the method to a small range of problems for which correlations are not important and the derivations can be performed (Parisi, 1988; Bishop, 2006; Blei et al., 2017). To extend variational inference to general inverse problems, Kucukelbir et al. (2017) used a Gaussian family in variational inference to create a method called automatic differential variational inference (ADVI), which has been applied to travel time tomography (Zhang and Curtis, 2020a) and earthquake slip inversion (Zhang and Chen, 2022), and extended to the family of sums (mixtures) of multiple Gaussians by Zhao and Curtis (2024). By using a sequence of invertible and differential transforms (called normalizing flows), Rezende and Mohamed (2015) proposed normalizing flow variational inference in which flows (functions, or simply, relationships) are designed which convert a simple initial distribution to an arbitrarily complex distribution that approximates the posterior pdf. In geophysics and related fields the method has been applied to travel time tomography (Zhao et al., 2022a), seismic imaging (Siahkoohi et al., 2020, 2022a), seismic data interpolation (Kumar et al., 2021), transcranial ultrasound tomography (Orozco et al., 2023) and cascading hazards estimation (Li et al., 2023).

By using a set of samples of parameter values (called particles) to represent the density of an approximating pdf, Liu and Wang (2016) introduced a method called Stein variational gradent descent (SVGD), which itera-

tively updates those particles by minimizing the KL divergence so that the final particle density provides an approximation to the posterior pdf. SVGD has been demonstrated to be an efficient method in a range of geophysical applications, such as travel time tomography (Zhang and Curtis, 2020a), full waveform inversion (FWI) (Zhang and Curtis, 2020b, 2021; Lomas et al., 2023; Wang et al., 2023), earthquake source inversion (Smith et al., 2022), hydrogeological inversion (Ramgraber et al., 2021), post-stack seismic inversion (Izzatullah et al., 2023) and neural network based seismic tomography (Agata et al., 2023). However the method becomes inefficient and inaccurate in high dimensional problems because of the finite number of particles and the practical limitation of computational cost (Ba et al., 2022). To reduce this issue, Gallego and Insua (2018) introduced the stochastic SVGD (sSVGD) method by combining SVGD and McMC: the efficiency of this method has recently been demonstrated when it was used to estimate the first Bayesian solution for a fully nonlinear, 3D FWI problem (Zhang et al., 2023).

Despite these theoretical and practical advances, variational inference has not been widely used in geophysics. This is partly because the method is not easily accessible to non-specialists, and also because there is no common code framework to perform geophysical inversions using the method. In this study we therefore present a Python variational inversion package (VIP), which includes ADVI, SVGD and sSVGD, to make it more straightforward to solve geophysical inverse problems using variational inference methods. The package provides complete implementations of 2D travel time tomography and 2D full waveform inversion problems, including test results for users to check that their implementation is correct. Users can also solve other inverse problems by supplying their own forward functions and gradient calculation codes. In addition, to solve large inverse problems the package is designed in a scalable way such that it can be deployed on a desktop computer as well as in modern high performance computational (HPC) facilities.

In the following section we describe the concept of variational inference, and algorithmic details of ADVI, SVGD and sSVGD. In section 3 we provide an overview of the VIP package, and in section 4 we demonstrate VIP using examples of 2D travel time tomography and 2D full waveform inversion. We thus show that VIP is an efficient, scalable, extensible and user-friendly package that will enable users to solve geophysical inverse problems using variational methods. Making these methods more tractable for practitioners should allow them to be tested on a wide range of problems.

2 Theoretical background

2.1 Variational inference

Variational inference solves Bayesian inference problems using optimisation. To achieve this, we first define a simplified family of pdf's $Q = \{q(\mathbf{m})\}$, for example, the family of all Gaussian distributions. The method then seeks an optimal approximation $q^*(\mathbf{m})$ to the posterior probability distribution $p(\mathbf{m}|\mathbf{d}_{\rm obs})$ within this family by minimizing the KL divergence between $q(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d}_{\rm obs})$:

$$q^*(\mathbf{m}) = \arg\min_{q\in Q} \mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\rm obs})] \tag{2}$$

The KL divergence measures the difference between two probability distributions:

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

where $\log p(\mathbf{m},\mathbf{d}_{\rm obs})$ is the joint distribution of model $\mathbf{m}$ and data $\mathbf{d}_{\rm obs}$. The expectations are calculated with respect to the known pdf q, and we have used Bayes' theorem to expand the posterior pdf $p(\mathbf{m}|\mathbf{d}_{\rm obs})$ in the second line of equation (3). It can be shown that the KL divergence is non-negative and only equals zero when $q(\mathbf{m}) = p(\mathbf{m}|\mathbf{d}_{\rm obs})$ (Kullback and Leibler, 1951). Because the evidence term $\log p(\mathbf{d}_{\rm obs})$ is computationally intractable, the KL divergence cannot be calculated directly. We therefore rearrange the above equation by moving the evidence term and the KL divergence onto the same side:

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

Given that the KL divergence is non-negative, the left-hand side defines a lower bound on the evidence, which is therefore called the evidence lower bound (ELBO):

$$\begin{aligned}\mathrm{ELBO}[q] &= \log p(\mathbf{d}_{\rm obs}) - \mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\rm obs})] \\ &= \mathrm{E}_q[\log p(\mathbf{m},\mathbf{d}_{\rm obs})] - \mathrm{E}_q[\log q(\mathbf{m})]\end{aligned} \tag{5}$$

The latter epxression can be estimated in practice using numerical methods because it does not involve the intractable evidence term. Since the evidence $\log p(\mathbf{d}_{\rm obs})$ is a constant for a specific problem, minimizing the KL-divergence is equivalent to maximizing the ELBO. Variational inference in equation (2) can therefore be expressed as:

$$q^*(\mathbf{m}) = \arg\max_{q\in Q} \mathrm{ELBO}[q(\mathbf{m})] \tag{6}$$

In variational inference, the choice of the variational family Q is important because it determines both the accuracy of the approximation and the complexity of the optimisation problem. A good choice should be a family which is rich enough to approximate the posterior pdf accurately or at least provides the information that we seek, but simple enough such that the optimisation problem is tractable. Different choices of family may also allow different types of algorithm to be developed. In the VIP package we implement three different algorithms, ADVI, SVGD and sSVGD to solve inverse problems.

2.2 Automatic differential variational inference (ADVI)

ADVI uses the family of (transformed) Gaussians to solve variational inference problems (Kucukelbir et al.,

2017). The transform arises because physical model parameters describe quantities that often have hard bounds, while Gaussian variables have infinite support. We therefore first transform the physical parameters into an unconstrained space using an invertible transform $T : \boldsymbol{\theta} = T(\mathbf{m})$. In this unconstrained space the joint distribution $p(\mathbf{m}, \mathbf{d}_{\text{obs}})$ becomes:

$$p(\boldsymbol{\theta}, \mathbf{d}_{\text{obs}}) = p(\mathbf{m}, \mathbf{d}_{\text{obs}})|\text{det}\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})| \tag{7}$$

where $\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})$ is the Jacobian matrix of the inverse of T which accounts for the effects of changes in hypervolume between the unconstrained and constrained parameter spaces. In this unconstrained space define a Gaussian variational family

$$q(\boldsymbol{\theta}; \boldsymbol{\zeta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8}$$

where $\boldsymbol{\zeta}$ represents variational parameters, that is, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. To ensure that the covariance matrix $\boldsymbol{\Sigma}$ is positive semi-definite, we use a Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ where $\mathbf{L}$ is a lower triangular matrix, to reparameterise $\boldsymbol{\Sigma}$.

With the above definition, the variational problem in equation (6) becomes:

$$\begin{aligned}\boldsymbol{\zeta}^* &= \arg\max_{\boldsymbol{\zeta}} \text{ELBO}[q(\boldsymbol{\theta}; \boldsymbol{\zeta})] \\ &= \arg\max_{\boldsymbol{\zeta}} \text{E}_q[\log p(\boldsymbol{\theta}, \mathbf{d}_{\text{obs}})] - \text{E}_q[\log q(\boldsymbol{\theta}; \boldsymbol{\zeta})] \\ &= \arg\max_{\boldsymbol{\zeta}} \text{E}_q[\log p\big(T^{-1}(\boldsymbol{\theta}), \mathbf{d}_{\text{obs}}\big) + \log|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|] \\ &\quad - \text{E}_q[\log q(\boldsymbol{\theta}; \boldsymbol{\zeta})]\end{aligned} \tag{9}$$

This optimisation problem can be solved by using a gradient ascent algorithm. As shown in Kucukelbir et al. (2017), the gradients of the ELBO with respect to variational parameters $\boldsymbol{\mu}$ and $\mathbf{L}$ can be calculated using:

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\text{ELBO} = \text{E}_{N(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\big[&\nabla_{\mathbf{m}}\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta}) \\ &+ \nabla_{\boldsymbol{\theta}}\log|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\big]\end{aligned} \tag{10}$$

$$\begin{aligned}\nabla_{\mathbf{L}}\text{ELBO} = \text{E}_{N(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\big[&\big(\nabla_{\mathbf{m}}\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta}) \\ &+ \nabla_{\boldsymbol{\theta}}\log|det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\big)\boldsymbol{\eta}^{\mathrm{T}}\big] + (\mathbf{L}^{-1})^{\mathrm{T}}\end{aligned} \tag{11}$$

where $\boldsymbol{\eta}$ is a random variable distributed according to the standard normal distribution $N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})$. The expectations can be estimated using Monte Carlo (MC) integration, which in practice only requires a low number of samples because the optimisation is performed over many iterations so that statistically the gradients will lead to convergence towards the correct solution (Kucukelbir et al., 2017). The variational problem in equation (9) can now be solved by using gradient ascent methods. In the VIP package we implement four optimisation algorithms: stochastic gradient descent (SGD), ADAGRAD (Duchi et al., 2011), ADADELTA (Zeiler, 2012) and ADAM (Kingma and Ba, 2014). The final approximation to the Bayesian solution can be obtained by transforming $q(\boldsymbol{\theta}; \boldsymbol{\zeta}^*)$ back to the original space.

For transform T we implement a commonly-used logarithmic transform (Team et al., 2016; Zhang and Curtis, 2020a)

$$\begin{aligned}\theta_i &= T(m_i) = \log(m_i - a_i) - \log(b_i - m_i) \\ m_i &= T^{-1}(\theta_i) = a_i + \frac{(b_i - a_i)}{1 + exp(-\theta_i)}\end{aligned} \tag{12}$$

where m_i and θ_i represent the i^{th} parameter in the original and transformed space respectively, and a_i and b_i are the lower and upper bound on m_i. The final approximation obtained using ADVI is therefore limited in complexity by the Gaussian distribution $q(\boldsymbol{\theta}; \boldsymbol{\zeta}^*)$ and the transform T. Note that if no transform is performed, the method approximates the posterior pdf using a Gaussian distribution directly.

2.3 Stein variational gradient descent (SVGD)

Instead of using a specific form of pdf (for example, the Gaussian distribution in ADVI) in variational inference, it is also possible to use the density of a set of samples to represent the approximating probability distribution. SVGD is one such method in which the set of samples are called particles. In SVGD those particles are iteratively updated by minimizing the KL divergence so that the density of the final set of particles is distributed according to the posterior probability distribution. If we define the set of particles as $\{\mathbf{m}_i\}$, SVGD updates each particle using a smooth transform:

$$T(\mathbf{m}_i) = \mathbf{m}_i + \epsilon\boldsymbol{\phi}(\mathbf{m}_i) \tag{13}$$

where $\mathbf{m}_i$ is the i^{th} particle, $\boldsymbol{\phi}(\mathbf{m}_i)$ is a smooth vector function which describes the perturbation direction, and ϵ is the magnitude of the perturbation. When ϵ is sufficiently small, the transform is invertible since the Jacobian of the transform is close to an identity matrix. Denote $q(\mathbf{m})$ as the pdf represented by the set of particles, and $q_T(\mathbf{m})$ as the transformed probability distribution of $q(\mathbf{m})$ using equation (13). In order to reduce the KL divergence between $q_T(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$, we first calculate the gradient of the KL divergence with respect to ϵ, which is found to be (Liu and Wang, 2016):

$$\nabla_{\epsilon}\text{KL}[q_T||p]\,|_{\epsilon=0} = -\text{E}_q\left[trace\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m})\right)\right] \tag{14}$$

where $\mathcal{A}_p$ is the Stein operator defined as $\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_{\mathbf{m}}\log p(\mathbf{m}|\mathbf{d}_{\text{obs}})\boldsymbol{\phi}(\mathbf{m})^T + \nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})$. This equation implies that one can obtain the steepest descent direction of the KL-divergence by maximizing the right-hand expectation $\text{E}_q\left[trace\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m})\right)\right]$, and consequently the KL divergence can be reduced by stepping a small distance in that direction. Iteratively re-calculating equation (14) and stepping in each revised direction locates a minimum in the KL divergence.

The optimal direction $\boldsymbol{\phi}^*$ that maximizes the expectation $\text{E}_q\left[trace\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m})\right)\right]$ in equation (14) can be found using kernels. Assume $x, y \in X$ and define a mapping ϕ from X to a space where an inner product $\langle , \rangle$ is defined (called a Hilbert space); a *kernel* is a function that satisfies $k(x, y) = \langle\phi(x), \phi(y)\rangle$. Given a kernel function $k(\mathbf{m}', \mathbf{m})$, the optimal $\boldsymbol{\phi}^*$ can be calculated using (see details in Liu and Wang, 2016):

$$\boldsymbol{\phi}^* \propto \text{E}_{\{\mathbf{m}'\sim q\}}[\mathcal{A}_p k(\mathbf{m}', \mathbf{m})] \tag{15}$$

In the VIP package, we implement a commonly-used kernel function, the radial basis function (RBF):

$$k(\mathbf{m}, \mathbf{m}') = \exp[-\frac{\|\mathbf{m} - \mathbf{m}'\|^2}{2h^2}] \tag{16}$$

where h is a scale factor that controls the magnitude of similarity between the two particles based on their distance apart. Given equations (14) and (15), the KL divergence can be minimized by iteratively applying the transform in equation (13) with the optimal $\boldsymbol{\phi}^*$ to a set of initial particles:

$$\mathbf{m}_i^{l+1} = T(\mathbf{m}_i^l) = \mathbf{m}_i^l + \epsilon^l \boldsymbol{\phi}_l^*(\mathbf{m}_i^l) \tag{17}$$

where l represents the l^{th} iteration. Note that the expectation in equation (15) can be estimated using the particles' mean value, so we can compute $\boldsymbol{\phi}_l^*$ using:

$$\begin{aligned}\boldsymbol{\phi}_l^*(\mathbf{m}) &= \frac{1}{n}\sum_{j=1}^{n}\left[\mathcal{A}_p k(\mathbf{m}_j^l, \mathbf{m})\right] \\ &= \frac{1}{n}\sum_{j=1}^{n}\left[k(\mathbf{m}_j^l, \mathbf{m})\nabla_{\mathbf{m}_j^l}\log p(\mathbf{m}_j^l|\mathbf{d}_{\text{obs}})\right. \\ &\left. + \nabla_{\mathbf{m}_j^l} k(\mathbf{m}_j^l, \mathbf{m})\right]\end{aligned} \tag{18}$$

where n is the number of particles. For sufficiently small $\{\epsilon^l\}$ the transform is invertible, and the process converges to the posterior distribution asymptotically as $n \to \infty$ (Liu and Wang, 2016). Note that even though the posterior distribution $p(\mathbf{m}_j^l|\mathbf{d}_{\text{obs}})$ is unknown in practice, we can always calculate its value up to an unknown constant for a specific model. As a result, its gradient $\nabla_{\mathbf{m}_j^l}\log p(\mathbf{m}_j^l|\mathbf{d}_{\text{obs}})$ can be obtained, and hence the $\boldsymbol{\phi}_l^*$.

The first term in equation (15) is the kernel weighted average of gradients of the posterior pdf from all particles, and drives particles toward high probability areas. For the RBF kernel the second term becomes $\sum_j \frac{\mathbf{m}-\mathbf{m}_j}{\sigma^2}k(\mathbf{m}_j, \mathbf{m})$ which move particles away from its neighbouring particles. This term therefore acts as a repulsive force that prevents particles from collapsing to a single mode. SVGD balances the drive towards high probabilities and the repulsive force such that the density of particles moves towards the posterior pdf.

Note that the scale factor h in the RBF kernel controls the weighting value of particles. As suggested in several studies (Liu and Wang, 2016; Zhang and Curtis, 2020a), we take h as $\tilde{d}/\sqrt{2\log n}$ where $\tilde{d}$ is the median of pairwise distances between all particles. This choice enables that for particle $\mathbf{m}_i$ the contribution form its own gradient is balanced from all other particles as $\sum_{j\neq i} k(\mathbf{m}_i, \mathbf{m}_j) \approx n\exp(-\frac{1}{2h^2}\tilde{d}^2) = 1$. If $h \to 0$, the method reduces to independent gradient ascent for each particle.

In SVGD the accuracy of estimation increases with the number of particles. For one single particle the method becomes a standard gradient ascent method toward the model with maximum a posterior (MAP) pdf value. This implies that even for a small number of particles SVGD can still produce an accurate parameter estimate as MAP estimation has been demonstrated to be an effective method in practice. Thus, in practice, one can start from a small number of particles and gradually increase the particles to produce more accurate estimates of the uncertainty.

2.4 Stochastic SVGD

Although SVGD has been applied in many fields (Gong et al., 2019; Zhang and Curtis, 2020a; Pinder et al., 2020; Ramgraber et al., 2021; Ahmed et al., 2022), the method can produce biased results in high dimensional problems because of the finite number of particles and the limitation of computational cost in practice (Ba et al., 2022). In order to further improve accuracy of the method, Gallego and Insua (2018) proposed a variant of SVGD, called stochastic SVGD (sSVGD), which combines SVGD and McMC by adding a Gaussian noise term to the dynamics of SVGD. By doing this sSVGD becomes an McMC method with multiple interacting Markov chains, and since every set of particle values can be regarded as a sample of the posterior pdf, the method can generate many samples that are distributed according to the posterior pdf. Under certain conditions (see below), sSVGD guarantees asymptotic convergence to the posterior pdf as the number of iterations tends to infinity, which standard SVGD with a finite number of particles cannot achieve. As a result sSVGD can produce more accurate results than the SVGD method, provided that the number of iterations is sufficient to remove effects of the distribution of samples near the start of the chain (the so-called burn-in period) (Gallego and Insua, 2018; Zhang et al., 2023).

To introduce sSVGD, we start from a stochastic differential equation (SDE). For a random variable $\mathbf{z}$, the SDE is defined as:

$$d\mathbf{z} = \mathbf{f}(\mathbf{z})dt + \sqrt{2\mathbf{D}(\mathbf{z})}d\mathbf{W}(t) \tag{19}$$

where $\mathbf{f}(\mathbf{z})$ is called the drift, $\mathbf{W}(t)$ is a Wiener process, and $\mathbf{D}(\mathbf{z})$ represents a positive semidefinite diffusion matrix. All continuous Markov process can be expressed as an SDE, and consequently one can construct a Markov chain by simulating the SDE (Oksendal, 2013). Assume $p(\mathbf{z})$ as the posterior distribution, an SDE that converges to the $p(\mathbf{z})$ can be constructed as (Ma et al., 2015):

$$\mathbf{f}(\mathbf{z}) = [\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})]\nabla\log p(\mathbf{z}) + \Gamma(\mathbf{z}) \tag{20}$$

where $\mathbf{Q}(\mathbf{z})$ is a skew-symmetric curl matrix, and $\Gamma_i(\mathbf{z}) = \sum_{j=1}^{d}\frac{\partial}{\partial \mathbf{z}_j}(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z}))$. To simulate this process, we can discretize the above equation using the Euler-Maruyama discretization:

$$\begin{aligned}\mathbf{z}_{t+1} = \mathbf{z}_t &+ \epsilon_t\left[\left(\mathbf{D}\left(\mathbf{z}_t\right) + \mathbf{Q}(\mathbf{z}_t)\right)\nabla\log p(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)\right] \\ &+ \mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{D}(\mathbf{z}_t))\end{aligned} \tag{21}$$

where $\mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{D}(\mathbf{z}_t))$ represents a Gaussian distribution with covariance $2\epsilon_t\mathbf{D}(\mathbf{z}_t)$. The gradient $\nabla\log p(\mathbf{z}_t)$ can be computed using the full data set, or using uniformly randomly selected minibatch data subsets which results in a stochastic gradient approximation. In either case the above process converges to the posterior distribution asymptotically as $\epsilon_t \to 0$ and $t \to \infty$ (Ma et al.,

2015). Matrices $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ can be adjusted to obtain faster convergence to the posterior distribution. For example, if we set $\mathbf{D} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$, one obtains stochastic gradient Langevin dynamics (Welling and Teh, 2011). If we construct an augmented space $\overline{\mathbf{z}} = (\mathbf{z}, \mathbf{x})$ by concatenating a moment term $\mathbf{x}$ to the state space $\mathbf{z}$, and set $\mathbf{D} = \mathbf{0}$ and $\mathbf{Q} = \begin{pmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}$ then the stochastic Hamiltonian Monte Carlo method can be derived (Chen et al., 2014).

In sSVGD we define an augmented space $\mathbf{z} = (\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_n)$ by concatenating the set of particles $\{\mathbf{m}_i\}$, and use equation (21) to generate samples from the posterior distribution $p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{m}_i|\mathbf{d}_{\text{obs}})$. Define a matrix $\mathbf{K}$

$$\mathbf{K} = \frac{1}{n}\begin{bmatrix} k(\mathbf{m}_1, \mathbf{m}_1)\mathbf{I}_{d\times d} & \dots & k(\mathbf{m}_1, \mathbf{m}_n)\mathbf{I}_{d\times d} \\ \vdots & \ddots & \vdots \\ k(\mathbf{m}_n, \mathbf{m}_1)\mathbf{I}_{d\times d} & \dots & k(\mathbf{m}_n, \mathbf{m}_n)\mathbf{I}_{d\times d} \end{bmatrix} \tag{22}$$

where $k(\mathbf{m}_i, \mathbf{m}_j)$ is a kernel function defined in equation (16) and $\mathbf{I}_{d\times d}$ is an identity matrix. According to the definition of kernel functions, the matrix $\mathbf{K}$ is positive definite (Gallego and Insua, 2018). By setting $\mathbf{Q}(\mathbf{z}_t) = \mathbf{0}$ and $\mathbf{D}(\mathbf{z}_t) = \mathbf{K}$, we obtain the stochastic SVGD algorithm:

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \epsilon_t[\mathbf{K}\nabla \log p(\mathbf{z}_t) + \nabla \cdot \mathbf{K}] + \mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{K}) \tag{23}$$

Note that without the noise term $\mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{K})$, the above equation becomes the standard SVGD method – compare equations (23) with equation (18), repeated here:

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \epsilon_t[\mathbf{K}\nabla \log p(\mathbf{z}_t) + \nabla \cdot \mathbf{K}] \tag{24}$$

sSVGD is therefore an McMC method that uses the gradients from SVGD to produce successive samples. According to equation (20), this process converges to $p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{m}_i|\mathbf{d}_{\text{obs}})$ asymptotically. Note that when n is sufficiently large, the noise term $\mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{K})$ becomes arbitrarily small. In such cases sSVGD and SVGD produce the same results.

The process defined in equation (23) requires samples to be generated from the distribution $\mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{K})$. In order to perform this efficiently, we first define a matrix $\mathbf{D_K}$

$$\mathbf{D_K} = \frac{1}{n}\begin{bmatrix} \overline{\mathbf{K}} & & \\ & \ddots & \\ & & \overline{\mathbf{K}} \end{bmatrix} \tag{25}$$

where $\overline{\mathbf{K}}$ is an $n \times n$ matrix with $\overline{\mathbf{K}}_{ij} = k(\mathbf{m}_i, \mathbf{m}_j)$. The matrix $\mathbf{D_K}$ can be constructed from $\mathbf{K}$ using $\mathbf{D_K} = \mathbf{P}\mathbf{K}\mathbf{P}^{\mathrm{T}}$ where $\mathbf{P}$ is a permutation matrix

$$\mathbf{P} = \left[\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ \hline 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ \hline \ddots & & \ddots & \ddots \\ \hline & 1 & & \\ & & 1 & \\ & & & \ddots \\ & & & 1 \end{array}\right] \tag{26}$$

The action of this permutation matrix on a vector $\mathbf{z}$ rearranges the order of the vector from the basis where the particles are listed sequentially to that where the first coordinates of all particles are listed, then the second, etc. With these definitions, a random sample $\boldsymbol{\eta}$ can be generated efficiently using

$$\begin{aligned} \boldsymbol{\eta} &\sim \mathcal{N}(\mathbf{0}, 2\epsilon_t\mathbf{K}) \\ &\sim \sqrt{2\epsilon_t}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathcal{N}(\mathbf{0}, \mathbf{K}) \\ &\sim \sqrt{2\epsilon_t}\mathbf{P}^{\mathrm{T}}\mathcal{N}(\mathbf{0}, \mathbf{D_K}) \\ &\sim \sqrt{2\epsilon_t}\mathbf{P}^{\mathrm{T}}\mathbf{L_{D_K}}\mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{27}$$

where $\mathbf{L_{D_K}}$ is the lower triangular Cholesky decomposition of matrix $\mathbf{D_K}$. Taking into account the fact that $\mathbf{D_K}$ is a block-diagonal matrix, $\mathbf{L_{D_K}}$ can be computed easily as only the lower triangular Cholesky decomposition of matrix $\overline{\mathbf{K}}$ is required. In practice this calculation is computationally negligible because the number of particles n is usually modest (< 1000). One can now use equation (23) to generate samples from the posterior distribution.

3 Code overview

The VIP package implements the suite of variational methods to solve geophysical inverse problems using the Python programming language. The package includes a set of specific forward and inverse problems such as 2D travel time tomography and 2D full waveform inversion, and also allows users to provide their own forward functions. In variational inference one needs to compute the gradient of the posterior pdf with respect to model parameters. We use the adjoint method to calculate the gradient in the case of seismic full waveform inversion (Lions, 1971; Tarantola, 1984; Tromp et al., 2005; Fichtner et al., 2006; Plessix, 2006), and the ray tracing method in the case of travel time tomography (Rawlinson and Sambridge, 2004). For user-specified forward problems it is required that users implement their own function that computes gradients.

The prior pdf is important in Bayesian inference as it provides information about model parameters independent of the data. The VIP package provides two commonly-used prior distributions: Uniform and Gaussian pdf's (note that these are only used as prior pdf's, and do not place any additional constraints on the variational families described above). To implement the Uniform distribution we employ two strategies. In the first strategy we impose hard constraints on model parameters, that is, for any parameter that assumes a value outside the distribution we reset the value to be the closest limit. Note that a similar strategy cannot be used in ADVI as the method assumes a Gaussian variational family which cannot be defined in a constrained space. The second strategy involves using equation (12) to transform model parameters into an unconstrained space and perform variational inversion in that space, which provides a more flexible way to employ a variety of variational families. In addition, users can provide their own prior distributions by implementing an appropriate pdf function (see details in the code documentation).

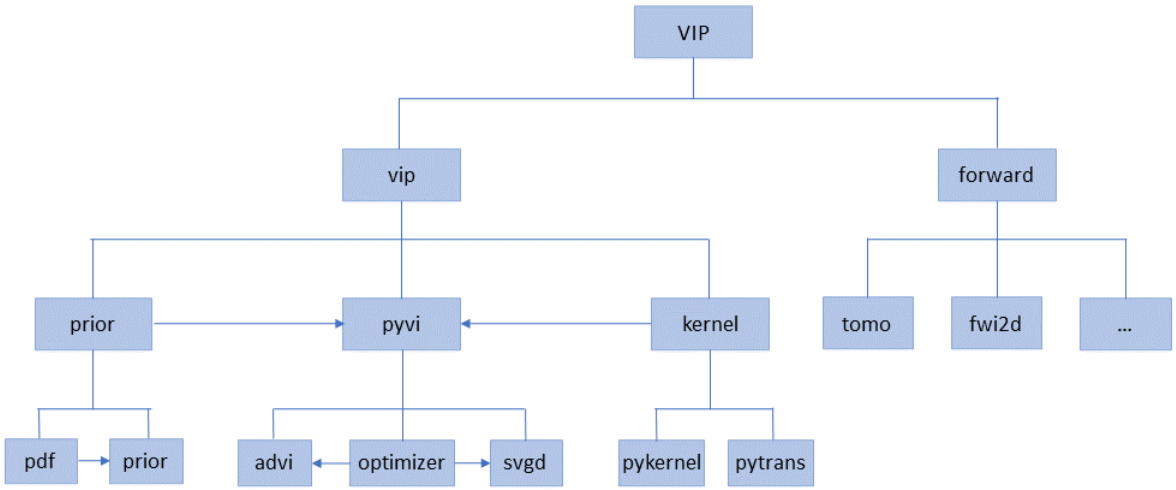

Figure 1 Code structure of VIP. Each rectangle represents a folder or file in the package. Users can implement their own forward functions similarly to the way this is implemented in examples *tomo* and *fwi2d*.

Python is a popular high-level interpreted programming language which suffers from slow execution for computationally intensive numerical simulations. We therefore implement time-consuming components of the code (e.g., the forward modelling functions) using Fortran and produce compiled C extensions for these codes using the Cython framework (Behnel et al., 2010). By doing this the code achieves C-like speeds. To further improve efficiency of the code, we use a Python library called Dask, which is designed for parallel and distributed computing, to parallelize the forward computation at the sample (particle) level (Rocklin et al., 2015). The package therefore provides an efficient, scalable and user-friendly implementation which can be deployed on a desktop as well as modern high performance computation facilities. Our aim is to implement a framework which can be used to solve various inverse problems, ranging from educational examples to complex, realistic studies.

Figure 1 shows the structure of VIP. The inversion code (*vip* in Figure 1) is implemented separately from forward modelling codes (*forward* in Figure 1), and only requires an interface of forward functions that returns logarithmic posterior pdf values and gradients (details can be found in the code documentation and in two examples *tomo* and *fwi2d*). Thus, users can easily combine their own forward functions with the package. In the *vip* code the prior distributions, kernel functions and variational algorithms are implemented in three different directories (*prior*, *kernel* and *pyvi* in Figure 1) so that the code can easily be extended to other prior pdfs, kernel functions and variational methods. For example, users can implement their own prior pdfs by adding a proper pdf function in the *pdf* code in the *prior* directory. Note that both SVGD and sSVGD methods are implemented in the *svgd* code.

4 Applications

4.1 Travel time tomography

As a first example we use the VIP package to solve a 2D tomographic problem. Specifically, we create Love wave group velocity maps of the British Isles using ambient seismic noise data recorded by 61 seismometers (blue triangles in Figure 2a). The geological setting and the main terrain boundaries of the British Isles are shown in Figure 2b. The ambient noise data were recorded in 2001-2003, 2006-2007 and in 2010 using three different subarrays. The two horizontal components of the data (N and E) were first rotated to the transverse and radial directions, and the obtained transverse data were cross correlated to produce Love waves between different station pairs. Travel times associated with group velocity at different periods between different station pairs are then estimated from those love waves. Details of the data processing procedures can be found in (Galetti et al., 2017). In this study we use a total number of 401 travel time measurements at 10 s period.

We parameterise the study region using a regular grid of 37 $\times$ 40 cells with a spacing of 0.33°in both longitude and latitude directions. The prior pdf for group velocity in each cell is set to be a Uniform distribution between 1.56 km/s to 4.8 km/s, of which the lower and upper bound were chosen to exceed the range of group velocities between all station pairs when assuming a great circle ray path (Zhao et al., 2022a). The likelihood function is chosen to be a Gaussian distribution to represent the data noise, which is estimated from independent travel time measurements by stacking randomly selected subsets of daily cross correlations (Galetti et al., 2017). In the inversion the predicted travel times are calculated using the fast marching method (Rawlinson and Sambridge, 2004).

We apply the above suite of methods to solve this tomographic problem, and compare the results with those obtained using the Metropolis-Hastings McMC

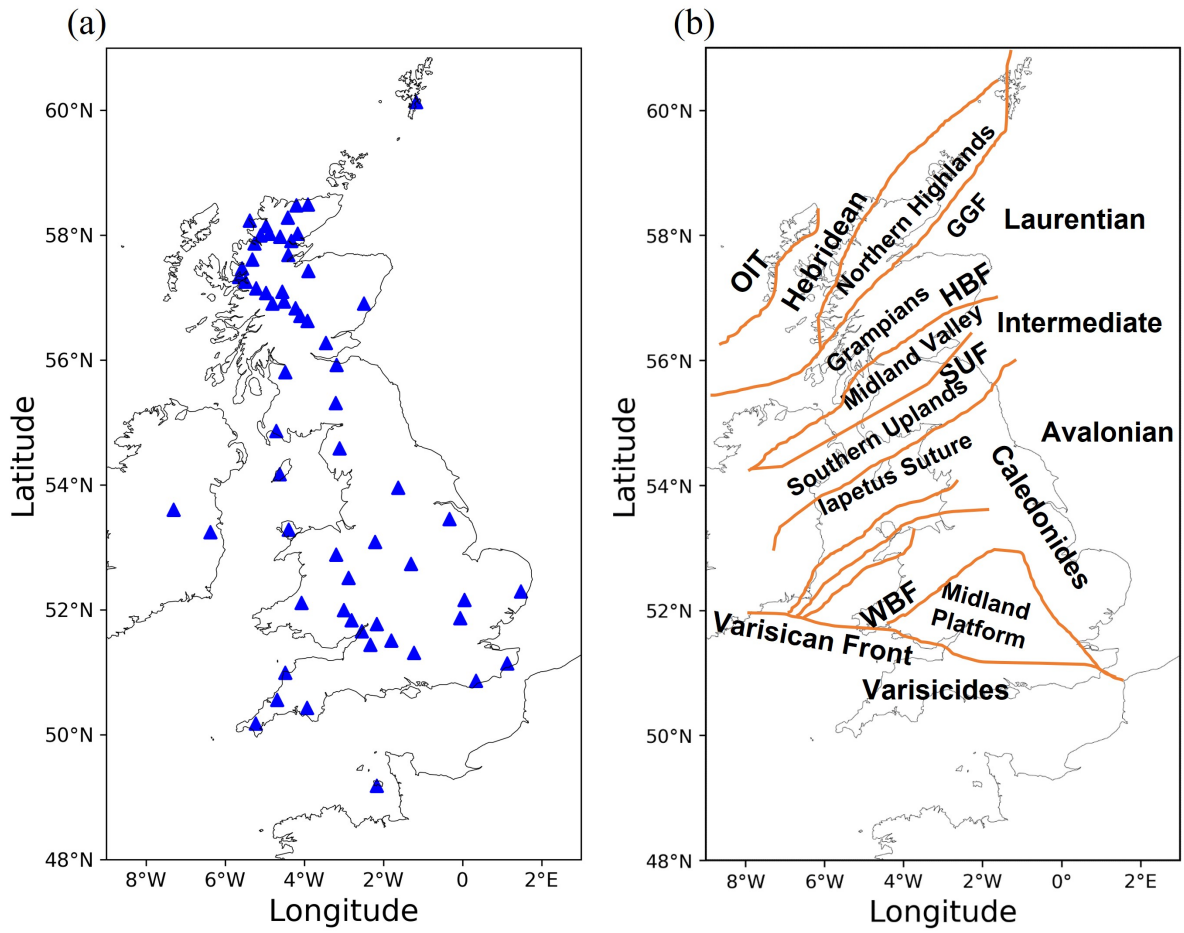

Figure 2 **(a)** Locations of seismometers (blue triangles) around British Isles used in this study. **(b)** Terrane boundaries in the British Isles from Galetti et al. (2017). Abbreviations are as follows: OIT, Outer Isles Thrust; GGF, Great Glen Fault; HBF, Highland Boundary Fault; SUF, Southern Uplands Fault; WBF, Welsh Borderland Fault System.

(MH-McMC) method (Zhao et al., 2022a). The Uniform prior distribution is implemented using the second strategy that transforms variables into an unconstrained space in variational inversions. For ADVI, we started the method with a standard Gaussian distribution in the unconstrained space, and performed 10,000 iterations at which point the misfit value ceases to decrease using the ADAM optimisation algorithm (Kingma and Ba, 2014). To visualize the results we generated 5,000 samples from the obtained Gaussian distribution and transformed them back to the original space to estimate posterior statistics. For SVGD, we generated 500 particles from the prior distribution and updated them using equation (18) for 3,000 iterations at which point the mean and standard deviation models became stable. The final particles are used to calculate the mean and standard deviation of the posterior distribution. For sSVGD, we started from 20 particles generated from the prior distribution, and updated them using equation (23) for 6,000 iterations after an additional burn-in period of 2,000 iteration, after which the average misfit value across all particles became approximately stationary. To reduce the memory and storage cost, we only retained samples every fourth iteration after the burn-in period, which results in a total of 30,000 samples.

Figure 3 shows the mean and standard deviation maps obtained using the suite of variational methods, as well as those obtained using the MH-McMC algorithm (Zhao et al., 2022a). Overall the results obtained using different methods show similar mean structures which have a good agreement with the known geology and previous tomographic studies in the British Isles (Nicolson et al., 2012, 2014; Galetti et al., 2017; Zhao et al., 2022a). For example, in the Scottish highlands the mean maps clearly exhibit high velocities (annotation 1 in Figure 3) which are consistent with the distribution of Lewisian and Dalradian complexes in this area. Similarly high velocities associated with the accretionary complex of the Southern Uplands (annotation 2) are clearly visible around 4°W, 55°N following a SW-NE trend. Between the Highland Boundary Fault and the Southern Uplands Fault a similar trend of low velocity zone (annotation 3) is found in the Midland Valley. Low velocities are also observed in a number of sedimentary basins such as the East Irish Sea (4.5°W, 54°E - annotation 4), the Cheshire Basin (2.5°W, 52.5°E - annotation 6), the Anglian-London Basin (0°, 52°N - annotation 7), the Weald Basin (0°, 51°N - annotation 8) and the Wessex Basin (3°W, 50.5°N - annotation 9). By contrast, high velocities can be found in granitic intrusion regions, for example, in northwest Wales (around 4°W, 53°N - annotation 5) and Cornwall (around 4.5°W, 50.5°N - annotation 10). More detailed

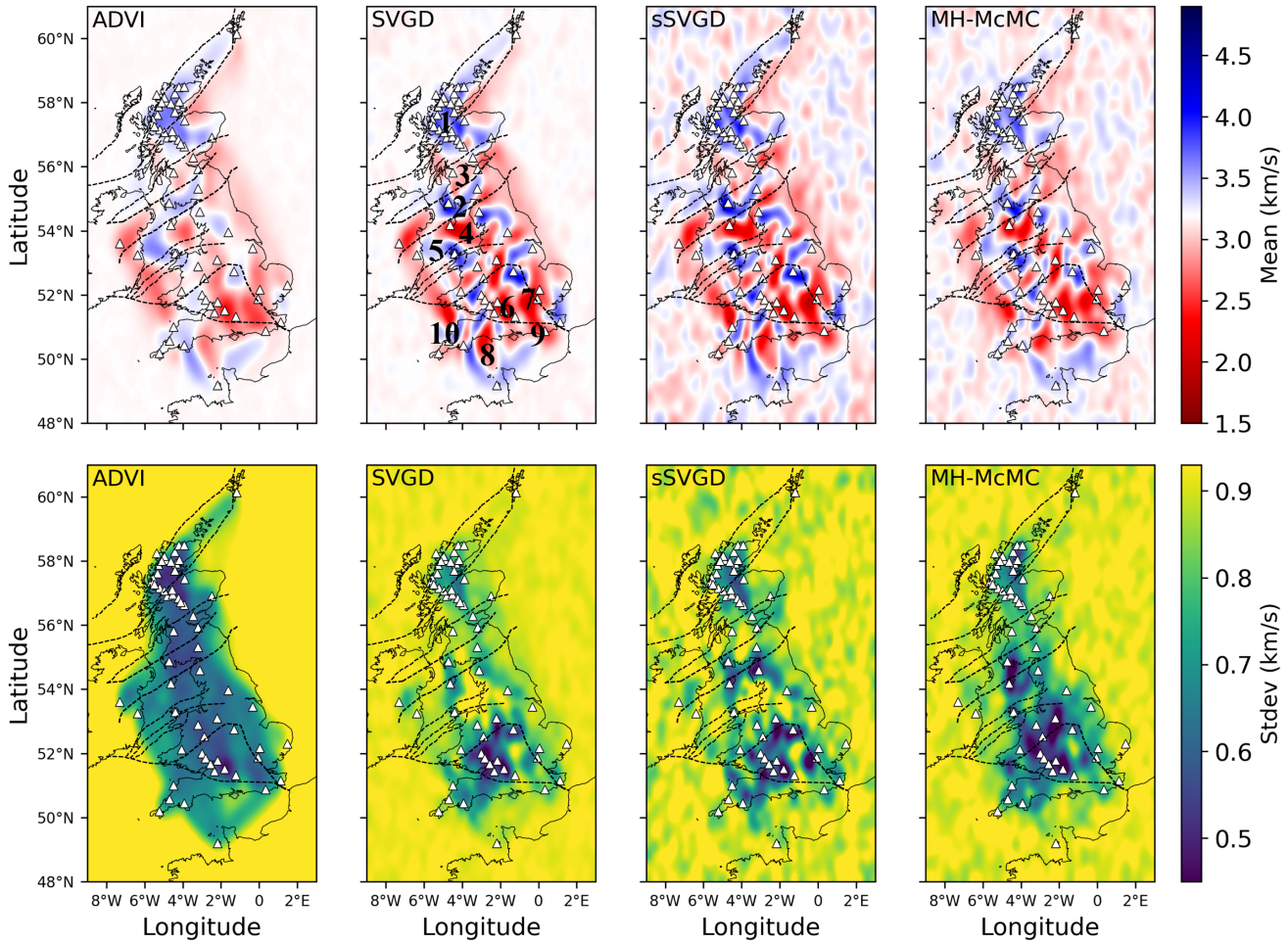

Figure 3 Mean (top row) and standard deviation (bottom row) maps of group velocity at 10 s period obtained using ADVI, SVGD, sSVGD and MH-McMC respectively. White triangles denote locations of seismometers. Black dashed lines show the Terrane boundaries in Figure 2. Black numbers are referred to in the main text.

discussion and interpretation of the velocity structures can be found in Galetti et al. (2017).

Among these results the mean map obtained using ADVI shows the smoothest structure, whereas other maps provide more detailed information. This has also been observed in previous studies (Zhang and Curtis, 2020a; Zhao et al., 2022a) and is likely caused by the limitation of implicit Gaussian assumption made in ADVI. In far offshore areas because few ray paths go through the open marine regions, the mean maps obtained using ADVI and SVGD show almost homogeneous velocity structure across these areas whose value is consistent with the mean of prior distribution. In comparison, the results obtained using sSVGD and MH-McMC exhibit more heterogeneous structures, which probably indicates that the two methods have not converged sufficiently. These areas are only loosely constrained by the data (or not at all) and hence have large posterior uncertainties requiring many more randomly generated samples in order to explore and represent the posterior distribution accurately compared to areas with tighter constraints from the data. Note that both sSVGD and MH-McMC involve random sampling of the posterior distribution, whereas samples obtained using SVGD are found deterministically by optimisation. As a result, SVGD produces smoother results (Zhang and Curtis, 2021; Zhang et al., 2023).

Overall the standard deviation maps obtained using SVGD, sSVGD and MH-McMC show similar structures. For example, the results show lower uncertainties in the Scottish highlands and southern England because of dense arrays in those areas. In the offshore areas the standard deviation is around 0.93 which is the standard deviation of the prior as no ray path goes through these regions. On the east side of the island just off the coast, although no seismometer is deployed, there are rays that travel through those areas (see details in Galetti et al., 2017), and consequently the standard deviation is smaller than that of the prior. There is a high uncertainty loop around the low velocity anomaly in the Anglian-London Basin (annotation 7 in Figure 3), which has also been observed in previous studies (Galetti et al., 2015, 2017) and reflects uncertainty in the shape of the anomaly. In addition, the East Irish Sea (annotation 4)

shows high uncertainties. This is probably because few ray paths go through this area due to its lower velocity, and consequently the area is not well constrained by the data. By contrast, the standard deviation map obtained using ADVI shows different features. Although in the Scottish highlands the results still show lower uncertainty, the rest of the area within the receiver array has almost the same uncertainty level with little variation. In addition, in the West Irish Sea and the North Sea area between Northern Scotland and Shetland Islands the results show lower uncertainties which are not observed in the results obtained using other methods. This suggests that ADVI can produce biased results because of its underlying Gaussian assumption as found in previous studies (Zhang and Curtis, 2020a).

Table 1 compares the number of forward simulations required by each method to obtain these results, which provides a good metric of the computation cost as the forward simulation is the most computationally expensive component of each method. Note that the three variational methods require computation of derivatives of the posterior pdf with respect to model parameters, which adds computational cost compared with the MH-McMC method. In this travel time tomography example the derivatives are calculated using ray paths, which are traced through the computed travel time field. This calculation requires a computation equivalent to approximately 0.08 forward simulations. We therefore compute the equivalent number of simulations by multiplying the number of simulations required by the three variational methods by 1.08, which are shown in the third column in Table 1.

The results indicate that ADVI is apparently the most efficient method as it only requires 10,000 simulations, but we have demonstrated that the method probably produces biased results. SVGD demands the highest computational cost among the three variational methods, while sSVGD requires about 10 times fewer simulations than SVGD. This makes sSVGD a good choice for practical applications as noted in Zhang et al. (2023). Nevertheless, all three variational methods are significantly more efficient than the basic MH-McMC method implemented here as a bench-mark, which required 15 millions simulations in total with 10 independent parallel chains.

We note that the above comparison depends on subjective assessment of the point of convergence for each method, so the absolute number of simulations required by each method may not be entirely accurate (especially the number used for the MH-McMC algorithm). Nevertheless the comparison at least provides insights into the relative computational cost of each method. A more careful and thorough comparison between the same MH-McMC method and variational methods can be found in Zhao et al. (2022a) which again demonstrated that variational methods were computationally efficient.

4.2 Full-waveform inversion

For the second example we use the VIP package to solve a 2D full waveform inversion problem. The input model is selected to be a part of the Marmousi model (Figure 4a, Martin et al., 2006), and is discretized using a regular 120×200 grid with a spacing of 20 m. Ten sources are equally distributed at 20 m water depth (red stars in Figure 4), and 200 receivers are equally spaced at the depth of 360 m on the seabed across the horizontal extent of the model. We simulate the waveform data using a time-domain finite difference method with a Ricker wavelet of 10 Hz central frequency, and added Gaussian noise to the data whose standard deviation is set to be 2 percent of the median of the maximum amplitude of each seismic trace. The gradients of the logarithm posterior pdf with respect to velocity are calculated using the adjoint method (Tarantola, 1988; Tromp et al., 2005; Fichtner et al., 2006; Plessix, 2006).

The prior distribution is set to be a Uniform distribution over an interval of 2 km/s at each depth (Figure 4b). To ensure that the rock velocity is higher than the velocity in the water, we imposed an extra lower bound of 1.5 km/s. For the likelihood function we use a Gaussian distribution to represent uncertainties on the waveform data:

$$p(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto \exp\left[-\frac{1}{2}\sum_i \left(\frac{d_i^{\text{obs}} - d_i(\mathbf{m})}{\sigma_i}\right)^2\right] \tag{28}$$

where i is the index of time samples, and σ_i is the standard deviation of that sample.

We apply SVGD and sSVGD to solve this full waveform inversion problem as we have demonstrated that these methods provide more accurate results than ADVI. For SVGD we used 600 particles that are initially generated from the prior distribution (an example is shown in Figure 4c) and updated them using equation (18) for 600 iterations. The final particles are used to calculate statistics of the posterior distribution. For sSVGD we generated 20 particles from the prior distribution and updated them for 4,000 iterations after an additional burn-in period of 2,000. Similarly, to reduce the memory and storage cost we only retain samples from every tenth iterations, which results in a total of 8,000 samples. Those final samples are then used to compute statistics of the posterior distribution.

Figure 5 shows the mean and standard deviation models obtained using SVGD and sSVGD. Overall the two methods produce similar results. For example, both mean models (Figure 5a and c) show similar structures to the true structure, especially in the shallow part ($<$ 1.5 km). In the deep part ($>$ 1.5 km) and close to the sides, the mean models appear to be less similar to the true structure because the waveform data are less sensitive to the velocity structure in those areas. However, the mean obtained using sSVGD is more similar to the true structure than that obtained using SVGD. This reflects the fact that sSVGD can produce more accurate results than SVGD in high dimensional spaces, which has also been observed in other studies (Gallego and Insua, 2018; Zhang et al., 2023). Note that similarly to the travel time tomography example above, the mean obtained using SVGD shows smoother structures than that obtained using sSVGD. This is likely because sSVGD is an McMC method which generates samples using stochastic sam-

Method	Number of simulations	Comparable number of simulations
ADVI	10,000	10,800
SVGD	1500,000	1620,000
sSVGD	160,000	172,800
MH-McMC	15,000,000	15,000,000

Table 1 A comparison of computational cost for ADVI, SVGD, sSVGD and MH-McMC.

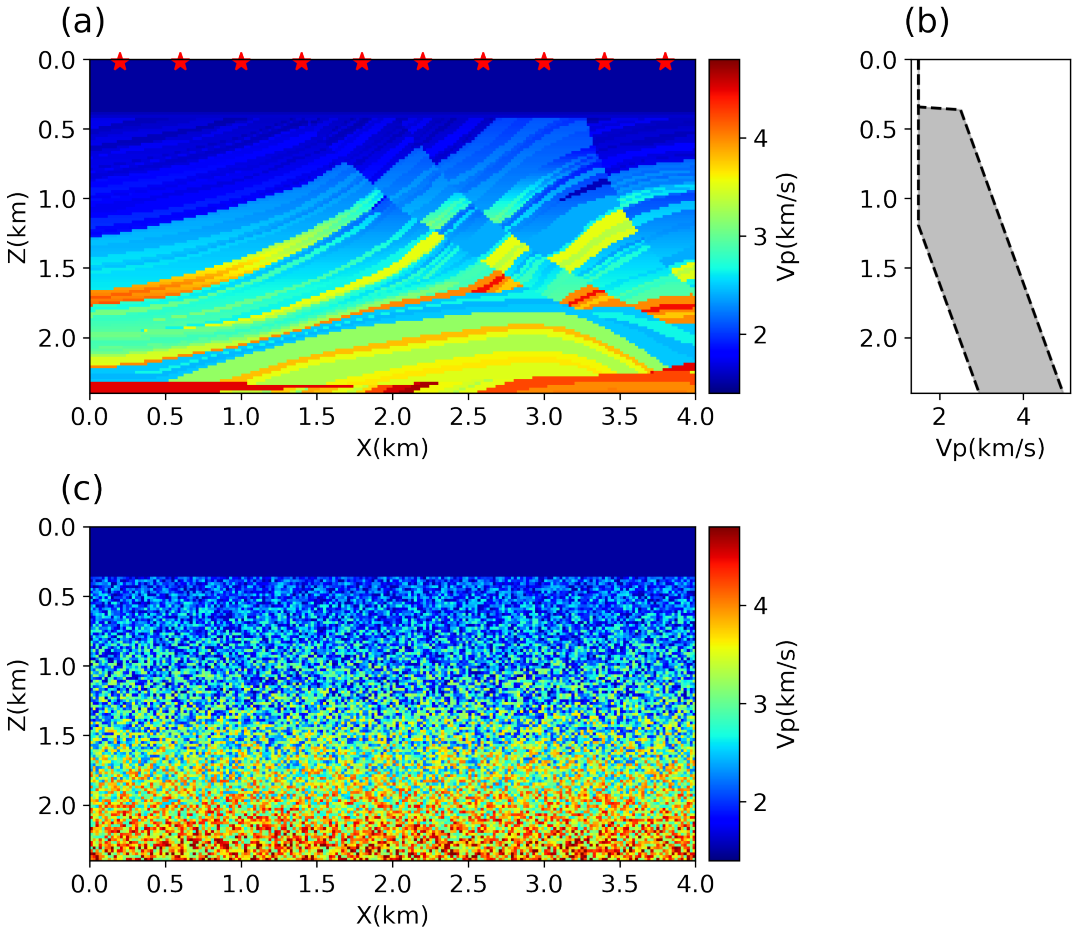

Figure 4 **(a)** The true structure used in the full waveform inversion example. Ten sources are located at the depth of 20 m (red stars) and 200 receivers (not shown) are equally spaced at the depth of 360 m on the seabed. **(b)** The prior distribution of seismic velocity, which is set to be a Uniform distribution with an interval of 2 km/s at each depth. An additional lower bound of 1.5 km/s is also imposed on the velocity to ensure that the rock velocity is higher than the velocity in water. **(c)** An example particle generated from the prior distribution.

pling, whereas in SVGD particles are obtained deterministically using optimisation. A similar phenomenon has also been observed in other studies when comparing results obtained using SVGD and sSVGD or McMC (Zhang and Curtis, 2021; Zhang et al., 2023).

Overall the standard deviation models show similar structural shapes to those in the mean model as has been observed in other studies (Gebraad et al., 2020; Zhang and Curtis, 2020b, 2021; Zhang et al., 2023). In the shallow part ($<$ 1.0 km) the results show lower uncertainties and in the deeper part the uncertainty is higher because of lower data coverage. Those higher velocity anomalies in the deeper part are clearly associated with lower standard deviations, which likely reflects that those anomalies have large influences on the waveform data and hence have lower uncertainty. Similarly to the mean structures, the standard deviations obtained using SVGD show smoother structures than are obtained using sSVGD. In addition, the magnitude of the standard deviation obtained using SVGD is slightly lower than that obtained using sSVGD, which is likely because SVGD can underestimate uncertainties in high dimensional spaces due to the limited number of posterior samples produced (Ba et al., 2022; Zhang et al., 2023).

To further understand the results we show marginal distributions obtained using SVGD and sSVGD along three vertical profiles whose locations are denoted by dashed black lines in Figure 5. Overall the results show broader distributions in the deeper part ($>$ 1 km) than in the shallow part as we have observed in the standard deviation models. Furthermore, the distributions

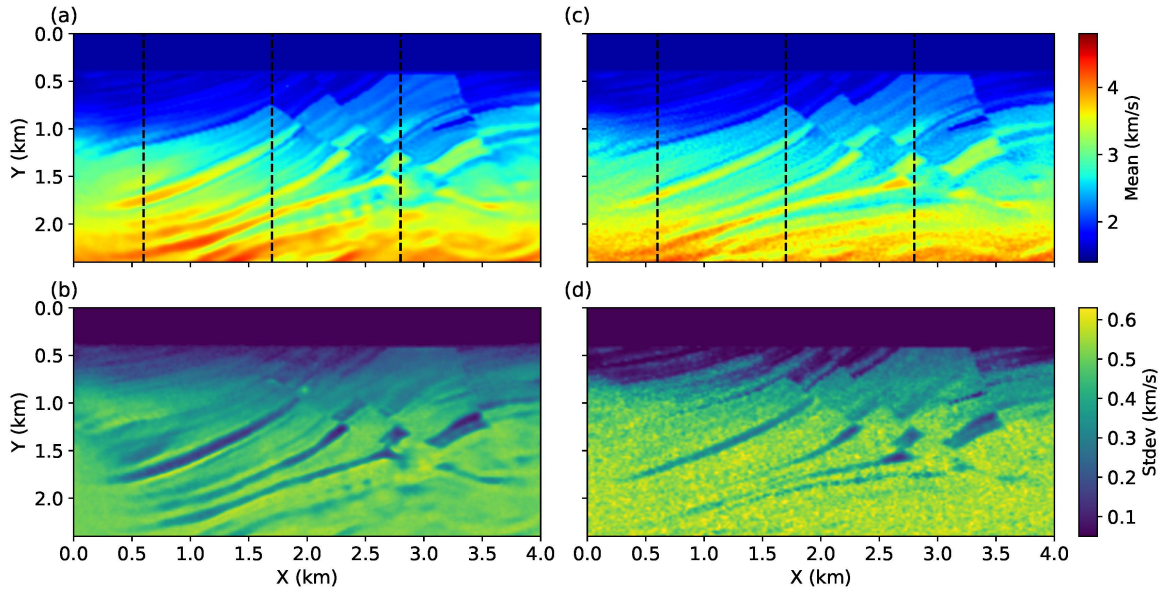

Figure 5 The mean (top row) and standard deviation (bottom row) obtained using SVGD (left panel) and sSVGD (right panel), respectively. Black dashed lines denote well log locations referred to in the main text.

Method	Number of simulations
SVGD	360,000
sSVGD	120,000

Table 2 Computational cost required by SVGD and sSVGD for FWI.

obtained using sSVGD are broader than those obtained using SVGD, which again demonstrates that SVGD can underestimate uncertainties. Note that in the results obtained using SVGD some true velocities lie outside the high probability area at large depths (> 1.5 km), whereas those obtained using sSVGD generally include the true velocity in values with non-zero uncertainty. This shows that SVGD can produce biased results for high dimensional problems as noted in several studies (Ba et al., 2022; Zhang et al., 2023).

Similarly to the above section we measure the computational cost required by each method using the number of forward and adjoint simulations (Table 2). Specifically, SVGD required 360,000 simulations to converge, while sSVGD used 120,000 simulations. This again demonstrates that sSVGD can be more computationally efficient than SVGD because sSVGD requires fewer particles yet generates many more samples. To give an overall idea of the computational cost, the above inversions required 49 hours for sSVGD using 40 AMD EPYC CPU cores, and 3 days for SVGD using 90 CPU cores.

5 Discussion

Although in the VIP package we only implemented 2D travel time tomography and 2D full waveform inversion, the code can easily be applied to other types of problems, and also to larger scale problems by using modern high performance computation (HPC) facilities. For example, users can implement 3D full waveform inversion by providing a 3D forward and adjoint simulation code (see more details in the code documentation, and an example in Zhang et al., 2023). In order to enable easy deployment on HPC facilities, the code provides a guide on how to parallelize the computation using the Sun Grid Engine queuing system. Other queuing systems can be implemented in a similar way.

Although we have demonstrated that sSVGD can generate more accurate results than SVGD in high dimensional problems and requires less computational cost in total, the method generally requires many more iterations. As a result, sSVGD may be less efficient than SVGD in wall clock time when a large number of CPU cores is available. This is why we implement SVGD in the VIP package as in practice it may be a better choice for low dimensional problems.

ADVI may become inefficient in a high dimensional space because of the increased size of the covariance matrix. To enable applications in such cases, we also implement a diagonal covariance matrix, that is, a mean-field approximation (Kucukelbir et al., 2017). In SVGD and sSVGD besides the radial basis function kernel used in above examples, the package also implements diagonal matrix-valued kernel functions which are constructed by combining a positive definite diagonal matrix $\mathbf{Q}$ and the radial basis function (Wang et al., 2019; Zhang and Curtis, 2021). The elements of $\mathbf{Q}$ can be set as the inverse of the variance calculated across particles (Zhang and Curtis, 2021).

To promote reproducibility and show how to use the code, we included several examples along with the code which can be used to reproduce those results obtained in the above section. We encourage interested readers to begin with these examples to familiarize themselves with the code. Finally, we note that VIP is actively being developed and expanded, and contributions from the community are welcome.

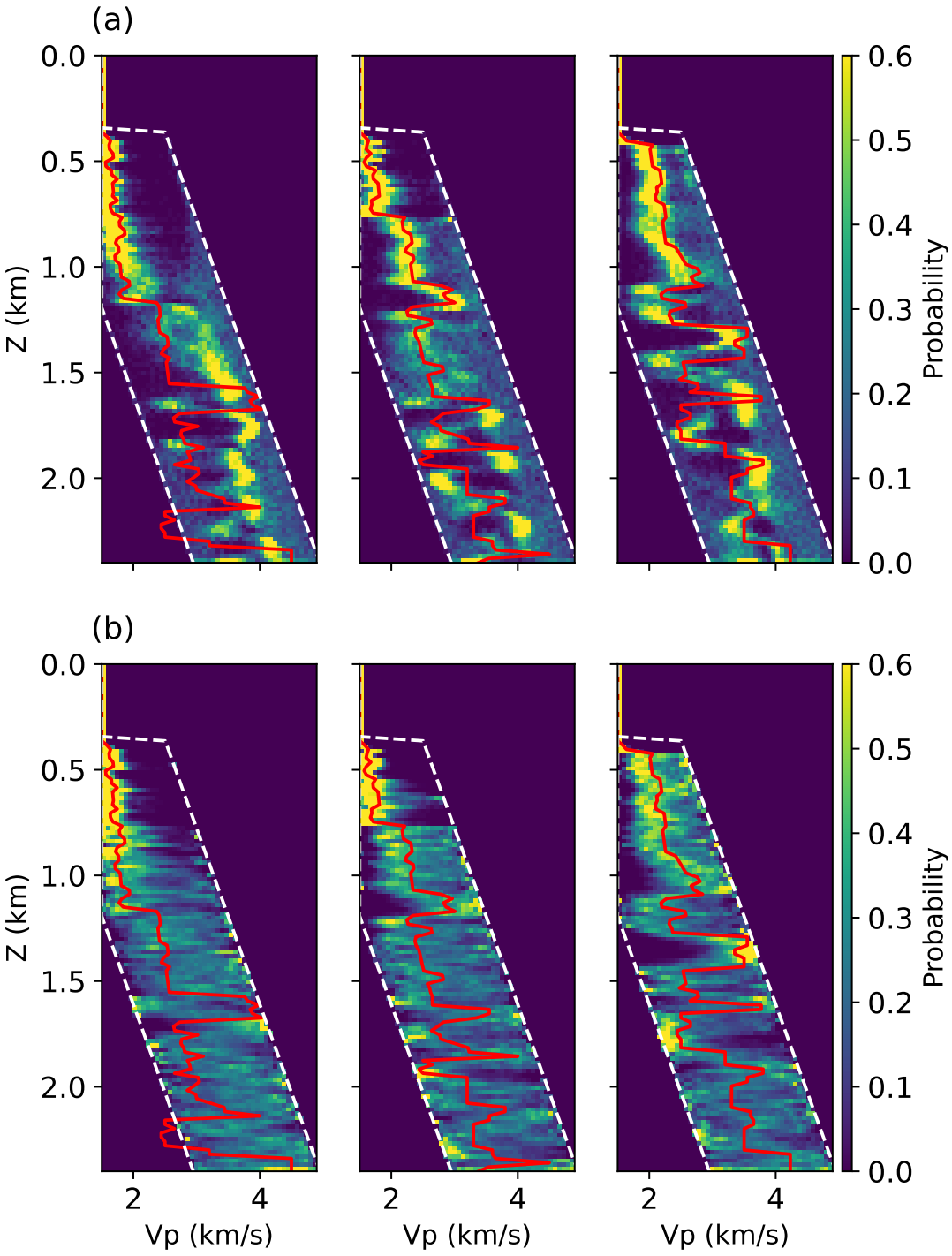

Figure 6 Marginal distributions at three well logs (black dashed lines in Figure 5) obtained using **(a)** SVGD and **(b)** sSVGD, respectively. Red lines show the true velocity profiles and white dashed lines show the lower and upper bound of the prior distribution.

6 Conclusion

VIP is a Python package which solves general inverse problems using variational inference methods, including automatic differential variational inference (ADVI), Stein variational gradient descent (SVGD) and stochastic SVGD (sSVGD). The package is designed to be easy enough for beginners to use, and efficient enough to solve complex inverse problems. In addition, VIP is implemented in a scalable way such that it can be deployed on a desktop as well as in high performance computation facilities. We demonstrated the package using two examples: 2D travel time tomography and 2D full waveform inversion. Users can also use the package to solve their own inverse problems by providing an appropriate forward modelling and gradient calculation code. We conclude that VIP can be used to solve a wide range of inverse problems in practice. The most recent release of the code can be downloaded from GitHub (https://github.com/xin2zhang/VIP) and a stable version is available on Zenodo (Zhang and Curtis, 2023).

Acknowledgements

The authors thank the Edinburgh Imaging Project sponsors (BP and Total), National Natural Science Foundation of China (42204055) and the Fundamental Research Funds for the Central Universities for supporting this research. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (http://www.ecdf.ed.ac.uk/). The authors also thank two anonymous reviewers for their valuable feedback which significantly improved the manuscript.

Data and code availability

The code and data used in this study are available in a Zenodo repository (Zhang and Curtis, 2023).

References

Afanasiev, M., Boehm, C., van Driel, M., Krischer, L., Rietmann, M., May, D. A., Knepley, M. G., and Fichtner, A. Modular and flexible spectral-element waveform modelling in two and three dimensions. *Geophysical Journal International*, 216(3):1675–1692, 2019. doi: 10.1093/gji/ggy469.

Agata, R., Shiraishi, K., and Fujie, G. Bayesian seismic tomography based on velocity-space Stein variational gradient descent for physics-informed neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. doi: 10.1109/TGRS.2023.3295414.

Ahmed, Z., Yunyue, L., and Arthur, C. Regularized seismic amplitude inversion via variational inference. *Geophysical Prospecting*, n/a(n/a), 2022. doi: 10.1111/1365-2478.13248.

Aki, K. and Lee, W. 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, 1976. doi: 10.1029/JB081i023p04381.

Andersen, K. E., Brooks, S. P., and Hansen, M. B. A Bayesian approach to crack detection in electrically conducting media. *Inverse Problems*, 17(1):121, 2001. doi: 10.1088/0266-5611/17/1/310.

Arnold, R. and Curtis, A. Interrogation theory. *Geophysical Journal International*, 214(3):1830–1846, 2018. doi: 10.1093/gji/ggy248.

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

Ba, J., Erdogdu, M. A., Ghassemi, M., Sun, S., Suzuki, T., Wu, D., and Zhang, T. Understanding the Variance Collapse of SVGD in High Dimensions. In *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=Qycd9j5Qp9J.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39, 2010. doi: 10.1109/MCSE.2010.118.

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

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Bodin, T. and Sambridge, M. Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, 178 (3):1411–1436, 2009. doi: 10.1111/j.1365-246X.2009.04226.x.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

Chen, T., Fox, E., and Guestrin, C. Stochastic Gradient Hamiltonian Monte Carlo. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1683–1691, Bejing, China, 22–24 Jun 2014. PMLR. https://proceedings.mlr.press/v32/cheni14.html.

Curtis, A. and Lomax, A. Prior information, sampling distributions, and the curse of dimensionality. *Geophysics*, 66(2):372–378, 2001. doi: 10.1190/1.1444928.

Dosso, S. E., Holland, C. W., and Sambridge, M. Parallel tempering for strongly nonlinear geoacoustic inversion. *The Journal of the Acoustical Society of America*, 132(5):3030–3040, 2012. doi: 10.1121/1.4757639.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. http://jmlr.org/papers/v12/duchi11a.html.

Fichtner, A., Bunge, H.-P., and Igel, H. The adjoint method in seismology: I. Theory. *Physics of the Earth and Planetary Interiors*, 157(1-2):86–104, 2006. doi: 10.1016/j.pepi.2006.03.016.

Fichtner, A., Zunino, A., and Gebraad, L. Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, 216(2):1344–1363, 2018. doi: 10.1093/gji/ggy496.

Galetti, E., Curtis, A., Meles, G. A., and Baptie, B. Uncertainty loops in travel-time tomography from nonlinear wave physics. *Physical review letters*, 114(14):148501, 2015. doi: 10.1103/physrevlett.114.148501.

Galetti, E., Curtis, A., Baptie, B., Jenkins, D., and Nicolson, H. 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, 2017. doi: 10.1093/gji/ggw286.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., and Stephenson, J. Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine and Petroleum Geology*, 26 (4):525–535, 2009. doi: 10.1016/j.marpetgeo.2009.01.003.

Gallego, V. and Insua, D. R. Stochastic gradient MCMC with repulsive forces. *arXiv preprint arXiv:1812.00071*, 2018.

Gebraad, L., Boehm, C., and Fichtner, A. Bayesian Elastic Full-Waveform Inversion Using Hamiltonian Monte Carlo. *Journal of Geophysical Research: Solid Earth*, 125(3):e2019JB018428, 2020. doi: 10.1029/2019JB018428.

Gong, C., Peng, J., and Liu, Q. Quantile Stein Variational Gradient Descent for Batch Bayesian Optimization. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2347–2356. PMLR, 09–15 Jun 2019. https://proceedings.mlr.press/v97/gong19b.html.

Green, P. J. Reversible jump Markov chain Monte Carlo computation and Byesian model determination. *Biometrika*, pages 711–732, 1995. doi: 10.1093/biomet/82.4.711.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97.

Hawkins, R. and Sambridge, M. Geophysical imaging using trans-dimensional trees. *Geophysical Journal International*, 203(2): 972–1000, 2015. doi: 10.1093/gji/ggv326.

Hukushima, K. and Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996. doi: 10.1143/JPSJ.65.1604.

Izzatullah, M., Alkhalifah, T., Romero, J., Corrales, M., Luiken, N., and Ravasi, M. Plug-and-Play Stein variational gradient descent for Bayesian post-stack seismic inversion. In *84th EAGE Annual Conference & Exhibition*, volume 2023, pages 1–5. European Association of Geoscientists & Engineers, 2023. doi: 10.3997/2214-4609.202310177.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Komatitsch, D., Tromp, J., Garg, R., Gharti, H. N., Nagaso, M., Oral, E., and et al. SPECFEM/specfem3d: SPECFEM3D v4.1.0. doi: 10.5281/zenodo.10413988.

Kubrusly, C. and Gravier, J. Stochastic approximation algorithms and applications. In *1973 IEEE conference on decision and control including the 12th symposium on adaptive processes*, pages 763–766. IEEE, 1973. doi: 10.1109/CDC.1973.269114.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(14):1–45, 2017. http://jmlr.org/papers/v18/16-107.html.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. http://www.jstor.org/stable/2236703.

Kumar, R., Kotsi, M., Siahkoohi, A., and Malcolm, A. Enabling uncertainty quantification for seismic data preprocessing using normalizing flows (NF)—An interpolation example. In *First International Meeting for Applied Geoscience & Energy*, pages 1515–1519. Society of Exploration Geophysicists, 2021. doi: 10.1190/segam2021-3583705.1.

Li, X., Bürgi, P. M., Ma, W., Noh, H. Y., Wald, D. J., and Xu, S. DisasterNet: Causal Bayesian Networks with Normalizing Flows for Cascading Hazards Estimation from Satellite Imagery. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4391–4403, 2023. doi: 10.1145/3580305.3599807.

Lions, J. L. *Optimal control of systems governed by partial differential equations*, volume 170. Springer, 1971.

Liu, Q. and Wang, D. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper_files/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

Lomas, A., Luo, S., Irakarama, M., Johnston, R., Vyas, M., and Shen, X. 3D Probabilistic Full Waveform Inversion: Application to Gulf of Mexico Field Data. In *84th EAGE Annual Conference & Exhibition*, volume 2023, pages 1–5. European Association of Geoscientists & Engineers, 2023. doi: 10.3997/2214-4609.202310720.

Ma, Y.-A., Chen, T., and Fox, E. A Complete Recipe for Stochastic Gradient MCMC. 28, 2015. https://proceedings.neurips.cc/paper_files/paper/2015/file/9a4400501febb2a95e79248486a5f6d3-Paper.pdf.

Malinverno, A. Parsimonious Byesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3):675–688, 2002. doi: 10.1046/j.1365-246X.2002.01847.x.

Malinverno, A., Leaney, S., et al. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data. In *2000 SEG Annual Meeting*. Society of Exploration Geophysicists, 2000. doi: 10.1190/1.1815943.

Martin, G. S., Wiley, R., and Marfurt, K. J. Marmousi2: An elastic upgrade for Marmousi. *The leading edge*, 25(2):156–166, 2006. doi: 10.1190/1.2172306.

Martin, J., Wilcox, L. C., Burstedde, C., and Ghattas, O. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012. doi: 10.1137/110845598.

McKean, S., Priest, J., Dettmer, J., Fradelizio, G., and Eaton, D. Separating Hydraulic Fracturing Microseismicity From Induced Seismicity by Bayesian Inference of Non-Linear Pressure Diffusivity. *Geophysical Research Letters*, 50(14):e2022GL102131, 2023. doi: 10.1029/2022GL102131.

Metropolis, N. and Ulam, S. The Monte Carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949. doi: 10.1080/01621459.1949.10483310.

Mosegaard, K. and Sambridge, M. Monte Carlo analysis of inverse problems. *Inverse problems*, 18(3):R29, 2002. doi: 10.1088/0266-5611/18/3/201.

Mosegaard, K. and Tarantola, A. Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7):12431–12447, 1995. doi: 10.1029/94JB03097.

Nawaz, M. A. and Curtis, A. Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies. *Geophysical Journal International*, 214(2):845–875, 2018. doi: 10.1093/gji/ggy163.

Nawaz, M. A. and Curtis, A. Rapid Discriminative Variational Byesian Inversion of Geophysical Data for the Spatial Distribution of Geological Properties. *Journal of Geophysical Research: Solid Earth*, 2019. doi: 10.1029/2018JB016652.

Nawaz, M. A., Curtis, A., Shahraeeni, M. S., and Gerea, C. VARIATIONAL BAYESIAN INVERSION OF SEISMIC ATTRIBUTES JOINTLY FOR GEOLOGICAL FACIES AND PETROPHYSICAL ROCK PROPERTIES. *GEOPHYSICS*, pages 1–78, 2020. doi: 10.1190/geo2019-0163.1.

Nicolson, H., Curtis, A., Baptie, B., and Galetti, E. Seismic interferometry and ambient noise tomography in the British Isles. *Proceedings of the Geologists' Association*, 123(1):74–86, 2012. doi: 10.1016/j.pgeola.2011.04.002.

Nicolson, H., Curtis, A., and Baptie, B. Rayleigh wave tomography of the British Isles from ambient seismic noise. *Geophysical Journal International*, 198(2):637–655, 2014. doi: 10.1093/gji/ggu071.

O'Hagan, A. and Forster, J. J. *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*, volume 2. Arnold, 2004.

Oksendal, B. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Orozco, R., Louboutin, M., Siahkoohi, A., Rizzuti, G., van Leeuwen, T., and Herrmann, F. Amortized Normalizing Flows for Transcranial Ultrasound with Uncertainty Quantification. *arXiv preprint arXiv:2303.03478*, 2023. doi: 10.48550/arXiv.2303.03478.

Parisi, G. *Statistical field theory*. Addison-Wesley, 1988. doi: 10.1063/1.2811677.

Pinder, T., Nemeth, C., and Leslie, D. Stein variational Gaussian processes. *arXiv preprint arXiv:2009.12141*, 2020. doi: 10.48550/arXiv.2009.12141.

Plessix, R.-E. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006. doi: 10.1111/j.1365-246X.2006.02978.x.

Ramgraber, M., Weatherl, R., Blumensaat, F., and Schirmer, M. Non-Gaussian Parameter Inference for Hydrogeological Models Using Stein Variational Gradient Descent. *Wa-*

ter Resources Research, 57(4):e2020WR029339, 2021. doi: 10.1029/2020WR029339.

Ramirez, A. L., Nitao, J. J., Hanley, W. G., Aines, R., Glaser, R. E., Sengupta, S. K., Dyer, K. M., Hickling, T. L., and Daily, W. D. Stochastic inversion of electrical resistivity changes using a Markov Chain Monte Carlo approach. *Journal of Geophysical Research: Solid Earth*, 110(B2), 2005. doi: 10.1029/2004JB003449.

Rawlinson, N. FMST: fast marching surface tomography package–Instructions. *Research School of Earth Sciences, Australian National University, Canberra*, 29:47, 2005.

Rawlinson, N. and Sambridge, M. Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics*, 69(5):1338–1350, 2004. doi: 10.1190/1.1801950.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015. doi: 10.48550/arxiv.1505.05770.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. doi: 10.1214/aoms/1177729586.

Roberts, G. O., Tweedie, R. L., et al. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. doi: 10.2307/3318418.

Rocklin, M. et al. Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*, volume 130, page 136. SciPy Austin, TX, 2015. doi: 10.25080/majora-7b98e3ed-013.

Rücker, C., Günther, T., and Wagner, F. M. pyGIMLi: An open-source library for modelling and inversion in geophysics. *Computers & Geosciences*, 109:106–123, 2017. doi: 10.1016/j.cageo.2017.07.011.

Sambridge, M. A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, page ggt342, 2013. doi: 10.1093/gji/ggt342.

Sambridge, M. and Mosegaard, K. Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3):3–1, 2002. doi: 10.1029/2000RG000089.

Sen, M. K. and Biswas, R. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. *Geophysics*, 82(3):R119–R134, 2017. doi: 10.1190/geo2016-0010.1.

Shen, W., Ritzwoller, M. H., Schulte-Pelkum, V., and Lin, F.-C. Joint inversion of surface wave dispersion and receiver functions: a Byesian Monte-Carlo approach. *Geophysical Journal International*, 192(2):807–836, 2012. doi: 10.1093/gji/ggs050.

Siahkoohi, A., Rizzuti, G., Witte, P. A., and Herrmann, F. J. Faster Uncertainty Quantification for Inverse Problems with Conditional Normalizing Flows. *arXiv preprint arXiv:2007.07985*, 2020. doi: 10.48550/arXiv.2007.07985.

Siahkoohi, A., Orozco, R., Rizzuti, G., and Herrmann, F. J. Wave-equation-based inversion with amortized variational Bayesian inference. *arXiv preprint arXiv:2203.15881*, 2022a. doi: 10.48550/arXiv.2203.15881.

Siahkoohi, A., Rizzuti, G., and Herrmann, F. J. Deep Bayesian inference for seismic imaging with tasks. *Geophysics*, 87(5): S281–S302, 2022b. doi: 10.1190/geo2021-0666.1.

Smith, J. D., Ross, Z. E., Azizzadenesheli, K., and Muir, J. B. HypoSVI: Hypocentre inversion with Stein variational inference and physics informed neural networks. *Geophysical Journal International*, 228(1):698–710, 2022. doi: 10.1093/gji/ggab309.

Tarantola, A. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266, 1984. doi: 10.1190/1.1441754.

Tarantola, A. Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation. In *Scattering and Attenuations of Seismic Waves, Part I*, pages 365–399. Springer, 1988. doi: 10.1007/bf01772605.

Tarantola, A. *Inverse problem theory and methods for model parameter estimation*, volume 89. SIAM, 2005. doi: 10.1137/1.9780898717921.

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

Tromp, J., Tape, C., and Liu, Q. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, 160(1):195–216, 2005. doi: 10.1111/j.1365-246X.2004.02453.x.

Valentine, A. P. and Sambridge, M. Gaussian process models—I. A framework for probabilistic continuous inverse theory. *Geophysical Journal International*, 220(3):1632–1647, 2020. doi: 10.1093/gji/ggz520.

Wang, D., Tang, Z., Bajaj, C., and Liu, Q. Stein variational gradient descent with matrix-valued kernels. In *Advances in neural information processing systems*, pages 7836–7846, 2019. doi: 10.48550/arXiv.1910.12794.

Wang, W., McMechan, G. A., and Ma, J. Re-weighted variational full waveform inversions. *Geophysics*, 88(4):1–61, 2023. doi: 10.1190/geo2021-0766.1.

Wathelet, M., Chatelain, J.-L., Cornou, C., Giulio, G. D., Guillier, B., Ohrnberger, M., and Savvaidis, A. Geopsy: A user-friendly open-source tool set for ambient vibration processing. *Seismological Research Letters*, 91(3):1878–1889, 2020. doi: 10.1785/0220190360.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Zaroli, C. Global seismic tomography using Backus–Gilbert inversion. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society*, 207(2):876–888, 2016. doi: 10.1093/gji/ggw315.

Zaroli, C. Seismic tomography using parameter-free Backus–Gilbert inversion. *Geophysical Journal International*, 218(1): 619–630, 2019. doi: 10.1093/gji/ggz175.

Zaroli, C., Koelemeijer, P., and Lambotte, S. Toward seeing the Earth's interior through unbiased tomographic lenses. *Geophysical Research Letters*, 44(22):11–399, 2017. doi: 10.1002/2017GL074996.

Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. doi: 10.48550/arXiv.1212.5701.

Zhang, C. and Chen, T. Bayesian slip inversion with automatic differentiation variational inference. *Geophysical Journal International*, 229(1):546–565, 2022. doi: 10.1093/gji/ggab438.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018a. doi: 10.1109/TPAMI.2018.2889774.

Zhang, X. and Curtis, A. Seismic tomography using variational inference methods. *Journal of Geophysical Research: Solid Earth*, 125(4):e2019JB018589, 2020a.

Zhang, X. and Curtis, A. Variational full-waveform inversion. *Geophysical Journal International*, 222(1):406–411, 2020b. doi: 10.1093/gji/ggaa170.

Zhang, X. and Curtis, A. Bayesian Full-waveform Inversion with Realistic Priors. *Geophysics*, 86(5):1–20, 2021. doi: 10.1190/geo2021-0118.1.

Zhang, X. and Curtis, A. Interrogating probabilistic inversion results for subsurface structural information. *Geophysical Journal International*, 229(2):750–757, 2022. doi: 10.1093/gji/ggab496.

Zhang, X., Curtis, A., Galetti, E., and de Ridder, S. 3-D Monte Carlo surface wave tomography. *Geophysical Journal International*, 215(3):1644–1658, 2018b. doi: 10.1093/gji/ggy362.

Zhang, X., Lomas, A., Zhou, M., Zheng, Y., and Curtis, A. 3-D Bayesian variational full waveform inversion. *Geophysical Journal International*, 234(1):546–561, 2023. doi: 10.1093/gji/ggad057.

Zhang, Z. and Curtis, A. Variational Inversion Package, 2023. doi: 10.5281/zenodo.10036815.

Zhao, X. and Curtis, A. Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference. *Journal of Geophysical Research: Solid Earth*, 129(1):e2023JB027789, 2024. doi: 10.1029/2023JB027789.

Zhao, X., Curtis, A., and Zhang, X. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228 (1):213–239, 2022a.

Zhao, X., Curtis, A., and Zhang, X. Interrogating Subsurface Structures using Probabilistic Tomography: an example assessing the volume of Irish Sea basins. *Journal of Geophysical Research: Solid Earth*, 127(4):e2022JB024098, 2022b. doi: 10.1029/2022JB024098.

Zhao, Z. and Sen, M. K. A gradient based MCMC method for FWI and uncertainty analysis. In *SEG Technical Program Expanded Abstracts 2019*, pages 1465–1469. Society of Exploration Geophysicists, 2019. doi: 10.1190/segam2019-3216560.1.

Zhdanov, M. S. *Geophysical inverse theory and regularization problems*, volume 36. Elsevier, 2002. doi: 10.1016/s0076-6895(02)x8037-8.

Zunino, A., Gebraad, L., Ghirotto, A., and Fichtner, A. HMCLab: a framework for solving diverse geophysical inverse problems using the Hamiltonian Monte Carlo method. *arXiv preprint arXiv:2303.10047*, 2023. doi: 10.1093/gji/ggad403.