Bayesian Inversion, Uncertainty Analysis and Interrogation using Boosting Variational Inference

Xuebin Zhao[1] **and Andrew Curtis**[1]

[1]School of Geosciences, University of Edinburgh, Edinburgh, United Kingdom

Key Points:

- We apply boosting variational inference to Bayesian inversion, which uses a mixture of Gaussians to approximate the posterior distribution.
- The method is shown to be efficient and accurate, and constructs a fully analytic expression for high-dimensional posterior distribution.
- The analytic solution allows extremely efficient methods to be used to answer scientific questions with minimum bias.

Corresponding author: Xuebin Zhao, xuebin.zhao@ed.ac.uk

Abstract

Geoscientists use observed data to estimate properties of the Earth's interior. This often requires non-linear inverse problems to be solved and uncertainties to be estimated. Bayesian inference solves inverse problems under a probabilistic framework, in which uncertainty is represented by a so-called posterior probability distribution. Recently, variational inference has emerged as an efficient method to estimate Bayesian solutions. By seeking the closest approximation to the posterior distribution within any chosen family of distributions, variational inference yields a fully probabilistic solution. It is important to define expressive variational families so that the posterior distribution can be represented accurately. We introduce *boosting variational inference* (BVI) as a computationally efficient means to construct a flexible approximating family comprising all possible finite mixtures of simpler component distributions. We use Gaussian mixture components due to their fully parametric nature and the ease to optimise. We apply BVI to seismic travel time tomography and full waveform inversion, comparing its performance with other methods. The results demonstrate that BVI achieves reasonable efficiency and accuracy while enabling the construction of a fully analytic expression for the posterior distribution. Samples that represent major components of uncertainty in the solution can be obtained analytically from each mixture component. We demonstrate that these samples can be used to solve an interrogation problem: to assess the size of a subsurface target structure. To the best of our knowledge, this is the first method in geophysics that provides both analytic and reasonably accurate solutions to fully non-linear, high-dimensional Bayesian full waveform inversion problems.

Plain Language Summary

This paper introduces an efficient method to solve non-linear problems in which Bayesian uncertainties in the solution are to be estimated given some observed data set. The method uses a flexible mathematical function which is optimised to best approximate the set of possible solutions. This enables a fully analytic expression to be estimated for the inversion results. We use the method to solve tomographic imaging problems using first seismic wave travel times, and then full waveform inversion. By interrogating the resulting distribution, we show how the answer to a specific scientific question of interest, "How large is a particular subsurface structure of interest?", can be found highly efficiently and with minimum bias.

1 Keywords

Bayesian Inference, Seismic Imaging, Probability Distribution, Uncertainty Analysis

2 Introduction

In many geophysical problems, information about the Earth is inferred using data recorded either on or beneath the Earth's surface, or in the oceans, atmosphere or near-Earth orbits. These properties of interest are usually described by so-called latent parameters, and it is often the case that observed data can be predicted approximately given values for those parameters. This calculation is called the *forward* problem, and the parameter-data relationship is usually non-linear. Yet typically in the same problem, no inverse relationship, which predicts the parameter values given the data, exists. The process of inferring the values of parameters is therefore formulated as an inverse problem. In practice, inverse problem solutions are always non-unique, so it is crucial to estimate the range of properties that are consistent with observations if solutions are to be interpreted in a reliable manner (Tarantola, 2005).

Geophysical inverse problems are often solved without estimating the true uncertainty structure. Usually such approaches seek a solution that best fits the observations, using a variant of the following procedure: the non-linear forward function is linearised around an initial reference Earth model (a set of parameter values) to yield approximate forward relationships. Using linear algebra, these approximations allow the parameter values to be perturbed so as to better fit recorded data. The process of linearisation and updating of parameter values is iterated using successive estimates as new reference models until convergence is observed. The final set of parameter values is used as a best estimate of the true model (Iyer & Hirahara, 1993).

Unfortunately, in many cases the result does not accurately represent the true Earth due to non-uniqueness of the inverse problem solution (Boyd & Vandenberghe, 2004), particularly in cases where the initial model is significantly different from the true solution. Moreover, within the above framework it is impossible to evaluate uncertainty in the inversion results that originates from non-linearity of the forward relations (Gallagher et al., 2009). It is therefore challenging to solve interrogation problems, in which the so-

lution is interpreted to answer scientific questions of interest (Arnold & Curtis, 2018; Ely et al., 2018; X. Zhang & Curtis, 2022; X. Zhao et al., 2022; Siahkoohi et al., 2022).

As an alternative, a suite of methods collectively referred to as *Bayesian inversion* or *Bayesian inference* allow statistics of the full uncertainty structure of the solution to be estimated. These methods employ Bayes' rule to update *prior* (initial) knowledge about the parameter values that is described probabilistically, using new information provided by the observed data. The result of the inversion is represented by the *posterior* probability distribution (or density) function (pdf): in principle this provides a complete solution which describes all parameter values that are consistent with the data, and quantifies their relative probabilities.

Bayesian inference often uses global search methods such as random sampling to characterise the family of values in parameter space that yield acceptable data fits (Rothman, 1986; Stoffa & Sen, 1991; Sen & Stoffa, 2013; Sambridge, 1999). Monte Carlo methods (Press, 1968; Anderssen & Seneta, 1971; Malinverno, 2002) and their variants, including Metropolis-Hastings Markov chain Monte Carlo (MH-McMC – Mosegaard & Tarantola, 1995), reversible-jump McMC (rj-McMC – Bodin & Sambridge, 2009; Bodin et al., 2012; Galetti et al., 2015, 2017; Biswas & Sen, 2022), informed proposal Monte Carlo (Khoshkholgh et al., 2021; Khoshkholgh, Zunino, & Mosegaard, 2022; Khoshkholgh, Orozova-Bekkevold, & Mosegaard, 2022), Hamiltonian Monte Carlo (HMC – Fichtner & Simutė, 2018; Gebraad et al., 2020; de Lima et al., 2023), Langevin dynamics McMC (Izzatullah et al., 2020; Siahkoohi et al., 2022), and others, have been studied extensively for various geophysical inversion problems. However, such methods still have notable issues that can become problematic in practical problems: (1) slow convergence, sometimes converging only in infinite time (Atchadé & Rosenthal, 2005; Andrieu & Thoms, 2008); (2) poor scalability to problems with many parameters due to the curse of dimensionality (Scales, 1996; Curtis & Lomax, 2001); and (3) parallelising some methods at the sample level is not possible (Neiswanger et al., 2013).

A different approach to finding Bayesian solutions is referred to as *variational inference*. In variational methods, a family of simple probability distributions (often referred to as the variational family) is defined, and an optimal member within this family is sought which best approximates the true (unknown) posterior pdf. This can be found by minimising the difference (or mathematically speaking, the distance) between the pos-

terior and variational distributions. The Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) is typically used for measuring the distance between two distributions. Thus, variational methods solve Bayesian problems using potentially efficient and parallelisable optimisation processes and offer well understood convergence criteria (Blei et al., 2017; C. Zhang et al., 2018).

In recent years, sophisticated variational algorithms have been proposed due to advances in computational power and the development of modern deep learning frameworks such as TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019), which enable tractable construction and learning of large scale probabilistic models. These methods either deterministically generate a set of posterior samples (Liu & Wang, 2016; Gallego & Insua, 2018) or directly model a parametric probability distribution to approximate the true posterior pdf (Kingma & Welling, 2014; Rezende & Mohamed, 2015; Kingma et al., 2016; Kucukelbir et al., 2017). In geophysics, novel variational inference methods were developed for rock physical interpretation and inversion of seismic data by Nawaz and Curtis (2018, 2019) and Nawaz et al. (2020). Since then the methodology has been applied to a variety of problems including travel time tomography (X. Zhang & Curtis, 2020a; X. Zhao et al., 2021; Levy et al., 2022), seismic denoising (Siahkoohi et al., 2021, 2023), seismic amplitude inversion (Zidan et al., 2022), earthquake hypocentre inversion (Smith et al., 2022), slip distribution inversion (Sun et al., 2023), full waveform inversion in 2D (X. Zhang & Curtis, 2020b; Urozayev et al., 2022; Wang et al., 2023) and in 3D (X. Zhang et al., 2023; Lomas et al., 2023), and survey or experimental design (Strutz & Curtis, 2023). In addition, a variety of other methods that train neural networks to emulate inverse operators, such that they produce an approximation to the posterior pdf of a problem given any recorded data set as input, could be regarded as variational methods (Devilee et al., 1999; Meier et al., 2007a, 2007b; A. K. Ray & Biswal, 2010; Shahraeeni & Curtis, 2011; Shahraeeni et al., 2012; de Wit et al., 2013; Käufl et al., 2014, 2016; Earp & Curtis, 2020; Earp et al., 2020; Cao et al., 2020; Lubo-Robles et al., 2021; X. Zhang & Curtis, 2021b; Hansen & Finlay, 2022; Bloem et al., 2023).

The performance of variational inference methods depends on the complexity and expressiveness of the predefined variational family. There is an inherent trade-off involved in selecting a tractable set of distributions: increasing the capacity of the variational family to approximate the posterior distribution usually also increases the complexity of the optimisation problem. In most variational methods, the approximating family is fixed

and constrained in ways which might exclude neighbourhoods surrounding the posterior distribution, preventing an accurate approximation to the true posterior distribution, no matter for how long the algorithm is run (F. Guo et al., 2016; Miller et al., 2017). This mismatch between the variational family and the true posterior pdf often results in underestimation of posterior variances of the model parameters and an inability to capture posterior correlations (Miller et al., 2017). For instance, the mean field approximation is commonly employed in variational methods in order to simplify the optimisation problem. This assumes a factorised structure for the variational distribution such as a Gaussian distribution with a diagonal covariance matrix. However, the method ignores correlation between different parameters and can therefore yield poor inversion results (Bishop, 2006; Blei et al., 2017; X. Zhang et al., 2023). The trend in defining an expressive variational family has mainly focused on designing more complex models, often using neural network based structures, to achieve greater flexibility. Examples of such models include *normalising flows* (Rezende & Mohamed, 2015) and their improved versions (Dinh et al., 2015; Kingma et al., 2016; Durkan et al., 2019; Kobyzev et al., 2019; Papamakarios et al., 2019). However, building effective variational models and solving the corresponding optimisation problems, which involve a large number of parameters to be optimised, pose significant challenges.

A mixture model is a weighted sum of component probability distributions, and is useful because a general mixture model has the capability to represent any complex probability distribution to any desired level of accuracy (Bishop, 1994, 2006). It is therefore reasonable to construct a variational family using a finite mixture of simple and parametric component distributions such as Gaussians. However, directly optimising a mixture model is a non-convex problem, so components can easily become trapped in suboptimal solutions. Additionally, it is challenging to determine the appropriate number of mixture components in advance.

Recently, a variational method called *Boosting Variational Inference* (BVI – F. Guo et al., 2016; Miller et al., 2017) has been investigated, which draws inspiration from classical gradient boosting techniques (Friedman, 2001; Meir & Rätsch, 2003). BVI starts by fitting a single component (a single variational distribution such as a Gaussian); this is equivalent to an existing method called automatic differential variational inference (ADVI: Kucukelbir et al., 2017). BVI then iteratively enhances that model by adding a new component distribution at each iteration. As more components are included, the posterior

approximation becomes progressively more accurate, in theory thereby improving the results offered by ADVI. An efficient, greedy algorithm is implemented by fixing the solution from the previous iteration, and optimising only the shape of the new component and its relative weight at each iteration. This approach avoids the need to design complex variational models a priori, but requires an additional optimisation for each added component. Similar to conventional mixture models, BVI is capable of capturing multimodality and incorporating rich covariance structures. However, unlike conventional methods, BVI simplifies the objective function by focusing solely on the optimisation of a single new component at each step (Locatello, Khanna, et al., 2018). This makes the optimisation process more manageable and facilitates the construction of an expressive variational family.

BVI was originally proposed in two independent works (Miller et al., 2017; F. Guo et al., 2016). Miller et al. (2017) employed the re-parametrisation trick (Kingma & Welling, 2014) to jointly optimise the variational parameters of the new component and the corresponding weight coefficient. However, this method is highly sensitive to initialisation: a new component should be initialised in a region that is under-represented by the previous components and an appropriate initial weight should be close to the proportion of the probability mass in that region. On the other hand, F. Guo et al. (2016) pointed out the non-convexity of jointly optimising these two parameters, making it challenging in general. They proposed a two-step approach where the new component is first optimised using typical gradient descent, and then the weight is optimised while keeping the new component fixed. Subsequently, Locatello, Khanna, et al. (2018) investigated the convergence properties of BVI from a modern optimisation viewpoint and established connections to the classic Frank-Wolfe framework (Frank & Wolfe, 1956; Jaggi, 2013). To ensure convergence, they imposed restrictions on the mixture components by using truncated distributions, such as truncated Gaussians. In follow-up work, Locatello, Dresdner, et al. (2018) relaxed this condition and proposed a modified objective function for variational optimisation, making BVI suitable for black box solvers (Ranganath et al., 2014). Giaquinto and Banerjee (2020) used parametric distribution models called normalising flows (Rezende & Mohamed, 2015) as mixture components, which improved the performance of existing flows based models. On the other hand, Campbell and Li (2019) proposed an alternative BVI scheme based on the *Hellinger distance* (Ghosal et al., 2000) instead of the KL divergence.

Previous studies in geophysics demonstrated that ADVI can be implemented efficiently and provides results that are straightforward to interpret. However, while ADVI provides an accurate posterior mean model, it tends to underestimate uncertainties (X. Zhang & Curtis, 2020a; X. Zhao et al., 2021). In this paper, our goal is to investigate whether the performance of ADVI can be improved while preserving its advantages by deploying the boosting strategy.

This paper is organised as follows. In section 2, we provide an introduction to variational Bayesian inversion and establish the BVI framework. We analyse the analytical properties of the posterior distribution and demonstrate the use of BVI for solving interrogation problems using representative samples obtained from BVI components. In subsequent sections we apply the method to two typical geophysical inversion problems: travel time tomography and full waveform inversion, and we compare the results to those obtained by using other existing methods. Finally, we discuss our findings and draw conclusions based on our study.

3 Methodology

3.1 Variational Bayesian Inversion

Bayesian inference solves inverse problems in a probabilistic manner by evaluating the so-called *posterior* probability distribution function (pdf) using Bayes' rule:

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

where $p(\mathbf{m})$ is the *prior* distribution of model parameters $\mathbf{m}$, which describes our knowledge about $\mathbf{m}$ prior to the inversion. The conditional probability $p(\mathbf{d}_{obs}|\mathbf{m})$ is the *likelihood* of observing data $\mathbf{d}_{obs}$ given an Earth model $\mathbf{m}$. The denominator $p(\mathbf{d}_{obs}) = \int_{\mathbf{m}} p(\mathbf{d}_{obs}|\mathbf{m})p(\mathbf{m})d\mathbf{m}$ is a normalisation constant called the *evidence*. By combining these three terms on the right hand side, we obtain the *posterior* distribution $p(\mathbf{m}|\mathbf{d}_{obs})$, which describes the probability of all possible models that are consistent with the observed data, prior information and physical forward functions used to evaluate the likelihood. Therefore, Bayesian inference provides a full inversion solution and quantifies the post inversion state of uncertainty.

Variational inference solves Bayesian problems by estimating the fixed but unknown posterior pdf. The variational goal is to select one optimal distribution $q^*(\mathbf{m})$ that best

approximates the posterior pdf within a family of known distributions (called the variational family) $\mathcal{Q}(\mathbf{m}) = \{q(\mathbf{m})\}$. This can be accomplished by finding the distribution $q(\mathbf{m})$ that minimises the following Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between the variational and posterior distributions:

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

The KL-divergence measures the distance between two distributions $q(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d}_{obs})$. It has the property $\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] \geq 0$, with equality only when $q(\mathbf{m}) = p(\mathbf{m}|\mathbf{d}_{obs})$. Evaluating the KL-divergence requires that the posterior probability $p(\mathbf{m}|\mathbf{d}_{obs})$ is calculated, which is infeasible in many problems since the evidence term $p(\mathbf{d}_{obs})$ is often intractable. However, it can be shown that minimising the KL-divergence is equivalent to maximising the *evidence lower bound* of $\log p(\mathbf{d}_{obs})$ (ELBO$[q(\mathbf{m})]$) defined as:

$$\mathrm{ELBO}[q(\mathbf{m})] = \mathbb{E}_{q(\mathbf{m})}[\log p(\mathbf{m}, \mathbf{d}_{obs})] - \mathbb{E}_{q(\mathbf{m})}[\log q(\mathbf{m})] \tag{3}$$

This only requires that the joint probability $p(\mathbf{m}, \mathbf{d}_{obs})$ is calculated, which is computationally tractable (Blei et al., 2017). By maximising equation 3 with respect to $q(\mathbf{m})$, we can estimate fully probabilistic solutions to Bayesian inverse problems using optimisation methods.

It is evident that the accuracy of variational inference depends on the expressiveness of the variational family $\mathcal{Q}(\mathbf{m})$. However, increasing the complexity of $\mathcal{Q}(\mathbf{m})$ to improve accuracy also tends to make the optimisation problem more challenging, or at least leads to higher-dimensional inverse problems. In the next section we will demonstrate how to mitigate this issue by employing mixtures of simpler distributions as the variational family.

3.2 Boosting Variational Inference

In boosting variational inference (BVI), we define the variational family to comprise the set of distributions that can be represented by a mixture of n simpler component distributions

$$q^n(\mathbf{m}) = \sum_{i=1}^{n} w_i g_i(\mathbf{m}) \tag{4}$$

where each $g_i(\mathbf{m})$ represents a chosen mixture component. The component pdfs are chosen to be parametric (meaning that an explicit formula describes their form, with pa-

rameters that define their shape). In this work we choose Gaussian component distributions $g_i(\mathbf{m}) = \mathcal{N}(\mathbf{m}; \mu_i, \Sigma_i)$ parametrised by a mean vector μ_i and a covariance matrix Σ_i. The weight w_i controls the magnitude of the contribution of each component $g_i(\mathbf{m})$, satisfying $0 \leq w_i \leq 1$ and $\sum_{i=1}^{n} w_i = 1$. Remarkably, the mixture in equation 4 can approximate any target distribution to any level of accuracy, even when using a simple base distribution $g_i(\mathbf{m})$ (Bishop, 1994; Meier et al., 2007b; Shahraeeni & Curtis, 2011; Earp & Curtis, 2020; Earp et al., 2020).

Directly maximising $\text{ELBO}[q^n(\mathbf{m})]$ with respect to the variational parameters $\{w_i, g_i(\mathbf{m}); i = 1, 2, ..., n\}$ is a non-convex problem so algorithms may converge to local minima at which one component dominates while the weights of other components become negligible (F. Guo et al., 2016). The gradient boosting approach (Friedman, 2001; Meir & Rätsch, 2003) can be used to solve this problem. The main idea is to sequentially add components to an ensemble, each being used to correct errors of its predecessors. Inspired by this, we determine an optimal variational distribution $q_n(\mathbf{m})$ through an iterative procedure, adding one new component distribution to the mixture model at each step. The procedure begins with a single component $q^1(\mathbf{m}) = g_1(\mathbf{m})$ with $w_1 = 1$. We fit $g_1(\mathbf{m})$ using a traditional variational objective function (Blei et al., 2017; C. Zhang et al., 2018). In each subsequent step $t = 2, 3, ..., n$, BVI adds one new component g_t to the mixture model, with weight $w_t \in [0, 1]$. The new distribution $q^t(\mathbf{m})$ is constructed by combining the previous mixture distribution $q^{t-1}(\mathbf{m})$, weighted by $(1-w_t)$, with the new component $g_t(\mathbf{m})$ weighted by w_t:

$$q^t(\mathbf{m}) = (1 - w_t)q^{t-1}(\mathbf{m}) + w_t g_t(\mathbf{m}) \tag{5}$$

We then maximise $\text{ELBO}[q^t(\mathbf{m})]$ with respect to w_t and g_t.

Since jointly optimising w_t and $g_t(\mathbf{m})$ is also a non-convex problem, we adopt a sequential approach which finds the optimal component $g_t(\mathbf{m})$ first, then finds the corresponding weight w_t. Based on equation 5, we treat the new mixture pdf $q^t(\mathbf{m})$ as a perturbation from the current distribution $q^{t-1}(\mathbf{m})$, where the component distribution $g_t(\mathbf{m})$ describes the shape of the perturbation and $w_t \in [0, 1]$ describes the size of the perturbation. We take the first-order Taylor expansion of $\text{ELBO}[q^t(\mathbf{m})]$ around $q^{t-1}(\mathbf{m})$:

$$\begin{aligned}\text{ELBO}[q^t(\mathbf{m})] &= \text{ELBO}[q^{t-1}(\mathbf{m}) + w_t g_t(\mathbf{m}) - w_t q^{t-1}(\mathbf{m})] \\ &= \text{ELBO}[q^{t-1}(\mathbf{m})] + w_t \left\langle g_t(\mathbf{m}), \nabla \text{ELBO}[q^{t-1}(\mathbf{m})] \right\rangle - w_t \left\langle q^{t-1}(\mathbf{m}), \nabla \text{ELBO}[q^{t-1}(\mathbf{m})] \right\rangle + o(w_t^2)\end{aligned} \tag{6}$$

where $\langle x(\theta), y(\theta) \rangle = \int x(\theta) y(\theta) d\theta$ calculates the inner product between functions $x(\theta)$ and $y(\theta)$. Term $\nabla \text{ELBO}[q^{t-1}(\mathbf{m})] = \log \dfrac{p(\mathbf{m}, \mathbf{d}_{obs})}{q^{t-1}(\mathbf{m})}$ is the functional gradient of the ELBO with respect to $q^{t-1}(\mathbf{m})$. In order to maximise $\text{ELBO}[q^t(\mathbf{m})]$ in equation 6, we must choose the $g_t(\mathbf{m})$ that maximises $\left\langle g_t(\mathbf{m}), \nabla \text{ELBO}[q^{t-1}(\mathbf{m})] \right\rangle$ since $q^{t-1}(\mathbf{m})$ is fixed. That is, we choose $g_t(\mathbf{m})$ to match the direction of $\nabla \text{ELBO}[q^{t-1}(\mathbf{m})]$. Then, we obtain $g_t(\mathbf{m})$ by

$$g_t(\mathbf{m}) = \underset{g_t(\mathbf{m})}{\operatorname{argmax}} \left\langle g_t(\mathbf{m}), \nabla \text{ELBO}[q^{t-1}(\mathbf{m})] \right\rangle = \underset{g_t(\mathbf{m})}{\operatorname{argmax}} \left\langle g_t(\mathbf{m}), \log \frac{p(\mathbf{m}, \mathbf{d}_{obs})}{q^{t-1}(\mathbf{m})} \right\rangle \tag{7}$$

Direct maximisation of the inner product in equation 7 is ill-posed and can lead to $g_t(\mathbf{m})$ degenerating into a narrow distribution or even a single point mass which only has non-zero probability value at the maximum of $\nabla \text{ELBO}[q^{t-1}(\mathbf{m})]$ – a degenerate probability distribution that has zero width. To solve this problem, we introduce an additional regularisation term that involves the entropy of $g_t(\mathbf{m})$, given by the negative scalar product of $g_t(\mathbf{m})$ and $\log g_t(\mathbf{m})$:

$$\begin{aligned} g_t(\mathbf{m}) &= \underset{g_t(\mathbf{m})}{\operatorname{argmax}} \left\langle g_t(\mathbf{m}), \nabla \text{ELBO}[q^{t-1}(\mathbf{m})] \right\rangle - \lambda \left\langle g_t(\mathbf{m}), \log g_t(\mathbf{m}) \right\rangle \\ &= \underset{g_t(\mathbf{m})}{\operatorname{argmax}} \mathbb{E}_{g_t(\mathbf{m})}[\log p(\mathbf{m}, \mathbf{d}_{obs})] - \mathbb{E}_{g_t(\mathbf{m})}[\log q^{t-1}(\mathbf{m})] - \lambda \mathbb{E}_{g_t(\mathbf{m})}[\log g_t(\mathbf{m})] \end{aligned} \tag{8}$$

where $\mathbb{E}_{g_t(\mathbf{m})}[\cdot]$ calculates the expectation of any function with respect to $g_t(\mathbf{m})$. Parameter λ is a regularisation factor that controls the weight of the entropy term. Entropy measures the uncertainty represented by any pdf, so by maximising the entropy we ensure that the pdf does not collapse to a narrow, effectively degenerate distribution. We refer to the objective function in equation 8 as the *residual evidence lower bound* ($\text{RELBO}[g_t(\mathbf{m})]$)

$$\text{RELBO}[g_t(\mathbf{m})] := \mathbb{E}_{g_t(\mathbf{m})}[\log p(\mathbf{m}, \mathbf{d}_{obs})] - \mathbb{E}_{g_t(\mathbf{m})}[\log q^{t-1}(\mathbf{m})] - \lambda \mathbb{E}_{g_t(\mathbf{m})}[\log g_t(\mathbf{m})] \tag{9}$$

The expectation terms and their gradients in both equations 3 and 9 can be estimated using Monte Carlo integration (details can be found in X. Zhao et al., 2021). Since we would normally perform many iterations to maximise these two equations, we can use a relatively small number of samples per iteration (even only a single sample – Kucukelbir et al., 2017). By maximising this objective function, we can find an optimal $g_t(\mathbf{m})$ at each step of the algorithm.

In equation 7, $\log \frac{p(\mathbf{m}, \mathbf{d}_{obs})}{q^{t-1}(\mathbf{m})}$ describes the residual discrepancy between the current variational distribution $q^{t-1}(\mathbf{m})$ and the joint probability distribution $p(\mathbf{m}, \mathbf{d}_{obs}) = p(\mathbf{d}_{obs}) p(\mathbf{m}|\mathbf{d}_{obs})$ which is equal to the unnormalised posterior distribution $p(\mathbf{m}|\mathbf{d}_{obs})$ according to equation 1. If $q^{t-1}(\mathbf{m})$ is proportional to the true (normalised) posterior pdf, i.e., $q^{t-1}(\mathbf{m}) \propto$

$p(\mathbf{m}|\mathbf{d}_{obs})$ everywhere in the parameter space, the above residual would be constant. However, in most situations this residual has peaks where the current variational distribution underestimates the posterior distribution, and has basins where $q^{t-1}(\mathbf{m})$ overestimates $p(\mathbf{m}|\mathbf{d}_{obs})$. By introducing a new component $g_t(\mathbf{m})$, we aim to add density to regions where $q^{t-1}(\mathbf{m})$ underestimates and (through the relative weighting scheme in equation 5) weaken regions where it overestimates the posterior pdf (this can be proven using information theory). The goal is to find an optimal $g_t(\mathbf{m})$ that maximises $(\mathbb{E}_{g_t(\mathbf{m})}[\log p(\mathbf{m}, \mathbf{d}_{obs})] - \mathbb{E}_{g_t(\mathbf{m})}[\log q^{t-1}(\mathbf{m})])$, which can be interpreted as minimising the cross entropy of $g_t(\mathbf{m})$ with respect to $p(\mathbf{m}, \mathbf{d}_{obs})$ and maximising that with respect to $q^{t-1}(\mathbf{m})$. In other words, $g_t(\mathbf{m})$ should be as close as possible to the (unnormalised) posterior distribution, and at the same time should be sufficiently different from the current approximation – it should capture the aspects of the posterior pdf that the current mixture distribution cannot yet approximate. This allows BVI to gradually improve the accuracy of the variational distribution by iteratively adding new components.

There are three commonly used methods to determine the weight coefficient $w_t \in [0, 1]$ for the new component in BVI. The first method uses an empirical formula to guarantee a series of weights for each additional component (Locatello, Dresdner, et al., 2018; Locatello, Khanna, et al., 2018):

$$w_t = \frac{2}{t+1}, \quad t = 1, 2, ..., n \tag{10}$$

Although this formula abandons the ideal of finding optimal weight coefficients, it provides a straightforward approach to update the weight. Any error caused by non-optimality of this can be corrected by the introduction of additional components to the mixture distribution.

The second method for updating weight coefficients involves a line search. The weight is updated by maximising $\mathrm{ELBO}[q^t(\mathbf{m})]$ (note this is not maximising $\mathrm{RELBO}[g_t(\mathbf{m})]$ with respect to w_t) (F. Guo et al., 2016):

$$w_t^{(k+1)} = w_t^{(k)} + \frac{b}{k} \nabla_{w_t} \mathrm{ELBO}[q^t(\mathbf{m})] \tag{11}$$

where superscripts (k+1) and (k) represent two consecutive iterations, and b is the initial step size decayed by $1/k$. The method to calculate $\nabla_{w_t} \mathrm{ELBO}[q^t(\mathbf{m})]$ is provided in Appendix Appendix A.

The third method, updates the weights for all components when each new component is added to the mixture model (Locatello, Dresdner, et al., 2018):

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \frac{b}{k}\nabla_{\mathbf{w}}\text{ELBO}[q^t(\mathbf{m})] \tag{12}$$

where $\mathbf{w} = [w_1, w_2, ..., w_t]^T$ is a vector containing the weights of all components. The gradient term can be calculated similarly to the line search method (Appendix Appendix A).

Once the weight coefficient is obtained the new mixture distribution $q^t(\mathbf{m})$ can be constructed by combining the new component $g_t(\mathbf{m})$ with the existing mixture distribution using Equation 5.

3.3 BVI using Gaussian Components

In this work, we use Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ as the mixture components. A mixture of Gaussians is capable of representing any target distributions (Bishop, 2006). For each component, we optimise a mean vector μ and a covariance matrix Σ by maximising the RELBO in equation 9, and below we test the three schemes to determine the weights. Once convergence is achieved, the obtained Gaussian component is added to form the new mixture distribution.

Considering that model parameters in many geophysical inverse problems are subject to hard constraints (e.g., seismic velocity must be greater than zero), and Gaussian distributions and their mixtures are defined in the unbounded space of Real numbers, we apply the inverse logistic function to transform the mixture distribution from the space of Real numbers into the constrained space (X. Zhang & Curtis, 2020a). This transformation is defined as:

$$\begin{cases} \mathbf{m} &= f(\mathbf{z}) = \mathbf{a} + \dfrac{\mathbf{b}-\mathbf{a}}{1+\exp(-\mathbf{z})} \\ \log p(\mathbf{m}|\mathbf{d}_{obs}) &= \log p(\mathbf{z}) - \log|\det \partial_{\mathbf{z}} f(\mathbf{z})| \\ &= \log \sum_i w_i \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i) - \log|\det \partial_{\mathbf{z}} f(\mathbf{z})| \end{cases} \tag{13}$$

where $\mathbf{m}$ and $\mathbf{z}$ are model parameters in the constrained and unconstrained spaces, respectively. Hyper-parameters $\mathbf{a}$ and $\mathbf{b}$ are lower and upper bounds on $\mathbf{m}$, and are fixed during optimisation. In the second equation, $p(\mathbf{z})$ is the mixture distribution obtained using BVI in the space of Real numbers. Term $|\det(\cdot)|$ calculates the absolute value of the determinant of the Jacobian matrix corresponding to this transform, which accounts

for the volume change. We use equation 13 to transform each parameter in vector $\mathbf{z}$ to that in $\mathbf{m}$, such that the corresponding Jacobian matrix is a diagonal matrix and its determinant is analytic and easy to calculate. This means that the correlation information of vector $\mathbf{m}$ is purely determined by the covariance matrices Σ_i (therefore we do not lose posterior correlation by applying this transform). As a result, the posterior distribution modelled using the proposed BVI algorithm, as well as its statistical properties, can be represented analytically.

As noted above, BVI becomes automatic differential variational inference (ADVI – Kucukelbir et al., 2017) when only a single Gaussian component is used. ADVI also provides an analytic approximation to the posterior distribution, and usually seems to estimate the mean model accurately. However, due to its theoretical assumption of a single Gaussian distribution in the unconstrained space, the method usually underestimates parametric uncertainty around the mean. By adding more Gaussian components we regard BVI as an iterative method to enhance the performance of ADVI.

Figure 1 shows a toy example that demonstrates the performance of BVI with Gaussian components. The target posterior distribution is a mixture of two Gaussian distributions: $p(x) = 0.5\mathcal{N}(x; -1, 0.4) + 0.5\mathcal{N}(x; 1, 0.6)$, represented by black line in Figure 1. To apply BVI, we first optimise the initial component by maximising the ELBO in equation 3, which is equivalent to a conventional variational problem. The dashed orange line in Figure 1 shows the first component after convergence. It is evident that this single Gaussian distribution fails to approximate the bimodal posterior distribution accurately, highlighting the limitations of ADVI, and variational methods in general when an inappropriate variational family that does not include the true posterior pdf is chosen.

We then iteratively add more Gaussian components to the mixture model by maximising the RELBO using equation 9. We compare the performance of the 3 different weight calculation methods in equations 10, 11 and 12. In each test, we boost the posterior distribution by adding 40 Gaussian components so as effectively to ensure full convergence of BVI. Although it looks redundant to use 40 Gaussian components to approximate a mixture of two Gaussian distributions, we generally do not know the true posterior distribution, so do not know when to stop the algorithm unless convergence is observed. The results using equations 10, 11 and 12 are shown by the dashed red, blue and

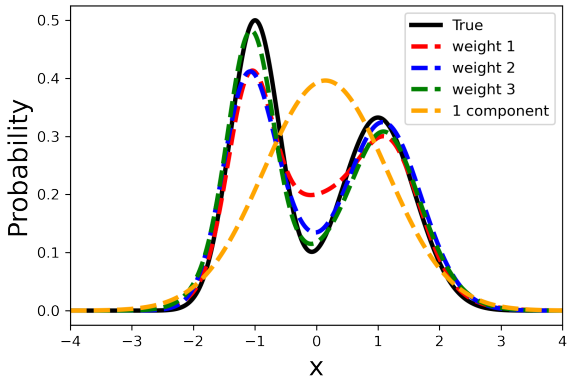

Figure 1. BVI results obtained using 3 different weight calculation methods. Black line represents the target distribution, while the dashed orange line shows the result from conventional variational inference without boosting, which uses a single Gaussian component (essentially the ADVI method). Dashed red, blue, and green lines correspond to the results obtained using different weight calculation methods in equations 10, 11 and 12. The last two methods yield better results but require additional computations.

green lines in Figure 1, respectively. All three methods provide a fair approximation to the true posterior distribution, with the first method performing the worst and the third method performing the best. However, the second and third methods require additional computations to estimate the gradient of the ELBO in equations 11 and 12, which involve evaluating the posterior distribution many times. This example demonstrates that even the simple fixed weight method significantly improves upon the initial variational solution (dashed orange line in Figure 1) without any additional computational complexity. Since we are interested in applying these methods to high-dimensional problems, minimising computational complexity is paramount if we are to find meaningful solutions. In the subsequent inversion examples, we therefore employ the fixed weight calculation method, but highlight that in other circumstances practitioners might prefer a different choice.

3.4 Probabilistic Interrogation using BVI

In scientific investigations, the ultimate goal is usually to answer some specific, low-dimensional questions. In geophysics, such questions are typically answered by interpreting imaging or inversion results, but this often leads to biased answer because human

interpretation is a subjective process and since usually only one single instance or statistic of a model is considered during interpretation. Interrogation theory (Arnold & Curtis, 2018) offers a systematic approach to answer high-level questions. It combines inverse theory, decision theory, elicitation theory and experimental design theory to optimise scientific investigations, with the overall goal to obtain the most informative answers to scientific inquiries. X. Zhao et al. (2022) and X. Zhang and Curtis (2022) exemplified the theory by answering a specific type of question: *what is the size of a subsurface body?* For a more comprehensive understanding of interrogation theory and its implementation, we recommend readers to refer to the above three papers.

In interrogation theory, a utility function $U(a)$ is defined which quantifies the net benefits associated with accepting any possible answer a. The optimal answer a^* is found by maximising this utility function within the space of possible answer: $a^* = \arg\max_{a \in \mathbb{A}} U(a)$. To reduce the complexity of this maximisation problem, Arnold and Curtis (2018) introduced a target space $\mathbb{T}$ such that the scientific question Q can be answered directly within this space. They defined a target function $T(\mathbf{m})$ that maps the high dimensional model parameter $\mathbf{m}$ to a low dimensional target space parameter values t. A simplified utility function can then be defined as $U(a|t, E_d)$, where the utility function is conditioned on the experimental design E_d to account for the cost of conducting the experiment. One of the utility functions considered in Arnold and Curtis (2018) is a negative squared error function:

$$U(a|t, E_d) = U(a|t) = -(a - t)^2 \tag{14}$$

in which t is considered to represent the true state in the target space. The above utility function is useful when value t represents exactly the answer that we seek, and the utility is maximised when the estimated answer a is equal to the true state t. However, in problems with uncertain solutions a single set of values that represent the true state is never known, so it is necessary to find an optimal answer by maximising the utility on average across possible true states. This formulation leads to an analytical expression for the optimal answer:

$$\begin{aligned} a^* &= \mathbb{E}[T(\mathbf{m}|\boldsymbol{f}(\mathbf{m}), C)|\mathbf{d}_{obs}, E_d] \\ &= \sum_{\boldsymbol{f}(\mathbf{m}), C} \int_{\mathbf{m}} T(\mathbf{m}|\boldsymbol{f}(\mathbf{m}), C) p(\mathbf{m}, \boldsymbol{f}(\mathbf{m}), C|\mathbf{d}_{obs}, E_d)\ d\mathbf{m} \\ &= \int_{\mathbf{m}} T(\mathbf{m}) p(\mathbf{m}|\mathbf{d}_{obs})\ d\mathbf{m} \end{aligned} \tag{15}$$

where $\boldsymbol{f}(\mathbf{m})$ is the forward function that relates the model space and data space and C represents any particular choice of mathematical or computational algorithms used to solve the forward, inverse, and design problems. X. Zhao et al. (2022) showed that relying solely on the results from one single algorithm can lead to biased interrogations. Therefore, they combined different such algorithms to mitigate bias in the optimal answer. In this work we simplify the analysis by considering a single forward function, a single choice of algorithms to find the solution, and a fixed experimental design. This simplification allows us to omit $\boldsymbol{f}(\mathbf{m})$, C and E_d in the subsequent derivation, but it is easy to extend our conclusions to the cases involving multiple forward functions, computational algorithms, and experimental designs if desired. The third line of Equation 15 states that the optimal answer corresponds to the posterior expectation of the target function, and different forms for this expression result from different choices of utility function in equation 14 (Arnold & Curtis, 2018).

In previous works (X. Zhao et al., 2022; X. Zhang & Curtis, 2022), the target function was assumed to be deterministic, meaning that the target value was uniquely determined given a model sample $\mathbf{m}$. Consequently, uncertainty in the answer was attributed solely to uncertainty in the inversion process. In reality, the definition of the target function often incorporates knowledge from a variety of experts, which introduces human biases and uncertainties (O'Hagan et al., 2006; Polson & Curtis, 2010; Bond et al., 2012). In an interrogation example below, we show that biased judgments from different individuals can lead to incorrect answers. To address the uncertainty in the final answer caused by the deterministic target function in order to mitigate bias, we use fully probabilistic target functions.

Define a random variable τ to represent different states of possible target function values, with an associated probability distribution function $p(\tau)$. This approach characterizes the nondeterministic behaviour of the target function and addresses the inherent uncertainty. The optimal answer, which calculates the posterior mean of the summarized state τ, is given by

$$\begin{aligned} a^* = \mathbb{E}[\tau|\mathbf{d}_{obs}] &= \int_{\mathbf{m}}\int_{\tau} \tau p(\tau, \mathbf{m}|\mathbf{d}_{obs})\ d\mathbf{m}d\tau \\ &= \int_{\mathbf{m}}\int_{\tau} \tau p(\tau|\mathbf{m}, \mathbf{d}_{obs}) p(\mathbf{m}|\mathbf{d}_{obs})\ d\mathbf{m}d\tau \\ &= \int_{\mathbf{m}}\int_{\tau} \tau p(\tau|\mathbf{m}) d\tau p(\mathbf{m}|\mathbf{d}_{obs})\ d\mathbf{m} \end{aligned} \tag{16}$$

In the first line, we extend the deterministic target function from equation 15 to a probabilistic formulation using the law of total probability $p(x) = \int_y p(x, y)dy$. Following Siahkoohi et al. (2022), we assume that the target function value τ and the observed data $\mathbf{d}_{obs}$ are conditional independent given the model parameter $\mathbf{m}$, when using interrogation theory to solve a decision problem that maps specific information from the inversion results. This assumption leads to the third line in equation 16. The inner integral $\mathbb{E}[\tau|\mathbf{m}] := \int_\tau \tau p(\tau|\mathbf{m})d\tau$ captures uncertainty in the target function value which represents the uncertainty in the interrogation process, while the outer integral accounts for uncertainty in the inversion process. Note that the above conditional independence assumption does not hold when solving a design problem using interrogation theory, as the optimal answer, which is the best design in this context, depends on the different datasets that would be observed given any considered design (Arnold & Curtis, 2018; Strutz & Curtis, 2023).

To summarise, equation 16 can be viewed as a more general version of the original interrogation framework, achieved by considering a random variable τ with a probability distribution function $p(\tau)$ which allows for the incorporation of uncertainty in the target function. When $p(\tau)$ is defined as a Dirac delta function, denoted by $p(\tau) = \delta_{(\tau=T)}(\tau)$, where T represents the deterministic target function in equation 15, equation 16 reduces to equation 15. Thus, equation 16 encompasses the original framework as a special case when the target function is deterministic.

Monte Carlo integration can be used to evaluate equation 16. First, we draw random model samples from the posterior distribution $p(\mathbf{m}|\mathbf{d}_{obs})$. Given each posterior sample, we generate an ensemble of possible target function values from $p(\tau|\mathbf{m})$. By combining these target values, the posterior distribution of the answer a can be obtained, and the optimal answer a^* to the question Q is the expectation of this distribution.

In the previous sections we showed that BVI provides an analytic expression of the posterior distribution. Directly incorporating this analytic result into equations above using either the deterministic or probabilistic target function is unfortunately non-trivial because the definition of the target function often contains some conceptual process which is easier to evaluate using posterior samples and is difficult to formulate as an explicit expression. In an interrogation example provided below, the calculation of the target function requires the largest continuous body within a velocity model to be found, which is

not straightforward to perform using the analytic posterior expression. To address this, we propose an implicit approach. In the BVI framework the posterior distribution is approximated in the Real (unconstrained) space as a mixture of Gaussian distributions, and significant information is captured by the mean vectors μ_i of the set of components. We transform these mean vectors μ_i back to the constrained space using equation 13 after which the transformed vectors $\mathbf{m}_i$ can be treated as a set of representative samples, weighted by the coefficient w_i corresponding to each Gaussian component in BVI. We use these samples to partly represent the full posterior pdf, and the optimal answer in equations 15 and 16 can be approximated as

$$a^* = \int_{\mathbf{m}} T(\mathbf{m}) p(\mathbf{m}|\mathbf{d}_{obs})\ d\mathbf{m} \approx \sum_i w_i T(\mathbf{m_i}) \tag{17}$$

for the deterministic case, and

$$\begin{aligned} a^* &= \int_{\mathbf{m}} \int_{\tau} \tau p(\tau|\mathbf{m}) d\tau p(\mathbf{m}|\mathbf{d}_{obs})\ d\mathbf{m} \\ &\approx \sum_i w_i \int_{\tau} \tau p(\tau|\mathbf{m}_i) d\tau = \sum_i w_i \mathbb{E}[\tau|\mathbf{m}_i] \end{aligned} \tag{18}$$

for the probabilistic case. Since only tens of components are used in BVI to approximate the posterior distribution, the target function is calculated using the same number of samples from BVI. This computational simplicity is particularly important when the target function itself is computationally expensive to evaluate, especially in the case of interrogation using probabilistic target function, and as we show below, equations 17 and 18 can still enable accurate interrogation.

4 Travel Time Tomography

Seismic travel time tomography is a typical non-linear geophysical inverse problem used to image the Earth's interior. The underground seismic velocity structure is mapped using measured first-arrival travel times of waves travelling between source and receiver locations. In this section, we present two tomographic examples to demonstrate the performance of BVI.

4.1 Synthetic Example

The first example is a 2D synthetic test. Figure 2 shows the true velocity model, which consists of a circular low velocity anomaly of 1 km/s surrounded by a high velocity background of 2 km/s. White triangles show the locations of 16 receivers, and assum-

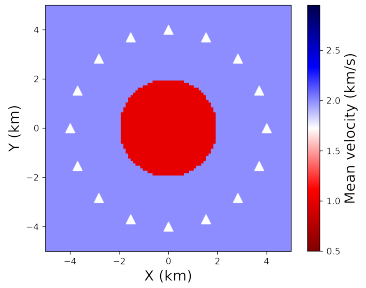

Figure 2. True velocity model of the 2D synthetic test. A low velocity circular anomaly with velocity 1 km/s is embedded within a background velocity of 2 km/s. White triangles show locations of 16 receivers (and sources), and travel times between each pair of locations form the observed data set in this example.

ing that each receiver can also be used as a virtual source using seismic interferometry (Campillo & Paul, 2003; Wapenaar, 2004; Curtis et al., 2006). 120 inter-receiver first-arrival travel times of waves that travel between each pair of receiver locations form the data set for this problem. For inversion we parametrise the model parameter $\mathbf{m}$ (the velocity vector) into 21 $\times$ 21 regular grid cells with a grid size of 0.5 km in both directions. We define a Uniform prior probability distribution bounded between 0.5 and 3.0 km/s for each grid cell. The likelihood function is assumed to be a diagonal Gaussian distribution with a data uncertainty $\sigma_d = 0.05$ s for all data points. We solve the forward problem to predict synthetic data using the fast marching method (FMM – Rawlinson & Sambridge, 2005).

For BVI we use a diagonal Gaussian distribution for all component distribution, and the empirical formula in equation 10 to calculate weight coefficients. The first component is obtained by maximising the ELBO in equation 3 which is equivalent to mean field ADVI (Kucukelbir et al., 2017). In subsequent BVI iterations, we sequentially optimise new components by maximising the RELBO in equation 9. We combine the obtained Gaussian components into a mixture distribution and transform it back to the constrained space using equation 13. The resulting distribution is used to approximate the true posterior distribution. For each component, we update the diagonal Gaussian distribution for 5000 iterations, and within each iteration 2 samples are used to approximate the RELBO (or ELBO for the first component) and its gradient using Monte Carlo integration. To test the convergence performance of BVI, we greedily add 10 components

by which point the statistics of the posterior pdf show no substantial change with each iteration, as shown in Figure 3.

Figures 3a and 3b show the mean and standard deviation maps of the posterior distribution obtained using BVI with different Gaussian components. All of these maps are calculated analytically from BVI solution without drawing any posterior samples, using equation 13. Within the receiver array, the mean models effectively recover the circular low velocity anomaly and are similar to the true velocity model shown in Figure 2, even with only 1 component, which corresponds to mean field ADVI as discussed previously. However, as expected the uncertainty map obtained using one component significantly underestimates uncertainties. As we introduce more components, the posterior uncertainties increase. The mean and standard deviation maps essentially converge such that no significant changes are observed after adding $6 - 7$ components. We observe two higher uncertainty loops in the uncertainty maps: inner one is located at the boundary of the low velocity anomaly and arises from variations in anomaly shapes and velocity values among the plausible models that fit the observed data, and the other loop corresponds to the lower average velocity loop between the receiver array and the central anomaly, potentially because the observed data exhibits lower sensitivity in this region, as observed in previous studies (Galetti et al., 2015; X. Zhang & Curtis, 2020a; X. Zhao et al., 2021).

Metropolis-Hastings Markov chain Monte Carlo (MH-McMC) was also run to estimate the solution for comparison. We ran 12 chains in parallel, each drawing 1 million samples to ensure convergence. After sampling, we discard the first 500,000 samples as the burn-in period, and retain every 50th sample from the remaining samples to approximate samples of the posterior distribution. This result serves as a reference solution for this tomographic problem. Figure 4 shows the mean and standard deviation maps obtained using MH-McMC. We find that the mean models obtained from BVI and MH-McMC show similar results, and the uncertainty maps from both methods exhibit similar loop-like higher uncertainty structures. However, the uncertainties from BVI are slightly lower than those from MH-MCMC, indicating that BVI still underestimates the true uncertainty to some extent. Nevertheless, since BVI yields results comparable to MH-MCMC (which is often assumed to provide the true solution), we conclude that BVI provides an approximately correct and, more importantly, fully analytic result. Furthermore, it significantly improves upon the results obtained using mean field ADVI.

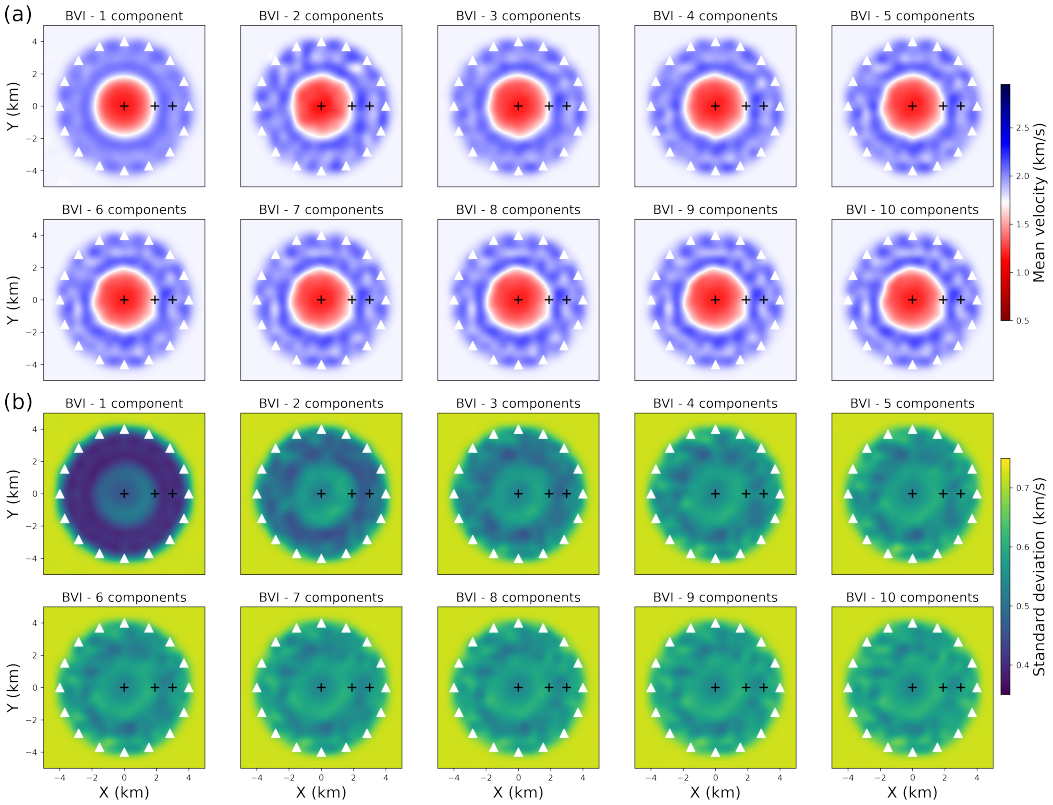

Figure 3. (a) Mean and (b) standard deviation maps of the posterior distribution obtained using BVI with different number of Gaussian components denoted in the title of each subfigure. White triangles show the 16 receiver locations and black crosses denote three specific locations whose marginal distributions are compared in Figure 5.

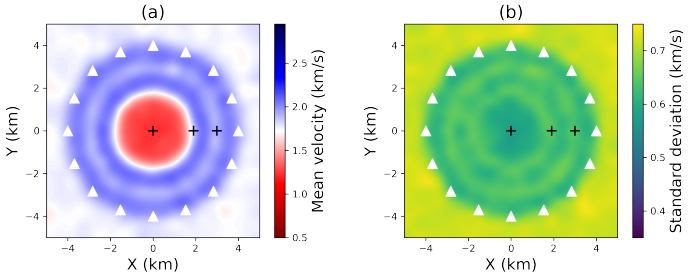

Figure 4. (a) Mean and (b) standard deviation maps obtained using MH-McMC. This result serves as the reference solution for this Bayesian tomographic problem.

In Figures 5a – 5c, we compare the marginal distributions of three representative points at (0, 0) km, (1.8, 0) km and (3.0, 0) km. These locations are denoted by black crosses in Figures 3 and 4. The first point lies within the low velocity anomaly, the second point is at the edge of the anomaly where the inner higher uncertainty loop is observed, and the last point is in the outer higher uncertainty loop. In each figure, the grey histogram shows the marginal distribution obtained using MH-McMC for reference, and dashed yellow line shows the Uniform prior pdf. For BVI, we can calculate the analytic marginal pdfs for these three points without drawing any samples. Results using 1 BVI component (mean field ADVI), 4 components, 7 components, and 10 components are depicted by blue, dashed green, dashed black, and red lines, respectively. It is evident that mean field ADVI underestimates the posterior uncertainties, particularly in Figures 5b and 5c. However, as we add more components to the mixture, the marginal pdfs become increasingly similar to those obtained from MH-McMC, especially for the third point in Figure 5c, where the red line perfectly matches the grey histogram. In Figure 5b the results obtained using BVI and McMC are a little different. While we treat the result from McMC as a reference solution for this non-linear problem, we never know the true solution because it is likely that the Monte Carlo solution has not converged in a problem of this dimensionality. Therefore, it is difficult to conclude which one of these two results is better. Nevertheless, we still observe that each new component corrects some of the residual from the previous distributions in the ensemble, apparently boosting the accuracy of the current variational distribution (hence the name, "boosting variational inference").

Figures 3 and 5 show that the results achieve a reasonable approximation to the true posterior distribution using only 7 components. Unfortunately, in real problems we do not have access to the true posterior distribution, and running a McMC test for high-dimensional problems is often infeasible. Consequently, it becomes challenging to decide when to stop adding more components. A viable approach is to monitor the convergence of the KL-divergence: after each BVI iteration, we estimate $\mathrm{KL}[q^t(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})]$ by drawing samples from the mixture distribution $q^t(\mathbf{m})$, and stop the BVI algorithm once $\mathrm{KL}[q^t(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})]$ ceases to decrease. However, accurately estimating the KL-divergence for high-dimensional problems is hindered by the curse of dimensionality. In this example, we therefore compared statistical properties that can be estimated more stably such as the mean, standard deviation, and marginal pdf of the current mixture distribution with those obtained

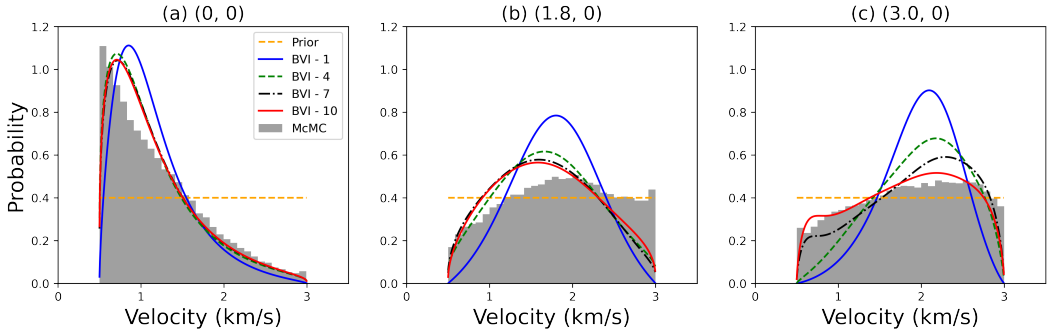

Figure 5. Marginal posterior distributions of velocity at three points located at (a) (0, 0) km, (b) (1.8, 0) km and (c) (3.0, 0) km, marked by three black crosses in Figures 3 and 4. In each figure, the grey histogram shows the marginal distribution obtained using MH-McMC, and dashed yellow line shows the prior distribution. Blue, dashed green, dashed black, and red lines show marginal distributions obtained using BVI with 1 component (corresponding to mean field ADVI), 4, 7 and 10 components, respectively.

from previous iterations. If no significant changes are observed, we assume that BVI has converged and refrain from adding more components.

4.2 Field Data Test

In a more complicated field data example we applied BVI to Love wave tomography of the British Isles. The British Isles have been extensively studied and well understood using ambient noise tomography with different inversion methods, including linearised inversion (Nicolson et al., 2012, 2014), rj-McMC (Galetti et al., 2017) and variational inference (X. Zhao et al., 2021, 2022). Therefore, this is a suitable real-data test case to evaluate the performance of the proposed method and analyse the results by comparison. We use part of the dataset created by Galetti et al. (2017): ambient noise data recorded by 61 seismometers located around the British Isles, as indicated by red triangles in Figure 6. The data were collected during three periods: 2001–2003, 2006–2007, and in 2010. The two horizontal components of the data were cross-correlated to compute Love waves between pairs of receiver stations. Subsequently, the first arrival travel times of group velocity were estimated at different periods ranging from 4 s to 15 s. Detailed information regarding the station network and data processing procedures can be

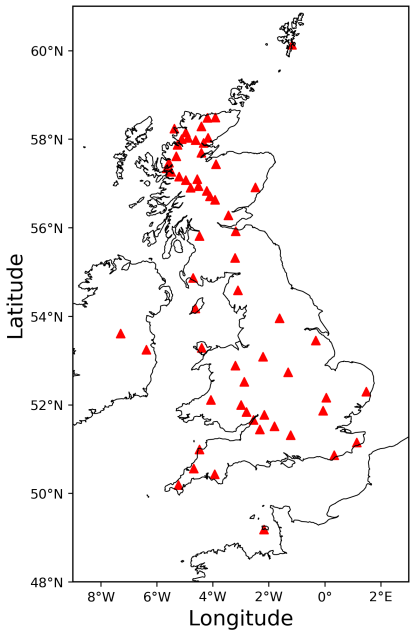

Figure 6. The location of 61 seismometers (red triangles) around the British Isles. The receiver stations are also treated as virtual sources for ambient noise interferometry to estimate inter-receiver first arrival travel times, which are used as the observed data in this test.

found in Galetti et al. (2017). For this test, we use the travel time measurements of Love waves at period of 10 s.

We parametrise the target region in Figure 6 into 37×40 regular grid cells with a spacing of $0.33°$ in both longitude and latitude directions. For each grid cell, we define a Uniform prior distribution ranging from 1.56 to 4.84 km/s: the average value of the Uniform distribution is obtained by measuring the average velocity across all valid ray paths by assuming a homogeneous medium, and the upper and lower bounds are chosen to exceed the range of velocities observed on the dispersion curves. The likelihood function is chosen to be a Gaussian distribution, and the travel time uncertainty for each inter-receiver path is estimated from the standard deviation of the estimated travel time of the corresponding station pair constructed by stacking randomly selected subsets of daily cross-correlations (Galetti et al., 2017).

Given this problem's higher dimensionality (1480) and non-linearity (due to higher noise and irregular data distribution) compared to the 2D synthetic test, BVI requires more components to converge to a reasonable approximation of the true posterior dis-

tribution. However, the greedy algorithm described in previous sections is time-consuming and does not fully use the computational power of modern compute clusters. To address this we propose an efficient implementation of BVI by running multiple independent runs in parallel, similar to McMC methods that often run independent chains in parallel. In this implementation, we start each independent BVI run from the second component, as the first component (corresponding to ADVI) has been shown to provide a stable (though inaccurate) result (Kucukelbir et al., 2017; X. Zhang & Curtis, 2020a). Each independent BVI run is initialised randomly and optimised separately, and after optimisation, the mixture distributions obtained from all runs are averaged to obtain the final approximation to the posterior distribution. This parallelisation approach allows BVI to take advantage of parallel computing capabilities while still providing analytic results.

We apply BVI and MH-McMC to this tomography problem for comparison. We run 4 independent BVI tests in parallel, and for each test, we sequentially add 5 components until the posterior distribution stops changing significantly. This results in a total of 20 Gaussian distributions used to model the posterior distribution. Again, we use a diagonal Gaussian distribution as the mixture component. Each component is updated for 5000 iterations with 2 samples used at each iteration. The weight coefficients for the mixture components are calculated using equation 10. After optimisation, we average the distributions obtained from the 4 runs to obtain the final results. To obtain results using MH-McMC, we run 10 Markov chains in parallel. Each chain consists of 1.5 million samples, with the first 1 million samples discarded as burn-in. We discard a large number of samples in the hope that the remaining samples are reasonably well distributed according to the posterior distribution. After the burn-in period every 100th sample is retained to approximate an ensemble of posterior samples.

Figures 7b and 7c show the average velocity (top row) and standard deviation (bottom row) maps of the Love wave tomography results obtained using BVI and MH-McMC. We also display the results obtained using mean field ADVI, which corresponds to the first component obtained from BVI, as shown in Figure 7a. The average velocity maps from the three methods exhibit similar features that are consistent with the known geology of the British Isles. For example, we observe a high velocity anomaly in the Scottish Highlands (6°W – 4°W and 57°N – 59°N), reflecting the crystalline metamorphic origin of the rocks in that region. A low velocity structure is observed in the area between 5°W – 3°W and 53°N – 55°N, which corresponds to the East Irish Sea sedimen-

tary basins. Several low velocity anomalies are also observed around the Midland Platform in southern England (3°W – 1°E and 50°N – 52°N), corresponding to various sedimentary basins such as the Cheshire Basin, the Anglian-London Basin, and the Wessex Basin.

The uncertainty models obtained from BVI and MH-McMC present similar patterns. For instance, lower uncertainties are observed in regions with densely placed receiver arrays such as across the Highlands and southern England. A higher uncertainty loop is observed around the East Irish Sea (4°W and 54°N) since a wide variety of different anomaly shapes and velocity values fit the observed travel time data, which is consistent with the findings from previous studies (Galetti et al., 2017; X. Zhao et al., 2021).

The results obtained from BVI and MH-McMC are similar to those from other variational methods: normalising flows and Stein variational gradient descent (SVGD) in X. Zhao et al. (2021). However, there are some small differences in the structures observed in Figures 7b and 7c compared to those obtained from rj-McMC in Galetti et al. (2017), which can be attributed to the different parametrisations used in that work (Voronoi cells versus regular cells). In the rj-McMC study (Galetti et al., 2017), 16 chains and 3 million samples per chain were used to ensure convergence. In this test, 10 chains and 1.5 million samples were used for MH-McMC. The presence of some non-smooth structures in Figure 7c compared to the smooth structures in the synthetic test (Figure 4) suggests that the chains may not have fully converged even after 1.5 million samples, and that 10 chains might not be sufficient to explore all possible parameter subspaces that fit the data. In X. Zhao et al. (2021), full rank ADVI was also applied to this problem. However, both full rank ADVI in that work and mean filed ADVI here, exhibit strong biases in the uncertainty results, with lower uncertainty than the McMC results observed everywhere inside the receiver array. In conclusion, since similar solutions have been obtained by multiple different methods, it can be assumed that BVI is capable of providing a reasonable estimate of the posterior distribution with an analytic expression, while also improving performance compared to mean field ADVI.

Table 1 compares the computational costs associated with several different methods, measured in terms of the required number of forward evaluations, since forward simulation is the most expensive part in each inversion. The computational costs of full rank ADVI, normalising flows and SVGD are obtained from X. Zhao et al. (2021), while the

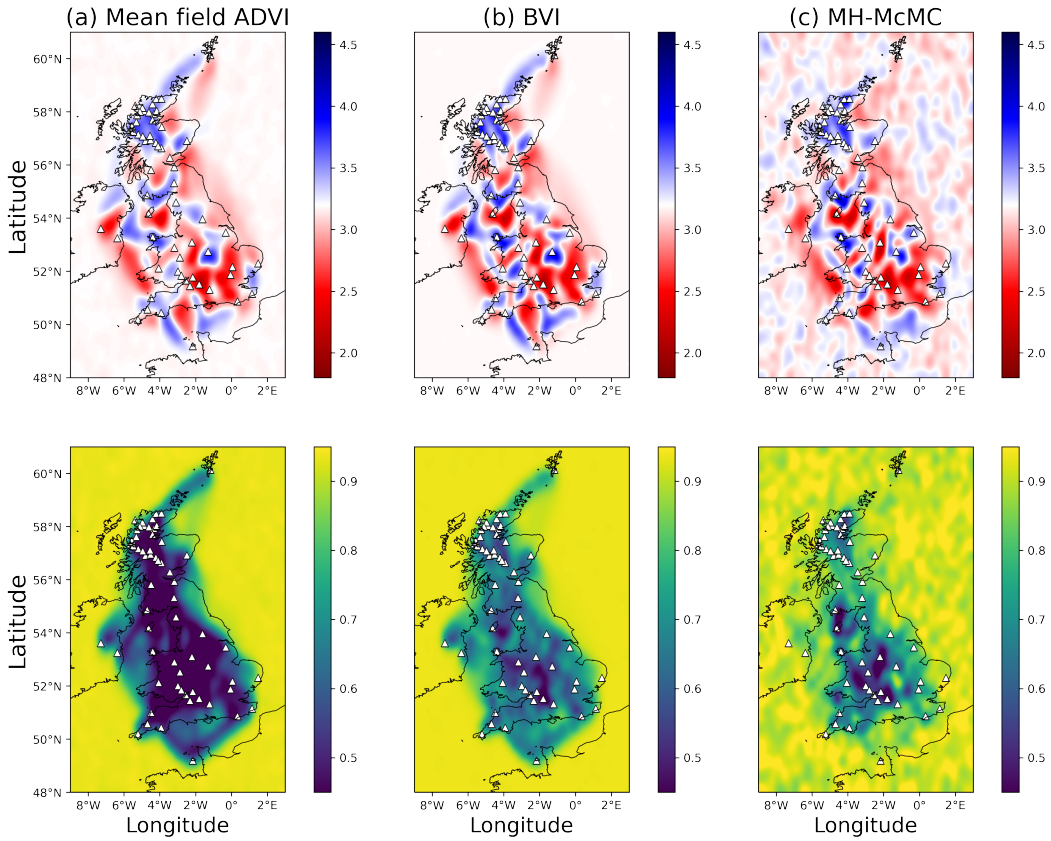

Figure 7. Mean (top row) and standard deviation (bottom row) maps of the Love wave tomography results of the British Isles using (a) mean field ADVI, (b) BVI, and (c) MH-McMC. White triangles show the locations of the receivers used in this example.

cost of rj-McMC is obtained from Galetti et al. (2017). For BVI, four parallel tests with five components are run, each updated for 5000 iterations with two samples per iteration. However, since the first component (mean field ADVI) is very stable, it only needs to be trained once, resulting in a total of 170,000 forward evaluations for BVI and 10,000 for mean field ADVI. MH-McMC consists of 10 chains with 1.5 million samples each, resulting in a total of 15 million samples. It should be noted that the comparison depends on subjectively detecting the convergence of each method and may not reflect the minimum possible computational cost. X. Zhao et al. (2021) showed that the same MH-McMC with 2 million samples only provides a few of the main features in the mean velocity model and hardly provides any useful information in the standard deviation map. This removes the possibility that our subjective assessment of when the Monte Carlo method had converged led to the large number of samples attributed to the method above, and justifies that the number of samples used for MH-McMC is reasonable. It is also true that significantly more efficient Monte Carlo methods may exist for this problem. Nevertheless, the significantly different numbers in Table 1 provide valuable insights into the amount of computation that we and other authors judged necessary to approach convergence for each method. Both mean field ADVI and full rank ADVI have the lowest computational costs, but they also provide biased results. Normalising flows are slightly more efficient than BVI, but they require a sophisticated design of flow structures and often rely on neural networks (Dinh et al., 2015, 2017; Kingma et al., 2016; Papamakarios et al., 2017; Durkan et al., 2019), which can be challenging or even impossible to train for high-dimensional problems such as full waveform inversion. BVI has a simpler structure, and each component is optimised sequentially, making it more attractive for large scale datasets with higher dimensionality in real applications. SVGD is the most expensive variational method, but it still offers a significant reduction in cost compared to these two Monte Carlo methods. The huge numbers of samples used in the latter methods indicate a significant efficiency improvement offered by variational inference for solving large scale and high dimensional inverse problems.

5 Full Waveform Inversion

5.1 Bayesian FWI Implementation

Seismic full waveform inversion (FWI) is a powerful technique to image subsurface structures using full waveform information in seismic data (Tarantola, 1984; Tromp et

Table 1. Number of forward evaluations required for different methods to provide the Love wave tomography results across the British Isles. The results for full rank ADVI, normalising flows and SVGD are from X. Zhao et al. (2021), while the result for rj-McMC is from Galetti et al. (2017).

Method	Forward Evaluations
Mean field ADVI	10,000
Full rank ADVI	10,000
Normalizing Flows	100,000
BVI	170,000
SVGD	600,000
MH-McMC	15,000,000
RJ-McMC	48,000,000

al., 2005). It is a highly non-linear and non-unique problem. Traditional linearised inversion methods can not reliably offer accurate solutions or effectively estimate the uncertainties in the inversion results. As a result, it is important to employ fully non-linear inversion methods for FWI.

FWI problems typically have high dimensionality, and the forward modelling step, in which synthetic seismic waveforms are computed for a given velocity model, is usually expensive. To address these challenges, several efficient Monte Carlo methods have been applied to FWI (Qin et al., 2016; A. Ray et al., 2016; Visser et al., 2019; P. Guo et al., 2020; Gebraad et al., 2020; Z. Zhao & Sen, 2021; Biswas & Sen, 2022; de Lima et al., 2023). Alternatively, in recent years variational methods have also been introduced to address the computational challenges of Bayesian FWI (X. Zhang & Curtis, 2021a; Wang et al., 2023; X. Zhang et al., 2023; Lomas et al., 2023). However, none of these methods provide an accurate and analytic approximation to the posterior probability distribution. In this section, we apply the BVI method to Bayesian FWI, to test its ability to provide analytic results efficiently.

We demonstrate the preceding BVI algorithm using a 2D acoustic FWI example. The true velocity model is a truncated Marmousi model (Martin et al., 2006), as shown in Figure 8a. The density is assumed to be constant. The velocity field is discretized us-

ing 110 × 250 square grid cells with side length 20 m. Twelve sources are placed along the surface at 400 m intervals (shown by red stars in Figure 8a), and 250 receivers are placed along the seabed at a depth of 200 m (white line in Figure 8a). The observed waveform data are obtained by solving the 2D acoustic wave equation using the finite difference method, and the total simulation time is 4 s with a sample interval of 2 ms. The source is a Ricker wavelet with a dominant frequency of 5 Hz. Figure 8c shows this waveform dataset.

For inversion, we use a Uniform prior distribution for the velocity model at each depth, with lower and upper bounds shown in Figure 8b. Velocity in the water layer is fixed at the true value during inversion. Therefore, there are 25,000 free parameters to be inverted, corresponding to the subsurface velocity model. We use the finite difference method to solve the forward function, and the adjoint-state method to calculate the data-model gradient (Fichtner et al., 2006; Plessix, 2006). The likelihood function is chosen to be a diagonal Gaussian function with a constant data error of 0.05 for each data point.

In this test, we compare BVI with 3 different variational methods: mean field ADVI, Stein variational gradient descent (SVGD) and stochastic SVGD (sSVGD). Stochastic SVGD is an extension of SVGD that incorporates a noise term to enhance the efficiency and accuracy of SVGD for large-scale inference problems (Gallego & Insua, 2018). It effectively converts the variational SVGD method to a Markov chain, showing that the divide between these methodological approaches can be bridged. SSVGD has recently been applied to a 3D FWI problem (X. Zhang et al., 2023). For mean field ADVI we use a diagonal Gaussian distribution to model the posterior distribution in the unconstrained space (Kucukelbir et al., 2017). A total of 50,000 hyper-parameters (means and variances in each cell) are updated for 10,000 iterations, and 5 samples per iteration are used. For SVGD, we randomly select 600 samples from the prior distribution and update them for 600 iterations. Once convergence is achieved, these samples are used to approximate statistics of the posterior distribution. For sSVGD, the algorithm starts with 24 random samples drawn from the prior distribution. These samples are then updated for 10,000 iterations, with the first 5,000 iterations discarded as the burn-in period. In this algorithm every sample value evaluated can be retained post burn-in, so all remaining samples are used to approximate the posterior distribution. For BVI, four parallel runs are performed, and each run contains six diagonal Gaussian distributions. This results in a total of 24

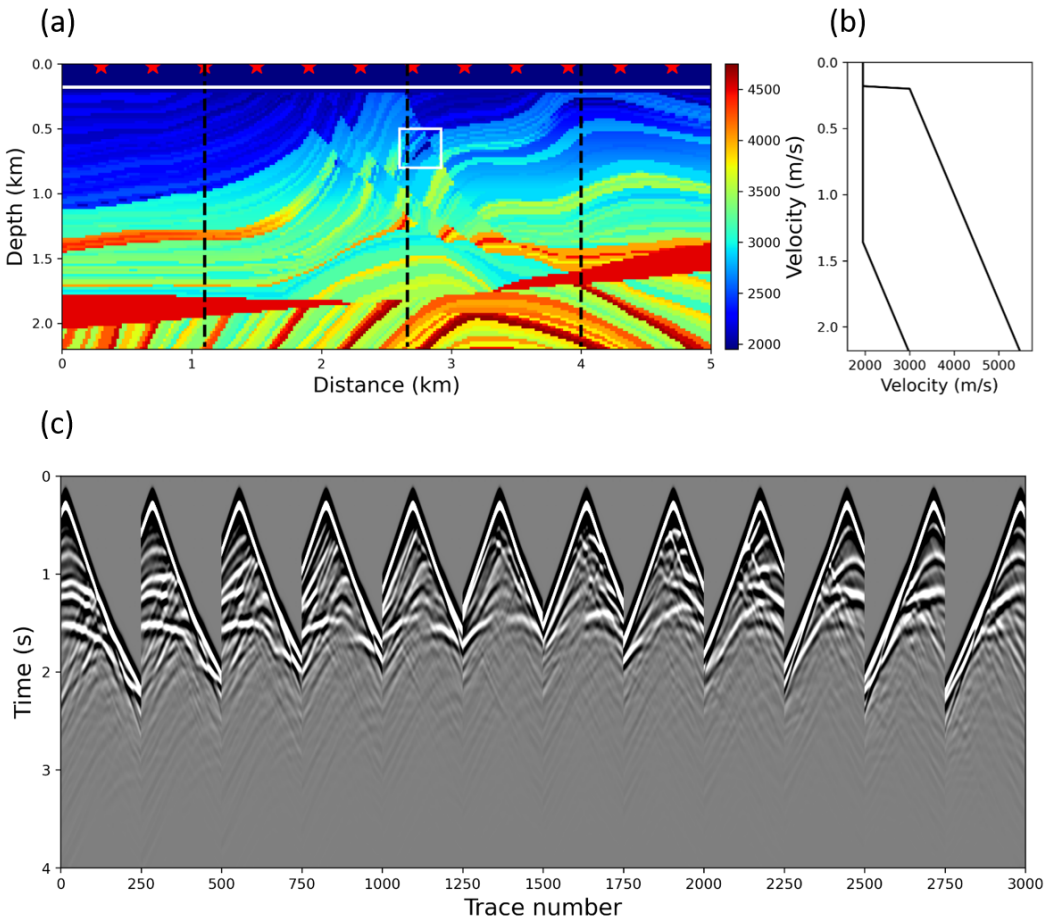

Figure 8. (a) The true Marmousi P wave velocity model with source locations indicated by red stars and receiver line marked by white line. Three dashed black lines display the locations of three well logs discussed in the main text. (b) Upper and lower bounds for the Uniform prior probability distribution for P wave velocity at each depth. (c) Twelve common shot gathers used as the observed data in this test.

Gaussian components used to approximate the posterior distribution. Each component is updated for 5,000 iterations, and two samples per iteration are used.

Figures 9a – 9d display the inversion results obtained using the aforementioned methods. The first two rows show the mean and standard deviation maps of the posterior distribution, while the third row displays the relative error, which is calculated by dividing the absolute error between the true and mean models by the standard deviation model. The mean velocity models from the 4 methods exhibit a similar pattern and generally resemble the true model. For example, within the white box in Figure 8a, we observe a low velocity structure in the true and mean velocity models. However, all four mean velocity maps fail to capture some of the fine-scale structures present in the true model. This can be attributed to the low dominant frequency used in this test (5 Hz). Among the four methods, the mean velocities obtained using mean field ADVI and SVGD appear smoother compared to those obtained using BVI and sSVGD. This observation is consistent with the results obtained in the previous example of 2D synthetic travel time tomography, where the posterior distribution obtained using MH-McMC (Figure 4) is smoother than that obtained using BVI (Figure 3). In the case of BVI, since we use a diagonal Gaussian distribution as the component distribution, each model parameter is updated independently. Every new component is initialised randomly to enhance the current posterior pdf by boosting it on either the lower or higher velocity end, and is optimised to approximate the posterior distribution within a local region in the parameter space, introducing a degree of variation between iterations. Hence, the results obtained from BVI exhibit less smoothness, despite the fact that results obtained from its first component (ADVI) are smooth. As for sSVGD, the introduction of a noise term during each iteration perturbs the samples, leading to increased randomness (X. Zhang et al., 2023). The result may therefore become smoother as a larger number of samples and iterations are used.

The standard deviation obtained using mean field ADVI significantly differs from the other three results and tends to be underestimated. Moreover, a majority of the relative errors are larger than 3, indicating inaccuracy of the results. However, the uncertainty map still exhibits similar geometrical structures compared to the mean and true velocity models. Therefore, we consider ADVI to be an efficient method that provides a fairly accurate mean model but biased uncertainties due to its restrictive theoretical assumptions (X. Zhang & Curtis, 2020a; X. Zhao et al., 2021). Similarly to the mean

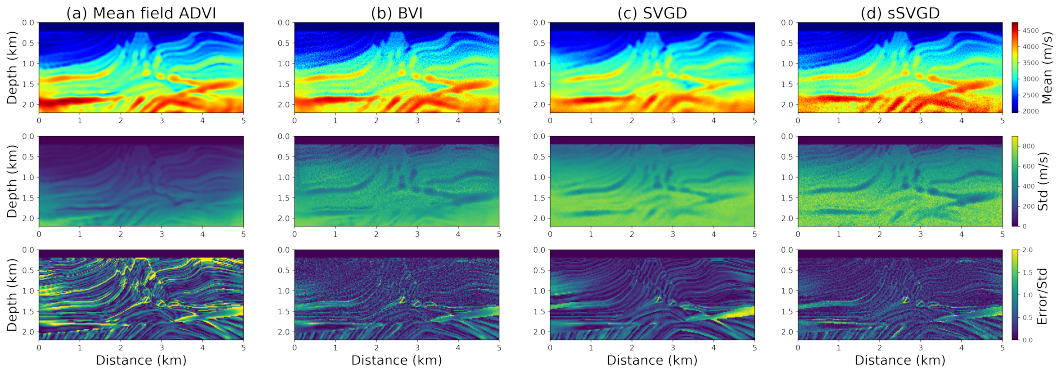

Figure 9. Mean (top row), standard deviation (middle row) and relative error (bottom row) of the posterior distribution for the 2D acoustic FWI test obtained using (a) mean field ADVI, (b) BVI, (c) SVGD and (d) sSVGD. The relative error is the absolute error between the mean and true models divided by the corresponding standard deviation.

velocities, SVGD yields a smoother standard deviation map compared to BVI and sSVGD. In other aspects, the results obtained using these three methods are similar, with errors distributed around two standard deviations. For example, we observe lower uncertainties and higher relative errors at locations with higher velocity anomalies (such as the higher velocity layer at a depth of 1.3 km depth and a distance between 0 – 2 km). Additionally, higher uncertainties are observed at layer boundaries, which is consistent with our observations in the two travel time tomography examples and correspond to uncertainty loops in previous studies (Galetti et al., 2015), especially in the shallower subsurface where data exhibits higher sensitivity. However, the uncertainty values obtained using BVI are slightly smaller compared to the other two methods. We attribute this to two main factors. First, the use of a diagonal Gaussian distribution in BVI tends to underestimate the uncertainty information compared to a Gaussian distribution with a full covariance matrix (Kucukelbir et al., 2017). This underestimation of posterior uncertainties is also evident in Figures 3 and 4. On the other hand, SVGD and sSVGD employ a repulsive force between different samples (Liu & Wang, 2016; Gallego & Insua, 2018): this pushes samples away from each other such that they can explore different parameter subspaces (while still approximating the posterior pdf with sample density). In cases where samples are sparsely distributed within the parameter space, as is the case in this test with 600 samples for SVGD and 24 samples for sSVGD, the repulsive force might push samples towards the corners of parameter hyperspace to maximise the objective

function. This leads to the results in Figures 9c and 9d with higher uncertainties. A similar phenomenon was observed in Love wave tomography using SVGD (X. Zhao et al., 2021). Given the absence of a true solution to this Bayesian FWI problem, it is challenging to determine which method provides a more accurate result. Nevertheless, obtaining similar results using three methods based on two different theoretical frameworks lends credibility to these findings.

For better comparison, Figures 10a – 10d display the marginal pdfs obtained using ADVI, BVI, SVGD and sSVGD, respectively, along three vertical profiles marked by dashed black lines in Figure 8a. Each row shows the marginal distributions along one profile using the four methods. Red lines show the true velocity profiles and black lines show the mean velocity profiles obtained using each method. Similarly to the mean and standard deviation maps in Figure 9, ADVI provides accurate mean velocity profiles but underestimates posterior uncertainties, as evidenced by the narrower marginal pdfs compared to the other three methods. As discussed in the Methodology section, BVI boosts the results obtained from ADVI by using multiple Gaussian components to approximate the posterior distribution. This effect can be observed when comparing Figures 10a and 10b: BVI explores the parameter space that was not adequately represented by ADVI, resulting in wider (and potentially more accurate) marginal distributions. This is particularly noticeable at a depth of 1.2 km within the two white rectangular boxes in the second row in Figures 10a and 10b, where the true velocity value exceeds the prior upper bound (deliberately, to check performance in anomalous cases in which prior distributions are mis-specified). The posterior pdf obtained using BVI is concentrated closer to the upper bound of the prior distribution compared to ADVI. The marginal pdfs obtained using BVI and sSVGD are highly similar and slightly different from those obtained using SVGD. The results from SVGD are sparser (due to limited number of samples) and smoother. In the shallower part of the second row of Figure 10c (indicated by red arrow), the higher probability region of the posterior pdf from SVGD is located close to the prior bound and deviates from the true value. This might be caused by either the limited number of samples or the relatively large step size used in SVGD, which pushes samples towards the boundary of the parameter space by the repulsive force. At a depth of 1.7 km in the third row of Figure 10c (indicated by white arrow), the mean velocity value deviates from the true value since SVGD fails to provide a sufficiently high resolution result to recover this high velocity anomaly compared to BVI and sSVGD. Ad-

ditionally, as indicated by three dashed white boxes in the second row, the posterior distributions from SVGD and sSVGD cover a larger parameter space than that from BVI, especially around the high velocity region. Consequently, we observe higher standard deviation values in Figures 9c and 9d compared to Figure 9b. However, in this region, the mean velocity model obtained using BVI is more similar to the true model, which might indicate higher accuracy compared to both SVGD and sSVGD. This demonstrates that higher uncertainties provided by SVGD and sSVGD might be less convincing due to effects of the the repulsive force, as previously discussed and observed in X. Zhao et al. (2021).

Finally, we compare the computational cost of the four methods in Table 2. In FWI, the forward simulation and data-model gradient calculation are much more expensive compared to those in travel time tomography. Therefore, the number of gradient evaluations provides a fair comparison. For mean field ADVI the model is updated for 10,000 iterations using 5 samples per iteration, resulting in 50,000 evaluations. In the case of BVI, we run 4 parallel tests, each containing 6 Gaussian components. However, we do not need to optimise the first component 4 times, thus a total of 21 Gaussian distributions are used. For each component, we use 5000 iterations and 2 samples per iteration. Therefore, BVI requires 210,000 gradient simulations. It is worth noting that the number of simulations used to optimise each component for BVI is smaller than that for ADVI, even though they have the same hyper-parameters (mean and standard deviation for a diagonal Gaussian distribution). This is because in BVI we do not require full convergence of each component. As long as new components fill some of the gap (residual) between the current mixture distribution and the true posterior distribution, this improves the current result. By adding more components, BVI gradually improves the posterior approximation. For sSVGD and SVGD, they require 240,000 and 360,000 gradient evaluations, respectively. Overall, ADVI is the cheapest method, but it produces biased results. BVI requires more computations to improve the biased results from ADVI, and is slightly more efficient than sSVGD. More importantly, BVI provides an analytic solution to the posterior distribution, while sSVGD only provides posterior samples. SVGD is the most expensive method, and it only provides 600 samples, which is far from sufficient to approximate such a high dimensional (25,000) space in this test.

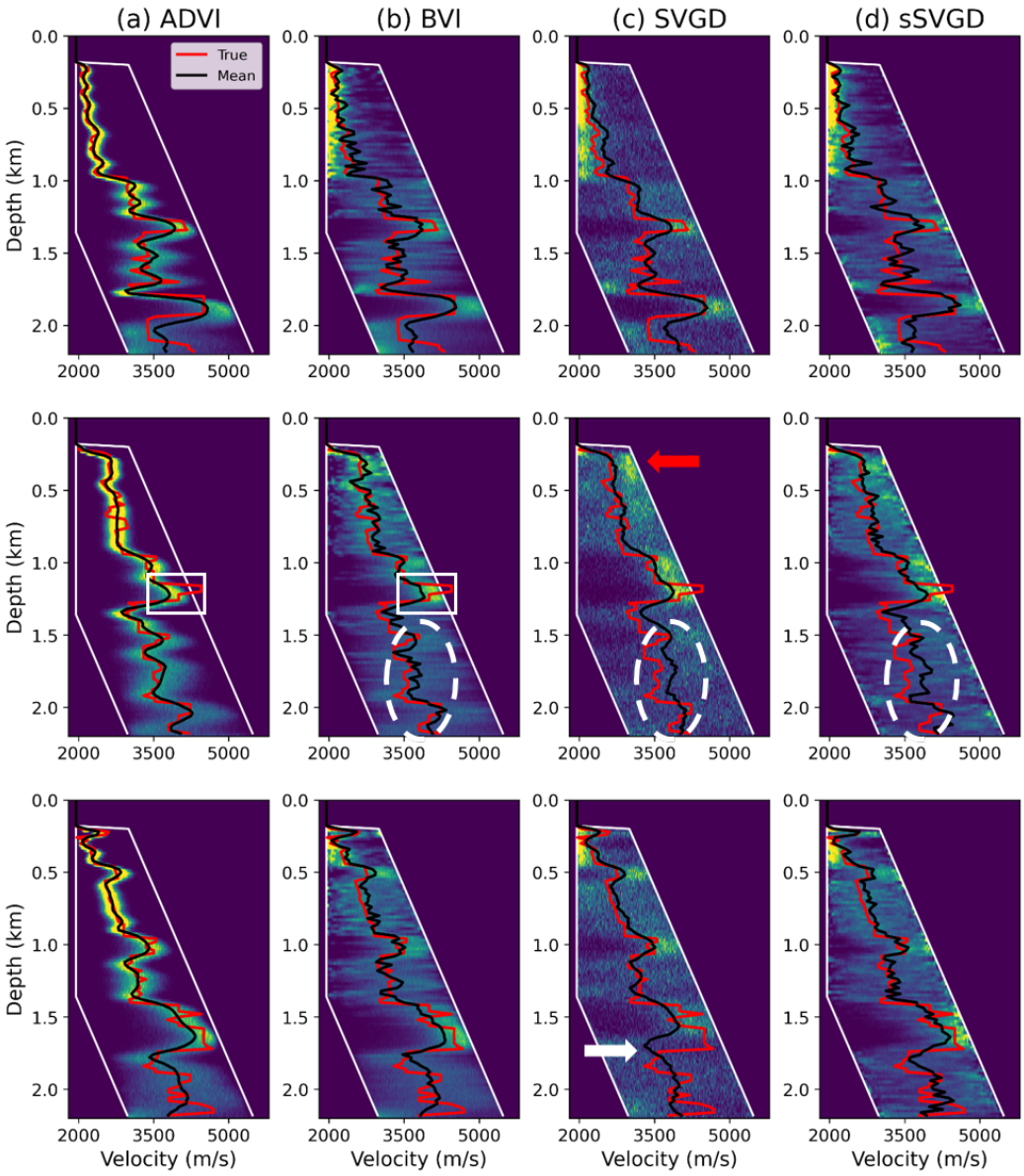

Figure 10. Marginal posterior distributions along vertical profiles at three locations (represented by black dashed lines in Figure 8a) obtained using (a) mean field ADVI, (b) BVI, (c) SVGD and (d) sSVGD, respectively. Each row displays the marginal distribution along one profile. In each figure, two white lines show the prior bounds at each depth, the black line shows the mean velocity model, and the red line shows the true velocity model.

Table 2. Number of forward and gradient evaluations for mean field ADVI, BVI, SVGD and sSVGD applied to the 2D FWI test. The values represent an indication of the computational cost of each method, as the evaluation of data-model gradients is the most computationally expensive part of this test.

Method	Number of Gradient Evaluations
ADVI	50,000
BVI	210,000
SVGD	360,000
sSVGD	240,000

5.2 Interrogating FWI Results

We demonstrate the interrogation theory in section 2.4 by using the FWI results to answer a specific question: *what is the size of the low velocity volume within the white box in Figure 8a?* Such inquiries are common in the geoscience community and are used, for example, to estimate the volume of a sedimentary basin or the size of a reservoir for oil and gas exploration and for CO_2 storage (Fletcher & Ponnambalam, 1996; Burshtein, 2006; Romdhane & Querendez, 2014; X. Zhao et al., 2022; X. Zhang & Curtis, 2022). We therefore refer to the low velocity volume as a reservoir hereafter. Figures 11a and 11b show the posterior mean and standard deviation maps inside the white box, obtained using BVI.

Previously, volume-related questions were answered using interrogation theory with a deterministic target function in X. Zhao et al. (2022) and X. Zhang and Curtis (2022). Here we provide a brief overview of the procedure. We first introduce a mask to restrict the region used to calculate the low velocity anomalies, as illustrated by the dashed black box in Figures 11a and 11b. Other low velocity bodies outside of this mask are assumed to be unrelated to the anomaly of interest and are ignored during the interrogation process. Considering a reservoir should be a continuous geological body in space, we define the target function to be the area of the largest continuous low velocity body inside the mask. To evaluate this function, we need to distinguish between low velocity and high velocity cells, which can be accomplished by introducing a threshold value: cells with

velocity values below the threshold are classified as low velocity, others are classified as not low velocity.

We use the same data-driven method as X. Zhao et al. (2022) to calculate the threshold value with minimal bias. First, we select a set of points from the inversion results that are most likely to belong to the low velocity reservoir since they have low mean velocity values and low standard deviation values (indicated by white stars in Figures 11a and 11b), and another set of points likely to be outside of the reservoir (represented by black crosses in Figures 11a and 11b). Then we calculate the average marginal cumulative density function (cdf) of the low velocity white stars accumulating as velocity increases, and of the high velocity black crosses accumulating as velocity decreases. The intersection point of these cdfs is the threshold value that discriminates low from high velocities with minimal bias according to the prior information provided by the locations of white stars and black crosses. The corresponding threshold value is illustrated by the blue line in Figure 11c. Given this value we can classify each cell as either a low or high velocity cell, find the largest continuous low velocity body inside the mask, and calculate its size which is the target function value. Figure 12d shows the posterior distribution of the target function (reservoir size) obtained using this threshold value. According to equation 15, the optimal answer is the mean of the target function values from all posterior samples, and is denoted by dashed black line in Figure 12d. For comparison, the true size is denoted by red line in Figure 12d.

The above method calculates the threshold value and the target function deterministically. As stated in section 2.4, this does not consider the uncertainty introduced by human bias, which may result in different sets of low and high velocity cells being selected by different experts, thus different threshold values and different target functions, potentially biasing reservoir size estimates. We therefore also apply interrogation with a probabilistic target function, which is defined by a probabilistic threshold value in this example. We implement this by randomly selecting a subset of the grid cells from each of the low and high velocity cells in Figures 11a and 11b. This random selection simulates possible variation in the selection by different experts. We also consider other cells situated on the boundaries of the low velocity body, as indicated by red dots in Figures 11a and 11b which in fact contain valuable information about reservoir shape and velocity values (Galetti et al., 2015), and incorporate the information provided by these cells to calculate the probabilistic threshold value. To do that, we randomly select a sub-

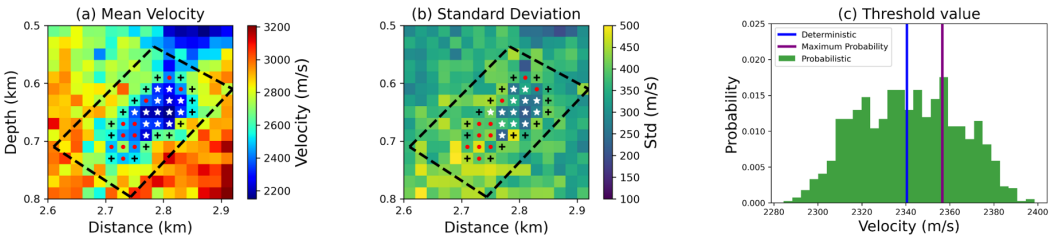

Figure 11. (a) Mean and (b) standard deviation maps of the posterior distribution obtained using BVI within the white box in Figure 8a. Black dashed box shows the mask inside which we calculate the area of the largest continuous low velocity body, which serves as the target function. White stars and black crosses denote cells that are most likely to be inside and outside the reservoir, respectively. Red dots denote cells predominantly located on the reservoir boundaries, where uncertainty remains regarding their classification as low or high velocities. (c) Threshold values to discriminate low and high velocities. Green histogram shows the probabilistic threshold value established in the main text. Blue line shows the deterministic threshold value obtained using the white stars and black crosses only, and purple line shows the maximum probability threshold value from the green histogram.

set of cells marked by those red dots, and assign cells that are directly connected to the cells marked by the white stars as low velocity cells (inside the reservoir) and the remaining cells as high velocity cells (outside the reservoir). This can be interpreted as a misclassification of low and high velocity cells at the boundaries of the reservoir, again simulating possible human bias and subjective choice. We use these randomly selected cells to calculate the threshold value. The above procedure is repeated 1000 times, resulting in a probability distribution over the threshold value represented by the green histogram in Figure 11c.

We perform interrogation using the green histogram in Figure 11c as the stochastic threshold value, which then defines the probabilistic target function. For each posterior model sample (velocity model obtained from BVI), we draw 100 random threshold values from the green histogram and calculate the size of the largest continuous low velocity body corresponding to each threshold value. The resulting distribution of 100 reservoir sizes values incorporates the uncertainty in the target function, so we repeat this process for each posterior model sample. Figure 12a shows the distribution of the target function values, and the optimal answer calculated using equation 16 is denoted

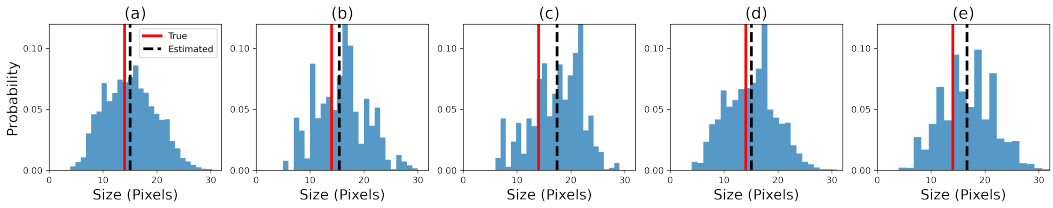

Figure 12. Posterior distributions of the target function by interrogation with probabilistic target function obtained using (a) full BVI inversion results, (b) 24 representative samples from BVI components, and (c) 40 random samples from full BVI inversion results. Panels (d) and (e) show the posterior target functions obtained using deterministic target functions whose threshold values are represented by the blue and purple lines in Figure 11c. In each figure, the red line denotes the true answer to this question, and black dashed line denotes the optimal answer obtained using interrogation theory.

by the dashed black line. This represents the interrogation results (with the probabilistic target function) obtained using the full posterior distribution from BVI. We also construct a solution using only the representative samples obtained from BVI components to perform interrogation, and the posterior target function is displayed in Figure 12b. The optimal answer is calculated using equation 18 (black dashed line in Figure 12b). Since we only use the mean vectors of the Gaussian components to obtain those representative samples, without considering the corresponding covariance matrices, the uncertainty of the posterior target function might be underestimated. Nevertheless, this still provides an accurate optimal answer while significantly reducing the number of target function evaluations. Additionally, we randomly choose 40 posterior samples from the full BVI inversion result and conduct probabilistic interrogation on these: the resulting posterior histogram is displayed in Figure 12c. In comparison to Figure 12b, the optimal answer obtained from this set of 40 samples is notably inaccurate, whereas of the order of ten representative samples from BVI provide almost equal interrogation results to the full posterior solution. This proves the value of these representative samples.

To simulate the bias that may be introduced by using a deterministic target function for example one defined by a single expert, we choose the maximum probability value from the green histogram in Figure 11c as the threshold value (denoted by purple line in Figure 11c). This value falls within the high probability region and can be treated as a reasonable threshold value obtained from one expert. We perform interrogation using

this single threshold value, and the result is displayed in Figure 12e. The optimal answer (black dashed line) shows a larger error and deviates more from the true answer than any estimate other than that from 40 random samples of the model posterior distribution in Figure 12c.

Overall, the comparison of the five histograms in Figure 11 reveals that interrogation using deterministic target functions may yield biased results due to the subjective nature of human interpretation. This bias can be mitigated by using probabilistic target functions. Note that the optimal answer using the deterministic target function in Figure 12d also provides an accurate result, and the posterior target function is similar to that in Figure 12a. However, we usually do not know the true answer to our question in real problems, and therefore have no means to prioritise the answer from one interpretation over any other. Probabilistic interrogation considers the subjectivity from different experts, and provides a more convincing answer. The optimal answer obtained using representative samples from BVI components is accurate, which proves that these samples capture a key portion of the uncertainty information in the inversion results. In contrast, randomly selected posterior samples fail to adequately represent this uncertainty. Therefore, subsequent uncertainty analysis tasks, especially those that are computationally intractable to perform for thousands of posterior samples (e.g., reservoir simulation), could be more efficiently carried out using the representative samples obtained from BVI.

6 Discussion

In BVI, the variational distribution is built by a mixture of simpler component distributions that are added sequentially using a greedy algorithm. This differs from traditional approaches that jointly optimise weight coefficients and component distributions in a mixture model. Such approaches are generally non-convex and challenging to implement (Bishop, 2006; F. Guo et al., 2016). Additionally, deciding in advance the number of components to construct the mixture distribution is difficult. In our work, we use an arbitrary number of Gaussian distributions as mixture components, adding components until little further benefit is obtained, and the resulting approximation to the posterior distribution can be represented analytically by equation 13.

Our synthetic travel time tomography examples illustrate how BVI progressively enhances the accuracy of the posterior approximation, and provide a reliable criterion

for assessing the algorithm's convergence. The application to a field dataset to perform Love wave tomography of the British Isles provides convincing results that are consistent with known geology and previous studies (Nicolson et al., 2012, 2014; Galetti et al., 2017; X. Zhao et al., 2021). Having established the method's credibility, we apply BVI and three other methods to an FWI problem, namely mean field ADVI, SVGD and stochastic SVGD (sSVGD). ADVI strongly underestimates the uncertainties, whereas the other three methods independently offer similar, thus hopefully approximately correct results.

As stated by the No Free Lunch theorem (Wolpert & Macready, 1997), no method is better than any other method when averaged across all problems, so there is no possibility to find a 'best' method in general. However, for a particular class of problems it is possible to find better or worse suited algorithms from different points of view. In all of our examples, ADVI yields biased uncertainty results, but provides an accurate mean velocity map and is the most computationally efficient method. The first component of BVI can be regarded as equivalent to ADVI, and so establishes an estimate of the mean. BVI then introduces additional components to better approximate uncertainty in the true posterior distribution, trading off with a higher computational cost.

Table 2 shows that BVI and sSVGD have similar computational costs. However, sSVGD provides higher uncertainties compared to BVI, as shown in Figures 9a and 9d. It is difficult to determine which method is more accurate since they employ fundamentally different approaches to explore the parameter space and to avoid degenerating into a single point mass at the maximum a posteriori (MAP) model. Specifically, sSVGD and SVGD employ a repulsive force in their objective functions to push samples away from each other (Liu & Wang, 2016; Gallego & Insua, 2018). BVI (as well as some other variational methods such as normalising flows and ADVI) maximises ELBO explicitly in its objective function

$$\mathrm{ELBO}[q(\mathbf{m})] = \mathbb{E}_{q(\mathbf{m})}[\log p(\mathbf{d}_{obs}|\mathbf{m})] - \mathrm{KL}[q(\mathbf{m})||p(\mathbf{m})] \tag{19}$$

Therefore, maximising the ELBO involves maximising the expectation of the log-likelihood and minimising the KL divergence between the variational distribution $q(\mathbf{m})$ and the prior distribution $p(\mathbf{m})$. The latter encourages $q(\mathbf{m})$ to explore the full prior space, rather than being restricted to the vicinity of the MAP. Consequently, it increases the complexity of the results (Blei et al., 2017; Wang et al., 2023). Additionally, using a sampling-based method such as sSVGD makes it is easier to calculate higher-order statistical informa-

tion such as correlations between different parameters (X. Zhang et al., 2023), whereas BVI, using diagonal Gaussian components as in this paper, struggles to capture such information. One possible improvement is to use Gaussian components with a full covariance matrix, but this can be computationally cumbersome for high-dimensional problems such as FWI, as it requires $D(D+1)/2$ hyper-parameters for a D-dimensional covariance matrix. Considering that only a few pairs of variables may exhibit significant posterior correlations (e.g., neighbouring cells), a feasible approach is to approximate the full covariance matrix using a low-rank plus diagonal approach (Miller et al., 2017). Regardless, BVI is parametric which allows as many samples as needed to be generated post optimisation, whereas this is difficult for sampling based methods (SVGD and sSVGD). More importantly, BVI provides an analytic representation of the posterior pdf, and all inversion results presented in this paper are obtained using analytic calculations without drawing any samples (except for the probabilistic interrogation example). On the other hand, SVGD is the most expensive method, and it provides only a limited number (hundreds) of samples which may not be sufficient to represent key properties of the posterior distribution.

Normalising flows are another variational method which can effectively model posterior correlations between different parameters (Dinh et al., 2015, 2017; Kingma et al., 2016; Papamakarios et al., 2017). The trend in the field of normalising flows is to develop deeper and more complex flows to achieve greater flexibility. It has been demonstrated that normalising flows outperform ADVI (X. Zhao et al., 2021), making them a promising choice for improving BVI. By using probability distributions modelled by normalizing flows as the component distributions in BVI, we might capture posterior correlations and create a wider, rather than strictly deeper, model that enhances the capabilities of existing normalising flows while reducing the complexity for designing flows structures, albeit at the expense of greedy optimisation (Giaquinto & Banerjee, 2020).

Gaussian processes (GP) is another class of methods that use Gaussian distributions to approximate the probability distribution of model parameters. GP is a form of stochastic process, and can be regarded as a way to define a Gaussian distribution over functions (for example, to define Gaussian distributions for velocity values at every subsurface location). It is commonly used as a non-parametric regression method that predicts model parameters and the corresponding uncertainties within a continuous region. A. Ray and Myer (2019), A. Ray (2021) and Blatter et al. (2021) used GP together with

a trans-dimensional McMC sampling scheme to perform inversion. In those works GP was used as a regression method to build a finely discretized or even spatially continuous (infinite-dimensional) model vector $\mathbf{m}$, which can be viewed as a random sample from an infinite-dimensional multivariate Gaussian distribution, given parameter values at some known locations. The obtained model was used to calculate the synthetic data to further update the GP. Valentine and Sambridge (2020a, 2020b) used GP to solve linear (or weakly non-linear) inverse problems. The inversion result can be expressed as a GP which represents the posterior distribution in function space. Due to the nature of GP, these works assume a Gaussian prior distribution for the model parameter at each location and a linear forward function (as in Valentine & Sambridge, 2020a). Such assumptions are not necessary for BVI as described in this paper.

Making use of the analyticity of BVI results can be challenging, but we have developed an implicit approach to address this issue. Our approach involves selecting one representative sample from each BVI component: leveraging the fact that a parametric and symmetric Gaussian distribution is used as the component distribution. We simply adopt the mean vector as a representative sample, allowing us to obtain tens of samples directly that partially represent the posterior distribution for uncertainty analysis. Considering that we also obtain a diagonal covariance matrix for each component, it is easily possible to incorporate the information from the covariance matrix into these representative samples (for example, by selecting a number of component samples that is proportional to the weight of that component and combining all such samples). This would capture more detail from the posterior distribution and improve the effectiveness of uncertainty analysis.

In our interrogation example, we show that the optimal answer obtained using the representative samples is accurate and comparable to that obtained using full inversion results. This is particular attractive when implementing probabilistic interrogation as proposed in this paper, or when the evaluation of the target function is computationally expensive. For example, if our goal is to estimate CO_2 saturation of a reservoir using FWI results, the target function might involve reservoir simulation or (non-linear) rock physics inversion to convert seismic velocity values into CO_2 saturation. Calculating the target function for thousands of posterior samples could then be prohibitively expensive. In such cases, we can simply use the representative samples obtained from BVI components for analysis. Moreover, storing a large set of posterior samples on disk

and loading them into memory can be extremely demanding, especially for 3D FWI problems (X. Zhang et al., 2023; Lomas et al., 2023), to which the use of representative samples from BVI components provides a practical solution. Finally, it is important to note that obtaining these representative samples would be challenging without the analytic expression of the posterior distribution provided by BVI, which provides these samples directly.

7 Conclusion

We have presented boosting variational inference (BVI) as a powerful variational method for solving fully non-linear Bayesian geophysical inverse problems. BVI constructs a flexible approximating family using a mixture of simple component distributions, with the Gaussian distribution chosen specifically for its ease of optimising and its parametric nature. The components are optimised sequentially using a greedy algorithm, progressively improving the accuracy of the posterior approximation as more components are added. We have demonstrated the effectiveness of BVI through applications to seismic travel time tomography and full waveform inversion (FWI). By comparing the results obtained using BVI with other variational and Monte Carlo sampling methods, we conclude that BVI is capable of providing efficient and accurate inversion results. One key advantage of BVI is its ability to provide an analytic expression for the posterior probability distribution function, which provides a low number of representative samples that partially represent the posterior uncertainty. We have introduced a probabilistic framework that uses these samples to solve an interrogation problem - answering a specific scientific question by interrogating the probabilistic inverse problem solution. The result demonstrates that the representative samples yield similar accuracy compared to that obtained using the full posterior distribution. This approach reduces the computation for subsequent uncertainty analysis, making it promising for large scale problems.

8 Open Research

Both synthetic and field data, and software used in this study are available at Edinburgh DataShare (`https://datashare.ed.ac.uk/handle/10283/8528`, X. Zhao & Galetti, 2023). Software used for the variational methods as well as the 2D McMC can be found at PyMC3 website (`https://docs.pymc.io/en/v3/`, Salvatier et al., 2016).

Software used to perform Automatic Differentiation can be found at PyTorch website (`https://pytorch.org/`, Paszke et al., 2019).

Acknowledgments

The authors thank Erica Galetti (Galetti et al., 2017) for providing traveltime data used in the field data test, and thank PyMC3 (Salvatier et al., 2016) and PyTorch (Paszke et al., 2019) developers for providing software used in this paper. The authors thank Edinburgh Imaging Project (EIP) sponsors (BP and TotalEnergies) for supporting this research. AC thanks Muhammad Atif Nawaz for contributory discussions. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... others (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Anderssen, R., & Seneta, E. (1971). A simple statistical estimation procedure for Monte Carlo inversion in geophysics. *Pure and Applied Geophysics*, *91*(1), 5–13.

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, *18*, 343–373.

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

Atchadé, Y. F., & Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, *11*(5), 815–828.

Bishop, C. M. (1994). Mixture density networks.

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

Biswas, R., & Sen, M. K. (2022). Transdimensional 2d full-waveform inversion and uncertainty estimation. *arXiv preprint arXiv:2201.09334*.

Blatter, D., Ray, A., & Key, K. (2021). Two-dimensional Bayesian inversion of magnetotelluric data using trans-dimensional Gaussian processes. *Geophysical Journal International*, *226*(1), 548–563.

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.

Bloem, H., Curtis, A., & Tetzlaff, D. (2023). Introducing conceptual geological information into Bayesian tomographic imaging. *Basin Research*.

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

Bodin, T., Sambridge, M., Rawlinson, N., & Arroucau, P. (2012). Transdimensional tomography with unknown data noise. *Geophysical Journal International*, *189*(3), 1536–1556.

Bond, C., Lunn, R., Shipton, Z., & Lunn, A. (2012). What makes an expert effective at interpreting seismic images? *Geology*, *40*(1), 75–78.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Burshtein, L. (2006). Statistical estimation of parameters of size distribution of oil fields in poorly explored sedimentary basins. *Russian Geology and Geophysics*, *47*(9), 1013–1023.

Campbell, T., & Li, X. (2019). Universal boosting variational inference. *Advances in Neural Information Processing Systems*, *32*.

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

Cao, R., Earp, S., de Ridder, S. A., Curtis, A., & Galetti, E. (2020). Near-real-time near-surface 3D seismic velocity and uncertainty models by wavefield gradiometry and neural network inversion of ambient seismic noise. *Geophysics*, *85*(1), KS13–KS27.

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.

de Lima, P. D. S., Corso, G., Ferreira, M. S., & de Araújo, J. M. (2023). Acoustic full waveform inversion with Hamiltonian Monte Carlo method. *Physica A: Statistical Mechanics and its Applications*, *617*, 128618.

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.

de Wit, R. W., Valentine, A. P., & Trampert, J. (2013). Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks. *Geophysical Journal International*, *195*(1), 408–422.

Dinh, L., Krueger, D., & Bengio, Y. (2015). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. (2019). Neural spline flows. In *Advances in neural information processing systems* (pp. 7509–7520).

Earp, S., & Curtis, A. (2020). Probabilistic neural network-based 2d travel-time tomography. *Neural Computing and Applications*, *32*(22), 17077–17095.

Earp, S., Curtis, A., Zhang, X., & Hansteen, F. (2020). Probabilistic neural network tomography across Grane field (North Sea) from surface wave dispersion data. *Geophysical Journal International*, *223*(3), 1741–1757.

Ely, G., Malcolm, A., & Poliannikov, O. V. (2018). Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method. *Geophysics*, *83*(2), R63–R75.

Fichtner, A., Bunge, H.-P., & Igel, H. (2006). The adjoint method in seismology: I. theory. *Physics of the Earth and Planetary Interiors*, *157*(1-2), 86–104.

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

Fletcher, S., & Ponnambalam, K. (1996). Estimation of reservoir yield and storage distribution using moments analysis. *Journal of hydrology*, *182*(1-4), 259–275.

Frank, M., & Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, *3*(1-2), 95-110.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

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.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., & Stephenson, J. (2009). 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.

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

Gebraad, L., Boehm, C., & Fichtner, A. (2020). Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo. *Journal of Geophysical Research: Solid Earth*, *125*(3), e2019JB018428.

Ghosal, S., Ghosh, J. K., & Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 500–531.

Giaquinto, R., & Banerjee, A. (2020). Gradient boosted normalizing flows. *Advances in Neural Information Processing Systems*, *33*, 22104–22117.

Guo, F., Wang, X., Fan, K., Broderick, T., & Dunson, D. B. (2016). Boosting variational inference. *Advances in Neural Information Processing Systems*.

Guo, P., Visser, G., & Saygin, E. (2020). Bayesian trans-dimensional full waveform inversion: synthetic and field data application. *Geophysical Journal International*, *222*(1), 610–627.

Hansen, T. M., & Finlay, C. C. (2022). Use of machine learning to estimate statistics of the posterior distribution in probabilistic inverse problems—an application to airborne em data. *Journal of Geophysical Research: Solid Earth*, *127*(11), e2022JB024703.

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

Izzatullah, M., Baptista, R., Mackey, L., Marzouk, Y., & Peter, D. (2020). Bayesian seismic inversion: Measuring Langevin MCMC sample quality with kernels. In *SEG international exposition and annual meeting*.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning* (pp. 427–435).

Käufl, P., P. Valentine, A., W. de Wit, R., & Trampert, J. (2016). Solving probabilistic inverse problems rapidly with prior samples. *Geophysical Journal International*, *205*(3), 1710–1728.

Käufl, P., Valentine, A. P., O'Toole, T. B., & Trampert, J. (2014). A framework for fast probabilistic centroid-moment-tensor determination—inversion of regional static displacement measurements. *Geophysical Journal International*, *196*(3), 1676–1693.

Khoshkholgh, S., Orozova-Bekkevold, I., & Mosegaard, K. (2022). Evolution of the stress and strain field in the Tyra field during the Post-Chalk deposition and seismic inversion of fault zone using informed-proposal Monte Carlo. *Applied Computing and Geosciences*, *14*, 100085.

Khoshkholgh, S., Zunino, A., & Mosegaard, K. (2021). Informed proposal monte carlo. *Geophysical Journal International*, *226*(2), 1239–1248.

Khoshkholgh, S., Zunino, A., & Mosegaard, K. (2022). Full-waveform inversion by informed-proposal Monte Carlo. *Geophysical Journal International*, *230*(3), 1824–1833.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems* (pp. 4743–4751).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kobyzev, I., Prince, S., & Brubaker, M. A. (2019). Normalizing flows: An introduction and review of current methods. *arXiv preprint arXiv:1908.09257*.

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.

Levy, S., Laloy, E., & Linde, N. (2022). Variational Bayesian inference with complex geostatistical priors using inverse autoregressive flows. *Computers & Geosciences*, 105263.

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information process-*

ing systems (pp. 2378–2386).

Locatello, F., Dresdner, G., Khanna, R., Valera, I., & Rätsch, G. (2018). Boosting black box variational inference. *Advances in Neural Information Processing Systems*, *31*.

Locatello, F., Khanna, R., Ghosh, J., & Ratsch, G. (2018). Boosting variational inference: an optimization perspective. In *International conference on artificial intelligence and statistics* (pp. 464–472).

Lomas, A., Luo, S., Irakarama, M., Johnston, R., Vyas, M., & Shen, X. (2023). 3D probabilistic full waveform inversion: Application to Gulf of Mexico field data. In *84th eage annual conference & exhibition* (Vol. 2023, pp. 1–5).

Lubo-Robles, D., Ha, T., Lakshmivarahan, S., Marfurt, K. J., & Pranter, M. J. (2021). Exhaustive probabilistic neural network for attribute selection and supervised seismic facies classification. *Interpretation*, *9*(2), T421–T441.

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

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

Meier, U., Curtis, A., & Trampert, J. (2007a). Fully nonlinear inversion of fundamental mode surface waves for a global crustal model. *Geophysical Research Letters*, *34*(16).

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.

Meir, R., & Rätsch, G. (2003). An introduction to boosting and leveraging. In *Advanced lectures on machine learning: Machine learning summer school 2002 canberra, australia, february 11–22, 2002 revised lectures* (pp. 118–183). Springer.

Miller, A. C., Foti, N. J., & Adams, R. P. (2017). Variational boosting: Iteratively refining posterior approximations. In *International conference on machine learning* (pp. 2420–2429).

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

12431–12447.

Nawaz, 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.

Nawaz, A., & Curtis, A. (2019). Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties. *Journal of Geophysical Research: Solid Earth*, *124*(6), 5867–5887.

Nawaz, A., Curtis, A., Shahraeeni, M. S., & Gerea, C. (2020). Variational Bayesian inversion of seismic attributes jointly for geological facies and petrophysical rock properties. *Geophysics*, *85*(4), 1–78.

Neiswanger, W., Wang, C., & Xing, E. (2013). Asymptotically exact, embarrassingly parallel mcmc. *arXiv preprint arXiv:1311.4780*.

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.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). Uncertain judgements: eliciting experts' probabilities.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*.

Papamakarios, G., Pavlakou, T., & Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in neural information processing systems* (pp. 2338–2347).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Plessix, R.-E. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, *167*(2), 495–503.

Polson, D., & Curtis, A. (2010). Dynamics of uncertainty in geological interpretation. *Journal of the Geological Society*, *167*(1), 5–10.

Press, F. (1968). Earth models obtained by Monte Carlo inversion. *Journal of Geophysical Research*, *73*(16), 5223–5234.

Qin, H., Xie, X., Vrugt, J. A., Zeng, K., & Hong, G. (2016). Underground structure defect detection and reconstruction using crosshole GPR and Bayesian waveform inversion. *Automation in Construction*, *68*, 156–169.

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

Rawlinson, N., & Sambridge, M. (2005). The fast marching method: an effective tool for tomographic imaging and tracking multiple phases in complex layered media. *Exploration Geophysics*, *36*(4), 341–350.

Ray, A. (2021). Bayesian inversion using nested trans-dimensional Gaussian processes. *Geophysical Journal International*, *226*(1), 302–326.

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.

Ray, A. K., & Biswal, S. (2010). An efficient method of effective porosity prediction using an unconventional attribute through multi-attribute regression and probabilistic neural network: A case study in a deep-water gas field, East Coast of India. In *Seg technical program expanded abstracts 2010* (pp. 1413–1417). Society of Exploration Geophysicists.

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

Romdhane, A., & Querendez, E. (2014). CO2 characterization at the Sleipner field with full waveform inversion: Application to synthetic and real data. *Energy procedia*, *63*, 4358–4365.

Rothman, D. H. (1986). Automatic estimation of large residual statics corrections. *Geophysics*, *51*(2), 332–346.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, *2*, e55.

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

Scales, J. A. (1996). Uncertainties in seismic inverse calculations. *Lecture Notes in Earth Sciences, Berlin Springer Verlag*, *63*, 79–97.

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

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.

Siahkoohi, A., Rizzuti, G., & Herrmann, F. J. (2022). Deep bayesian inference for seismic imaging with tasks. *Geophysics*, *87*(5), S281–S302.

Siahkoohi, A., Rizzuti, G., Louboutin, M., Witte, P. A., & Herrmann, F. J. (2021). Preconditioned training of normalizing flows for variational inference in inverse problems. *arXiv preprint arXiv:2101.03709*.

Siahkoohi, A., Rizzuti, G., Orozco, R., & Herrmann, F. J. (2023). Reliable amortized variational inference with physics-based latent distribution correction. *Geophysics*, *88*(3), 1–137.

Smith, J. D., Ross, Z. E., Azizzadenesheli, K., & Muir, J. B. (2022). Hyposvi: Hypocentre inversion with stein variational inference and physics informed neural networks. *Geophysical Journal International*, *228*(1), 698–710.

Stoffa, P. L., & Sen, M. K. (1991). Nonlinear multiparameter optimization using genetic algorithms; inversion of plane-wave seismograms. *Geophysics*, *56*(11), 1794–1810.

Strutz, D., & Curtis, A. (2023). Variational bayesian experimental design for geophysical applications. *arXiv preprint arXiv:2307.01039*.

Sun, L., Wang, L., Xu, G., & Wu, Q. (2023). A new method of variational bayesian slip distribution inversion. *Journal of Geodesy*, *97*(1), 10.

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

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

Tromp, J., Tape, C., & Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, *160*(1), 195–216.

Urozayev, D., Ait-El-Fquih, B., Hoteit, I., & Peter, D. (2022). A reduced-order variational bayesian approach for efficient subsurface imaging. *Geophysical Journal International*, *229*(2), 838–852.

Valentine, A. P., & Sambridge, M. (2020a). Gaussian process models—i. a framework for probabilistic continuous inverse theory. *Geophysical Journal International*, *220*(3), 1632–1647.

Valentine, A. P., & Sambridge, M. (2020b). Gaussian process models—ii. lessons for discrete inversion. *Geophysical Journal International*, *220*(3), 1648–1656.

Visser, G., Guo, P., & Saygin, E. (2019). Bayesian transdimensional seismic full-waveform inversion with a dipping layer parameterization. *Geophysics*, *84*(6), R845–R858.

Wang, W., McMechan, G. A., & Ma, J. (2023). Re-weighted variational full waveform inversions. *Geophysics*, *88*(4), 1–61.

Wapenaar, K. (2004). Retrieving the elastodynamic green's function of an arbitrary inhomogeneous medium by cross correlation. *Physical review letters*, *93*(25), 254301.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82.

Zhang, C., Bütepage, J., Kjellström, H., & Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, *41*(8), 2008–2026.

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

Zhang, X., & Curtis, A. (2020b). Variational full-waveform inversion. *Geophysical Journal International*, *222*(1), 406-411.

Zhang, X., & Curtis, A. (2021a). Bayesian full-waveform inversion with realistic priors. *Geophysics*, *86*(5), 1–20.

Zhang, X., & Curtis, A. (2021b). Bayesian geophysical inversion using invertible neural networks. *Journal of Geophysical Research: Solid Earth*, e2021JB022320.

Zhang, X., & Curtis, A. (2022). Interrogating probabilistic inversion results for subsurface structural information. *Geophysical Journal International*, *229*(2), 750–757.

Zhang, X., Lomas, A., Zhou, M., Zheng, Y., & Curtis, A. (2023). 3-D Bayesian variational full waveform inversion. *Geophysical Journal International*, *234*(1), 546–561.

Zhao, X., Curtis, A., & Zhang, X. (2021). Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, *228*(1), 213–239.

Zhao, X., Curtis, A., & Zhang, X. (2022). Interrogating subsurface structures using probabilistic tomography: an example assessing the volume of Irish Sea basins. *Journal of Geophysical Research: Solid Earth*, *127*(4), e2022JB024098.

Zhao, X., & Galetti, E. (2023). Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference, [dataset]. *Edinburgh DataShare*. doi: https://datashare.ed.ac.uk/handle/10283/8528

Zhao, Z., & Sen, M. K. (2021). A gradient-based markov chain Monte Carlo method for full-waveform inversion and uncertainty analysis. *Geophysics*, *86*(1), R15–R30.

Zidan, A., Li, Y., & Cheng, A. (2022). Regularized seismic amplitude inversion via variational inference. *Geophysical Prospecting*, *70*(9), 1507–1527.

Appendix A Derivation and calculation for ∇ELBO

In this Appendix, we derive the gradient of ELBO$[q^t(\mathbf{m})]$ with respect to the weight coefficient w_t in equation 11 and the numerical method used for its calculation.

Substitute equation 5 into 3, and this gradient term can be written as

$$\begin{aligned}
\nabla_{w_t}\text{ELBO}[q^t(\mathbf{m})] &= \nabla_{w_t}\mathbb{E}_{q^t(\mathbf{m})}[\log p(\mathbf{m},\mathbf{d}_{obs}) - \log q^t(\mathbf{m})] \\
&= \nabla_{q^t}\mathbb{E}_{q^t(\mathbf{m})}[\log p(\mathbf{m},\mathbf{d}_{obs}) - \log q^t(\mathbf{m})]\nabla_{w_t}\left((1-w_t)q^{t-1}(\mathbf{m}) + w_t g_t(\mathbf{m})\right) \\
&= \int (\log p(\mathbf{m},\mathbf{d}_{obs}) - \log q^t(\mathbf{m}))(g_t(\mathbf{m}) - q^{t-1}(\mathbf{m}))d\mathbf{m} \\
&= \mathbb{E}_{g_t(\mathbf{m})}[\log \frac{p(\mathbf{m},\mathbf{d}_{obs})}{q^t(\mathbf{m})}] - \mathbb{E}_{q^{t-1}(\mathbf{m})}[\log \frac{p(\mathbf{m},\mathbf{d}_{obs})}{q^t(\mathbf{m})}]
\end{aligned} \tag{A1}$$

which can be estimated using Monte Carlo integration by drawing samples from $g_t(\mathbf{m})$ and $q^{t-1}(\mathbf{m})$. Then we iteratively update w_t using stochastic gradient descent (equation 11).