



Effects of Conjugate Gradient Methods and Step-Length Formulas on the Multiscale Full Waveform Inversion in Time Domain: Numerical Experiments

YOSHAN LIU,¹  JIWEN TENG,¹ TAO XU,^{1,2} JOSÉ BADAL,³ QINYA LIU,⁴ and BING ZHOU⁵

Abstract—We carry out full waveform inversion (FWI) in time domain based on an alternative frequency-band selection strategy that allows us to implement the method with success. This strategy aims at decomposing the seismic data within partially overlapped frequency intervals by carrying out a concatenated treatment of the wavelet to largely avoid redundant frequency information to adapt to wavelength or wavenumber coverage. A pertinent numerical test proves the effectiveness of this strategy. Based on this strategy, we comparatively analyze the effects of update parameters for the nonlinear conjugate gradient (CG) method and step-length formulas on the multiscale FWI through several numerical tests. The investigations of up to eight versions of the nonlinear CG method with and without Gaussian white noise make clear that the *HS* (Hestenes and Stiefel in *J Res Natl Bur Stand Sect 5:409–436, 1952*), *CD* (Fletcher in *Practical methods of optimization vol. 1: unconstrained optimization, Wiley, New York, 1987*), and *PRP* (Polak and Ribière in *Revue Française Informat Recherche Operationelle, 3e Année 16:35–43, 1969*; Polyak in *USSR Comput Math Math Phys 9:94–112, 1969*) versions are more efficient among the eight versions, while the *DY* (Dai and Yuan in *SIAM J Optim 10:177–182, 1999*) version always yields inaccurate result, because it overestimates the deeper parts of the model. The application of FWI algorithms using distinct step-length formulas, such as the direct method (*Direct*), the parabolic search method (*Search*), and the two-point quadratic interpolation method (*Interp*), proves that the *Interp* is more efficient for noise-free data, while the *Direct* is more efficient for Gaussian white noise data. In contrast, the *Search* is less efficient because of its slow convergence. In general, the three step-length formulas are robust or partly insensitive to Gaussian white noise and the complexity of the model. When the initial velocity model deviates far from the real model or the data

are contaminated by noise, the objective function values of the *Direct* and *Interp* are oscillating at the beginning of the inversion, whereas that of the *Search* decreases consistently.

Key words: Full waveform inversion, nonlinear conjugate gradient method, step-length formulas, multiscale strategy, frequency-band selection strategy.

1. Introduction

Full waveform inversion (FWI) makes full use of both the amplitude and phase of the wave altogether to provide accurate seismic velocity models that can be used later to characterize earth structures or reservoirs containing natural resources (Tarantola 1984; Pratt 1990; Ravaut et al. 2004; Tromp et al. 2005; Liu and Tromp 2006; Liu and Gu 2012). It bridges the gap between the conventional seismic velocity analysis (in which phase information is mostly used) and the amplitude versus offset analysis (in which amplitude information is used). FWI is a method that aims at finding the best seismic velocity model to interpret the available data. In such process, synthetic data are calculated using an assumed model and compared against the observed data. If the fit is not acceptable, the model is perturbed, so that the synthetic data are regenerated and the procedure is repeated until to approach the convergence. It is a powerful working tool in seeking images and properties (such as velocity and impedance) of complex geological structures (Tarantola 1984; Plessix and Li 2013; Zhang et al. 2014; Zhou et al. 2015). This approach can be implemented either in time domain (Tarantola 1984, 1986, 1988; Mora 1987; Vigh and Starr 2008; Liu and Tromp 2008) or in frequency domain (Pratt and Worthington 1990; Zhou and

¹ State Key Laboratory of Lithospheric Evolution, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China. E-mail: ysliu@mail.iggcas.ac.cn; jwtenge@mail.iggcas.ac.cn; xutao@mail.iggcas.ac.cn

² CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing 100101, China.

³ Physics of the Earth, Sciences B, University of Zaragoza, Pedro Cerbuna 12, 50009 Saragossa, Spain. E-mail: badal@unizar.es

⁴ Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada. E-mail: liuqy@physics.utoronto.ca

⁵ Petroleum Geosciences, The Petroleum Institute, Abu Dhabi, United Arab Emirates. E-mail: bizhou@pi.ac.ae

Greenhalgh 2011). In the frequency domain, a direct solver based on an LU (lower and upper matrices) factorization (Davis and Duff 1997) instead of an iterative solver is efficient for FWI, because it allows obtain all solutions for all excitation sources simultaneously by factoring the impedance matrix only once (Pratt and Worthington 1990). Usually, only several frequency points or groups are needed to obtain a high-resolution reconstructed model. However, the direct solver faces a challenge in 3D because of its extremely large computer memory demand. Unlike frequency-domain FWI, time-domain FWI usually extrapolates the seismic wavefields explicitly, which avoids solving a large-scale algebraic equation. Since time-domain FWI needs to extrapolate the wavefield at all time sampling points and for all shots, and from this viewpoint, it is computationally inefficient. However, it can be easily extended to three dimensions, because it explicitly extrapolates the wavefields along the time axis (Liu and Tromp 2008).

FWI is a highly nonlinear inversion problem, so that its objective function is multimodal (Symes 2008). Although the global optimization methods, such as the stochastic Monte Carlo-based methods (Rothman 1985; Kvoren et al. 1991; Mosegaard and Tarantola 1995) tend to search a global minimum solution of FWI, the excessive cost required for convergence makes these methods go beyond the ordinary computation capabilities. Some feasible and alternative ways are the iterative local optimization methods, such as the steepest descent method, the nonlinear conjugate gradient method (Pratt and Worthington 1990; Song et al. 1995; Liao and McMechan 1996; Pratt 1999; Shipp and Singh 2002; Ravaut et al. 2004; Sirgue and Pratt 2004; Malinowski and Operto 2008; Mulder and Plessix 2008), the Gauss–Newton method (Pratt et al. 1998; Hu et al. 2009; Pan et al. 2016, 2017), the quasi-Newton method (Brossier et al. 2009), or the truncated Newton method (Métivier et al. 2014; Pan et al. 2016, 2017).

Although the iterative inversion methods are computationally feasible, the large number of local minima at all scales impedes FWI to converge to the vicinity of the global minimum. In particular, these iterative inversion methods may fail to invert seismic

data obtained from structurally complex models (e.g., the Marmousi model) due to the presence of numerous local minima of the objective function, unless the initial velocity model is already in the neighborhood of the global solution. To overcome such issue, Bunks et al. (1995) proposed a multiscale approach that can improve the performance of iterative local optimization methods, which decomposes the highly nonlinear optimization problem into several scales. Once the inversion problem has been decomposed by scales, the longer scale components are first inverted with the idea to greatly reduce the number of local minima (Fichtner et al. 2013; ten Kroode et al. 2013), and to get a good guess to the inversion for shorter scale components (higher frequency band). In frequency domain, the wavefields at different frequency slices can be naturally decomposed by scales, so that a careful selection of inversion frequency slices yields computationally efficient inversion schemes (Sirgue and Pratt 2004). Therefore, the frequency selection for the scale decomposition becomes the core part of the multiscale strategy. In contrast, time-domain FWI uses multiple frequencies simultaneously during the inversion process, which allows update a much wider range of wavenumber than using a single frequency at one time. As pointed out by Sirgue and Pratt (2004), the frequency bandwidth can adjust the range of wavelength or wavenumber. Naturally, we can design a frequency-band selection strategy that guarantees the ranges of wavelength corresponding to adjacent frequency bands to be less redundant which leads to an efficient inversion at each frequency band. In addition, the windowed input data also alternatively reduce the effects of local minima by focusing the inversion on different parts of the data (Sheng et al. 2006; Brenders et al. 2009).

Although FWI exhibits a great potential, many factors contribute to the inverted results. Pageot et al. (2013) presented a two-dimensional parametric analysis of frequency-domain FWI of teleseismic data for lithospheric imaging, to identify the main factors that impact on the quality of the inverted P- and S-wave velocity models. However, a similar study for the time-domain FWI has not been systematically conducted to date. In this paper, we investigate the factors that clearly impinge on FWI in time domain. In particular, these factors mainly

include update parameters for the nonlinear conjugate gradient method and step-length formulas, besides already known statements, such as the initial velocity model, local optimization methods, multiscale strategy, and data stacking. Below, we briefly restate time-domain FWI and then we propose an alternative frequency-band selection strategy with the help of a Wiener low-pass filter (Boonyasiriwat et al. 2009). Next, we consider three effective step-length formulas to optimize the nonlinear inversion process. After that, we perform a series of numerical tests to study the effects of the cited influencing factors on FWI. Finally, we draw interesting conclusions from all this computational work. Of course, the present work does not deplete the subject. Certainly, some other factors also affect the performance of FWI, such as the absence of low-frequency data (ten Kroode et al. 2013) and types of objective functions, etc. However, these other issues go beyond the scope of this paper, although deserve further attention in the future.

2. Theory

2.1. Full Waveform Inversion in Time Domain

For the sake of completeness, in this section, we briefly summarize the theory of full waveform inversion in time domain. FWI is traditionally expressed as the minimization of the sample-by-sample differences between the observed and simulated seismic data, so that a starting velocity model is updated iteratively. Usually, FWI in time domain tries to minimize the following L2 (least-squares) norm objective function:

$$E(\mathbf{c}) = \frac{1}{2} \sum_s \sum_r \int [\delta p(\mathbf{x}_r, t|\mathbf{x}_s)]^2 dt, \quad (1)$$

$$\delta p(\mathbf{x}_r, t|\mathbf{x}_s) = p^{\text{cal}}(\mathbf{x}_r, t|\mathbf{x}_s) - p^{\text{obs}}(\mathbf{x}_r, t|\mathbf{x}_s), \quad (2)$$

where $\delta p(\mathbf{x}_r, t|\mathbf{x}_s)$ is the data residuals; the subscripts s and r denote summations over sources and receivers, respectively; $p^{\text{cal}}(\mathbf{x}_r, t|\mathbf{x}_s)$ and $p^{\text{obs}}(\mathbf{x}_r, t|\mathbf{x}_s)$ are the observed and simulated data at the receiver position \mathbf{x}_r and the time instant t , respectively, which are excited by a source located at the position \mathbf{x}_s . The wavefields are computed through forward modeling

by solving the following constant density acoustic-wave equation:

$$\frac{\partial^2}{\partial t^2} p(\mathbf{x}, t|\mathbf{x}_s) = c^2 \nabla^2 p(\mathbf{x}, t|\mathbf{x}_s) + f(\mathbf{x}_s, t), \quad (3)$$

where $p(\mathbf{x}, t|\mathbf{x}_s)$ is the pressure field at the spatial location \mathbf{x} arising from a disturbance at the source location \mathbf{x}_s ; $c(\mathbf{x})$ is the velocity of medium at the location \mathbf{x} ; ∇^2 is the Laplace operator; and $f(\mathbf{x}_s, t)$ is the seismic source function. In all experiments, we adopt a central finite-difference stencil of the 16th-order accuracy in space and the second-order accuracy in time to extrapolate the source and receiver wavefields. We solve the second-order constant density acoustic-wave equation considering perfectly matched layers as absorbing boundary conditions (Liu et al. 2012) to suppress spurious reflections coming from artificial boundaries.

Although some estimation methods based on a random sampling of the model-space can be theoretically applied, it is a long and costly process, because it requires a large number of evaluations of the misfit function for each new model (Rothman 1985; Kvoren et al. 1991; Guitton et al. 2012). Currently, the local optimization methods are usually preferred due to its computational efficiency, although inherently are limited to local convergence and cannot guarantee a global solution (Pratt et al. 1998). For the local optimization methods, the computation of the gradient with respect to velocity model becomes the core part of FWI.

Usually, an efficient computation of the gradient is based on the so-called adjoint-state method (Plessix 2006). In the framework of the adjoint-state method, the gradient of the objective function (1) with respect to velocity model $c(\mathbf{x})$ is calculated by the zero-lag cross-correlation between forward-propagated wavefield and backward-projected wavefield residuals (Tarantola 1984; Boonyasiriwat et al. 2009):

$$\begin{aligned} g(\mathbf{x}) &= \frac{2}{c(\mathbf{x})} \sum_s \int \frac{\partial^2}{\partial t^2} p(\mathbf{x}, t|\mathbf{x}_s) q(\mathbf{x}, t|\mathbf{x}_s) dt \\ &= -\frac{2}{c(\mathbf{x})} \sum_s \int \frac{\partial}{\partial t} p(\mathbf{x}, t|\mathbf{x}_s) \frac{\partial}{\partial t} q(\mathbf{x}, t|\mathbf{x}_s) dt, \end{aligned} \quad (4)$$

where $p(\mathbf{x}, t | \mathbf{x}_s)$ denotes the forward-propagated wavefield, while $q(\mathbf{x}, t | \mathbf{x}_s)$ denotes the backward-projected wavefield residuals.

2.2. Local Optimization Methods

Once the gradient (4) is available, the nonlinear optimization problem (1) can be solved by the steepest descent (SD) method:

$$\mathbf{c}_{k+1} = \mathbf{c}_k - \alpha_k \mathbf{g}_k, \quad (5)$$

or the nonlinear conjugate gradient (CG) method

$$\mathbf{c}_{k+1} = \mathbf{c}_k + \alpha_k \mathbf{d}_k, \quad (6)$$

$$\mathbf{d}_k = -\mathbf{g}_k + \beta_k \mathbf{d}_{k-1}, \quad (7)$$

where α_k is the step-length at k th iteration; \mathbf{g}_k is the gradient; and β_k is a scalar parameter that has many variants. The different ways in which this parameter β can be chosen lead to distinct versions of the nonlinear CG method. At present, various choices for the nonlinear CG update parameter are available (Hager and Zhang 2006). These update parameters include the *HS* (Hestenes and Stiefel 1952), *FR* (Fletcher and Reeves 1964), *PRP* (Polak and Ribière 1969; Polyak 1969), *CD* (Fletcher 1987), *LS* (Liu and Storey 1991), *DY* (Dai and Yuan 1999), *HZ* (Hager and Zhang 2005), and *HZI* (Liu et al. 2014) versions. These update parameters are listed as follows:

$$\beta_k^{HS} = \frac{\mathbf{g}_k^T \mathbf{y}_k}{\mathbf{d}_{k-1}^T \mathbf{y}_k}, \quad (8a)$$

$$\beta_k^{FR} = \frac{\|\mathbf{g}_k\|^2}{\|\mathbf{g}_{k-1}\|^2}, \quad (8b)$$

$$\beta_k^{PRP} = \frac{\mathbf{g}_k^T \mathbf{y}_k}{\|\mathbf{g}_{k-1}\|^2}, \quad (8c)$$

$$\beta_k^{CD} = -\frac{\|\mathbf{g}_k\|^2}{\mathbf{d}_{k-1}^T \mathbf{g}_{k-1}}, \quad (8d)$$

$$\beta_k^{LS} = -\frac{\mathbf{g}_k^T \mathbf{y}_k}{\mathbf{d}_{k-1}^T \mathbf{g}_{k-1}}, \quad (8e)$$

$$\beta_k^{DY} = \frac{\|\mathbf{g}_k\|^2}{\mathbf{d}_{k-1}^T \mathbf{y}_k}, \quad (8f)$$

$$\beta_k^{HZ} = \left(\mathbf{y}_k - 2\mathbf{d}_{k-1} \frac{\|\mathbf{y}_k\|^2}{\mathbf{d}_{k-1}^T \mathbf{y}_k} \right)^T \frac{\mathbf{g}_k}{\mathbf{d}_{k-1}^T \mathbf{y}_k}, \quad (8g)$$

$$\beta_k^{HZ1} = \left(\mathbf{y}_k - \mathbf{d}_{k-1} \frac{\|\mathbf{y}_k\|^2}{\mathbf{d}_{k-1}^T \mathbf{y}_k} \right)^T \frac{\mathbf{g}_k}{\mathbf{d}_{k-1}^T \mathbf{y}_k}, \quad (8h)$$

where $\mathbf{y}_k = \mathbf{g}_k - \mathbf{g}_{k-1}$ is the gradient change; $\|\cdot\|$ represents the L2 norm. In fact, we use the following modified scalar parameter:

$$\beta_k^+ = \max\{0, \beta\}, \quad (9)$$

which ensures the convergence of the nonlinear CG method (Hager and Zhang 2006). However, the remaining question is which of the many available choices for the CG update parameter is the most effective and efficient. In the next section, we will take care of this issue through several numerical examples.

The optimization problem (1) can be also solved by the Gauss–Newton method:

$$\mathbf{c}_{k+1} = \mathbf{c}_k + \alpha_k \mathbf{d}_k, \quad (10)$$

$$\mathbf{d}_k = -\mathbf{H}_k^{-1} \mathbf{g}_k, \quad (11)$$

where \mathbf{H}_k is the approximate Hessian matrix. The inverse Hessian operator acts as deconvolution operator that accounts for the limited bandwidth of the seismic data and corrects for the loss of amplitude of poorly illuminated subsurface parameters. Hence, the Newton-based methods possess better convergence properties (superlinear to quadratic convergence rate). In general, the explicit computation of the Hessian matrix is always expensive, which goes beyond the capability of modern computer hardware, especially for 3D models with millions and even billions of unknowns.

Nevertheless, a practical scheme is the Limited-memory Broyden–Fletcher–Goldfarb–Shanno method, i.e., the L-BFGS approach (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970) proposed by Nocedal (1980), which appears to be one of the most robust and efficient limited-memory quasi-Newton algorithms. This quasi-Newton approach iteratively computes an estimation of the product of the approximate inverse Hessian matrix \mathbf{B}_k and the gradient \mathbf{g}_k using the history of the solution and gradient vectors.

One of the main benefits of this technique is that the approximate inverse Hessian matrix is never explicitly formed, thus involving significant memory savings (Nocedal 1980; Guitton and Symes 2003; Brossier et al. 2009). The product of the approximate- or quasi-inverse Hessian matrix and the gradient can be calculated using a recursive formula with information from the last m iterations, where m is any number supplied by the user. The application details can be seen in Nocedal (1980). At each iteration, we use the following initial inverse Hessian matrix given by Nocedal (1980):

$$\mathbf{B}_k^0 = \mathbf{y}_k^T \mathbf{s}_k / \mathbf{y}_k^T \mathbf{y}_k \mathbf{I}, \quad (12)$$

where \mathbf{I} is the identity matrix with the same dimension as \mathbf{B}_k^0 ; $\mathbf{s}_k = \mathbf{c}_k - \mathbf{c}_{k-1}$ is the model change. In the following experiments, the number of the stored \mathbf{y} and \mathbf{s} for corrections used in the L-BFGS approach is set to 10. Usually, an effective descent direction generated by the L-BFGS method must be well behaved, which is ensured by the Wolfe linear search (Nocedal and Wright 1999; Wu et al. 2015). Here, we also consider this special case. In this case, the L-BFGS method fails to generate an effective descent direction to decrease the objective function value (i.e., a nonpositive definite Hessian matrix). We use the negative gradient as the descent direction when the sufficient descent condition $\mathbf{g}_k^T \mathbf{d}_k < 0$ is not satisfied (Hu and Wang 2014). Another alternative method is the truncated Newton method (Métivier et al. 2014; Pan et al. 2016, 2017). At each iteration, the model update is computed as an approximate solution of the Newton equations through a linear iterative solver (such as a conjugate gradient solver). This iterative solver only requires computing Hessian-vector products in a matrix-free fashion. It is also not necessary to form the Hessian operator explicitly. Although this method is a better approximation to the inverse Hessian than the L-BFGS method, it needs much more computation.

In summary, the descent direction of the above three local optimization methods (SD, CG, and L-BFGS) can be formulated by means of the following generalized expressions:

$$\mathbf{d}_k = \begin{cases} -\mathbf{g}_k, & \text{(SD)} \\ -\mathbf{g}_k + \beta_k \mathbf{d}_{k-1}, & \text{(CG)} \\ -\mathbf{B}_k \mathbf{g}_k, & \text{(L - BFGS)} \end{cases}, \quad (13)$$

where \mathbf{B}_k is the approximate inverse Hessian of the L-BFGS method.

2.3. Inversion Step-Length Formulas

As for local optimization methods, the descent direction must be scaled by a proper scalar, i.e., an appropriate step-length, to ensure the declining value of the objective function at each iteration. FWI is an inversion problem that requires intensive computation, so an efficient and cost-effective step-length formula is extremely important in this context. In this study, we consider three cost-saving step-length formulas.

2.3.1 Direct Method

Pica et al. (1990) derived an optimum step-length formula for time-domain FWI. The formula can be written as follows:

$$\alpha = - \frac{\sum_s \sum_r \int [p^{\text{cal}}(\mathbf{c}_k + \alpha_t \mathbf{d}_k) - p^{\text{cal}}(\mathbf{c}_k)] \cdot \delta p(\mathbf{x}_r, t | \mathbf{x}_s) dt}{\sum_s \sum_r \int [p^{\text{cal}}(\mathbf{c}_k + \alpha_t \mathbf{d}_k) - p^{\text{cal}}(\mathbf{c}_k)]^2 dt} \alpha_t, \quad (14)$$

where α_t is a test step-length; \mathbf{c}_k is the velocity model at k th iteration; and $\delta p(\mathbf{x}_r, t | \mathbf{x}_s)$ is the data residuals. The test step-length should be satisfy the following condition:

$$\max(|\alpha_t \mathbf{d}_k|) \leq \frac{1}{100} \max(\mathbf{c}_k). \quad (15)$$

To compute the optimum step-length, only an extra forward modeling is required.

2.3.2 Parabolic Search Method

At current state, i.e., \mathbf{c}_k at k th iteration, by applying the Taylor series expansion of α up to the second order, the objective function (1) can be approximated as follows:

$$\begin{aligned} E(\mathbf{c}_k + \alpha \mathbf{d}_k) &\approx E(\mathbf{c}_k) + \alpha [\nabla E(\mathbf{c}_k)]^T \mathbf{d}_k + \frac{\alpha^2}{2} \mathbf{d}_k^T \nabla^2 E(\mathbf{c}_k) \mathbf{d}_k \\ &= E(\mathbf{c}_k) + \alpha \left. \frac{\partial E(\mathbf{c}_k + \alpha \mathbf{d}_k)}{\partial \alpha} \right|_{\alpha=0} \\ &\quad + \left. \frac{\alpha^2 \partial^2 E(\mathbf{c}_k + \alpha \mathbf{d}_k)}{2 \partial \alpha^2} \right|_{\alpha=0} \\ &= c + b\alpha + a\alpha^2. \end{aligned} \quad (16)$$

This expression can be approximated by fitting a parabola and finding its minimum. Then, if the values of the objective function at three states are available and these values satisfy the following condition (Vigh et al. 2009)

$$\begin{cases} E(\mathbf{c}_k + \alpha_{t1}\mathbf{d}_k) < E(\mathbf{c}_k) \\ E(\mathbf{c}_k + \alpha_{t2}\mathbf{d}_k) > E(\mathbf{c}_k + \alpha_{t1}\mathbf{d}_k), \\ 0 < \alpha_{t1} < \alpha_{t2} \end{cases}, \quad (17)$$

which ensures that the desired step-length is bracketed in the interval $[0, \alpha_{t2}]$. The optimum step-length can be calculated under the condition of minimum ($\alpha_{\text{opt}} = -b/2a$) through the determination of the unknown parameters a and b . Once we find two step-lengths satisfying the condition (17), we have the following relationship:

$$\begin{cases} E(\mathbf{c}_k) = c \\ E(\mathbf{c}_k + \alpha_{t1}\mathbf{d}_k) = c + b\alpha_{t1} + a\alpha_{t1}^2 \\ E(\mathbf{c}_k + \alpha_{t2}\mathbf{d}_k) = c + b\alpha_{t2} + a\alpha_{t2}^2 \end{cases}. \quad (18)$$

After solving the equation system (18), the optimum step-length is given by the following:

$$\begin{aligned} \alpha_{\text{opt}} &= -\frac{b}{2a} \\ &= -\frac{\alpha_{t2}^2 E(\mathbf{c}_k + \alpha_{t1}\mathbf{d}_k) - \alpha_{t1}^2 E(\mathbf{c}_k + \alpha_{t2}\mathbf{d}_k) + E(\mathbf{c}_k) (\alpha_{t2}^2 - \alpha_{t1}^2)}{\alpha_{t1} E(\mathbf{c}_k + \alpha_{t2}\mathbf{d}_k) - \alpha_{t2} E(\mathbf{c}_k + \alpha_{t1}\mathbf{d}_k) + E(\mathbf{c}_k) (\alpha_{t2} - \alpha_{t1})}. \end{aligned} \quad (19)$$

To compute the optimum step-length (19), at least two times extra forward modeling are needed, because the satisfaction of the condition (17) may need more number of extra forward modeling instead of just two times. Figure 1 outlines the principle of the parabolic search method. Although the linear search method satisfying the Wolfe condition for determining the step-length (Nocedal and Wright 1999) ensures an effective descent direction generated by the L-BFGS method, the Wolfe linear search involves the extra computation of gradient with the model updated by a test step-length. It violates the cost-effective purpose, because FWI inversion is itself an expensive process. In this paper, we implement the standard parabolic search method or parabola fitting, as Brossier et al. (2009) and Vigh et al. (2009) made.

2.3.3 Two-Point Quadratic Interpolation Method

Unlike the parabolic search method, the two-point quadratic interpolation uses not only the value of the

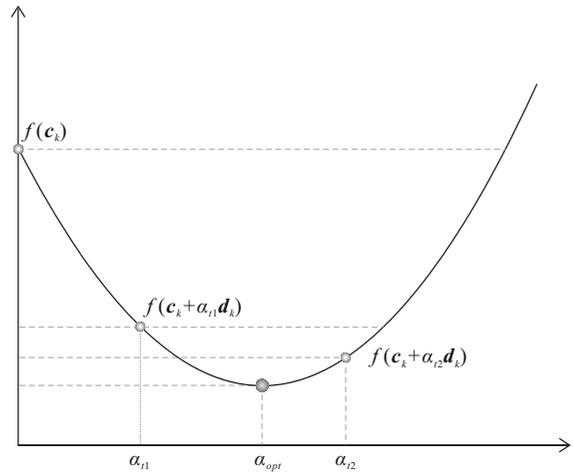


Figure 1

Illustration that schematically outlines the principle of the parabolic search step-length formula

misfit function but also the gradient at current state. Comparing the terms of the right sides of the expression (16), we obtain the following relationships:

$$\begin{cases} c = E(\mathbf{c}_k) \\ b = [\nabla E(\mathbf{c}_k)]^T \mathbf{d}_k = \frac{\partial E(\mathbf{c}_k + \alpha \mathbf{d}_k)}{\partial \alpha} \Big|_{\alpha=0} \end{cases}. \quad (20)$$

If the value of the objective function calculated with the velocity model updated by a test step-length α_t is available, we obtain

$$\begin{cases} E(\mathbf{c}_k) = c \\ E'(\mathbf{c}_k) = b \\ E(\mathbf{c}_k + \alpha_t \mathbf{d}_k) = c + b\alpha_t + a\alpha_t^2 \end{cases}. \quad (21)$$

Here, the prime represents the derivative with respect to α . With the help of the relationships (20) and (21), the three undetermined coefficients are as follows:

$$\begin{cases} a = \frac{E(\mathbf{c}_k + \alpha_t \mathbf{d}_k) - E(\mathbf{c}_k) - b\alpha_t}{\alpha_t^2} \\ b = [\nabla E(\mathbf{c}_k)]^T \mathbf{d}_k \\ c = E(\mathbf{c}_k) \end{cases}. \quad (22)$$

Applying Eq. (16) and the condition of minimum ($\alpha_{\text{opt}} = -b/2a$), we can obtain the optimum step-length

$$\alpha = -\frac{b}{2a} = -\frac{b\alpha_t}{[E(\mathbf{c}_k + \alpha_t \mathbf{d}_k) - E(\mathbf{c}_k) - b\alpha_t]} \alpha_t. \quad (23)$$

For this step-length formula, the test step-length α_t is crucial. Although Tape et al. (2007) suggested

the test step-length as $\alpha_t = -\frac{2E(c_k)}{b}$, i.e., the trough of the parabolic curve, it may cannot reduce the objective function value, because this test step-length cannot ensure that the desired step-length is bracketed in the interval $[0, \alpha_t]$. In this study, we first start from a small step-length to check whether it satisfies the following condition:

$$\begin{cases} \alpha_t > 0 \\ f(c_k + \alpha_t d_k) \geq f(c_k) \end{cases} \quad (24)$$

If the condition (24) is not satisfied, then we increase the test step-length α_t until to satisfy. Once this condition is met, the interval $[0, \alpha_t]$ brackets the optimum step-length. The principle of the two-point quadratic interpolation method is illustrated in Fig. 2. To compute the optimum step-length, the two-point quadratic interpolation method needs at least one extra forward modeling.

In terms of computational efficiency, at each iteration, the two-point quadratic interpolation method (hereafter named simply *Interp*) may be slightly costlier than the step-length of the direct method (hereafter named *Direct*) as above analysis. However, the parabolic search method (hereafter named simply *Search*) is most inefficient as it needs at least two times extra forward modeling. In the next section, we investigate the efficiency, accuracy, and robustness of these step-length formulas. To improve efficiency, we only carry out forward modeling from

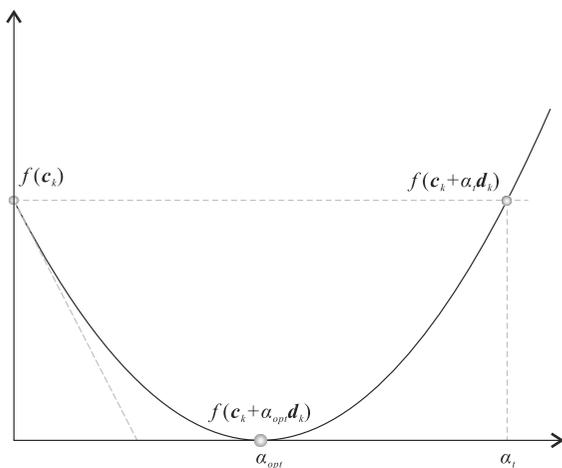


Figure 2

Illustration that schematically outlines the principle of the two-point quadratic interpolation step-length formula

a certain number of evenly distributed shot points to determine the optimum step-length for a test step-length.

To measure the accuracy (error) of the inverted results, we use the following mean absolute percentage error (MAPE):

$$\varepsilon = \frac{100}{N} \sum_n \left| \frac{c_{\text{true}}^n - c_{\text{inv}}^n}{c_{\text{true}}^n} \right|, \quad (25)$$

where N is the total number of grids of the discretized model; $||$ represents the absolute operator; c_{true} and c_{inv} are the real and inverted velocity models, respectively; and n is the index of the grid point. The smaller the MAPE, the more accurate the inverted result.

2.4. Frequency-Band Selection Strategy

Bunks et al. (1995) were the first to propose a multiscale approach to mitigate the nonlinearity of FWI in time domain. They adopted a finite-impulse response Hamming-window filter for low-pass filtering the seismic source wavelet and data. This multiscale scheme performs the inversion sequentially from low-frequency data to high-frequency data. As a result, the multiscale FWI is more likely to reach the global minimum (Sirgue and Pratt 2004; Boonyasiriwat et al. 2009), because the misfit function at low-frequency band is more linear with respect to the slowness than at high-frequency band (Bunks et al. 1995). However, the Hamming-window function is a leaky low-pass filter, which means that the Hamming-windows low-pass filter may contribute to the objective function value. Boonyasiriwat et al. (2009) proposed a Wiener low-pass filter that is more efficient than the Hamming-window low-pass filter. The amplitudes of leaked high-frequency components with the Wiener filter are several orders of magnitudes smaller than those obtained with the Hamming filter. In this paper, we adopt the Wiener filter for inversion. To further improve the efficiency of FWI, we utilize a larger time interval at low-frequency band and a smaller time interval at high-frequency band according to the Courant number.

Initially, Sirgue and Pratt (2004) proposed a frequency-band selection scheme for FWI in

frequency domain. The idea behind this scheme is to reduce the redundancy of information in wavenumber coverage as much as possible by selecting fewer frequencies. This scheme depends on the maximum effective offset presented in the surface seismic survey. The larger the range of offsets, the fewer frequency slices are required. Later, Boonyasiriwat et al. (2009) extended this scheme to time-domain FWI with the Wiener low-pass filter (Boonyasiriwat et al. 2009). With this filter, the user-defined parameter is only the dominant frequency of the target wavelet.

Here, we present an alternative frequency-band selection scheme using the Wiener low-pass filter. According to the idea of the separation of frequency ranges in the multiscale strategy, seismic data and wavelet must be decomposed at different scales (such as different wavelength components), which helps FWI to converge to a global minimum. Time-domain FWI uses multiple frequencies simultaneously during the inversion, which allows update a much wider range of wavenumbers than using only a single frequency at one time, as happens with frequency-domain FWI. As a result, a strong redundancy in the wavenumber domain is presented in the data. If the range of the velocity values of the model is fixed, the range of the wavelength components is determined by the source wavelet spectrum, since the frequency bandwidth can adjust the range of wavelength or wavenumber (Sirgue and Pratt 2004). Given that low, intermediate, and high frequencies correspond to long, intermediate, and short wavelengths, respectively, reducing the frequency redundancy is more significant to produce effective inversion at each scale. Therefore, we need decompose the data and wavelet into different frequency bands to reduce such redundancy. This inspires us to design a frequency-band selection strategy based on the decomposition of the source wavelet spectrum into different frequency bands, so that adjacent frequency bands have less redundant information in frequency to adapt to wavelength coverage.

If we consider that the dominant amplitude of the source spectrum is bounded by its frequency bandwidth, i.e., half the maximum amplitude in the frequency spectrum, the frequency components whose amplitudes are smaller than half the maximum

amplitude are then considered to have insignificant contributions to the recovery of wavelength. Figure 3 shows the amplitude spectra of the Ricker wavelet with dominant frequency of 1.07, 4.85, and 22 Hz; the dashed lines indicate half the maximum amplitudes, respectively. As can be seen, these amplitude spectra for adjacent frequency bands intersect just at half the maximum amplitude for the higher frequency band, thus giving rise to less redundancy in frequency. In this example, the two small shaded regions A and B illustrate the redundant information in frequency that can be discarded. This is due to how the seismic data are band-limited naturally. Each frequency component of the data has a different amplitude or strength, which results in a band-limited range of the recovered wavelength components. Strong frequency components of the data largely contribute to the wavelength update, whereas weak frequency components (especially at the low and high ends of a frequency band) provide weak contributions (Boonyasiriwat et al. 2009). Consequently, the frequency-band selection strategy should make that the overlapped region of wavelength recovered from

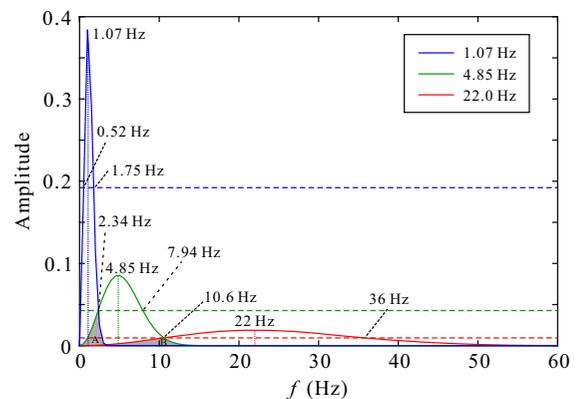


Figure 3

Multiscale strategy for selection of frequency bands. The blue, green, and red solid lines represent the amplitude spectra of the Ricker wavelet with dominant frequency of 1.07, 4.85, and 22 Hz, respectively; while the blue, green, and red dashed lines indicate the half the maximum amplitude of each spectrum, respectively. These amplitude spectra for adjacent frequency bands intersect just at half the maximum amplitude for the higher frequency band. These intersection points are marked in the illustration by their respective spectral frequencies: 1.75, 7.94, and 36 Hz. Analogously, the crossing points between two adjacent spectra are also marked by their respective frequencies: 2.34 and 10.6 Hz. The two small shaded regions A (around 2.3 Hz) and B (around 10 Hz) illustrate the redundant information in frequency

two consecutive frequency bands to be minimal. This is equivalent to implement the strategy of Sirgue and Pratt (2004) in time-domain.

Continuing with the example of the Ricker wavelet, its frequency band can be determined by two frequency points, i.e., $f_{l,h} = c_{l,h}f_0$, where f_0 is the dominant frequency of the target Ricker wavelet. At half the maximum spectral amplitude, the lower frequency point is set as $f_l = c_l f_0$, while the higher frequency point is set as $f_h = c_h f_0$, being $c_l = 0.482$ and $c_h = 1.637$. The detailed deduction of these two constants can be seen in Appendix 1 or the paper of Wang (2015). Both frequency points define the frequency bandwidth of the Ricker wavelet. To proceed to an adequate frequency-band selection, i.e., the frequency or wavelength information contained in adjacent frequency bands be less redundant, we set the equality

$$A_{i-1}(c_l f_0^i) = A_i(c_h f_0^i), \quad (26)$$

where f_0^i is the dominant frequency of the target Ricker wavelet within the higher frequency band used in the Wiener low-pass filter; A_{i-1} and A_i are the amplitude spectra of the two adjacent frequency bands; i denotes the frequency band index. This equality requires that the frequency spectrum of the lower frequency band intersects the frequency spectrum of the higher frequency band just at the lower frequency point $f_l^i = c_l f_0^i$ (of the higher frequency band) that defines half the maximum spectral amplitude. After deduction (see Appendix 2), the frequency-band selection strategy is performed by the following recursive formula:

$$f_0^{i-1} = f_0^i / c_0, \quad (27)$$

where $c_0 = 4.533$ is a constant (see Appendix 2); f_0^{i-1} is the dominant frequency of the target Ricker wavelet within the lower frequency band. This relationship expresses the recursive relation between the dominant frequencies of the target Ricker wavelets within the adjacent frequency bands, thus allowing the effective frequency selection to implement the multiscale strategy. With the help of the above formula (27), the data are decomposed into several frequency bands, so that the adjacent frequency bands have less redundancy in wavelength or wavenumber information. The gist of this scheme lies in

establishing the intersection between the lower and higher frequency bands just at the point where the amplitude is half the maximum amplitude of the higher frequency band.

Usually, the seismic data and wavelet are decomposed into three wavelength intervals, i.e., long, intermediate, and short components (sometime in more wavelength components). In this study, we take the dominant frequency within the highest frequency band as a real wavelet. Figure 3 schematically outlines the frequency-band selection strategy. The dominant frequency of the real wavelet is 22 Hz. By applying the above-depicted strategy, it can be observed that the crossing point between the third and second frequency bands is at 10.6 Hz just where the amplitude is half the maximum amplitude within the third frequency band. Analogously, the crossing point between the second and first frequency bands is at 2.34 Hz where the amplitude is half the maximum amplitude within the second frequency band. The small shaded domains show the overlapped regions A (around 2.3 Hz) and B (around 10 Hz) between adjacent frequency bands, which highlights that the redundancy in frequency information is very small. Figure 4 shows the wavelength coverage of the three frequency bands (dominant Ricker frequencies of 1.07, 4.85, and 22 Hz) in a homogeneous medium with compressional velocity value of 1.5 km/s. It is clear that the wavelength coverage of the highest frequency band is very narrow, while the one of the lowest frequency band is relatively wide. The former can provide a good constrain on the fine structure of the model, while the latter can also describe the macro or background model. In general, the wavelength components within the three frequency bands present less redundancy. The effectiveness of the frequency-band selection strategy will be investigated in the next section.

3. Numerical Examples

3.1. Effectiveness of the Frequency-Band Selection Strategy

In this section, we first investigate the effectiveness of the frequency-band selection strategy proposed in this study compared with the strategy

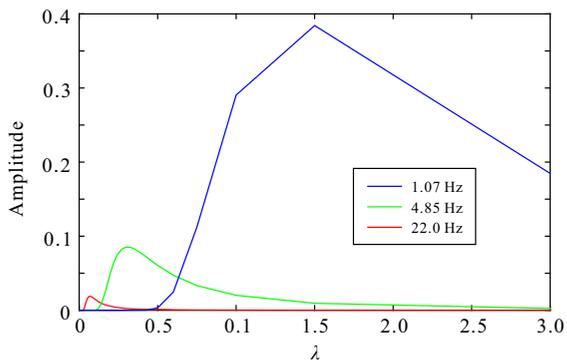


Figure 4

Wavelength coverage of three frequency bands (dominant frequencies of 1.07, 4.85, and 22 Hz) in a homogenous medium with compressional velocity of 1.5 km/s

proposed by Boonyasiriwat et al. (2009). We adopt the Marmousi model to perform this investigation. The model is shown in Fig. 5a and consists of 383×142 grid-cells in the horizontal and vertical directions, respectively. Both the horizontal and vertical grid spacings are 10 m. Up to 383 seismic receivers are evenly deployed on the surface with fixed-spread acquisition geometry. The seismic source is located at the depth of 0.05 km and is modeled by a Ricker wavelet with dominant frequency of 22 Hz. The synthetic data come from 38 shots separated by an interval of 0.1 km. The recording length is 3.2 s and the sampling interval is 0.8 ms.

Figure 5b shows a Gaussian smoothed version of the Marmousi model that is taken as the initial velocity model. The initial model is first computed by smoothing the real velocity model with a 2D Gaussian function of the vertical correlation and horizontal correlation ranges of 0.5 km, and then by implementing a 1D Gaussian function of the horizontal correlation range of 1 km. It can be seen that the initial velocity model (Fig. 5b) deviates substantially from the real model. Therefore, it allows us to investigate the effectiveness of the frequency-band selection strategy. According to the definition of the MAPE (Eq. 25), the MAPE related to the initial model (Fig. 5b) is 10.8%.

Essentially, the strategy of Boonyasiriwat et al. (2009) required the continuity of the vertical wavenumber, whereas our strategy emphasizes the less redundancy in adjacent frequency bands.

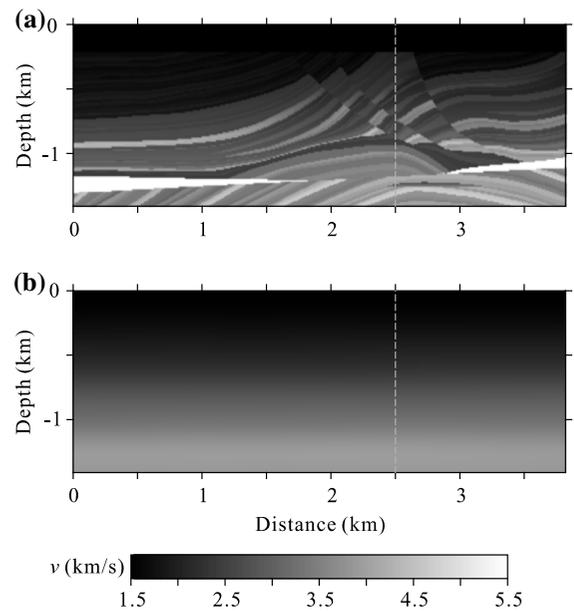


Figure 5

Marmousi model. **a** Real velocity model; **b** a Gaussian smoothed version of the Marmousi model as the initial velocity model. The dashed gray line represents a cut for further analysis

Although three frequency bands are preferable, because the low, intermediate and high frequencies, respectively, constrain the long, intermediate, and short components of the model, enough low-frequency information is unavailable in real data. To mimic realistic cases, we just consider the intermediate- and high-frequency bands. Based on the frequency-band selection strategy proposed in this study (relation 27), now the seismic data and wavelet can be decomposed into two scales with the Wiener low-pass filter (Boonyasiriwat et al. 2009), namely: (3.1–10.6 Hz) and (10.6–36 Hz). Correspondingly, according to the strategy of Boonyasiriwat et al. (2009), the seismic data and wavelet can also be decomposed into two frequency bands, namely: (2.5–8.5 Hz) and (10.6–36 Hz). Here, we use half the maximum spectral amplitude to define the frequency bandwidth of the seismic wavelet. We recommend the readers going to the paper of Boonyasiriwat et al. (2009) for implementation details. In contrast, the two frequency bands obtained by Boonyasiriwat et al. (2009) have a frequency gap (i.e., jumping from 8.5 to 10.6 Hz), while our strategy produces two continuous frequency bands.

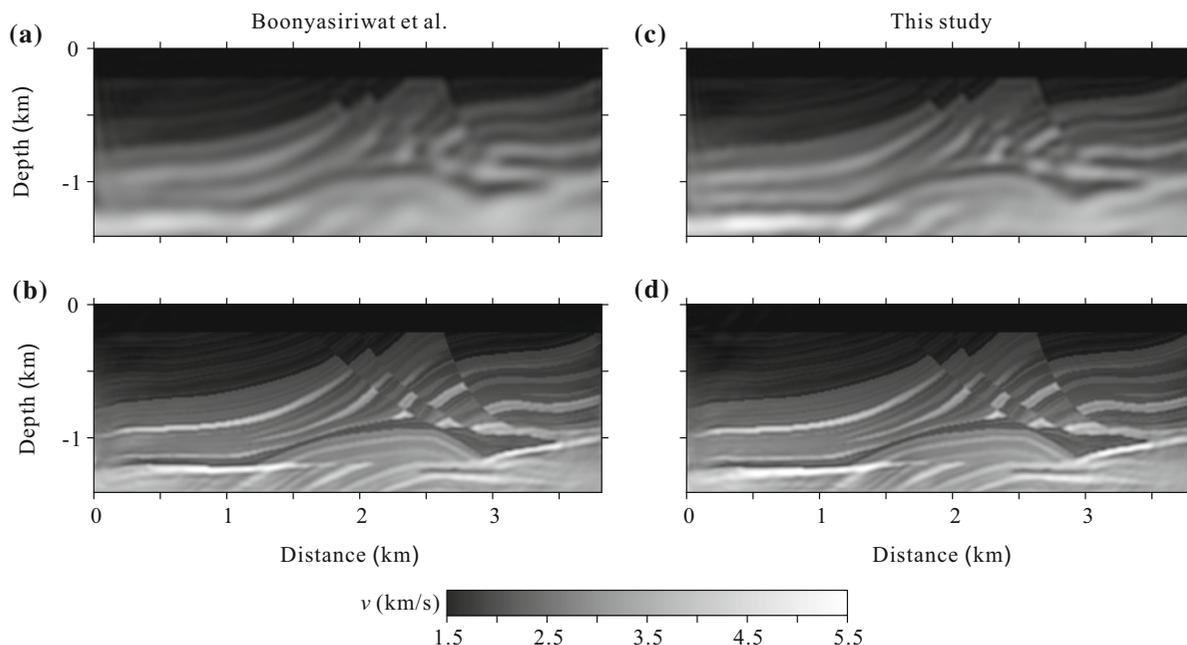


Figure 6

Multiscale velocity models obtained by FWI with different frequency-band selection strategies. The initial velocity model is shown in Fig. 5b. From left to right inverted results provided by the strategies of Boonyasiriwat et al. (2009) and proposed in this study operating with two frequency bands (from top to bottom a–b and c–d, respectively). The inversion method is the L-BFGS method. The step-length formula is the direct method

Furthermore, the strategy of Boonyasiriwat et al. (2009) adopts more unrealistic low-frequency information (2.5 Hz) than our strategy. As this strategy repeatedly uses the low-frequency components of the data in the inversion, some frequency components overlap is allowed. The overlapped region of the frequency bands generated by our strategy is very small, so that it can be ignored (Fig. 3).

Based on the two frequency-band selection strategies mentioned above, we carry out the time-domain full waveform inversion with the Marmousi model. Here, we adopt the L-BFGS method as inversion method and the *Direct* as step-length formula. Throughout this paper, the stopping criterion is that the relative change rate in the objective function value is less than 0.0001, or that the number of iterations (at each scale or frequency band) exceeds 400. Because the inversion process may be unstable when the initial velocity model deviates far from the real model or the data are contaminated by noise, and hence, 20% increase in the objective function value is allowed.

The inverted velocity models obtained by FWI with different frequency-band selection strategies are shown in Fig. 6. From left to right, we display the inverted results provided by each one of the two frequency-band selection strategies, while from top to bottom, we show the results concerning the first and second frequency scales, respectively. Because the frequency-band selection strategy of Boonyasiriwat et al. (2009) uses lower frequency information than the strategy presented in this study (Eq. 27), therefore, the inverted result obtained by the former has a slightly lower resolution than that of the latter at the first frequency band. Thus, the former generates a more accurate background velocity model than the latter. Intuitively, both strategies yield almost identical high-resolution results at the second frequency band. To compare their accuracy, we calculate their MAPEs. The MAPEs corresponding to the final results obtained by the two strategies are 5.817 and 6.258%, respectively. Compared with the MAPE in relation to the initial model (10.8%), the MAPEs associate with the inverted results reveal a significant

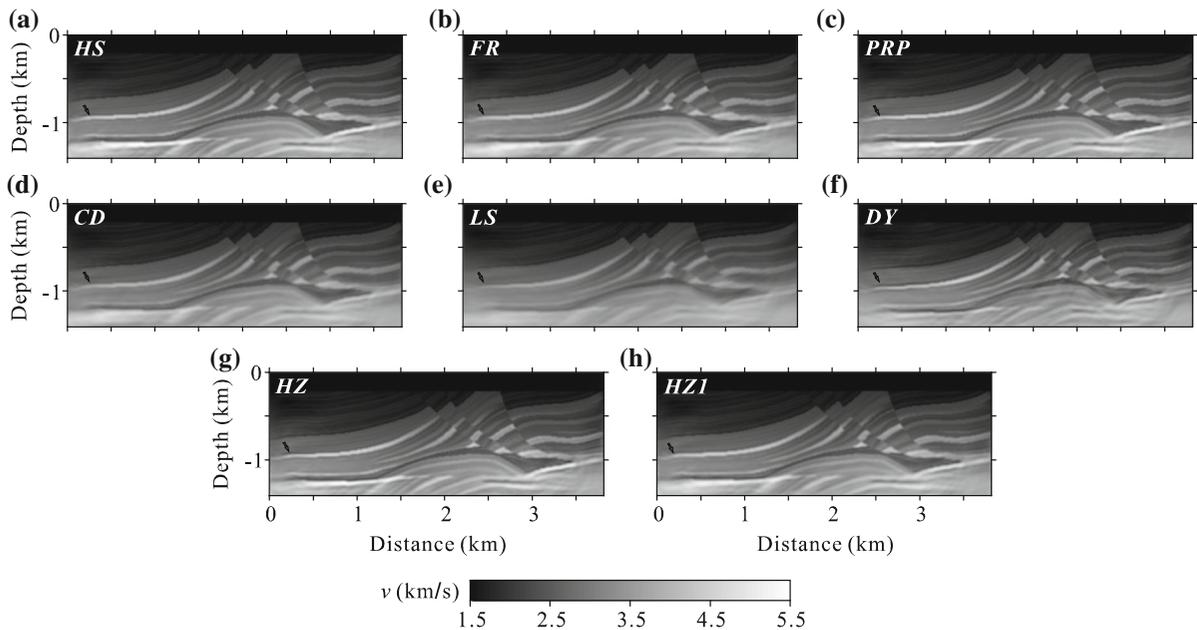


Figure 7

Multiscale velocity models obtained by FWI with different versions (eight update parameters) of the nonlinear CG method. The initial velocity model is shown in Fig. 5b. The acronym on the *top-left corner* of each image makes reference to the inversion with the nonlinear CG method used for computation (see the text in Sect. 2.2). The inversion method is the L-BFGS method. The step-length formula is the direct method

decrease, which demonstrates the effectiveness of the frequency-band selection strategy proposed in this study. Although the frequency-band selection strategy of Boonyasiriwat et al. (2009) obtains a slightly more accurate result, it uses unrealistic low-frequency information.

3.2. Analyzing Distinct CG Update Versions

As we already indicated, at present, we have eight possible options to choose the update parameter β for applying the CG method (Eqs. 8a–h). Here, with the purpose of investigating their respective behaviors in the context of FWI, we taking advantage of the previously synthesized shot gathers from the Marmousi model reproduced in Fig. 5a. In the inversion process, we adopt the *Direct* as our step-length formula. In the following sections, we always adopt the frequency-band selection proposed in this study to decompose seismic data and wavelet by two frequency bands.

3.2.1 Noise-Free Data

The inverted velocity models obtained by FWI with distinct versions (eight update parameters) of the nonlinear CG method are shown in Fig. 7a–h. The acronym on the top-left corner of each image makes reference to the inversion with the used CG method for computation. In this experiment, the basic algorithm is completely identical except different nonlinear CG update parameters. This means that the smaller the number of iterations, the higher the computational efficiency. Obviously, the solution contributed by the *LS* has a slightly lower resolution, whereas the *DY* severely overestimates the deeper parts of the model. To confirm this point, the velocity profiles at the horizontal location of 2.5 km are shown in Fig. 8. It can be seen that the velocity profile obtained by the *DY* always focuses on wrong spatial position, while the profile obtained by the *LS* always has smaller amplitude. In contrast, the other versions obtain comparatively high-resolution results,

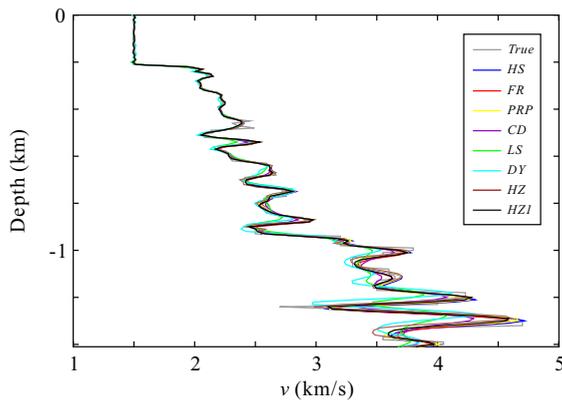


Figure 8

Velocity-depth profiles concerning the Marmousi model and with reference at the horizontal location of 2.5 km (dashed gray line in Fig. 5). Gray line denotes the real velocity model, while other lines represent inverted results obtained by the eight versions of the nonlinear CG method, respectively

which recovery the fine structure of the Marmousi model quite faithfully. The high-velocity layers in the bottom-left corner of the inverted models are slightly distorted, which may be due to inaccurate initial velocity model (marked by arrows in the illustration).

To quantitatively check the computational efficiency and accuracy of different versions of the nonlinear CG method, we use the number of iterations to analyze efficiency and the MAPE (Eq. 25) to analyze the model misfit related to each of the

implemented algorithms. Figure 9a shows the value of the respective error functions (E) versus the number of iterations (in parentheses, top-right inset). One can observe that the FR converges fastest followed by the HS and PRP , while the LS is costliest followed by the CD and HZ (Fig. 9a). The DY and HZI reveal comparable mediate convergence. Even though the number of iterations reaches 400, the LS still does not converge. It demonstrates the inefficiency of the LS (Fig. 9a). In addition, the objective function values corresponding to some versions are oscillating at the beginning of each frequency band, such as the DY and HZ , which may be due to the less-accurate initial velocity model.

Figure 9b shows the MAPEs in correspondence with the inverted models depending on the version of the nonlinear CG method. It can be seen that the CD , HS , and PRP are comparatively accurate followed by the FR , HZI , LS , and HZ , while the DY is the most inexact relatively because of focusing on wrong spatial location. In general, compared with the MAPE related to the initial model (10.8%), the MAPEs that associate with the inverted results are significantly decreased, which illustrates the effectiveness of the eight versions of the nonlinear CG method. When taking jointly efficiency and accuracy into consideration, it is not difficult to conclude that the HS , CD , and PRP are more efficient than the other versions. In

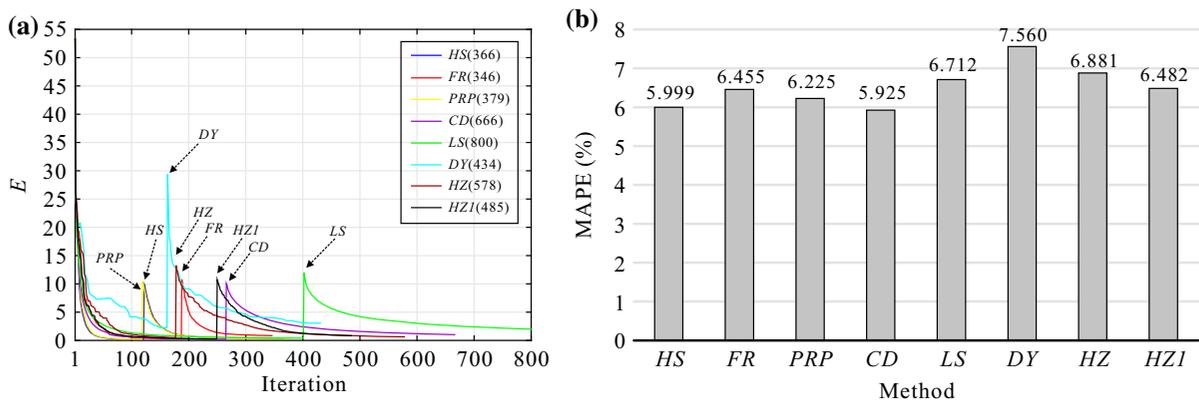


Figure 9

Statistics of the error functions (E) and computational accuracy for different versions of the nonlinear CG method used with the Marmousi model shown in Fig. 5. **a** Error functions (E) versus the number of iterations. The numbers in parentheses (top-right inset) give the total number of iterations. **b** Histogram of the MAPEs for the inverted velocity models depending on the version of the nonlinear CG method. The numbers over bars are the corresponding MAPEs

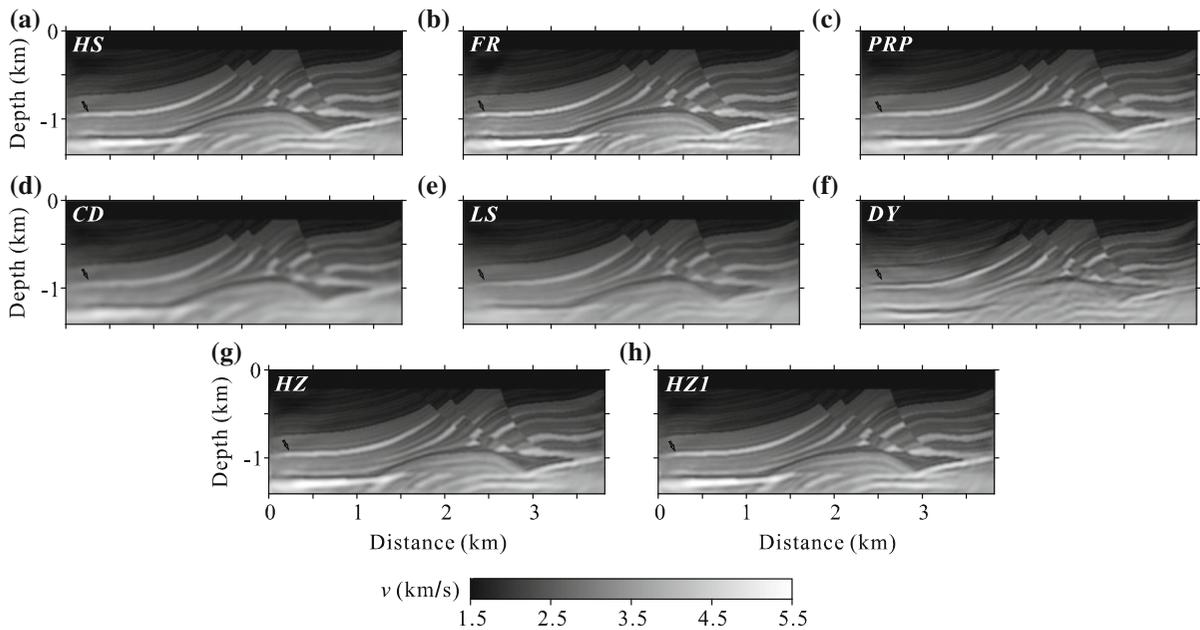


Figure 10
Same as Fig. 7 using noisy data with signal-to-noise ratio (SNR) of 20 dB

addition, this experiment also proves the effectiveness of the frequency-band selection strategy proposed in this study.

3.2.2 Noisy Data

In the previous experiment, we draw the conclusion just from the noise-free data. Unfortunately, the data are always contaminated by stochastic or/coherent noise in real cases. To check the robustness of the distinct versions of the nonlinear CG method, we added Gaussian white noise to the data set generated from the Marmousi model to achieve a significantly decreasing signal-to-noise ratio (SNR) of 20 dB. The acquisition geometry and physical parameters are the same as in Sect. 3.1.

Figure 10 shows the inverted velocity models obtained by FWI with distinct versions (eight update parameters) of the nonlinear CG method. It can be observed that the *DY* still overestimates the deeper parts just as that obtained in the noise-free data. In general, the *CD* and *LS* generate relatively low-resolution images, while the other versions obtain comparatively high-resolution results. At the first

glance, it is difficult to distinguish from each other. To evaluate their efficiency and accuracy, Fig. 11 shows the convergence curves and MAPEs in correspondence with each nonlinear CG method. Figure 11a shows the convergence curves. Obviously, the *LS* still converges slowest followed by the *FR*, *HZ*, and *HZI*. The *HS*, *CD*, and *DY* converge after a small number of iterations. The *PRP* reveals a moderate convergence. In addition, the inversion with noisy data is prone to instability compared with the convergence curves from the noise-free data (Fig. 9). Although the objective function values are significantly oscillating within the first 100 iterations, they eventually decrease to a comparable level. In particular, the objective function value of the *DY* has the largest increase within first 100 iterations. This illustrates that it is possible to obtain a more accurate result at the cost of allowing local increase in the objective function. In this study, we allow 20% increase in the objective function.

To compare the respective accuracy of each version, we show the MAPEs in Fig. 11b. The *CD*, *HS*, and *PRP* are comparatively accurate followed by the *LS*, *HZI*, *HZ*, and *FR*, which is similar to the

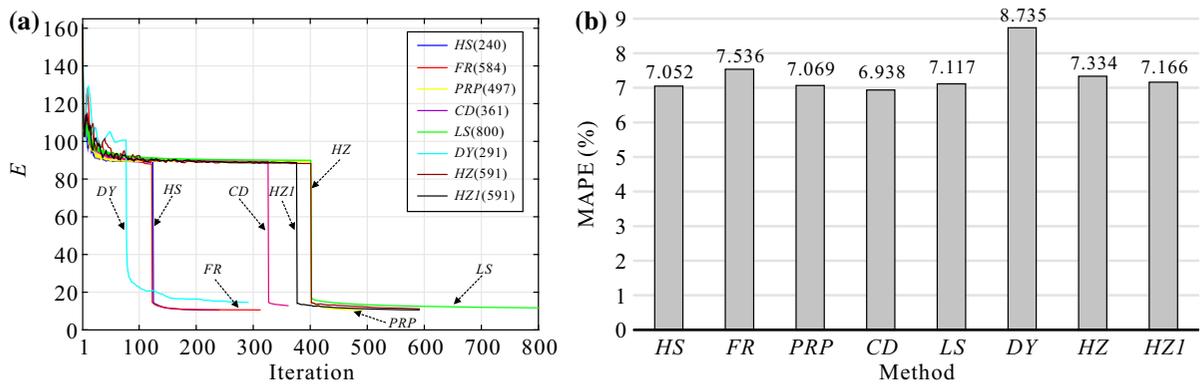


Figure 11
Same as Fig. 9 using noisy data with signal-to-noise ratio (SNR) of 20 dB

noise-free case. The *DY* is still the most inaccurate. In general, the MAPEs related to the noisy data (Fig. 11b) are consistently larger than those obtained with noise-free data (Fig. 9b). It indicates that the noise can prevent the inversion from converging to a more accurate result. Once again, when taking jointly efficiency and accuracy into consideration, it can be concluded that the *HS*, *CD*, and *PRP* are relatively efficient among all versions.

3.3. Comparison of Step-Length Formulas with Noise-Free Data

In this section, we investigate the behaviors of the three step-length formulas given previously (Eqs. 16, 21, and 26), i.e., the so-called direct method (*Direct*), the parabolic search method (*Search*), and the two-point quadratic interpolation method (*Interp*). In terms of the inversion methods, the L-BFGS method is the well-known most efficient one among the SD, CG, and L-BFGS methods; therefore, we adopt it as our inversion method in the following sections. Throughout the following numerical experiments, we adopt the same stopping criterion defined in Sect. 3.1. To explore the sensitivity of the three step-length formulas to the complexity of the model, we first consider both a simple structure in geometry and velocity (like a basin-shaped model) and a more complex one (for instance, the Marmousi model) with noise-free data.

3.3.1 Basin-Shaped Model

First, we designed a basin-shaped model as real model (Fig. 12a), from which we constructed a smoothed version as the initial model (Fig. 12b) with the MAPE of 8.425%. The model consists of 321×201 grid-cells in the horizontal and vertical directions, respectively. Both the horizontal and vertical grid spacings are 10 m. Up to 321 seismic receivers are evenly deployed on the surface with fixed-spread acquisition geometry. The seismic source is located at the depth of 0.05 km and is modeled by a Ricker wavelet with dominant frequency of 30 Hz. The synthetic data are acquired from 32 shots separated by an interval of 0.1 km. The recording length is 2 s and the sampling interval is 1 ms. As before, based on the frequency-band selection strategy proposed in this study (relation 27), the data and seismic wavelet are decomposed into two scales with the Wiener low-pass filter (Boonyasiriwat et al. 2009), namely: (4.3–14.4 Hz) and (14.4–49.1 Hz).

Figure 13 shows the multiscale velocity models recovered from the smoothed basin-shaped model using the L-BFGS method with different step-length formulas. From left to right, we display the inverted results provided by each of these formulas, while from top to bottom, we show the results concerning the first and second frequency scales, respectively. As can be seen, the results obtained with any of these

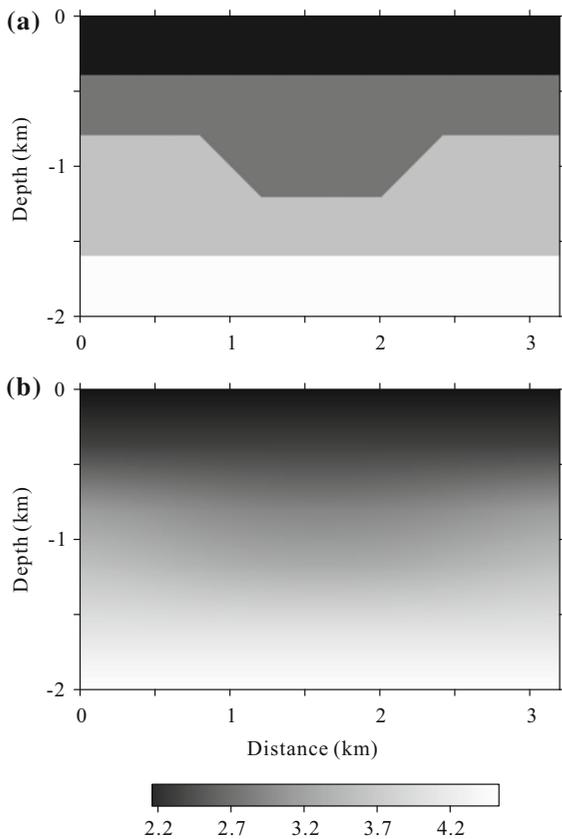


Figure 12

Basin-shaped model. **a** Real velocity model; **b** initial velocity model

formulas (*Direct*, *Search*, and *Interp*) are perfectly comparable. Hence, it is difficult to distinguish from each other intuitively. In particular, the two ends of the third layer (deepest) are somewhat raised (such as is marked by arrows in the illustration) due to the poor initial velocity model and insufficient data coverage.

To appreciate the efficiency and accuracy of the three step-length formulas, we resort to the respective error (E) functions versus the number of iterations and also to the MAPEs. Figure 14 shows these error functions. The numbers in parentheses denote the total number of iterations performed in each case. Both the *Direct* and *Interp* converge to almost an identical minimum after smaller number of iterations, while the *Search* needs more number of iterations to converge. The MAPEs corresponding to the final inverted results obtained by the three step-length formulas are 3.577, 3.185, and 3.166%, respectively.

The MAPEs have been significantly decreased, which demonstrates that the model is well reconstructed. It can be observed that both the *Search* and *Interp* reveal similar accuracy. In contrast, the *Direct* is slightly less accurate. When taking efficiency and accuracy into consideration, the *Interp* is the most efficient algorithm assuming a simple model.

3.3.2 Marmousi Model

To investigate the sensitivity of the three step-length formulas (*Direct*, *Search* and *Interp*) to the complexity of the model, we carry out the same numerical test starting from the Marmousi model. The real and initial velocity models are those already shown in Fig. 5. The acquisition geometry and physical parameters are identical to those in Sect. 3.1.

Figure 15 shows the multiscale velocity models obtained using the L-BFGS method with different step-length formulas (*Direct*, *Search* and *Interp*). From left to right, we display the inverted results provided by these step-length formulas, while from top to bottom, we show the results related to the first and second frequency scales, respectively. It is easy to distinguish that the result provided by the *Interp* exhibits a slightly lower resolution at the second frequency scale when compared with those obtained by the *Direct* and *Search*. In general, the results obtained with the three step-length formulas can depict the fine structures integrating the Marmousi model at the second frequency band.

Again, the respective error functions and the MAPEs allow us to appreciate the efficiency and accuracy provided by the three step-length formulas. Figure 16 shows the error function (E) versus the number of iterations. The numbers in parentheses denote the total number of iterations. Now, the *Search* still needs the maximum number of iterations to converge, while the *Interp* needs least iterations to meet the predefined stopping criterion. In contrast, the *Direct* lies between them. The MAPEs obtained by the three step-length formulas are 6.258, 6.139, and 6.298%, respectively. Compared with the MAPE in relation to the initial model (10.8%), the Marmousi model has successfully been reconstructed. In this Marmousi model, both the *Direct* and *Interp* yield comparable results, while the *Search* obtains a

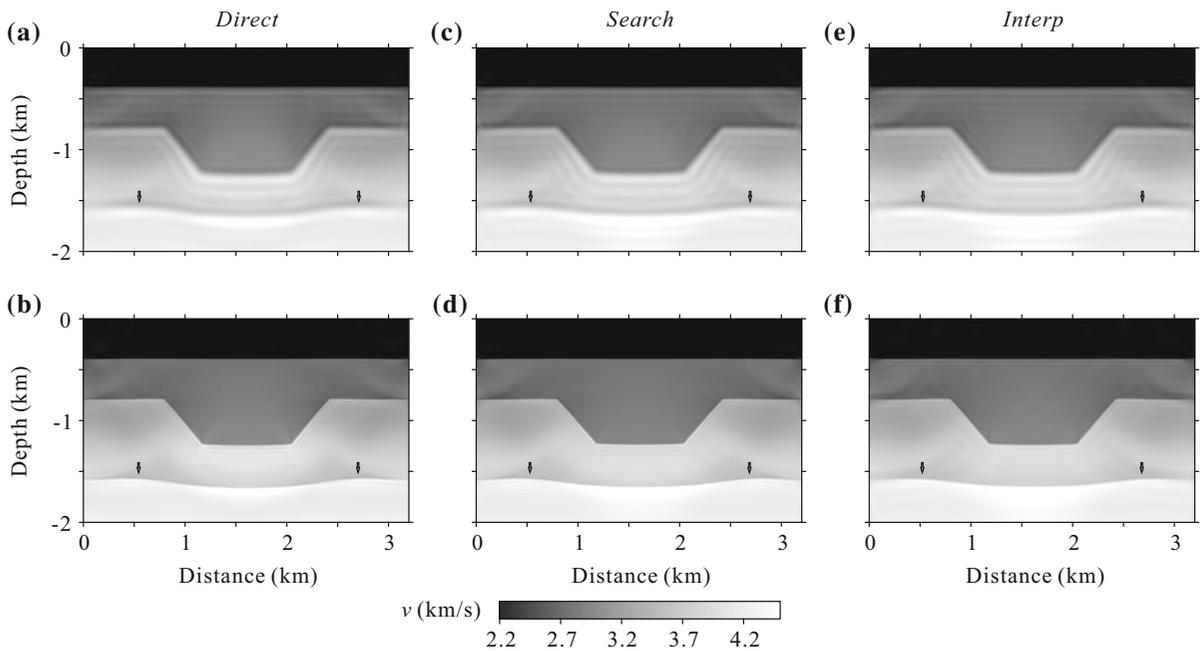


Figure 13

Multiscale velocity models recovered from the basin-shaped model using the three step-length formulas. From left to right inverted results provided by the direct method (*Direct*), the parabolic search method (*Search*), and the two-point quadratic interpolation method (*Interp*) operating with two frequency bands (from top to bottom **a–b**, **c–d** and **e–f**, respectively). The initial basin-shaped model is shown in Fig. 12b. The inversion method is the L-BFGS method

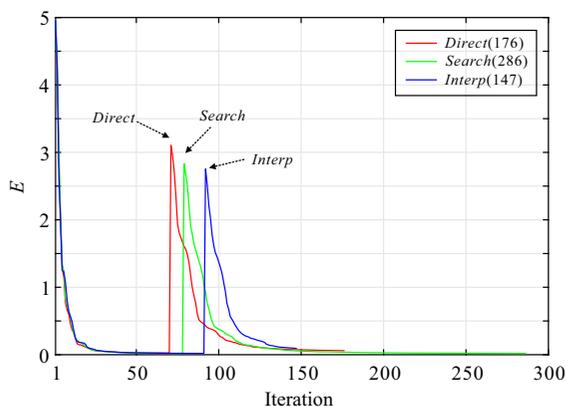


Figure 14

Error functions (E) versus the number of iterations in relation to the basin-shaped model and the used step-length formulas, namely: the direct method (*Direct*), the parabolic search method (*Search*), and the two-point quadratic interpolation method (*Interp*). The numbers in parentheses (*top-right inset*) give the total number of iterations. The initial basin-shaped model is shown in Fig. 12b. The inversion method is the L-BFGS method

slightly more accurate result at the cost of more number of iterations.

In summary, the results obtained with the three step-length formulas and two very different models, with remarkable differences as to its structural complexity, reflect the different behaviors of the three step-length formulas. Keeping the efficiency and accuracy in mind, the *Interp* is more efficient followed by the *Direct* even when dealing with a complex model, while the *Search* is slightly less efficient, because it needs more number of iterations to converge.

3.4. Comparison of Step-Length Formulas with Noisy Data

In all the previous examples, we have handled noise-free data. In this section, we investigate the robustness of the three step-length formulas with noisy data.

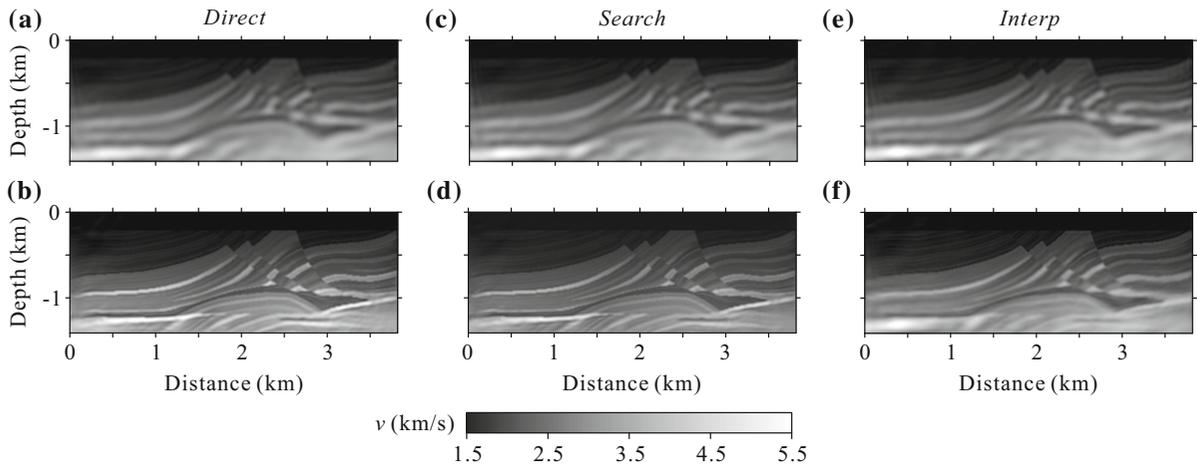


Figure 15

Multiscale velocity models recovered from the Marmousi model using the three step-length formulas. From left to right inverted results provided by the direct method (*Direct*), the parabolic search method (*Search*), and the two-point quadratic interpolation method (*Interp*) operating with two frequency bands (from top to bottom **a–b**, **c–d** and **e–f**, respectively). The initial velocity model is shown in Fig. 5b. The inversion method is the L-BFGS method

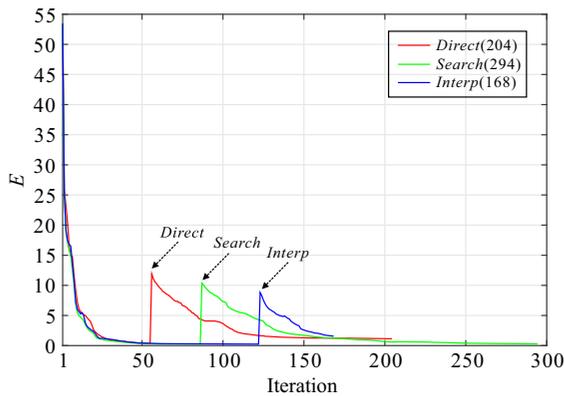


Figure 16

Error functions (E) versus the number of iterations in relation to the Marmousi model (Fig. 5) and the used step-length formulas, namely: the direct method (*Direct*), the parabolic search method (*Search*), and the two-point quadratic interpolation method (*Interp*). The numbers in parentheses (top-right inset) give the total number of iterations. The initial Marmousi model is shown in Fig. 5b. The inversion method is the L-BFGS method

3.4.1 Basin-Shaped Model

Here, we added Gaussian white noise to the data set generated from the basin-shaped model to achieve a decreasing signal-to-noise ratio (SNR) of 30 dB. The acquisition geometry and physical parameters are the same as in Sect. 3.3.1.

Figure 17 shows the multiscale velocity models obtained using the three step-length formulas and noisy data. In general, the inverted results obtained with the noisy data have slightly lower resolution compared to those obtained with the noise-free data (Fig. 13). In addition, the results obtained with the noisy data are clearly noisier, which is due to the travel-time mismatches caused by Gaussian white noise. Intuitively, the results obtained by the three step-length formulas are comparable at each frequency band. This supports that the three step-length formulas are robust or partly insensitive to Gaussian white noise.

To compare their efficiency and accuracy, we also display the error functions and the MAPEs in relation to each step-length formula. Figure 18 allows appreciating the efficiency and accuracy of the three step-length formulas through the respective error functions (E) versus the number of iterations. Compared with Fig. 14, the values of the objective function increase from 5 to about 60 at least initially, although they decrease quickly to a comparable value after a determined number of iterations. Within the first 25 iterations, the objective function values of the *Direct* and *Interp* are oscillating. In addition, the number of iterations becomes smaller due to the existence of noise. Like the tests in Sect. 3.3, both the *Direct* and

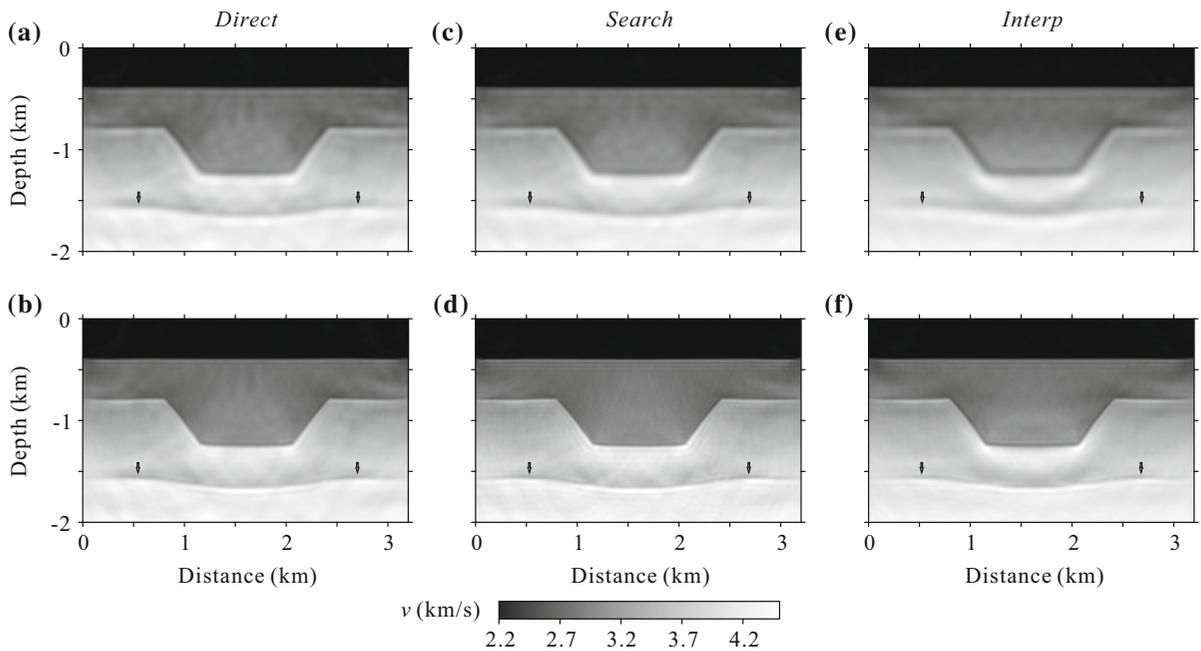


Figure 17
Same as Fig. 13 using noisy data with signal-to-noise ratio (SNR) of 30 dB

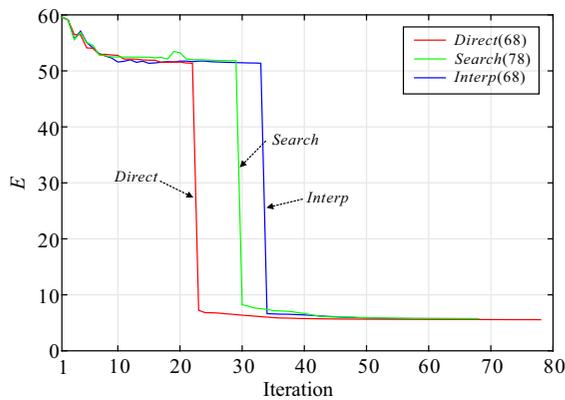


Figure 18
Same as Fig. 14 using noisy data with signal-to-noise ratio (SNR) of 30 dB

Interp can converge after smaller number of iterations, while the *Direct* needs slightly more iterations to meet the predefined stopping criterion. In this experiment, the MAPEs in correspondence with the final inverted results when using the three step-length formulas are 3.946, 3.960, and 4.042%, respectively. For this noisy data, both the *Direct* and *Search* generate more accurate results than the *Interp*. In

contrast, the *Direct* is more efficient than the *Search* and *Interp*, because it needs a relatively small number of iterations and obtains a more accurate result.

3.4.2 Marmousi Model

We continue check the robustness of the three step-length formulas in a complex model with Gaussian white noise data, i.e., the Marmousi model. Thus, we adopt the noisy data set used in Sect. 3.2. The acquisition geometry and physical parameters are the same as in Sect. 3.1. In this experiment, we still select the L-BFGS method as inversion method.

Figure 19 shows the multiscale velocity models obtained using the three step-length formulas and noisy data. The results obtained by the three step-length formulas generate comparable high-resolution results. This supports that the three step-length formulas are still robust for noisy data even in a complex model. In general, compared with Fig. 15, the results are now noisier at each frequency scale, as expected, which is due to the travel-time mismatches caused by Gaussian white noise. In particular, the shape of the gas and oil cap obtained by the *Interp*

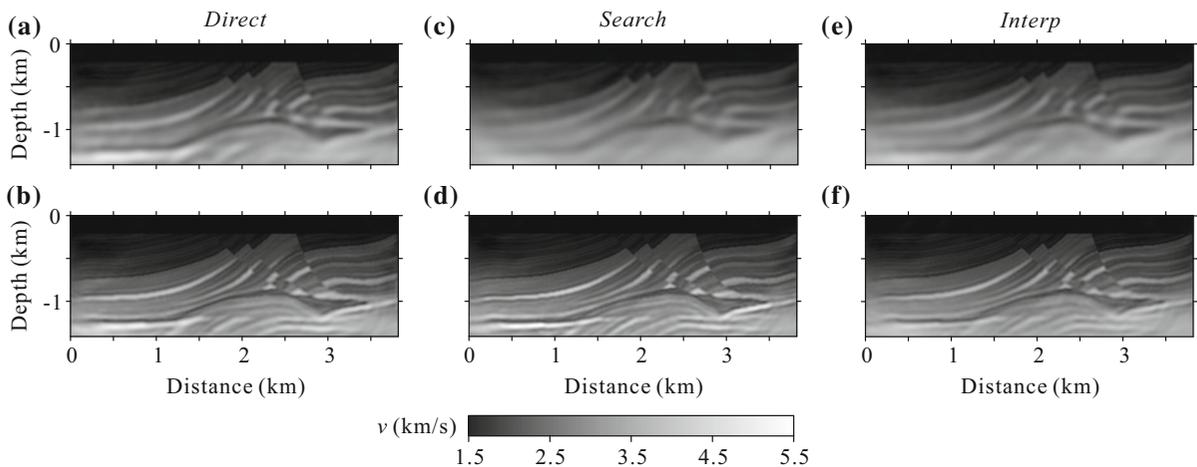


Figure 19

Same as Fig. 15 using noisy data with signal-to-noise ratio (SNR) of 20 dB

(Fig. 19f) is slightly less clear than those obtained by the *Direct* and *Search* (Fig. 19b, d).

Figure 20 allows comparing the efficiency and accuracy of the three step-length formulas through the respective error functions (E) versus the number of iterations. Compared with Fig. 16, the values of the objective function increase from 55 to more than 160 at least initially. In addition, the objective function values of the *Direct* and *Interp* are oscillating within the first 20 iterations, while the objective function values of the *Search* are consistently decrease. Once again, the number of iterations becomes smaller, because noise prevents the objective function value from further decreasing. The

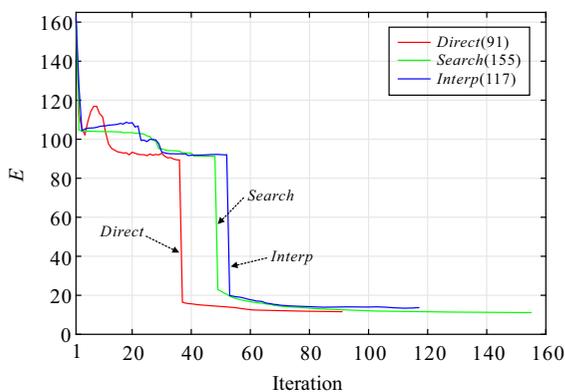


Figure 20

Same as Fig. 16 using noisy data with signal-to-noise ratio (SNR) of 20 dB

MAPEs corresponding to the final inverted results when using the three step-length formulas are 7.521, 7.887, and 7.772%, respectively. In general, the results obtained with the noisy data (Fig. 19) are consistently inaccurate than those obtained with the noise-free data (Fig. 15). In particular, the accuracy of the *Search* and *Interp* is comparable. Like the tests with the basin-shaped model, the *Direct* is more efficient than the *Search* and *Interp* because of relatively small number of iterations and high accuracy.

In summary, the above tests performed with different step-length formulas demonstrate that the *Interp* is more efficient from the viewpoint of the computational efficiency and accuracy with noise-free data, while the *Direct* is more efficient with noisy data. The *Search* is less efficient than both the *Direct* and *Interp*, because it always needs more number of iterations and needs at least two times extra forward modeling to estimate the optimum step-length.

4. Conclusions

First, we have implemented an alternative frequency separation strategy that aims at reducing the redundant information in frequency to adapt to the wavelength or wavenumber coverage. Then, we have performed a variety of numerical experiments with

the target of analyzing the effects of the update parameters for the nonlinear conjugate gradient method and three commonly used step-length formulas on FWI. We have analyzed up to eight different versions of the nonlinear conjugate gradient (CG) method, each distinguished by its respective update parameter. With noise-free and noisy data, the comparison of the inverted results using all these versions demonstrates that the *HS*, *CD*, and *PRP* nonlinear CG methods are more efficient among these versions.

Finally, three commonly used step-length formulas for FWI have been tested, identified as the direct (*Direct*) method, the parabolic search (*Search*) method, and the two-point quadratic interpolation (*Interp*) method. For noise-free data, the numerical tests prove that the *Interp* is more efficient than the others both with simple and complex models, while the *Search* is slightly less efficient, because it converges very slowly, and the *Direct* lies between them. The three step-length formulas were also applied to data contaminated by Gaussian white noise to obtain high-resolution images, thus proving the robustness of all of them. In general, the noisy data generate less accurate results compared to the noise-free data. For noisy data, the *Direct* is more efficient than the others both with simple and complex models, while the *Search* is still less efficient, because it needs more number of iterations, and the *Interp* lies between them. When the initial model deviates far from the real model or the data are contaminated by noise, the objective function values of the *Direct* and *Interp* are oscillating at the beginning of the inversion.

Acknowledgements

The authors are thankful to the Computer Simulation Laboratory of IGGCAS for allocation of computing time. We thank Colin Farquharson and two anonymous reviewers for insightful comments and suggestions, which greatly improved the manuscript. The authors thank the useful discussions with Qiancheng Liu. We are grateful to Jinhai Zhang for his assistance and the facilities given in the course of this work. We gratefully acknowledge the financial

support for this work contributed by the National key research and development program of China (Grants Nos. 2016YFC0600101, 2016YFC0600201 and 2016YFC0600402), the China Earthquake Administration (Grant No. 201408023), the National Natural Science Foundation of China (Grants Nos. 41604076, 41674102, 41604075, 41674095, 41522401, 41174075, 41474068, and 41374062), and the first class general financial grant from China Postdoctoral Science Foundation (Grant No. 2016M600128).

Appendix I

Frequency Width of the Ricker wavelet

The expression of the Ricker wavelet in time domain is as follows:

$$R(t) = \left(1 - 2\pi^2 f_0^2 (t - t_0)^2\right) \exp\left(-\pi^2 f_0^2 (t - t_0)^2\right), \quad (28)$$

where t is time; t_0 is the delayed time; f_0 is the dominant frequency. The Fourier transform of the Ricker wavelet is as follows:

$$F(\omega) = \frac{\sqrt{\pi}\omega^2}{\omega_c^3} \exp\left(-\frac{\omega^2}{\omega_c^2} + i\omega t_0\right), \quad (29)$$

where ω is the angular frequency; $\omega_c = 2\pi f_0$ is the angular frequency corresponding to the maximum amplitude. The amplitude spectrum is

$$A(\omega) = \frac{\sqrt{\pi}\omega^2}{\omega_c^3} \exp\left(-\frac{\omega^2}{\omega_c^2}\right). \quad (30)$$

To determine the frequency width of the Ricker wavelet, we take the first derivative of (29) with respect to ω , and after denoting the maximum amplitude of the Ricker wavelet by $A(\omega_c)$, we obtain

$$A(\omega) = \frac{1}{2}A(\omega_c). \quad (31)$$

Substituting (31) into (30), we have

$$\frac{\omega^2}{\omega_c^2} \exp\left(-\frac{\omega^2}{\omega_c^2}\right) = \frac{1}{2}e^{-1}, \quad (32)$$

The solution of (32) is equivalent to the root of the following equation:

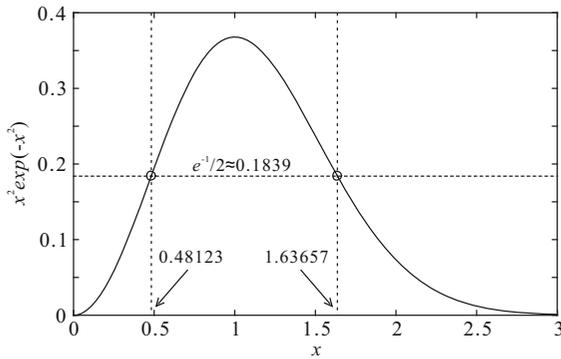


Figure 21

Lambert function (solid line). The two circles indicate the points intersected by the horizontal dashed line that marks the value of $e^{-1}/2$, which defines the frequency bandwidth of the Ricker wavelet

$$x^2 \exp(-x^2) = \frac{1}{2} e^{-1}, \quad (33)$$

where $x = \omega/\omega_c$ and 'e' is Euler's number. The solution of (33) leads to the well-known Lambert W function (Lambert, 1758), and with the help of the graphical method (see Fig. 21), the two roots of (33) are

$$\begin{cases} \omega_l/\omega_c = c_l \\ \omega_h/\omega_c = c_h \end{cases}, \quad (34)$$

where $c_l = 0.482$ and $c_h = 1.637$ are two constants. These two limit frequencies define the frequency width of the Ricker wavelet.

Appendix 2

Frequency Band Selection Strategy

According to multiscale strategy, the seismic data and wavelets are decomposed by scale. To proceed to an adequate frequency-band selection, so that the frequency or wavelength information contained in adjacent frequency bands is less redundant, we set the equality

$$A_{i-1}(c_l f_0^i) = A_i(c_l f_0^i), \quad (35)$$

where c_l is the constant given in Appendix 1; f_0^i is the dominant frequency of the target Ricker wavelet within the higher frequency bands; A_{i-1} and A_i are the amplitude spectra of the target Ricker wavelet

within the lower and higher frequency bands, respectively. The subscript or superscript i denotes the frequency band index. Substituting (30) into (35), we obtain the following equation:

$$\left(\frac{\omega_c^i}{\omega_c^{i-1}}\right)^3 \exp\left(-c_l^2 \left(\frac{\omega_c^i}{\omega_c^{i-1}}\right)^2\right) = \exp(-c_l^2), \quad (36)$$

where ω_c^{i-1} and ω_c^i are the most energetic angular frequencies corresponding to the lower and higher frequency bands, respectively. The solution of (36) is equivalent to the root of the following equation:

$$x^3 \exp(-c_l^2 x^2) = \exp(-c_l^2), \quad (37)$$

where $x = \frac{\omega_c^i}{\omega_c^{i-1}}$. Also with the help of the graphical method (see Fig. 22), the two roots of (37) and, therefore, of (36) are

$$\omega_c^i/\omega_c^{i-1} = c_{1,2}, \quad (38)$$

where $c_1 = 1$ and $c_2 = 4.533$. Obviously, c_2 is the desired solution, so that

$$f_0^{i-1} = f_0^i/4.533. \quad (39)$$

Here, f_0^{i-1} is the dominant frequency within the lower frequency band and f_0^i is the dominant frequency within the higher frequency band. This relationship expresses the recursive relation between dominant frequencies of the target wavelet within

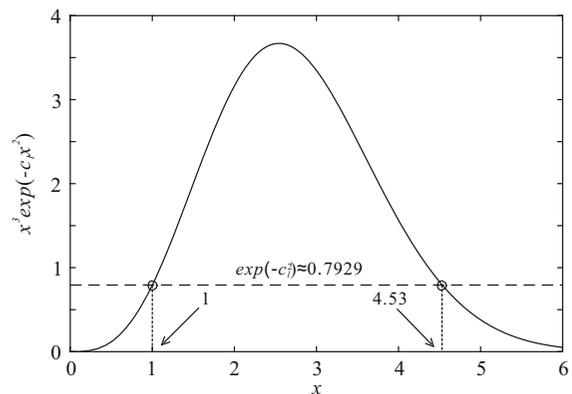


Figure 22

$x^3 \exp(-c_l^2 x^2)$ function (solid line). Here, the value c_l corresponds to the lower frequency point of the two points that define the frequency bandwidth of the Ricker wavelet (see Fig. 21). The two circles indicate the points intersected by the horizontal dashed line that marks the value of $\exp(-c_l^2)$, which are the two roots of the equation $x^3 \exp(-c_l^2 x^2) = \exp(-c_l^2)$

adjacent frequency bands, thus allowing the correct frequency selection to implement the multiscale strategy.

REFERENCES

- Boonyasirawat, C., Valasek, P., Routh, P., Cao, W., Schuster, G. T., & Macy, B. (2009). An efficient multiscale method for time-domain waveform tomography. *Geophysics*, *74*, WCC59–WCC68.
- Brenders, A., Pratt, R., Charles, S. (2009). Waveform tomography of 2-D seismic data in the Canadian foothills—data preconditioning by exponential time-damping. In *71st annual international conference and exhibition, EAGE, extended abstracts U041*.
- Brossier, R., Operto, S., & Virieux, J. (2009). Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *Geophysics*, *74*, WCC105–WCC118.
- Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, *6*, 7690.
- Bunks, C., Salek, F. M., Zaleski, S., & Chavent, G. (1995). Multiscale seismic waveform inversion. *Geophysics*, *60*, 1457–1473.
- Dai, Y. H., & Yuan, Y. (1999). A nonlinear conjugate gradient method with a strong global convergence property. *SIAM Journal on Optimization*, *10*, 177–182.
- Davis, T. A., & Duff, I. S. (1997). An unsymmetric pattern multifrontal method for sparse LU factorization. *SIAM Journal on Matrix Analysis and Applications*, *18*, 140–158.
- Fichtner, A., Trampert, J., Cupillard, P., Saygin, E., Taymaz, T., Capdeville, Y., et al. (2013). Multiscale full waveform inversion. *Geophysical Journal International*, *194*, 534–556.
- Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, *13*, 317322.
- Fletcher, R. (1987). *Practical methods of optimization vol. 1: Unconstrained optimization*. New York: Wiley.
- Fletcher, R., & Reeves, C. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*, 149–154.
- Goldfarb, D. (1970). A family of variable metric updates derived by variational means. *Mathematics of Computation*, *24*, 2326.
- Guittou, A., Ayeni, G., & Díaz, E. (2012). Constrained full-waveform inversion by model reparameterization. *Geophysics*, *77*, R117–R127.
- Guittou, A., & Symes, W. W. (2003). Robust inversion of seismic data using the Huber norm. *Geophysics*, *68*, 1310–1319.
- Hager, W. W., & Zhang, H. C. (2005). A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on Optimization*, *16*, 170–192.
- Hager, W. W., & Zhang, H. C. (2006). A survey of nonlinear conjugate gradient methods. *Pacific Journal of Optimization*, *2*, 335–358.
- Hestenes, M. R., & Stiefel, E. L. (1952). Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards, Section, 5*, 409–436.
- Hu, W., Abubakar, A., & Habashy, T. M. (2009). Simultaneous multifrequency inversion of full-waveform seismic data. *Geophysics*, *74*, R1–R14.
- Hu, P., & Wang, Z. (2014). A non-monotone line search combination technique for unconstrained optimization. *The Open Electrical and Electronic Engineering Journal*, *8*, 218–221.
- Kvoren, Z., Mosegaard, K., Landa, E., Thore, P., & Tarantola, A. (1991). Monte Carlo estimation and resolution analysis of seismic background velocities. *Journal of Geophysical Research*, *96*, 20289–20299.
- Lambert, J. H. (1758). *Observationes variae in Mathesin Puram. Acta Helvetica, physico mathematico anatomico botanico medica*, *3*, 128–168.
- Liao, Q., & McMechan, G. A. (1996). Multifrequency viscoacoustic modeling and inversion. *Geophysics*, *61*, 1371–1378.
- Liu, Q., & Gu, Y. J. (2012). Seismic imaging: From classical to adjoint tomography. *Tectonophysics*, *566–567*, 31–66.
- Liu, Y., Liu, S., Zhang, M., & Ma, D. (2012). An improved perfectly matched layer absorbing boundary condition for second order elastic wave equation (in Chinese). *Progress in Geophysics*, *27*, 2113–2122.
- Liu, Y., & Storey, C. (1991). Efficient generalized conjugate gradient algorithms, part 1: Theory. *Journal of Optimization Theory and Applications*, *69*, 129–137.
- Liu, Q., & Tromp, J. (2006). Finite-frequency kernels based on adjoint methods. *Bulletin of the Seismological Society of America*, *96*, 2383–2397.
- Liu, Q., & Tromp, J. (2008). Finite-frequency sensitivity kernels for global seismic wave propagation based upon adjoint methods. *Geophysical Journal International*, *174*, 265–286.
- Liu, H., Wang, H., & Ni, Q. (2014). On Hager and Zhang's conjugate gradient method with guaranteed descent. *Applied Mathematics and Computation*, *236*, 400–407.
- Malinowski, M., & Operto, S. (2008). Quantitative imaging of the Permo-Mesozoic complex and its basement by frequency domain waveform tomography of wide-aperture seismic data from the Polish basin. *Geophysical Prospecting*, *56*, 805–825.
- Métivier, L., Breteau, F., Brossier, R., Virieux, J., & Operto, S. (2014). Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures. *Geophysical Prospecting*, *62*, 123.
- Mora, P. (1987). Nonlinear two-dimensional elastic inversion of multioffset seismic data. *Geophysics*, *54*, 1211–1228.
- Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*, *100*, 12431–12447.
- Mulder, W., & Plessix, R. E. (2008). Exploring some issues in acoustic full waveform inversion. *Geophysical Prospecting*, *56*, 827–841.
- Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, *95*, 339–353.
- Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. Berlin: Springer.
- Pageot, D., Operto, S., Vallée, M., Brossier, R., & Virieux, J. (2013). A parametric analysis of two-dimensional elastic full waveform inversion of teleseismic data for lithospheric imaging. *Geophysical Journal International*, *193*, 1479–1505.
- Pan, W., Innanen, K. A., Margrave, G. F., Fehler, M. C., Fang, X., & Li, J. (2016). Estimation of elastic constants for HTI media using Gauss-Newton and full-Newton multiparameter full-waveform inversion. *Geophysics*, *81*, R275–R291.
- Pan, W., Innanen, K. A., Liao, W. (2017). Accelerating Hessian-free Gauss-Newton full-waveform inversion via l-BFGS preconditioned conjugate-gradient algorithm. *Geophysics*, *82*, R49–R64.
- Pica, A., Diet, J. P., & Tarantola, A. (1990). Nonlinear inversion of seismic reflection data in a laterally invariant medium. *Geophysics*, *55*, 284–292.

- Plessix, R. É. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, *167*, 495–503.
- Plessix, R. É., & Li, Y. (2013). Waveform acoustic impedance inversion with spectral shaping. *Geophysical Journal International*, *195*, 301–314.
- Polak, E., & Ribière, G. (1969). Note sur la convergence de directions conjuguées. *Revue Francaise Informat Recherche Opertionelle*, *3e Année*, *16*, 35–43.
- Polyak, B. T. (1969). The conjugate gradient method in extreme problems. *USSR Computational Mathematics and Mathematical Physics*, *9*, 94–112.
- Pratt, R. G. (1990). Frequency-domain elastic wave modeling by finite differences: A tool for cross-hole seismic imaging. *Geophysics*, *55*, 626–632.
- Pratt, R. G. (1999). Seismic waveform inversion in the frequency domain, part I: Theory and verification in a physical scale model. *Geophysics*, *64*, 888–901.
- Pratt, R. G., Shin, C., & Hicks, G. (1998). Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International*, *13*, 341–362.
- Pratt, R. G., & Worthington, M. H. (1990). Inverse theory applied to multisource cross-hole tomography: Part I. Acoustic wave-equation method. *Geophysical Prospecting*, *38*, 287–310.
- Ravaut, C., Operto, S., Imbrota, L., Virieux, J., Herrero, A., & dell'Aversana, P. (2004). Multiscale imaging of complex structures from multifold wide-aperture seismic data by frequency-domain full-wavefield inversions: Application to a thrust belt. *Geophysical Journal International*, *159*, 1032–1056.
- Rothman, D. (1985). Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics*, *50*, 2784–2796.
- Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, *24*, 647656.
- Sheng, J., Leeds, A., Buddensiek, M., & Schuster, G. T. (2006). Early arrival waveform tomography on near-surface refraction data. *Geophysics*, *71*, U47–U57.
- Shipp, R., & Singh, S. C. (2002). Two-dimensional full wavefield inversion of wide-aperture marine seismic streamer data. *Geophysical Journal International*, *151*, 325–344.
- Sirgue, L., & Pratt, R. G. (2004). Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies. *Geophysics*, *69*, 231–248.
- Song, Z. M., Williamson, P. R., & Pratt, R. G. (1995). Frequency-domain acoustic-wave modeling and inversion of crosshole data: Part II—Inversion method, synthetic experiments, and real-data results. *Geophysics*, *60*, 796–809.
- Symes, W. (2008). Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, *56*, 765–790.
- Tape, C., Liu, Q., & Tromp, J. (2007). Finite-frequency tomography using adjoint methods—methodology and examples using membrane surface waves. *Geophysical Journal International*, *168*, 1105–1129.
- Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, *49*, 1259–1266.
- Tarantola, A. (1986). A strategy for nonlinear inversion of seismic data. *Geophysics*, *51*, 1893–1903.
- Tarantola, A. (1988). Theoretical background for inversion of seismic waveforms, including elasticity and attenuation. *Pure and Applied Geophysics*, *128*, 365–399.
- ten Kroode, F., Bergler, S., Corsten, C., de Maag, J. W., Strijbos, F., & Tijhof, H. (2013). Broadband seismic data—The importance of low frequencies. *Geophysics*, *78*, WA3–WA14.
- Tromp, J., Tape, C., & Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, *160*, 195–216.
- Vigh, D., & Starr, E. W. (2008). 3D prestack plane-wave, full-waveform inversion. *Geophysics*, *73*, VE135–VE144.
- Vigh, D., Starr, W., & Kapoor, J. (2009). Developing earth models with full waveform inversion. *The Leading Edge*, *28*, 432–435.
- Wang, Y. (2015). The Ricker wavelet and the Lambert W function. *Geophysical Journal International*, *200*, 111–115.
- Wu, S., Wang, Y., Zheng, Y., & Chang, X. (2015). Limited-memory BFGS based least-squares pre-stack Kirchhoff depth migration. *Geophysical Journal International*, *202*, 738–747.
- Zhang, Y., Ratcliffe, A., Roberts, G., & Duan, L. (2014). Amplitude-preserving reverse time migration: From reflectivity to velocity and impedance inversion. *Geophysics*, *79*, S271–S283.
- Zhou, W., Brossier, R., Operto, S., & Virieux, J. (2015). Full waveform inversion of diving and reflected waves for velocity model building with impedance inversion based on scale separation. *Geophysical Journal International*, *202*, 1535–1554.
- Zhou, B., & Greenhalgh, S. A. (2011). Computing the sensitivity kernel for 2.5-D seismic waveform inversion in heterogeneous, anisotropic median. *Pure and Applied Geophysics*, *168*, 1729–1748.