Uncertainty quantification for inverse problems with weak PDE-constraints

Zhilong Fang[1], Curt Da Silva[2], Rachel Kuske[3], and Felix J. Herrmann[1,4]

[1]Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia

[2]Department of Mathematics, University of British Columbia

[3]School of Mathematics, Georgia Institute of Technology

[4]School of Earth and Atmospheric Sciences, Georgia Institute of Technology

(December 19, 2017)

Running head: **UQ with weak PDE-constraints**

ABSTRACT

In a statistical inverse problem, the objective is a complete statistical description of unknown parameters from noisy observations in order to quantify uncertainties of the parameters of interest. We consider inverse problems with partial-differential-equation-constraints, which are applicable to a variety of seismic problems. Bayesian inference is one of the most widely-used approaches to precisely quantify statistics through a posterior distribution, incorporating uncertainties in observed data, modeling kernel, and prior knowledge of the parameters. Typically when formulating the posterior distribution, the partial-differential-equation-constraints are required to be exactly satisfied, resulting in a highly nonlinear forward map and a posterior distribution with many local maxima. These drawbacks make it difficult to find an appropriate approximation for the posterior distribution. Another complicating

factor is that traditional Markov chain Monte Carlo methods are known to converge slowly for realistically sized problems. In this work, we relax the partial-differential-equation-constraints by introducing an auxiliary variable, which allows for Gaussian deviations in the partial-differential-equations. Thus, we obtain a new bilinear posterior distribution consisting of both data and partial-differential-equation misfit terms. We illustrate that for a particular range of variance choices for the partial-differential-equation misfit term, the new posterior distribution has fewer modes and can be well-approximated by a Gaussian distribution, which can then be sampled in a straightforward manner. Since it is prohibitively expensive to explicitly construct the dense covariance matrix of the Gaussian approximation for intermediate to large-scale problems, we present a method to implicitly construct it, which enables efficient sampling. We apply this framework to two-dimensional seismic inverse problems with $1,800$ and $92,455$ unknown parameters. The results illustrate that our framework can produce comparable statistical quantities to those produced by conventional Markov chain Monte Carlo type methods while requiring far fewer partial-differential-equation solves, which are the main computational bottlenecks in these problems.

INTRODUCTION

Inverse problems with partial-differential-equation (PDE) constraints arise in many applications of geophysics (Tarantola and Valette, 1982a; Pratt, 1999; Haber et al., 2000; Epanomeritakis et al., 2008; Virieux and Operto, 2009). The goal of these problems is to infer the values of unknown spatial distributions of physical parameters (e.g., sound speed, density, or electrical conductivity) from indirectly measured data, where the underlying physical model is described by a PDE (e.g., the Helmholtz equation or Maxwell's equations). The most challenging aspects of these problems arise from the fact that they are typically multimodal, with many spurious local minima (Biegler et al., 2012), which can inhibit gradient-based optimization algorithms from estimating the true parameters successfully.

This multimodality stems in part from the fact that the observed data are measured on a small subset of the entire boundary of the domain (Bui-Thanh et al., 2013) and the nonlinear parameter-to-data forward map (van Leeuwen and Herrmann, 2013; van Leeuwen and Herrmann, 2015). One approach to dealing with the multimodality is to formulate the inverse problem as a deterministic optimization problem that aims at minimizing the misfit between the predicted and observed data in an appropriate norm, while also adding a regularization term that may eliminate the nonconvexity in certain situations (Virieux and Operto, 2009; Martin et al., 2012). The result of this deterministic approach is an estimate of the model parameters that is consistent with the observed data and contains few unwanted features. Since observed data typically contain measurement noise and modeling errors, we are not only interested in an estimate that best fits the data, but also in a complete statistical description of the unknown model parameters (Tarantola and Valette, 1982b; Osypov et al., 2013). To that end, statistical approaches, and in particular the Bayesian

inference method, are desirable and necessary. Unlike in the deterministic case, the solution produced by Bayesian inference is a posterior probability density function (PDF), which incorporates uncertainties in the observed data, the forward modeling map, and one's prior knowledge of the parameters. Once we can tractably compute the posterior distribution, we can extract various statistical properties of the unknown parameters.

Bayesian inference methods have been applied to a number of PDE-constrained geophysical statistical inverse problems (Martin et al., 2012; Bui-Thanh et al., 2013; Zhu et al., 2016; Ely et al., 2017). In these reported works, the PDE is typically treated as a strict constraint when formulating the posterior PDF, i.e., the field variables should always exactly satisfy the PDE. This leads to the so-called reduced or adjoint-state method (Plessix, 2006; Hinze et al., 2008) that eliminates the field variables by solving the PDE, resulting in a posterior PDF with multiple modes. To study the posterior PDF, Markov chain Monte Carlo (McMC) type methods, including the Metropolis-Hasting based methods (Haario et al., 2006; Stuart et al., 2016; Ely et al., 2017), the stochastic Newton-type method (Martin et al., 2012), and the randomize-then-optimize (RTO) method (Bardsley et al., 2015) sample the posterior PDF by drawing samples from a proposal distribution followed by an accept or reject step. To compute the accept/reject ratio, these methods have to evaluate the posterior PDF for each sample, which leads to solving a large number of computationally expensive PDEs. Moreover, according to the scaling analysis by Roberts et al. (2001), McMC type methods require a significantly larger number of samples to reach a status of convergence for large-scale problems in comparison with small-scale problems, which is well known as the curse of dimensionality. These difficulties preclude the straightforward applications of these methods to large-scale problems with more than 10^6 unknown parameters.

Contributions

In this work, we present a new formulation of the posterior distribution that subsumes the conventional reduced formulation as a special case. Instead of treating the PDE as a strict constraint and eliminating the field variables by solving the PDE, we relax the PDE-constraint by introducing the field variables as auxiliary variables. This idea is similar to the method of van Leeuwen and Herrmann (2015) applied to deterministic PDE-constrained optimization problems, in which the PDE misfit is treated as a penalty term in the misfit function and weighted by a penalty parameter. Moreover, the idea of relaxing the PDE-constraint is also widely-used in weather forecasting applications (Fisher et al., 2005) for the sake of improving the stability of results. In the field of seismic exploration, Fang et al. (2015) and Fang et al. (2016) first introduced a method to construct a posterior PDF with weak acoustic wave-equation constraints. In the following study by Lim (2017), the authors showed that for small-scale problems, this posterior PDF can be sampled by the randomize maximum likelihood-McMC (RML-McMC) method. We demonstrate that the conventional reduced posterior PDF is a special case of our new formulation. By exploiting the structure of the new posterior PDF, we show that, with an appropriate penalty parameter, the new posterior PDF can be approximated by a Gaussian distribution, which is centered at the maximum a posteriori (MAP) estimate that maximizes the posterior PDF. To construct this Gaussian approximation, we exploit the local derivative information of the posterior PDF and formulate the covariance matrix as a PDE-free operator, which allows us to compute the matrix-vector product without the requirement of computing a large number of additional PDEs. By avoiding an explicit formulation of the covariance matrix, which would be impractical to compute and store, we can apply a recently proposed bootstrap type method (Efron, 1981, 1992) — the so-called randomize-then-optimize method (Bardsley et al., 2014) — to affordably

draw samples from this surrogate distribution.

We apply our new computational framework to several seismic wave-equation based inverse problems ranging in size and complexity of the underlying parameters. Our first example compares our sampling method with a benchmark method — the randomize maximum likelihood (RML) method (Chen and Oliver, 2012) — to validate our Gaussian approximation on a simple model with 1800 unknown parameters. Next, we apply our computational framework to a more complex model with 92,455 unknown parameters to test the feasibility of the approach to more realistically sized problems.

This paper is organized into three major sections. The first introduces the derivation of the posterior PDF and the corresponding sampling method in a general setting. The second section introduces each component in the general framework when applied to the full-waveform inversion type problems. The final section presents the results of the application of our framework to several numerical inverse problems for velocity models with different size and complexity.

BAYESIAN FRAMEWORK FOR INVERSE PROBLEMS WITH A WEAK PDE-CONSTRAINT

In a PDE-constrained inverse problem, the goal is to infer the unknown discretized n_{grid}-dimensional physical model parameters $\mathbf{m} \in \mathbb{R}^{n_{\text{grid}}}$ from n_{data}-dimensional noisy observed data $\mathbf{d} \in \mathbb{C}^{n_{\text{data}}}$. As the noisy data are stochastic in nature, so are the inversion results obtained from them. Bayesian inference is a widely-used approach that seeks to estimate the posterior PDF of the unknown parameters $\mathbf{m}$ by incorporating the statistics of the measurement and modeling error and one's prior knowledge of the underlying model. Mathematically,

Bayesian inference applies Bayes' law to formulate the posterior PDF $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ of the model parameters $\mathbf{m}$ given the observed data $\mathbf{d}$ by combining a likelihood PDF and a prior PDF as

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \rho_{\text{like}}(\mathbf{d}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m}), \tag{1}$$

where the likelihood PDF $\rho_{\text{like}}(\mathbf{d}|\mathbf{m})$ describes the probability of observing the data $\mathbf{d}$ given the model parameters $\mathbf{m}$, and the prior PDF $\rho_{\text{prior}}(\mathbf{m})$ describes one's prior beliefs in the unknown model parameters. The proportionality constant depends on the observed data $\mathbf{d}$, which are fixed. Once we have a computationally tractable estimate of the posterior PDF, we can apply certain sampling methods to draw samples from the posterior PDF, which can then be used to compute statistical properties of interest such as the MAP estimate, the mean value, the model covariance matrix, the model standard deviation (STD), and the marginal distributions of $\mathbf{m}$ (Kaipio and Somersalo, 2006; Matheron, 2012). The primary issue for statisticians is to construct the posterior PDF and design methods that can draw samples from it with affordable cost.

The posterior PDF

To motivate the derivation of the new posterior PDF, it is helpful to start with the conventional formulation of the posterior PDF for PDE-constrained inverse problems. The so-called reduced approach eliminates the PDE-constraint by solving the PDE, which leads to the following nonlinear forward modeling map $F(\mathbf{m})$:

$$F(\mathbf{m}) = \mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}. \tag{2}$$

Here the vector $\mathbf{q} \in \mathbb{C}^{n_{\text{grid}}}$ represents the discretized (known) source term. The matrix $\mathbf{A}(\mathbf{m}) \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{grid}}}$ denotes the discretized PDE operator and the operator $\mathbf{P} \in \mathbb{R}^{n_{\text{data}} \times n_{\text{grid}}}$

samples the data $\mathbf{d}$ from the vector of field variables $\mathbf{u}$, which is the solution of the PDE $\mathbf{u} = \mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$. In real-world seismic applications, the observed data always contain both correlated measurement noise arising from environmental disturbances and modeling errors, which are difficult to precisely quantify and model. One popular approach to simplify the problem is to assume that the combined measurement and modeling noise $\epsilon \in \mathbb{C}^{n_{\text{data}}}$ is drawn from a Gaussian distribution with zero mean and covariance matrix Γ_{noise} (Bui-Thanh et al., 2013; Osypov et al., 2013; Bardsley et al., 2015), i.e., $\epsilon \sim \mathcal{N}(0, \Gamma_{\text{noise}})$, independent of $\mathbf{m}$. The assumption of Gaussianity results in distributions that are relatively easy to model and sample, thereby providing a rich source of tractable examples (Kaipio and Somersalo, 2006). With this assumption in mind and the additional assumption that the prior distribution of the model $\mathbf{m}$ is also Gaussian with the mean model parameters $\tilde{\mathbf{m}}$ and the covariance matrix Γ_{prior}, we arrive at the following posterior distribution:

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\left(-\frac{1}{2}\|F(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right), \tag{3}$$

where the symbol $\|\,\cdot\,\|_{\Gamma^{-1}_{\text{noise}}}$ denotes the weighted ℓ_2-norm with the weighting matrix $\Gamma^{-1}_{\text{noise}}$. There are several challenges in computing various quantities associated with the posterior distribution in equation 3. In order to obtain the MAP estimate $\mathbf{m}_*$, we need to solve the following deterministic optimization problem:

$$\begin{aligned} \mathbf{m}_* &= \arg\max_{\mathbf{m}} \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) = \arg\min_{\mathbf{m}} -\log \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \\ &= \arg\min_{\mathbf{m}} \frac{1}{2}\|F(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}, \end{aligned} \tag{4}$$

which can be solved by the so-called adjoint-state method. As noted in van Leeuwen and Herrmann (2013), the nonlinear forward modeling map $F(\mathbf{m})$ results in the objective function $-\log \rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ being highly oscillatory with respect to the model parameters $\mathbf{m}$, which yields many local minima. To find the globally optimal solution, a sufficiently close initial

model is necessary, which may be difficult to obtain in real-world scenarios. As mentioned previously, the nonlinear parameter-to-data map also results in computational difficulties when sampling the posterior distribution in equation 3. Specifically, the tradeoff between designing a proposal distribution that is well tuned to the true posterior distribution and one that is computationally cheap to sample is not a straightforward choice to make. As one models a proposal that is easier to sample, typically the price to pay is having to draw more samples until convergence is reached.

These challenges result from the nonlinear forward modeling map $F(\mathbf{m})$ induced by the strict PDE-constraint in the optimization problem in equation 4. To overcome these difficulties, van Leeuwen and Herrmann (2013) and van Leeuwen and Herrmann (2015) proposed a penalty formulation to solve deterministic PDE-constrained optimization problems, wherein they relax the strict PDE-constraint by penalizing the data misfit function by a weighted PDE misfit with a penalty parameter λ. This results in the following joint optimization problem with respect to both the model parameters $\mathbf{m}$ and the field variables collected in the vector $\mathbf{u}$:

$$\arg\min_{\mathbf{m},\mathbf{u}} f_{\text{pen}}(\mathbf{m},\mathbf{u}) = \frac{1}{2}\|\mathbf{P}\mathbf{u}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{\lambda^2}{2}\|\mathbf{A}(\mathbf{m})\mathbf{u}-\mathbf{q}\|^2 + \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}. \quad (5)$$

The authors note that the problem in equation 5 is a separable nonlinear least-squares problem, in which the optimization with respect to $\mathbf{u}$ is a linear data-fitting problem when $\mathbf{m}$ is fixed. In van Leeuwen and Herrmann (2015), the authors eliminate the field variables $\mathbf{u}$ by the variable projection method (Golub and Pereyra, 2003) in order to avoid the high memory costs involved in storing a unique field variable for each individual source. The variable projection method also eliminates the dependence of the objective function in equation 5 on $\mathbf{u}$. As for each input parameter $\mathbf{m}$, there is a unique $\overline{\mathbf{u}}(\mathbf{m})$ satisfying $\nabla_{\mathbf{u}} f_{\text{pen}}(\mathbf{m},\mathbf{u})|_{\mathbf{u}=\overline{\mathbf{u}}(\mathbf{m})} = 0$,

which has the closed form solution

$$\overline{\mathbf{u}}(\mathbf{m}) = \big(\lambda^2 \mathbf{A}(\mathbf{m})^\top \mathbf{A}(\mathbf{m}) + \mathbf{P}^\top \Gamma_{\text{noise}}^{-1} \mathbf{P}\big)^{-1} \big(\lambda^2 \mathbf{A}(\mathbf{m})^\top \mathbf{q} + \mathbf{P}^\top \Gamma_{\text{noise}}^{-1} \mathbf{d}\big), \tag{6}$$

where the symbol $\top$ denotes the (complex-conjugate) transpose. As noted by van Leeuwen and Herrmann (2013) and van Leeuwen and Herrmann (2015), minimizing the objective function (equation 5) with a carefully selected λ is less prone to being trapped in suboptimal local minima, because the inversion is carried out over a larger search space (implicitly through $\overline{\mathbf{u}}(\mathbf{m})$) and it is therefore easier to fit the observed data for poor starting models compared to the conventional reduced formulation in equation 4.

Motivated by the penalty approach to solving the deterministic inverse problems, we propose a more generic posterior PDF for statistical PDE-constrained inverse problems. As before, we relax the PDE-constraint by introducing the field variables $\mathbf{u}$ as auxiliary variables, i.e., we have

$$\rho_{\text{post}}(\mathbf{u}, \mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{u}, \mathbf{m}) \rho(\mathbf{u}, \mathbf{m}), \tag{7}$$

where the conditional PDF $\rho(\mathbf{d}|\mathbf{u}, \mathbf{m})$ now describes the probability of observing the data $\mathbf{d}$ given the field variables $\mathbf{u}$ and model parameters $\mathbf{m}$. To formulate the joint PDF $\rho(\mathbf{u}, \mathbf{m})$, we apply the standard conditional decomposition (Sambridge et al., 2006)

$$\rho(\mathbf{u}, \mathbf{m}) = \rho(\mathbf{u}|\mathbf{m}) \rho_{\text{prior}}(\mathbf{m}). \tag{8}$$

Hence,

$$\rho_{\text{post}}(\mathbf{u}, \mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{u}, \mathbf{m}) \rho(\mathbf{u}|\mathbf{m}) \rho_{\text{prior}}(\mathbf{m}). \tag{9}$$

The implication of the reduced formulation is that the field variables $\mathbf{u}$ satisfy the PDE strictly—i.e., $\mathbf{u} = \mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$. This adherence to the PDE corresponds to solutions $\mathbf{u}$ that satisfy the PDE $\mathbf{A}(\mathbf{m})\mathbf{u} = \mathbf{q}$ with the probability density $\rho(\mathbf{u}|\mathbf{m}) = \delta\big(\mathbf{A}(\mathbf{m})\mathbf{u} - \mathbf{q}\big)$, where

$\delta()$ denotes the Kronecker delta function (Dummit and Foote, 2004). Conversely, replacing the constraint by a quadratic penalty term allows for a Gaussian error with zero mean and covariance matrix $\lambda^{-2}\mathbf{I}$ in the PDE misfit $\mathbf{A}(\mathbf{m})\mathbf{u}-\mathbf{q}$. This yields

$$\rho(\mathbf{u}|\mathbf{m}) = (2\pi)^{-\frac{n_{\text{grid}}}{2}} \det\left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda^2}{2}\|\mathbf{A}(\mathbf{m})\mathbf{u}-\mathbf{q}\|^2\right). \tag{10}$$

Indeed, for a given $\mathbf{m}$, this distribution $\rho(\mathbf{u}|\mathbf{m})$ with respect to $\mathbf{u}$ is a Gaussian distribution with a mean of $\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$ and a covariance matrix of $\lambda^{-2}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}$. The conditional probability $\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})$ is the joint PDF with respect to both the model parameters $\mathbf{m}$ and field variables $\mathbf{u}$. Because the control or model variables $\mathbf{m}$ are of primary interest, we eliminate the dependence of the joint PDF on the auxiliary variables $\mathbf{u}$ by marginalizing over $\mathbf{u}$:

$$\begin{aligned}\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) &= \int \rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})\mathrm{d}\mathbf{u}\\ &\propto \det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \det\left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\right)^{\frac{1}{2}}\\ &\quad\times \exp\Big(-\frac{1}{2}\big(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m})-\mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m})-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big)\Big),\end{aligned} \tag{11}$$

where

$$\begin{aligned}\mathbf{H}(\mathbf{m}) &= -\nabla^2_{\mathbf{u}} \log \rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})\big|_{\mathbf{u}=\overline{\mathbf{u}}(\mathbf{m})}\\ &= \lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m}) + \mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{P},\end{aligned} \tag{12}$$

and $\overline{\mathbf{u}}(\mathbf{m})$ is given by equation 6. A complete derivation of the marginal PDF $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ is given in Appendix A . As in the deterministic case, the posterior PDF corresponding to the conventional reduced formulation (cf. equation 3) can also be derived from the marginal PDF $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ in equation 11. Indeed, as $\lambda\to\infty$, we have

$$\lim_{\lambda\to\infty}\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\big(-\frac{1}{2}\|\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big), \tag{13}$$

which, as expected, is the posterior PDF corresponding to the conventional reduced formulation. The derivation of equation 13 can be found in Appendix A in more details.

To illustrate the advantages of our penalty formulation for the posterior PDF (cf. equation 11) over the conventional reduced formulation (cf. equation 3), we conduct a simple experiment adopted from Esser et al. (2016). We invert for the sound speed given partial measurements of the pressure field and generate data with a known band-limited source. The data are recorded at four receivers (see Figure 1a) and is contaminated with Gaussian noise (see Figure 1b). As in Esser et al. (2016), we define the forward operator $F(m)$ as

$$F(m) = \mathbf{P}\mathbf{A}^{-1}(m)\mathbf{q} = \begin{pmatrix} \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_1-\mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \\ \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_2-\mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \\ \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_3-\mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \\ \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_4-\mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \end{pmatrix}, \tag{14}$$

where the operator $\mathcal{F}$ denotes the temporal Fourier transform and ω denotes the angular frequency. The vectors $\mathbf{x}_i$, $i = 1, ..., 4$ and $\mathbf{x}_s$ denote the receiver and source locations, respectively, and the scalar m represents the slowness of the medium, i.e., $m = \frac{1}{v}$ where v is the velocity. The source and receiver locations are in Figure 1a denoted by the symbols $(*)$ and (∇). With the forward model defined in equation 14, we formulate the posterior distribution for the reduced formulation and the penalty formulation by choosing a Gaussian distribution with a mean of $5 \cdot 10^{-4}$ s/m and a standard deviation of $1.2 \cdot 10^{-4}$ s/m for the prior distribution $\rho_{\text{prior}}(m)$. Finally, we add a Gaussian noise with a variance of 4×10^{-4} to the observed data, resulting in a noise-to-signal ratio of $\|\text{noise}\|_2/\|\text{signal}\|_2 = 0.24$.

[Figure 1 about here.]

Figure 2 depicts the posterior PDFs of the reduced and penalty formulations for $\lambda =$ 10, 50, and 250. As expected, local maxima are present in the PDF of the reduced formulation and in the PDFs of the penalty formulation when the λ values become too large. As λ

increases from $\lambda = 10$, where the posterior is unimodol, local maxima appear for larger $\lambda = 250$, only one of which has strong statistical significance. For $\lambda = 250$, the resulting PDF is close to the one yielded by the reduced formulation, which corresponds to $\lambda \to \infty$. From this stylized example, the strongly relaxed formulation appears unimodal and with low bias. As we will demonstrate in later sections, being less prone to local maxima reduces the computational cost associated with sampling these distributions.

[Figure 2 about here.]

Selection of λ

Before discussing computationally efficient sampling schemes, we first propose a method to select values for λ, which will balance the tradeoff between the unimodality of the distribution and its deviation from the reduced-formulation PDF.

To arrive at a scheme to select λ, we focus on two terms of the negative logarithm function of the posterior PDF in equation 11:

$$\begin{aligned}\phi(\mathbf{m}) =& -\log \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \\ =& \frac{1}{2}\log\det\left(\mathbf{I} + \frac{1}{\lambda^2}\Gamma_{\text{noise}}^{-\frac{1}{2}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma_{\text{noise}}^{-\frac{\top}{2}}\right) \\ &+ \frac{1}{2}\left(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right),\end{aligned} \tag{15}$$

namely the determinant

$$\phi_1(\mathbf{m}) = \frac{1}{2}\log\det\left(\mathbf{I} + \frac{1}{\lambda^2}\Gamma_{\text{noise}}^{-\frac{1}{2}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma_{\text{noise}}^{-\frac{\top}{2}}\right), \tag{16}$$

and the misfit

$$\phi_2(\mathbf{m}) = \frac{1}{2}\left(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right). \tag{17}$$

These equations imply that we should avoid situations in which $\lambda \to 0$ and $\lambda \to \infty$. When $\lambda \to 0$, the optimal variable $\overline{\mathbf{u}}(\mathbf{m})$ tends to fit the data. As a result, $\phi_2(\mathbf{m}) \to \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}$, which means that the observed data are not informative to the unknown parameters and have few contributions to the posterior distribution. Furthermore, as $\lambda \to 0$, the nonlinear determinant function $\phi_1(\mathbf{m}) \to \infty$ and will dominate the overall function $\phi(\mathbf{m})$, which results in a highly nonlinear mapping $\mathbf{m} \to \phi(\mathbf{m})$. On the other hand, when $\lambda \to \infty$, we find that $\phi_1(\mathbf{m}) \to 0$ and $\phi_2(\mathbf{m}) \to \frac{1}{2}\|\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}$ in which case the misfit $\phi(\mathbf{m})$ converges to the nonlinear reduced formulation. Considering both facts, we want to find an appropriate λ, so that $\phi_1(\mathbf{m})$ is relatively small compared to $\phi_2(\mathbf{m})$, thus ensuring enough information from the observed data, while $\phi_2(\mathbf{m})$ is still less likely to contain local minima.

Based on spectral arguments, van Leeuwen and Herrmann (2015) proposed a scaling for the penalty parameter according to the largest eigenvalue μ_1 of the matrix $\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}$. Relative to μ_1, a penalty parameter $\lambda^2 \gg \mu_1$ can be considered large while $\lambda^2 \ll \mu_1$ is considered small. As a result, λ chosen much greater than this reference scale—i.e., when $\lambda^2 \gg \mu_1$, the minimizers for field variables $\overline{\mathbf{u}}(\mathbf{m})$ will converge to the solution of the wave equation $\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$ with a convergence rate of $\mathcal{O}(\lambda^{-2})$, and therefore our penalty misfit approaches the reduced misfit. A similar consideration applies when $\lambda^2 \ll \mu_1$. After extensive parameter testing, we found that choosing $\lambda^2 = 0.01\mu_1$ strikes the right balance so that the posterior PDF is less affected by local maxima compared to the reduced formulation while the determinant term $\phi_1(\mathbf{m})$ remains negligible compared to $\phi_2(\mathbf{m})$. With this choice of λ, we can therefore neglect the $\phi_1(\mathbf{m})$ term, as it is small relative to $\phi_2(\mathbf{m})$, and consider an approximate posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ consisting only of

the $\phi_2(\mathbf{m})$ term, i.e., we now have the approximate equality

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \approx \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\left(-\frac{\lambda^2}{2}\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 - \frac{1}{2}\|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right). \tag{18}$$

This approximation results in a posterior PDF that is much easier to evaluate. From here on out, we consider $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ as our PDF of interest in the subsequent sampling considerations.

Sampling method

Given this choice of λ, yielding the PDF in equation 18, the computational considerations of drawing samples from this approximate distribution are paramount in designing a tractable method. McMC-type methods are unfortunately computationally unfeasible for high-dimensional problems, owing to the relatively large number of the expensive evaluations of posterior distributions needed to converge adequately. For this reason, we follow an alternative approach — widely used in Bayesian inverse problems with PDE-constraints (Bui-Thanh et al., 2013; Zhu et al., 2016) — where we approximate the target posterior PDF by a Gaussian PDF. To construct this Gaussian PDF, we first find the MAP estimate $\mathbf{m}_*$ by solving

$$\mathbf{m}_* = \arg\min_{\mathbf{m}} -\log\left(\overline{\rho}_{\text{post}}\left(\mathbf{m}|\mathbf{d}\right)\right). \tag{19}$$

Next, we use the local second-order derivative information of the posterior PDF at the MAP point — i.e., the Hessian $\mathbf{H}_{\text{post}} = -\nabla^2_{\mathbf{m}} \log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ — to construct the covariance matrix of the Gaussian PDF, which yields the Gaussian distribution $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$. Afterwards, we draw samples from the Gaussian distribution from which we compute statistical quantities. We incorporate this sampling strategy with the proposed posterior PDF with weak PDE-constraints and obtain the following Bayesian framework as shown in Algorithm 1:

Algorithm 1 Bayesian framework for inverse problems with weak PDE-constraints

1. Set Γ_{noise}, Γ_{prior}, prior mean model $\tilde{\mathbf{m}}$, and a value for the penalty parameter λ;
2. Formulate the posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ with equation 18;
3. Find the MAP estimate $\mathbf{m}_*$ by minimizing $-\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$;
4. Compute the Hessian matrix $\mathbf{H}_{\text{post}} = -\nabla^2_{\mathbf{m}} \log\left(\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})\right)$ at the MAP estimate $\mathbf{m}_*$ and define the Gaussian PDF $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$;
5. Draw n_{smp} samples from the Gaussian PDF $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$;
6. Compute statistical properties of interest from the n_{smp} samples.

Compared to McMC type methods, the additional evaluations of the posterior PDF are not needed once we calculate the MAP estimate $\mathbf{m}_*$, which significantly reduces the computational cost. However, the accuracy of samples drawn from this surrogate PDF strongly depends on the accuracy of the Gaussian approximation in a neighborhood of $\mathbf{m}_*$, which is related to our choice of λ. To illustrate this dependence, we continue the example shown in Figure 2 and compare the Gaussian approximation of the reduced formulation (i.e., $\lambda = \infty$) to the penalty formulation for different values of λ, plotted in Figure 3. In this case the largest eigenvalue is $\mu_1 = 10^4$ and the corresponding is $\lambda = 10$. Clearly, when selecting $\lambda = 10$, the Gaussian approximation is relatively close to the true PDF, whereas increasing λ decreases the accuracy of the Gaussian approximation.

[Figure 3 about here.]

Armed with an accurate Gaussian approximation to the unimodal PDF (for an appropriate choice of λ), we are now in a position to draw samples from the Gaussian distribution $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$. For small-scale problems, an explicit expression of the Hessian $\mathbf{H}_{\text{post}}$ is available. Hence, we can draw samples $\mathbf{m}_{\text{s}}$ from the distribution $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$ by utilizing

the Cholesky factorization of the Hessian $\mathbf{H}_{\text{post}}$—i.e., $\mathbf{H}_{\text{post}} = \mathbf{R}_{\text{post}}^{\top}\mathbf{R}_{\text{post}}$ — as follows (Rue, 2001):

$$\mathbf{m}_{\text{s}} = \mathbf{m}_{*} + \mathbf{R}_{\text{post}}^{-1}\mathbf{r}, \tag{20}$$

where the matrix $\mathbf{R}_{\text{post}}$ is an upper triangular matrix and the vector $\mathbf{r}$ is a random vector drawn from the n_{grid}-dimensional standard Gaussian distribution $\mathcal{N}(0, \mathcal{I}_{n_{\text{grid}} \times n_{\text{grid}}})$. Nevertheless, for large-scale problems, constructing and storing an explicit expression of the Hessian $\mathbf{H}_{\text{post}}$ is infeasible. Typically, to avoid the construction of the explicit Hessian matrix, the Hessian $\mathbf{H}_{\text{post}}$ is constructed as a matrix-free implicit operator, and we only have access to computing the matrix-vector product with an arbitrary vector. As a result, we need a matrix-free sampling method to draw samples. In the following section, we will develop the implementation details for a suitable sampling method for seismic wave-equation-constrained inverse problems.

UNCERTAINTY QUANTIFICATION FOR SEISMIC WAVE-EQUATION-CONSTRAINED INVERSE PROBLEMS

Wave-equation based inversions, where the coefficients of the wave-equation are the unknowns, are amongst the most computationally challenging inverse problems as they typically require wave-equation solves for multiple source experiments on a large domain where the wave travels many wavelengths. Additionally, these inverse problems, also known as full-waveform inversion (FWI) in the seismic community (Pratt, 1999), involve fitting oscillatory predicted and observed data, which can result in parasitic local minima.

Motivated by the penalty formulation with its weak PDE-constraints and the results presented above presumably, we will derive a time-harmonic Bayesian inversion framework

that is capable of handling relatively large-scale problems where the wave propagates about 30 wavelengths, a moderate number for realistic problems. Given acoustic seismic data, collected at the surface from sources that fire along the same surface, our aim is to construct the statistical properties of the spatial distribution of the acoustic wave velocity. With these properties, we are able to conduct uncertainty quantification for the recovered velocity model. The subsections are organized roughly in according to the main steps outlined in Algorithm 1.

Steps 1 and 2: designing the posterior and prior PDF

To arrive at a Bayesian formulation for wave-equation based inverse problems with weak constraints, we consider monochromatic seismic data collected at n_{rcv} receivers locations from n_{src} seismic source locations and sampled at n_{freq} frequencies. Hence, the observed data $\mathbf{d} \in \mathbb{C}^{n_{\text{data}}}$ with $n_{\text{data}} = n_{\text{rcv}} \times n_{\text{src}} \times n_{\text{freq}}$. As before, the synthetic data are obtained by applying, for each source, the projection operator $\mathbf{P} \in \mathbb{C}^{n_{\text{rcv}} \times n_{\text{grid}}}$. For the i^{th} source and j^{th} frequency, the time-harmonic wave equation corresponds to the following discretized Helmholtz system:

$$\mathbf{A}_j(\mathbf{m})\mathbf{u}_{i,j} = \mathbf{q}_{i,j} \quad \text{with} \quad \mathbf{A}_j(\mathbf{m}) = \Delta + \omega_j^2 \,\text{diag}\left(\mathbf{m}^{-2}\right). \tag{21}$$

In this expression, the $\mathbf{q}_{i,j}$'s are the monochromatic sources, the symbol Δ refers to the discretized Laplacian, ω represents the angular frequency, and $\mathbf{m} \in \mathbb{R}^{n_{\text{grid}}}$ denotes the vector with the discretized velocities. With a slight abuse of notation, this vector appears as the elementwise reciprocal square on the diagonal. To discretize this problem, we use the Helmholtz discretization from Chen et al. (2013).

If we consider the data from all sources and frequencies simultaneously, the posterior

PDF for the weak-constrained penalty formulation becomes

$$\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}}\|\mathbf{P}\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{d}_{i,j}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{\lambda^2}{2}\|\mathbf{A}_j(\mathbf{m})\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\|^2\right) \times \exp\left(-\frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right). \tag{22}$$

Aside from choosing a proper value for the penalty parameter λ, another crucial component of the posterior PDF in equation 22 is the choice of the prior PDF. From a computational perspective, a suitable prior should have a bounded variance and one should be able to draw samples with moderate cost (Bui-Thanh et al., 2013). More specifically, this results in having computationally feasible access to (matrix-free) actions of the square-root of the prior covariance operator or its inverse on random vectors. To meet this requirement, we utilize Gaussian smoothness priors (Matheron, 2012), which provide a flexible way to describe random fields and are commonly employed in Bayesian inference (Lieberman et al., 2010; Martin et al., 2012; Bardsley et al., 2015). Following Lieberman et al. (2010), we construct a Gaussian smoothness prior $\rho_{\text{prior}}(\mathbf{m}) \propto \mathcal{N}(\tilde{\mathbf{m}}, \Gamma_{\text{prior}})$ with a reference mean model $\tilde{\mathbf{m}}$ and a covariance matrix Γ_{prior} given by

$$\Gamma_{\text{prior}}(k,l) = a\exp\left(\frac{-\|\mathbf{s}_k - \mathbf{s}_l\|^2}{2b^2}\right) + c\delta_{k,l}. \tag{23}$$

In this expression for the covariance, the vectors $\mathbf{s}_k = (z_k, x_k)$ and $\mathbf{s}_l = (z_l, x_l)$ denote the k^{th} and l^{th} spatial coordinates corresponding to the k^{th} and l^{th} elements in the vector $\mathbf{m}$, respectively. The parameters a, b, and c control the correlation strength, variance, and spatial correlation distance. The variance of the k^{th} element m_k is $\text{var}(m_k) = \Gamma_{\text{prior}}(k,k) = a + c$. The parameter c also ensures that the prior covariance matrix remains numerically well-conditioned (Martin et al., 2012). Clearly, when the distance between $\mathbf{s}_k$ and $\mathbf{s}_l$ is large—i.e., $\frac{\|\mathbf{s}_k - \mathbf{s}_l\|^2}{2b^2} \gg 1$, the cross-covariance $\Gamma_{\text{prior}}(k,l)$ vanishes quickly.

Steps 3 and 4: Gaussian approximation

After setting up the posterior PDF, we need to construct its Gaussian approximation $\mathcal{N}(\mathbf{m}_*, \mathbf{H}_{\text{post}}^{-1})$, corresponding to steps 3 and 4 in Algorithm 1. In order to achieve this objective, first we compute the MAP estimate of the posterior PDF $\mathbf{m}_*$, which is equivalent to minimizing the negative logarithm $-\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$:

$$\begin{aligned}\mathbf{m}_* &= \arg\min_{\mathbf{m}} -\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \\ &= \arg\min_{\mathbf{m}} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \frac{1}{2} \|\mathbf{P}\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{d}_{i,j}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{\lambda^2}{2} \|\mathbf{A}_j(\mathbf{m})\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\|^2 \\ &\qquad + \frac{1}{2} \|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}.\end{aligned} \tag{24}$$

Note that the objective function $-\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ is analogous to the cost function of the deterministic optimization problem (van Leeuwen and Herrmann, 2015). Using similar techniques as in the aforementioned work, we can express the gradient $\mathbf{g}$ for this objective as

$$\mathbf{g} = \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \lambda^2 \mathbf{G}^\top_{i,j} \big(\mathbf{A}_j(\mathbf{m})\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\big) + \Gamma^{-1}_{\text{prior}}(\mathbf{m} - \tilde{\mathbf{m}}), \tag{25}$$

where the sparse Jacobian matrix $\mathbf{G}_{i,j} = \big(\nabla_{\mathbf{m}} \mathbf{A}_j(\mathbf{m})\big)\overline{\mathbf{u}}_{i,j}(\mathbf{m})$. Following van Leeuwen and Herrmann (2013), we use the limited-memory-Broyden–Fletcher–Goldfarb–Shanno method (l-BFGS, Nocedal and Wright, 2006) to solve the optimization problem in equation 24 to find the MAP estimate $\mathbf{m}_*$.

Once we have computed $\mathbf{m}_*$, we focus on approximating the posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ by a Gaussian distribution centered at $\mathbf{m}_*$. For simplicity, we omit the dependence of $\mathbf{A}_j(\mathbf{m})$ and $\overline{\mathbf{u}}_{i,j}(\mathbf{m})$ on $\mathbf{m}$. A necessary component in this process is computing the Hessian $\mathbf{H}_{\text{post}} = -\nabla^2_{\mathbf{m}} \log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$, which is given by

$$\begin{aligned}\mathbf{H}_{\text{post}} &= \mathbf{H}_{\text{like}} + \Gamma^{-1}_{\text{prior}} \\ &= \sum_{i}^{n_{\text{src}}} \sum_{j}^{n_{\text{freq}}} \lambda^2 \mathbf{G}^\top_{i,j} \mathbf{G}_{i,j} - \mathbf{S}^\top_{i,j} \left(\mathbf{P}^\top \Gamma^{-1}_{\text{noise}} \mathbf{P} + \lambda^2 \mathbf{A}_j^\top \mathbf{A}_j\right)^{-1} \mathbf{S}_{i,j} + \Gamma^{-1}_{\text{prior}},\end{aligned} \tag{26}$$

where

$$\mathbf{S}_{i,j} = \lambda^2(\nabla_{\mathbf{m}}\mathbf{A}_j^\top)\,(\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j}) + \lambda^2\mathbf{A}_j^\top\mathbf{G}_{i,j}, \tag{27}$$

and the optimal wavefield $\overline{\mathbf{u}}_{i,j}$ is computed by equation 6.

The full Hessian $\mathbf{H}_{\text{post}}$ is a dense $n_{\text{grid}} \times n_{\text{grid}}$ matrix, which is prohibitive to construct explicitly when n_{grid} is large, even in the two-dimensional setting. On the other hand, it is also prohibitive if we formulate the Hessian $\mathbf{H}_{\text{post}}$ as a traditional implicit operator, which requires $2 \times n_{\text{src}} \times n_{\text{freq}}$ PDE solves to compute each matrix-vector product $\mathbf{H}_{\text{post}}\overline{\mathbf{m}}$ with any vector $\overline{\mathbf{m}}$, according to the expression in equation 26. Since the posterior covariance matrix is the inverse of the Hessian, we need to invert the square root of the Hessian operator in order to generate random samples. With the implicit Hessian operator, we would need to employ an iterative solver such as LSQR or CG (Golub and Van Loan, 2012), and the total number of PDE solves required is therefore proportional to $2 \times n_{\text{smp}} \times n_{\text{iter}} \times n_{\text{src}} \times n_{\text{freq}}$. As a result, this type of approach requires an extremely large computational cost when drawing sufficiently many samples. As a remedy, we exploit the structure of the Hessian matrix $\mathbf{H}_{\text{post}}$ to find an approximation that can be constructed and applied in a computationally efficient manner.

To exploit the structure of the Hessian matrix $\mathbf{H}_{\text{post}}$, we will focus on the Hessian matrix $\mathbf{H}_{\text{like}}$ of the likelihood term, as we already have discussed the matrix Γ_{prior} in the previous section. Based on equations 26 and 27, $\mathbf{H}_{\text{like}}$ consists of three components — the matrices $\mathbf{P}^\top\Gamma_{\text{noise}}^{-1}\mathbf{P} + \lambda^2\mathbf{A}_j^\top\mathbf{A}_j$, $(\nabla_{\mathbf{m}}\mathbf{A}_j^\top)(\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j})$, and $\mathbf{G}_{i,j}$. We first consider the Jacobian $(\nabla_{\mathbf{m}}\mathbf{A}_j^\top)(\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j})$ and specifically the PDE misfit $\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j}$. When the PDE misfit is approximately zero, this overall term is also expected to be small. As the MAP estimate $\mathbf{m}_*$ simultaneously minimizes the data misfit, PDE misfit, and model penalty term, the PDE

misfit is expected to be small when model and observational errors are not too large. Thus, we obtain a good approximation $\tilde{\mathbf{H}}_{\text{like}}$ of the full Hessian $\mathbf{H}_{\text{like}}$, which corresponds to the Gauss-Newton Hessian derived by van Leeuwen and Herrmann (2015):

$$\tilde{\mathbf{H}}_{\text{like}} = \sum_{i}^{n_{\text{src}}} \sum_{j}^{n_{\text{freq}}} \lambda^2 \mathbf{G}_{i,j}^{\top} \mathbf{G}_{i,j} - \lambda^4 \mathbf{G}_{i,j}^{\top} \mathbf{A}_j (\mathbf{P}^{\top} \Gamma_{\text{noise}}^{-1} \mathbf{P} + \lambda^2 \mathbf{A}_j^{\top} \mathbf{A}_j)^{-1} \mathbf{A}_j^{\top} \mathbf{G}_{i,j}. \tag{28}$$

Consequently, we can use the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ to construct the Gaussian distribution that approximates the posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$:

$$\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \approx \rho_{\text{Gauss}}(\mathbf{m}) = \mathcal{N}(\mathbf{m}_*, \tilde{\mathbf{H}}_{\text{post}}^{-1}) = \mathcal{N}(\mathbf{m}_*, (\tilde{\mathbf{H}}_{\text{like}} + \Gamma_{\text{prior}}^{-1})^{-1}). \tag{29}$$

The Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ has a compact expression derived from the Sherman–Morrison–Woodbury formula (Golub and Van Loan, 2012):

$$\tilde{\mathbf{H}}_{\text{like}} = \sum_{i}^{n_{\text{src}}} \sum_{j}^{n_{\text{freq}}} \mathbf{G}_{i,j}^{\top} \mathbf{A}_j^{-\top} \mathbf{P}^{\top} (\Gamma_{\text{noise}} + \frac{1}{\lambda^2} \mathbf{P} \mathbf{A}_j^{-1} \mathbf{A}_j^{-\top} \mathbf{P}^{\top})^{-1} \mathbf{P} \mathbf{A}_j^{-1} \mathbf{G}_{i,j}. \tag{30}$$

We shall see that this expression provides a factored formulation to implicitly construct the Gauss-Newton Hessian, which does not require any additional PDE solves to compute matrix-vector products. In order to construct the implicit Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$, three matrices are necessary — $\mathbf{A}_j^{-\top} \mathbf{P}^{\top} \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{rcv}}}$, $\mathbf{G}_{i,j} \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{grid}}}$, and $\Gamma_{\text{noise}} + \frac{1}{\lambda^2} \mathbf{P} \mathbf{A}_j^{-1} \mathbf{A}_j^{-\top} \mathbf{P}^{\top} \in \mathbb{C}^{n_{\text{rcv}} \times n_{\text{rcv}}}$. For each frequency, constructing the matrix $\mathbf{A}_j^{-\top} \mathbf{P}^{\top}$ requires n_{rcv} PDE solves. As described in the previous section, the matrix $\mathbf{G}_{i,j}$ is sparse and driven by the corresponding wavefields $\overline{\mathbf{u}}_{i,j}$, whose computational cost approximately equals to one PDE solve for each source and each frequency. The computational complexity of inverting the matrix $\Gamma_{\text{noise}} + \frac{1}{\lambda^2} \mathbf{P} \mathbf{A}_j^{-1} \mathbf{A}_j^{-\top} \mathbf{P}^{\top}$ is $\mathcal{O}(n_{\text{rcv}}^3)$. Since $n_{\text{rcv}} \ll n_{\text{grid}}$, inverting this matrix is much cheaper than solving a PDE. Thus, to construct the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$, we only need to solve $n_{\text{freq}} \times (n_{\text{src}} + n_{\text{rcv}})$ PDEs. With the computed matrix $\mathbf{A}_j^{-\top} \mathbf{P}^{\top}$ and wavefield $\overline{\mathbf{u}}_{i,j}$, the action of the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ on any vector $\overline{\mathbf{m}}$ can

be performed with several efficient matrix-vector multiplications related to the matrices $\mathbf{A}_j^{-\top}\mathbf{P}^\top$, $\mathbf{G}_{i,j}$ and $\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top$. Compared to the conventional approach, this new factored formulation of the implicit operator does not require additional PDE solves to compute a matrix-vector multiplication once it is constructed. In general, we have that $n_{\text{rcv}} \ll n_{\text{smp}} \times n_{\text{iter}}$ and using this implicit Gauss-Newton Hessian operator to draw n_{smp} samples is significantly cheaper than the conventional approach. Another advantage of using this operator arises from the fact that the computations of the necessary matrices corresponding to different frequencies are independent from each other. As a result, we can compute and store these matrices in parallel for different frequencies, allowing us to speed up our computations in a distributed computing environment. Furthermore, the expression in equation 30 provides a natural decomposition of the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ as follows:

$$\begin{aligned}
&\tilde{\mathbf{H}}_{\text{like}} = \mathbf{R}_{\text{like}}^\top \mathbf{R}_{\text{like}}, \\
&\mathbf{R}_{\text{like}} = \begin{bmatrix}
(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_1^{-1}\mathbf{A}_1^{-\top}\mathbf{P}^\top)^{-\frac{1}{2}}\mathbf{P}\mathbf{A}_1^{-1}\mathbf{G}_{1,1} \\
\cdots \\
(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top)^{-\frac{1}{2}}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{G}_{i,j} \\
\cdots \\
(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_{n_{\text{freq}}}^{-1}\mathbf{A}_{n_{\text{freq}}}^{-\top}\mathbf{P}^\top)^{-\frac{1}{2}}\mathbf{P}\mathbf{A}_{n_{\text{freq}}}^{-1}\mathbf{G}_{n_{\text{src}},n_{\text{freq}}}
\end{bmatrix}.
\end{aligned} \tag{31}$$

Similarly to the factored formulation of the implicit Gauss-Newton Hessian, we can construct the factor $\mathbf{R}_{\text{like}}$ as an implicit operator once we have computed the matrix $\mathbf{A}_j^{-\top}\mathbf{P}^\top$ and wavefield $\overline{\mathbf{u}}_{i,j}$. We will use this implicit operator $\mathbf{R}_{\text{like}}$ for the sampling method introduced in the next subsection.

Steps 5 and 6: sampling the Gaussian PDFs

The covariance matrix $\tilde{\mathbf{H}}_{\text{post}}^{-1}$ is a dense $n_{\text{grid}} \times n_{\text{grid}}$ matrix, and the construction of its Cholesky factorization involves $\mathcal{O}(n_{\text{grid}}^3)$ operations. Both of these facts prohibit us from applying the Cholesky factorization method to sample the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ for large-scale problems. Here we propose to use the so-called optimization-driven Gaussian simulators (Papandreou and Yuille, 2010; Orieux et al., 2012) or the randomize-then-optimize (RTO) method (Solonen et al., 2014) to sample the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$. This method belongs to the classical bootstrap method (Efron, 1981, 1992; Kitanidis, 1995), and it does not require the explicit formulation of the Hessian matrix as well as the expensive Cholesky factorization. To outline this method, we first use equation 29 to divide $\tilde{\mathbf{H}}_{\text{post}}$ into $\tilde{\mathbf{H}}_{\text{like}}$ and $\Gamma_{\text{prior}}^{-1}$. The Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ has the factorization in equation 31, and we can also compute the Cholesky factorization of the prior covariance matrix $\Gamma_{\text{prior}} = \mathbf{R}_{\text{prior}}^{\top}\mathbf{R}_{\text{prior}}$ with an upper-triangular matrix $\mathbf{R}_{\text{prior}}$. Substituting these two factorizations into equation 29, we can rewrite the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ as follows:

$$\rho_{\text{Gauss}}(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\|\mathbf{R}_{\text{like}}\mathbf{m} - \mathbf{R}_{\text{like}}\mathbf{m}_*\|^2 - \frac{1}{2}\|\mathbf{R}_{\text{prior}}^{-\top}\mathbf{m} - \mathbf{R}_{\text{prior}}^{-\top}\mathbf{m}_*\|^2\right). \tag{32}$$

Papandreou and Yuille (2010) and Solonen et al. (2014) noted that independent realizations from the distribution in equation 32 can be computed by repeatedly solving the following linear data-fitting optimization problem:

$$\begin{aligned} \mathbf{m}_{\text{s}} &= \arg\min_{\mathbf{m}} \|\mathbf{R}_{\text{like}}\mathbf{m} - \mathbf{R}_{\text{like}}\mathbf{m}_* - \mathbf{r}_{\text{like}}\|^2 + \left\|\mathbf{R}_{\text{prior}}^{-\top}\mathbf{m} - \mathbf{R}_{\text{prior}}^{-\top}\mathbf{m}_* - \mathbf{r}_{\text{prior}}\right\|^2, \\ \mathbf{r}_{\text{like}} &\sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{data}} \times n_{\text{data}}}), \\ \mathbf{r}_{\text{prior}} &\sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{grid}} \times n_{\text{grid}}}). \end{aligned} \tag{33}$$

This optimization problem can be solved by iterative solvers such as LSQR and PCG, which does not require the explicit expression for the matrix $\mathbf{R}_{\text{like}}$ but merely an operator that

can compute the matrix-vector product. As a result, we can use our implicit formulation of $\mathbf{R}_{\text{like}}$ in equation 31 to solve the optimization problem in equation 33 and draw samples from the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$. The pseudo code of the RTO method to draw samples from the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ is shown in Algorithm 2, which realizes step 5 in Algorithm 1. Because the overall sampling strategy consists of the Gaussian approximation and the RTO method, we will refer to the proposed method as GARTO in the rest of the paper.

Algorithm 2 Sample $\rho_{\text{Gauss}}(\mathbf{m})$ by the RTO method

1. Start with the MAP estimate $\mathbf{m}_*$, covariance matrices Γ_{noise} and Γ_{prior};
2. Formulate the operator $\mathbf{R}_{\text{like}}(\mathbf{m}_*)$ by equation 31, and compute the Cholesky factorization of $\Gamma_{\text{prior}} = \mathbf{R}_{\text{prior}}^{\top}\mathbf{R}_{\text{prior}}$;
3. for s = 1:n_{smp}
4. Generate $\mathbf{r}_{\text{prior}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{grid}} \times n_{\text{grid}}})$ and $\mathbf{r}_{\text{like}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{data}} \times n_{\text{data}}})$;
5. Solve $\mathbf{m}_{\text{s}} = \arg\min_{\mathbf{m}} \|\mathbf{R}_{\text{like}}\mathbf{m} - \mathbf{R}_{\text{like}}\mathbf{m}_* - \mathbf{r}_{\text{like}}\|^2 + \left\|\mathbf{R}_{\text{prior}}^{-\top}\mathbf{m} - \mathbf{R}_{\text{prior}}^{-\top}\mathbf{m}_* - \mathbf{r}_{\text{prior}}\right\|^2$;
6. end

A benchmark method: the randomized maximum likelihood method

We have proposed a computationally efficient algorithm — GARTO — that can approximately sample the target distribution in equation 22 without additional PDE solves once the MAP estimate and the Gauss-Newton Hessian operator are computed. However, due to the loss of accuracy caused by the Gaussian approximation, it is important to investigate the accuracy of the GARTO method by comparing it to a benchmark method that can sample the target distribution in equation 22 regardless of the computational cost. The randomized maximum likelihood (RML) (Chen and Oliver, 2012) method is a viable candidate as a benchmark

because it has the capability to draw samples that are good approximations of those from the target distribution for weakly nonlinear problems (Bardsley et al., 2014, 2015). Indeed, as previously discussed, the target distribution with weak PDE-constraints that the GARTO method aims to sample is less prone to the nonlinearity with a carefully selected λ.

To draw a sample, the RML method performs a bootstraping-like method (Efron, 1981, 1992) that first samples the data and prior model and then computes the resulting MAP. More precisely, in order to draw a sample from the target distribution in equation 22, the RML method solves the following nonlinear optimization problem:

$$\begin{aligned}\mathbf{m}_s = \arg\min_{\mathbf{m}} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \Big(& \frac{1}{2} \| \Gamma_{\text{noise}}^{-\frac{1}{2}} \mathbf{P} \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \Gamma_{\text{noise}}^{-\frac{1}{2}} \mathbf{d}_{i,j} - \mathbf{r}_{i,j}^{(1)} \|^2 \\ & + \frac{1}{2} \| \lambda \mathbf{A}_j(\mathbf{m}) \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \lambda \mathbf{q}_{i,j} - \mathbf{r}_{i,j}^{(2)} \|^2 \Big) \\ & + \frac{1}{2} \| \Gamma_{\text{prior}}^{-\frac{1}{2}} \mathbf{m} - \Gamma_{\text{prior}}^{-\frac{1}{2}} \tilde{\mathbf{m}} - \mathbf{r}^{(3)} \|^2,\end{aligned} \tag{34}$$

where the vectors $\mathbf{r}_{i,j}^{(1)}$, $\mathbf{r}_{i,j}^{(2)}$, and $\mathbf{r}^{(3)}$ are random realizations from the standard norm distribution $\mathcal{N}(0, \mathbf{I})$. We refer interested readers to Chen and Oliver (2012) for more details about the RML method. As a result of this approach, the computational cost of drawing one sample by the RML method is approximately equivalent to solving one FWI problem, which is significantly more expensive than the GARTO method. Therefore, we are only able to conduct an comparison with the RML method on a small-scale problem in the following section.

NUMERICAL EXPERIMENTS

Influence of the penalty parameter λ

The feasibility of the proposed Bayesian framework relies on the accuracy of the approximations in equations 18 and 29, both of which depend on the selection of λ. To get a better understanding of the influence of this parameter on these approximations, we will work with a relatively simple 2D velocity model parameterized by a single varying parameter v_0—i.e., the velocity is given by $v(z) = v_0 + 0.75z\,\mathrm{m/s}$ and z increases with vertical depth. We simulate data with a single source and single frequency for $v_0 = 2000\,\mathrm{m/s}$ using a grid spacing of 25 m. The frequency of the data is 5 Hz, which is suitable for avoiding numerical dispersion on this particular grid spacing. We place our source at (z, x) coordinates $(50\,\mathrm{m}, 50\,\mathrm{m})$ and record the data at 200 receivers located at the depth of 50 m with a sampling interval of 25 m. We do not simulate the free surface in this example. After the simulation, we add 10% Gaussian noise to the observed data. Because the prior distribution is independent of λ, we only investigate its influence on the negative log-likelihood of the associated PDFs in this experiment. We abbreviate the negative log-likelihood with "NLL".

Figure 4 shows the NLL for the reduced approach (cf. equation 3) as well as for the penalty approach (cf. equation 11) for various values of λ as a function of v_0. As discussed previously, we select values of λ^2 proportional to the largest eigenvalue μ_1 of the matrix $\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma_{\text{noise}}^{-1}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}$, i.e., $\lambda^2 = 10^{-10}\mu_1$, $10^{-6}\mu_1$, $10^{-4}\mu_1$, $10^{-2}\mu_1$, $10^{0}\mu_1$, and $10^{2}\mu_1$. From this figure, we observe that, when λ is large, i.e., $\lambda = 10^2\mu_1$ and $10^0\mu_1$, the NLL exhibits several local minima, and, as λ increases, it converges to the reduced approach formulation, as expected. We also note that for small λ, i.e., $\lambda = 10^{-10}\mu_1$, the resulting NLL is monotonically decreasing, which is due to the fact that the determinant term in

equation 15 dominates the NLL. Additionally, in between these two extreme values for λ (i.e., when $\lambda = 10^{-6}\mu_1$, $10^{-4}\mu_1$, and $10^{-2}\mu_1$), the resulting NLLs are unimodal. This observation implies that with a carefully selected λ, the posterior distribution with weak PDE-constraints potentially contains less local maxima.

[Figure 4 about here.]

To investigate the influence of the parameter λ on the accuracy of the approximations in equations 18 and 29, we plot in Figure 5 the NLL corresponding to the true (cf. equation 11) $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$, its approximation neglecting the determinant term $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ (cf. equation 18), and the Gaussian approximation $\rho_{\text{Gauss}}(\mathbf{m})$ (cf. equation 29) for various values of λ. For simplicity, we refer to these three different functions as ψ_1, ψ_2, and ψ_3, respectively. From Figure 5a, we observe that when $\lambda = 10^{-10}\mu_1$, ψ_2 fails to adequately approximate ψ_1—i.e., the approximation in equation 18 fails — because the determinant term in equation 11 dominates the negative logarithm function. As λ increases, the determinant term becomes negligible, and ψ_2 becomes a reasonable approximation of ψ_1, as shown in Figures 5b to 5f. However, among the five various selections of λ, only when $\lambda^2 = 10^{-2}\mu_1$ does ψ_3 adequately approximate ψ_2. This occurs because when λ is relatively large, ψ_2 contains a number of nonoptimal local minima, resulting in ψ_2 being poorly modeled by its Gaussian approximation. Additionally, when $\lambda < 10^{-2}\mu_1$, the term $\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j}$ in equation 27 is not negligible, resulting in the Gauss-Newton Hessian being a poor approximation of the full Hessian. These results imply that the proposed criterion — i.e., $\lambda^2 = 10^{-2}\mu_1$ — provides a reasonable choice for the parameter λ, which can simultaneously satisfy both the approximations in equations 18 and 29. As a result, the corresponding posterior distribution is less prone to the local maxima and can be appropriately approximated by the Gaussian

distribution in equation 29, which ensures the feasibility of the proposed framework.

[Figure 5 about here.]

A 2D layered model

In this section, we develop an example to compare the accuracy of the statistical quantities produced by the GARTO method relative to those produced by the RML method. Considering the large computational cost required by the RML method, we use a small three-layer model as our baseline model, as shown in Figure 6a, whose size is 1500 m $\times$ 3000 m. We discretize the velocity model with a grid spacing of 50 m, yielding 1800 unknown parameters. At the surface, we place sixty sources and receivers with a horizontal sampling interval of 50 m. To control the computational cost and ensure the feasibility of the RML method, we use a Ricker wavelet centered at 6 Hz to simulate the observed data with frequencies of 5, 6, and 7 Hz. We use an absorbing boundary condition at the surface, so that no surface-related multiples are included. After the simulation, we add 15% Gaussian noise to the observed data resulting in a covariance matrix of $\Gamma_{\text{noise}} = 175^2\mathbf{I}$. To set up the prior distribution, we first construct the monotonically increasing model shown in Figure 6b as the prior mean model, which corresponds to the well-known observation that the velocity of the Earth, in general, increases with depth. Following the strategy used by Bardsley et al. (2015) that ensures the prior covariance matrix is well-conditioned, we construct the prior covariance matrix by selecting $a = 0.1\,\text{km}^2/\text{s}^2$, $b = 0.65$, and $c = 0.01\,\text{km}^2/\text{s}^2$, resulting in a prior standard deviation of 0.33 km/s that meets the maximal difference between the true and prior mean models. We select the penalty parameter λ for each frequency according to the proposed selection criterion, resulting in $\lambda = 13,\ 12,$ and 11, respectively. To compute

reliable statistical quantities and while also ensuring that the RML method terminates in a reasonable amount of time, we use both the GARTO and the RML methods to generate $10,000$ samples from the posterior distribution. Based on our experience, generating $10,000$ samples is sufficient for both methods to produce stable results in this case.

[Figure 6 about here.]

Given that the computational overhead introduced by these methods is negligible compared to the number of PDEs that need to be solved, we use the number of PDEs as our performance metric. To generate $10,000$ samples, the RML method needs to solve $10,000$ nonlinear optimization problems. To solve each nonlinear optimization problem, following van Leeuwen et al. (2014), we stop the optimization when the relative misfit difference between two iterations drops below 10^{-3} resulting in 100 l-BFGS iterations. During each l-BFGS iteration, we have to solve $2 \times 3 \times 60$ PDEs to compute the objective and gradient. As a result, the RML method requires $360 \times 100 \times 10000 = 360$ million PDE solves to draw the $10,000$ samples. Contrary to the RML method, the GARTO method requires significantly fewer PDE solves. The total number of PDE solves required by the GARTO method is $36,360$, which includes $36,000$ PDE solves to find the MAP estimate and another 360 PDE solves to construct the Gauss-Newton operator. With the MAP estimate and the Gauss-Newton operator in hand, the GARTO method uses the RTO approach to sample the $10,000$ samples without involving any additional PDE solves as we explained above. Therefore, neglecting the costs associated with solving the least-squares systems, compared to the RML method, the GARTO method requires only $\frac{1}{10000}$th the computational budget to generate the same number of samples.

In addition to the significant computational speedups introduced by the GARTO method,

the GARTO method also generates samples that yield similar statistical quantities as those produced by the RML method. For instance, the posterior mean models obtained by the two methods are shown in Figure 7. From Figures 7a and 7b, we observe that aside from some slight differences in the second and third layers, the two results are roughly identical, with an average pointwise relative difference of 1.5%. Both results provide acceptable reconstructions of the original velocity model, despite the fact that the data are noisy. We also use the two posterior mean models to compute the predicted data and compare them with the observed data in Figures 9a and 9b, which faithfully match the observed data, aside from the noise. The pointwise standard deviations computed from both methods, shown in Figure 8, result in estimates that are visually quite similar throughout the entire model. The average relative difference between the standard deviations produced by both methods is 6%, an acceptable error level, and results from the Gaussian approximation in GARTO. Figure 10 depicts the 95% confidence intervals obtained with the two methods at three vertical cross-sections through the model. The confidence intervals obtained by the two methods are virtually identical, as are the pointwise marginal distributions shown in Figure 11. All these results illustrate that, compared to the RML method, the proposed GARTO method can produce accurate estimates of statistical quantities, while requiring a fraction of the computational costs.

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

BG Model

To demonstrate the effectiveness of our method when applied to a more realistic problem, we conduct an experiment on a 2D slice from the 3D BG Compass model shown in Figure 12a. This is a synthetic velocity model created by BG Group, which contains a large degree of variability and has been widely used to evaluate performances of different FWI methods (Li et al., 2012; van Leeuwen and Herrmann, 2013). The model is 2000 m deep and 4500 m wide, with a grid size of 10 m resulting in 92,455 unknown parameters. Following van Leeuwen and Herrmann (2013), we use a Ricker wavelet with a central frequency of 15 Hz to simulate 91 sources at the depth of 50 m with a horizontal sampling interval of 50 m. As before, we do not model the free-surface, so that the data do not contain free-surface related multiples. We place 451 equally spaced receivers at the same depth as the sources to record the data, which contain 30 equally spaced frequency components ranging from 2 Hz to 31 Hz. This results in 1,231,230 observed data values. To mimic a real-world noise level, we corrupt the synthetic observations with 15% random Gaussian noise, leading to $\Gamma_{\text{noise}} = 46^2\mathbf{I}$. To construct the prior distribution, we first set its mean model (Figure 12b) by smoothing the true model followed by horizontal averaging. Second, we construct the covariance matrix of the prior distribution utilizing the fact that we have the true 3D model, which contains 1800 2D slices. We regard these 2D slices as 1800 2D samples, from which we compute the pointwise standard deviation. After horizontal averaging, we obtain the prior standard deviation shown in Figure 12c. With the prior standard deviation, we select $a = 0.02\text{km}^2/\text{s}^2$ and $b = 19.5$ to construct a well-conditioned covariance matrix with a correlation length of 60 m (Bardsley et al., 2015). The parameter c in equation 23 is calculated according to the

standard deviation and the parameters a and b. Finally, we compute the penalty parameter λ for each frequency (listed in Table 1) according to the criterion introduced earlier in order to obtain a posterior distribution that is less prone to local maxima. Considering both the computational resources and the accuracy of the inverted statistical quantities, we will use the GARTO method to draw 2000 samples according to Bardsley et al. (2015). Compared to the previous example, which had a much simpler model, this model contains a significantly larger number of unknown parameters and data points and is a better proxy for real-world problems.

[Figure 12 about here.]

[Table 1 about here.]

During the inversion, we use 200 l-BFGS iterations to compute the MAP estimate plotted in Figure 13a with the same stopping criterion as in the previous example. Compared to the true model, we observe that most of the observable velocity structures are reconstructed in the MAP estimate, aside from some small measurement noise related imprints near the boundary of the model. We also observe that the shallow parts ($z \leq 1400\,\mathrm{m}$) of the BG model, where the turning waves exist (for a maximal offset of 4500 m (Virieux and Operto, 2009)) are better recovered relative to the deep parts ($z \geq 1400\,\mathrm{m}$), where the only received energy arises from reflected waves. This implies that the portion of the data corresponding to the turning waves is more informative to the velocity reconstruction than that of the reflected waves, which is a well-known observation in seismic inversions.

After obtaining the MAP estimate, we construct the Gauss-Newton Hessian operator and apply the RTO method to generate 2000 samples. This allows us to compute the

posterior standard deviation (Figure 13b) and compare it with the prior standard deviation (Figure 12c). To have a better understanding of the information that the data introduce, we also compute the relative difference $\frac{\mathrm{STD}_{\mathrm{post}}(m_k)-\mathrm{STD}_{\mathrm{prior}}(m_k)}{|\mathrm{STD}_{\mathrm{prior}}(m_k)|}$ between the posterior and prior standard deviations (Figure 13c). In the shallow parts of the model ($z \leq 1400\,\mathrm{m}$), the posterior standard deviation decreases between $10\% - 50\%$ compared to the prior standard deviation, while in the deep parts ($z \geq 1400\,\mathrm{m}$), the reduction in standard deviation is smaller than 3%. This observation is consistent with the notion that, owing to the amplitude decay of propagating waves, the data place more constraints on the velocity variations in the shallow parts of the model compared to the deep parts. Additionally, considering the areas where the turning waves and the reflected waves exist, this observation also implies that the portion of the data corresponding to the turning waves can reduce more uncertainties in the recovered velocity compared to the reflected waves. To further evaluate this inversion result, we compare the prior model, the MAP estimate of the posterior, and the true velocity at three different cross sections in Figure 14 (i.e., $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$), in which we also plot the prior standard deviation (red patch) and posterior standard deviation (blue patch). In the shallow region of the model, the MAP estimates closely match the true model, while they diverge in the deeper region in a more pronounced manner. This is again consistent with the notion that the data are more informative in the shallow area of the model compared to the deeper areas.

[Figure 13 about here.]

[Figure 14 about here.]

We also consider the pointwise posterior marginal distribution generated by the posterior and prior distributions to further understand the results of the GARTO method.

Figure 15 compares these distributions at four different locations, $(x, z) = (2240\text{m}, 40\text{m})$, (2240m, 640m), (2240m, 1240m), and (2240m, 1840m). The posterior distribution is more concentrated than the prior distribution in the shallow regions of the model, while in the deep parts, the two distributions are almost identical. Therefore, the recovered velocity in the shallow parts is more reliable than in the deep parts.

[Figure 15 about here.]

To verify our statistical results, we also utilize the RML method as a baseline approach. For this example, drawing a single sample with the RML method using 200 l-BFGS iterations requires at least 1.09 million PDE solves, which is computationally daunting. As a result, we only generate 10 samples via the RML method and compare them to the 95% confidence intervals (i.e., the blue patch) obtained by the GARTO method in Figure 16. From these figures, it is clear that the majority of the 10 samples lie in the blue patch. Moreover, the variation of the ten samples also matches the 95% confidence interval. In this case, we conclude that the estimated confidence intervals are likely accurate approximations of the true confidence intervals.

[Figure 16 about here.]

DISCUSSION

When the underlying model is given by PDE-constraints consisting of multiple experiments, one is forced to make various approximations and tradeoffs in order to develop a computationally tractable procedure. There are a large number of discretized parameters, corresponding to a discretized 2D or 3D function, and one must necessarily avoid having to explicitly construct

dense covariance matrices of the size of the model grid squared, whose construction requires a large number of PDE solves. Moreover, each evaluation of the posterior distribution involves the solution of multiple PDEs, a computationally expensive affair. Methods that require tens or hundreds of thousands of such posterior evaluations, such as McMC-type methods, do not scale to realistically sized problems. The original PDE-constrained formulation of Bayesian inference, while theoretically convenient, results in a posterior that cannot be appropriately approximated by a Gaussian distribution, whereas the relaxation of the exact PDE-constraints results in a posterior that is much more amenable to Gaussian approximation. Ideally, one would like to parameterize the distribution over the joint model/field variables $(\mathbf{m}, \mathbf{u})$ and estimate the variances accordingly. Hidden in this notation, however, is the fact that in real-world problems, we have hundreds or potentially thousands of source experiments, each of which corresponds to a different $\mathbf{u}$. Storing all of these fields and updating them explicitly becomes prohibitive memory-wise, even on a large cluster. As a result, our approach aims to keep the benefits of relaxing the PDE-constraints while still resulting in a computationally feasible algorithm.

The initial results illustrate that with the specific selection of λ, i.e., $\lambda^2 = 0.01\mu_1$, the relaxed formulation of the posterior PDF is less prone to local maxima, which enables us to analyze it via an arguably accurate Gaussian approximation. Once we can manipulate the covariance matrix of the Gaussian approximation through the implicit PDE-free formulation, we have access to a variety of statistical quantities, including the pointwise standard deviation, the confidence intervals, and the pointwise marginal distributions, in a tractable manner. We can use these quantities to assess the uncertainties of our inversion results and to identify the areas with high/low reliability in the model. This information is important and useful for the subsequent processing and interpretation. A straightforward example is that it allows us

to assess the reliability of the visible image features obtained by the subsequent procedure of imaging, as in Ely et al. (2017).

While the initial results are promising, some aspects of the proposed framework warrant further investigation. Although numerical examples illustrate the feasibility of the proposed method for the case with the selection of $\lambda^2 = 0.01\mu_1$, it does not guarantee the feasibility of the proposed approach for PDFs arising from other choices of λ. For other selections of λ, the posterior PDFs can significantly differ from a Gaussian PDF, which makes approximately sampling them challenging. Potentially other sampling schemes can be explored for these distributions.

While we have shown the feasibility of the proposed sampling method for 2D problems, the application of the proposed sampling method to 3D problems is still challenging from the perspective of memory usage. To satisfy the $\mathcal{O}\big(n_{\text{grid}} \times (n_{\text{rcv}} + n_{\text{src}}) \times n_{\text{freq}}\big)$ storage requirement for formulating the implicit Gauss-Newton Hessian operator, a large high-performance cluster with enough computational workers and memory is needed to store all of the necessary matrices in parallel.

CONCLUSIONS

We have described a new Bayesian framework for partial-differential-equation-constrained inverse problems. In our new framework, we introduce the field variables as auxiliary variables and relax the partial-differential-equation-constraint. By integrating out the field parameters, we avoid having to store and update the high number of field parameters, and exploit the fact that the new distribution is Gaussian in the field variables for every fixed model. We propose a method for selecting an appropriate penalty parameter such that the new posterior

distribution can be approximated by a Gaussian distribution, which is more accurate than the conventional formulation.

We apply the new formulation to the seismic inverse problems and derive each component of the general framework. For this specific application, we use a partial-differential-equation-free Gauss-Newton Hessian to formulate the Gaussian approximation of the posterior distribution. We also illustrate that with this Gauss-Newton Hessian, the Gaussian approximation can be sampled without the requirement of the explicit formulation of the Gauss-Newton Hessian and its Cholesky factorization.

Our proposed method compares favorably to the existing randomized maximum likelihood method for generating different statistics on a simple layered model and the more challenging BG Compass model. Compared to the randomized maximum likelihood method, our method produces results that are quite visually similar, while requiring significantly less partial-differential-equation solves, which are the computational bottleneck in these problems. We expect that these methods will scale reasonably well to 3D models, where traditional methods have only begun to scratch the surface of the problem.

While in this paper we utilize the Gaussian smoothness prior distribution, indeed, the proposed sampling method can also handle other choices of prior distributions, as long as they have sparse covariance matrices. In the future, we can investigate the incorporation of the proposed method with other kinds of prior distributions.

ACKNOWLEDGMENTS

This research was carried out as part of the SINBAD II project with the support of the member organizations of the SINBAD Consortium. The authors wish to acknowledge Dr. James

Gunning from CSIRO for his valuable discussion and suggestions. The authors also wish to acknowledge the SENAI CIMATEC Supercomputing Center for Industrial Innovation, with support from BG Brasil and the Brazilian Authority for Oil, Gas and Biofuels (ANP), for the provision and operation of computational facilities and the commitment to invest in Research & Development.

APPENDIX A

MARGINAL DISTRIBUTION

To derive the marginal PDF $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ in equation 11, we start with the joint PDF

$$\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{u},\mathbf{m})\rho(\mathbf{u}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m}). \tag{A-1}$$

As the noise in the data are assumed to be Gaussian, we have

$$\rho(\mathbf{d}|\mathbf{u},\mathbf{m}) = (2\pi)^{-\frac{n_{\text{data}}}{2}} \det(\Gamma_{\text{noise}})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\mathbf{P}\mathbf{u}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}}\right). \tag{A-2}$$

Substituting equations 10 and A-2 into equation A-1, we arrive at

$$\begin{aligned}\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d}) \propto & (2\pi)^{-\frac{n_{\text{grid}}}{2}} \det\left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\right)^{\frac{1}{2}} \\ & \times \exp\Big(-\frac{1}{2}\big(\lambda^2\|\mathbf{A}(\mathbf{m})\mathbf{u}-\mathbf{q}\|^2 + \|\mathbf{P}\mathbf{u}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big)\Big).\end{aligned} \tag{A-3}$$

Clearly, for fixed $\mathbf{m}$, the PDF $\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})$ with respect to $\mathbf{u}$ is rendered into a Gaussian PDF with a mean value $\overline{\mathbf{u}}(\mathbf{m})$ given by equation 6 and a covariance matrix $\Gamma(\mathbf{m})$ given by

$$\begin{aligned}\Gamma(\mathbf{m})^{-1} = \mathbf{H}(\mathbf{m}) &= -\nabla^2_{\mathbf{u}} \log \rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})\big|_{\mathbf{u}=\overline{\mathbf{u}}(\mathbf{m})} \\ &= \lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m}) + \mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{P}.\end{aligned} \tag{A-4}$$

Given this identity, we can integrate $\mathbf{u}$ out and derive the following closed form expression for the marginal PDF of $\mathbf{m}$:

$$\begin{aligned}\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) &= \int \rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})d\mathbf{u} \\ &= (2\pi)^{\frac{n_{\text{grid}}}{2}} \det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \rho_{\text{post}}\big(\mathbf{m},\overline{\mathbf{u}}(\mathbf{m})|\mathbf{d}\big) \\ &\propto \det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \det\big(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\big)^{\frac{1}{2}} \\ &\quad \times \exp\Big(-\frac{1}{2}\big(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m})-\mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m})-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big)\Big).\end{aligned} \tag{A-5}$$

To derive the limit of the marginal posterior PDF, equation A-5, as $\lambda \to \infty$, we insert equation A-4 into the term $\det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \det\big(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\big)^{\frac{1}{2}}$ and obtain

$$\begin{aligned}&\det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \det\big(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\big)^{\frac{1}{2}} \\ =&\det\left(\mathbf{I}+\frac{1}{\lambda^2}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\right)^{-\frac{1}{2}} \\ =&\det\left(\mathbf{I}+\frac{1}{\lambda^2}\Gamma^{-\frac{1}{2}}_{\text{noise}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma^{-\frac{\top}{2}}_{\text{noise}}\right)^{-\frac{1}{2}}.\end{aligned} \tag{A-6}$$

As a result, we arrive at the following expression for the marginalized posterior distribution

$$\begin{aligned}\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto &\det\left(\mathbf{I}+\frac{1}{\lambda^2}\Gamma^{-\frac{1}{2}}_{\text{noise}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma^{-\frac{\top}{2}}_{\text{noise}}\right)^{-\frac{1}{2}} \\ &\times \exp\left(-\frac{1}{2}\big(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m})-\mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m})-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big)\right).\end{aligned} \tag{A-7}$$

When $\lambda \to \infty$, we have

$$\begin{aligned}&\lim_{\lambda\to\infty} \det\big(\mathbf{I}+\frac{1}{\lambda^2}\Gamma^{-\frac{1}{2}}_{\text{noise}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top\Gamma^{-\frac{\top}{2}}_{\text{noise}}\big)^{-\frac{1}{2}} = 1, \text{ and} \\ &\lim_{\lambda\to\infty} \overline{\mathbf{u}}(\mathbf{m}) = \mathbf{A}(\mathbf{m})^{-1}\mathbf{q},\end{aligned} \tag{A-8}$$

which leads to

$$\lim_{\lambda\to\infty} \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\big(-\frac{1}{2}\|\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big). \tag{A-9}$$

As expected, the marginal PDF of the penalty formulation (cf. equation A-5), converges to the conventional reduced formulation (cf. equation 3) as $\lambda \to \infty$.

REFERENCES

Bardsley, J. M., A. Seppänen, A. Solonen, H. Haario, and J. Kaipio, 2015, Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography: SIAM/ASA Journal on Uncertainty Quantification, **3**, 1136–1158.

Bardsley, J. M., A. Solonen, H. Haario, and M. Laine, 2014, Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems: SIAM Journal on Scientific Computing, **36**, A1895–A1910.

Biegler, L. T., T. F. Coleman, A. Conn, and F. N. Santosa, 2012, Large-scale optimization with applications: Part i: Optimization in inverse problems and design: Springer Science & Business Media, **92**.

Bui-Thanh, T., O. Ghattas, J. Martin, and G. Stadler, 2013, A computational framework for infinite-dimensional Bayesian inverse problems part i: The linearized case, with application to global seismic inversion: SIAM Journal on Scientific Computing, **35**, A2494–A2523.

Chen, Y., and D. S. Oliver, 2012, Ensemble randomized maximum likelihood method as an iterative ensemble smoother: Mathematical Geosciences, **44**, 1–26.

Chen, Z., D. Cheng, W. Feng, and T. Wu, 2013, An optimal 9-point finite difference scheme for the Helmholtz equation with PML: International Journal of Numerical Analysis and Modeling, **10**, 389–410.

Dummit, D. S., and R. M. Foote, 2004, Abstract algebra: Wiley Hoboken, **3**.

Efron, B., 1981, Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods: Biometrika, **68**, 589–599.

———, 1992, Bootstrap methods: another look at the jackknife, *in* Breakthroughs in statistics: Springer, 569–593.

Ely, G., A. Malcolm, and O. V. Poliannikov, 2017, Assessing uncertainties in velocity models

and images with a fast nonlinear uncertainty quantification method: Geophysics, **83(2)**, 1–50.

Epanomeritakis, I., V. Akçelik, O. Ghattas, and J. Bielak, 2008, A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion: Inverse Problems, **24**, 034015.

Esser, E., L. Guasch, T. van Leeuwen, A. Y. Aravkin, and F. J. Herrmann, 2016, Total-variation regularization strategies in full-waveform inversion: arXiv preprint arXiv:1608.06159.

Fang, Z., C. Lee, C. Silva, F. Herrmann, and R. Kuske, 2015, Uncertainty quantification for wavefield reconstruction inversion: Presented at the 77th EAGE Conference and Exhibition 2015.

Fang, Z., C. Y. Lee, C. Da Silva, F. J. Herrmann, and T. Van Leeuwen, 2016, Uncertainty quantification for wavefield reconstruction inversion using a PDE-free semidefinite Hessian and randomize-then-optimize method, *in* SEG Technical Program Expanded Abstracts 2016: Society of Exploration Geophysicists, 1390–1394.

Fisher, M., M. Leutbecher, and G. Kelly, 2005, On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation: Quarterly Journal of the Royal Meteorological Society, **131**, 3235–3246.

Golub, G., and V. Pereyra, 2003, Separable nonlinear least squares: the variable projection method and its applications: Inverse Problems, **19**, R1.

Golub, G. H., and C. F. Van Loan, 2012, Matrix computations: Johns Hopkins University Press, **3**.

Haario, H., M. Laine, A. Mira, and E. Saksman, 2006, Dram: efficient adaptive MCMC: Statistics and Computing, **16**, 339–354.

Haber, E., U. M. Ascher, and D. Oldenburg, 2000, On optimization techniques for solving nonlinear inverse problems: Inverse problems, **16**, 1263.

Hinze, M., R. Pinnau, M. Ulbrich, and S. Ulbrich, 2008, Optimization with PDE constraints: Springer Science & Business Media, **23**.

Kaipio, J., and E. Somersalo, 2006, Statistical and computational inverse problems: Springer Science & Business Media.

Kitanidis, P. K., 1995, Recent advances in geostatistical inference on hydrogeological variables: Reviews of Geophysics, **33**, 1103–1109.

Li, X., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2012, Fast randomized full-waveform inversion with compressive sensing: Geophysics, **77(3)**, A13–A17.

Lieberman, C., K. Willcox, and O. Ghattas, 2010, Parameter and state model reduction for large-scale statistical inverse problems: SIAM Journal on Scientific Computing, **32**, 2523–2542.

Lim, S., 2017, Bayesian inverse problems and seismic inversion: PhD thesis, University of Oxford.

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

Matheron, G., 2012, Estimating and choosing: an essay on probability in practice: Springer Science & Business Media.

Nocedal, J., and S. J. Wright, 2006, Numerical optimization: Springer-Verlag New York.

Orieux, F., O. Féron, and J.-F. Giovannelli, 2012, Sampling high-dimensional Gaussian distributions for general linear inverse problems: Signal Processing Letters, IEEE, **19**, 251–254.

Osypov, K., Y. Yang, A. Fournier, N. Ivanova, R. Bachrach, C. E. Yarman, Y. You, D. Nichols, and M. Woodward, 2013, Model-uncertainty quantification in seismic tomography: method and applications: Geophysical Prospecting, **61**, 1114–1134.

Papandreou, G., and A. L. Yuille, 2010, Gaussian sampling by local perturbations: Advances in Neural Information Processing Systems, 1858–1866.

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**, 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**, 888–901.

Roberts, G. O., J. S. Rosenthal, et al., 2001, Optimal scaling for various Metropolis-Hastings algorithms: Statistical science, **16**, 351–367.

Rue, H., 2001, Fast sampling of Gaussian Markov random fields: Journal of the Royal Statistical Society: Series B (Statistical Methodology), **63**, 325–338.

Sambridge, M., K. Gallagher, A. Jackson, and P. Rickwood, 2006, Trans-dimensional inverse problems, model comparison and the evidence: Geophysical Journal International, **167**, 528–542.

Solonen, A., A. Bibov, J. M. Bardsley, and H. Haario, 2014, Optimization-based sampling in ensemble Kalman filtering: International Journal for Uncertainty Quantification, **4**.

Stuart, G., W. Yang, S. Minkoff, and F. Pereira, 2016, A two-stage Markov chain Monte Carlo method for velocity estimation and uncertainty quantification, *in* SEG Technical Program Expanded Abstracts 2016: Society of Exploration Geophysicists, 3682–3687.

Tarantola, A., and B. Valette, 1982a, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics, **20**, 219–232.

———, 1982b, Inverse problems = quest for information: Journal of Geophysics, **50**, 159–170.

van Leeuwen, T., A. Y. Aravkin, and F. J. Herrmann, 2014, Comment on: "application of the variable projection scheme for frequency-domain full-waveform inversion" (M. Li, J. Rickett, and A. Abubakar, Geophysics, 78, no. 6, R249–R257): Geophysics, **79(3)**, X11–X17.

van Leeuwen, T., and F. J. Herrmann, 2013, Fast waveform inversion without source-encoding: Geophysical Prospecting, **61**, 10–19.

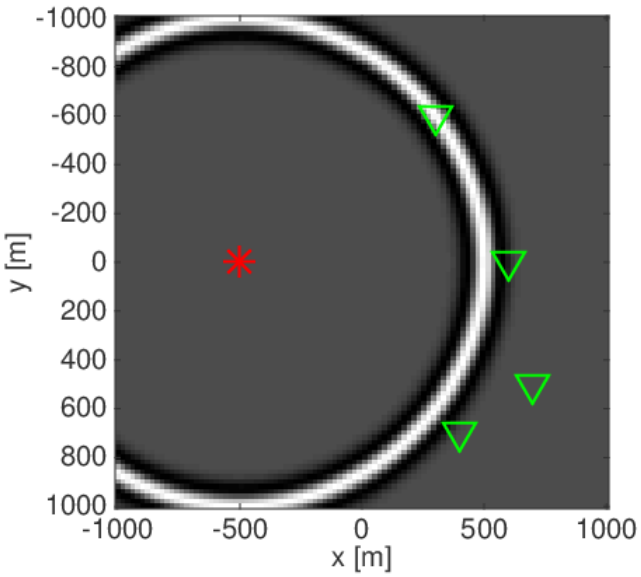
———, 2013, Mitigating local minima in full-waveform inversion by expanding the search space: Geophysical Journal International, **195**, 661–667.

———, 2015, A penalty method for PDE-constrained optimization in inverse problems: Inverse Problems, **32**, 015007.

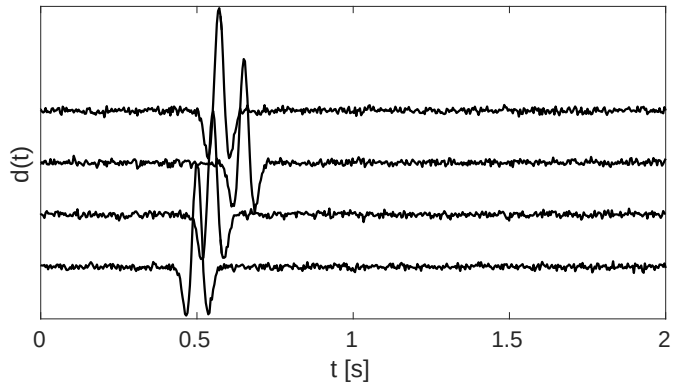
Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74(6)**, WCC1–WCC26.

Zhu, H., S. Li, S. Fomel, G. Stadler, and O. Ghattas, 2016, A Bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration: Geophysics, **81(5)**, R307–R323.

LIST OF FIGURES

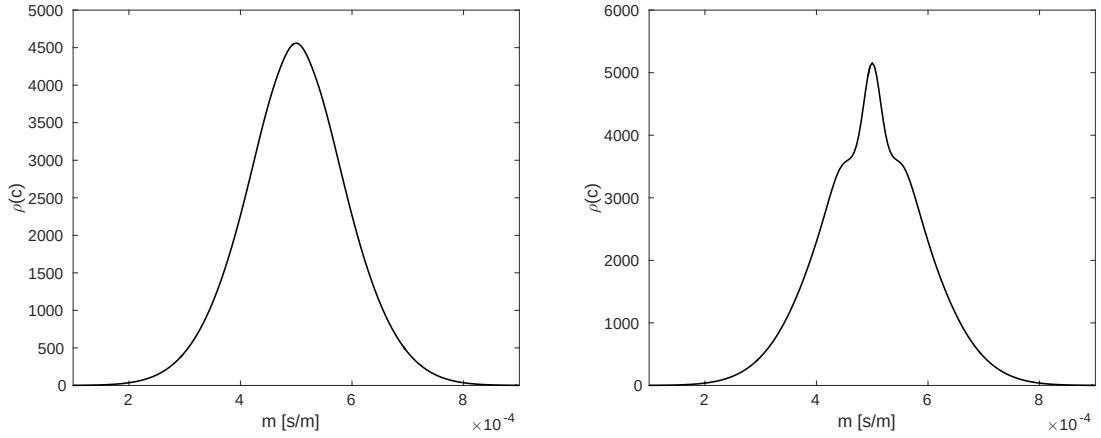


(a) Snap shot



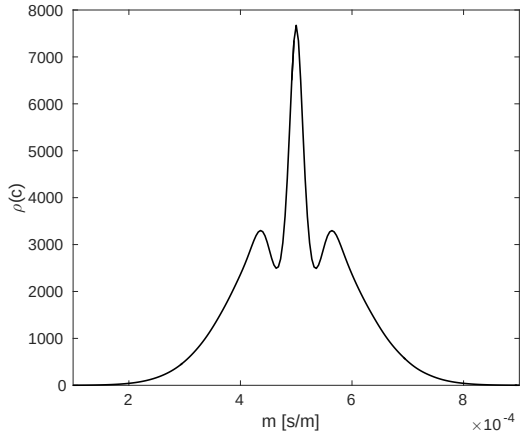
(b) Data

Figure 1: (a) Snapshot of the time-domain wavefield generated by a single source ($*$); (b) recorded time-domain data at four receiver locations (∇).

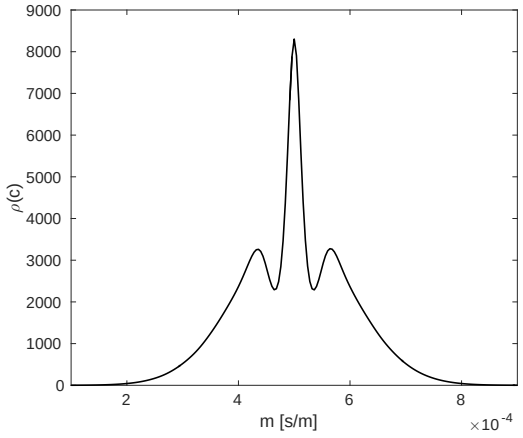


(a) $\lambda = 10$

(b) $\lambda = 50$

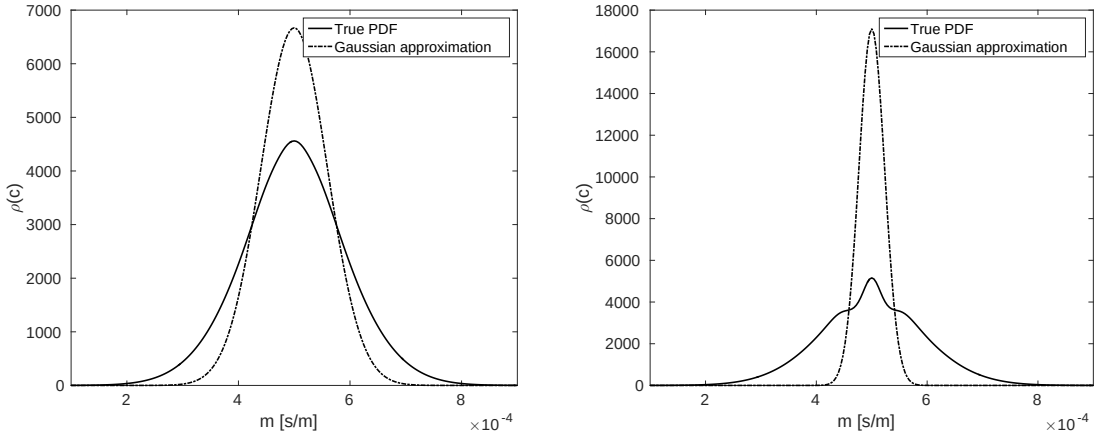


(c) $\lambda = 250$



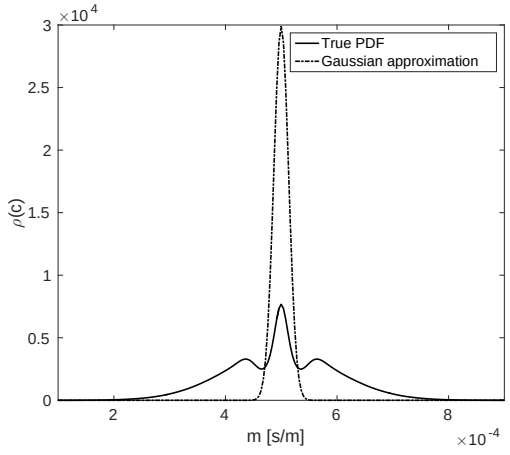
(d) Reduced

Figure 2: The four posterior PDFs corresponding to the penalty formulations with $\lambda = 10$ (a), 50 (b), and 250 (c), and the reduced formulation (d).

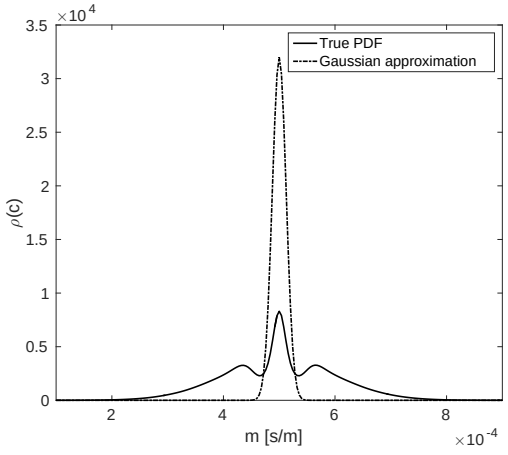


(a) $\lambda = 10$

(b) $\lambda = 50$



(c) $\lambda = 250$



(d) Reduced

Figure 3: The Gaussian approximations of the four posterior PDFs associated with the model in equation 14.

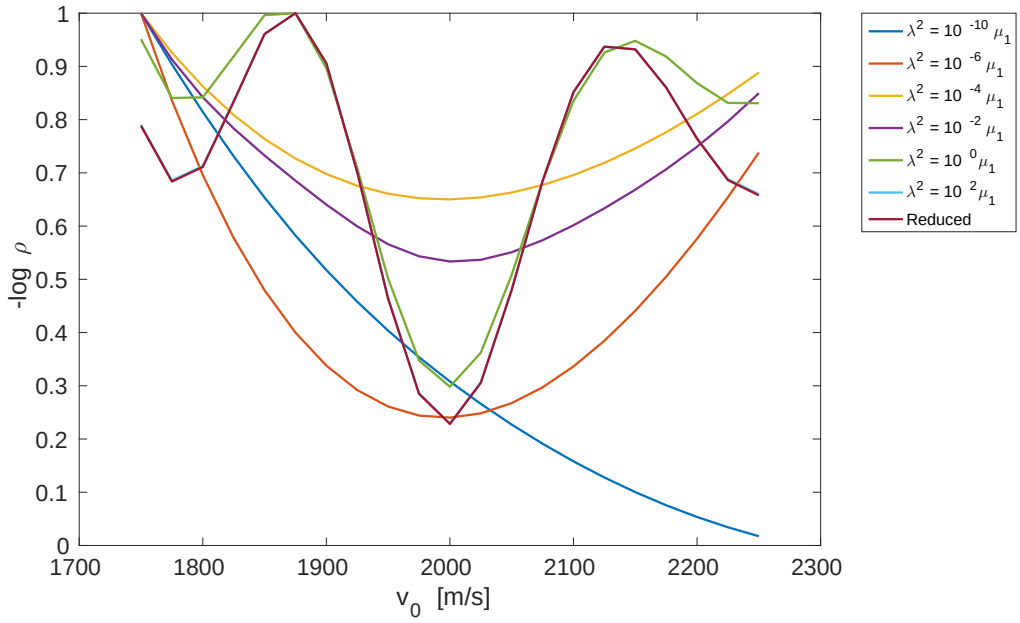
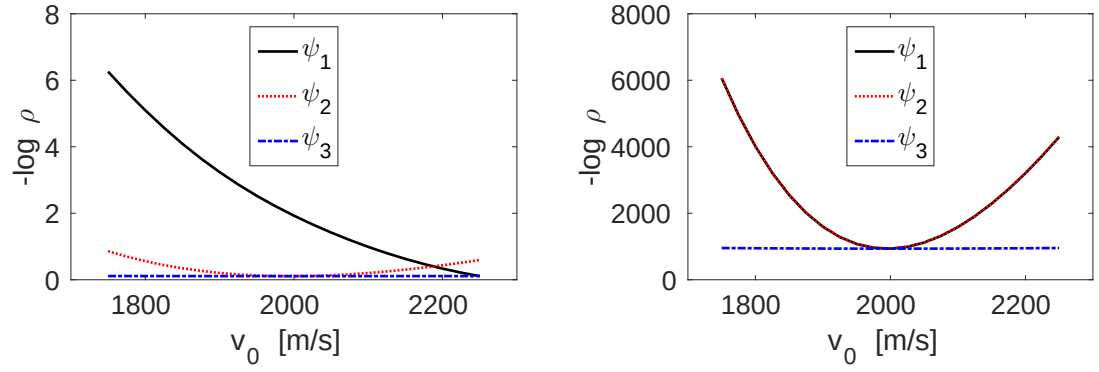
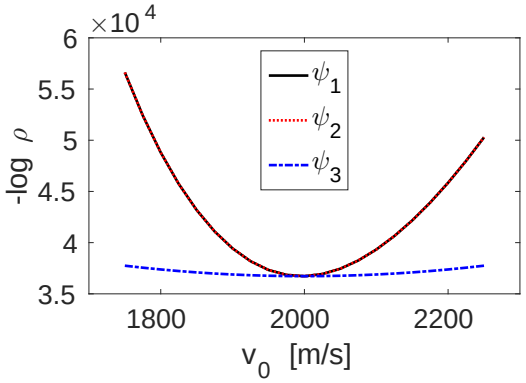


Figure 4: The comparison of the negative log-likelihood functions.

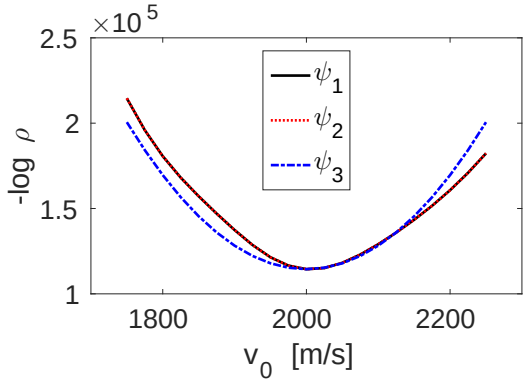


(a) $\lambda^2 = 10^{-10}\mu_1$

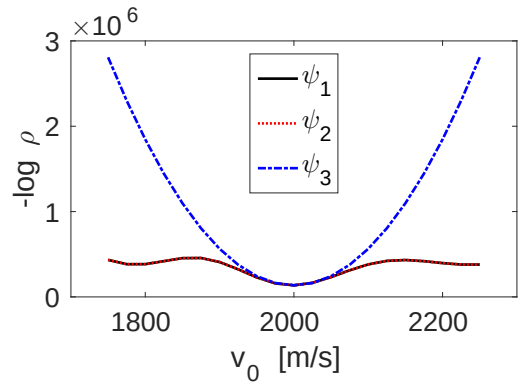
(b) $\lambda^2 = 10^{-6}\mu_1$



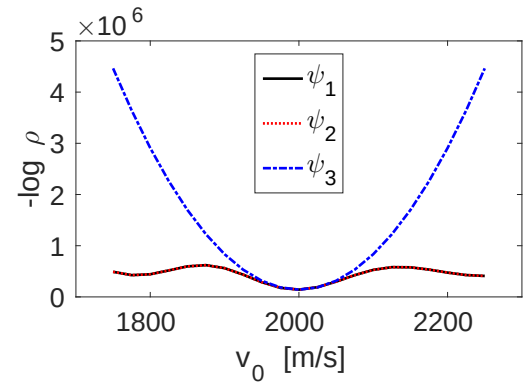
(c) $\lambda^2 = 10^{-4}\mu_1$



(d) $\lambda^2 = 10^{-2}\mu_1$

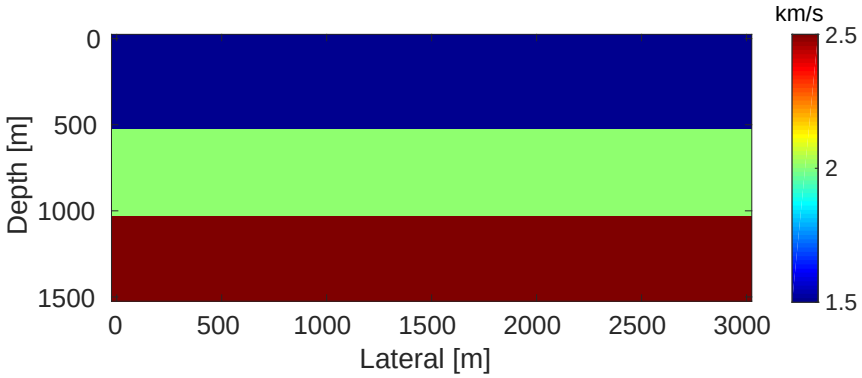


(e) $\lambda^2 = 10^{0}\mu_1$

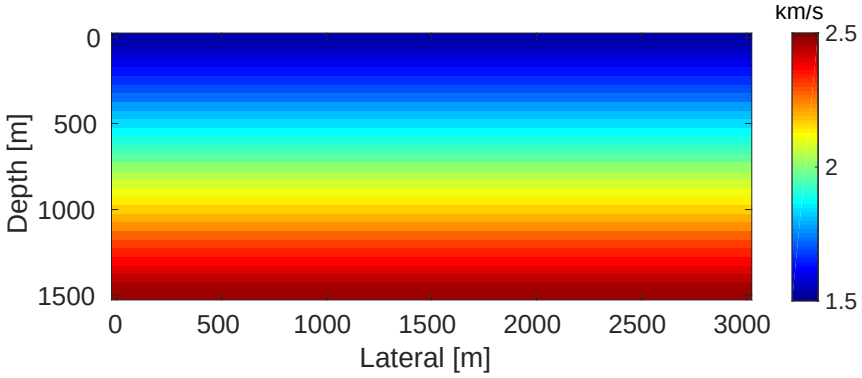


(f) $\lambda^2 = 10^{2}\mu_1$

Figure 5: The comparison of the negative log-likelihood functions of the true distribution (ψ_1), the approximate distribution (ψ_2), and the Gaussian approximation distribution (ψ_3) with the values of $\lambda^2 = 10^{-10}\mu_1$, $10^{-6}\mu_1$, $10^{-4}\mu_1$, $10^{-2}\mu_1$, $10^{0}\mu_1$, and $10^{2}\mu_1$.

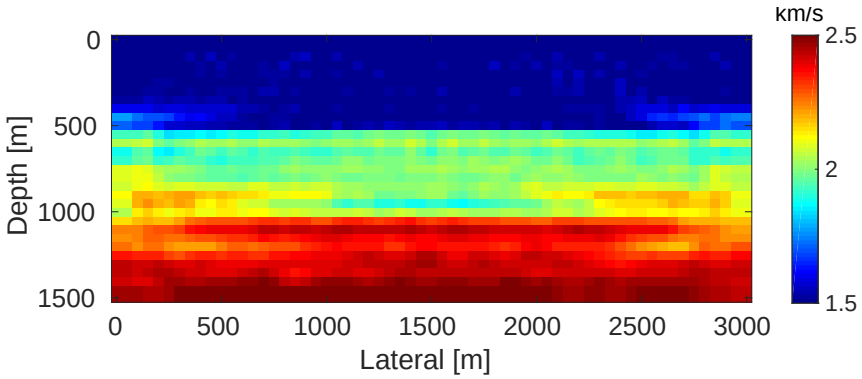


(a) True model

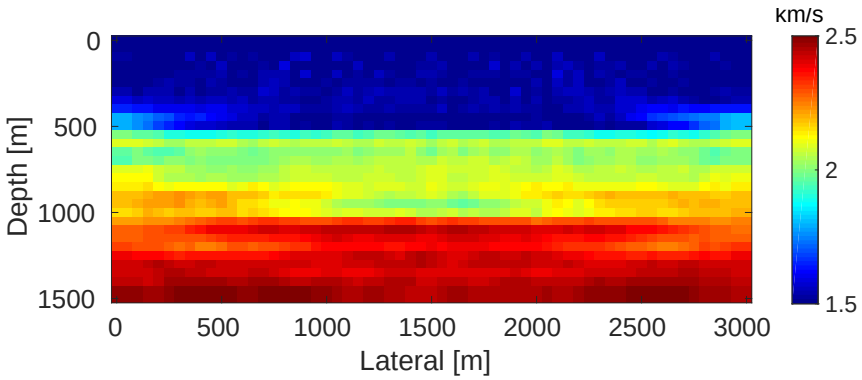


(b) Prior mean model

Figure 6: The true model (a) and the prior mean model (b).

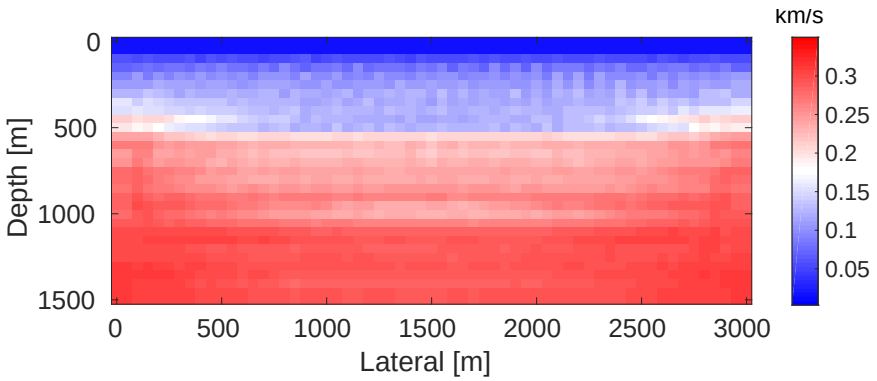


(a) GARTO

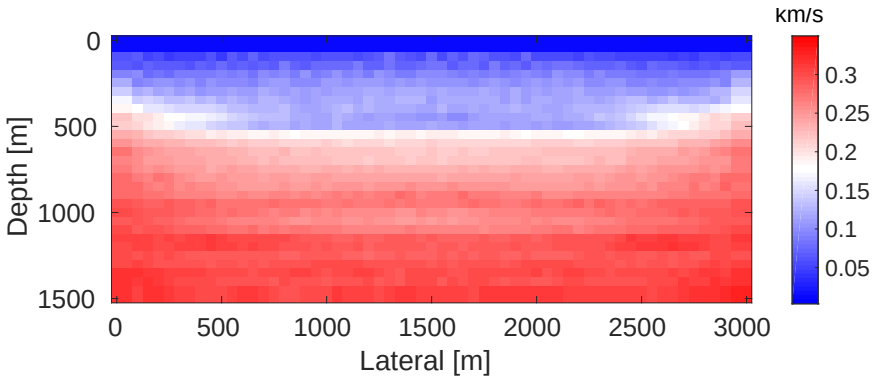


(b) RML

Figure 7: The posterior mean models obtained by the GARTO method (a) and the RML method (b).

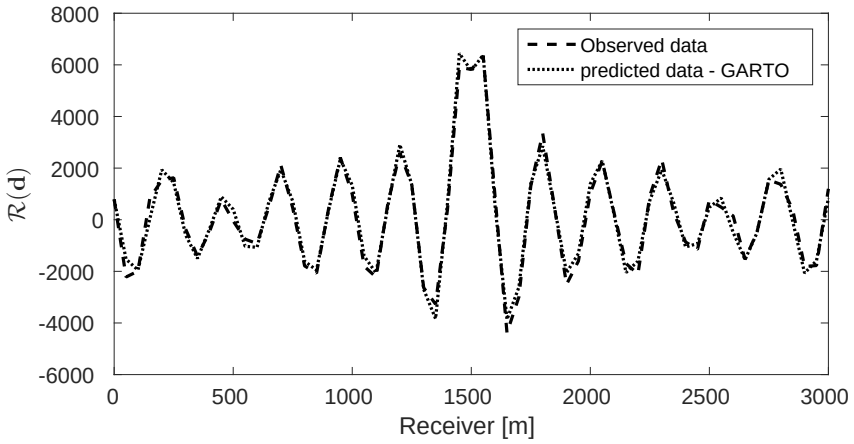


(a) GARTO

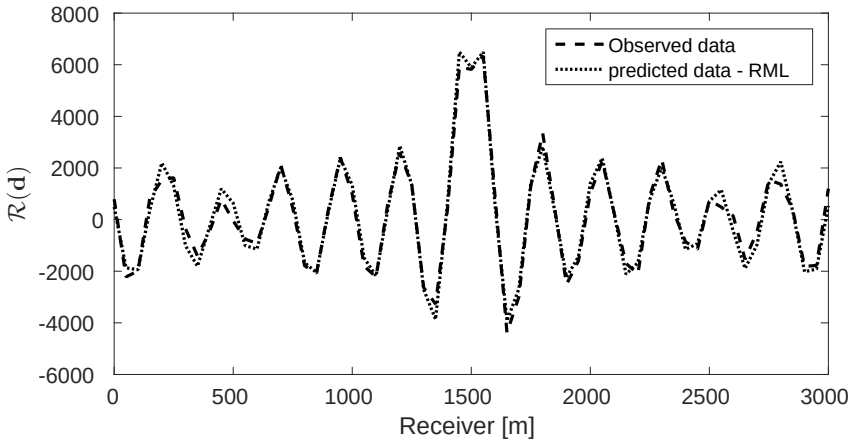


(b) RML

Figure 8: The posterior standard deviations obtained by the GARTO method (a) and the RML method (b).

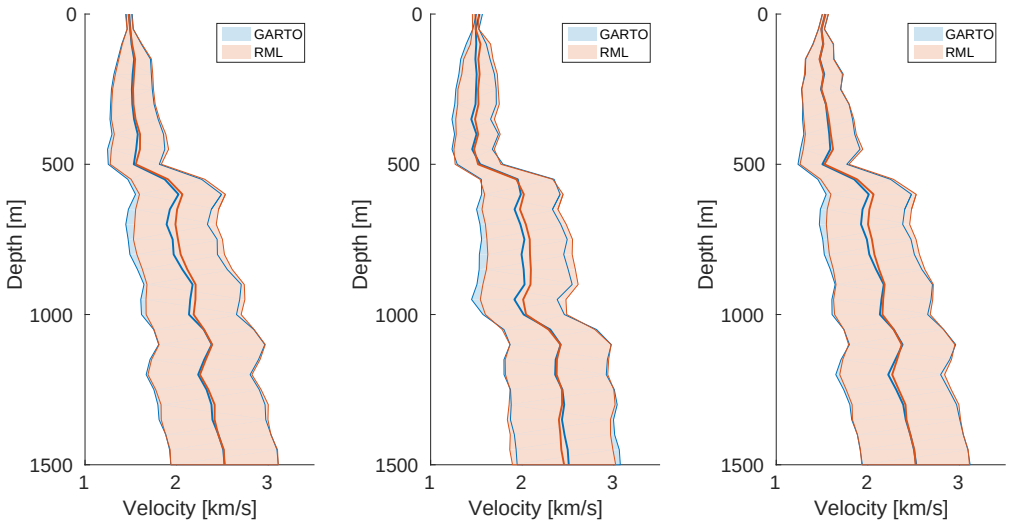


(a) GARTO



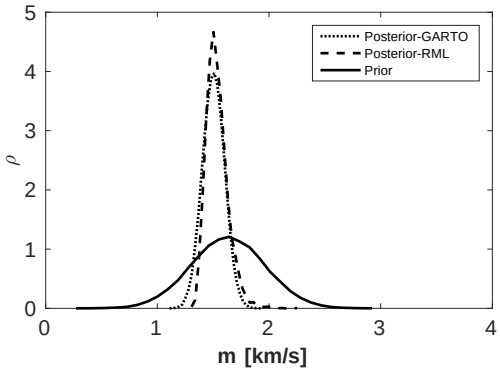
(b) RML

Figure 9: The comparison between the observed data and the predicted data with the posterior mean models obtained by the GARTO method (a) and the RML method (b).

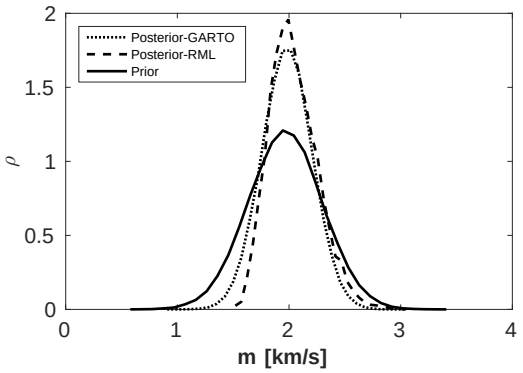


(a) $x = 500\,\mathrm{m}$ (b) $x = 1500\,\mathrm{m}$ (c) $x = 2500\,\mathrm{m}$

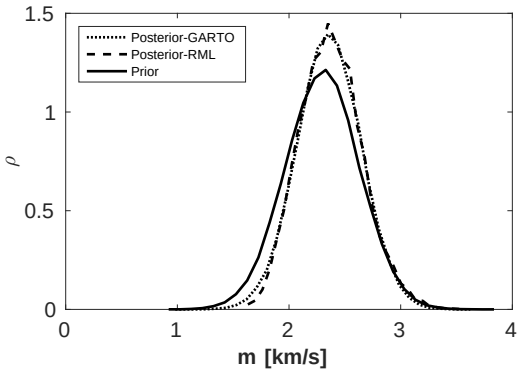
Figure 10: The mean (line) and 95% confidence interval (background patch) comparisons of the GARTO method (blue) and the RML method (red) at $x = 500\,\mathrm{m}$, $1500\,\mathrm{m}$, and $2500\,\mathrm{m}$. The similarity between these two results implies that the confidence intervals obtained with the GARTO method is a good approximation of the ones obtain with the RML method.



(a) $x = 1500\,\mathrm{m}$, $z = 200\,\mathrm{m}$

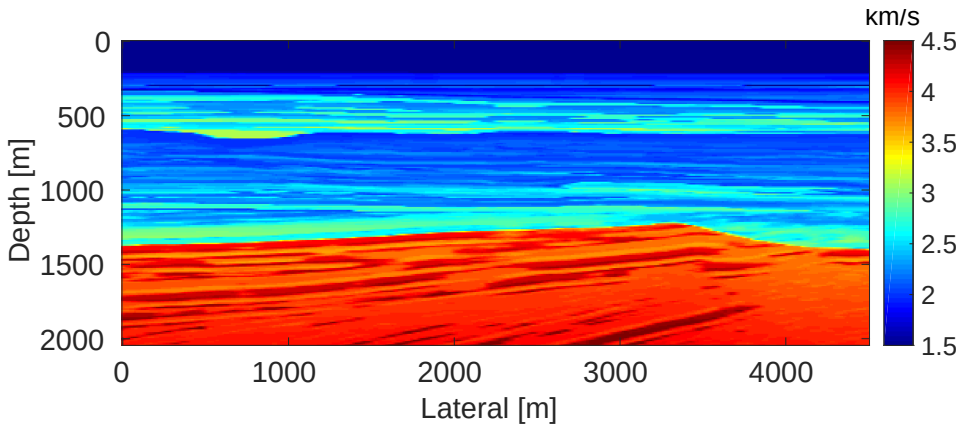


(b) $x = 1500\,\mathrm{m}$, $z = 700\,\mathrm{m}$

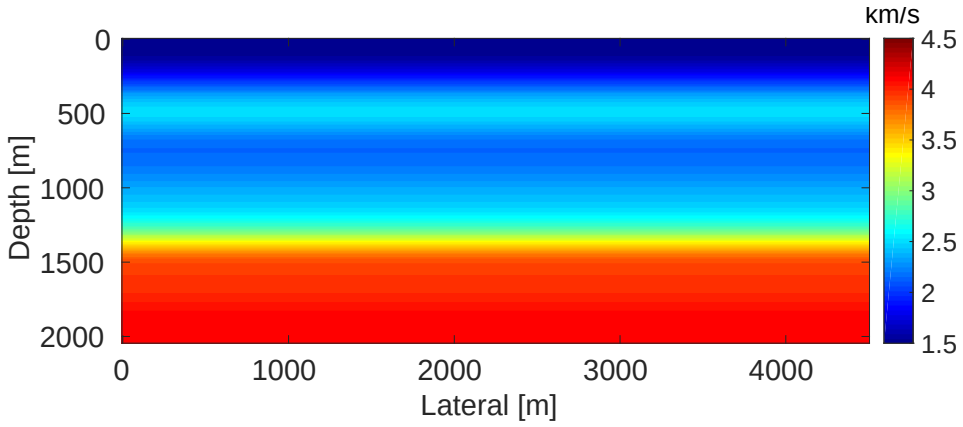


(c) $x = 1500\,\mathrm{m}$, $z = 1200\,\mathrm{m}$

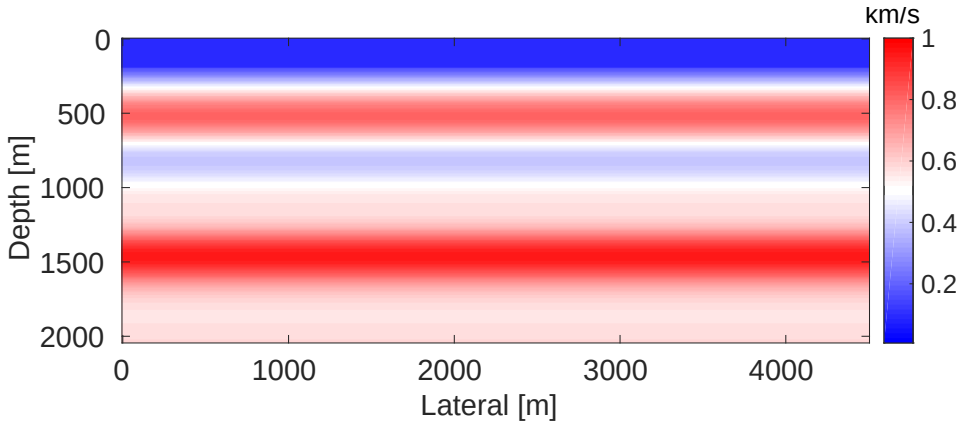
Figure 11: The comparison of the prior marginal distribution (solid line) and the posterior marginal distributions obtained by the GARTO method (dotted line) and the RML method (dashed line) at the locations of $(x, z) = (1500\mathrm{m}, 200\mathrm{m})$, $(1500\mathrm{m}, 700\mathrm{m})$, and $(1500\mathrm{m}, 1200\mathrm{m})$.



(a) True model

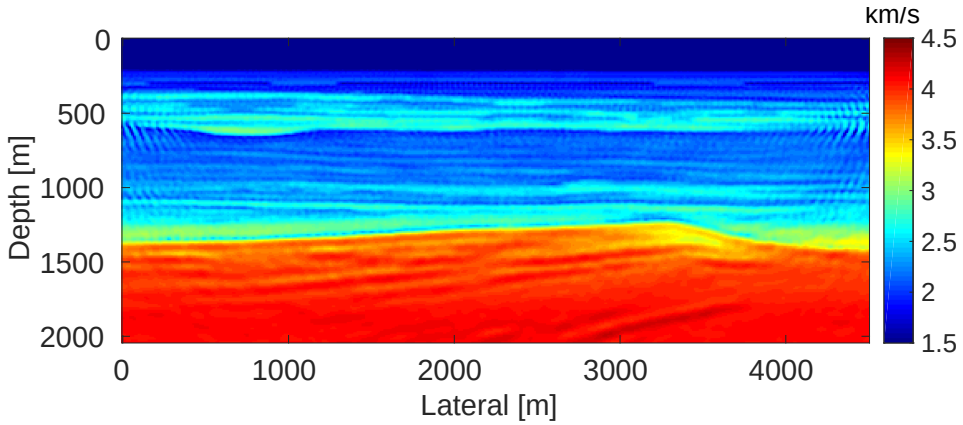


(b) Prior mean model

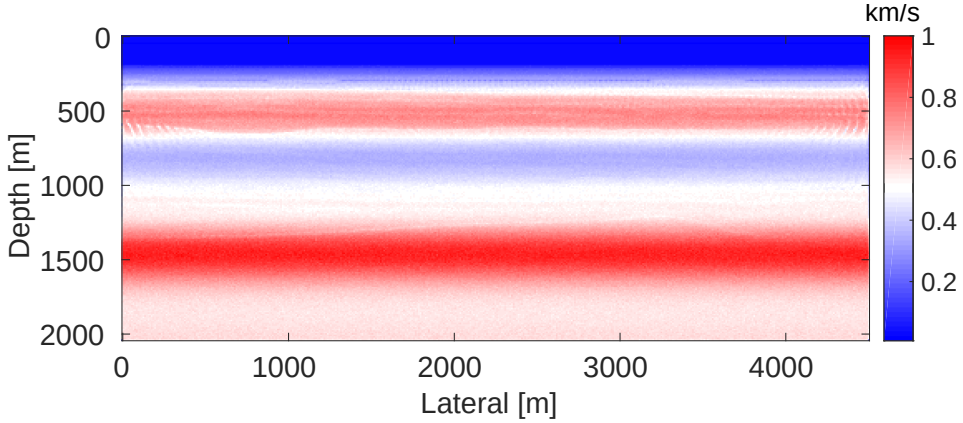


(c) Prior standard deviation

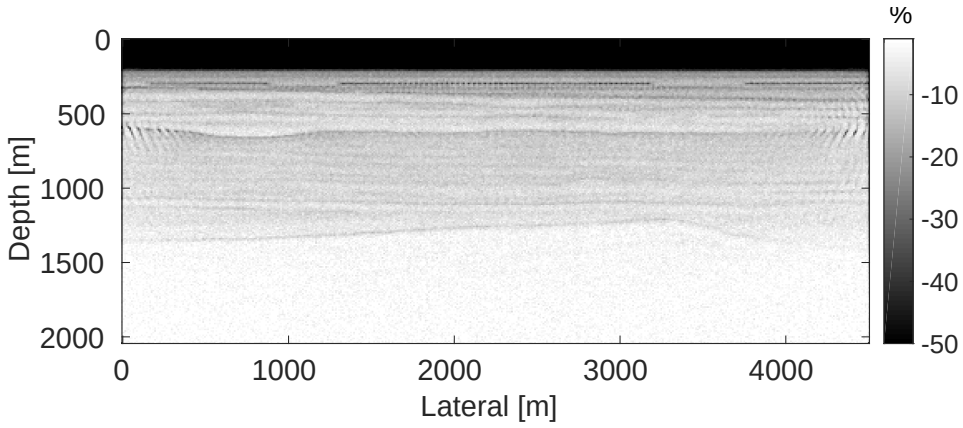
Figure 12: The true model (a), the prior mean model (b), and the prior standard deviation (c).



(a) Posterior MAP

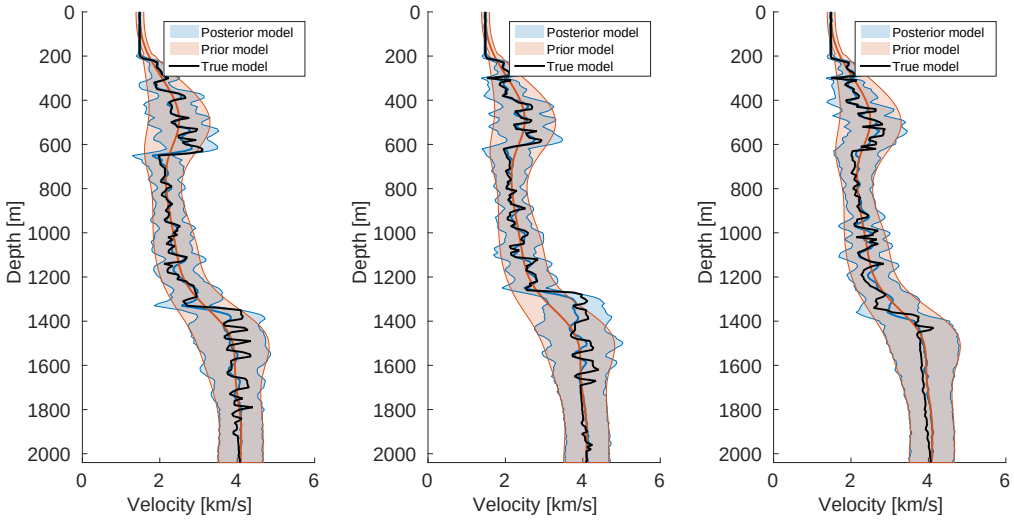


(b) Posterior standard deviation



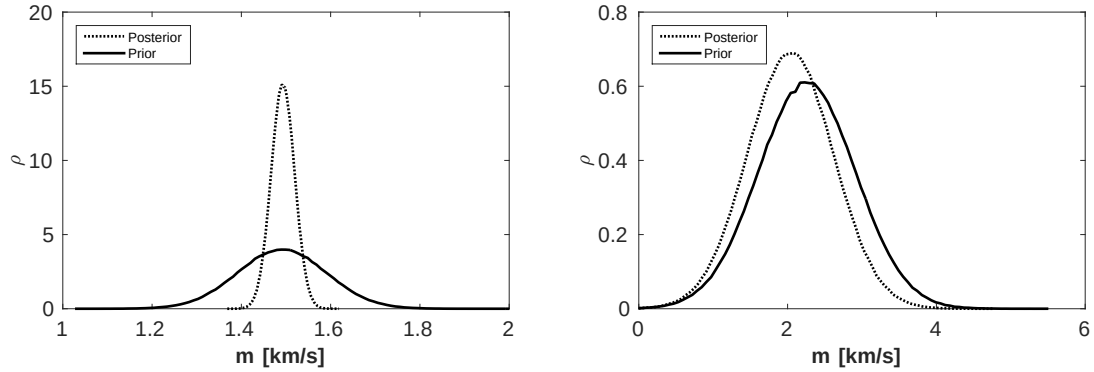
(c) Relative difference between the posterior and prior standard deviations

Figure 13: The posterior MAP estimate (a), the posterior standard deviation (b), and the relative difference $\frac{\mathrm{STD}_{\mathrm{post}}(m_k)-\mathrm{STD}_{\mathrm{prior}}(m_k)}{|\mathrm{STD}_{\mathrm{prior}}(m_k)|}$ between the posterior and the prior standard deviations (c).



(a) $x = 1000\,\mathrm{m}$ (b) $x = 2500\,\mathrm{m}$ (c) $x = 4000\,\mathrm{m}$

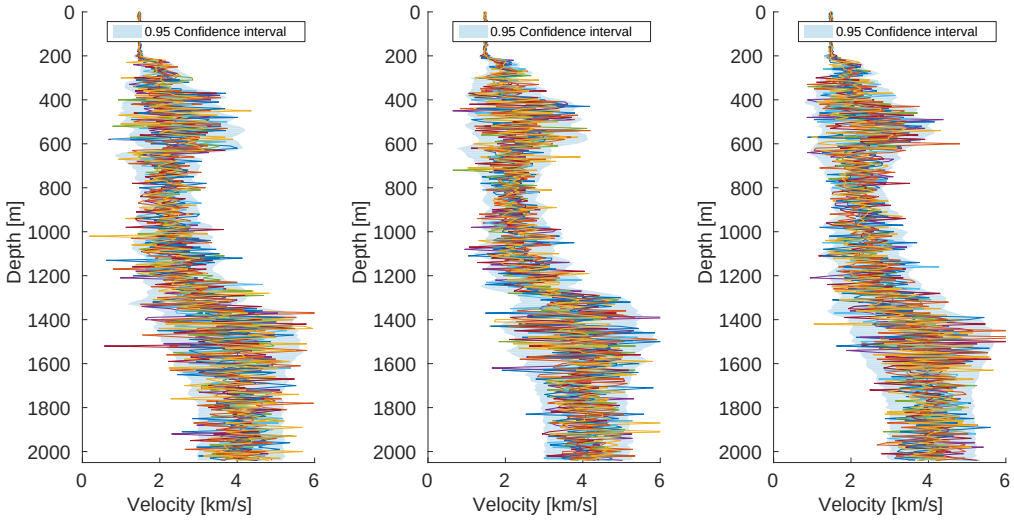
Figure 14: The mean and standard deviation comparisons of the posterior (blue) and the prior (red) distributions at $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$.



(a) $x = 2240\,\mathrm{m}$, $z = 40\,\mathrm{m}$ (b) $x = 2240\,\mathrm{m}$, $z = 640\,\mathrm{m}$

(c) $x = 2240\,\mathrm{m}$, $z = 1240\,\mathrm{m}$ (d) $x = 2240\,\mathrm{m}$, $z = 1840\,\mathrm{m}$

Figure 15: The comparison of the prior (solid line) and the posterior (dotted line) marginal distributions at the locations of $(x, z) =$ (2240m, 40m), (2240m, 640m), (2240m, 1240m), and (2240m, 1840m).



(a) $x = 1000\,\mathrm{m}$ (b) $x = 2500\,\mathrm{m}$ (c) $x = 4000\,\mathrm{m}$

Figure 16: The 95% confidence intervals and the 10 realizations via the RML method at $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$.

LIST OF TABLES

Frequency	2	3	4	5	6	7	8	9	10	11
λ	37.8	29.3	24.4	20.5	17.6	15.2	13.5	12.1	10.9	10.1
Frequency	12	13	14	15	16	17	18	19	20	21
λ	9.3	8.8	8.3	7.9	7.5	7.1	6.9	6.5	6.4	6.1
Frequency	22	23	24	25	26	27	28	29	30	31
λ	5.9	5.7	5.5	5.4	5.2	5.1	4.9	4.8	4.7	4.6

Table 1: The selection of λ for each frequency