

An Introduction to Variational inference in Geophysical Inverse Problems

Xin Zhang, Muhammad Atif Nawaz, Xuebin Zhao, Andrew Curtis

School of Geosciences, University of Edinburgh, Edinburgh, EH9 3FE, United Kingdom

This article is a preprint version of the manuscript. The final version is published in Advances in Geophysics: <https://doi.org/10.1016/bs.agph.2021.06.003>

Chapter 1

An Introduction to Variational Inference in Geophysical Inverse Problems

Xin Zhang^{*,1}, Muhammad Atif Nawaz*, Xuebin Zhao*, and Andrew Curtis*

**School of Geosciences, University of Edinburgh, Edinburgh, EH9 3FE, United Kingdom*

ABSTRACT

In a variety of scientific applications we wish to characterize a physical system using measurements or observations. This often requires us to solve an inverse problem, which usually has non-unique solutions so uncertainty must be quantified in order to define the family of all possible solutions. Bayesian inference provides a powerful theoretical framework which defines the set of solutions to inverse problems, and variational inference is a method to solve Bayesian inference problems using optimization while still producing fully probabilistic solutions. This chapter provides an introduction to variational inference, and reviews its applications to a range of geophysical problems, including petrophysical inversion, travel time tomography and full-waveform inversion. We demonstrate that variational inference is an efficient and scalable method which can be deployed in many practical scenarios.

KEYWORDS

Inverse problem, Bayesian inference, Variational inference, Uncertainty quantification, Petrophysical inversion, Seismic tomography, Full-waveform inversion

1. x.zhang2@ed.ac.uk

SUMMARY OF NOTATION

ADVI	automatic differential variational inference
c	a subset of variables (clique)
C	a set of cliques, i.e. $c \in C$
\det	determinant
\mathbf{d}_{obs}	observed data vector
ELBO	evidence lower bound
EM	Expectation-Maximization
$F(q; \Theta)$	evidence lower bound of probability distribution q defined as a function of parameters Θ
E_q	expectation with respect to probability distribution q
\mathbf{F}_θ	a normalizing flow parameterized by θ
FWI	full-waveform inversion
GM	Gaussian mixture
μ	mean of a Gaussian distribution
Σ	covariance matrix of a Gaussian distribution
HMC	Hamiltonian Monte Carlo
\mathbf{J}	Jacobian matrix
$k(\cdot, \cdot)$	a scalar kernel function
$\mathbf{K}(\cdot, \cdot)$	a matrix-valued kernel function
KL	Kullback-Leibler divergence
$L(\Theta; \mathbf{d}_{\text{obs}})$	logarithmic evidence defined as a function of parameters Θ
\mathbf{L}	a lower-triangular matrix
\mathbf{m}	vector of model parameters
MAP	maximum a posterior
MC	Monte Carlo
McMC	Markov chain Monte Carlo
MH-McMC	Metropolis-Hastings Markov chain Monte Carlo
MRF	Markov random field
NN	a neural network function
$N(\cdot \mathbf{0}, \mathbf{I})$	a standard normal distribution with zero mean and an identity covariance matrix
pdf	probability density function
$p(\mathbf{d}_{\text{obs}} \mathbf{m})$	likelihood function
$p(\mathbf{d}_{\text{obs}})$	normalization factor in Bayes' theorem (evidence)
$p(\mathbf{m}, \mathbf{d}_{\text{obs}})$	joint probability density function of \mathbf{m} and \mathbf{d}_{obs}
$p(\mathbf{m})$	prior probability density function

$p(\mathbf{m} \mathbf{d}_{\text{obs}})$	posterior probability density function
$p(\mathbf{m} \mathbf{d}_{\text{obs}}, \Theta)$	posterior probability density function parameterized with Θ
Φ	a smooth vector function
ψ, ϕ	scalar functions
q	a pobability density function used to approximate the posterior pdf
Q	$q \in Q$
RBF	radial basis function
κ	geological facies
γ	geological rock properites
\mathcal{A}_p	Stein operator defined on probability distribution p
SVGD	Stein variational gradient descent
T	an invertible transform
VFWI	variational full-waveform inversion

1 INTRODUCTION

In a variety of scientific applications scientists often wish to characterize a physical system using measurements or observations which do not represent the system directly. A simplified model of the system is defined which includes a physical relation that predicts measurements or observations for any particular values of the model parameters. One then seeks parameter values that match the measurements or observations. This process is called inversion, and the physical relation that predicts observations that would be made if any particular set of parameter values were true is called the *forward* function. In this article we focus on Geophysics. Geophysicists often need to characterize properties of the Earth's interior using measurements such as seismic, gravitational or electromagnetic data. Subsurface properties are usually parameterized such that one can construct a forward function that predicts corresponding data, and the inverse problem is therefore a parameter estimation problem (Aki & Lee, 1976; Tarantola, 2005).

Due to nonlinearity of the physical relation, insufficient data coverage and noise in the data, the inverse problem almost always has non-unique solutions, as infinitely many sets of parameter values fit the observed data to within their measurement uncertainties. This family of values defines uncertainty in the inverse problem solution. In order to reduce this uncertainty, any available *prior* information about parameters (information known independently of the geophysical data) is usually imposed on the solution, and remaining uncertainties in the estimated parameters must be described (Tarantola, 2005).

Inverse problems are often solved using optimisation methods by seeking parameter values that minimize misfits between observed data and the data predicted by the forward function. Since most inverse problems are under-determined, some form of regularization is often imposed on the model. This process is well-established for linear problems in which the system reduces to solving a set of linear simultaneous equations (Aster, Borchers, & Thurber, 2018). This approach can also be applied to nonlinear problems by linearising (approximating) the nonlinear physics around a reference model and solving that linearised problem for the parameter values. The process of linearising and solving the problem is iterated until the misfit or update to the values is sufficiently small (Aki & Lee, 1976; Aster et al., 2018; Constable, Parker, & Constable, 1987; Dziewonski & Woodhouse, 1987; Iyer & Hirahara, 1993; Tarantola, 2005; Tarantola & Valette, 1982). However, since the regularization is often ad-hoc in the sense that it does not correspond to genuine prior information, the results can be biased and valuable information can be concealed in the process (Zhdanov, 2002). In addition this method cannot provide accurate uncertainty estimate for nonlinear problems, nor even for linear problems with complex data uncertainty distributions.

Bayesian inference provides a different way to solve inverse problems and quantify uncertainties. In Bayesian inference the prior information is represented

by a probability density function (pdf) and is updated with new information from the data to produce a probability density function that describes all information post inversion, called a *posterior* pdf. According to Bayes' theorem, the posterior pdf can be expressed as:

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

where \mathbf{m} is a vector of model parameter values, \mathbf{d}_{obs} is the observed data, and $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ is the posterior pdf; $p(\mathbf{m})$ represents the prior pdf which describes information independent of data, $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ is called the *likelihood* which represents the probability of observing data \mathbf{d}_{obs} given parameters \mathbf{m} which in turn depends on the forward function, and $p(\mathbf{d}_{\text{obs}})$ is a normalization factor called the *evidence*. The term *inference* indicates that the prior information is combined with uncertainties in the measured data and forward function to infer the posterior pdf.

A common way to solve Bayesian inference problems is to use Markov chain Monte Carlo (McMC). In McMC one constructs a set (chain) of successive samples of \mathbf{m} drawn from the posterior pdf by taking a structured random walk through a parameter space (Brooks, Gelman, Jones, & Meng, 2011); those samples can thereafter be used to calculate useful statistics of that pdf, e.g. the mean and standard deviation. The Metropolis-Hastings algorithm is one such method (Hastings, 1970; Metropolis & Ulam, 1949) and has been applied to a range of geophysical applications (Andersen, Brooks, & Hansen, 2001; Gallagher, Charvin, Nielsen, Sambridge, & Stephenson, 2009; Malinverno, 2002; Malinverno, Leaney, et al., 2000; Mosegaard & Sambridge, 2002; Mosegaard & Tarantola, 1995; Oh & Kwon, 2001; Ramirez et al., 2005; Sambridge & Mosegaard, 2002). The method has been generalized to trans-dimensional inversion called *reversible-jump* McMC, in which the number of parameters (the dimensionality of parameter space) can vary in the inversion and consequently the parameterization itself can be adapted to the data and the prior information (Green, 1995; Green & Hastie, 2009). Reversible-jump McMC has been applied to various geophysical applications, including vertical seismic profile inversion (Malinverno et al., 2000), electrical resistivity inversion (Galetti & Curtis, 2018; Malinverno, 2002), electromagnetic inversion (Minsley, 2011; Ray, Alumbaugh, Hoversten, & Key, 2013), surface wave dispersion inversion (Bodin et al., 2012; Shen, Ritzwoller, Schulte-Pelkum, & Lin, 2012; Young, Rawlinson, & Bodin, 2013), travel time tomography (Bodin & Sambridge, 2009; Galetti, Curtis, Baptie, Jenkins, & Nicolson, 2017; Galetti, Curtis, Meles, & Baptie, 2015; Hawkins & Sambridge, 2015; Piana Agostinetti, Giacomuzzi, & Malinverno, 2015; X. Zhang, Curtis, Galetti, & de Ridder, 2018; X. Zhang, Hansteen, Curtis, & de Ridder, 2020; X. Zhang, Roy, Curtis, Nowacki, & Baptie, 2020) and full-waveform inversion (Guo, Visser, & Saygin, 2020; Ray, Kaplan, Washbourne, & Albertin, 2017; Ray, Sekar, Hoversten, & Albertin, 2016; Sen & Biswas, 2017). However, due to its random-walk behaviour the method becomes inefficient in

high dimensional space (e.g., $> 1,000$). Other more advanced McMC methods have been introduced to geophysics to solve high dimensional problems, such as Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, & Roweth, 1987; Fichtner, Zunino, & Gebraad, 2018; Gebraad, Boehm, & Fichtner, 2020; Kotsi, Malcolm, & Ely, 2020; Sen & Biswas, 2017), Langevin Monte Carlo (Roberts, Tweedie, et al., 1996; Siahkoohi, Rizzuti, & Herrmann, 2020), stochastic Newton McMC (J. Martin, Wilcox, Burstedde, & Ghantas, 2012; Z. Zhao & Sen, 2019), and parallel tempering (Dosso, Holland, & Sambridge, 2012; Hukushima & Nemoto, 1996; Sambridge, 2013). Nevertheless, these methods remain intractable for large datasets and high dimensionality because of their extremely high computational cost.

Variational inference solves Bayesian inference problems in a different way: one seeks an optimal approximation to the posterior pdf within a predefined family of (simplified) probability distributions. This is achieved by minimizing a measure of the difference between the posterior pdf and the approximating pdf, for example the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951). Since the method uses optimization rather than random sampling, it can be computationally more efficient than McMC and provide better scaling to high dimensionality. The methods can also be applied to large datasets by dividing the dataset into minibatches and using stochastic optimization techniques (Kubrusly & Gravier, 1973; Robbins & Monro, 1951). By contrast, stochastic optimisation cannot be applied to McMC because it breaks the detailed balance which is required by most McMC methods.

In variational inference the choice of the variational family determines the accuracy of the approximation and the complexity of the optimization problem. A good choice should be rich enough to approximate complex distributions and simple enough such that the optimization problem can be efficiently solved. A common choice is to use a *mean-field* approximation in which the parameters are assumed to be mutually independent (Bishop, 2006; Blei, Kucukelbir, & McAuliffe, 2017; Parisi, 1988; C. Zhang, Bütepage, Kjellström, & Mandt, 2018). The optimisation problem can then be solved efficiently using a coordinate ascent algorithm (Bishop, 2006; Blei et al., 2017) which has been applied in geophysics to invert for spatial distributions of geological facies using seismic data (Nawaz & Curtis, 2018, 2019; Nawaz, Curtis, Shahraeeni, & Gerea, 2020).

Despite its wide application in practice, the mean-field method ignores correlations between parameters and requires tedious model-specific mathematical derivations and implementation. This restricts the method to a narrow range of inverse problems for which the derivations can be performed. To make variational inference applicable to general inverse problems, a variety of "black box" methods have been proposed based on different variational families, for example, the mean-field approximation (Ranganath, Gerrish, & Blei, 2014; Ranganath, Tran, & Blei, 2016), Gaussian distributions (Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2017) and probability transforms (Q. Liu & Wang, 2016; Marzouk, Moseley, Parno, & Spantini, 2016; Rezende & Mohamed, 2015; Tran, Ran-

ganath, & Blei, 2015). These methods are quite general and can be applied to a wide range of applications, for example in geophysics they have been applied to travel time tomography (X. Zhang & Curtis, 2020a; X. Zhao, Curtis, & Zhang, 2020), full-waveform inversion (X. Zhang & Curtis, 2020b) and seismic image denoising (Siahkoohi, Rizzuti, Witte, & Herrmann, 2020).

This chapter aims to give a brief introduction to variational inference and its applications in geophysics. In the following sections we first introduce the concepts of variational inference, and then describe four different variational methods: mean-field variational inference, automatic differential variational inference (ADVI), normalizing flows and Stein variational gradient descent (SVGD). The first of these shows how the structure of some inference problems can be exploited to obtain highly efficient variational methods of solution, whereas the latter three methods make few assumptions about the problem structure. In section 3 we demonstrate how these methods have been applied to a range of different applications, including petrophysical inversion, travel time tomography and full-waveform inversion. We conclude the chapter by discussing some limitations and possible improvements to the variational methodology.

2 VARIATIONAL INFERENCE

Variational inference uses optimization to solve Bayesian inference problems. First a family of known probability distributions $Q = \{q(\mathbf{m})\}$ is defined. For example, Q could be the family of all Gaussians, or sums of Gaussians. The variational method then seeks the best approximation to the posterior pdf $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ within that family by minimizing the KL-divergence between $q(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$:

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

$q^*(\mathbf{m})$ is then used as an approximation to the posterior pdf. The KL-divergence is a measure of difference between two pdfs, and can be expressed as:

$$\text{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})] = E_q[\log q(\mathbf{m})] - E_q[\log p(\mathbf{m}|\mathbf{d}_{\text{obs}})] \quad (3)$$

where the expectations are taken with respect to the known pdf $q(\mathbf{m})$. The KL-divergence is nonnegative and only equals zero when $q = p$ (Kullback & Leibler, 1951). Expanding the posterior pdf $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ using Bayes' theorem,

$$\text{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})] = E_q[\log q(\mathbf{m})] - E_q[\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})] + \log p(\mathbf{d}_{\text{obs}}) \quad (4)$$

The evidence term $\log p(\mathbf{d}_{\text{obs}})$ is computationally intractable in many problems: it is the marginal pdf over \mathbf{d}_{obs} of the joint distribution $p(\mathbf{m}, \mathbf{d}_{\text{obs}})$, so the evidence calculation requires an integral of the forward function over the full prior pdf on \mathbf{m} to be evaluated. This is often impossible. Therefore we move the evidence term to the left-hand side and reverse the sign of the equation:

$$\log p(\mathbf{d}_{\text{obs}}) - \text{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})] = E_q[\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})] - E_q[\log q(\mathbf{m})] \quad (5)$$

Given that the KL-divergence is nonnegative, the left-hand side defines a lower bound for the evidence, called the evidence lower bound (ELBO):

$$\begin{aligned} \text{ELBO}[q] &= \log p(\mathbf{d}_{\text{obs}}) - \text{KL}[q(\mathbf{m}) || p(\mathbf{m} | \mathbf{d}_{\text{obs}})] \\ &= E_q [\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})] - E_q [\log q(\mathbf{m})] \end{aligned} \quad (6)$$

Since the second line of equation 6 does not involve the intractable evidence term, it can be computed in practice by analytical or numerical methods. In addition because the evidence term $\log p(\mathbf{d}_{\text{obs}})$ is a constant for a given problem, minimizing the KL-divergence is equivalent to maximizing the ELBO. Variational inference in equation 2 can therefore be expressed as:

$$q^*(\mathbf{m}) = \arg \max_{q \in Q} \text{ELBO}[q(\mathbf{m})] \quad (7)$$

In variational inference the choice of family Q is important because it determines the accuracy of the approximation and the complexity of the optimization. A good choice should be flexible enough to approximate the posterior pdf accurately, but simple enough for efficient optimization. Depending on different choices of the family, different variational methods have been proposed. In the following sections we describe several such methods.

2.1 Mean field approximation

For problems that have particular types of structures, extremely efficient variational methods can be derived to find solutions. In this section we look at problems that have known, structured probabilistic relationships amongst the variables.

Exact Bayesian inference requires evaluation of the evidence – the denominator in Bayes theorem (equation 1). As the model dimensionality increases, the cost of this calculation escalates exponentially. Thus, exact inference becomes infeasible for many-parameter models and for all practical purposes one needs to resort to approximate inference.

Stochastic sampling-based inference, such as the commonly used Markov-chain Monte Carlo (McMC) method, is only asymptotically exact, i.e., sampling distributions in high-dimensional (henceforth, simply large) models converge to the true distribution only theoretically as sampling continues to infinite time. Instead of approximating the true distribution by a finite number of samples, one may consider other approximation schemes such as limiting the dimensionality of probabilistic dependence among variables (Nawaz & Curtis, 2016). One such scheme is the mean field approximation which provides an efficient method to model probabilistic dependence in high dimensional problems by exploiting structure in the probabilistic dependence among various variables, and replacing at least some probabilistic dependence in the model by an effective random field that is defined by a set of scalar potential functions ψ_i , each of which is

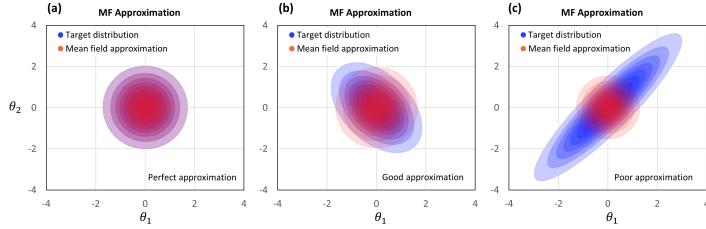


FIGURE 1 Examples of mean field approximation to bivariate Gaussian distributions with (a) zero correlation, (b) weak correlation and (c) strong correlation between the two parameters.

defined over just a few variables. Thus, the intractable joint posterior distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ over all of the variables under the mean field approximation assumes a factorized form

$$p(\mathbf{m}|\mathbf{d}_{\text{obs}}) \cong q(\mathbf{m}) = \frac{1}{Z} \prod_i \psi_i(m_i) \quad (8)$$

where Z is the normalization constant and $q(\mathbf{m})$ is the factorized approximation of $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. Such a factorized approximation allows computationally efficient inference in large models. In its simplest form, each random variable is regarded as independent of the others, and the only source of mutual interaction (or correlation) among several variables is a random field - a structured set of probabilistic relationships among various parameters of interest at multiple locations. Figure 1 shows examples of the mean field approximation to different bivariate Gaussian distributions. While the method can provide accurate approximations to distributions that have zero or weak correlation between parameters (Figure 1a and b), it fails to produce accurate estimates of distributions with strong correlations (e.g., Figure 1c). Thus, a naive implementation of the mean field approximation cannot be used to infer posterior distributions with strong correlations.

A more common approach is to capture correlations among pairs of variables which results in the so called *Ising* or *Potts* model depending on whether each variable can take two or more possible states, respectively. Modelling of pairwise dependence among variables imposes smoothness constraints that can be described by second order statistics, e.g. using covariance matrix. These models, however, ignore higher-order dependence structure beyond pairs of variables, e.g. multiple-point statistics. Nawaz and Curtis (2019) introduced higher-order mean field inference method that makes use of structure of dependence among variables to capture most of the significant higher-order correlations among them while still allowing computationally efficient inference. Factorization of joint

distribution in this case takes the form

$$q(\mathbf{m}) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{m}_c) \quad (9)$$

where ψ_c represents potential functions (called *clique* potentials) defined over some subsets c of variables called a *clique* and C represents the set of cliques in the graph. Full probabilistic dependence among variables \mathbf{m}_c within a clique c is honoured, however, it is assumed that a variable in c may interact with other variables outside c only through an effective field defined by the functional form of clique potentials ψ_c . According to the Hammersley-Clifford theorem (Besag, 1974; Hammersley & Clifford, 1971), the joint distribution $q(\mathbf{m})$ over all \mathbf{m} may be expressed as a Gibbs distribution which takes the form

$$q(\mathbf{m}) = \frac{1}{Z} \exp \left\{ -\frac{1}{r} \sum_{c \in C} E_c(\mathbf{m}_c) \right\} \quad (10)$$

where $E_c(\mathbf{m}_c)$ represents the energy function that associates low energy states correspond to high probability configurations of \mathbf{m}_c , and r is a constant. The clique potentials ψ_c , therefore, take the function form

$$\psi_c(\mathbf{m}_c) = \exp \left\{ -\frac{E_c(\mathbf{m}_c)}{r} \right\} \quad (11)$$

A factorized distribution that takes the form of the Gibbs distribution is commonly known as a Markov random field. The quality of the mean field approximation can be determined using some distance measure between the true (unknown) posterior distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ and its factorized approximation $q(\mathbf{m})$. This may be achieved by using the KL divergence (equation 4) which we then minimize, showing that mean field inference is a special form of variational inference where the approximating distribution takes a factorized form. Mean field inference commonly employs iterative optimization methods to perform probabilistic inference in an optimization framework without stochastic sampling while still providing full probabilistic results, as described below.

In order to estimate the intractable constant $p(\mathbf{d}_{\text{obs}})$ in Bayes' theorem under the above simplified model, we denote its logarithm as a function of parameters Θ as $L(\Theta; \mathbf{d}_{\text{obs}})$ and refer to it as the log evidence. Any choice of the auxiliary distribution q defines a lower bound $F(q; \Theta)$ (the ELBO in equation 6) on the log-evidence $L(\Theta; \mathbf{d}_{\text{obs}})$ (Beal, 2003; Nawaz & Curtis, 2018; Neal & Hinton, 1998), such that

$$L(\Theta; \mathbf{d}_{\text{obs}}) = F(q; \Theta) + \text{KL}(q(\mathbf{m}) || p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta)) \quad (12)$$

where the lower bound $F(q; \Theta)$ is also called *variational free energy* or simply *free energy*. It has its origin in statistical physics where it corresponds to the negative of Gibbs free energy (Feynman, 1972), and $\text{KL}(q(\mathbf{m}) || p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta)) \geq 0$ is

the KL divergence (also called relative-entropy) between $q(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta)$ as defined above. For a factorisable distribution q , $F(q; \Theta)$ assumes a closed-form expression in terms of marginal distributions of q (Nawaz & Curtis, 2018).

Although $L(\Theta; \mathbf{d}_{\text{obs}})$ is intractable, its lower bound $F(q; \Theta)$ may be estimated for a suitably chosen family of approximate pdf's q . This suggests that an iterative scheme may be devised to estimate $L(\Theta; \mathbf{d}_{\text{obs}})$ by successively updating q and Θ in each iteration. For example, a variational form of the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) may be used to approximate $L(\Theta; \mathbf{d}_{\text{obs}})$ in an iterative fashion such that its lower bound $F(q; \Theta)$ is increased which effectively decreases $\text{KL}(q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta))$ for a given set of parameters Θ in each iteration.

Algorithm 1: Mean field approximation

Input: An initial distribution q^0 from an approximating family $q(\mathbf{m})$;
 An initial set of parameters Θ^0 of the (unkonwn) posterior
 distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta)$.

Output: Updated Θ of the true posterior pdf $p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta)$.

```

 $l \leftarrow 0$ 
while  $l < N$  do
    Calculate  $F(q^l; \Theta^l)$  (e.g. see Nawaz & Curtis, 2018)
    E-step:
         $q^{l+1} \leftarrow \arg \max_q F(q; \Theta^l)$ 
    M-step:
         $\Theta^{l+1} \leftarrow \arg \max_{\Theta} F(q^{l+1}; \Theta)$ 
         $\delta \leftarrow \text{abs}(\Theta^{l+1} - \Theta^l)$ 
        if  $\delta$  is sufficiently small then
            exit
        else
             $l \leftarrow l + 1$ 
        end
end
return  $\Theta^N$ .
```

The E-step of the EM algorithm at any iteration l updates the variational distribution $q(\mathbf{m})$ by maximizing the free-energy $F(q; \Theta)$ with respect to q while keeping the parameters Θ^l fixed such that

$$q^{l+1} = \arg \max_q F(q; \Theta^l) \quad (13)$$

where the bracketed superscripts refer to the iteration number. Nawaz and Curtis (2018) showed that the E-step of the EM algorithm can be solved using a message passing algorithm, called belief propagation (BP) (Pearl, 1982), or its variant, the loopy belief propagation (LBP) (Mariethoz & Caers, 2014; Yedidia, Freeman, & Weiss, 2003). The M-step of the EM algorithm at any iteration l computes an updated set of parameters Θ^{l+1} by maximizing the free-energy $F(q; \Theta)$ with

respect to Θ while keeping the variational distribution q fixed at its value q^{l+1} estimated during the E-step, such that

$$\Theta^{l+1} = \arg \max_{\Theta} F(q^{l+1}; \Theta) \quad (14)$$

In summary, at the end of $(l + 1)^{th}$ iteration the E-Step of the EM algorithm yields the free energy $F(q^{l+1}, \Theta^l)$ equal to $L(\Theta^l; \mathbf{d}_{\text{obs}})$ which is the upper bound of $F(q, \Theta^l)$. Therefore, the E-step improves the estimate of the posterior distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}}, \Theta)$ while the M-step improves the estimate of parameters Θ , such that the combined E-M steps are guaranteed not to decrease the estimate of log evidence $L(\Theta; \mathbf{d}_{\text{obs}})$ during any iteration of the EM algorithm. On convergence, the EM algorithm yields the best mean field approximation $q(\mathbf{m})$ of the true intractable posterior distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. We summarize the method in Algorithm 1.

2.2 Automatic differential variational inference

The mean-field approximation allows highly efficient variational methods to be derived, at the expense of losing full correlation information between parameters. Such methods require model-specific derivations and implementations, which restricts them to those types of problems for which approximation applies. In this section we describe a method called automatic differential variation inference (ADVI) which can be applied to a general class of inverse problems, and which is made efficient by introducing a different approximation (Kucukelbir et al., 2017).

The key idea behind ADVI is to use a Gaussian variational family. Gaussians are defined over the entire set of real numbers whereas in reality model parameters often have hard bound constrains (for example seismic velocity is greater than zero). To apply ADVI to constrained variables we first transform those variables into an unconstrained space using an invertible transform T : $\Theta = T(\mathbf{m})$. In this space the joint pdf $p(\mathbf{m}, \mathbf{d}_{\text{obs}})$ becomes:

$$p(\Theta, \mathbf{d}_{\text{obs}}) = p(\mathbf{m}, \mathbf{d}_{\text{obs}}) |\det \mathbf{J}_{T^{-1}}(\Theta)| \quad (15)$$

where $\mathbf{J}_{T^{-1}}(\Theta)$ is the Jacobian matrix of the inverse of transform T , and $|\cdot|$ denotes absolute value. Define a Gaussian family in this unconstrained space,

$$q(\Theta; \zeta) = N(\Theta | \mu, \Sigma) = N(\Theta | \mu, \mathbf{L}\mathbf{L}^T) \quad (16)$$

where ζ represents variational parameters, that is the mean vector μ and the covariance matrix Σ . To ensure the covariance matrix is positive semidefinite, we use a Cholesky factorization $\Sigma = \mathbf{L}\mathbf{L}^T$ where \mathbf{L} is a lower-triangular matrix, to reparameterize the covariance matrix. If Σ is a diagonal matrix, q reduces to a mean-field approximation as described in section 2.1.

Within this Gaussian family the variational problem in equation 7 becomes:

$$\begin{aligned}\zeta^* &= \arg \max_{\zeta} \text{ELBO}[q(\theta; \zeta)] \\ &= \arg \max_{\zeta} E_q[\log p(T^{-1}(\theta), \mathbf{d}_{\text{obs}}) + \log|det \mathbf{J}_{T^{-1}}(\theta)|] - E_q[\log q(\theta; \zeta)]\end{aligned}\quad (17)$$

This optimisation problem can be solved by gradient-based optimisation methods, for example gradient ascent. In order to calculate the gradients of the ELBO with respect to variational parameters ζ , we first transform the Gaussian distribution $q(\theta; \zeta)$ to a standard Normal distribution $N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})$ by using the transform $\boldsymbol{\eta} = R(\theta) = \mathbf{L}^{-1}(\theta - \boldsymbol{\mu})$. The problem thereafter becomes:

$$\begin{aligned}\zeta^* &= \arg \max_{\zeta} \text{ELBO}[q(\theta; \zeta)] \\ &= \arg \max_{\zeta} E_{N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})}[\log p(T^{-1}R^{-1}(\boldsymbol{\eta}), \mathbf{d}_{\text{obs}}) + \log|det \mathbf{J}_{T^{-1}}(R^{-1}(\boldsymbol{\eta}))|] \\ &\quad - E_q[\log q(\theta; \zeta)]\end{aligned}\quad (18)$$

where the first expectation in equation 18 is calculated with respect to a standard Normal distribution. There is no Jacobian term appearing in equation 18 according to the rules of integration by substitution. For example for any function $h(\theta)$,

$$\begin{aligned}E_q[h(\theta)] &= \int h(\theta)q(\theta; \zeta)d\theta \\ &= \int h(R^{-1}(\boldsymbol{\eta}))q(R^{-1}(\boldsymbol{\eta}); \zeta)|det \mathbf{J}_{R^{-1}}(\boldsymbol{\eta})|d\boldsymbol{\eta} \\ &= \int h(R^{-1}(\boldsymbol{\eta}))N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})d\boldsymbol{\eta} \\ &= E_{N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})}[h(R^{-1}(\boldsymbol{\eta}))]\end{aligned}\quad (19)$$

The second expectation in equation 18 does not need to be transformed because the expectation has an analytic form. In fact this expectation is called the *entropy* of q , written $H[q(\theta; \zeta)]$:

$$\begin{aligned}H[q(\theta; \zeta)] &= -E_q[\log q(\theta; \zeta)] \\ &= \frac{k}{2} + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log|det \mathbf{L}\mathbf{L}^T|\end{aligned}\quad (20)$$

where k is the dimension of vector θ .

The gradients of the ELBO with respect to variational parameters can be calculated by exchanging the derivative and expectation according to the dominant convergence theorem (Çinlar, 2011) which allows the derivatives to be calculated inside the expectations, and by applying the chain rule:

$$\nabla_{\mu} \text{ELBO} = E_{N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})}[\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{\text{obs}}) \nabla_{\theta} T^{-1}(\theta) + \nabla_{\theta} \log|det \mathbf{J}_{T^{-1}}(\theta)|]\quad (21)$$

The gradients of the ELBO with respect to \mathbf{L} can be written similarly:

$$\begin{aligned}\nabla_{\mathbf{L}} \text{ELBO} = \mathbb{E}_{N(\boldsymbol{\eta}|\mathbf{0}, \mathbf{I})} & [(\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{\text{obs}}) \nabla_{\boldsymbol{\theta}} T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log |\det \mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|) \boldsymbol{\eta}^T] \\ & + (\mathbf{L}^{-1})^T \quad (22)\end{aligned}$$

The expectations can be estimated using Monte Carlo (MC) integration which provides noisy, unbiased estimates of the expectations. The accuracy of MC integration increases with the number of samples, but in practice a low number or even a single sample can be sufficient at each iteration since optimisations are usually performed over many iterations so that statistically they will converge towards the solution (Kucukelbir et al., 2017). The variational problem in equation 18 can therefore be solved by standard gradient-based optimisation methods, by gradient ascent. The final approximation $q(\mathbf{m})$ can then be obtained by transforming the solution $q^*(\boldsymbol{\theta})$ back to the constrained parameter space, either numerically or analytically depending on the form of transform T . By combining with the automatic differential technique (Baydin, Pearlmutter, Radul, & Siskind, 2018; Wengert, 1964) the whole process can be conducted automatically, hence the name "automatic differential". The procedure is summarized in Algorithm 2.

Algorithm 2: Automatic differential variational inference (ADVI)

Input: The joint pdf $p(\mathbf{m}, \mathbf{d}_{\text{obs}})$ in a constrained space, which can be estimated for any particular value of \mathbf{m} ; a transform $T : \boldsymbol{\theta} = T(\mathbf{m})$ which transforms \mathbf{m} into an unconstrained variable $\boldsymbol{\theta}$ and an initial Gaussian distribution $q^0(\boldsymbol{\theta}; \boldsymbol{\mu}^0, \mathbf{L}^0)$ in the unconstrained space.

Output: A distribution $q^N(\mathbf{m})$ that approximates the posterior pdf.

for $l \leftarrow 1$ to N **do**

 Calculate gradients $\nabla_{\boldsymbol{\mu}^{l-1}} \text{ELBO}$ and $\nabla_{\mathbf{L}^{l-1}} \text{ELBO}$ using equation 21 and 22.

 Update $\boldsymbol{\mu}$ and \mathbf{L}

$$\begin{aligned}\boldsymbol{\mu}^l &= \boldsymbol{\mu}^{l-1} + \epsilon^l \nabla_{\boldsymbol{\mu}^{l-1}} \text{ELBO} \\ \mathbf{L}^l &= \mathbf{L}^{l-1} + \epsilon^l \nabla_{\mathbf{L}^{l-1}} \text{ELBO} \quad (23)\end{aligned}$$

 where ϵ^l is the step size at the l^{th} iteration.

end

Transform $q^N(\boldsymbol{\theta}; \boldsymbol{\mu}^N, \mathbf{L}^N)$ to $q^N(\mathbf{m})$ using T .

Note that the final approximation is determined by the Gaussian distribution $q^*(\boldsymbol{\theta})$ in the unconstrained space and by the transform T . Unfortunately the optimal transform is difficult to determine because it depends on the unknown properties of the posterior distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. A commonly-used transform is:

$$\begin{aligned}\theta_i &= T(m_i) = \log(m_i - a) - \log(b - m_i) \\ m_i &= T^{-1}(\theta_i) = a_i + \frac{(b_i - a_i)}{1 + \exp(-\theta_i)} \quad (24)\end{aligned}$$

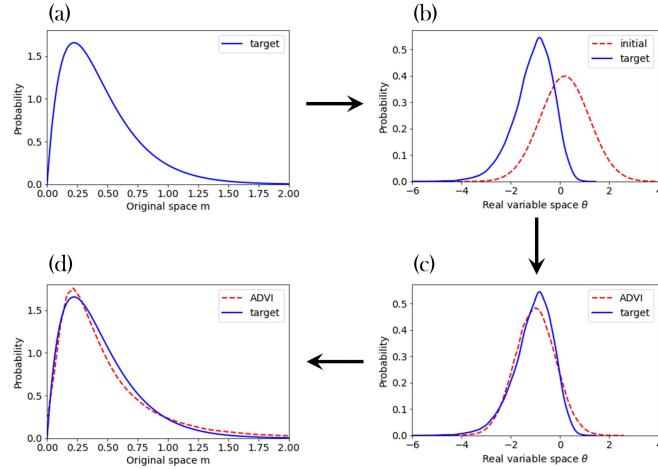


FIGURE 2 A 1D example of ADVI. **(a)** the target (posterior) pdf in the original positive half space. **(b)** The target pdf in the transformed unconstrained space (blue line) and an initial Gaussian approximation (red dashed line). **(c)** and **(d)** show the target pdf (blue line) and the approximation obtained using ADVI (red dashed line) in the unconstrained space and the original space, respectively.

where m_i represents the i^{th} parameter in the original constrained space, θ_i is the transformed unconstrained variable, and a_i and b_i are the lower and upper bound on m_i respectively (Team et al., 2016). The final approximation is then limited by the Gaussian distribution $q^*(\theta)$ and the transform in equation 24.

Figure 2 shows a 1D example of the ADVI. The true target (posterior) pdf is defined in the positive half space (Figure 2a). An initial Gaussian distribution is defined in the transformed unconstrained space (Figure 2b) and updated using the gradient ascent method (Figure 2c). The final approximation is obtained by transforming the obtained Gaussian distribution back to the original space (Figure 2d). Since the true distribution is a non-Gaussian distribution in both original and transformed space, the obtained approximation is different to the true distribution. This indicates that ADVI can produce biased results for non-Gaussian posterior pdfs.

Note that in very high dimensional space ADVI may become inefficient because of the large size of the full covariance matrix (the number of variables is proportional to the square of dimensionality). In such cases if correlation between certain parameters can be ignored, a diagonal covariance matrix or a sparse covariance matrix may be used to reduce computational cost. Due to the Gaussian variational family, ADVI cannot provide accurate approximations to multimodal distributions. However, further improvements are made possible by using a mixture of Gaussian distributions (Arenz, Zhong, & Neumann, 2018; Zobay et al., 2014).

2.3 Normalizing flows

The approximation to the posterior pdf obtained using ADVI is limited by a Gaussian distribution in the unconstrained space, and a fixed transform T that is used to transform that Gaussian distribution to the original parameter space. It should be possible to improve the approximation by finding a more suitable transform. This idea leads to a method called normalizing flows, in which a series of invertible and differential transforms (called flows) are applied to an initial known distribution (e.g. a Gaussian distribution); the flows are optimized to produce an improved approximation to the posterior pdf (Rezende & Mohamed, 2015).

Let \mathbf{m}_0 be a random vector variable which has a simple and analytically known pdf $q_0(\mathbf{m}_0)$, for example a Gaussian distribution, and apply an invertible transform \mathbf{F}_Θ (parameterized by Θ) such that $\mathbf{m}_1 = \mathbf{F}_\Theta(\mathbf{m}_0)$. The pdf of transformed variable \mathbf{m}_1 can be expressed as:

$$q_1(\mathbf{m}_1) = q_0(\mathbf{m}_0) \left| \det \frac{\partial \mathbf{F}_\Theta}{\partial \mathbf{m}_0} \right|^{-1} \quad (25)$$

where $\frac{\partial \mathbf{F}_\Theta}{\partial \mathbf{m}_0}$ is the Jacobian matrix of the transform \mathbf{F}_Θ . The pdf q_0 is called an initial distribution and the transform \mathbf{F}_Θ is referred to as a normalizing flow which pushes the simple and known pdf q_0 to a target pdf q_1 . Depending on the form of the normalizing flow, the initial pdf can be manipulated in different ways, for example it can be expanded, contracted, rotated or its location can be shifted to produce different target pdfs.

In Bayesian inference the goal is to estimate the posterior pdf, that is, to find a normalizing flow \mathbf{F}_Θ such that the pdf q_1 is close to the posterior pdf. However, in general it is difficult to construct a single flow that transforms a simple distribution to the posterior distribution given that real posterior pdfs often have complex forms (which a priori we do not know). Instead this ideal transform can be approximated by combining multiple simple flows and successively applying equation 25.

Assume we have K flows, $\mathbf{F}(\theta_0), \mathbf{F}(\theta_1), \dots, \mathbf{F}(\theta_{K-1})$, and successively apply them to the initial variable \mathbf{m}_0 :

$$\mathbf{m}_K = \mathbf{F}_{\theta_{K-1}} \cdot \mathbf{F}_{\theta_{K-2}} \cdots \mathbf{F}_{\theta_1} \cdot \mathbf{F}_{\theta_0}(\mathbf{m}_0) \quad (26)$$

where \mathbf{m}_K is the variable after the combined transformation. The pdf of \mathbf{m}_K can be obtained using equation 25:

$$q_K(\mathbf{m}_K) = q_0(\mathbf{m}_0) \prod_{i=0}^{K-1} \left| \det \frac{\partial \mathbf{F}_{\theta_i}}{\partial \mathbf{m}_i} \right|^{-1} \quad (27)$$

Hereafter for simplicity we use the notation $\Theta = (\theta_0, \theta_1, \dots, \theta_{K-1})$ and \mathbf{F}_Θ to represent the chain of transforms: $\mathbf{F}_\Theta = \mathbf{F}_{\theta_{K-1}} \cdot \mathbf{F}_{\theta_{K-2}} \cdots \mathbf{F}_{\theta_1} \cdot \mathbf{F}_{\theta_0}$, and

use $|det \frac{\partial \mathbf{F}_\Theta}{\partial \mathbf{m}_0}| = \prod_{i=0}^{K-1} |det \frac{\partial \mathbf{F}_{\Theta_i}}{\partial \mathbf{m}_i}|$. By using a series of transforms equation 27 improves the expressibility of the combined transformation, so that more complex final distributions can be created. Note that if we use an analytically known initial distribution and construct transforms such that their Jacobian determinants are also analytically known, the final distribution is also analytic.

To approximate the posterior pdf using the distribution $q_K(\mathbf{m}_K)$ obtained from normalizing flows, we optimize the flow parameters Θ by maximizing the ELBO as in equation 7. This results in a variational problem:

$$\begin{aligned}\Theta^* &= \arg \max_{\Theta} \text{ELBO}[q_K(\mathbf{m}_K)] \\ &= \arg \max_{\Theta} \mathbb{E}_{q_K} [\log p(\mathbf{m}_K, \mathbf{d}_{\text{obs}}) - \log q_K(\mathbf{m}_K)]\end{aligned}\quad (28)$$

According to the change of variables theorem, for any function $h(\mathbf{m}_K)$ the expectation with respect to $q_K(\mathbf{m}_K)$ can be expressed as:

$$\int h(\mathbf{m}_K) q_K(\mathbf{m}_K) d\mathbf{m}_K = \int h(\mathbf{m}_k) q_0(\mathbf{m}_0) d\mathbf{m}_0 \quad (29)$$

Combining equation 27 and 29 with equation 28 gives

$$\begin{aligned}\Theta^* &= \arg \max_{\Theta} \text{ELBO}[q_K(\mathbf{m}_K)] \\ &= \arg \max_{\Theta} \mathbb{E}_{q_0} [\log p(\mathbf{m}_K, \mathbf{d}_{\text{obs}}) - \log q_0(\mathbf{m}_0) + \log |det \frac{\partial \mathbf{F}_\Theta}{\partial \mathbf{m}_0}|]\end{aligned}\quad (30)$$

where the expectation is taken with respect to the initial distribution $q_0(\mathbf{m}_0)$. This problem can be solved using standard gradient-based optimization methods, for example, gradient ascent. Similarly to the gradient computations in ADVI, the gradients of ELBO with respect to Θ can be obtained by exchanging the expectations and derivatives and by applying the chain rule:

$$\nabla_\Theta \text{ELBO} = \mathbb{E}_{q_0} \left[\nabla_{\mathbf{m}_K} \log p(\mathbf{m}_K, \mathbf{d}_{\text{obs}}) \nabla_\Theta \mathbf{m}_K + \nabla_\Theta \log |det \frac{\partial \mathbf{F}_\Theta}{\partial \mathbf{m}_0}| \right] \quad (31)$$

As in ADVI the expectation can be calculated using MC integration over a small number of samples, and the resulting gradients can be used to solve the optimization problem using gradient ascent methods. The final approximation can be obtained using equation 27 with the optimal parameters Θ^* . The procedure is summarized in Algorithm 3.

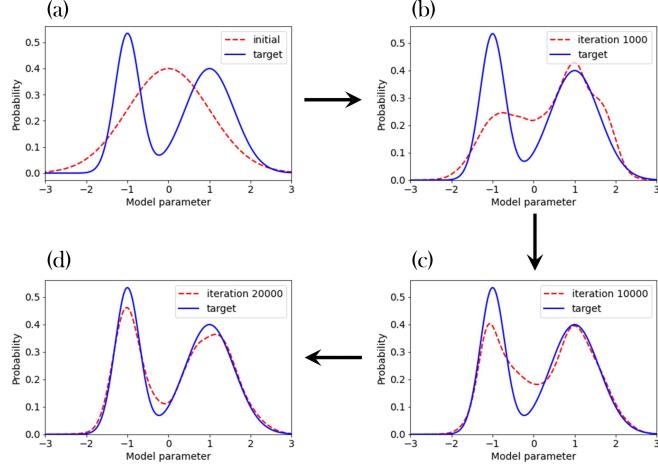


FIGURE 3 A 1D example using normalizing flows. (a) the true or target pdf (blue line) and the initial pdf (red dashed line). (b), (c) and (d) show the estimated pdfs after 1000, 10000 and 20000 iterations of gradient ascent respectively.

Algorithm 3: Normalizing flows

Input: An initial pdf $q_0(\mathbf{m}_0)$; the joint pdf $p(\mathbf{m}, \mathbf{d}_{\text{obs}})$; a series of flows $\mathbf{F}_\Theta = \mathbf{F}_{\Theta_{K-1}} \cdot \mathbf{F}_{\Theta_{K-2}} \cdots \mathbf{F}_{\Theta_1} \cdot \mathbf{F}_{\Theta_0}$ parameterized by $\Theta = (\Theta_0, \Theta_1, \dots, \Theta_{K-1})$.

Output: A distribution $q(\mathbf{m})$ that approximates the posterior pdf.
Initialize Θ with Θ^0 .

for $l \leftarrow 1$ to N **do**

 Calculate gradients $\nabla_{\Theta^{l-1}} \text{ELBO}$ using equation 31.

 Update Θ

$$\Theta^l = \Theta^{l-1} + \epsilon^l \nabla_{\Theta^{l-1}} \text{ELBO} \quad (32)$$

 where ϵ^l is the step size at the l^{th} iteration.

end

Obtain final approximation $q(\mathbf{m}) = q_0(\mathbf{m}_0) |\det \frac{\partial \mathbf{F}_{\Theta^N}}{\partial \mathbf{m}_0}|$.

As described in section 2.2, in ADVI we apply two transforms: one transforms constrained variables to unconstrained variables and the other transforms a Gaussian distribution to a standard Gaussian distribution. The first transform is fixed, while the parameters of the latter transform (the mean and covariance matrix of the Gaussian distribution) are optimized such that they maximise the ELBO between the Gaussian distribution and the posterior pdf in the unconstrained space (equation 18). Thus, ADVI is in fact a single normalizing flow.

To construct a flexible normalizing flow for practical applications, several conditions are required: the flows must be 1) invertible, and 2) expressive

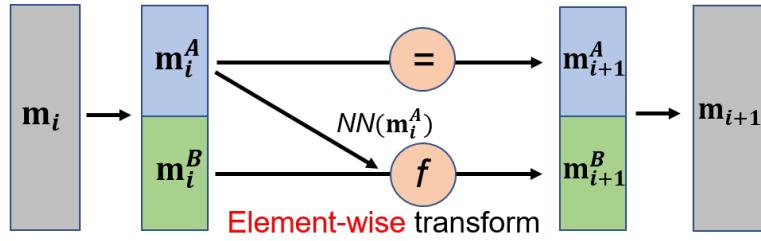


FIGURE 4 An illustration of a coupling flow. The input vector \mathbf{m}_i is first divided into two halves \mathbf{m}_i^A and \mathbf{m}_i^B . The former half \mathbf{m}_i^A is copied as the first half \mathbf{m}_{i+1}^A of the output. It is also input to a neural network which outputs hyperparameters of an element-wise function f ; this function f transforms the second half \mathbf{m}_i^B to \mathbf{m}_{i+1}^B which together with \mathbf{m}_{i+1}^A forms the output \mathbf{m}_{i+1} .

enough to represent any desired pdf; 3) The forward and backward transform and associated Jacobian determinant must be able to be computed efficiently. A simple example of such flows is planar flow:

$$\mathbf{m}_{i+1} = \mathbf{m}_i + \mathbf{u} h(\mathbf{w}^T \mathbf{m}_i + b) \quad (33)$$

where \mathbf{u} and \mathbf{w} are vectors, b is a scalar and h is a smooth function (Rezende & Mohamed, 2015): $h(x) = \tanh(x)$ is usually used. The determinant of the Jacobian matrix of this flow is:

$$\det \frac{\partial \mathbf{m}_{i+1}}{\partial \mathbf{m}_i} = 1 + \mathbf{u}^T h'(\mathbf{w}^T \mathbf{m}_i + b) \mathbf{w} \quad (34)$$

The planar flow essentially expands or contracts a distribution along the direction perpendicular to the hyperplane $\mathbf{w}^T \mathbf{m}_i + b = 0$, and can be interpreted as a neural network with one hidden layer and one hidden unit (Kingma & Dhariwal, 2018).

Figure 3 shows a 1D example using planar flows. The true target (posterior) pdf is a multimodal distribution (blue line in Figure 3a). We use a standard normal distribution as the initial distribution and a normalizing flows model with 10 planar flows in equation 33. The model parameters are updated using gradient ascent with gradients calculated using equation 31. Figure 3b, c and d show the estimated pdfs after 1000, 10000 and 20000 iterations respectively. The initial pdf is gradually reshaped and finally produces an accurate approximation to the true pdf.

It becomes difficult to use planar flows to approximate complex posterior distributions in high dimensionality in the sense that each planar flow is a neural network with the necessarily simple structure of only one hidden layer and one hidden unit. To improve the expressiveness of the sequence of flows in equation 27, many different forms of flow have been proposed (an overview is given in X. Zhao et al., 2020). One such flow is constructed from an invertible neural network with a specific design to enable invertibility and fast computation of the Jacobian determinant (Behrmann, Grathwohl, Chen, Duvenaud, & Jacobsen,

2019; Dinh, Sohl-Dickstein, & Bengio, 2016; Greydanus, Dzamba, & Yosinski, 2019; Kingma & Dhariwal, 2018). In this study we describe an invertible neural network called a *coupling flow* (Dinh et al., 2016). In a coupling flow an input vector \mathbf{m}_i is divided into two half vectors \mathbf{m}_i^A and \mathbf{m}_i^B , and the output halves \mathbf{m}_{i+1}^A and \mathbf{m}_{i+1}^B are obtained using (see Figure 4):

$$\begin{aligned}\mathbf{m}_{i+1}^A &= \mathbf{m}_i^A \\ \mathbf{m}_{i+1}^B &= f(\mathbf{m}_i^B; NN(\mathbf{m}_i^A))\end{aligned}\tag{35}$$

where $NN(\mathbf{m}_i^A)$ represents any neural network which takes \mathbf{m}_i^A as input, and f transforms \mathbf{m}_i^B to \mathbf{m}_{i+1}^B and is an invertible, element-wise bijection function parameterized by the output of the neural network. The two halves \mathbf{m}_{i+1}^A and \mathbf{m}_{i+1}^B are combined to obtain the output vector \mathbf{m}_{i+1} . This transform can be easily inverted

$$\begin{aligned}\mathbf{m}_i^A &= \mathbf{m}_{i+1}^A \\ \mathbf{m}_i^B &= f^{-1}(\mathbf{m}_{i+1}^B; NN(\mathbf{m}_i^A))\end{aligned}\tag{36}$$

and the Jacobian determinant of the transform can also be calculated using

$$\det \frac{\partial \mathbf{m}_{i+1}}{\partial \mathbf{m}_i} = \det \begin{bmatrix} \frac{\partial \mathbf{m}_{i+1}^A}{\partial \mathbf{m}_i^A} & \frac{\partial \mathbf{m}_{i+1}^A}{\partial \mathbf{m}_i^B} \\ \frac{\partial \mathbf{m}_{i+1}^B}{\partial \mathbf{m}_i^A} & \frac{\partial \mathbf{m}_{i+1}^B}{\partial \mathbf{m}_i^B} \end{bmatrix} = \det \frac{\partial \mathbf{m}_{i+1}^B}{\partial \mathbf{m}_i^B}\tag{37}$$

where we have used the fact that $\frac{\partial \mathbf{m}_{i+1}^A}{\partial \mathbf{m}_i^A} = \mathbf{I}$ and $\frac{\partial \mathbf{m}_{i+1}^A}{\partial \mathbf{m}_i^B} = \mathbf{0}$. Since the function f is an element-wise function, the matrix $\frac{\partial \mathbf{m}_{i+1}^B}{\partial \mathbf{m}_i^B}$ is a diagonal matrix whose determinant can be calculated efficiently.

In practice a series of successive coupling flows are used to improve the expressiveness of the overall transform. To ensure that all elements in the input vector \mathbf{m}_i are modified, the locations of the two outputs \mathbf{m}_{i+1}^A and \mathbf{m}_{i+1}^B are exchanged before feeding into the next flow. The function f can be any element-wise functions which is invertible and differentiable, and many choices of f can be used in practice (De Cao, Aziz, & Titov, 2020; Dinh, Krueger, & Bengio, 2014; Dinh et al., 2016; Durkan, Bekasov, Murray, & Papamakarios, 2019a, 2019b; Kingma & Dhariwal, 2018).

Note that instead of coupling flows, other designs of invertible neural networks can also be used in normalizing flows, for example invertible residual networks (Behrmann et al., 2019), neural ordinary differential equations (R. T. Chen, Rubanova, Bettencourt, & Duvenaud, 2018; Grathwohl, Chen, Bettencourt, Sutskever, & Duvenaud, 2018) or Hamiltonian neural networks (Greydanus et al., 2019). Further research that performs fair comparisons between these networks would be a useful contribution.

2.4 Stein variational gradient descent

In normalizing flows a series of analytical invertible transforms are applied to a simple initial distribution and are optimized by maximizing the ELBO between the final transformed distribution and the posterior distribution. In practice construction of effective analytic transforms can be a difficult task. Instead of using analytical transforms, Stein variational gradient descent (SVGD) uses a smooth transform whose analytical form remains unknown, and successively applies it to an initial probability distribution represented by a set of parameter-space samples which are referred to as particles (Q. Liu & Wang, 2016). Similarly to normalizing flows, the transforms are optimized to minimize the KL-divergence between the transformed distribution and the posterior distribution so that the final set of particles are distributed according to the posterior.

In SVGD a smooth transform is used:

$$T(\mathbf{m}) = \mathbf{m} + \epsilon \Phi(\mathbf{m}) \quad (38)$$

where \mathbf{m} is a d -dimensional vector, $\Phi(\mathbf{m}) = [\phi_1, \dots, \phi_d]$ is a smooth d -dimensional vector function which describes the perturbation direction and ϵ is the magnitude of the perturbation. When ϵ is sufficiently small, the transform T is invertible as the Jacobian matrix is close to an identity matrix. Define q as an initial distribution and q_T as the transformed distribution, the gradient of KL-divergence between q_T and the posterior pdf p with respect to ϵ can be calculated as:

$$\nabla_\epsilon \text{KL}[q_T || p] |_{\epsilon=0} = -E_q[\text{trace}(\mathcal{A}_p \Phi(\mathbf{m}))] \quad (39)$$

where \mathcal{A}_p is the Stein operator such that $\mathcal{A}_p \Phi(\mathbf{m}) = \nabla_{\mathbf{m}} \log p(\mathbf{m} | \mathbf{d}_{\text{obs}}) \Phi(\mathbf{m})^T + \nabla_{\mathbf{m}} \Phi(\mathbf{m})$ (Q. Liu & Wang, 2016). This implies that by maximizing the right-hand side expectation we obtain the steepest direction of change in the KL-divergence; the KL-divergence can therefore be minimized by iteratively stepping a small distance in that direction.

The optimal Φ^* which maximizes the expectation in equation 39 can be found using kernels. Assume $x, y \in X$ and define a mapping φ from X to an inner product space; a *kernel* is a function which satisfies $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ where $\langle \cdot, \cdot \rangle$ represents an inner product (Gretton, 2013). The optimal Φ^* is found to be:

$$\Phi^* \propto E_{\{\mathbf{m}' \sim q\}}[\mathcal{A}_p k(\mathbf{m}', \mathbf{m})] \quad (40)$$

where $k(\mathbf{m}', \mathbf{m})$ is a kernel function (Q. Liu & Wang, 2016).

Given equation 40, the KL-divergence can be minimized by iteratively applying the transform in equation 38 with the optimal Φ^* to an initial distribution. For example, define an initial distribution q_0 , and apply the transform $T_0(\mathbf{m}) = \mathbf{m} + \epsilon \Phi_0^*(\mathbf{m})$ where $\Phi_0^*(\mathbf{m})$ is given in equation 40. This produces a new distribution $q_{[T_0]}$ which decreases the KL-divergence. This process is

iterated to obtain an approximation to the posterior:

$$\begin{aligned} T_l(\mathbf{m}) &= \mathbf{m} + \epsilon_l \Phi_l^*(\mathbf{m}) \\ q_{l+1} &= q_l[T_l] \end{aligned} \tag{41}$$

where the subscript l denotes the l^{th} iteration. If the perturbation magnitude $\{\epsilon_l\}$ is sufficiently small, that is, the transform is invertible at each iteration, the process should eventually converge to the posterior distribution.

In practice since the posterior distribution $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ and its gradient with respect to model \mathbf{m} are analytically unknown (and are needed in the definition of the Stein operator \mathcal{A}_p), we cannot obtain the analytical form of the optimal Φ^* and consequently the optimal transform T . Fortunately the unnormalized posterior distribution can usually be estimated at a set of samples $\{\mathbf{m}_1, \dots, \mathbf{m}_n\}$ distributed approximately according to the posterior pdf, which enables us to estimate the optimal Φ^* numerically, for example using the mean value taken over the set of samples. Thus in SVGD we use a set of samples $\{\mathbf{m}_i\}$ (the particles) to represent the approximate distribution q and to approximate the optimal Φ^* using the particles mean. Each particle is then updated using the estimated Φ^* . This results in Algorithm 4.

Algorithm 4: Stein Variational gradient descent (SVGD)

Input: An initial pdf q_0 ; the posterior pdf $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ which can be estimated up to a normalising constant for any particular value of \mathbf{m} .
Output: A set of particles $\{\mathbf{m}_i\}$ whose density approximates the posterior pdf.
Draw a set of particles $\{\mathbf{m}_i^0\}_{i=1}^n$ from q_0 ;
for $l \leftarrow 1$ to N **do**

$$\begin{aligned} \Phi_{q_l, p}^*(\mathbf{m}) &= \frac{1}{n} \sum_{j=1}^n \left[k(\mathbf{m}_j^l, \mathbf{m}) \nabla_{\mathbf{m}_j^l} \log p(\mathbf{m}_j^l | \mathbf{d}_{\text{obs}}) + \nabla_{\mathbf{m}_j^l} k(\mathbf{m}_j^l, \mathbf{m}) \right] \\ \mathbf{m}_i^{l+1} &= \mathbf{m}_i^l + \epsilon^l \Phi_{q_l, p}^*(\mathbf{m}_i^l) \end{aligned} \tag{42}$$

where ϵ^l is the step size at the l^{th} iteration.

end

Since SVGD uses particles to approximate the posterior pdf, the accuracy of the method increases with the number of particles. For sufficiently small $\{\epsilon_l\}$ the method converges to the posterior distribution asymptotically with the number of particles. On the other hand for one single particle the method reduces to a standard gradient ascent method towards the model with maximum a posterior (MAP) pdf value if the gradient $\nabla_{\mathbf{m}} k(\mathbf{m}, \mathbf{m})$ vanishes (which is valid for many kernels, including the radial basis function kernel described below). This suggests that in practice we can start from a small number of particles and gradually increase the particles to produce more accurate results. In comparison

to other particle-based methods, for example, sequential Monte Carlo (Smith, 2013), SVGD requires fewer samples to achieve the same accuracy which makes it more efficient (Q. Liu & Wang, 2016). It is also important to notice that sequential Monte Carlo is a stochastic sampling method, whereas SVGD is a deterministic sampling method.

The kernel function enables interactions between particles and strongly affects the efficiency of the method. We first describe a simple, commonly-used kernel function, the *radial basis function* (RBF)

$$k(\mathbf{m}, \mathbf{m}') = \exp\left[-\frac{\|\mathbf{m} - \mathbf{m}'\|^2}{2\sigma^2}\right] \quad (43)$$

where σ is a scale factor which intuitively controls the interaction intensity between pairs of particles based on their distance apart.

With a RBF kernel the first term of Φ^* in equation 42 is the weighted average of gradients of the posterior pdf from all particles, in which the weights are determined by particle distances and the scale factor σ . This term drives particles towards a local high probability area. The second term of Φ^* becomes $\sum_j \frac{\mathbf{m} - \mathbf{m}_j}{\sigma^2} k(\mathbf{m}_j, \mathbf{m})$ which pushes the particle \mathbf{m} away from its neighbouring particles with high kernel values. The two terms therefore contribute in different ways to arrange particles to represent the posterior pdf: the first term drives particles towards a local high probability area, whereas the second term acts as a *repulsive force* which prevents particles from collapsing to a single mode. These terms balance such that the limiting distribution is the posterior pdf provided that the derivative of the kernels (the second term of Φ^* in equation 42) does not vanish. Note that when $\sigma \rightarrow 0$, the method becomes independent gradient ascent for each particle as the kernel value and its derivative between any two particles vanish.

Figure 5 shows a 1D example using SVGD with a RBF kernel. The target pdf is the same multimodal distribution in Figure 3a (blue line). We start from 1,000 particles generated from a standard Normal distribution (red histograms in Figure 5a) and iteratively update them using equation 42. Figure 5b, c and d show the histograms of those particles after 5, 100 and 500 iterations respectively. After 100 iterations the method has almost converged to the true distribution. This example shows that SVGD arranges particles to represent the posterior pdf optimally.

Kernel functions can be generalized to matrix forms and used in SVGD instead of scalar kernel functions. By doing this one can inject information about correlations between the different parameters in \mathbf{m} into the method. Assuming a matrix-valued kernel function \mathbf{K} , the Φ^* in equation 42 becomes:

$$\Phi_{q_l, p}^*(\mathbf{m}) = \frac{1}{n} \sum_{j=1}^n \left[\mathbf{K}(\mathbf{m}_j^l, \mathbf{m}) \nabla_{\mathbf{m}_j^l} \log p(\mathbf{m}_j^l | \mathbf{d}_{\text{obs}}) + \mathbf{K}(\mathbf{m}_j^l, \mathbf{m}) \nabla_{\mathbf{m}_j^l} \right] \quad (44)$$

where $\mathbf{K}(\mathbf{m}_j^l, \mathbf{m}) \nabla_{\mathbf{m}_j^l}$ represents matrix multiplication (Wang, Tang, Bajaj, &

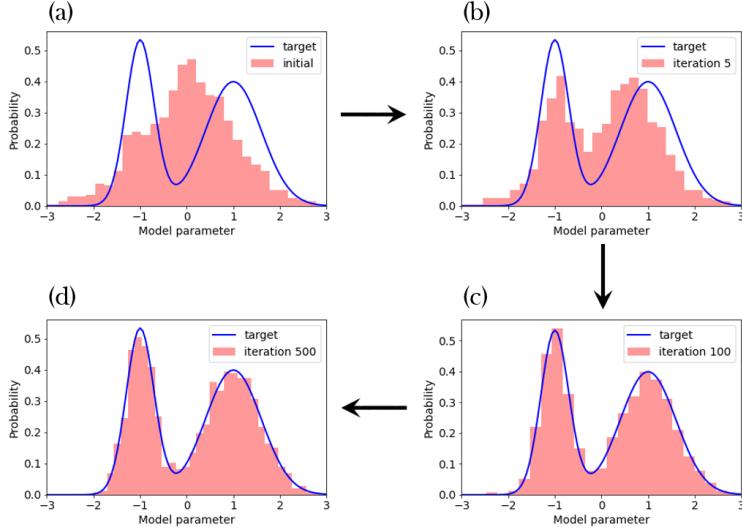


FIGURE 5 A 1D example using SVGD. (a) The true pdf (blue line) and the histogram of 1000 initial particles (red) which are generated from a Gaussian distribution. (b), (c) and (d) show the histograms of the particles after 5, 100 and 500 iterations respectively.

Liu, 2019). A possible choice of a matrix-valued kernel is:

$$\mathbf{K}(\mathbf{m}', \mathbf{m}) = \mathbf{Q}^{-1} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{m} - \mathbf{m}'\|_{\mathbf{Q}}^2\right) \quad (45)$$

where \mathbf{Q} is a positive definite matrix, $\|\mathbf{m} - \mathbf{m}'\|_{\mathbf{Q}}^2 = (\mathbf{m} - \mathbf{m}')^T \mathbf{Q} (\mathbf{m} - \mathbf{m}')$ and σ is a scaling parameter. This kernel is essentially a RBF kernel with a preconditioning matrix \mathbf{Q} . Wang et al. (2019) showed that by setting \mathbf{Q} to be the average Hessian matrix of all particles, the method converges faster than a scalar RBF kernel. Other choice of \mathbf{Q} include the inverse of the covariance matrix calculated from the particles, or the inverse of the diagonal covariance (variance) matrix.

3 APPLICATIONS

3.1 Petrophysical inversion

In this section, we present an application of variational inference using the mean field approximation for joint estimation of geological facies κ and petrophysical rock properties γ using information derived from seismic data that are referred to as seismic attributes \mathbf{d} . These attributes represent elastic rock properties that may directly be inverted from seismic waveform data such as P- and S-wave impedances (I_p and I_s), velocities (V_p and V_s) and their ratios (V_p/V_s).

Examples of petrophysical properties γ of interest include porosity (ϱ), clay volume (V_{cl}) and water saturations (S_w). Geological facies refer to well-defined discrete classes of lithology and fluid types that are in principle distinctively distinguishable from seismic and well data. Petrophysical rock properties and facies together represent the unknown model parameters, i.e. $\mathbf{m} \equiv \{\gamma, \kappa\}$.

Estimation of rock properties from seismic attributes is a non-unique inverse problem. Usually the solution can be better constrained if the spatial distribution of geological facies is known (Nawaz et al., 2020). For this reason, we would like to infer the rock properties γ and facies κ jointly from the seismic attributes \mathbf{d} along with their associated uncertainty of prediction. In terms of probability theory, we seek the posterior distribution $p(\gamma, \kappa|\mathbf{d})$ of unknown model parameters γ and κ conditioned on the attribute data \mathbf{d} . According to Bayes' theorem

$$p(\gamma, \kappa|\mathbf{d}) = \frac{p(\mathbf{d}|\gamma, \kappa)p(\gamma|\kappa)p(\kappa)}{p(\mathbf{d})} \quad (46)$$

where $p(\kappa)$ represents the prior distribution of facies κ , $p(\gamma|\kappa)$ represents the conditional prior distribution of the petrophysical properties γ given the facies κ , $p(\mathbf{d}|\gamma, \kappa)$ represents the data likelihood given γ and κ , and $p(\mathbf{d})$ represents the marginal distribution of data \mathbf{d} . Since the data \mathbf{d} is observed, $p(\mathbf{d})$ is a constant that normalizes the posterior distribution.

The joint distribution $p(\kappa)$ of facies is modelled as a Markov random field (MRF) with pair-wise correlations, which according to equation (9) is given by

$$p(\kappa) = \frac{1}{Z} \prod_{i,j} \psi_{ij}(\kappa_i, \kappa_j) \quad (47)$$

where the potential functions $\psi_{ij}(\kappa_i, \kappa_j)$ define how probable it is to find the facies κ_i and κ_j in locations i and j in the model, and may be estimated by scanning a training image (Mariethoz & Caers, 2014) and building histograms for various combinations of facies over various neighbouring locations.

The conditional prior distribution $p(\gamma|\kappa)$ of γ given κ is usually modelled using well logs that have been up-scaled to the dominant seismic wavelength (Grana & Della Rossa, 2010), and the likelihood $p(\mathbf{d}|\gamma, \kappa)$ is usually modelled using rock physics models (Grana, 2018; Grana & Della Rossa, 2010) calibrated with the well data and local geological information. We adopt a different approach: we model both the conditional prior $p(\gamma|\kappa)$ and the likelihood $p(\mathbf{d}|\gamma, \kappa)$ jointly using up-scaled well-logs in the form of a joint distribution $p(\mathbf{d}, \gamma|\kappa, \Theta)$ of elastic attributes \mathbf{d} and petrophysical properties γ given the facies κ , parameterized by Θ . Equation (46) may then be written as

$$p(\gamma, \kappa|\mathbf{d}, \Theta) = \frac{p(\mathbf{d}, \gamma|\kappa, \Theta)p(\kappa)}{p(\mathbf{d}|\Theta)} \quad (48)$$

Thus, we do not use a rock physics model explicitly. However, if only limited well data is available, rock physics models may be used to augment the existing data.

We use a Gaussian mixture (GM) distribution to model $p(\mathbf{d}, \boldsymbol{\gamma} | \boldsymbol{\kappa}, \Theta)$ that is defined as a linear combination of Gaussian kernels, usually referred to as the components of the mixture distribution. A GM distribution is a universal approximator of pdfs: given a sufficient number of Gaussian kernels with appropriate parameters, a GM can approximate any complex pdf to any desired non-zero accuracy (McLachlan & Peel, 2004). The GM distribution for rock properties d_i and γ_i given facies κ_i at a location i may be expressed as

$$p(d_i, r_i | \kappa_i, \Theta) = \sum_{t=1}^{T_k} \alpha_{t,k} g_{t,k}(d_i, r_i), \quad \forall i \quad (49)$$

where T_k is the number of mixture components (which may be different for each facies k), $\alpha_{t,k}$ is the component weight, and $g_{t,k}(d_i, \gamma_i)$ is the Gaussian kernel for the t^{th} component given by

$$g_{t,k}(d_i, r_i) = N\left(\begin{bmatrix} \boldsymbol{\mu}_d \\ \boldsymbol{\mu}_r \end{bmatrix}_{t,k}, \begin{bmatrix} \boldsymbol{\Sigma}_{d,d} & \boldsymbol{\Sigma}_{d,r} \\ \boldsymbol{\Sigma}_{r,d} & \boldsymbol{\Sigma}_{r,r} \end{bmatrix}_{t,k}\right), \quad \forall i \quad (50)$$

where N represents the pdf of the Normal distribution, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are means and block covariance matrices of the kernel with subscripts indicating the components with respect to the data \mathbf{d} and the petrophysical properties $\boldsymbol{\gamma}$.

Since the joint conditional distribution $p(\mathbf{d}, \boldsymbol{\gamma} | \boldsymbol{\kappa}, \Theta)$ of seismic attributes \mathbf{d} and rock properties $\boldsymbol{\gamma}$ given facies $\boldsymbol{\kappa}$ (and the distribution parameters Θ) is modelled as a GM distribution, and the prior distribution of facies $p(\boldsymbol{\kappa})$ is modelled as a MRF, the overall model of the joint distribution $p(\mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\kappa} | \Theta)$ of the data \mathbf{d} and unknown model parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\kappa}$ represents a Gaussian mixture - Markov random field (GM-MRF) given by

$$p(\boldsymbol{\gamma}, \boldsymbol{\kappa} | \mathbf{d}, \Theta) = \frac{p(\mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\kappa} | \Theta)}{p(\mathbf{d} | \Theta)} \cong \frac{1}{Z'} \prod_i p(d_i, r_i | \kappa_i, \Theta) \prod_{(i,j)} \psi_{ij}(\kappa_i, \kappa_j) \quad (51)$$

where $p(\mathbf{d} | \Theta)$ has been absorbed in the normalization constant Z' on the right-hand side. This demonstrates that although we only assumed that the prior distribution $p(\boldsymbol{\kappa})$ on facies $\boldsymbol{\kappa}$ is a MRF, the posterior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\kappa} | \mathbf{d}, \Theta)$ and the joint distribution $p(\mathbf{d}, \boldsymbol{\gamma}, \boldsymbol{\kappa} | \Theta)$ then also turn out to be MRFs. This is a consequence of the conditional independence assumption on the rock properties \mathbf{d} and $\boldsymbol{\gamma}$ that is invoked in the mean-field approximation. The factorization of the posterior distribution in equation (51) is instrumental in making inference tractable for real-scale models using, for example, the EM method of inference as described in section 2.1.

3.1.1 Results

We now show the application of the joint inversion method to estimate the spatial distribution of petrophysical rock properties and geological facies from well data

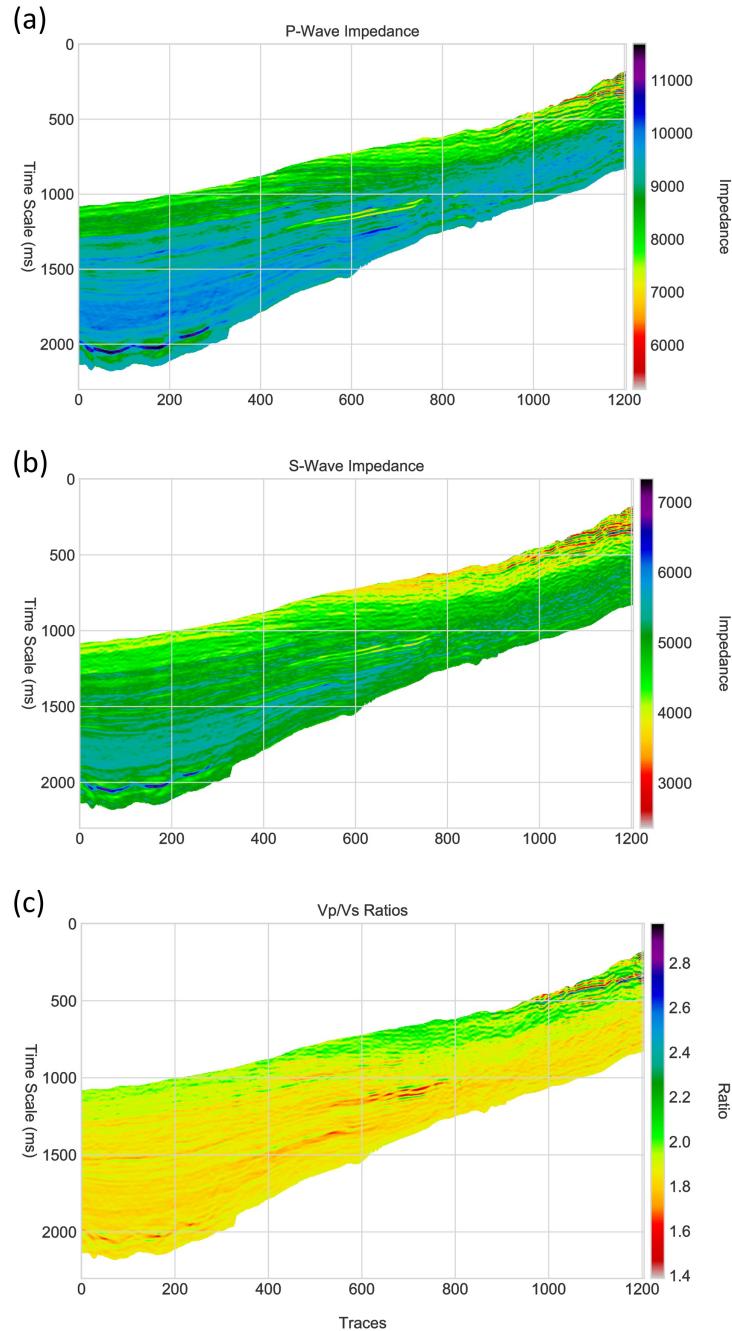


FIGURE 6 Seismic attributes (a) P-wave impedance, (b) S-wave impedance, and (c) Vp/Vs ratio, derived from a selected 2D section of waveform seismic data using a deterministic inversion method. These attributes are used as inputs to our method for the joint inversion of geological facies and petrophysical rock properties.

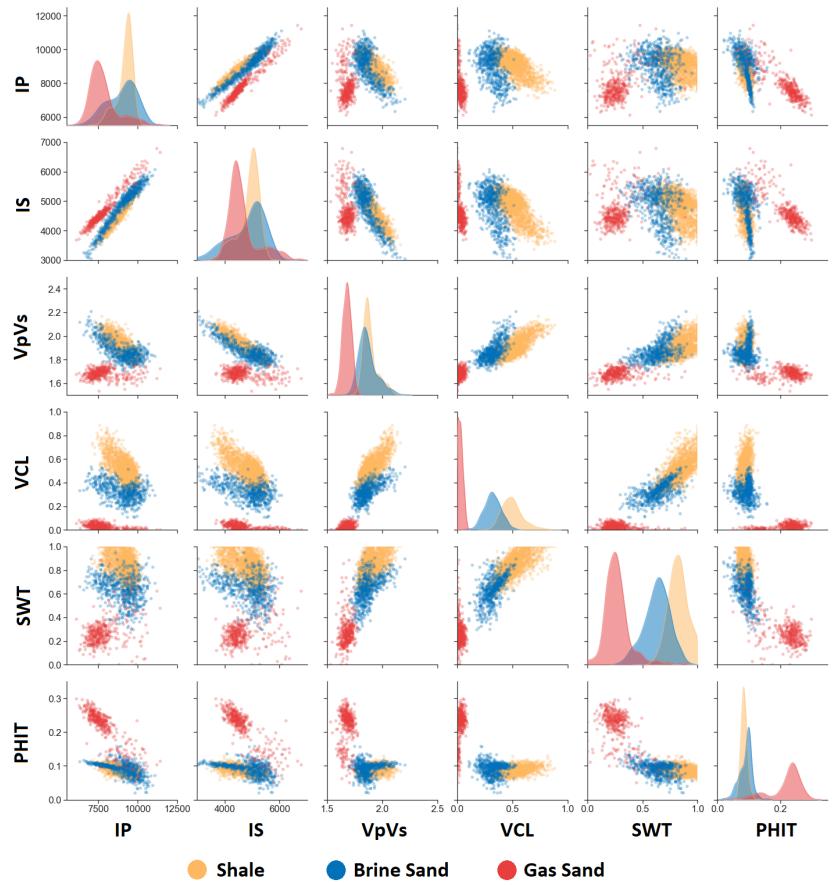


FIGURE 7 Seismic attributes P-wave impedance (IP), S-wave impedance (IS), and Vp/Vs ratio (VpVs), and petrophysical properties clay volume (VCL), water saturations (SWT) and porosity (PHIT) of three geological facies: Shale, Brine Sand and Gas Sand obtained from the well log data.

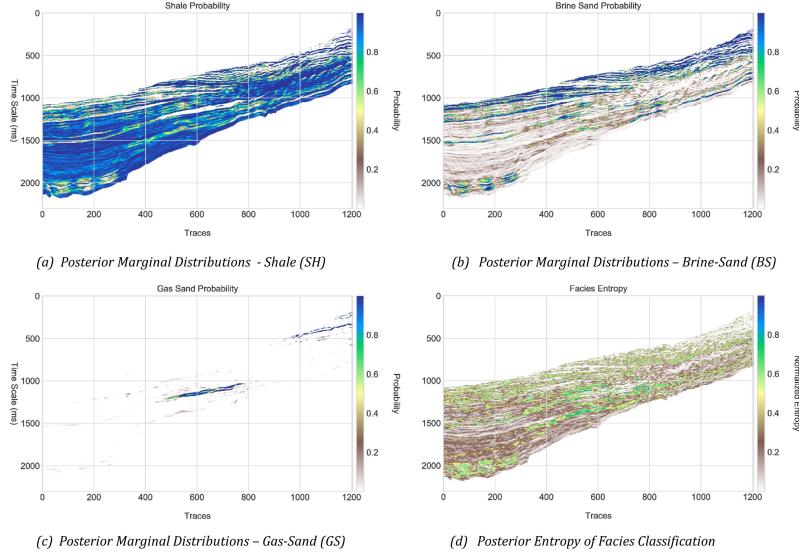


FIGURE 8 Cell-wise posterior marginal distributions of (a) shale, (b) brine-sand, (c) gas-sand, and (d) the posterior marginal entropy of facies classification scaled between 0.0 and 1.0. Yellow colour represents high probability or entropy (value=1.0) and dark blue colour represents low probability or entropy (value=0.0).

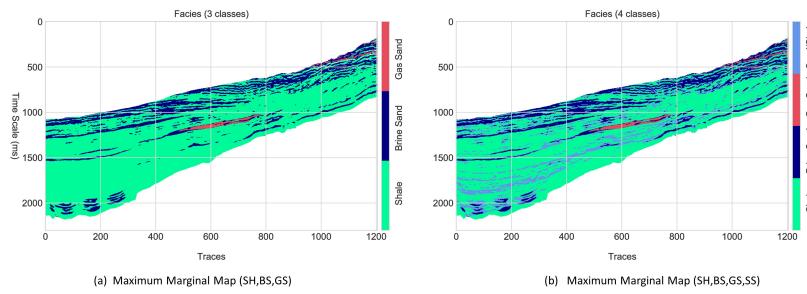


FIGURE 9 Cell-wise maps of facies with maximum marginal distribution. (a) Map of the three inverted facies: Shale (SH: shown in yellow), brine-sand (BS: blue) and gas-sand (GS: red). (b) Map with an additional facie “Shale/Sand” (SS: brown) identified from high entropy layers in Figure 8(d).

and seismic attributes from a gas field in the North Sea. This example is based on that in Nawaz et al. (2020), where the available data includes vertical 2D sections of seismic attributes: I_p , I_s , and V_p/V_s (Figure 6), and well logs from two wells that are located on the available 2D seismic section. The seismic attributes were available from a previous deterministic inversion of seismic waveform data. We are interested in classifying the seismic attribute data into three geological facies: shale, brine-sand and gas-sand, which are identified from the well log data (Figure 7). We notice from the top-left 3x3 sub-plots in Figure 7 that there is a significant overlap between the shale and brine-sand elastic properties. However, we may notice from rest of the sub-plots that these shale and brine sand may be resolved better when elastic properties are analyzed jointly with the petrophysical properties of interest: V_{cl} , S_w and ϱ . This forms the geophysical basis for our approach to jointly invert facies and petrophysical properties from seismic attributes (elastic rock properties). Further, it may also be noticed that since well logs are recorded at a much higher resolution than seismic data, a higher number of facies could be identified from the well log data (e.g. silt, sandy-shale and shaly-sand). However, we limited our analysis to the three main facies (shale, brine-sand and gas-sand) because we hypothesized at this stage that any further sub-division of shale and sand may not be identifiable from the seismic data due to limited resolution. However, contrary to our hypothesis, we later found that seismic data could resolve at least one more facies (shaly-sand or sandy-shale) as we describe below.

We used the EM method to invert the available elastic seismic attributes jointly for the spatial distributions of facies and petrophysical rock properties. The estimated marginal posterior distributions (under the mean field approximation) of the three facies and the entropy (a measure of uncertainty) of these distributions scaled between 0.0 and 1.0 is shown in Figure 8. The entropy is mostly low except at the transitions between different facies, but it appears to be high within some layers too. Since gas-sands typically have well discriminated properties, high entropy within some layers indicates the presence of a mixture of brine-sand and shale lithology that is not well discriminated. Figure 9(a) shows the facies map with maximum marginal probability in each model cell for the three inverted facies: shale, brine-sand, and gas-sand. Figure 9(b) shows the facies map with an additional facies defined as a combination of non-discriminated shale-sand identified to exist in the cells where entropy is greater than a cut-off value of 0.5 (i.e. 50% of the scaled entropy range from 0.0 to 1.0). Even though we inverted for 3 facies, the entropy of the marginal posterior distributions identifies that an additional facies may also be interpreted as shaly-sand or sandy-shale shown in brown colour in Figure 9(b).

The inverted petrophysical properties along with their standard deviations are shown in Figure 10. The seismic attribute inversion results are compared with the well data for verification and are shown in Figure 11. The measured well logs are shown in solid-black curves for reference. The solid-red curves are the input seismic attributes along the borehole in columns 1-3 and are means

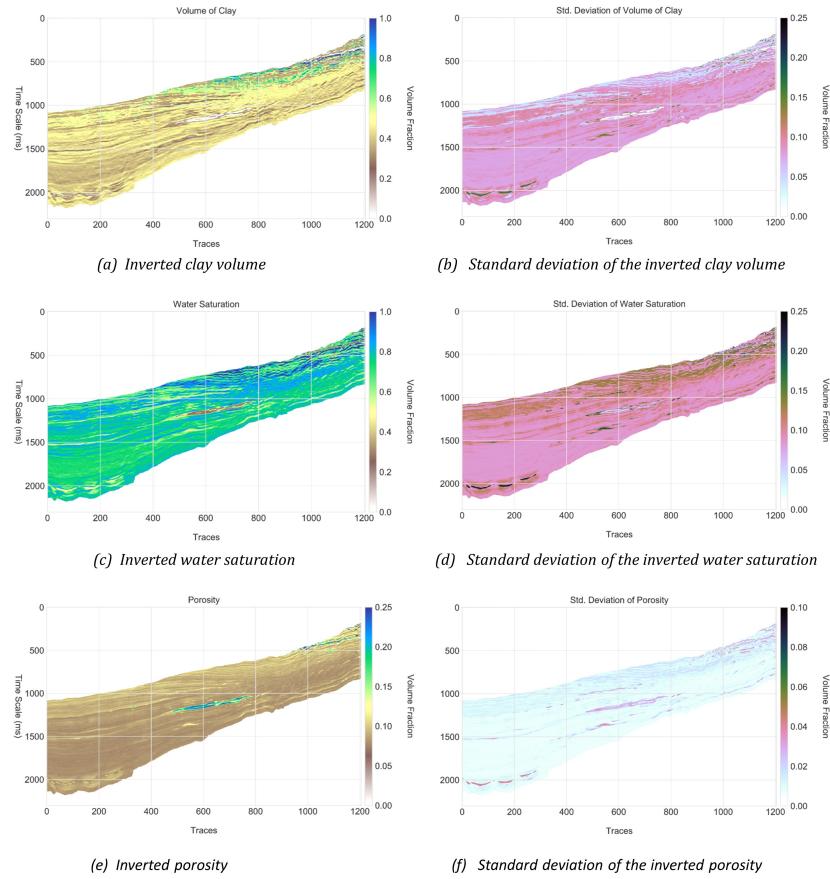


FIGURE 10 Cell-wise maps of petrophysical properties and their associated standard deviations (std.). (a) clay volume (V_{cl}) and (b) its std., (c) water saturation (S_w) and (d) its std., and (e) porosity (ϱ) and (f) its std. Yellow colour represents high values and dark blue colour represents low values of the respective properties.

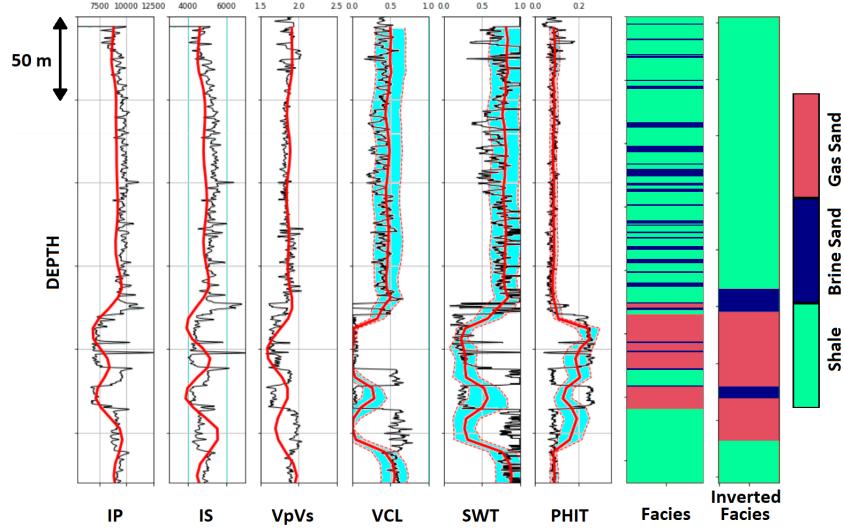


FIGURE 11 The inverted results compared with the well log data. Solid black lines show the measured well log data. Red lines in columns 1-3 are the input seismic attributes. Columns 4-6 show the mean (red lines) and two standard deviations (blue shaded area) of the inverted posterior distribution of petrophysical properties. The rightmost two columns show the measured facies and inverted facies, respectively.

of the posterior distribution of petrophysical properties in columns 4-6. The blue shaded regions bounded by the dashed-red curves in columns 3-4 represent the two standard deviations of the posterior distribution of corresponding rock properties. The mean inverted petrophysical properties clearly identify the gas reservoir characterized by lower V_{cl} and S_w , and higher ϱ compared to the non-reservoir rocks.

3.2 Travel time tomography

In this section we explore applications of variational inference methods to seismic travel time tomography based on examples in X. Zhang and Curtis (2020a) and X. Zhao et al. (2020). We image a simple 2D velocity structure that has been studied previously using Monte Carlo methods (Galetti et al., 2015). The velocity structure contains a circular low velocity anomaly with a 2 km radius and 1 km/s velocity within a homogeneous background of 2 km/s velocity (Figure 12a). 16 receivers are equally distributed around the low velocity anomaly approximating a circular acquisition geometry with a 4 km radius. Each receiver is also treated as a virtual source to simulate a typical ambient noise tomographic experiment (Curtis, Gerstoft, Sato, Snieder, & Wapenaar, 2006; Shapiro et al., 2005). Travel times between each receiver pair are calculated using the fast marching method

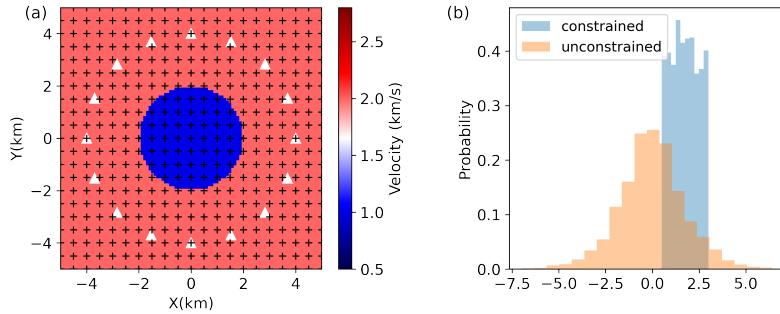


FIGURE 12 (a) The target structure and receiver geometry (while triangles). Each receiver also acts as a virtual source to simulate the scenario in ambient noise tomography (e.g., Shapiro et al., 2005). Black crosses denote the location of grid points used in inversion – the wave velocity at each location is described by one parameter. (b) The prior distribution in the original space (blue histogram) and the transformed space (orange histogram) – as described by 2000 samples. .

over a 101×101 gridded discretisation in space of each modeled velocity structure (Rawlinson & Sambridge, 2004), and the travel times through the target structure are used as data to infer the velocity structure.

For inversion we use a regular 21×21 grid of cells to parameterize the velocity structure (black pluses in Figure 12a). The likelihood function is set to a Gaussian distribution with 0.05 s standard deviation which represents the uncertainty on observed travel times. For each cell the prior pdf of the velocity is set to be a Uniform distribution between 0.5 km/s and 3 km/s (blue histogram in Figure 12b). To understand the characteristics of different methods we compare the posterior pdfs obtained using four methods: ADVI, Normalizing flows, SVGD and Metropolis-Hastings McMC (MH-McMC). In order to handle the hard constraints imposed by the prior information in variational methods, we transform the constrained velocity into an unconstrained space using equation 24. The orange histogram in Figure 12b shows the prior distribution in the transformed space. For all inversions, travel times are calculated using the fast marching method over a 41×41 grid interpolated from the lower spatial resolution properties. The gradients of the posterior pdf with respect to velocity are calculated by tracing rays backwards from each receiver to (virtual) sources using the spatial gradients of travel time fields.

In ADVI the initial Gaussian distribution in the unconstrained space is simply set to be a standard Gaussian distribution $N(\theta|\mathbf{0}, \mathbf{I})$, and updated using the ADAGRAD algorithm (Duchi, Hazan, & Singer, 2011) for 10,000 iterations using the gradients from equation 21 and 22. The final Gaussian distribution is transformed back to the original space, from which 5,000 samples are generated to visualize the final results.

For normalizing flows we use 6 coupling flows which each use rational quadratic splines (Durkan et al., 2019b) for the bijective function. Each bijective

function is parameterized by the output of a fully connected neural network, which contains 2 hidden layers each of which contains 100 hidden units with Rectified Linear Unit activation functions. The prior pdf is used as the initial distribution and is first transformed into the unconstrained space, and normalizing flows are applied in this space. The flows are updated using 3,000 iterations, and at each iteration the expectation in equation 31 is estimated using 10 samples. After the process we generate 2,000 samples from the initial (prior) distribution and transform them through the analytic flows (including the transform in equation 24) to obtain the final set of samples, whose density provides an approximation of the posterior pdf.

For SVGD we use a RBF kernel in which the scale factor σ is chosen to be $\tilde{d}/\sqrt{2\log n}$ where \tilde{d} is the median of pairwise distances between all particles. This choice is suggested by Q. Liu and Wang (2016) based on the intuition that $\sum_{j \neq i} k(\mathbf{m}_i, \mathbf{m}_j) \approx n \exp(-\frac{1}{h}\tilde{d}^2) = 1$, such that for particle \mathbf{m}_i the contribution from its own gradient is balanced by the influence from all other particles. We generate 800 particles from the prior distribution and first transform them into the unconstrained space. Those particles are then updated using equation 42 for 500 iterations and transformed back to the original space.

To demonstrate the convergence properties of these variational methods we compare the results with those obtained using the well-tested and robust method of MH-McMC (Metropolis & Ulam, 1949). Gaussian perturbations are used as the proposal distribution. We use a total of 6 chains, each of which contains 2,000,000 iterations with a burn-in period of 1,000,000 iterations. To reduce the correlation effects between successive samples we only retain every 50th sample after the burn-in period. This results in a total of 120,000 samples which are used to calculate statistics of the estimated posterior pdf.

3.2.1 Results

Figure 13 shows mean and standard deviation models obtained using the suite of methods. Overall the mean models obtained using different methods show similar features. For example, all models show a low velocity anomaly as in the target structure. The velocity of the mean (1.2 km/s) is slightly higher than the target value (1.0 km/s), but since this value was found by four independent methods this indicates that the mean value of the posterior pdf is genuinely lies at higher values than the target. Between the location of the receiver array and the low velocity anomaly there is a slightly lower velocity loop, and since the means from different methods show consistent features, the means probably reveal the true structure of the mean of the posterior distribution. The mean velocity structure does not necessarily need to be similar to the true velocity structure as it is the point-wise mean calculated from different samples. The circular shape of the mean velocity structure obtained from normalizing flows (Figure 13c) is less symmetric compared to those obtained using other methods. In normalizing flows a chain of non-linear transforms are optimized to directly

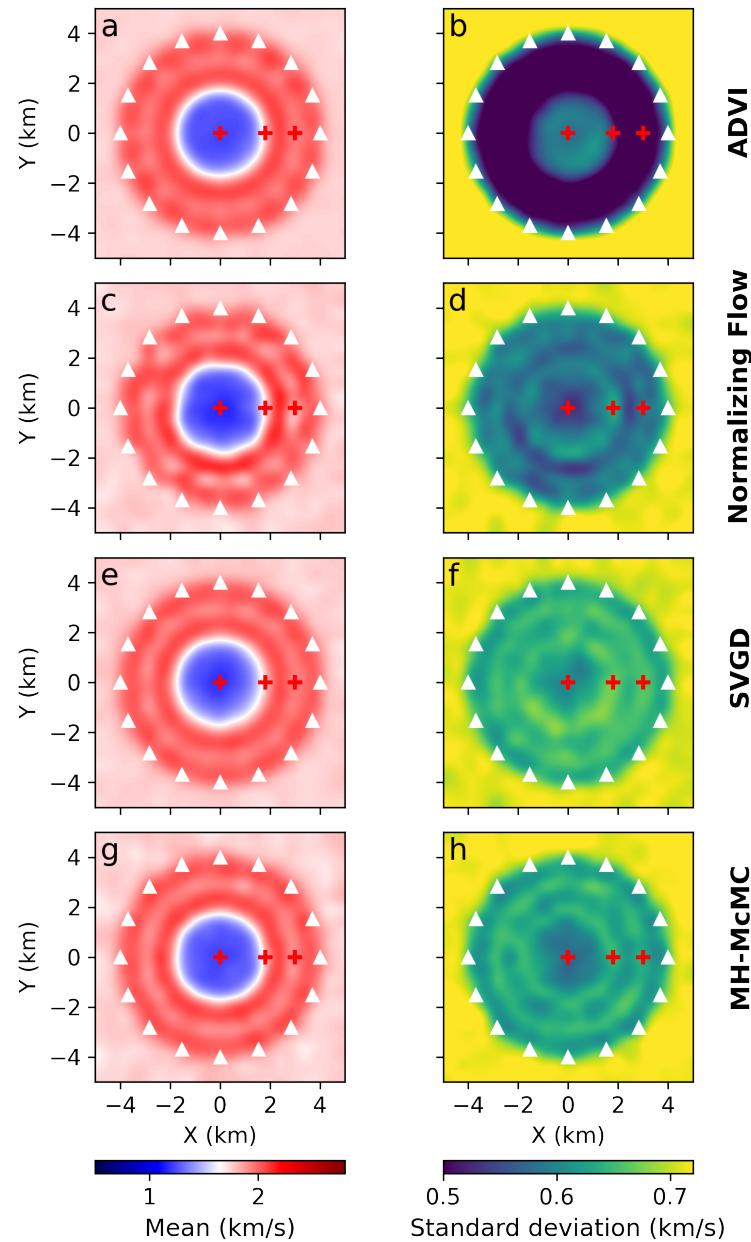


FIGURE 13 The mean (**left panel**) and standard deviation models (**right panel**) obtained using ADVI, Normalizing flows, SVGD and MH-McMC respectively. Red pluses are referred to in the main text and in Figure 14.

reshape an initial distribution towards the posterior distribution. It is highly likely that the high number of parameters in those transforms have non-unique solutions, some of which are not globally optimal. Converging to one of the latter solutions is likely to be the cause of the irregularity in the results.

The standard deviation models obtained from normalizing flows, SVGD and MH-McMC show very similar features (Figure 13d,f and h). For example, the middle low velocity anomaly has lower standard deviation suggesting that the low velocity anomaly is well constrained. There are two high uncertainty loops: one around the middle low velocity anomaly and the other one between the low velocity anomaly and the receiver array. The inner loop has also been observed in seismic tomographic results obtained using reversible jump McMC which is due to the uncertainty caused by the trade-off between the velocity of the anomaly and its shape (Galetti et al., 2015; X. Zhang et al., 2018). The latter high uncertainty loop is associated with the lower velocity loop in the mean velocity model. This is probably caused by the lower ray path coverage in this region, so that the mean velocity tends towards the mean of the prior (1.75 km/s) which is lower than the true value and the uncertainty is higher. In comparison the standard deviation from ADVI shows different results: higher uncertainty at the location of the middle low velocity anomaly and lower uncertainty between the low velocity anomaly and the receiver array (Figure 13b). Instead of the double high uncertainty loops exhibited by the other results, the standard deviation only shows a slightly higher uncertainty loop around the middle low velocity anomaly. This difference is probably caused by the fact that in ADVI we use a Gaussian distribution to approximate the posterior pdf, whereas in practice the posterior pdf often assumes non-Gaussian shapes due to the nonlinear relationship between velocity structure and data. Note that outside of the receiver array all standard deviations show high uncertainties because there is no ray coverage.

To further analyse the results in Figure 14 we show marginal distributions obtained using different methods at three locations (red pluses in Figure 13): point (0,0) km at the middle of the velocity structure, point (1.8,0) km and point (3.0,0) km which lie in the two high uncertainty loops. Due to symmetries of the system the marginal distributions at the three locations should reflect properties of most of the single-parameter marginal distributions. At point (0,0) km the marginal distributions are all very similar and show a distribution concentrated at one side of the prior distribution (Figure 14a, d, g and j). At point (1.8,0) km and (3.0,0) km the marginal distributions from normalizing flows (Figure 14e and f), SVGD (Figure 14h and i) and MH-McMC (Figure 14k and l) show similar features and are close to the prior distribution. This suggests that those regions are poorly constrained by the data and explains the double high uncertainty loops observed in the standard deviation structure. Note that the marginal distributions from SVGD and normalizing flows are less smooth than those obtained using MH-McMC. In SVGD this is caused by the lower number of samples used to approximate the distribution, whereas in normalizing flows it is due to the non-uniqueness of the variational optimization problem. In comparison the

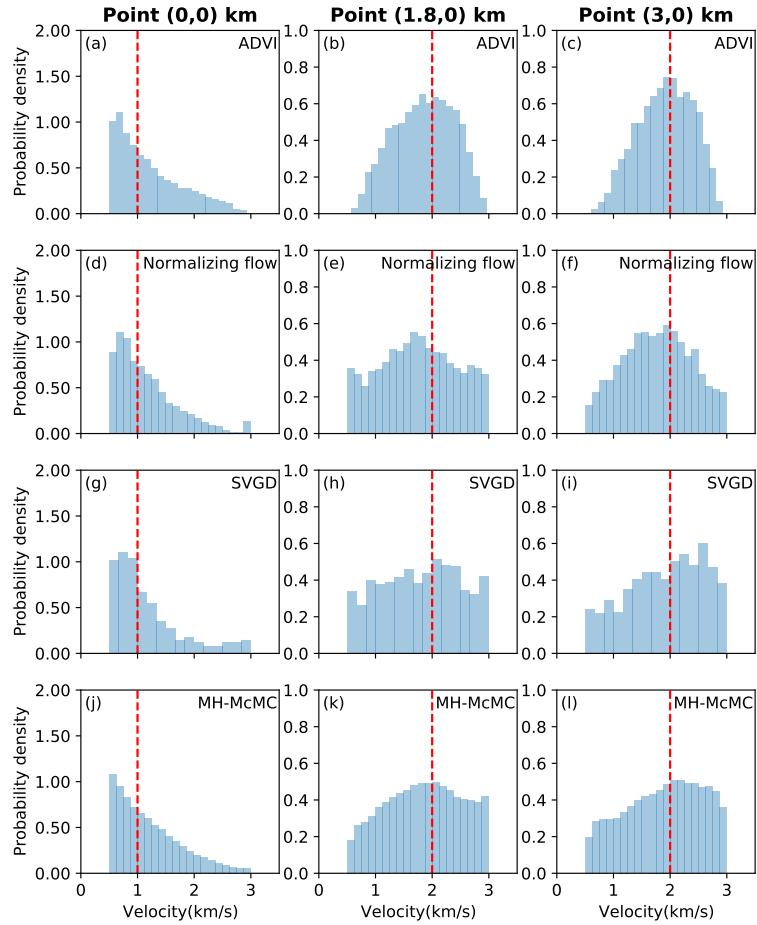


FIGURE 14 The marginal distributions at three locations: (0,0) km (**left panel**), (1.8,0) km (**middle panel**) and (3,0) km (**right panel**) obtained using ADVI, Normalizing flows, SVGD and MH-McMC respectively. Red line denotes the location of the true value.

TABLE 1.1 The comparison of computational cost for all 4 methods. The third column shows the equivalent number of simulations – these numbers are calculated as described in the main text.

Method	Number of simulations	Equivalent number of simulations
ADVI	10,000	10,870
Normalizing flows	30,000	32,609
SVGD	400,000	434,782
MH-McMC	12,000,000	12,000,000

marginal distributions at point (1.8, 0) km and (3.0,0) km obtained using ADVI show Gaussian-like distributions due to the implicit (transformed) Gaussian assumption which fails to describe the true uncertainty structure.

3.2.2 Computational cost

In Table 1.1 we summarize the number of forward simulations required by each method, which provides a good metric of the computational cost since for each method the forward simulation is the most time-consuming part. The results show that ADVI is the cheapest variational method, but we demonstrated above that it may provide biased results due to the implicit Gaussian assumption. Normalizing flows is slightly less efficient than ADVI, but produced significantly more accurate results above. SVGD requires approximately ten times more simulations than normalizing flows, but provides the most accurate results among the three variational methods. In comparison MH-McMC requires far more simulations than all three variational methods; however, the comparison is not fair in this case since MH-McMC only requires forward function evaluations, whereas the variational methods also require derivatives of the (logarithm of the) posterior pdf with respect to parameters (which in turn involves calculating derivatives of forward function with respect to parameters). In these travel time examples, derivatives were calculated using ray paths, which were traced through the travel time fields calculated by the fast marching method. For each forward simulation, calculating derivatives required a computation equivalent to approximately $f = 0.08$ forward simulations. A fairer comparison with the Monte Carlo method is therefore given in column 3 of Table 1 which shows the 'equivalent' number of simulations for each method, obtained by multiplying the number of simulations for the three variational methods by 1.08. In this case because of the efficient computation of derivatives, it does not increase the computation cost of variational methods significantly. Clearly this comparison will vary for different types of problems, since factor f will also vary. We demonstrate this below for waveform inversion problems for which f is approximately 2 (Q. Liu & Tromp, 2006; Tarantola, 1988).

Note that the above comparison is only valid for this specific example and does not necessarily provide general guidance for the practical choice of algorithms. For example, although ADVI provides biased results, it can still be useful

for weakly nonlinear problems in scenarios where efficiency is important and a Gaussian distribution is sufficient for uncertainty analysis. For very high dimensional problems such as 3D tomography and full-waveform inversion, ADVI can become inefficient as the full covariance matrix may require extremely large memory. In the above example, normalizing flows would be a good choice given that it produces reasonably accurate results yet requires the same order of computational cost as ADVI. However we note that normalizing flows may require more human interaction as it has many hyperparameters to tune – which specific flow to use, how many flows to use, and if invertible neural networks are used then the structure of the neural network needs elaborate design. For very high dimensional problems we may require large neural networks, so the training time cannot be neglected and may even dominate the whole calculation. SVGD solves variational inference problem using a set of samples, which provides a flexible way to approximate complex probability distributions but at the price of an increased number of forward function evaluations. The method is fully parallelizable which makes it more efficient in real time when combined with modern parallel computer architecture. However it remains unclear how the method performs in very high dimensional space, as it might be impossible to use hundreds of samples to approximate the posterior pdf meaningfully in high dimensional problems.

In this example we only compared the computational cost of variational methods with MH-McMC. In practice there are many ways to make Monte Carlo methods more efficient, for example reversible-jump McMC (Bodin & Sambridge, 2009; Green, 1995; Malinverno, 2002), Hamiltonian Monte Carlo (Duane et al., 1987; Fichtner, Zunino, & Gebraad, 2018; Neal et al., 2011), Langevin Monte Carlo (Girolami & Calderhead, 2011; Roberts et al., 1996), Sequential Monte Carlo (J. S. Liu & Chen, 1998; Smith, 2013), slice sampling (Neal, 2003), physics informed Monte Carlo (Khoshholgh, Zunino, & Mosegaard, 2020) and parallel tempering (Earl & Deem, 2005; Hukushima & Nemoto, 1996; Sambridge, 2013). Nevertheless, Monte Carlo methods cannot be parallelized within a Markov chain, several of these Monte Carlo methods require calculation of gradients of the forward function which introduces an additional factor f to the cost as described above, and the methods often become intractable for large datasets which are usually expensive to simulate. In contrast, variational methods can be parallelized at the sample level in each iteration – for example gradient calculation in ADVI, normalizing flows and SVGD can be fully parallelized. In addition variational methods can be applied to large datasets by using stochastic optimization (Kubrusly & Gravier, 1973; Robbins & Monro, 1951) and distributed optimization, which is likely to make variational methods more efficient in practice for some types of problems.

In travel time tomography the gradients of posterior pdf with respect to model parameters can be calculated efficiently using the travel time field obtained in the forward simulation. In the case that gradients are difficult to calculate, MH-McMC may be more efficient than both variational methods and many other

Monte Carlo methods since MH-McMC does not require gradient information. We also note that our comparison above depends on subjective assessments of the point of convergence of each method, so the absolute number of simulations required by each method may not be accurate. Nevertheless they at least provide a reasonable insight into the computational efficiency of each method.

3.3 Full waveform inversion

Full waveform inversion (FWI) uses filtered versions of full seismic recordings to characterize properties of the subsurface, and can produce high resolution images of the Earth's interior (Gauthier, Virieux, & Tarantola, 1986; Pratt, 1999; Tarantola, 1984, 1988; Tromp, Tape, & Liu, 2005). The method has been used at industrial scale (Prieux, Brossier, Operto, & Virieux, 2013; Warner et al., 2013), regional scale (P. Chen, Zhao, & Jordan, 2007; Fichtner, Kennett, Igel, & Bunge, 2009; Tape, Liu, Maggi, & Tromp, 2009) and global scale (Bozdağ et al., 2016; Fichtner, van Herwaarden, et al., 2018; French & Romanowicz, 2014). Due to the high nonlinearity and nonuniqueness of the problem, in traditional optimization-based methods a good starting model is required to avoid converging to incorrect solutions. A variety of misfit functions that can reduce multimodalities in the posterior pdf have also been proposed (Bozdağ, Trampert, & Tromp, 2011; Brossier, Operto, & Virieux, 2010; Fichtner, Kennett, Igel, & Bunge, 2008; Gee & Jordan, 1992; Luo & Schuster, 1991; Métivier, Brossier, Mérigot, Oudet, & Virieux, 2016; Van Leeuwen & Mulder, 2010; Warner & Guasch, 2016). In addition, to quantify uncertainties in the solution Monte Carlo methods have recently been used to solve FWI problems (Biswas & Sen, 2017; Gebraad et al., 2020; Guo et al., 2020; Ray et al., 2017, 2016; Z. Zhao & Sen, 2019). We now use variational inference methods, specifically SVGD to solve FWI problems probabilistically, which we refer to as variational full waveform inversion or VFWI, based on examples in X. Zhang and Curtis (2020b, 2021).

3.3.1 Transmission seismic FWI with strong prior information

We first apply SVGD to a transmission FWI problem in which seismic data are recorded on a receiver array that lies above the structure to be imaged given earthquake-like sources located underneath the structure. We use a 2D fully elastic target structure and data acquisition setup that is identical to that used by Gebraad et al. (2020) such that the results obtained by SVGD can be fairly compared to those that Gebraad et al. (2020) obtained using Hamiltonian Monte Carlo (HMC). Figure 15 shows the target V_p, V_s and density model. 7 sources with random moment tensors are located at the bottom of the region. Similarly to Gebraad et al. (2020) we use a Ricker wavelet source-time function with a dominant frequency of 50 Hz. 19 receivers are located at the depth of 10 m with a regular spacing of 12.5 m. The model is discretised using a regular 200×100 grid of cells, within which a 180×60 sub-grid of cells have free parameters (black dashed box in Figure 15). This leads to a total of $180 \times 60 \times 3 = 32,400$

free parameters. The waveform data are modelled using a fourth-order variant of the staggered-grid finite difference scheme (Gebraad et al., 2020; Virieux, 1986). The gradients of the likelihood function with respect to velocities and density are computed using the adjoint method (Fichtner, Bunge, & Igel, 2006; Q. Liu & Tromp, 2006; Plessix, 2006; Tarantola, 1988).

To reduce the complexity of the inverse problem and guide both methods towards to correct solution we use a strong prior information as in Gebraad et al. (2020): Uniform distributions in the interval of 2000 ± 100 m/s for Vp, 800 ± 50 m/s for Vs and 1500 ± 100 kg/m³ for density. For the likelihood function, we assume a Gaussian distribution with a diagonal covariance matrix:

$$p(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto \exp\left[-\frac{1}{2} \sum_i \left(\frac{d_i^{\text{obs}} - d_i(\mathbf{m})}{\sigma_i}\right)^2\right] \quad (52)$$

where i is the index of time samples and σ_i is the standard deviation of that data point. To keep the inverse problem identical to that in Gebraad et al. (2020), we set σ_i to be $1 \mu\text{m}^2$ and did not add any noise to the waveform data. The effects of different values of σ_i on the solution are analysed in Gebraad et al. (2020).

Similarly to the previous section we use a RBF kernel for SVGD whose scale factor is determined from the median of pairwise distances between all particles. We generated 600 particles from the prior distribution and transformed them into an unconstrained space using equation 24. Those particles are then updated using equation 42 for 600 iterations, and are finally transformed back to the original space.

Figure 16 shows the mean and standard deviation structures obtained using SVGD. The mean Vs model shows similar features to the true velocity structure, for example the bottom high velocity structure and tilted layers above that structure. The horizontal layers at the shallow part (< 80 m) are not as clearly observable as those in the true velocity structure, which probably reflects the limits of the resolution of the data. By contrast, the mean Vp model only recovers the bottom large scale structure. This is probably because when a simple unweighted L2 norm misfit function is used, seismic waveforms are more sensitive to Vs than to Vp due to the higher amplitudes of shear waves. Figure 17 shows kernels (gradients of the misfit function) of Vp, Vs and density calculated using the mean models in Figure 16. The magnitude of Vs and density kernels are significantly higher than that of Vp. As a result, the Vp structure is not well constrained by the data. The mean density model clearly shows horizontal and tilted layers except that the value of the lower density tilted layers is smaller than the true value. In comparison the bottom high density structure is not present in the mean model which is probably because seismic waveforms are mainly sensitive to spatial gradients of density.

Overall the standard deviation models show similar features to their associated mean structure. For example, the standard deviation model of Vp shows lower uncertainty at the location of the large scale high velocity structure. The

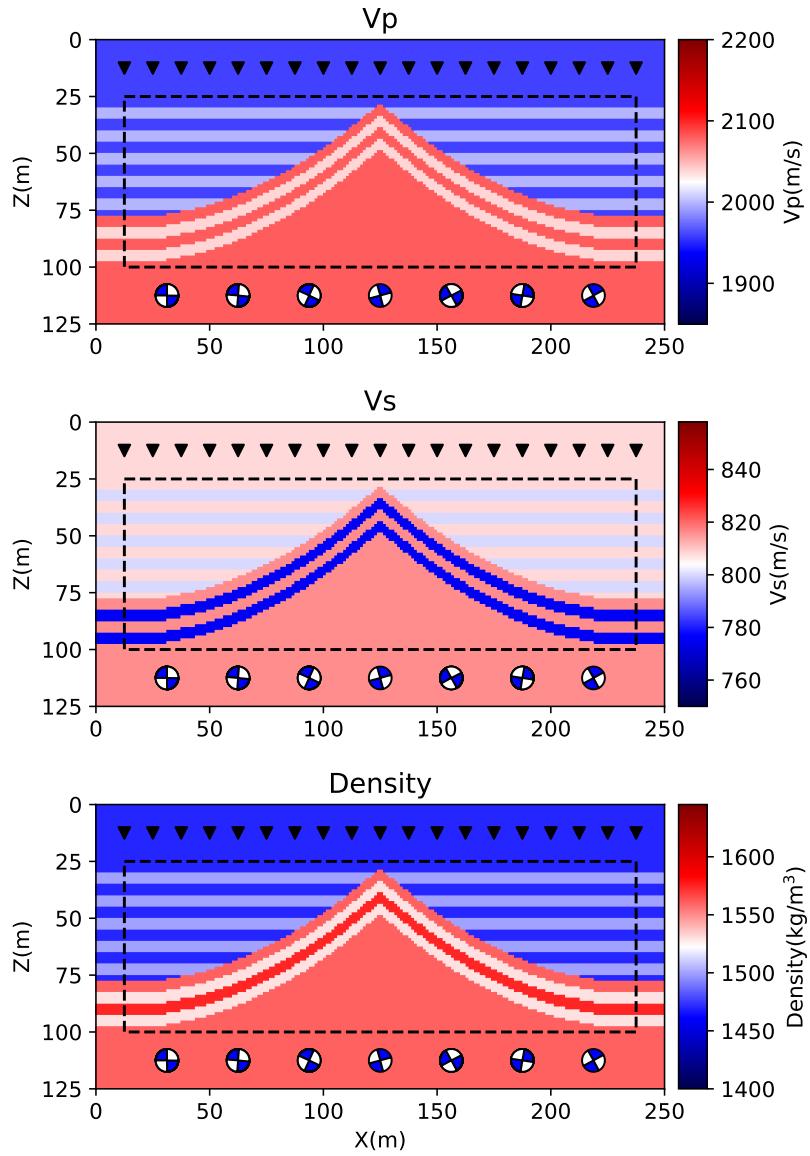


FIGURE 15 The target structure of Vp , Vs and density. Sources are located at the bottom of the model with random moment tensors and receivers are located at the near surface (black triangles). The black dashed line indicate the area that has free parameters. This inverse problem setup is identical to that in Gebraad et al. (2020).

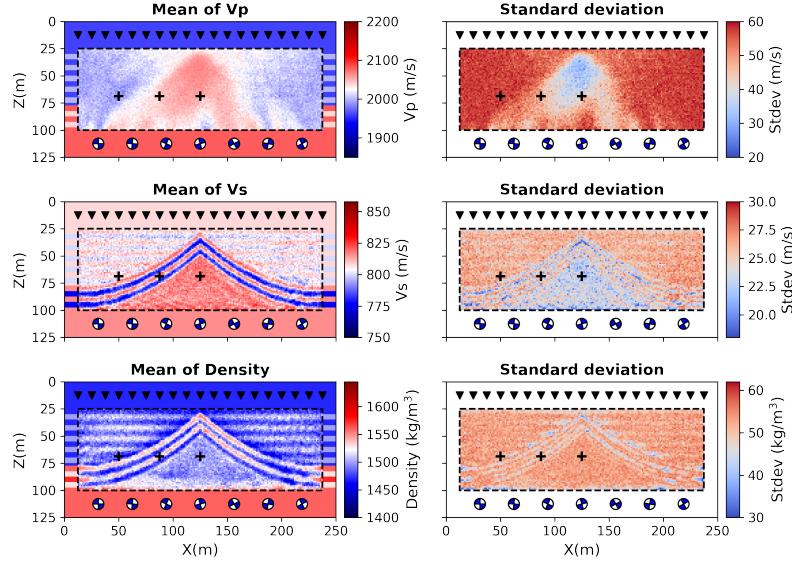


FIGURE 16 The mean and standard deviation of V_p , V_s and density obtained using SVGD. Black crosses are referred in the main text.

V_s standard deviation model shows lower uncertainties at the location of the horizontal high velocity layers and the bottom high velocity structure. There are high uncertainties at the boundaries of tilted layers, which suggests that the location of velocity layers are not well-constrained. Note that a similar phenomenon observed in the travel time tomography examples in the previous section. Similarly there are high uncertainties at those boundaries in the standard deviation model of density. Due to the fact that seismic waveforms are mainly sensitive to density spatial gradients, the bottom high density structure has high uncertainty.

To explore the effects that the number of particles have on the results, in Figure 18 we show the mean and standard deviation models of V_s obtained using 400 particles and 600 particles respectively. As expected, the results show that when using 600 particle, we can obtain more accurate results. For example, the mean V_s model obtained using 400 particles only shows the bottom high velocity structure and the tilted layers. The shallow horizontal layers are smeared into each other. Similarly the standard deviation model does not show much structure in the shallow part compared with that obtained using 600 particles. This shows that the accuracy of the results of SVGD improves with the number of particles (Q. Liu, 2017).

To validate the results obtained using SVGD, we compared the results with those obtained using HMC by Gebraad et al. (2020) (Figure 19). The mean and standard deviation structures obtained using HMC are very similar to those

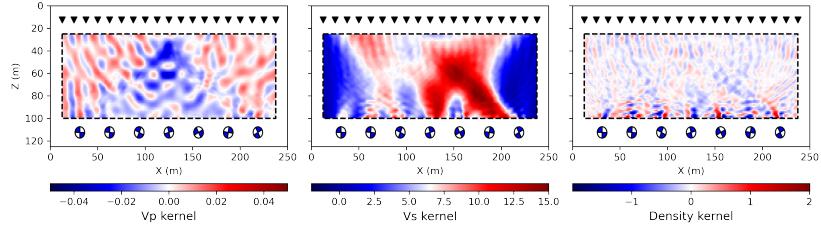


FIGURE 17 The kernel (gradients of the misfit function) of Vp, Vs and density calculated using mean structures in Figure 16. The magnitude of Vs and density kernels are significantly higher than that of Vp kernel.

obtained using SVGD. For example, the mean Vp only shows the bottom large scale structure whereas the mean Vs successfully recovers the true structure. The mean density shows the horizontal and tilted layers and fails to find the bottom high density structure. The standard deviations also show similar features to associated mean structures. Since the two methods are completely different, it is highly likely that these results represent the true solution to this specific FWI problem. Note that the results from SVGD are smoother than those from HMC, which is probably caused by undersampling of both methods and lack of convergence of HMC (Gebraad et al., 2020).

To further analyse the results, in Figure 20 we show marginal distributions of Vp, Vs and density obtained using SVGD at three points (black crosses in Figure 19): (50, 68.75) m, (87.5, 68.75) m and (125, 68.75) m. Overall the results show high probability around the true value. At X=50 m the marginal distributions are wider than those at the other locations, which indicates high uncertainties at this location. At X=125 m the true value of density deviates from the values with highest probability as we have observed in the mean model due to the fact that seismic waveforms are mainly sensitive to density spatial gradients. Note that the marginal distributions show nonsmoothness due to the undersampling of the posterior pdf (a small number of particles).

Since SVGD is based on particles, the method can be computationally expensive. For example, the above example requires $600 \times 600 = 360,000$ forward and adjoint simulations; whereas HMC took approximately 130,000 forward and adjoint simulations. Although in this case it appears that HMC is slightly more efficient, in the above example HMC has clearly not fully converged. While SVGD can be easily parallelized, it is difficult to parallel a Markov chain due to the dependence between successive samples (Neiswanger, Wang, & Xing, 2013). Also in practice HMC often requires deliberate and tedious tuning to construct an efficient Markov chain (see discussions in Gebraad et al., 2020) so the actual computational cost may be significantly higher than the number of samples reported above. In contrast SVGD is much easier to tune by using adaptive gradient ascent methods (Duchi et al., 2011; Q. Liu & Wang, 2016). In

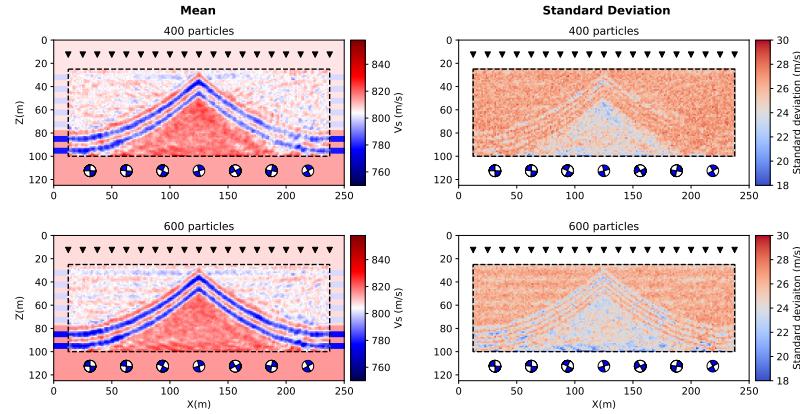


FIGURE 18 The mean (left) and standard deviation (right) of Vs obtained using SVGD with 400 and 600 particles respectively.

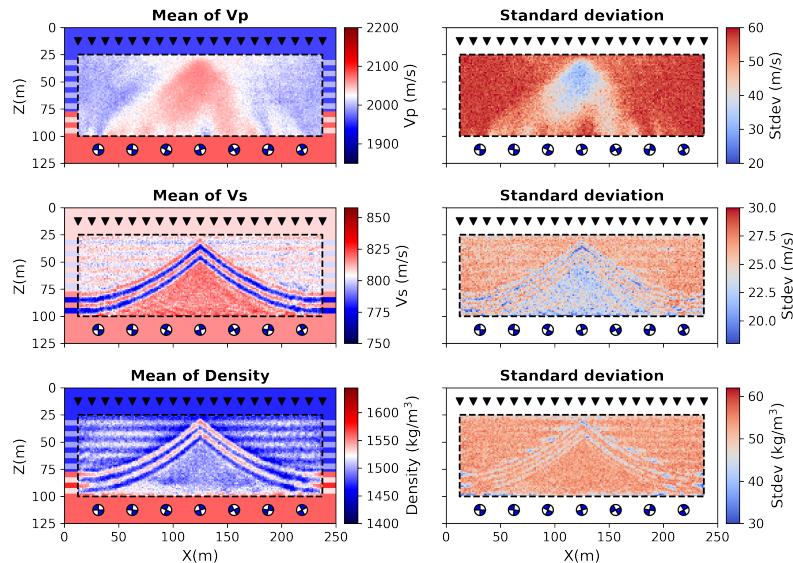


FIGURE 19 The mean and standard deviation of Vp, Vs and density obtained using HMC from Gebraad et al. (2020).

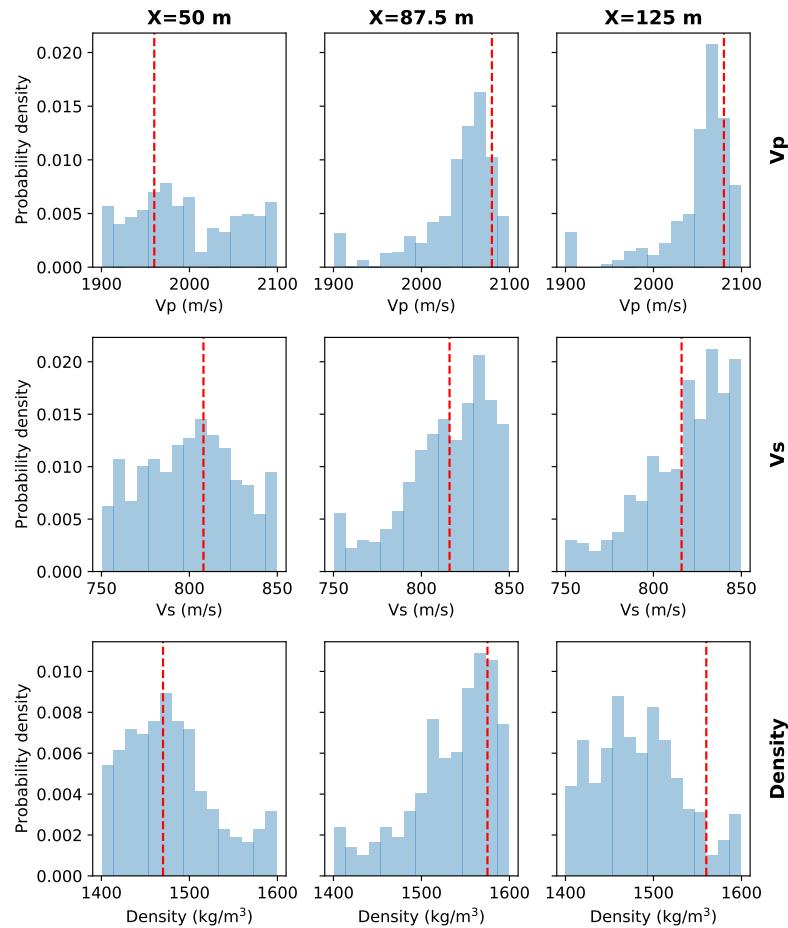


FIGURE 20 The marginal distributions of V_p , V_s and density obtained using SVGD at the depth of 68.75 m and at $X=50$ m, 87.5 m and 125 m respectively. Red lines denote the true values.

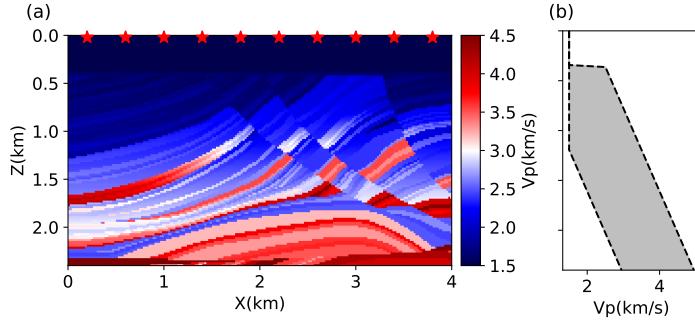


FIGURE 21 (a) Part of the Marmousi model (G. S. Martin et al., 2006) which is used as the target velocity structure. 10 sources (stars) are located at the depth of 20 m and 200 receivers (not shown) are equally spaced at a depth of 360 m across the horizontal extent of the model (this depth represents the seabed). (b) Prior distribution used in the inversion: a Uniform distribution with an width of 2 km/s at each depth. Note that an extra lower bound is also imposed to the velocity to ensure that the rock velocity is higher than the velocity in water (1.5 km/s).

addition SVGD can be performed on large datasets by using stochastic optimization by dividing large datasets into minibatches (Q. Liu & Wang, 2016). The same technique cannot be used in McMC methods because it breaks the detailed balance required by McMC. To give an idea about the overall computational cost required by SVGD, the above example took 6 days of computation parallelized across 16 Intel Xeon cores.

3.3.2 Reflection seismic FWI with realistic prior information

In the previous section we applied SVGD to a transmission FWI problem with known, double-couple (earthquake-like) sources and strong prior information on parameters. Unfortunately such strong prior information about sources and parameters is never available in practice. To explore the applicability of the method in practice, in this section we apply SVGD to seismic reflection data generated by known near-surface sources with more practically realistic prior information.

We solve a 2D acoustic FWI problem using the waveform data generated from a part of the Marmousi model (G. S. Martin et al., 2006). The model is discretised in space using a 200×100 regular grid of cells. 10 sources are located at 20 m depth and 200 receivers are located at the 360 m depth (which represents the seabed) across the full horizontal extent of the model with a regular spacing of 20 m (Figure 21). Similarly to the previous section, the waveform data are generated using the finite difference method and the gradients of the posterior pdf with respect to velocity parameters are computed using the adjoint method.

Instead of using strong prior information (a Uniform distribution over an interval of 0.2 km/s) as in the previous section, we impose ten times weaker prior information to the velocity: a Uniform distribution over an interval width

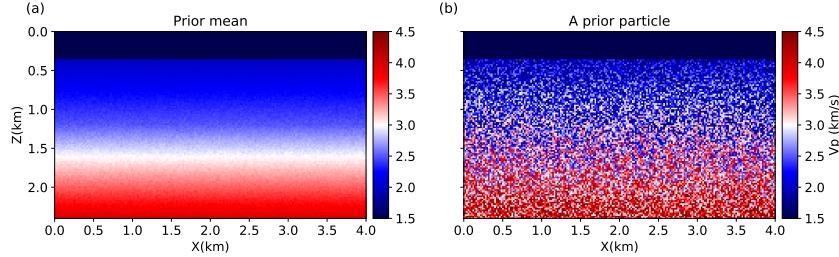


FIGURE 22 (a) The mean of the prior distribution and (b) a random sample generated from the prior pdf.

of 2 km/s at each depth (Figure 21b). We also impose a lower bound on the velocity to ensure that the rock velocity is higher than that in the water (1.5 km/s). The velocity of the water layer is fixed to be the true velocity (1.5 km/s) in the inversion as is standard in practical marine seismic FWI. Figure 22 shows the mean of the prior distribution and a random particle generated from the prior distribution. We simulate waveform data using a Ricker wavelet with a dominant frequency of 10 Hz. Uncorrelated Gaussian noise with a standard deviation of 0.1 amplitude units is added to the data. For the likelihood function we use the same Gaussian distribution as described in equation 52 where σ_i is set to be the true value. For standard optimisation-based FWI this problem is difficult because the reference parameter values from which the inversion begins (which in practice would normally be the mean structure) is very different from the target.

X. Zhang and Curtis (2021) showed that one can improve accuracy of the inversion results by performing an inversion using low frequency data first, and using the results of the low frequency inversion as the starting distribution for high frequency inversions. Therefore, we first perform SVGD on low frequency data generated by a Ricker wavelet with a dominant frequency of 4 Hz with the same Gaussian noise as above added to the data (a standard deviation of 0.1). The inversion is conducted using 600 particles that are initially generated from the prior distribution (e.g., Figure 22b) and the matrix kernel described in equation 44 where $\mathbf{Q}^{-1} = \text{diag}(\text{var}(\mathbf{m}))$ and $\text{var}(\mathbf{m})$ is the variance computed across those particles. For parameters with higher variance this kernel applies higher weights to the posterior gradients, and also enables more distant interactions with other particles. As in the previous section we first transform those particles to an unconstrained space using equation 24 and update them using equation 42 for 600 iterations. Those particles are then used as the starting particles for the high frequency inversion and are updated for another 300 iterations. The mean and standard deviation are calculated after transforming those particles back to the original parameter space.

Figure 23 shows the mean and standard deviation structures obtained using

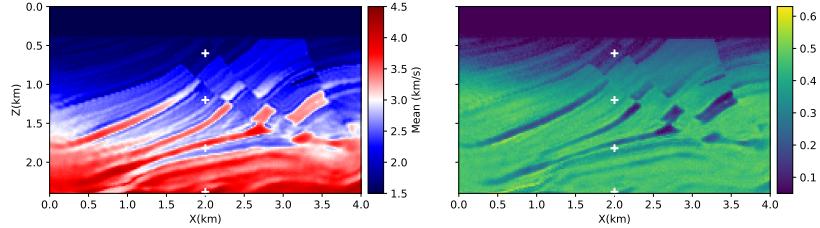


FIGURE 23 The mean structure and its point-wise standard deviation obtained using SVGD given high frequency data and using particles from low frequency inversion as the starting distribution.

the above strategy. A comparison of the results to those obtained using only low frequency data and only high frequency data is discussed in X. Zhang and Curtis (2021). Overall the mean exhibits similar features to the target structure, except that the deeper part (> 2 km) is slightly different from the target structure because of the poor illumination. The standard deviation has qualitatively similar features to the mean as we observed in the previous section. For example, in the near surface (< 1 km) the low velocity anomalies are associated with lower uncertainty, and in the deeper part (> 1 km) there are lower uncertainties at the location of high velocity anomalies. This phenomenon probably reveals the fact that waves spend comparatively longer in low velocity area which results in higher sensitivity. Note that due to the stronger prior information and better data coverage, the shallower part (< 1 km) has lower uncertainty compared to the deeper part.

To further analyse the results, in Figure 24 we show marginal distributions at four locations (white pluses in Figure 23): (2.0, 0.6) km, (2.0, 1.2) km, (2.0, 1.8) km and (2.0, 2.4) km. Overall the true velocity values are around the high probability area, except that at the depth of 2.4 km the true value slightly deviates from the value with highest probability because of the poor illumination. At the deeper locations (1.8 and 2.4 km) the marginal distributions show complex, multimodal distributions which reflects the complexity of this inverse problem.

The above inversion took about 10,055 CPU hours for the total 900 iterations and required approximately 111.7 hours to run on 90 Intel Xeon CPU cores. In practice for larger datasets the method can be implemented using stochastic minibatch optimization. In addition, since the method does not require strong prior information, it could also be used to provide a good starting model for standard linearised FWI by using a small part of a large dataset. In addition, one may be able to perform the method on data types that require lower computational cost first, e.g. travel time tomography, and use those results as the starting distribution for VFWI to improve efficiency.

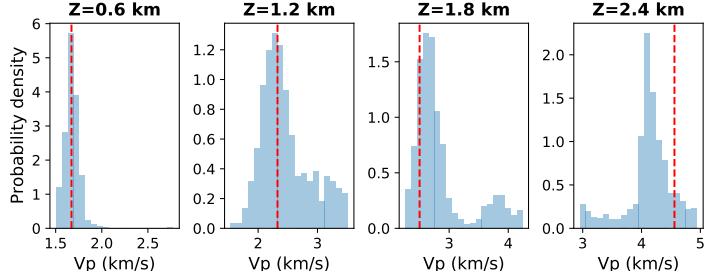


FIGURE 24 The marginal distributions at the horizontal location of 2 km and the depth of 0.6 km, 1.2 km, 1.8 km and 2.4 km obtained using SVGD given high frequency data and using particles from low frequency inversion as the starting distribution. Red lines denotes the true values.

4 DISCUSSION

In this study prior probabilities are simply set as Uniform distributions. While Uniform prior probabilities are simple to impose and are useful to explore properties of different methods, they may cause complex posterior pdfs which are hard to explore. In practice in cases where we have more knowledge about the subsurface, a more informative prior distribution should ideally be used. For example some prior regularization can be used to produce smoother models (MacKay, 2003), or a Gaussian process may be used to inject prior information with adaptable complexity into inference scheme (Ray & Myer, 2019). Neural networks can also be used to encode geological information into prior distributions (Laloy, Héroult, Lee, Jacques, & Linde, 2017; Mosser, Dubrule, & Blunt, 2020). For likelihood functions we simply used Gaussian distributions with a known, fixed data noise level. In practice this noise level might be estimated from data using the maximum likelihood method (Ray et al., 2016; Sambridge, 2013) or a variety of other methods (Bensen, Ritzwoller, & Yang, 2009; Nicolson, Curtis, & Baptie, 2014; Nicolson, Curtis, Baptie, & Galetti, 2012; Weaver, Hadzioannou, Larose, & Campillo, 2011; Yao & Van Der Hilst, 2009). It may also be possible to estimate the noise level in variational methods using a hierarchical Bayesian formulation (Ranganath, Tran, & Blei, 2016). To further improve the results a non-Gaussian likelihood function might also be used at little or no additional cost to the method.

In above examples we used a fixed regular grid of cells to parameterize the subsurface which can cause overfitting or underfitting of the data. For instance, in the travel time tomography example we observed a lower velocity loop with high uncertainty between the middle velocity anomaly and the receiver array (Figure 13), which may be caused by overfitting as there is no such structure in the true model. To resolve this issue, an optimal grid might be sought. This can be achieved by applying a series of different grids and selecting the best one based on Bayesian or other model selection theories (Arnold & Curtis, 2018;

Curtis & Snieder, 1997; Walter & Pronzato, 1997). For example, the ELBO calculated implicitly in variational methods can be used as a model selection criterion (Bernardo et al., 2003; McGrory & Titterington, 2007; Sato, 2001). However, we note that the statistical theory behind such a design criterion is currently under explored, especially compared to McMC methods: in McMC a variety of well-established methods are available to perform model selection, for example reversible-jump McMC (Green, 1995), sequential Monte Carlo (Smith, 2013) and nested sampling (Feroz & Hobson, 2008; Skilling, 2004). Further research is required to develop appropriate model selection in variational inference. Apart from regularly gridded cells, we note that other more advanced parameterizations can be used in variational methods to provide more flexibility, such as Voronoi cells (Bodin & Sambridge, 2009; X. Zhang et al., 2018), wavelet parameterization (Fang & Zhang, 2014; Hawkins & Sambridge, 2015; X. Zhang & Zhang, 2015), Johnson-Mehl tessellation (Belhadj, Romary, Gesret, Noble, & Figliuzzi, 2018) and Delaunay and Clough-Tocher parameterizations (Curtis & Snieder, 1997; Hawkins, Bodin, Sambridge, Choblet, & Husson, 2019).

While we focused on variational inference using KL-divergence to measure difference between two distributions, it is also possible to use other measures of divergence. For example, Minka (2013) proposed the expectation propagation method by using KL-divergence in the other direction, that is $\text{KL}[p||q]$ rather than $\text{KL}[q||p]$. Other more general divergences, such as α -divergence (Amari, 1985) and f -divergence (Ali & Silvey, 1966) have also been used employed within variational inference (Bamler, Zhang, Opper, & Mandt, 2017; Hernandez-Lobato et al., 2016; Li & Turner, 2016; Wang, Liu, & Liu, 2018). Stein's discrepancy provides another measure of difference between two distributions (Gorham & Mackey, 2015; Q. Liu, Lee, & Jordan, 2016; Stein et al., 1972) and can also be used to develop variational methods (Y. Liu, Ramachandran, Liu, & Peng, 2017; Ranganath, Altosaar, Tran, & Blei, 2016).

Since the ELBO is a nonconvex objective function, variational inference can converge to a local optimum. For instance, in our travel time tomography example the results obtained using normalizing flows show irregularities and non-smoothness, which likely reflects convergence to a local optimum. To reduce this issue, more advanced optimization methods can be used, for example variational tempering (Mandt, McInerney, Abrol, Ranganath, & Blei, 2016), the trust-region method (Theis & Hoffman, 2015) or population variational inference (Kucukelbir & Blei, 2014). Different variational methods may also be combined together to increase robustness. For example, the probability distribution obtained using ADVI can be used as a starting distribution for normalizing flows and SVGD.

Monte Carlo sampling methods and variational inference are different methods that can be used to solve similar problems. Monte Carlo methods are usually applied using Markov chains, which generate a chain of samples that approximately follow the posterior pdf; variational inference seeks an optimal approximation to the posterior pdf within a predefined family of probability

distributions. Monte Carlo methods are well-understood and are guaranteed to converge to the true posterior pdf asymptotically as the number of samples tends to infinity (Robert & Casella, 2013), whereas the theoretical aspects of accuracy and convergence of variational inference are still unknown. The two methods can be used together to combine the merits of both. For example, a variational approximation can be used to build proposal distributions for Metropolis-Hastings algorithms to improve their efficiency (De Freitas, Højen-Sørensen, Jordan, & Russell, 2001), or McMC steps can be incorporated into variational inference to improve accuracy (Salimans, Kingma, & Welling, 2015). Further research on the interface between the two methods is certainly an interesting topic.

We have applied variational inference methods to petrophysical inversion, 2D travel time tomography and 2D FWI, and demonstrated their efficiency in solving these problems. However, it remains a challenge to apply variational methods to very high dimensional inverse problems, e.g. 3D FWI. In such cases the forward modelling itself is usually computationally extremely expensive. For methods like normalizing flows we may end up with very large neural networks, which can occupy huge memory and become very difficult to train. For SVGD we are likely to need many more particles than used herein, which may demand more resources than one can afford. In addition kernel metrics used in SVGD may become inefficient in high dimensional space due to the curse of dimensionality (Wainwright, 2019). Therefore further work is required to explore the properties of variational methods in a range of high dimensional, practical applications.

5 CONCLUSION

In this chapter we reviewed the basic concepts of variational inference, and discussed four specific methods: mean-field approximation, automatic differential variational inference (ADVI), normalizing flows and Stein variational gradient descent (SVGD). Mean-field approximations can provide very efficient methods, but they assumes mutually independent parameters. ADVI uses a Gaussian distribution to approximate the posterior distribution, again leading to a reasonably efficient method but results that may be biased. Both normalizing flows and SVGD use a series of invertible transforms to transform an initial distribution to an approximation to the posterior distribution. Normalizing flows use a series of analytical invertible transforms, whereas SVGD uses an implicit transform to rearrange a set of particles from an initial distribution to represent the posterior distribution. We reviewed previous applications of the methods to a range of different examples: petrophysical inversion, travel time tomography and full-waveform inversion (FWI). In travel time tomography example we compared the results from ADVI, normalizing flows and SVGD with those obtained using Monte Carlo methods. The results show that ADVI is the cheapest method but provides biased results due to the implicit Gaussian assumption. In comparison, normalizing flows and SVGD can provide more accurate approximations to the results from the Monte Carlo method. Normalizing flows further improved

efficiency of the inversion compared with SVGD. To further demonstrate variational methods, we applied SVGD to full-waveform inversion (FWI) problems and demonstrated that SVGD can produce accurate results to FWI problems, similar to those from Monte Carlo where the comparison has been made. We conclude that variational inference is an efficient and valuable tool to solve Geophysical inverse problems. We also note that variational inference is still in a phase of rapid development, for example, to solve the variational optimization problem more efficiently and to make the method more feasible to large scale inverse problems, so the method may become more accurate and more efficient in the near future.

Glossary

forward function	a function that predicts data for any particular values of model parameters
inversion	the process that infers the value of model parameters from measurements or observations
prior pdf	a probability density function of model parameters which describes information that is independent of the data
likelihood function	a probability density function that defines the probability of observing certain data given a specific set of model parameters
posterior pdf	a probability density function which describes the uncertainty of model parameters by combining the prior information and the information from the data the probability distribution of observed data marginalized over the model parameters
evidence	a lower bound for the evidence
ELBO	a method that uses Bayes' theorem to infer the posterior probability distribution of model parameters given the observed data
Bayesian inference	a method that uses optimization to solve Bayesian inference problem
variational inference	the Kullback-Leibler divergence is a measure of difference between two probability distributions
KL-divergence	a family of probability density functions from which one seeks an optimal approximation to the posterior probability density function
variational family	probability density functions that assume mutually independent parameters
mean field approximation	automatic differential variational inference, a method that seeks an optimal Gaussian distribution to approximate the posterior probability distribution
ADVI	an invertible transform which transforms an initial distribution to a target distribution
normalizing flow	Stein variational gradient descent, a method that optimizes a set of model samples to approximate the posterior probability distribution
SVGD	

References

- Aki, K., & Lee, W. (1976). Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. a homogeneous initial model. *Journal of Geophysical research*, 81(23), 4381–4399.
- Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1), 131–142.
- Amari, S.-i. (1985). α -divergence and α -projection in statistical manifold. In *Differential-geometrical methods in statistics* (pp. 66–103). Springer.
- Andersen, K. E., Brooks, S. P., & Hansen, M. B. (2001). A bayesian approach to crack detection in electrically conducting media. *Inverse Problems*, 17(1), 121.
- Arenz, O., Zhong, M., & Neumann, G. (2018). Efficient gradient-free variational inference using policy search. In *International conference on machine learning* (pp. 234–243).
- Arnold, R., & Curtis, A. (2018). Interrogation theory. *Geophysical Journal International*, 214(3), 1830–1846.
- Aster, R. C., Borchers, B., & Thurber, C. H. (2018). *Parameter estimation and inverse problems*. Elsevier.
- Bamler, R., Zhang, C., Opper, M., & Mandt, S. (2017). Perturbative black box variational inference. *arXiv preprint arXiv:1709.07433*.
- Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18.
- Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference* (Unpublished doctoral dissertation). UCL (University College London).
- Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., & Jacobsen, J.-H. (2019). Invertible residual networks. In *International conference on machine learning* (pp. 573–582).
- Belhadj, J., Romary, T., Gesret, A., Noble, M., & Figliuzzi, B. (2018). New parameterizations for bayesian seismic tomography. *Inverse Problems*, 34(6), 065007.
- Bensen, G., Ritzwoller, M., & Yang, Y. (2009). A 3-D shear velocity model of the crust and uppermost mantle beneath the United States from ambient seismic noise. *Geophysical Journal International*, 177(3), 1177–1196.
- Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., . . . others (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7, 453–464.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192–225.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Biswas, R., & Sen, M. (2017). 2D full-waveform inversion and uncertainty estimation using the reversible jump Hamiltonian Monte Carlo. In *Seg technical program expanded abstracts 2017* (pp. 1280–1285). Society of Exploration Geophysicists.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.
- Bodin, T., & Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, 178(3), 1411–1436.
- Bodin, T., Sambridge, M., Tkalcic, H., Arroucau, P., Gallagher, K., & Rawlinson, N. (2012). Trans-

- dimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth*, 117(B2).
- Bozdağ, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., ... Pugmire, D. (2016). Global adjoint tomography: first-generation model. *Geophysical Journal International*, 207(3), 1739–1766.
- Bozdağ, E., Trampert, J., & Tromp, J. (2011). Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2), 845–870.
- Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.
- Brossier, R., Operto, S., & Virieux, J. (2010). Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3), R37–R46.
- Chen, P., Zhao, L., & Jordan, T. H. (2007). Full 3D tomography for the crustal structure of the Los Angeles region. *Bulletin of the Seismological Society of America*, 97(4), 1094–1120.
- Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*.
- Çınlar, E. (2011). *Probability and stochastics* (Vol. 261). Springer Science & Business Media.
- Constable, S. C., Parker, R. L., & Constable, C. G. (1987). Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, 52(3), 289–300.
- 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., & Snieder, R. (1997). Reconditioning inverse problems using the genetic algorithm and revised parameterization. *Geophysics*, 62(5), 1524–1532.
- De Cao, N., Aziz, W., & Titov, I. (2020). Block neural autoregressive flow. In *Uncertainty in artificial intelligence* (pp. 1263–1273).
- De Freitas, N., Höjen-Sørensen, P., Jordan, M. I., & Russell, S. (2001). Variational MCMC. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 120–127).
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.
- Dinh, L., Krueger, D., & Bengio, Y. (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.
- Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.
- Dosso, S. E., Holland, C. W., & Cambridge, M. (2012). Parallel tempering for strongly nonlinear geoacoustic inversion. *The Journal of the Acoustical Society of America*, 132(5), 3030–3040.
- Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2), 216–222.
- Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul), 2121–2159.
- Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. (2019a). Cubic-spline flows. *arXiv preprint arXiv:1906.02145*.
- Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. (2019b). Neural spline flows. *arXiv preprint arXiv:1906.04032*.
- Dziewonski, A. M., & Woodhouse, J. H. (1987). Global images of the Earth's interior. *Science*, 236(4797), 37–48.
- Earl, D. J., & Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23), 3910–3916.

- Fang, H., & Zhang, H. (2014). Wavelet-based double-difference seismic tomography with sparsity regularization. *Geophysical Journal International*, 199(2), 944–955.
- Feroz, F., & Hobson, M. P. (2008). Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2), 449–463.
- Feynman, R. (1972). *Statistical mechanics, a set of lectures, california, institute of technology*. WA Benjamin, Inc. Advanced Book Program Reading, Massachusetts.
- 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., Kennett, B. L., Igel, H., & Bunge, H.-P. (2008). Theoretical background for continental- and global-scale full-waveform inversion in the time–frequency domain. *Geophysical Journal International*, 175(2), 665–685.
- Fichtner, A., Kennett, B. L., Igel, H., & Bunge, H.-P. (2009). Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophysical Journal International*, 179(3), 1703–1725.
- Fichtner, A., van Herwaarden, D.-P., Afanasiev, M., Simutė, S., Krischer, L., Çubuk-Sabuncu, Y., ... others (2018). The collaborative seismic earth model: generation 1. *Geophysical research letters*, 45(9), 4007–4016.
- Fichtner, A., Zunino, A., & Gebraad, L. (2018). Hamiltonian monte carlo solution of tomographic inverse problems. *Geophysical Journal International*, 216(2), 1344–1363.
- French, S., & Romanowicz, B. (2014). Whole-mantle radially anisotropic shear velocity structure from spectral-element waveform tomography. *Geophysical Journal International*, 199(3), 1303–1327.
- Galetti, E., & Curtis, A. (2018). Transdimensional electrical resistivity tomography. *Journal of Geophysical Research: Solid Earth*, 123(8), 6347–6377.
- Galetti, E., Curtis, A., Baptie, B., Jenkins, D., & Nicolson, H. (2017). Transdimensional love-wave tomography of the British Isles and shear-velocity structure of the east Irish Sea Basin from ambient-noise interferometry. *Geophysical Journal International*, 208(1), 36–58.
- Galetti, E., Curtis, A., Meles, G. A., & Baptie, B. (2015). Uncertainty loops in travel-time tomography from nonlinear wave physics. *Physical review letters*, 114(14), 148501.
- 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.
- Gauthier, O., Virieux, J., & Tarantola, A. (1986). Two-dimensional nonlinear inversion of seismic waveforms: Numerical results. *Geophysics*, 51(7), 1387–1403.
- 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. doi: 10.1029/2019JB018428
- Gee, L. S., & Jordan, T. H. (1992). Generalized seismological data functionals. *Geophysical Journal International*, 111(2), 363–390.
- Girolami, M., & Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2), 123–214.
- Gorham, J., & Mackey, L. (2015). Measuring sample quality with Stein’s method. In *Advances in neural information processing systems* (pp. 226–234).
- Grana, D. (2018). Joint facies and reservoir properties inversion. *Geophysics*, 83(3), M15–M24.
- Grana, D., & Della Rossa, E. (2010). Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. *Geophysics*, 75(3), O21–O37.

- Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., & Duvenaud, D. (2018). Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 711–732.
- Green, P. J., & Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, 155(3), 1391–1403.
- Gretton, A. (2013). Introduction to RKHS, and some simple kernel algorithms.
- Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian neural networks. In *Advances in neural information processing systems* (pp. 15379–15389).
- 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.
- Hammersley, J. M., & Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unpublished manuscript*, 46.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.
- Hawkins, R., Bodin, T., Sambridge, M., Choblet, G., & Husson, L. (2019). Trans-dimensional surface reconstruction with different classes of parameterization. *Geochemistry, Geophysics, Geosystems*, 20(1), 505–529.
- Hawkins, R., & Sambridge, M. (2015). Geophysical imaging using trans-dimensional trees. *Geophysical Journal International*, 203(2), 972–1000.
- Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., & Turner, R. (2016). Black-box alpha divergence minimization. In *International conference on machine learning* (pp. 1511–1520).
- Hukushima, K., & Nemoto, K. (1996). Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6), 1604–1608.
- Iyer, H., & Hirahara, K. (1993). *Seismic tomography: Theory and practice*. Springer Science & Business Media.
- Khoshkhohlg, S., Zunino, A., & Mosegaard, K. (2020). Informed proposal monte carlo. *arXiv preprint arXiv:2005.14398*.
- Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems* (pp. 10215–10224).
- Kotsi, M., Malcolm, A., & Ely, G. (2020). Time-lapse full-waveform inversion using hamiltonian monte carlo: A proof of concept. In *Seg technical program expanded abstracts 2020* (pp. 845–849). Society of Exploration Geophysicists.
- Kubrusly, C., & Gravier, J. (1973). Stochastic approximation algorithms and applications. In *1973 ieee conference on decision and control including the 12th symposium on adaptive processes* (pp. 763–766).
- Kucukelbir, A., & Blei, D. M. (2014). Population empirical bayes. *arXiv preprint arXiv:1411.0292*.
- 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.
- Laloy, E., Héault, R., Lee, J., Jacques, D., & Linde, N. (2017). Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Advances in water resources*, 110, 387–405.
- Li, Y., & Turner, R. E. (2016). R\’enyi divergence variational inference. *arXiv preprint arXiv:1602.02311*.
- Liu, J. S., & Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443), 1032–1044.

- Liu, Q. (2017). Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems* (pp. 3115–3123).
- Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning* (pp. 276–284).
- Liu, Q., & Tromp, J. (2006). Finite-frequency kernels based on adjoint methods. *Bulletin of the Seismological Society of America*, 96(6), 2383–2397.
- Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in neural information processing systems* (pp. 2378–2386).
- Liu, Y., Ramachandran, P., Liu, Q., & Peng, J. (2017). Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*.
- Luo, Y., & Schuster, G. T. (1991). Wave-equation traveltime inversion. *Geophysics*, 56(5), 645–653.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
- Malinverno, A. (2002). Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3), 675–688.
- Malinverno, A., Leaney, S., et al. (2000). A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data. In *2000 seg annual meeting*.
- Mandt, S., McInerney, J., Abrol, F., Ranganath, R., & Blei, D. (2016). Variational tempering. In *Artificial intelligence and statistics* (pp. 704–712).
- Mariethoz, G., & Caers, J. (2014). *Multiple-point geostatistics: stochastic modeling with training images*. John Wiley & Sons.
- Martin, G. S., Wiley, R., & Marfurt, K. J. (2006). Marmousi2: An elastic upgrade for Marmousi. *The leading edge*, 25(2), 156–166.
- Martin, J., Wilcox, L. C., Burstedde, C., & Ghattas, O. (2012). A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3), A1460–A1487.
- Marzouk, Y., Moselhy, T., Parno, M., & Spantini, A. (2016). An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*.
- McGrory, C. A., & Titterington, D. (2007). Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51(11), 5352–5367.
- McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
- Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., & Virieux, J. (2016). Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysical Journal International*, 205(1), 345–377.
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American statistical association*, 44(247), 335–341.
- Minka, T. P. (2013). Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*.
- Minsley, B. J. (2011). A trans-dimensional bayesian markov chain monte carlo algorithm for model assessment using frequency-domain electromagnetic data. *Geophysical Journal International*, 187(1), 252–272.
- Mosegaard, K., & Sambridge, M. (2002). Monte carlo analysis of inverse problems. *Inverse problems*, 18(3), R29.
- Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7), 12431–12447.
- Mosser, L., Dubrule, O., & Blunt, M. J. (2020). Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. *Mathematical Geosciences*, 52(1), 53–79.
- Nawaz, M. A., & Curtis, A. (2016). Bayesian inversion of seismic attributes for geological facies

- using a hidden markov model. *Geophysical Journal International*, ggw411.
- Nawaz, M. A., & Curtis, A. (2018). Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies. *Geophysical Journal International*, 214(2), 845–875.
- Nawaz, M. 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*.
- Nawaz, M. A., Curtis, A., Shahraeeni, M. S., & Gerea, C. (2020). Variational Bayesian inversion of seismic attributes jointly for geological facies and petrophysical rock properties. *GEOPHYSICS*, 1-78. Retrieved from <https://doi.org/10.1190/geo2019-0163.1> doi: 10.1190/geo2019-0163.1
- Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705–741.
- Neal, R. M., & Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Springer.
- Neal, R. M., et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11), 2.
- 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.
- Oh, S.-H., & Kwon, B.-D. (2001). Geostatistical approach to bayesian inversion of geophysical data: Markov chain monte carlo method. *Earth, planets and space*, 53(8), 777–791.
- Parisi, G. (1988). *Statistical field theory*. Addison-Wesley.
- Pearl, J. (1982). *Reverend bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science . . .
- Piana Agostinetti, N., Giacomuzzi, G., & Malinverno, A. (2015). Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling. *Geophysical Journal International*, 201(3), 1598–1617.
- 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.
- Pratt, R. G. (1999). Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model. *Geophysics*, 64(3), 888–901.
- Prieux, V., Brossier, R., Operto, S., & Virieux, J. (2013). Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. part 1: Imaging compressional wave speed, density and attenuation. *Geophysical Journal International*, 194(3), 1640–1664.
- Ramirez, A. L., Nitao, J. J., Hanley, W. G., Aines, R., Glaser, R. E., Sengupta, S. K., . . . Daily, W. D. (2005). Stochastic inversion of electrical resistivity changes using a markov chain monte carlo approach. *Journal of Geophysical Research: Solid Earth*, 110(B2).
- Ranganath, R., Altonaar, J., Tran, D., & Blei, D. M. (2016). Operator variational inference. *arXiv preprint arXiv:1610.09033*.
- Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics* (pp. 814–822).
- Ranganath, R., Tran, D., & Blei, D. (2016). Hierarchical variational models. In *International conference on machine learning* (pp. 324–333).
- Rawlinson, N., & Sambridge, M. (2004). Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics*, 69(5), 1338–1350.

- Ray, A., Alumbaugh, D. L., Hoversten, G. M., & Key, K. (2013). Robust and accelerated Bayesian inversion of marine controlled-source electromagnetic data using parallel tempering. *Geophysics*, 78(6), E271–E280.
- Ray, A., Kaplan, S., Washbourne, J., & Albertin, U. (2017). Low frequency full waveform seismic inversion within a tree based Bayesian framework. *Geophysical Journal International*, 212(1), 522–542.
- Ray, A., & Myer, D. (2019). Bayesian geophysical inversion with trans-dimensional gaussian process machine learning. *Geophysical Journal International*, 217(3), 1706–1726.
- Ray, A., Sekar, A., Hoversten, G. M., & Albertin, U. (2016). Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian trans-dimensional algorithm. *Geophysical Journal International*, 205(2), 915–937.
- Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.
- Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.
- Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.
- Roberts, G. O., Tweedie, R. L., et al. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4), 341–363.
- Salimans, T., Kingma, D., & Welling, M. (2015). Markov chain Monte Carlo and variational inference: Bridging the gap. In *International conference on machine learning* (pp. 1218–1226).
- Sambridge, M. (2013). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, ggt342.
- Sambridge, M., & Mosegaard, K. (2002). Monte carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3), 3–1.
- Sato, M.-A. (2001). Online model selection based on the variational bayes. *Neural computation*, 13(7), 1649–1681.
- Sen, M. K., & Biswas, R. (2017). Transdimensional seismic inversion using the reversible jump hamiltonian monte carlo algorithm. *Geophysics*, 82(3), R119–R134.
- Shapiro, N. M., Campillo, M., Stehly, L., & Ritzwoller, M. H. (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science*, 307(5715), 1615–1618.
- Shen, W., Ritzwoller, M. H., Schulte-Pelkum, V., & Lin, F.-C. (2012). Joint inversion of surface wave dispersion and receiver functions: a Bayesian Monte-Carlo approach. *Geophysical Journal International*, 192(2), 807–836.
- Siahkoohi, A., Rizzuti, G., & Herrmann, F. J. (2020). Uncertainty quantification in imaging and automatic horizon tracking—a bayesian deep-prior based approach. In *Seg technical program expanded abstracts 2020* (pp. 1636–1640). Society of Exploration Geophysicists.
- Siahkoohi, A., Rizzuti, G., Witte, P. A., & Herrmann, F. J. (2020). Faster uncertainty quantification for inverse problems with conditional normalizing flows. *arXiv preprint arXiv:2007.07985*.
- Skilling, J. (2004). Nested sampling. In *Aip conference proceedings* (Vol. 735, pp. 395–405).
- Smith, A. (2013). *Sequential Monte Carlo methods in practice*. Springer Science & Business Media.
- Stein, C., et al. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*.
- Tape, C., Liu, Q., Maggi, A., & Tromp, J. (2009). Adjoint tomography of the southern California crust. *Science*, 325(5943), 988–992.
- Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8), 1259–1266.

- Tarantola, A. (1988). Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation. In *Scattering and attenuations of seismic waves, part i* (pp. 365–399). Springer.
- Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation* (Vol. 89). SIAM.
- Tarantola, A., & Valette, B. (1982). Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics*, 20(2), 219–232.
- Team, S. D., et al. (2016). Stan modeling language users guide and reference manual. *Technical report*.
- Theis, L., & Hoffman, M. (2015). A trust-region method for stochastic variational inference with applications to streaming data. In *International conference on machine learning* (pp. 2503–2511).
- Tran, D., Ranganath, R., & Blei, D. M. (2015). The variational Gaussian process. *arXiv preprint arXiv:1511.06499*.
- Tromp, J., Tape, C., & Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, 160(1), 195–216.
- Van Leeuwen, T., & Mulder, W. (2010). A correlation-based misfit criterion for wave-equation traveltime tomography. *Geophysical Journal International*, 182(3), 1383–1394.
- Virieux, J. (1986). P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics*, 51(4), 889–901.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* (Vol. 48). Cambridge University Press.
- Walter, E., & Pronzato, L. (1997). *Identification of parametric models from experimental data*. Springer Verlag.
- Wang, D., Liu, H., & Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. *arXiv preprint arXiv:1810.11943*.
- Wang, D., Tang, Z., Bajaj, C., & Liu, Q. (2019). Stein variational gradient descent with matrix-valued kernels. In *Advances in neural information processing systems* (pp. 7836–7846).
- Warner, M., & Guasch, L. (2016). Adaptive waveform inversion: Theory. *Geophysics*, 81(6), R429–R445.
- Warner, M., Ratcliffe, A., Nangoo, T., Morgan, J., Umpleby, A., Shah, N., ... others (2013). Anisotropic 3D full-waveform inversion. *Geophysics*, 78(2), R59–R80.
- Weaver, R. L., Hadzioannou, C., Larose, E., & Campillo, M. (2011). On the precision of noise correlation interferometry. *Geophysical Journal International*, 185(3), 1384–1392.
- Wengert, R. E. (1964). A simple automatic derivative evaluation program. *Communications of the ACM*, 7(8), 463–464.
- Yao, H., & Van Der Hilst, R. D. (2009). Analysis of ambient noise energy distribution and phase velocity bias in ambient noise tomography, with application to SE tibet. *Geophysical Journal International*, 179(2), 1113–1132.
- Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2003). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8, 236–239.
- Young, M. K., Rawlinson, N., & Bodin, T. (2013). Transdimensional inversion of ambient seismic noise for 3D shear velocity structure of the Tasmanian crust. *Geophysics*, 78(3), WB49–WB62.
- Zhang, 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. (2021). Bayesian full-waveform inversion with realistic priors. *arXiv preprint arXiv:2104.04775*.
- Zhang, X., Curtis, A., Galetti, E., & de Ridder, S. (2018). 3-D Monte Carlo surface wave tomography. *Geophysical Journal International*, 215(3), 1644–1658.
- Zhang, X., Hansteen, F., Curtis, A., & de Ridder, S. (2020). 1D, 2D and 3D Monte Carlo ambient noise tomography using a dense passive seismic array installed on the north sea seabed. *Journal of Geophysical Research: Solid Earth*, 125(2), e2019JB018552. doi: 10.1029/2019JB018552
- Zhang, X., Roy, C., Curtis, A., Nowacki, A., & Baptie, B. (2020, 05). Imaging the subsurface using induced seismicity and ambient noise: 3-D tomographic Monte Carlo joint inversion of earthquake body wave traveltimes and surface wave dispersion. *Geophysical Journal International*, 222(3), 1639–1655. doi: 10.1093/gji/ggaa230
- Zhang, X., & Zhang, H. (2015). Wavelet-based time-dependent travel time tomography method and its application in imaging the Etna volcano in Italy. *Journal of Geophysical Research: Solid Earth*, 120(10), 7068–7084.
- Zhao, X., Curtis, A., & Zhang, X. (2020). Bayesian seismic tomography using normalizing flows. *EarthArXiv*. doi: <https://doi.org/10.31223/X53K6G>
- Zhao, Z., & Sen, M. K. (2019). A gradient based MCMC method for FWI and uncertainty analysis. In *SEG technical program expanded abstracts 2019* (pp. 1465–1469). Society of Exploration Geophysicists.
- Zhdanov, M. S. (2002). *Geophysical inverse theory and regularization problems* (Vol. 36). Elsevier.
- Zobay, O., et al. (2014). Variational bayesian inference with gaussian-mixture approximations. *Electronic Journal of Statistics*, 8(1), 355–389.