Square-root variable metric based elastic full-waveform inversion – Part 1: theory and validation

Qiancheng Liu 1, Daniel Peter1 and Carl Tape 2

1 Division of Physical Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: daniel.peter@kaust.edu.sa
2 Geophysical Institute and Department of Geosciences, University of Alaska Fairbanks, Fairbanks, AK 99775, USA

SUMMARY
Full-waveform inversion (FWI) has become a powerful tool in inverting subsurface geophysical properties. The estimation of uncertainty in the resulting Earth models and parameter trade-offs, although equally important to the inversion result, has however often been neglected or became prohibitive for large-scale inverse problems. Theoretically, the uncertainty estimation is linked to the inverse Hessian (or posterior covariance matrix), which for massive inverse problems becomes impossible to store and compute. In this study, we investigate the application of the square-root variable metric (SRVM) method, a quasi-Newton optimization algorithm, to FWI in a vector version. This approach allows us to reconstruct the final inverse Hessian at an affordable storage memory cost. We conduct SRVM-based elastic FWI on several elastic models in regular, free-surface and practical cases. Comparing the results with those obtained by the state-of-the-art L-BFGS (limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm, we find that the proposed SRVM method performs on a similar, highly efficient level as L-BFGS, with the advantage of providing additional information such as the inverse Hessian needed for uncertainty quantification.

Key words: Inverse theory; Waveform inversion; Computational seismology; Seismic tomography.

1 INTRODUCTION
Seismic waves carry valuable information about Earth’s internal structures by probing the media properties on their passage from the origin to receiver locations. In exploration as well as regional- and global seismology, seismic measurements on the surface represent the main source of structural information. Over the last decades, full-waveform inversion (FWI) has become a well-known method in estimating subsurface parameters quantitatively (Lailly 1983; Tarantola 1984; Pratt & Worthington 1990). FWI outputs high-quality results in both academic and industrial applications (Tape et al. 2009, 2010; Fichtner et al. 2009; Plessix 2009; Sirgue et al. 2010; Zhu et al. 2012; Warner et al. 2013; Fichtner et al. 2013; Operto et al. 2015; Bozdag et al. 2016). However, these valuable outputs require to solve a large-scale inverse problem using massive seismic datasets.

Since its advent, FWI has been the focus of many research studies and impressive advances have been made. However, its uncertainty estimation is still left behind with only a few studies trying to tackle this issue. Fichtner & Trampert (2011a,b) introduce the concept of ‘point-spread function’ (PSF) and formulate the Fourier components of the Hessian kernels to analyse the resolution of FWI. Martin et al. (2012) present a Hessian-based Markov Chain Monte Carlo (MCMC) method to sample the posterior information of a large-scale seismic inversion. An (2012) approximates the resolution operators with a Gaussian function to calculate the spatially varying length. Trampert et al. (2012) obtain the resolution lengths for some tomographic models by using random sampling. Bui-Thanh et al. (2013) take both the data uncertainties and prior model information into account to formulate the seismic inversion in a Bayesian inference framework. Fichtner & Leeuwen (2015) use stochastic probing to the Hessian operators to analyse the resolution lengths for seismic tomography. Zhu et al. (2016) adopt the Bayesian inference framework to estimate the model uncertainties in FWI by using the PSFs of the Hessian. Most of these studies above are based on the Hessian or inverse Hessian which is closely related to the posterior covariance matrix.
In this research, we aim to approximate the Hessian or inverse Hessian more directly by using the square-root variable metric (SRVM) method. The concept of variable metric (VM) was introduced by Davidon (1959) and then simplified by Fletcher & Powell (1963) as the Davidson–Fletcher–Powell (DFP) algorithm. Another VM algorithm, the dual of DFP, is the well-known Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. The modified version of BFGS, limited-memory BFGS (L-BFGS, Nocedal 1980), has become the most popular optimization method in applied mathematics. In contrast, SRVM (Williamson 1975; Morf & Kailath 1975; Hull & Tapley 1977), which preserves the positive definiteness of the inverse Hessian in a square-root form, rarely attracts attention. The reason might be that SRVM is considered to be more expensive and complicated than L-BFGS, without further benefits. Tarantola (2005) unburies SRVM and highlights the fact that this method allows for a model resolution analysis and the sampling of the posterior model distribution (Luo 2012). Furthermore, SRVM overcomes the shortcomings of DFP such as mandatory exact line search and susceptibility to round-off errors (Williamson 1975; Hull & Tapley 1977).

This paper represents the first part of our research focusing on the theoretical development and application of the SRVM method in elastic FWI. In a second paper, we will address the upcoming resolution analysis and uncertainty estimation. The SRVM algorithm presented here is written in a matrix-free vector version to make it memory-affordable and applicable to large-scale problems. The vector-version SRVM workflow is easy to implement in a recursive manner and can be adopted on high-performance computing (HPC) facilities. In the SRVM-based elastic FWI, we will use the inexact line search for effectiveness as SRVM is able to overcome the shortcoming of mandatory exact line searches. To evaluate its performance, we will compare it to the widely popular L-BFGS-based elastic FWI for reference.

We arrange this paper as follows: first, we briefly review the theory of classic SRVM and then formulate it in a vector-version workflow. Second, we present basic methods on elastic FWI and discuss how to incorporate SRVM in these frameworks. Finally, we conduct the SRVM-based elastic FWI on four numerical examples in regular, free-surface and practical cases, and then compare the convergence and model results to the state-of-the-art L-BFGS outputs. As we hoped, the SRVM performance is comparable to L-BFGS in our applications. The primary advantage of SRVM over L-BFGS is in its ability to quantify uncertainties, a topic addressed in a companion paper (Liu & Peter 2019).
2 THEORY

Let us first review the SRVM method briefly and then express the SRVM algorithm in a storage-efficient vector version.

2.1 Review of the square-root variable metric method

In Newton’s method, we approximate a function $f(x)$ around $x_0$ in a second-order Taylor series as

$$f(m_0 + \Delta m) \approx f(m_0) + g^T \Delta m + \frac{1}{2} \Delta m^T H \Delta m,$$

where the gradient $g = \nabla f(m_0)$ is a vector and the Hessian $H = \nabla^2 f(m_0)$ a matrix for higher dimensions. The perturbation of this approximation with respect to $\Delta m$ reads

$$\nabla f(m_0 + \Delta m) \approx g + H \Delta m.$$

To meet the objective of optimization, we set the left-hand side of eq. (2) to be zero, yielding

$$\Delta m \approx -\mu H^{-1} g,$$

where $\mu$ is a scalar representing the search step. Eq. (3) is also known as the Newton update and involves knowledge about the inverse Hessian. Furthermore, we see that the inverse Hessian can be considered a pre-conditioner on the gradient vector. Based on eq. (3), the optimal pre-conditioner matrix is the inverse Hessian. Finding the inverse Hessian however is often an expensive operation and thus modifications have been formulated based on approximations of the Hessian, resulting in so-called quasi-Newton methods. The original idea of Davidon (1959) was to use differences in the gradients to approximate the Hessian. Following the standard DFP formula, we update the inverse Hessian $B = H^{-1}$ as following:

$$B_{k+1} = B_k - \frac{B_k \Delta g_k \Delta g_k^T B_k}{\Delta g_k^T B_k \Delta g_k} + \frac{\Delta m_k \Delta m_k^T}{\Delta g_k^T \Delta g_k},$$

where $\Delta m_k = m_{k+1} - m_k$.
where $\Delta g_k = g_{k+1} - g_k$. We repeat the updating procedure in eq. (4) until the minimization process converges. The DFP algorithm is effective, but soon it was overtaken by L-BFGS due to the following reason: the DFP formula is sensitive to the line-search strategy and round-off errors. If the line search is not implemented very accurately or $B$ is singular at some point, a non-positive inverse Hessian will arise in DFP. In such cases, the formula cannot guarantee a strict decrease in the minimization optimization and one must restart $B$ from the identity matrix, setting $B = I$.

To circumvent such shortcomings and to ensure that the matrix $B$ is always positive, Williamson (1975) observed that eq. (4) could be expressed in a square-root form:

$$ S_{k+1} S_{k+1}^T = S_k \left( I - \frac{R_k}{P_k} S_k^T y_k y_k^T S_k \right) S_k^T, $$

where $B_{k+1} = S_{k+1} S_{k+1}^T$ and $B_k = S_k S_k^T$ are rewritten in terms of a square-root matrix $S$ of $B$, and introducing the auxiliary vectors $y_k = \mu_k g_k + \Delta g_k$ and scalars $P_k = y_k^T B_k y_k$, with $\mu_k$ being the search step scalar value at the $k$th iteration. For simplification, we also rewrite the term within brackets in a square-root form as

$$ A_k A_k^T = I - \frac{R_k}{P_k} S_k^T y_k y_k^T S_k, $$

where $A_k = I - (v_k / P_k) S_k^T y_k y_k^T S_k$. The constant $v_k$ is determined as

$$ v_k = \frac{1 \pm \sqrt{1 - Q_k / P_k}}{Q_k / P_k}, $$

with $Q_k = y_k^T S_k S_k^T y_k$, such that the relationship in eq. (6) holds. In eq. (7), we need to choose the minus sign to prevent $v_k$ from being singular at $Q / P = 0$; we also observe that if $Q / P > 1$, $v_k$ is imaginary, resulting in a non-real matrix $S$. Hull & Tapley (1977) suggest using $v_k = 1$ for $Q / P > 1$, because $0 \leq v_k \leq 1$ when $-\infty \leq Q / P \leq 1$. We finally update the square root of the inverse Hessian as following

$$ S_{k+1} = S_k \left( I - (v_k / P_k) S_k^T y_k y_k^T S_k \right). $$
Figure 4. Inverted $V_3$ models by L-BFGS and SRVM-based FWIs of two-Sphere, Checkerboard, Marmousi and Overthrust, respectively.

Figure 5. Comparison of data misfits by L-BFGS- and SRVM-based FWIs for two-Sphere, Checkerboard, Marmousi and Overthrust models, respectively.
Figure 6. Comparison of $V_p$ model misfits by L-BFGS- and SRVM-based FWIs on two-Sphere, Checkerboard, Marmousi and Overthrust models, respectively.

Figure 7. Comparison of $V_S$ model misfits by L-BFGS- and SRVM-based FWIs on two-Sphere, Checkerboard, Marmousi and Overthrust models, respectively.

By this, the updating workflow of SRVM has been briefly described. We notice that eq. (8) is still in a matrix form. Although we can define $S$ by an analytic Cholesky decomposition (Carlson 1973), the memory requirement of $S$ is still at half of matrix $B$ in eq. (4), which remains difficult to store for large-scale, practical problems. In the next section, we will rewrite the SRVM method in a vector-version form to alleviate these storage costs and make the implicit memory requirements of the algorithm more affordable.
Figure 8. Inverted $V_p$ models by L-BFGS- and SRVM-based FWIs of free-surface Checkerboard and Marmousi models, respectively.

Figure 9. Inverted $V_s$ models by L-BFGS- and SRVM-based FWIs of free-surface Checkerboard and Marmousi models, respectively.

Figure 10. Comparison of data misfits by L-BFGS- and SRVM-based FWIs for free-surface Checkerboard and Marmousi models, respectively.

Figure 11. Comparison of $V_p$ model misfits by L-BFGS and SRVM based FWIs for free-surface Checkerboard and Marmousi models, respectively.
Figure 12. Comparison of $V_S$ model misfits by L-BFGS and SRVM for free-surface Checkerboard and Marmousi models, respectively.

Figure 13. The data and model misfits of the free-surface Marmousi by the second-stage L-BFGS-based FWI, which continues from the inverted models by SRVM-based FWI.

Figure 14. A practical Marmousi-X model with variable $V_p/V_S$ ratio and a strip intrusion into $V_S$. The $V_p/V_S$ ratio decreases linearly from 2 to 1.5 with depth. Top and bottom: true and initial $V_p$ and $V_S$ models, respectively.
2.2 Square-root variable metric method in vector version

For practical geophysical problems, such as seismic 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:

**Algorithm 1 SRVM workflow**

1. for $k \leftarrow 0$ to $n$ do
2. $\Delta g_k = g_{k+1} - g_k$ ▷ Gradient update
3. $y_k = \mu_k g_k + \Delta g_k$
4. $P_k = y_k^T S_k y_k$ ▷ Scaling factors
5. $Q_k = y_k^T S_k^T y_k$
6. $\nu_k = \frac{1 - \sqrt{1 - Q_k / P_k}}{Q_k / P_k}$
7. $S_{k+1} = S_k \left( I - \frac{\nu_k}{P_k} S_k y_k y_k^T S_k \right)$ ▷ Matrix update
8. $B_{k+1} = S_{k+1}^T S_{k+1} = S_k \left( I - \frac{\nu_k}{P_k} S_k y_k y_k^T S_k \right) \left( I - \frac{\nu_k}{P_k} S_k y_k y_k^T S_k \right)^T S_k$
9. end for
In Algorithm (1), we observe that the term $y_k^T S_k$ and its transpose often occur. Also, by setting $\beta = S_k^T \Delta g_k$, we can write a vector-version SRVM workflow as following:

**Algorithm 2 SRVM vector-version workflow**

$$
\begin{align*}
1: & \text{ for } k \leftarrow 0 \text{ to } n \text{ do} \\
2: & \Delta g_k = g_{k+1} - g_k \\
3: & y_k = \mu g_k + \Delta g_k \\
4: & w_k = S_k^T y_k \\
5: & \beta = S_k^T \Delta g_k \\
6: & P_k = w_k^T \beta \\
7: & Q_k = w_k^T w_k \\
8: & v_k = \frac{1}{2} \nu \beta^T P_k^{-1} \\
9: & S_{k+1} = S_k \left( 1 - \frac{\nu}{\beta} w_k w_k^T \right) \\
10: & B_{k+1} = S_{k+1} S_{k+1}^T = S_k \left( 1 - \frac{\nu}{\beta} w_k w_k^T \right) \left( 1 - \frac{\nu}{\beta} v_k v_k^T \right) S_k^T \\
11: & \text{ end for}
\end{align*}
$$

Note that we do not need to store $\beta$ because it is simply a temporary term per iteration. Both algorithms implicitly assume that we begin from an initial model $m_0$, which is a vector containing $M$ parameters. In Algorithm (2), we observe that every entity is at size $M \times M$. Like L-BFGS, we choose an estimation of the initial inverse Hessian as $B_0 = I$, which is obviously symmetric positive-definite, and leads to $S_0 = I$ as well. This way, we do not need to explicitly store $S$ in memory but instead only store one vector $w_k$ and one scalar $v_k/P_k$ for all the iteration steps to reconstruct or extract elements from $S$. Furthermore, when applying the vector-version SRVM algorithm to elastic FWI, we only implement a matrix-vector multiplication to find out the search direction by $p_{k+1} = -S_{k+1} S_{k+1}^T g_{k+1}$, where $p_{k+1}$ is also a vector.

### 3 ELASTIC FULL-WAVEFORM INVERSION

In general, an FWI is composed of three recursive steps: first, the computation of the gradient, e.g. by the adjoint method; second, an update of the gradient to approach a search direction, chosen by the optimization method; and finally, the determination of a step length along the search direction, e.g. through line searching and use as initial model for the next consecutive iteration. This final step also involves a convergence criteria which determines whether to continue or stop the recursive algorithm. Let us describe these three steps in more detail for the investigation done of the SRVM method in this study.

#### 3.1 Sensitivity kernels

According to Tarantola (1984, 1987) and Tromp et al. (2005), the waveform misfit function in a least-squares sense measuring the goodness of fit reads

$$
f(m) = \frac{1}{2} \sum_{r=1}^{N} \int_0^T |s(x_r, t; m) - d(x_r, t)|^2 dt, \tag{9}
$$

where $d(x_r, t)$ is the observed seismogram at receiver $x_r$ with record duration $T$ for $N$ receivers in total, $s(x_r, t; m)$ the corresponding synthetic trace computed for a given model $m$. We can characterize an isotropic, elastic model $m$ using a parametrization of $(\rho, \kappa, \mu)$, with $\rho$ being density, $\kappa$ bulk modulus and $\mu$ shear modulus (Tarantola 2005; Tromp et al. 2005; Luo et al. 2009). Then, the perturbation of the misfit function $\delta f$ is related to the logarithmic parameters as

$$
\delta f(m) = \int (K_\rho \delta \ln \rho + K_\mu \delta \ln \mu + K_\kappa \delta \ln \kappa) d^3 x. \tag{10}
$$

The sensitivity kernels, i.e. the Fréchet derivatives, with respect to $(\rho, \mu, \kappa)$ can be expressed as (Tromp et al. 2005)

$$
K_\rho(x) = - \int_0^T \rho(x) \partial_t s^T(x, T - t) \partial_t s(x, t) dt \tag{11}
$$

$$
K_\mu(x) = - \int_0^T 2 \mu(x) \partial_t D^T(x, T - t) \partial_t D(x, t) dt \tag{12}
$$
\[
K_s(x) = - \int_0^T \kappa(x) \left[ \nabla \cdot \mathbf{s}^\dagger(x, T-t) \right] [\mathbf{s}(x, t)] dt
\] (13)

where \( \mathbf{s}^\dagger \) denotes the adjoint wavefield, \( \mathbf{D} = \frac{1}{2} \left[ \nabla \mathbf{s} + (\nabla \mathbf{s})^T \right] - \frac{1}{2} (\nabla \cdot \mathbf{s}) \mathbf{I} \) and \( \mathbf{D}^T \) the traceless strain deviator and its adjoint, respectively. The adjoint wavefield is generated by backward propagating the data residual \( \mathbf{f} = \mathbf{s}(x, t; \mathbf{m}) - \mathbf{d}(x, t) \). In this paper, we ignore the challenging density inversion (Wu & Aki 1985; Tarantola 1986; Forgues & Lambaré 1997; Virieux & Operto 2009), and only focus on the inversion of \( P \)- and \( S \)-wave velocities with their sensitivity kernels as (Tromp et al. 2005)

\[
K_{1P}(x) = 2 \left( \frac{k + \frac{4}{3} \mu}{k} \right) K_s(x)
\] (14)

\[
K_{1S}(x) = 2 \left( K_{1S}(x) - \frac{4}{3} \frac{k}{\mu} K_s(x) \right)
\] (15)
in which \( K_{1S} \) and \( K_s \) have been given above by eqs (12) and (13), respectively.

### 3.2 SRVM method in elastic FWI

Elastic FWI falls into the category of ‘multiparameter’ inversions. The choice of parameterization for such inversions becomes important as it relates to the trade-off between different parameters. The sensitivity kernels above calculate the Fréchet derivatives and when combining them with an optimization algorithms such as L-BFGS, we can seek for the minimum of the misfit function of elastic FWI in eq. (9) for large-scale problems. However, both a blurring effect as well as a coupling effect between \( P \) and \( S \) waves due to crosstalk are present in the adjoint operation. These effects are incorporated in the Hessian and can in principle be suppressed by a proper approximation to the inverse Hessian, by either direct or iterative approaches. Furthermore, considering the quadratic term, i.e. the Hessian, in the inversion can significantly accelerate the convergence rate and improve the final result of an FWI (Virieux & Operto 2009). Even more important, the (inverse) Hessian contains valuable information for a resolution analysis and uncertainty estimation in waveform tomography (Sambridge & Mosegaard 2002; Fichtner & Trampert 2011a,b). We will discuss this further in a following paper highlighting the second part of this research study. In this first part presented here, we focus on the application of SRVM to elastic FWI and the comparison of its performance against the L-BFGS algorithm.

In numerical optimization, the L-BFGS algorithm (Liu & Nocedal 1989) is considered as the most efficient quasi-Newton method in solving unconstrained, non-linear optimization problems (Nocedal 1992). The algorithm uses as input the current gradient, approximates with a limited amount of computer memory the inverse Hessian, and finally outputs a new search direction by using the matrix-vector product of the two. The L-BFGS method uses mainly vector operations, thus is well suited for optimization problems with a large number of unknowns.

Unlike the L-BFGS method, the SRVM method derives from the DFP algorithm, the dual of the BFGS algorithm. In the square-root form, the SRVM method overcomes two main limitations of the conventional DFP method, i.e. a mandatory, adequate line search and the susceptibility to round-off errors (Fletcher & Powell 1963; Morf & Kailath 1975). Expressing the SRVM method in a vector version as in Algorithm (2) makes it attractive for large-scale applications. We notice though that the SRVM algorithm becomes more complicated and slightly more expensive than the L-BFGS one in its numerical implementation. As a member of the quasi-Newton method family, the SRVM method is also expected to demonstrate a fast convergence rate in elastic FWI. A difference in convergence rate between the two algorithms would mainly stem from a difference in the approximation to the (inverse) Hessian used in the search direction update and in the estimation of the step length for the model update.

For the inversions, we merge the gradients concerning \( P \)- and \( S \)-wave velocities into a single vector to enter the workflow stated by Algorithm (2). Note that here we do not output the approximate inverse Hessian \( \mathbf{SS}^T \), but the search direction by \( \mathbf{p} = -\mathbf{SS}^T \mathbf{g} \) directly. The search direction \( \mathbf{p} \) must furthermore satisfy the following relationship with the gradient \( \mathbf{g} \)

\[
\mathbf{p}^T \mathbf{g} < 0
\] (16)

to ensure a decrease of the misfit function. This relationship in eq. (16) holds as long as the matrix \( \mathbf{SS}^T \) stays positive definite. This criterion is easy to implement without significant additional costs as it requires only one additional vector product.

### 3.3 Inexact line search and Wolfe conditions

Once a new search direction \( \mathbf{p}_k \) is obtained in the iteration, we need to determine the step length \( \mu_k \) which satisfies

\[
\mu_k = \arg \min_{\mu \geq 0} f(m_k + \mu_k \mathbf{p}_k)
\] (17)

This determination of the step length involves additional several forward simulations to assess the behaviour of the misfit function to \( \mu \). For L-BFGS, any line search method will be valid. Among this, the inexact line search provides an efficient way to compute an acceptable step...
length that reduces the misfit function sufficiently rather than exactly (Nocedal & Wright 2006). As previously mentioned, SRVM overcomes one shortcoming of the DFP algorithm where only the exact line search works. We will therefore apply the safeguarded backtracking line search method (Nocedal & Wright 2006; Modrak & Tromp 2016) to the SRVM-based elastic FWI. By letting \( \phi(\mu) = f(m + \mu p) \), we accept the step length once it meets the following inequalities according to the Wolfe conditions:

\[
\phi(\mu) \leq \phi(0) + c_1 \mu \phi'(0)
\]

\[
\phi'(\mu) \geq c_2 \phi'(0)
\]

where the constants \( c_1 \) and \( c_2 \) must be \( 0 < c_1 < c_2 < 1 \). The first inequality is known as the Armijo condition or the sufficient decrease condition and the second is known as the curvature condition. Together, they are referred to as the Wolfe conditions. The suggested values for \( c_1 \) and \( c_2 \) are 0.0001 and 0.9, respectively (Nocedal & Wright 2006). With the Wolfe conditions, a typical backtracking line search in average takes 1.22 trials for each FWI iteration. We set some upper limit for the number of trials, e.g. 10. When the backtracking trials exceed this number, we consider to restart the algorithm or stop the FWI. In our numerical experiments, we notice that once the quasi-Newton-based search direction \( p \) loses the property of descent after numerous trials, it is time to stop the FWI.

4 NUMERICAL EXAMPLES

To investigate the performance of the SRVM algorithm, we verify the SRVM-based elastic FWI using two synthetic models, i.e. a two-Sphere and Checkerboard model, as well as two more realistic synthetic models, i.e. the Marmousi and Overthrust model. The former two examples mimic the horizontal or vertical cross-section in global or regional geophysics, the latter two correspond to the widely used benchmarks in exploration geophysics. All the models have 500 evenly placed receivers at the top. There are 32, 32, 32 and 50 sources for the two-Sphere, Checkerboard, Marmousi and Overthrust models, respectively. The sources and receivers are at a 10-m depth. The numerical simulations are done by SPECFEM2D (Komatitsch & Vilotte 1998) and workflows are running with Seisflows (Modrak & Tromp 2016). We will take the L-BFGS algorithm as a performance benchmark because it is one of the most competitive algorithms in solving FWI (Modrak & Tromp 2016). For additional assessments beside the regular data case where the model is included in a homogeneous full-space without boundary reflections, we also consider a free-surface scenario with surface waves, as well as a practical scenario with variable \( V_P/V_S \) ratios and different \( V_P \) and \( V_S \) structures using a flat-spectrum source function.

4.1 Whole-space model benchmark

We will test our SRVM method in the time domain for 2D elastic FWIs on the two-Sphere, Checkerboard, Marmousi and Overthrust models. The goal is to compare the results with those obtained by the L-BFGS algorithm, regarding the quality of the inverted model and the convergence behaviour. We only consider the simultaneous elastic inversion on \( P \)- and \( S \)-wave velocities, with the density model kept constant. Since these tests are performed in a synthetic benchmark, we will be able to assess the final model results against the true models, which were used to calculate the observation data. The true elastic models are shown in Fig. 1, the (smoothed) initial models in Fig. 2, the inverted \( V_P \) models in Fig. 3 and the inverted \( V_S \) models in Fig. 4. The comparisons between corresponding data and model misfits by SRVM and L-BFGS are shown in Figs 5–7, respectively.

We start both the two-Sphere and Checkerboard models from the same initial model shown in the top two rows of Fig. 2. Their velocities increase linearly with depth \( z \) following the trends \( V_P(z) = 1700 + \frac{1}{5} z \) and \( V_S(z) = 1250 + \frac{1}{5} z \). The true two-Sphere model has two prominent \( \pm 10 \) per cent spherical perturbations to investigate the different effects of fast and slow anomalies on the inversion performance. The true Checkerboard model consists of \( \pm 5 \) per cent perturbations along a rectangular checkerboard pattern with respect to their background velocities. This regular pattern will allow to intuitively apprehend the resolution and smearing of the inversion results. As more realistic Earth models, we chose the Marmousi and Overthrust model, taken from their original definitions but with a constant ratio \( V_P/V_S = 1.5 \), close to a Poisson solid. We blur the starting models of Marmousi and Overthrust by convolving the true models with a Gaussian function wide enough to erase any imprint of the layered discontinuities.

The source functions are Ricker wavelets with 2-Hz, 4-Hz, 4-Hz and 4-Hz dominant frequencies for the two-Sphere, Checkerboard, Marmousi and Overthrust models, respectively. In our benchmarks, we employ relatively high frequencies (such as 4-Hz) in the inversion without a multiscale approach in Bunks et al. (1995) to directly investigate the inversion algorithms rather than a mixture between inversion and workflow strategy. We extend the models with convolutional perfect matched layers (Komatitsch & Martin 2007). The same source wavelet is used for modelling the observed data as well as for performing elastic FWI. Also, no (white) noise was added to the observed data. This avoids inversion complications due to unknown source descriptions and puts the benchmarks in an ideal situation, thus intentionally committing the inverse crime. In real-world inversions, the source wavelet would be mostly unknown, but could be estimated by solving a linearized source inversion problem (Pratt 1999; Virieux & Operto 2009). Both vertical and horizontal components of particle displacement are recorded and used in the inversion algorithms.

To allow straightforward comparisons, SRVM- and L-BFGS-based FWIs share identical setups. For L-BFGS, one needs to store the history of gradients and updates for several past iterations (Nocedal & Wright 2006). Modrak & Tromp (2016) observe that L-BFGS works...
well in FWI with a history of the past 3–7 gradients and model updates. A change of history size among this range of values only causes minor differences in the final results. We here set the history size of L-BFGS to 5. For SRVM, we keep all history, i.e. vectors \( w_k \) and scalars \( v_k/P_k \) of eq. (10) over all iterations. A limited-memory SRVM version could in principle be developed, but was left for future research. The full history enables us to access the complete inverse Hessian after the SRVM-based FWI converges (Tarantola 2005). The search directions and iterative model updates are determined together with the step lengths calculated by the backtracking line search, as explained in the previous section. We let all FWI procedures iteratively converge until the stop criterion is met.

The inversion results for \( V_P \) and \( V_S \) models using both the SRVM and L-BFGS algorithm are shown in Figs 3 and 4, respectively. We evaluate their convergence performances according to data and model misfits in Figs 5–7. For the cross-comparison between the \( V_P \) and \( V_S \) model misfits, we observe that the \( V_S \) convergence curve decreases faster than the \( V_P \) one. A reason might be that the radiation pattern of \( V_P \) scatterers is homogeneous, while the radiated energies of \( V_S \) scatterers mainly distribute at large scattering angles. Thus, the low-wavenumber components contained in the \( V_S \) gradient will benefit more from the waveform inversion (Wu & Aki 1985; Virieux & Operto 2009). For the comparison between (data and model) misfit curves by SRVM and L-BFGS, we notice two significant behaviours: (i) the misfit curves by SRVM decrease in an undulating way while those by L-BFGS decrease more smoothly; and (ii) the misfit curves by SRVM exhibit a stronger decrease at the beginning, with a slightly higher end value only in the two-Sphere model. For the more realistic Marmousi and Overthrust models, SRVM misfits achieve nearly identical final values as L-BFGS. However, for these more realistic results SRVM needs fewer iterations. The mathematical details behind these differences would require further study but go beyond the scope of this paper. Nevertheless, we find that with the similar performances between SRVM and L-BFGS algorithms in FWI, the SRVM has the advantage that it can further be used to conduct a resolution analysis and assessment of the uncertainty estimation of the resulting velocity models. All information to reconstruct the inverse Hessian is included in the stored vectors \( w_k \) and scalars \( v_k/P_k \) of eq. (10) per iteration. The vector \( w_k \) has the same size as the model parameter vector \( m \) and thus becomes affordable to store even for large-scale inversions. The resolution analysis and uncertainty quantification will be focus in Part 2 of our accompanying research paper.

### 4.2 Free-surface model benchmark

In this benchmark, we want to investigate the effect of surface waves upon inversion results. Surface waves are usually observed with high-amplitude, dispersive characteristics in land surveys. These waves decay exponentially with depth and most of their strong energies distribute within the depth of half a wavelength away from the surface (Rayleigh 1885; Miller & Pursey 1955), increasing the non-linearity of FWI (Yuan et al. 2015). Although surface waves have been used in some studies such as near-surface geophysics and regional or global seismology (Socco et al. 2010), they are usually treated as noise and removed in deep-structure imaging or inversion. Here, we test the inversions with free-surface data as it makes FWI more challenging and therefore further validates the SRVM approach in FWI.

The FWI model results by SRVM and L-BFGS algorithms are shown in Figs 8 and 9, respectively, using elastic FWIs with free-surface data. Compared to previous results with full-space models, we find a strong impact on the inverted models on both parameters. The \( V_S \) resulting models are slightly better inverted than the \( V_P \) one, exhibiting a deeper depth range. This difference between model parameters might be explained by that the seismograms dominated by surface waves are more sensitive to \( V_S \) information (Aki & Richards 2002). Comparing the inverted results between the SRVM and L-BFGS schemes, we observe that the SRVM-based FWI performs better in the Checkerboard test, but worse in the Marmousi test. Data and model misfits in Figs 10–12 show that although SRVM has a faster convergence rate, the algorithm stops too early for the free-surface Marmousi test.

More details are presented in Figs S5 and S6 in the Supporting Information, where a single source–receiver trace is plotted to observe the fit between true and inverted results. We find that even with a free-surface boundary, the initial synthetics of the Checkerboard data do not suffer much from non-linearity problems while this becomes more pronounced in the initial traces of the Marmousi case. The non-linearity mainly arises from the short wavelengths of surface waves. The SRVM algorithm which keeps all iteration history becomes tainted by a strong initial model offset, while the L-BFGS algorithm which only keeps a limited number of iteration gradients can better overcome such an initial model offset. However, we notice that the data and model misfits converge much faster using the SRVM scheme. To explore this behaviour further, we employ a combined scheme where we continue the inversion with the L-BFGS scheme, starting from the final inverted model by SRVM. We show the corresponding data and model misfits in Fig. 13. This second stage L-BFGS inversion becomes faster by 20 iterations than the conventional L-BFGS for the Marmousi case tested and leads to smaller final values for both data and model misfits. This suggests that also a combination of SRVM and L-BFGS schemes could help to improve the efficiency in elastic FWI, which will be subject of future work.

### 4.3 Practical whole-space Marmousi-X benchmark

We will test our SRVM algorithm, with the L-BFGS one as reference, in a practical elastic FWI scenario. We use a whole-space Marmousi-X model and a flat-spectrum Ormsby source function. The Marmousi-X model in Fig. 14 has a variable \( V_P/V_S \) ratio and a strip intrusion of 1800 m s\(^{-1}\) into \( V_S \). The \( V_P/V_S \) ratio decreases linearly from 2 to 1.5 with depth due to the increased degree of compaction in sediments (Bromirski et al. 1992). The initial model in Fig. 14 is obtained by smoothing the true model. The amplitude spectrum of the Ormsby source function is [0.5, 2.5, 10, 15]Hz. There are 32 shots and 500 receivers evenly placed at a 10 m depth. A more realistic but challenging scenario...
is to invert for inhomogeneous density together with $V_p$ and $V_S$. To avoid complications caused by density inversion (Virieux & Operto 2009), we only consider the inversion for $V_p$ and $V_S$.

A point to consider in this scenario is the sensitivity of SRVM- and L-BFGS-based elastic FWIs to variable $V_p/V_S$ ratios together with different parameter structures. The $V_p/V_S$ ratio is relatively higher at the shallow part, decreasing with depth. One difficulty of multiparameter inversion is the trade-off between parameters. When two parameters share very similar structures, the inversion becomes easier than otherwise. Fig. 15 shows the inverted $V_p$ and $V_S$ models, with Fig. 16 displaying the corresponding data and model misfits. More details can be found in Fig. S8 in the Supporting Information, in which a single source–receiver trace is plotted to observe the fit between true and inverted results. For both SRVM and L-BFGS schemes, the Marmousi-X model can be inverted well. The strip intrusions are clearly present in the two inverted $V_S$ results, without obvious leakage into inverted $V_p$ results. We find from the data and model misfit comparisons that SRVM and L-BFGS algorithms have similar performances within this scenario.

5 CONCLUSIONS
FWI has become a powerful technique to invert for Earth’s structure, given the sensitivity of seismic signals to geophysical properties in the subsurface. Nevertheless, inversion schemes are challenged by the non-uniqueness of the inverse problem and the multiscale feature in realistic applications. This paper deals with the first part of our research about a non-linear, gradient optimization scheme using the SRVM algorithm in elastic FWI. The SRVM algorithm overcomes the shortcomings of the conventional DFP algorithm, enables an inexact line search and exhibits robustness to round-off errors. We formulate the SRVM algorithm in a vector version to make its implicit memory requirements affordable for realistic large-scale applications. We furthermore incorporate a search step using the backtracking line search under Wolfe conditions in elastic FWI. This new inversion scheme was investigated with numerical examples under regular, free-surface and pseudo-practical scenarios. As a performance benchmark, we compare SRVM-based elastic FWI results to those obtained by an L-BFGS-based elastic FWI. The comparisons demonstrate that the SRVM and L-BFGS algorithms have similar performance, with SRVM results even slightly better for several test cases.

A main advantage of using the SRVM algorithm is the possibility of reconstructing the associated inverse Hessian, which comes as a side product of the history-keeping gradient scheme. This will be further investigated in the accompanying Part 2 of this research for resolution analysis and uncertainty estimation in elastic FWI.

ACKNOWLEDGEMENTS
The authors are grateful to editor Jean Virieux and two anonymous reviewers for improving the initial manuscript. This work was supported by the King Abdullah University of Science & Technology (KAUST) Office of Sponsored Research (OSR) under award No. UAPN#2605-CRG4. Computational resources were provided by the Information Technology Division and Extreme Computing Research Center (ECRC) at KAUST.

REFERENCES
Aki, K. & Richards, P.G., 2002. Quantitative Seismology. University Science Books.
An, M., 2012. A simple method for determining the spatial resolution of a general inverse problem, Geophys. J. Int., 191(2), 849–864.
Bizoń, E., Peter, D., Lefèvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N. & Pugmire, D., 2016. Global adjoint tomography: first-generation model, Geophys. J. Int., 207, 1739–1766.
Bromirski, P., Frazer, L. & Duennebier, F., 1992. Sediment shear Q from airgun OBS data, Geophys. J. Int., 110(3), 465–485.
Bui-Thanh, T., Ghattas, O. & Stadler, G., 2013. A computational framework for infinite-dimensional bayesian inverse problems. Part I: the linearized case, with application to global seismic inversion, SIAM J. Sci. Comput., 35(6), A2494–A2523.
Bunks, C., Saleck, F.M., Zaleski, S. & Chavent, G., 1995. Multiscale seismic waveform inversion, Geophysics, 60(5), 1457–1473.
Carlson, N., 1973. Fast triangular formulation of the square root filter, AIAA journal, 11(9), 1259–1265.
Davidson, W.C., 1959. Variable metric method for minimization, A.E.C. Research and Development Report, ANL-5990.
Fichtner, A., Kennett, B.L.N., Igel, H. & Bunge, H.P., 2009. Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods, Geophys. J. Int., 179(3), 1703–1725.
Fichtner, A. & Leeuw, T.v., 2015. Resolution analysis by random probing, J. geophys. Res., 120(8), 5549–5573.
Fichtner, A. & Trampert, J., 2011a. Hessian kernels of seismic data functions based upon adjoint techniques, Geophys. J. Int., 185(2), 775–798.
Fichtner, A. & Trampert, J., 2011b. Resolution analysis in full waveform inversion, Geophys. J. Int., 187(3), 1604–1624.
Fichtner, A., Trampert, J., Cupillard, P., Saygin, E., Taymaz, T., Capdeville, Y. & Villase nor, A., 2013. Multiscale full waveform inversion, Geophys. J. Int., 194(1), 534–556.
Fletcher, R. & Powell, M.J., 1963. A rapidly convergent descent method for minimization, Comput. J., 6(2), 163–168.
Fouga, E. & Lambaré, G., 1997. Parameterization study for acoustic and elastic ray plus born inversion, J. Seism. Explor., 6(2-3), 253–277.
Hull, D.G. & Tapley, B.D., 1977. Square-root variable-metric methods for minimization, J. Optim. Theor. Appl., 21(3), 251–259.
Komatitsch, D. & Martin, R., 2007. An unsplit convolutional Perfectly Matched Layer improved at grazing incidence for the seismic wave equation, Geophysics, 72(5), SM155–SM167.
Komatitsch, D. & Vilotte, J.P., 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.
Lailly, P., 1983. The seismic inverse problem as a sequence of before-stack migrations, in Proceedings of the Conference on Inverse Scattering, Theory and Application Expanded Abstracts, pp. 206–220, Society of Industrial and Applied Mathematics, Philadelphia, PA, USA.
Liu, D.C. & Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization, Math. Program., 45(1), 503–528.
Liu, Q. & Peter, D., 2019. Square-root variable metric based elastic full-waveform inversion? Part 2: uncertainty estimation, Geophys. J. Int., 218(2), 1100–1120.
Luo, Y., 2012. Seismic imaging and inversion based on spectral-element and adjoint methods, PhD thesis, Department of Geosciences.
Luo, Y., Zhu, H., Nissen-Meyer, T., Morency, C. & Tromp, J., 2009. Seismic modeling and imaging based upon spectral-element and adjoint methods, *Leading Edge*, 28(5), 568–574.

Martin, J., Wilcox, L.C., Burstedde, C. & Ghatts, O., 2012. A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion, *SIAM J. Sci. Comput.*, 34(3), A1460–A1487.

Miller, G. & Pursey, H., 1955. On the partition of energy between elastic waves in a semi-infinite solid, *Proc. R. Soc. Lond.: Math. Phys. Eng. Sci.*, 233, 55–69.

Modrak, R. & Tromp, J., 2016. Seismic waveform inversion best practices: regional, global and exploration test cases, *Geophys. J. Int.*, 206(3), 1864–1889.

Morf, M. & Kailath, T., 1975. Square-root algorithms for least-squares estimation, *IEEE Trans. Autom. Control*, 20(4), 487–497.

Nocedal, J., 2002. Updating quasi-newton matrices with limited storage, *Math. Comput.*, 35(151), 773–782.

Nocedal, J., 1992. Theory of algorithms for unconstrained optimization, *Acta Numer.*, 1, 199–242.

Nocedal, J. & Wright, S., 2006. *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer-Verlag, 2nd edn.

Operto, S., Minissi, A., Brossier, R., Combe, L., MÉtivier, L., Monteller, V., Ribodetti, A. & Virieux, J., 2015. Efficient 3-d frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to valhall in the visco-acoustic vertical transverse isotropic approximation, *Geophys. J. Int.*, 202(2), 1362–1391.

Plessix, R.E., 2009. Three-dimensional frequency-domain full-waveform inversion with an iterative solver, *Geophysics*, 74(6), WC53–WCC61.

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.

Pratt, R.G. & Worthington, M., 1990. Inverse theory applied to multi-source cross-hole tomography, Part 1: acoustic wave-equation method, *Geophys. Prospect.*, 38(3), 287–310.

Rayleigh, L., 1885. On waves propagated along the plane surface of an elastic solid, *Proc. Lond. Math. Soc.*, 1(1), 4–11.

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.*, 40(3), 1–29.

SIRgue, L., Barkved, O.I., Delling, J., Egen, J., Albertin, U. & Kommedal, J.H., 2010. Thematic set: Full waveform inversion: the next leap forward in imaging at valhall, *First Break*, 28(4), 65–70.

Socco, L.V., Foti, S. & Boiiero, D., 2010. Surface-wave analysis for building near-surface velocity models—established approaches and new perspectives, *Geophysics*, 75(5), A83–A102.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2009. Adjoint tomography of the southern California crust, *Science*, 325, 988–992.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2010. Seismic tomography of the southern California crust based on spectral-element and adjoint methods, *Geophys. J. Int.*, 180, 433–462.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysica*, 49, 1259–1266.

Tarantola, A., 1986. A strategy for nonlinear elastic inversion of seismic reflection data, *Geophysics*, 51(10), 1893–1903.

Tarantola, A., 1987. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier Science Publ. Co., Inc.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics.

Trampert, J., Fichtner, A. & Ritsema, J., 2012. Resolution tests revisited: the power of random numbers, *Geophys. J. Int.*, 192(2), 676–680.

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

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

Warner, M. et al., 2013. Anisotropic 3d full-waveform inversion, *Geophysics*, 78(2), R59–R80.

Williamson, W.E., 1975. Square-root algorithms for function minimization, *J. Comput. Appl. Math.*, 13(1), 107–109.

Wu, R.-S. & Aki, K., 1985. Scattering characteristics of elastic waves by an elastic heterogeneity, *Geophysics*, 50(4), 582–595.

Yuan, Y.O., Simons, F.J. & Bozdağ, E., 2015. Multiscale adjoint waveform tomography for surface and body waves, *Geophysics*, 80(5), R281–R302.

Zhu, H., Bozdağ, E., Peter, D. & Tromp, J., 2012. Structure of the European upper mantle revealed by adjoint tomography, *Nat. Geosci.*, 5(7), 493–498.

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

**Supporting Information**

Supplementary data are available at *GJI* online.

We provide detailed trace comparisons about SRVM- and L-BFGS-based FWIs. The traces are selected from a relatively far offset (as shown in the schematic diagrams in Fig. S1), where the initial synthetics are more likely to get cycle-skipped. Assuming $X$ is the horizontal dimension, we locate the source at $1/10$th of $X$ and the receiver at $9/10$th of $X$. Each figure corresponds to a specific test model, as indicated by their captions.

**Figure S1.** Schematic sketch of the configuration for the following single source–receiver traces. The offset is relatively large, so the data are more likely to suffer from cycle-skipping. With the traces from the initial and true models for reference, we compare and evaluate the performances of L-BFGS- and SRVM-based FWIs.

**Figure S2.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for two-Sphere model.

**Figure S3.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for Checkerboard model.

**Figure S4.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for Marmousi model.

**Figure S5.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for Overthrust model.

**Figure S6.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for free-surface Checkerboard model.

**Figure S7.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for free-surface Marmousi model.

**Figure S8.** Example of single source–receiver trace: synthetics by L-BFGS- and SRVM-based FWIs for Marmousi-X model.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.