Acoustic Full-Waveform Inversion and its Uncertainty Estimation based on a vector-version Square-Root Variable Metric method

Qiancheng Liu

Division of Physical Sciences and Engineering (PSE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.

E-mail: qiancheng.liu@kaust.edu.sa

December 2018

Abstract. Although in FWI the gradient optimisation-based methods have been well developed, the uncertainty estimation methods are also essential but still left behind. The evaluation of FWI uncertainty involves the Hessian-related posterior covariance matrix, which is prohibitive to compute and store for practical problems. To confront this issue, we propose to run FWI with square-root-variable-metric (SRVM), a quasi-Newton method similar to L-BFGS, in a matrix-free vector version. The vector-version SRVM is memory-affordable even for large-scale problems. The size of one SRVM vector is identical to that of the parameter model, and the number of SRVM vectors equals the total iteration number. We validate SRVM within acoustic FWI, with L-BFGS for reference. After the SRVM-based FWI converges, we can access the approximated inverse data-misfit Hessian and then the posterior covariance from the stored SRVM scalar and vector series. To reconstruct the inverse Hessian more efficiently, we factorise its eigenvalues and eigenvectors via randomised singular value decomposition (RSVD). We can qualify the inversion uncertainty through the posterior standard deviation, the 2D prior/posterior random samplings. We demonstrate our methods with numerical examples.

Keywords: computational seismology, seismic tomography, waveform inversion, uncertainty estimation

1. Introduction

Full waveform inversion (FWI) plays a vital role in exploring the subsurface structures. In most geophysical applications, FWI is introduced as an iterative, local optimization problem that attempts to minimize the residuals between observed and synthetic data. As an ill-posed inverse problem, FWI has non-unique solutions, i.e., uncertainty [Tarantola, 2005]. Although methods in FWI the gradient-based optimisation have been well developed, the uncertainty estimation methods are still left behind. Several methods in geophysics have been proposed in the FWI community. We divide them into two

categories: deterministic methods such as optimization-based ones [Virieux et al., 2009], and probabilistic approaches such as Bayesian inversion [Tarantola, 2005].

The deterministic methods on FWI have been well developed in tackling several different issues, such as convergence rate, cycling-skipping and multi-parameters [Tromp et al., 2005, Virieux et al., 2009, Wu and Alkhalifah, 2016, Wu and Alkhalifah, 2017, da Silva and Yao, 2017, Yao et al., 2018]. The probability inversion methods in FWI are also essential. From the viewpoint of Bayesian inference, the solution of the inverse problem should not be limited to some single set of inverted values, but be represented by a probability density function (PDF) to quantify the posterior model uncertainty [Tarantola, 2005].

Estimation of the resolution or uncertainty of the seismic inversion has a long history in geophysics. This information can be analyzed with mathematical tools such as posterior covariance matrix [Backus and Gilbert, 1968, Backus and Gilbert, 1970]. The posterior covariance matrix is closely related to Hessian [Fichtner and Trampert, 2011a, Fichtner and Trampert, 2011b, Fichtner and van Leeuwen, 2015]. For practical problems with millions of parameters, it is unfeasible to store such vast matrices. To handle large-scale inverse problems, [Zhang and McMechan, 1995] modify the classic inversion algorithm with LSQR. [An, 2012] evaluates the spatial resolution lengths with a Gaussian approximation to the resolution matrix. [Trampert et al., 2012] sample the tomographic models for resolution lengths with random probing. [Fichtner and van Leeuwen, 2015] analyze the direction-dependent resolution lengths of waveform tomography by autocorrelating the randomly sampled Hessian. [Rawlinson et al., 2014] carry out a recent review of the uncertainty estimations.

With the development of matrix probing theories in applied mathematics, randomised singular-value decomposition (RSVD) [Liberty et al., 2007, Halko et al., 2011] comes into the sight of geophysicists. [Bui-Thanh et al., 2013] formulate waveform tomography in a Bayesian inference workflow, deriving an approximation to the posterior covariance by decomposing the data-misfit Hessian into eigenvalues and eigenvectors with RSVD. [Zhu et al., 2016] use the RSVD to improve the efficiency of the Hessian-computing in acoustic FWI. Similar researches using Bayesian inference based nonlinear sampling for uncertainty estimation can also be found in [Mosegaard and Tarantola, 1995, Gouveia and Scales, 1998, Sambridge and Mosegaard, 2002, Osypov et al., 2013]. However, the application of such methods to large-scale tomography problems is prohibitive due to the involved large number of unknown parameters. Also, some ensemble-based approaches based on the Kalman Filter (KF) theory [Kalman et al., 1960, Evensen, 1994] have been applied to tomography problem by [Jordan, 2015, Thurin et al., 2017].

In this paper, we investigate the application of the Square-Root Variable Metric (SRVM) method [Williamson, 1975, Morf and Kailath, 1975, Hull and Tapley, 1977] in FWI and its uncertainty estimation. The SRVM method, which preserves the positive definiteness of the inverse data-misfit Hessian in a square-root form during the inversion process, overcomes the shortcomings of Davidon-Fletcher-Powell (DFP)

methods such as mandatory exact line search and being sensitive to the round-off error [Williamson, 1975, Hull and Tapley, 1977]. [Tarantola, 2005] unburies and presents SRVM to us due to its potential in posterior model samplings [Luo, 2012]. Theoretically, the SRVM method is capable of collecting up the information of the second-order derivative from the initial to the final inverted models. We apply the vector-version SRVM into FWI. There are one SRVM vector and one SRVM scalar per iteration in FWI, and the size of an SRVM vector is the same as that of the parameter model. After the SRVM-based FWI converges, we can retrieve the second-order information from the stored SRVM vectors and scalars in a recursive manner. SRVM has already approached the inverse Hessian in a low-rank form. To facilitate the posterior information retrieval more, we incorporate SRVM with RSVD.

The paper is organized as follows. First, we briefly review the theory of SRVM, and formulate it into a memory-affordable vector version. Then, we discuss the implementation of SRVM into FWI. Next, we reconstruct the inverse data-misfit Hessian via stored SRVM scalars and vectors, and incorporate RSVD into SRVM. Afterward, we discuss how to perform uncertainty estimation and draw random samples from the posterior covariance. Finally, we verify our methods above on numerical examples to demonstrate their applications in FWI.

2. THEORIES

2.1. Review of the Square-Root Variable-Metric (SRVM) method

In Newtons method, we use a second-order Taylor approach to expand function $f(\mathbf{m})$ around $\mathbf{m}_0$ as follows:

$$f(\mathbf{m}_0 + \Delta\mathbf{m}) \approx f(\mathbf{m}_0) + \mathbf{g}^T \Delta\mathbf{m} + \frac{1}{2} \Delta\mathbf{m}^T \mathbf{H} \Delta\mathbf{m}, \quad (1)$$

with $\mathbf{g} = \nabla f(\mathbf{m}_0)$ and $\mathbf{H} = \nabla^2 f(\mathbf{m}_0)$. The gradient of this approximation concerning $\Delta\mathbf{m}$ reads

$$\Delta f(\mathbf{m}_0 + \Delta\mathbf{m}) \approx \mathbf{g}^T + \mathbf{H} \Delta\mathbf{m}. \quad (2)$$

When $f(\mathbf{m})$ reaches its minimum, the left-hand side of Eq. (2) becomes zero, yielding

$$\Delta\mathbf{m} \approx -\mathbf{H}^{-1} \mathbf{g}^T. \quad (3)$$

Eq. (3) represents the Newton updating. Following the standard Davidon-Fletcher-Powell (DFP) formula, we update the inverse Hessian $\mathbf{B} = \mathbf{H}^{-1}$ as follows:

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{\mathbf{B}_k \Delta\mathbf{g}_k \Delta\mathbf{g}_k^T \mathbf{B}_k}{\Delta\mathbf{g}_k^T \mathbf{B}_k \Delta\mathbf{g}_k} + \frac{\Delta\mathbf{m}_k \Delta\mathbf{m}_k^T}{\Delta\mathbf{g}_k^T \Delta\mathbf{m}_k} \quad (4)$$

with $\Delta\mathbf{g}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. Eq. (4) repeats recursively until the minimization process converges. $\mathbf{B}_0$, the initial guess about the inverse Hessian, is usually chosen as an identity matrix $\mathbf{I}$ for convenience. To ensure $\mathbf{B}$ is always positive-definite, [Williamson, 1975] observed that Eq. (4) can be expressed in a square-root form:

$$\mathbf{S}_{k+1} \mathbf{S}_{k+1}^T = \mathbf{S}_k \left(\mathbf{I} - (1/P_k) \mathbf{S}_k^T \mathbf{y}_k \mathbf{y}_k^T \mathbf{S}_k \right) \mathbf{S}_k^T, \quad (5)$$

where $\mathbf{B}_{k+1} = \mathbf{S}_{k+1} \mathbf{S}_{k+1}^T$, $\mathbf{B}_k = \mathbf{S}_k \mathbf{S}_k^T$, $\mathbf{y}_k = \mu_k \mathbf{g}_k + \Delta \mathbf{g}_k$, $P_k = \mathbf{y}_k^T \mathbf{B}_k \Delta \mathbf{g}_k^T$, with μ_k being the one-dimension search step at the k th iteration and $\Delta \mathbf{g}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ [Huang, 1970]. We furthermore clarify Eq. (5) in a square-root form as

$$\mathbf{A}_k \mathbf{A}_k^T = \mathbf{I} - (v_k / P_k) \mathbf{S}_k^T \mathbf{y}_k \mathbf{y}_k^T \mathbf{S}_k, \quad (6)$$

The constant v_k is determined as

$$v_k = \frac{1 \pm (1 - Q_k / P_k)^{1/2}}{Q_k / P_k}, \quad (7)$$

where $Q_k = \mathbf{y}_k^T \mathbf{S}_k \mathbf{S}_k^T \mathbf{y}_k$, such that the relationship in Eq. (6) holds. In Eq. (7), we choose the minus sign to prevent v_k from being singular at $Q/P = 0$; we also observe that once $Q/P > 1$, v_k becomes imaginary, resulting in a no real matrix $\mathbf{S}$. [Hull and Tapley, 1977] suggest using $v_k = 1$ for $Q/P > 1$, because when $-\infty \leq Q/P \leq 1$, $0 \leq \eta \leq 1$. By here the updating workflow of SRVM has been briefly stated. The square root of the inverse Hessian updates as following

$$\mathbf{S}_{k+1} = \mathbf{S}_k \left(\mathbf{I} - (v_k / P_k) \mathbf{S}_k^T \mathbf{y}_k \mathbf{y}_k^T \mathbf{S}_k \right). \quad (8)$$

However, we notice that Eq. (8) remains in a matrix form. Even after an analytic Cholesky decomposition [Carlson, 1973], the memory requirement of $\mathbf{S}$ is still at the half of $\mathbf{B}$ in Eq. (4), non-storable as well. In the next section, we will formulate how to reduce the matrix-version SRVM into a vector version to make the memory requirement affordable.

2.2. Vector-version Square-Root Variable-Metric (SRVM)

For practical geophysics problems such as tomography at regional or global scales, matrices containing tens of millions of parameters are unfeasible to store or update. First, let us briefly summarize the matrix-version SRVM into the following workflow:

$$\begin{aligned} \Delta \mathbf{g}_k &= \mathbf{g}_{k+1} - \mathbf{g}_k \\ \mathbf{y}_k &= \mu_k \mathbf{g}_k + \Delta \mathbf{g}_k \\ P_k &= \mathbf{y}_k^T \mathbf{S}_k \mathbf{S}_k^T \Delta \mathbf{g}_k \\ Q_k &= \mathbf{y}_k^T \mathbf{S}_k \mathbf{S}_k^T \mathbf{y}_k \\ v_k &= \frac{1 - (1 - Q_k / P_k)^{1/2}}{Q_k / P_k} \\ \mathbf{S}_{k+1} &= \mathbf{S}_k \left(\mathbf{I} - \frac{v_k}{P_k} \mathbf{S}_k^T \mathbf{y}_k \mathbf{y}_k^T \mathbf{S}_k \right) \\ \mathbf{B}_{k+1} &= \mathbf{S}_{k+1} \mathbf{S}_{k+1}^T = \mathbf{S}_k \left(\mathbf{I} - \frac{v_k}{P_k} \mathbf{S}_k^T \mathbf{y}_k \mathbf{y}_k^T \mathbf{S}_k \right) \left(\mathbf{I} - \frac{v_k}{P_k} \mathbf{S}_k^T \mathbf{y}_k \mathbf{y}_k^T \mathbf{S}_k \right)^T \mathbf{S}_k^T \end{aligned} \quad (9)$$

From Alg. (9) we observe that the term $\mathbf{y}_k^T \mathbf{S}_k$ and its transpose widely exist. Then, by setting $\beta_k = \mathbf{S}_k^T \Delta \mathbf{g}_k$, we obtain a vector-version SRVM workflow as following

$$\begin{aligned}
\Delta \mathbf{g}_k &= \mathbf{g}_{k+1} - \mathbf{g}_k \\
\mathbf{y}_k &= \mu_k \mathbf{g}_k + \Delta \mathbf{g}_k \\
\mathbf{w}_k &= \mathbf{S}_k^T \mathbf{y}_k \\
\beta_k &= \mathbf{S}_k^T \Delta \mathbf{g}_k \\
P_k &= \mathbf{w}_k^T \beta_k \\
Q_k &= \mathbf{w}_k^T \mathbf{w}_k \\
v_k &= \frac{1 - (1 - Q_k / P_k)^{1/2}}{Q_k / P_k} \\
\mathbf{S}_{k+1} &= \mathbf{S}_k \left(\mathbf{I} - \frac{v_k}{P_k} \mathbf{w}_k \mathbf{w}_k^T \right) \\
\mathbf{B}_{k+1} &= \mathbf{S}_{k+1} \mathbf{S}_{k+1}^T = \mathbf{S}_k \left(\mathbf{I} - \frac{v_k}{P_k} \mathbf{w}_k \mathbf{w}_k^T \right) \left(\mathbf{I} - \frac{v_k}{P_k} \mathbf{w}_k \mathbf{w}_k^T \right)^T \mathbf{S}_k^T
\end{aligned} \tag{10}$$

Given $\mathbf{m}_0$ as a vector containing M parameters, in Alg. (10), we observe that there remains one matrix $\mathbf{S}$ of size M by M . Similar with L-BFGS, we choose an initial guess of the inverse Hessian as $\mathbf{B}_0 = \mathbf{I}$, which is surely symmetric positive-definite, such that $\mathbf{S}_0 = \mathbf{I}$ as well. We only need to store one vector $\mathbf{w}_k$ and one scalar v_k / P_k for all the iterative history step to reconstruct and extract elements from $\mathbf{S}$. When applying vector-version SRVM to FWI, with $\mathbf{B}_0 = \mathbf{I}$, we implement the matrix-vector multiplication to determine the search direction by $\mathbf{p}_{k+1} = -\mathbf{S}_{k+1} \mathbf{S}_{k+1}^T \mathbf{g}_{k+1}$ as

$$\begin{aligned}
\delta &= -\mathbf{S}_{k+1}^T \mathbf{g}_{k+1} = \prod_{i=1}^k \left(\frac{v_i}{P_i} \mathbf{w}_i \mathbf{w}_i^T - \mathbf{I} \right) \mathbf{g}_{k+1} = \prod_{i=1}^k \left[\frac{v_i}{P_i} (\mathbf{w}_i^T \mathbf{g}_{k+1}) \mathbf{w}_i - \mathbf{g}_{k+1} \right], \\
\mathbf{p}_{k+1} &= \mathbf{S}_{k+1} \mathbf{S}_{k+1}^T \mathbf{g}_{k+1} = \mathbf{S}_{k+1} \delta = \prod_{i=1}^k \left(\mathbf{I} - \frac{v_i}{P_i} \mathbf{w}_i \mathbf{w}_i^T \right) \delta = \prod_{i=1}^k \left[\delta - \frac{v_i}{P_i} (\mathbf{w}_i^T \delta) \mathbf{w}_i \right],
\end{aligned} \tag{11}$$

in which both $\mathbf{w}_i^T \mathbf{g}_{k+1}$ and $\mathbf{w}_i^T \delta$ are scalars. The combination of Algs. (10) and (11) indicates that only vector and scalar operations occur in the updating of search direction $\mathbf{p}_{k+1}$.

3. FULL-WAVEFORM INVERSION (FWI) with SRVM

FWI consists of three recursive steps: (i) obtain a gradient by the adjoint method; (ii) update the gradient direction to get the search direction by the optimization method; (iii) determine a step length along the search direction through line searching.

3.1. Sensitivity kernel

The waveform misfit function in a least-squares sense measuring the goodness of fit [Tarantola, 1984, Tarantola, 1987, Tromp et al., 2005] reads

$$\chi(\mathbf{m}) = \frac{1}{2} \sum \int_0^T \|\mathbf{s}(\mathbf{x}_r, t, \mathbf{m}) - \mathbf{d}(\mathbf{x}_r, t)\|^2 dt, \tag{12}$$

where $\mathbf{d}(\mathbf{x}_r, t)$ is the observed data at receivers $\mathbf{x}_r$, $\mathbf{s}(\mathbf{x}_r, t, \mathbf{m})$ the corresponding synthetic data with $\mathbf{m}$ being a given model, and T the recording duration. In acoustic FWI, we measure the pressure as the observable and define $\mathbf{m}$ as the P-wave velocity.

According to the adjoint method [Tromp et al., 2005], the gradient (sensitivity kernel) of the misfit function concerning the P-wave velocity v can be expressed as

$$g(\mathbf{x}) = -\frac{1}{v(\mathbf{x})^3} \int_0^T \mathbf{s}^\dagger(\mathbf{x}, T-t) \cdot \frac{\partial^2}{\partial t^2} \mathbf{s}(\mathbf{x}, t) dt, \quad (13)$$

where $\mathbf{s}^\dagger(\mathbf{x}, T-t)$ denotes the adjoint field generated by the backward-propagated data residual $\mathbf{s}(\mathbf{x}_r, t, \mathbf{m}) - \mathbf{d}(\mathbf{x}_r, t)$, and $\mathbf{s}(\mathbf{x}, t)$ the forward-propagated source wavefield.

3.2. SRVM into FWI

When the gradient is ready in FWI, one can use a quasi-Newton method to update it for the search direction. The quasi-Newton method assumes that a function can be locally approximated quadratically around the optimum [Nocedal and Wright, 2006]. The consideration of the quadratic term (the Hessian) can significantly accelerate the converging rate and improve the final inverted result of FWI [Virieux and Operto, 2009]. Also, the (inverse) Hessian contains valuable information for furthermore resolution analysis or uncertainty estimation in waveform tomography [Sambridge and Mosegaard, 2002, Fichtner and Trampert, 2011a, Fichtner and Trampert, 2011b].

The L-BFGS algorithm [Liu and Nocedal, 1989] is considered the most efficient quasi-Newton method in solving unconstrained nonlinear optimization problems [Nocedal and Wright, 2006]. With this algorithm, we can use a limited amount of gradient series to approximate the inverse Hessian. L-BFGS runs in vector operations, well suitable for optimization with a massive number of variables. Unlike L-BFGS originating from BFGS algorithm, the SRVM method originates from the DFP algorithm, the dual of BFGS algorithm. In a square-root form, SRVM overcomes two limitations of the conventional DFP method: mandatory adequate line search and sensitive to roundoff error [Fletcher and Powell, 1963, Morf and Kailath, 1975]. Its vector version in Alg. (10) makes feasible the large-scale applications. As a member of the quasi-Newton family, SRVM is also expected to speed up the convergence behavior of FWI. Note that here we output not the approximate inverse Hessian $\mathbf{SS}^T$ but the search direction by $\mathbf{p} = -\mathbf{SS}^T \mathbf{g}$ directly, as stated in Alg. (11). The search direction $\mathbf{p}$ must satisfy this relationship with the gradient $\mathbf{g}$ as

$$\mathbf{p}^T \mathbf{g} < 0, \quad (14)$$

to make sure such a decrease.

3.3. Backtracking line search

After obtaining the search direction, we need to search for the step length which satisfies

$$\mu_k = \arg \min_{\mu \geq 0} f(\mathbf{m}_k + \mu \mathbf{p}_k). \quad (15)$$

Determination of the step length involves additional forward modelings to assess the sensitivity of the misfit function to μ . We choose the backtracking line search method due to its high efficiency (Nocedal & Wright 2006). This inexact line search provides

an efficient way to compute an acceptable step length that reduces the misfit function sufficiently rather than exactly. We accept the step length μ once it meets the following inequalities:

$$\phi(\mu) \leq \phi(0) + c_1 \mu \phi'(0), \quad (16)$$

$$\phi'(\mu) \geq c_2 \phi'(0), \quad (17)$$

where $0 < c_1 < c_2 < 1$. The first inequality is known as the "Armijo condition" or the "sufficient decrease condition" and the second known as the "curvature condition". They two are referred to as the Wolfe conditions together [Nocedal and Wright, 2006]. The suggested values for c_1 and c_2 are, respectively, 0.0001 and 0.9 [Nocedal and Wright, 2006].

4. POSTERIOR ANALYSIS

4.1. Constructing the inverse Hessian via SRVM

Let us get back to Eq. (4), in which $\mathbf{B}_0 = \mathbf{I}$ acts as a stabilizer to ensure a sufficiently stable $\mathbf{B}_k$ during iterative updating [Nocedal and Wright, 2006]. Eq. (4) indicates that the history information about the gradients and the model-updates only contributes to $\mathbf{B}_{k+1} - \mathbf{B}_0 = \mathbf{B}_{k+1} - \mathbf{I}$, because $\mathbf{I}$ acts as a stabilizer, containing no information about seismic wavefields. At the convergence of FWI, we may retrieve the approximated inverse Hessian by

$$\mathbf{H}^{-1} = \mathbf{B}_{k+1} - \mathbf{I}. \quad (18)$$

Given the stored scalar series v/P and vector series $\mathbf{w}$ in Alg. (10) over the past k iterations, we can reconstruct $\mathbf{B}_{k+1}$ recursively via

$$\begin{aligned} \mathbf{S}_{i+1} &= \mathbf{S}_i \left(\mathbf{I} - \frac{v_i}{P_i} \mathbf{w}_i \mathbf{w}_i^T \right) \\ \mathbf{B}_{i+1} &= \mathbf{S}_{i+1} \mathbf{S}_{i+1}^T = \mathbf{S}_i \left(\mathbf{I} - \frac{v_i}{P_i} \mathbf{w}_i \mathbf{w}_i^T \right) \left(\mathbf{I} - \frac{v_i}{P_i} \mathbf{w}_i \mathbf{w}_i^T \right)^T \mathbf{S}_i^T \end{aligned} \quad (19)$$

through which the full reconstruction and storage of $\mathbf{B}_{k+1}$ is still impractical. Fortunately, it is feasible to extract column (row) elements from $\mathbf{B}_{k+1}$ via a pulse probing vector $\{0, 0, \dots, 0, 1, 0, \dots, 0, 0\}$, in which the locates at the target column (row) index. However, it is not easy to extract the diagonal elements, which requires $N * k$ loops, with N being the number of parameters and k being the number of iterations in acoustic FWI, because we can only extract one-column (row) elements per impulse sampler.

4.2. RSVD into SRVM

The SRVM method has already been a low-rank approach to reconstruct $\mathbf{H}^{-1}$, but its efficiency in information extraction is not so satisfactory. To facilitate the extraction operations, we turn to help from RSVD [Liberty et al., 2007, Halko et al., 2011], which provides an efficient solution to factorise a large matrix into its corresponding eigenvalues and eigenvectors. Compared with the classic RSVD [Liberty et al., 2007], the new

approach in [Halko et al., 2011] only probes a matrix with one set of independent random vectors. Given a symmetric $Nm \times Nm$ matrix $\mathbf{Z}$ and a set of random vectors $\mathbf{X}$ of $Nm \times Nr$, with being the estimated rank order of $\mathbf{Z}$, we carry out RSVD in the workflow below:

$$\begin{aligned} \mathbf{Y} &= \mathbf{Z}\mathbf{X} \\ \mathbf{Q}\mathbf{R} &= \mathbf{Y} \\ (\mathbf{Q}^T\mathbf{X})\mathbf{A} &= \mathbf{Q}^T\mathbf{Y} \\ \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T &= \mathbf{A} \\ \mathbf{V} &= \mathbf{Q}\mathbf{U} \\ \mathbf{Z} &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \end{aligned} \tag{20}$$

in which the first to the fourth lines correspond to *Sampling $\mathbf{Z}$ with $\mathbf{X}$* , *QR decomposition on $\mathbf{Y}$* , *Solve for $\mathbf{A}$* , and *SVD on $\mathbf{A}$* , respectively. When merging RSVD into SRVM, we only need to run Nr random probings using $\mathbf{X}$ in Alg. (19) to obtain $\mathbf{Y}$ directly, and then run Alg. (20) to output

$$\mathbf{H}^{-1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{21}$$

The resulting $\mathbf{V}$ and $\mathbf{\Lambda}$ are the eigenvectors and eigenvalues of $\mathbf{H}^{-1}$. Note that its rank order Nr complies with the total iteration number $Niter$, because SRVM has already expressed $\mathbf{B}$ in Alg. (10) in a low-rank form and $Niter$ is its corresponding rank order.

4.3. Practical assessment of the posterior covariance

Appendix A provides an probability method based viewpoint on the objective function. However, from a practical point of view, we do not use Eq. (A.9) but the following form [Rawlinson et al., 2014, Modrak and Tromp, 2016] instead

$$f(\mathbf{m}) = \frac{1}{2} \left((\mathbf{d} - \mathbf{g}(\mathbf{m}))^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{m})) + \varepsilon \delta \mathbf{m}^T \mathbf{C}_m^{-1} \delta \mathbf{m} + \eta \delta \mathbf{m}^T \mathbf{D}^T \mathbf{D} \delta \mathbf{m} \right), \tag{22}$$

where $\delta \mathbf{m} = \mathbf{m} - \mathbf{m}_0$, $\mathbf{C}_d$ is the data covariance matrix, $\mathbf{C}_m$ is the prior model covariance matrix, $\mathbf{D}$ is the second derivative smoothing operator [Trinh et al., 2017], ε and η are tuning factors supplied by users. When $\partial f(\mathbf{m})/\partial \mathbf{m} = 0$, the solution to Eq. (22) occurs as

$$\widetilde{\mathbf{m}} = \mathbf{m}_0 + \left(\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \varepsilon \mathbf{C}_m^{-1} + \eta \mathbf{D}^T \mathbf{D} \right)^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{G} \mathbf{m}_0), \tag{23}$$

with $\mathbf{G} = \partial \mathbf{g}(\mathbf{m})/\partial \mathbf{m}$ being the Fréchet derivative. We take the term $\mathbf{D}^T \mathbf{D}$ in Eq. (23) as smoothing factors, and then define the posterior covariance in a practical form as

$$\mathbf{C}_M = \left(\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \varepsilon \mathbf{C}_m^{-1} \right)^{-1}, \tag{24}$$

which slightly differs from Eq. (A.9) with an additional prefactor ε . In the approaches of [Bui-Thanh et al., 2013, Zhu et al., 2016], they both first calculate the data misfit Hessian $\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}$ and then evaluate the posterior covariance. $\mathbf{C}_d^{-1}$ measures the uncertainty of the data. When only taking its diagonal elements, we impose $\mathbf{C}_d^{-1}$ on the data as uncertainty-dependent weighting functions.

Note that we may determine the magnitude of $\mathbf{C}_m^{-1}$ but can hardly do the same on $\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}$ with the tuning factor ε . When $\varepsilon \rightarrow \infty$, $\mathbf{C}_M \rightarrow 0$. This result is misleading. When the magnitude of $\varepsilon \mathbf{C}_m^{-1}$ is much higher than $\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}$, $\mathbf{C}_M \rightarrow \mathbf{C}_m$. The model misfit will dominate the inversion. When $\varepsilon \rightarrow 0$, $\varepsilon \mathbf{C}_m^{-1}$ is taken as a stabilizer, $\mathbf{C}_M \rightarrow (\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G})^{-1}$, the inverse will be mainly dominated by the data misfit. The last case is more desirable than the others [Rawlinson et al., 2014]. The idea that the posterior covariance is closely related with the inverse Hessian in the vicinity of $\tilde{\mathbf{m}}$ can also be found in [Fichtner and Trampert, 2011a, Fichtner and van Leeuwen, 2015]. In this case, to well connect the prior and the posterior covariances, we may precondition $\mathbf{G}$ by multiplying $\mathbf{C}_m^{1/2}$. Thus, the final posterior covariance becomes

$$\mathbf{C}_M = \mathbf{C}_m \mathbf{H}^{-1}, \quad (25)$$

where $\mathbf{H}^{-1}$ can be found in Eqs. (18) and (21), in its original and SVD forms, respectively. The prior covariance $\mathbf{C}_m$ indicating the known information before waveform tomography can be estimated from undersurface orientations or well logs [Fomel and Claerbout, 2003, Zhu et al., 2016]. We leave the precise estimation about prior covariance as an open question in this paper. We take the diagonal of $\mathbf{C}_m$ as constants for the sake of simplified implementation. The square-root diagonal of $\mathbf{C}_M$, known as the standard variance, provides a quantitative measure of the posterior distribution.

4.4. Sampling prior and posterior distributions

Once obtaining a low-rank SVD approximation to the posterior covariance, we can draw a posterior Gaussian random sample on $\mathbf{C}_M$. The distribution sampling on the prior and posterior distributions can be, respectively, expressed as following [Tarantola, 2005]:

$$\begin{aligned} \mathbf{m}_{prior} &= \mathbf{m}_0 + \mathbf{C}_m^{1/2} \mathbf{n}, \\ \mathbf{m}_{post} &= \tilde{\mathbf{m}} + \mathbf{C}_M^{1/2} \mathbf{n}, \end{aligned} \quad (26)$$

where $\mathbf{n}$ is a 2D Gaussian random sampler in the 2D case, or a 3D Gaussian random sampler in the 3D case. has zero mean and identity covariance matrix [Tarantola, 2005]. We compute the square root of the posterior covariance matrix $\mathbf{C}_M^{1/2}$ in Eq. (25) by

$$\mathbf{C}_M^{1/2} = \mathbf{C}_m^{1/2} \mathbf{H}^{-1/2} = \mathbf{C}_m^{1/2} \mathbf{V} \mathbf{\Lambda}^{1/2} \mathbf{V}^T, \quad (27)$$

where $\mathbf{V}$ and $\mathbf{\Lambda}$ are the eigenvectors and eigenvalues of $\mathbf{H}^{-1}$, respectively. This way, we can assess the prior and posterior model uncertainties through visual comparisons of the random samplings on the prior and posterior distributions.

5. NUMERICAL EXAMPLES

In this part, we demonstrate our methods with the 2D acoustic Marmousi model. This demonstration includes: (i) the performance comparison of SRVM and L-BFGS in their applications to acoustic FWI; (ii) the eigenvectors and eigenvalues of inverse data-misfit Hessian via the RSVD algorithm; (iii) standard deviation map extracted from

the diagonals of posterior covariance; (iv) posterior model generation and the visualized comparisons between the prior and posterior distributions.

5.1. 2D Marmousi benchmark

The 2D acoustic Marmousi model is 9200m long and 3000m deep. Figs. 1a and 1b show the true and initial models, respectively. The acquisition system locates 10m below the surface, with 32 shots and 500 receivers evenly distributed. The source function is a 4-Hz Ricker wavelet. The time step is $9.0\text{e-}4$, and the total step number is 7500. We test this Marmousi FWI benchmark on a HPC (high-performance computing) cluster.

In FWI, L-BFGS is usually considered as the state-of-the-art optimization method in FWI [Modrak and Tromp, 2016], so we take it as reference for the SRVM-based FWI testing. Figs. 1c and 1d show the L-BFGS and SRVM based FWI inverted results, respectively. Fig. 2a shows that the SRVM-based FWI stops at the 32nd iteration, with data-misfit minimum being $2.04\text{e-}2$; the L-BFGS-based FWI stops at the 45th iteration, with data-misfit minimum being $1.86\text{e-}2$. By comparing the data-misfit curve, we find that although the SRVM-based FWI converges slower than the L-BFGS-based one at the early stage, the former catches up and then has a faster convergence speed. Similar performances can be observed in Fig. 2b regarding the model-misfit convergence comparison.

After the SRVM-based FWI is done, we can retrieve the inverse data-misfit Hessian from the stored SRVM vector and scalar series. The memory consumption for the SRVM scalars is negligible. One SRVM vector is at a memory-affordable size of Nm , with Nm being the size of the model. We have $Niter$ such vectors totally, with $Niter$ being the total iteration number. $Niter$ is far less than Nm . The vector-version SRVM based inverse data-misfit Hessian requires a storage size of $Nm*Niter$ rather than $Nm*Nm$, making it memory-affordable even for large-scale problems.

With the SRVM vectors and scalars ready, the information extraction algorithm in Alg. (19) is not so efficient, so we attempt to make it more convenient by using Alg. (20), namely, the SRVM-RSVD algorithm. The number of random samplers needs in the RSVD algorithm [Halko et al., 2011] equals the rank of the target matrix, which is $Niter$ in the vector-version SRVM. Therefore, we generate 32 independent random samplers to carry out the spectrum decomposition by RSVD. Due to that RSVD is an efficient algorithm, the computational cost of SRVM-RSVD is very cheap compared to that of one seismic modeling. Fig. 3 shows the eigenvalue curve, and Fig. 4 shows the selected eigenvectors for illustration. We observe from Fig. 4 that as the order of the eigenvector increase, the principal energy distributions gradually moves from bottom to top. We understand this trend with the help of Hessian, most of whose energy distribution should be distributed at the shallow part. In our methods, SRVM vectors and scalars try to approach the inverse Hessian, so it is reasonable to have this trend. Also, the spectrum decomposition in SVD form implies the feasibility to approach the data-misfit Hessian, which we will show next.

Eq. (25) expresses an approximated relationship between the prior and posterior covariances, in which the inverse Hessian acts as a big filter on the prior one. Assuming the unknowns are independent of each other, we use a homogeneous diagonal approximation to the prior covariance. Here, we make a simplified estimation by directly taking the square-rooted diagonal of inverse data-misfit Hessian as the 2D standard deviation map, as shown in Fig. 5. Note that both the magnitude and the appearance of our map resemble those in [Trinh et al., 2017], who use the Ensemble Kalman Filter (EnKF) method.

Fig. 5. provides a quantitative way to measure the uncertainties of the inversion results. The uncertainty map is related with the inverse Hessian, whose energy distribution fades away from top to bottom. We notice that some of the high-velocity structures have high uncertainties after inversion. For the high-velocity structures, seismic waves are easy to bend away rather than go through them, so these areas have less data coverage. For the marginal areas, for example, the PML zones, the uncertainty being zero seems reasonable. Because of the gradient-based method, we can not adjust areas Fréchet kernels always being zero. One can find similar explanations in [Kennett et al., 1988, Rawlinson et al., 2014].

We visualize difference between the prior and posterior. For an estimation about the exact prior distribution, we leave it as an open question, and set the same standard deviation 250m/s everywhere, so the term $\mathbf{C}_m^{1/2}\mathbf{n}$ ranges between [-1000 m/s, 1000 m/s]. Fig. 6 shows two different, independent 2D Gaussian random sampler [Tarantola, 2005]. We generate 1000 such independent samplers for the prior sampling with $\mathbf{C}_m^{1/2}\mathbf{n}$. We obtain the posterior sampling with $\mathbf{C}_M^{1/2}\mathbf{n} = \mathbf{C}_m^{1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}^T\mathbf{n}$. Fig. 7 shows three extracted prior distributions, and Fig. 8 shows three extracted posterior ones. By comparing Figs. 7 and 8, we can see that although the prior uncertainties are homogeneous everywhere, the posterior uncertainties become inhomogeneous due to the information gain from observed data. Fig. 9 shows detailed investigations about the prior and posterior distribution comparisons. From them we notice that both the prior and posterior distributions are in Gaussian shapes, but the posterior ones come into being more concentrated. Also, there is a shift between the prior and posterior ones. The information gain from observed data by FWI makes the concentration and shift.

A 2D panoramic view of $\mathbf{C}_m^{1/2}\mathbf{n}$ is similar with $\mathbf{n}$ but differs in magnitude. Fig. 10 shows four 2D panoramic views about the sampling of $\mathbf{C}_M^{1/2}\mathbf{n}$. Each sample is obtained by a matrix-vector product with different $\mathbf{n}$. We can see that although the prior samples look random, the posterior ones turn into being continuous because of the contribution from $\mathbf{H}^{-1/2}$. One interesting aspect we notice from $\mathbf{C}_M^{1/2}\mathbf{n}$ is that it has the pattern of the maximum a posteriori model. The variance of $\mathbf{C}_M^{1/2}\mathbf{n}$ between different models help us to analyse the posterior uncertainties.

6. Discussion and conclusions

We formulate the SRVM algorithm in a vector version to make it memory-affordable in FWI applications. A main advantage of using the vector-version SRVM algorithm is the possibility of reconstructing the associated inverse Hessian for resolution analysis and uncertainty estimation in FWI. After the SRVM-based FWI converges, the SRVM scalar and vector series inherently captures the information for the reconstruction of inverse data-misfit Hessian. To make the SRVM-related reconstruction more efficient we factorise it into eigenvalues and eigenvectors by using RSVD. Furthermore, we relate the inverse Hessian with the posterior covariance and assess the uncertainty of the posterior model parameters. Because the number of SRVM vectors and scalars only increase with the iteration number, we will explore its application to large-scale problems. Based on this acoustic study, we will extend our methods to elastic cases in the future.

Appendix A. Review of Bayesian Inference

Bayesian inference is a method to update the prior distribution to the posterior distribution by making use of the observed information [Tarantola, 2005]. In the framework of FWI, the Bayesian inference allows us to update the prior model covariance by using waveform tomography to estimate the posterior model covariance.

Let us start from the forward problem

$$\mathbf{d} = \mathbf{g}(\mathbf{m}), \quad (\text{A.1})$$

where $\mathbf{d}$ is the observed data, $\mathbf{m}$ the parameter model, and $\mathbf{g}(\cdot)$ the nonlinear operator. In this paper, Eq. (A.1) represent the numerical solution of acoustic wave-equation by using the spectral-element method [Komatitsch and Vilotte, 1998, Komatitsch and Tromp, 1999]. Given the knowledge of the a priori information, we express the prior probability distribution function (PDF) as multidimensional Gaussian distributions

$$\rho(\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_0)^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{m}_0)\right), \quad (\text{A.2})$$

where $\mathbf{m}_0$ is the prior mean model (initial model) and $\mathbf{C}_m$ is the prior model covariance. $\mathbf{C}_m$ is an indicator of the uncertainty of $\mathbf{m}_0$, which can be inferred from geophysical surveys, for example, well log samples. Similarly, we express the likelihood function in multidimensional Gaussian distributions as

$$\rho(\mathbf{d}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{g}(\mathbf{m}))^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{m}))\right), \quad (\text{A.3})$$

where $\mathbf{C}_d$ denotes the data uncertainty matrix, for example, the measurement inadequacy or theory error. The existence of $\mathbf{C}_d$ means that even when the parameter model $\mathbf{m}$ happens to be the true model, the synthetic data may still differ from the observed one.

According to the Bayesian Inference theory, the solution to an inverse problem yields a posteriori PDF as

$$\rho(\mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{m})\rho(\mathbf{m}), \quad (\text{A.4})$$

in which $\rho(\mathbf{m}|\mathbf{d})$ indicates the posterior PSF, $\rho(\mathbf{d}|\mathbf{m})$ the likelihood function, $\rho(\mathbf{m})$ the prior PSF. Here, $\rho(\mathbf{m})$ is related to the model misfit, and $\rho(\mathbf{d}|\mathbf{m})$ is related to the data misfit. Combining Eqs. (A.2) and (A.3) into (A.4), we have

$$\rho(\mathbf{m}|\mathbf{d}) \propto \exp(-f(\mathbf{m})), \quad (\text{A.5})$$

where $f(\mathbf{m})$ is the least-squares misfit function

$$f(\mathbf{m}) = \frac{1}{2} \left((\mathbf{d} - \mathbf{g}(\mathbf{m}))^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{g}(\mathbf{m})) + (\mathbf{m} - \mathbf{m}_0)^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{m}_0) \right). \quad (\text{A.6})$$

Following the form of Eq. (A.2), we rewrite Eq. (A.5) as

$$\rho(\mathbf{m}|\mathbf{d}) \propto \exp\left(-\frac{1}{2}(\mathbf{m} - \widetilde{\mathbf{m}})^T \mathbf{C}_M^{-1} (\mathbf{m} - \widetilde{\mathbf{m}})\right), \quad (\text{A.7})$$

where $\widetilde{\mathbf{m}}$ corresponds to the global minimum of Eq. (A.6), namely, the maximum a posteriori probability (MAP) estimate of Eq. (A.5); $\mathbf{C}_M$ is the posterior model covariance. $\mathbf{C}_M$ indicates the uncertainty of $\widetilde{\mathbf{m}}$. When the minimum of Eq. (A.6) reaches, $\widetilde{\mathbf{m}}$ reads [Tarantola, 2005]

$$\widetilde{\mathbf{m}} = \mathbf{m}_0 + \left(\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \mathbf{C}_m^{-1} \right)^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{G}\mathbf{m}_0), \quad (\text{A.8})$$

in which $\mathbf{G} = \partial \mathbf{g}(\mathbf{m}) / \partial \mathbf{m}$ denotes the Fréchet derivative. The direct calculation of the Fréchet kernel according to its definition is impossible for practical problems. An efficient solution to this problem is the adjoint method [Tromp et al., 2005], which estimate the Fréchet derivative by cross-correlating the forward and adjoint wavefields. We express the posterior covariance as

$$\mathbf{C}_M = \left(\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \mathbf{C}_m^{-1} \right)^{-1} = \left(\mathbf{H} + \mathbf{C}_m^{-1} \right)^{-1}, \quad (\text{A.9})$$

in which $\mathbf{H} = \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}$ is known as the Gauss-Newton approximation to the data-misfit Hessian [Pratt, 1999, Virieux et al., 2009]. By looking into Eq. (A.9), we realize that the critical step towards the uncertainty estimation is to efficiently quantify the second-order derivative, the Hessian or the inverse Hessian.

- [An, 2012] An, M., 2012, A simple method for determining the spatial resolution of a general inverse problem: *Geophysical Journal International*, **191**(2), 849–864.
- [Backus and Gilbert, 1968] Backus, G., and F. Gilbert, 1968, The resolving power of gross Earth data: *Geophysical Journal International*, **16**(2), 169–205.
- [Backus and Gilbert, 1970] ———, 1970, Uniqueness in the inversion of inaccurate gross Earth data: *Phil. Trans. R. Soc. Lond. A*, **266**(1173), 123–192.
- [Bui-Thanh et al., 2013] 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**(6), A2494–A2523.
- [Carlson, 1973] Carlson, N. A., 1973, Fast triangular formulation of the square root filter: *AIAA journal*, **11**(9), 1259–1265.
- [da Silva and Yao, 2017] da Silva, N. V., and G. Yao, 2017, Wavefield reconstruction inversion with a multiplicative cost function: *Inverse Problems*, **34**(1), 015004.
- [Evensen, 1994] Evensen, G., 1994, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics: *Journal of Geophysical Research: Oceans*, **99**(C5), 10143–10162.
- [Fichtner and Trampert, 2011a] Fichtner, A., and J. Trampert, 2011a, Hessian kernels of seismic data functionals based upon adjoint techniques: *Geophysical Journal International*, **185**(2), 775–798.
- [Fichtner and Trampert, 2011b] ———, 2011b, Resolution analysis in full waveform inversion: *Geophysical Journal International*, **187**(3), 1604–1624.
- [Fichtner and van Leeuwen, 2015] Fichtner, A., and T. van Leeuwen, 2015, Resolution analysis by random probing: *Journal of Geophysical Research: Solid Earth*, **120**(8), 5549–5573.
- [Fletcher and Powell, 1963] Fletcher, R., and M. J. Powell, 1963, A rapidly convergent descent method for minimization: *The computer journal*, **6**(2), 163–168.
- [Fomel and Claerbout, 2003] Fomel, S., and J. F. Claerbout, 2003, Multidimensional recursive filter preconditioning in geophysical estimation problems: *Geophysics*, **68**(2), 577–588.
- [Gouveia and Scales, 1998] Gouveia, W. P., and J. A. Scales, 1998, Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis: *Journal of Geophysical Research: Solid Earth*, **103**(B2), 2759–2779.
- [Halko et al., 2011] Halko, N., P.-G. Martinsson, and J. A. Tropp, 2011, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions: *SIAM review*, **53**(2), 217–288.
- [Huang, 1970] Huang, H.-Y., 1970, Unified approach to quadratically convergent algorithms for function minimization: *Journal of Optimization Theory and Applications*, **5**(6), 405–423.
- [Hull and Tapley, 1977] Hull, D. G., and B. D. Tapley, 1977, Square-root variable-metric methods for minimization: *Journal of optimization theory and applications*, **21**(3), 251–259.
- [Jordan, 2015] Jordan, M., 2015, Estimation of spatial uncertainties in tomographic images.
- [Kalman et al., 1960] Kalman, R. E., et al., 1960, A new approach to linear filtering and prediction problems: *Journal of basic Engineering*, **82**(1), 35–45.
- [Kennett et al., 1988] Kennett, B., M. Sambridge, and P. Williamson, 1988, Subspace methods for large inverse problems with multiple parameter classes: *Geophysical Journal International*, **94**(2), 237–247.
- [Komatitsch and Tromp, 1999] Komatitsch, D., and J. Tromp, 1999, Introduction to the spectral element method for three-dimensional seismic wave propagation: *Geophysical Journal International*, **139**(3), 806–822.
- [Komatitsch and Vilotte, 1998] Komatitsch, D., and J. P. Vilotte, 1998, The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures: *Bull. seism. Soc. Am.*, **88**(2), 368–392.
- [Liberty et al., 2007] Liberty, E., F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert, 2007, Randomized algorithms for the low-rank approximation of matrices: *Proceedings of the National*

- Academy of Sciences, **104**(51), 20167–20172.
- [Liu and Nocedal, 1989] Liu, D. C., and J. Nocedal, 1989, On the limited memory BFGS method for large scale optimization: *Mathematical programming*, **45**(1), 503–528.
- [Luo, 2012] Luo, Y., 2012, Seismic imaging and inversion based on spectral-element and adjoint methods: Ph.D. thesis.
- [Modrak and Tromp, 2016] Modrak, R., and J. Tromp, 2016, Seismic waveform inversion best practices: regional, global and exploration test cases: *Geophysical Journal International*, **206**(3), 1864–1889.
- [Morf and Kailath, 1975] Morf, M., and T. Kailath, 1975, Square-root algorithms for least-squares estimation: *IEEE Transactions on Automatic Control*, **20**(4), 487–497.
- [Mosegaard and Tarantola, 1995] Mosegaard, K., and A. Tarantola, 1995, Monte Carlo sampling of solutions to inverse problems: *Journal of Geophysical Research: Solid Earth*, **100**(B7), 12431–12447.
- [Nocedal and Wright, 2006] Nocedal, J., and S. Wright, 2006, *Numerical optimization*, 2nd ed.: Springer-Verlag. Springer Series in Operations Research and Financial Engineering.
- [Osypov et al., 2013] 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**(6), 1114–1134.
- [Pratt, 1999] Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model: *Geophysics*, **64**(3), 888–901.
- [Rawlinson et al., 2014] Rawlinson, N., A. Fichtner, M. Sambridge, and M. K. Young, 2014, Chapter one-seismic tomography and the assessment of uncertainty: *Advances in Geophysics*, **55**, 1–76.
- [Sambridge and Mosegaard, 2002] Sambridge, M., and K. Mosegaard, 2002, Monte Carlo methods in geophysical inverse problems: *Reviews of Geophysics*, **40**(3).
- [Tarantola, 1984] Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: *Geophysics*, **49**, 1259–1266.
- [Tarantola, 1987] ———, 1987, *Inverse problem theory: Methods for data fitting and model parameter estimation*: Elsevier Science Publ. Co., Inc.
- [Tarantola, 2005] ———, 2005, *Inverse problem theory and methods for model parameter estimation*: Society for Industrial and Applied Mathematics.
- [Thurin et al., 2017] Thurin, J., R. Brossier, and L. Métivier, 2017, Ensemble-based uncertainty estimation in full waveform inversion.
- [Trampert et al., 2012] Trampert, J., A. Fichtner, and J. Ritsema, 2012, Resolution tests revisited: the power of random numbers: *Geophysical Journal International*, **192**(2), 676–680.
- [Trinh et al., 2017] Trinh, P., R. Brossier, L. Métivier, J. Virieux, and P. Wellington, 2017, Bessel smoothing filter for spectral-element mesh: *Geophysical Journal International*, **209**(3), 1489–1512.
- [Tromp et al., 2005] Tromp, J., C. Tape, and Q. Liu, 2005, Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels: *Geophysical Journal International*, **160**(1), 195–216.
- [Virieux and Operto, 2009] Virieux, A., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: *Geophysics*, **74**(6), WCC1–WCC26.
- [Virieux et al., 2009] Virieux, J., S. Operto, H. Ben-Hadj-Ali, R. Brossier, V. Etienne, F. Sourbier, L. Giraud, and A. Haidar, 2009, Seismic wave modeling for seismic imaging: *The Leading Edge*, **28**(5), 538–544.
- [Williamson, 1975] Williamson, W. E., 1975, Square-root algorithms for function minimization: *AIAA Journal*, **13**(1), 107–109.
- [Wu and Alkhalifah, 2016] Wu, Z., and T. Alkhalifah, 2016, The optimized gradient method for full waveform inversion and its spectral implementation: *Geophysical Journal International*, **205**(3), 1823–1831.
- [Wu and Alkhalifah, 2017] ———, 2017, Efficient scattering-angle enrichment for a nonlinear inversion of the background and perturbations components of a velocity model: *Geophysical Journal International*, **210**(3), 1981–1992.
- [Yao et al., 2018] Yao, G., N. V. da Silva, M. Warner, and T. Kalinicheva, 2018, Separation of

migration and tomography modes of full-waveform inversion in the plane wave domain: *Journal of Geophysical Research: Solid Earth*, **123**(2), 1486–1501.

[Zhang and McMechan, 1995] Zhang, J., and G. A. McMechan, 1995, Estimation of resolution and covariance for large matrix inversions: *Geophysical Journal International*, **121**(2), 409–426.

[Zhu et al., 2016] 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.

Figure 1. (a) True V_p model used in the Marmousi testing for data generation. (b) Initial model obtained after Gaussian smoothers on (a). (c) Inverted FWI result by L-BFGS. (d) Inverted FWI result by SRVM.

Figure 2. Comparisons of (a) data-misfit and (b) model-misfit curves against iterations by SRVM (in red) and L-BFGS (in blue). We can see that the performance of SRVM is close to that of L-BFGS in acoustic FWI.

Figure 3. Eigenvalues extracted by RSVD algorithm from the stored 33 SRVM vectors in the acoustic Marmousi testing.

Figure 4. Eigenvectors factorized by RSVD algorithm from the stored 33 SRVM vectors in the Marmousi testing. We can see that as the eigen-order increases, the energy distributions of these eigenvectors gradually move from bottom to top.

Figure 5. Standard deviation from diagonals of the posterior covariance in the Marmousi testing. The area with higher value has more uncertainty, and vice versa.

Figure 6. Two independent 2D Gaussian random fields, with zero mean and identity covariance matrix, are used to sample the prior and posterior distributions. Note that we draw 1000 such samplers, and each of them is independent of the others. The 2D random fields are only for the sampling of distributions.

Figure 7. Prior distributions of Vp in acoustic Marmousi testing. Top row from left to right: Prior distributions against the depth of the Vp model at X=3100m, 4600m, 6900m. Bottom row from left to right: Prior distributions against the depth of the Vp deviation at X=3100m, 4600m, 6900m.

Figure 8. Posterior distributions of V_p in acoustic Marmousi testing. Top row from left to right: Posterior distributions against the depth of the V_p model at $X=3100\text{m}$, 4600m , 6900m . Bottom row from left to right: Posterior distributions against the depth of the V_p deviation at $X=3100\text{m}$, 4600m , 6900m .

Figure 9. Comparison of prior and posterior V_p distributions at different locations in the Marmousi testing. The individual locations are titled on each subplot.

Figure 10. Four random samples drawn from the posterior distributions about the Vp model deviations in the acoustic Marmousi testing. Because of the diagonal prior covariance being constant, the random samples from prior distributions should be similar to those in Fig. 6 such that we do not plot them out. The reason we call the figures sampled Vp deviation is that here we are plotting $\mathbf{C}_M^{1/2} \mathbf{n}$, which is one term in Eq. (26).