Pure Appl. Geophys. 178 (2021), 411–422

https://doi.org/10.1007/s00024-021-02662-w

Pure and Applied Geophysics

Adaptive Overcomplete Dictionary Learning-Based Sparsity-Promoting Regularization for Full-Waveform Inversion

Hongsun Fu,[1] Yan Zhang,[1] and Xiaolin Li[1]

Abstract—Full-waveform inversion (FWI) is a highly nonlinear and ill-posed inverse problem, which needs proper regularization to produce reliable results. Recently, sparsity and overcompleteness have been successfully applied to seismic data processing. In this study, we propose a novel adaptive sparsity-promoting regularization for FWI which combines the L-BFGS algorithm with an adaptive overcomplete dictionary learning method. The dictionary is learned from many small imaging patches taken from the optimal velocity model that is obtained by previous L-BFGS iterations. Our dictionary learning method tries to exploit the 2D geometric structure of the training patches in a more direct way and is simple to implement. We test our proposed method on a smoothed Marmousi model, a BG Compass model, and a SEG/EAGE salt model. Since total variation (TV) regularization plays an important role in FWI, the inversion results using the TV regularization method are also presented for comparison purposes. From these experiments, we conclude that the proposed method can achieve better performance than the FWI with the TV regularization method.

Keywords: Full-waveform inversion, sparsity-promoting regularization, overcomplete dictionary learning, total variation regularization.

1. Introduction

Full-waveform inversion (FWI) is a popular method for obtaining high-resolution seismic images by estimating the physical parameters (Tarantola 1984; Pratt et al. 1998; Kumar et al. 2019). Recently, the advancement of computational science has enabled the application of FWI to increasingly complex physics and multi-parameter problems, from acoustic wave equations to the more complex anisotropic (visco)elastic wave equations (Matharu and Sacchi 2019; Huang et al. 2020; Oh et al. 2020). FWI is still a challenging data-fitting procedure. It is usually formulated as a nonlinear least squares optimization problem. If the initial model is not sufficiently accurate, it is difficult to obtain a global optimal solution because of cycle-skipping. Most important, FWI is ill-posed and under-determined, which is sensitive to noise in data. One way to address these problems is adding regularization to guide the FWI toward to a reasonable solution.

The basis of a good regularization is the prior expressed by the regularizer, which can be smooth or non-smooth, non-adaptive or adaptive (data-driven) (Aghamiry et al. 2020). Tikhonov regularization with l^2-norm penalty is one of the most well-established methods that has been used widely in recent years. While Tikhonov regularization is known to be simple for use, it tends to produce overly smooth reconstructions and is unable to preserve important model attributes such as faults and other discontinuous structures.

Recently, in the realm of seismic imaging and inversion problems, l^1-norm regularization methods have been proved to be very successful with a multitude of variants. Among them, TV regularization is usually used to preserve sharp interfaces and obtain non-smooth solutions, where edges and discontinuities are reconstructed. Nowadays, several TV regularization methods have been reported to return more realistic discontinuous solutions in FWI (Yong et al. 2018).

However, TV regularization still has limitations in some respects (Chan et al. 2000). Specifically, the l^1-norm penalty of the gradient encourages the recovery of a velocity model in a piecewise constant form, which results in a reconstructed model with patchy or painting-like staircase artifacts. Then, many variants

[1] School of Science, Dalian Maritime University, Dalian, China. E-mail: fuhongsun@dlmu.edu.cn

of TV have appeared in some existing literature, such as asymmetric TV constraint (Esser et al. 2018), total generalized p-variation regularization scheme (Gao and Huang 2019), and high-order TV regularization scheme (She et al. 2019), etc. On the other hand, Loris et al. (2007) regularized the velocity model in the wavelet domain with a l^1-norm penalty, allowing sharp model discontinuities to be superimposed on a smooth known background. Li et al. (2012) proposed a modified Gauss–Newton method to solve the FWI problem using the sparsity-promoting regularization of the velocity perturbations in the curvelet domain, and obtained a solution that preserves the smooth component and accurately recovers both the locations and magnitudes of the spiky perturbations. Based on the ability of curvelets to sparsely represent geophysical models, Fu et al. (2020) proposed a new accelerated proximal gradient algorithm for solving the sparse optimization problem. Xue et al. (2017) introduced a l^1-norm sparse regularization scheme with seislet transform to improve the accuracy and robustness of FWI.

All these approaches greatly improve the quality of the inversion result by regularizing the sparsity of the model parameters (or its perturbation) over a carefully chosen transform domain. The basic idea is that the search for model parameters can be compactly expressed with a sparse set of expansion coefficients over a predefined transform domain/dictionary (Donoho 2006). Actually, restricting the model to a few representation coefficients does not necessarily lead to a geophysically plausible inversion result, since the seismic velocity models usually contain both smooth variations and sharp interfaces (Aghamiry et al. 2020), and it is difficult to accurately account for both of these two important ingredients of model parameter variations using a single specified basis. Hence, dictionary learning methods from a set of training models/images have attracted much attention, such as orthogonal dictionaries (Bao et al. 2013) and overcomplete dictionaries learned by the K-singular value decomposition (K-SVD) algorithm (Aharon et al. 2006). For example, Zhu et al. (2017) and Li and Harris (2018) used orthogonal dictionary learning to improve the robustness and efficiency of FWI. However, in most settings, compared to orthogonal dictionary learning, overcomplete dictionary learning can provide greater flexibility in modeling as well as better robustness to noise (see Lewicki and Sejnowski 2000; Elad 2010; Huang et al. 2019 for details). In recent years, the K-SVD method has been successfully utilized for adaptively learning the overcomplete dictionary in 2D seismic denoising (Chen 2017). In a typical scenario, the traditional patch-based dictionary learning methods convert 2D image/model patches into 1D vectors for further processing, thereby losing the inherent 2D geometric structure of natural images. Here a new overcomplete dictionary learning framework for FWI application is constructed by using the singular value decomposition (SVD) and a patch clustering method, which leads to the improved waveform inversion performance. The proposed dictionary learning algorithm not only incorporates the inherent 2D geometric structure of natural images into the dictionary atoms, but also makes the learning process easier and more direct.

In this paper, we develop a novel sparsity-promoting regularization for FWI called ASRI-FWI method. We combine the advantages of the L-BFGS algorithm and overcomplete dictionary learning-based l^1-norm regularization to guide the inversion process to obtain a satisfactory solution. In brief, our ASRI-FWI method is an iterative reconstruction process, which mainly consists of a conventional FWI with the L-BFGS algorithm and an artifact removal process with the sparse prior implemented by overcomplete dictionary learning. Our dictionary learning algorithm is inspired by ideas from Zeng et al. (2015) and can be understood as a generalized wavelet construction method. To be specific, we first build a special tree structure to partition the set of our velocity patches; the dictionary elements are then determined by the obtained subset partitions in the tree.

The remainder of this paper is organized as follows: in Sect. 2, we present the optimization problem associated with the Helmholtz equation in the frequency domain, and introduce the iterative method for the reconstruction of the velocity model. Section 3 elaborates the design of the proposed ASRI-FWI method in detail. In Sect. 4, the performances of our proposed ASRI-FWI method are verified by

extensive experimental results. Finally, conclusions and possible future extensions are proposed in Sect. 5.

2. Acoustic Full-Waveform Inversion in the Frequency Domain

FWI aims to obtain high-resolution, high-fidelity velocity models of the subsurface from measured wavefield data, which can be formulated as a nonlinear least squares optimization problem

$$\min_{\mathbf{v}}\{J(\mathbf{v}) := \frac{1}{2}\|F(\mathbf{v}) - d\|_2^2\}, \tag{1}$$

where $F(\cdot)$ is the nonlinear forward modeling operator, the observed data $d \in \mathbb{C}^{N_r N_s}$ are acquired from N_r receivers, N_s is the number of sources, the model $\mathbf{v} \in \mathbb{R}^{N_z \times N_x}$ is the acoustic velocity of interest, and N_x and N_z are the number of grid points in the lateral and vertical directions, respectively.

In the space-frequency domain, the operator $F(\mathbf{v})$ can be formally written $RA^{-1}(\mathbf{v})Q$. Here, $A^{-1}(\mathbf{v})$ is a discretized 2D Helmholtz operator $\omega^2/\mathbf{v}^2 + \nabla^2$ with a perfectly matched layer (PML) boundary condition for frequency ω related to the source function Q, and the operator R is a restriction of the solution of the Helmholtz equation to the surface where the data are recorded.

Since the observed wavefield data depend nonlinearly on the velocity parameters, the optimization problem (1) must be performed iteratively. Obviously, we can apply any gradient-like method, such as the conjugate gradient method, the steepest descent method, or quasi-Newton method, to solve this optimization problem. In this work, we use the L-BFGS algorithm (a popular quasi-Newton algorithm) as an important part for the ASRI-FWI method, due to its high precision and low storage requirement. Besides, the FWI problem is ill-posed and requires regularization to stabilize the solution.

3. Proposed Method

3.1. ASRI-FWI Method

Similar to the work of Bao et al. (2018), our FWI method consists of two nested loops.

In the inner loop step, we adopt the L-BFGS algorithm (Nocedal and Wright 2006; Byrd et al. 1995) to solve the optimization problem (1). For given initial model $\mathbf{v}_0$, the velocity parameter is updated as follows

$$\mathbf{v}_{j+1} = \mathbf{v}_j - \tau_j \mathcal{H}_j \mathbf{g}_j, \tag{2}$$

where the step length τ_j is computed by a line search that satisfies the weak Wolfe conditions, the vector $\mathbf{g}_j$ is the gradient of the objective function $J(\mathbf{v})$, and the symmetric and positive definite matrix $\mathcal{H}_j$ denotes an approximation to the inverse Hessian. Note that L-BFGS employs the model and gradient changes from the most recent iterations to estimate the Hessian matrix, which is cheaper on time. In our implementation, the computation of the gradient $\mathbf{g}_j$ is accomplished by using the adjoint-state method (Plessix 2006).

The outer loop step is to obtain an artifact-reduced result by using sparsity-promoting regularization with overcomplete dictionary learning, where the estimated result $\hat{\mathbf{v}}$ from the L-BFGS algorithm is taken as a degradation model. Then, we consider the following constrained optimization problem

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{v} - \hat{\mathbf{v}}\|_F^2, \text{subject to } \mathbf{v} = \mathbf{D}\mathbf{X}, \|\mathbf{x}_i\|_0 < T_0, \ \forall i, \tag{3}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix, $\mathbf{D}$ is referred to as the dictionary matrix, $\mathbf{X}$ denotes the matrix of the corresponding sparse coefficient vectors $\mathbf{x}_i$, $\|\cdot\|_0$ denotes the l^0-norm that counts the nonzero entries of a vector, and T_0 is a constant that controls the number of nonzero entries in $\mathbf{x}_i$. Obviously, the full dictionary learning can be achieved by updating the dictionary $\mathbf{D}$ and then iteratively computing the sparse matrix $\mathbf{X}$.

3.2. Dictionary Learning

A learning-based overcomplete dictionary can usually provide better sparse approximation properties (Beckouche and Ma 2014; Liu et al. 2018). Nowadays, the learned dictionaries from patch-based representation have been widely used for various signal and image processing problems, i.e., seismic data denoising, image deblurring, inpainting, etc. (see Liu et al. 2018; Mairal et al. 2014; Bruckstein et al. 2009, and references therein). In this research, we want to use dictionary learning to construct an overcomplete dictionary.

In FWI scenarios, we first define a patch extraction operator $\mathbf{E}_i : \mathbb{R}^{N_z N_x} \rightarrow \mathbb{R}^{\sqrt{m}\times\sqrt{m}}$, such that $\mathbf{E}_i \mathbf{v} = \mathbf{P}_i \in \mathbb{R}^{\sqrt{m}\times\sqrt{m}}$, $i = 1, 2, \ldots, n$ (total number of patches), where $\mathbf{P}_i$ corresponds to a $\sqrt{m} \times \sqrt{m}$ image patch (2D) in the estimated velocity. In practice, the overlapping image patches of size $\sqrt{m} \times \sqrt{m}$ are extracted with a shift of S spatial grids, and one can employ periodic boundary conditions for mathematical convenience. All the patches are arranged into a set, denoted by $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n\}$, that then serves as a training set. As is shown below, the elements of the dictionary are constructed in the form of patches $\mathbf{D}_k$ $(k = 1, 2, \ldots, K)$, such that $d_k = \mathrm{vec}(\mathbf{D}_k)$ are the columns (atoms) of $\mathbf{D}$. Each dictionary atom $\mathbf{D}_k$ is the linear combination of low-rank approximation of a suitable mean matrix, which is averages of subsets of $\mathbf{P}_i$ subsets with nonlocal similarity. This method lets us incorporate 2D geological features into the dictionary that cannot be simply found by using only the vectorized training patches.

An overcomplete dictionary $\mathbf{D}$ is a matrix of dimension $m \times K (K > m)$, which contains K column vectors or atoms of size m. In fact, an overcomplete dictionary is not unique, and different dictionary learning methods use different algorithms to solve the optimization problems in Eq. (3). K-singular value decomposition (K-SVD) is one of the greatest potential dictionary learning algorithms (Aharon et al. 2006), and it has been successfully used to learn the adaptive sparse dictionary in seismic denoising. However, K-SVD is very time-consuming (Liu et al. 2017). Here, we employ a top-bottom two-dimensional subspace partition (TTSP) algorithm (Zeng et al. 2015; Liu et al. 2018) for obtaining an overcomplete dictionary $\mathbf{D}$. The main steps of the TTSP algorithm are described as follows:

1. **Create the partition tree**. First, we construct a root node that includes all the training patches $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n\}$, and define the two non-symmetrical covariance matrices

$$\begin{aligned} \mathbf{L}_{\mathrm{cov}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}_i - \mathbf{C})(\mathbf{P}_i - \mathbf{C})^T, \\ \mathbf{R}_{\mathrm{cov}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}_i - \mathbf{C})^T(\mathbf{P}_i - \mathbf{C}), \end{aligned} \tag{4}$$

where $\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}_i$. Note that $\mathbf{L}_{\mathrm{cov}}$ and $\mathbf{R}_{\mathrm{cov}}$ have the same eigenvalue. Then, we compute the normalized maximum eigenvectors $\mathbf{u}$ and $\mathbf{w}$ of $\mathbf{L}_{\mathrm{cov}}$ and $\mathbf{R}_{\mathrm{cov}}$,

$$\begin{aligned} \mathbf{u} &= \arg\max_{\|\mathbf{y}\|_2=1} \mathbf{y}^T \mathbf{L}_{cov}\mathbf{y}, \\ \mathbf{w} &= \arg\max_{\|\mathbf{y}\|_2=1} \mathbf{y}^T \mathbf{R}_{cov}\mathbf{y} \end{aligned} \tag{5}$$

representing the main structures of the velocity patches not being captured by the mean matrix $\mathbf{C}$. Using $\mathbf{u}$ and $\mathbf{w}$, we can compute the one-dimensional projection representations of all image patches $s_i = \mathbf{u}^T\mathbf{P}_i\mathbf{w}$, $i = 1, \ldots, n$, and sort these numbers from smallest to largest, denoted by $\{s_{l_1}, s_{l_2}, \ldots, s_{l_n}\}$. As a measure of similarity between the training patches, these numbers are used to partition the set of $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n\}$ into two partial sets. For this purpose, we compute

$$\begin{aligned} \hat{k} = \arg\min_{1\le k\le n-1} \Bigg[&\sum_{i=1}^{k}\left(s_{l_i} - \frac{1}{k}\sum_{j=1}^{k} s_{l_j}\right)^2 \\ &+ \sum_{i=k+1}^{n}\left(s_{l_i} - \frac{1}{n-k}\sum_{j=k+1}^{n} s_{l_j}\right)^2\Bigg]. \end{aligned} \tag{6}$$

Using this $\hat{k}$, the partition $\{\mathbf{P}_{l_1}, \mathbf{I}_{l_2}, \ldots, \mathbf{P}_{l_{\hat{k}}}\}$ and $\{\mathbf{P}_{l_{\hat{k}+1}}, \mathbf{P}_{l_{\hat{k}+2}}, \ldots, \mathbf{P}_{l_n}\}$ are obtained. And the depth of the node is added one simultaneously. Once the number of velocity patches in this child node is less than the row number or column number of the velocity image patches, we will stop further partitioning. Let K be the total number of nodes in a binary tree.

2. **Determine the dictionary**. Now, for each leaf node k of the partition tree, i.e., for each subset of training patches $\{\mathbf{P}_i\}_{i\in\Lambda_k}$, where $\Lambda_k \subset \{1,2,\ldots,n\}$ is the subset of indices of these training patches, the mean matrix is computed by

$$\mathbf{C}_k = \frac{1}{|\Lambda_k|}\sum_{i\in\Lambda_k}\mathbf{P}_i. \tag{7}$$

Then, compute the normalized eigenvectors $\mathbf{u}_k$ and $\mathbf{w}_k$ to the maximal eigenvalue of $\mathbf{C}_k^T\mathbf{C}_k$ and $\mathbf{C}_k\mathbf{C}_k^T$, $k = 1,2,\ldots,K$. Let λ_k be the maximal singular value of the mean matrix $\mathbf{C}_k$, then $\lambda_k\mathbf{u}_k\mathbf{w}_k^T$ is the best rank-one approximation of the mean matrix $\mathbf{C}_k$. Hence, we initialize the first dictionary element

$$\widetilde{\mathbf{D}}_1 = \mathbf{u}_1\mathbf{w}_1^T \tag{8}$$

describing the main structure of the mean matrix $\mathbf{C} = \mathbf{C}_1$. That is, we get the first layer sub-dictionary. Then, for each pair of children nodes with index sets Λ_{2k} and Λ_{2k+1} from the same parent node, we define

$$\widetilde{\mathbf{D}}_k = \lambda_{2k}\mathbf{u}_{2k}\mathbf{w}_{2k}^T - \lambda_{2k+1}\mathbf{u}_{2k+1}\mathbf{w}_{2k+1}^T, \mathbf{D}_k = \frac{\widetilde{\mathbf{D}}_k}{\|\widetilde{\mathbf{D}}_k\|_F}, \tag{9}$$

thereby describing the difference of main structures of $\mathbf{C}_{2k}$ and $\mathbf{C}_{2k+1}$. Once $\mathbf{D}_k$ is calculated for all nodes, the final dictionary $\mathbf{D} = [\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_K]$ is determined, where $\mathbf{d}_k = \mathrm{vec}(\mathbf{D}_k) \in \mathbb{R}^m$. Figure 1 shows the process for dictionary learning by the TTSP algorithm.

By construction, our dictionary learning method can be regarded as a generalized wavelet approach, where the dictionary elements $\mathbf{D}_k$ for $k > 1$ are "wavelet functions" and $\mathbf{D}_1$ is a "scaling function". In the TTSP algorithm, the partition to the subsets of image patches is used to implement the nonlocal similarity prior. It can be seen that the TTSP algorithm is used to quickly top-bottom divide each leaf node into the left child and right child by the best rank-1 approximation of the mean (center) matrix pair. While the traditional K-SVD algorithm is an iterative method, which has the disadvantages of large calculation quantity and low accuracy (Zhou et al. 2014). Nonetheless, the point here is that a noisy result from L-BFGS is used to learn the dictionary. Using the TTSP algorithm, the inversion for stronger noise levels is still not satisfactory. To get around this limitation, the artifact-reduced step is not employed at every L-BFGS iteration, but rather, we perform several iterations for each L-BFGS step.

For given (noisy) training data $\mathbf{P} = [\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_n] \in \mathbb{R}^{m\times n}$ and $\mathbf{D} = [\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_K] \in \mathbb{R}^{m\times K}$, Eq. (3) can be formulated as

$$\min_{\mathbf{X}\in\mathbb{R}^{K\times n}} \|\mathbf{P}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_0, \tag{10}$$

where the columns $\mathbf{p}_i = \mathrm{vec}(\mathbf{P}_i) \in \mathbb{R}^m$ are the vectorized velocity patches, and λ is a regularization parameter. Greedy algorithms such as matching pursuit (MP) (Mallat and Zhang 1993) and orthogonal matching pursuit (OMP) (Donoho et al. 2012) can be used to find sparse representation of Eq. (10). Because of its faster convergence speed in empirical observations, we use OMP to find the sparse coefficient matrix $\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_K]^T \in \mathbb{R}^{K\times n}$, such that the set of velocity patches $\hat{\mathbf{P}} = [\hat{\mathbf{p}}_1,\hat{\mathbf{p}}_2,\ldots,\hat{\mathbf{p}}_n]$ is sparsely represented by $\mathbf{DX}$.

Next, we use $\mathbf{E}_i^T : \mathbb{R}^{\sqrt{m}\times\sqrt{m}} \to \mathbb{R}^{N_z N_x}$ to denote placing patch $\hat{\mathbf{P}}_i$ back into the corresponding position of the reconstructed velocity image. Then, the dictionary learning-based denoising model can be expressed as

$$\tilde{\mathbf{v}} = \sum_{i=1}^{n}\mathbf{E}_i^T\hat{\mathbf{P}}_i / \sum_{i=1}^{n}\mathbf{E}_i^T\mathbf{1}_{\sqrt{m}\times\sqrt{m}}, \tag{11}$$

where the operator / denotes the element-wise division of two vectors, and $\mathbf{1}_{\sqrt{m}\times\sqrt{m}}$ is an all-ones matrix of the same size as $\hat{\mathbf{P}}_i$.

3.3. Implementation Details

Above all, our inversion method is composed of two main parts: the L-BFGS algorithm and sparsity-promoting regularization. We summarize the pseudo-code of our ASRI-FWI method in algorithm 1.

Algorithm 1 ASRI-FWI method

Initialization:

Given initial velocity $\mathbf{v}_0$, parameters λ.

Outer loops: For $l = 1, 2, \cdots, L$

Inter loops: (L-BFGS step)

for $j = 0, 1, \cdots, J-1$

$\mathbf{v}_{j+1} = \mathbf{v}_j - \tau_j \mathcal{H}_j \mathbf{g}_j$

end for

$\hat{\mathbf{v}} = \mathbf{v}_J$

if $\hat{\mathbf{v}} < 0$

$\hat{\mathbf{v}} = 0$

end if

Sparsity-promoting regularization step:

Determine the dictionary D by TTSP algorithm (eq.(4)-(9))

Determine the coefficient matrix $\mathbf{X}$ by OMP algorithm

Reconstruct the velocity model $\tilde{\mathbf{v}}$ by eq.(11)

$\widetilde{\mathbf{v}}_l = \tilde{\mathbf{v}}$

Warm start $\mathbf{v}_0 = \tilde{\mathbf{v}}$

End For

Output inversion result $\widetilde{\mathbf{v}}_L$

The ASRI-FWI algorithm stops if the objective function $J(\mathbf{v})$ decrease is small enough ($\leqslant 1 \times 10^{-4}$), or the relative change of the velocity model is less than 1×10^{-3} between consecutive iterations, that is, $\|\widetilde{\mathbf{v}}_{l+1} - \widetilde{\mathbf{v}}_l\| / \|\widetilde{\mathbf{v}}_l\| < 1 \times 10^{-3}$. In practice, we can also terminate the inverse computation when the maximum number of iterations L of the outer loop reaches a pre-defined value. An appropriate choice of regularization parameter λ which controls the degree of sparseness is very important for all sparsity-promoting regularization. It is not the main mission here to delve into this issue, and we will choose a single value $\lambda = 1 \times 10^{-1}$ manually, which proves sufficient, for all the numerical results obtained by the proposed ASRI-FWI method.

4. Experiment

In this section, we present some inversion results that verify the performance of the proposed ASRI-FWI method. In all the experiments, we create noisy observed data by adding 5% white Gaussian noise, because noise is unavoidable in field seismic data. To make it fair, we handle the FWI with TV regularization (denoted as the TV-FWI method for convenience) using the same acquisition geometry as ASRI-FWI. And the TV-FWI method is also carried out by using the L-BFGS algorithm in the frequency domain. To be specific, the TV-FWI solution can be recovered by solving the following optimization problem

$$\min_{\mathbf{v}} \{J_{TV}(\mathbf{v}) := \frac{1}{2}\|F(\mathbf{v}) - d\|_2^2 + \alpha \|\nabla \mathbf{v}\|_1\}, \quad (12)$$

where α is a regularization parameter, for which values of the order of 10^{-3} work well.

To improve convergence and avoid trapping to local minima, all waveform inversions are performed sequentially in 12 overlapping frequency bands on the interval 2.9–26 Hz. In our computation, the parameters of ASRI-FWI are set as follows: we perform $J = 10$ iterations for each L-BFGS step, the sliding distance S is set to 6, the minimal number of velocity patches in a subset corresponding to one node is set to 16, and the patch size is 8×8, which means that $m = 64$. In the following experiments, the number of iterations of the outer loop is set to $L = 4$, and TV-FWI method runs for 40 iterations, so this helps to ensure a fair comparison between both inversion methods.

All the experiments are conducted in MATLAB 2017a environment on a desktop PC with a quad-core processor at 3.20 GHz and 16 GB of RAM. The peak signal-to-noise ratio (PSNR), structural similarity (SSIM), root-mean-square error (RMSE) and running time duration are used to quantitatively evaluate the performance of the FWI methods.

4.1. Marmousi Model

In this experiment, the Marmousi velocity model in Fig. 2a is used for numerical tests. The velocity model is discretized over a $N_z \times N_x = 260 \times 819$ grid with a spacing of 4 m. We put $N_s = 77$ sources and $N_r = 231$ receivers near the surface, and both sources and receivers are evenly distributed in the horizontal direction. To ensure convergence of the iterative scheme, the starting model (see Fig. 2b) is obtained by smoothing the original velocity model with a

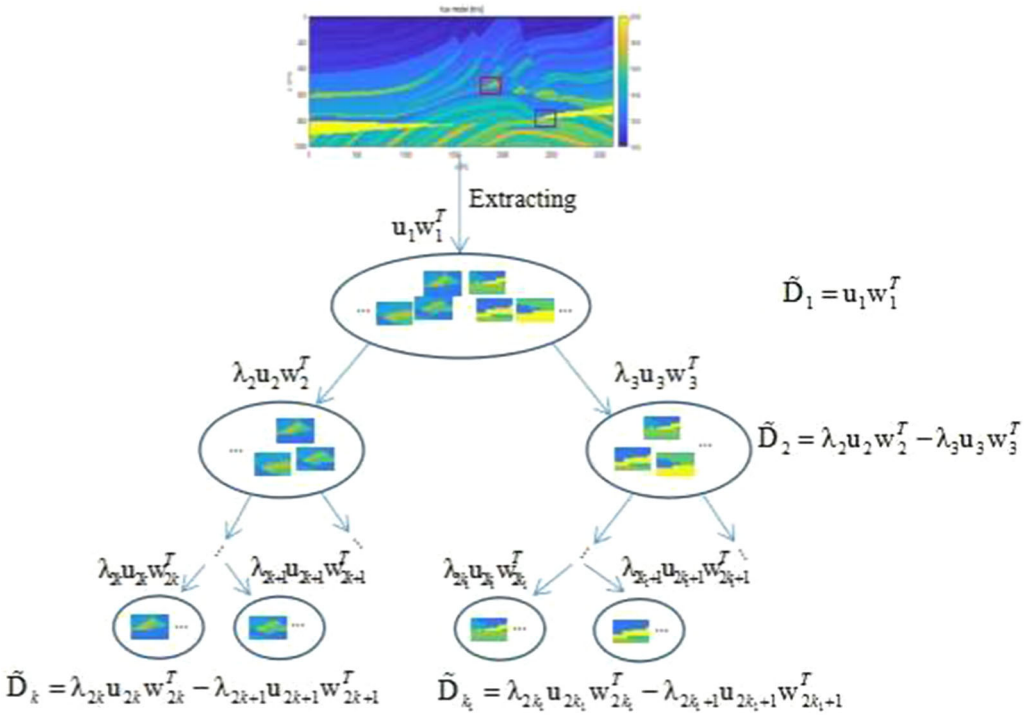


Figure 1
The process for dictionary learning by the TTSP algorithm

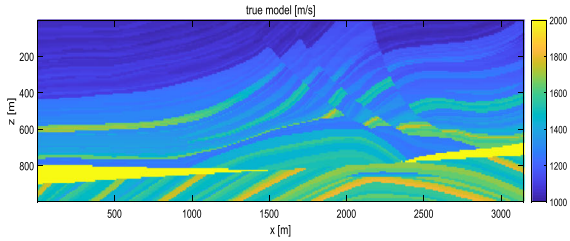
Gaussian kernel of varying widths. The inversion results are shown in Fig. 2c, d. In both cases, the uppermost and the low-velocity region of the Marmousi velocity model are reconstructed reasonably well. However, the velocity magnitudes and edges in the deeper part of the model are significantly improved by making use of the proposed ASRI-FWI method.

Figure 3 shows the comparison of depth velocity profiles (the Marmousi model) of the two inversion results. We also observe that the ASRI-FWI method reconstructs both the smooth variations and sharp interfaces more accurately. Compared to the TV-FWI method, the increase in running time caused by the dictionary learning step is not significant (see Table 1). From Table 1, it is clear to observe that our method achieves the much better performance for all quantitatively evaluate metrics (PSNR, SSIM and RMSE).

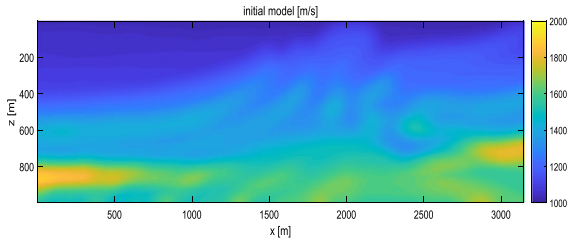
4.2. BG Compass Model

In the second example, we test our method on a part of the BG Compass synthetic benchmark model. The true velocity is depicted in Fig. 4a. This model is rescaled to $N_z \times N_x = 205 \times 701$ with a spacing of 10 m. We place 64 sources and 192 receivers, all regularly spaced along the top of the velocity model. Figure 4b shows the initial model without any lateral information, where the velocity magnitudes increase linearly with depth. The results for the TV-FWI method and our method are shown in Fig. 4c, d, respectively. A comparison of vertical velocity profiles at different locations on the horizon is given in Fig. 5. The quantitative evaluation of the BG Compass model is given in Table 1.

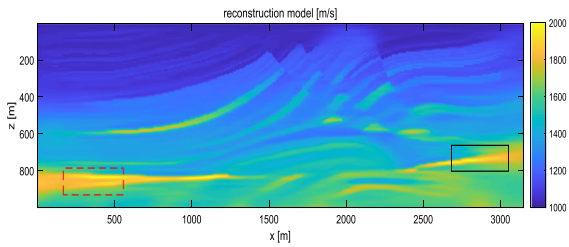
We make the following observations from these visual effects and performance metrics. First, compared with the TV-FWI method, our proposed method not only improves the clarity of the inversion image, but also preserves more detail information. Second, consistent with the visual effects, our



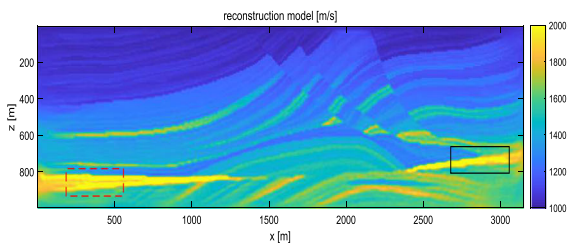
(a) The true velocity model



(b) The initial velocity model



(c) Inversion result using TV regularization



(d) Inversion result using ASRI method

Figure 2
Inversion results of the Marmousi model

approach has the best test scores for all metrics, and the improvements are very satisfying, but the computing time is almost the same as the TV-FWI method.

4.3. SEG/EAGE Salt Model

We finally apply our method to the 2D SEG/EAGE salt model. The dimensions of the model are

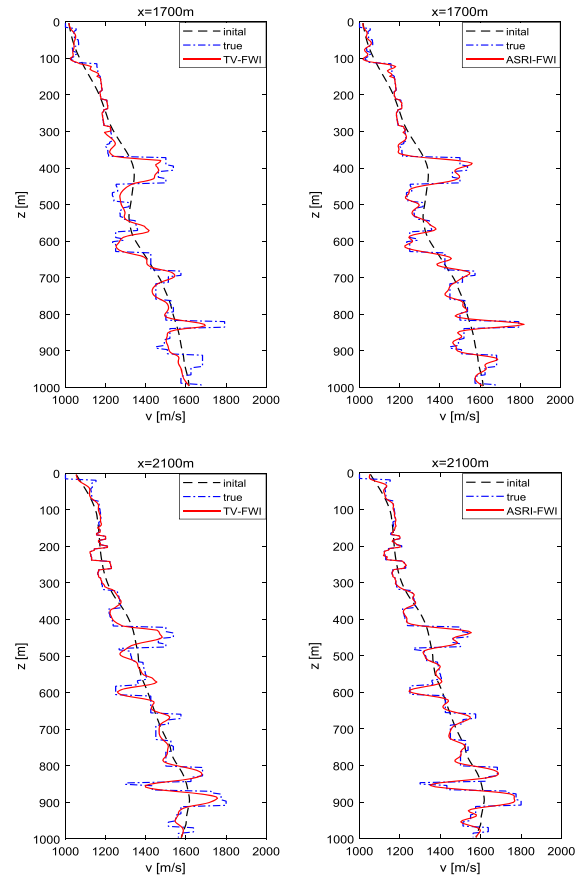


Figure 3
The vertical velocity profiles extracted from the Marmousi model

640 m × 1280 m, with grid spacings of 5 m, which is shown in Fig. 6a. The model is composed of an isolated salt dome and several faults. The lower and upper bounds of velocity are 1500 m/s and 4500 m/s, respectively. The initial model is shown in Fig. 6b. The data is generated for 31 equispaced sources and 93 receivers on the surface.

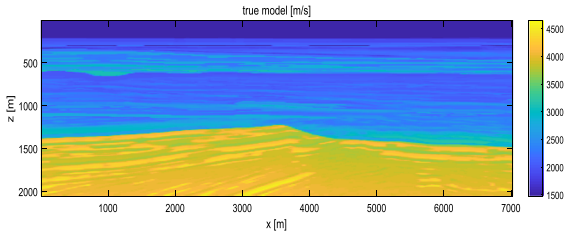
In general, the salt structure is one of the most challenging objects to recover, particularly when given a poor initial estimate. Although the initial estimate is rather poor, the reconstruction deteriorates only slightly compared with that of the two previous examples. If we examine the inversion results (Fig. 6c, d) carefully, at certain places, such as the area represented by the rectangular box, the TV-FWI results have spurious thin layers, while our method maintains the edges better. The comparison of the vertical velocity profiles of the two inversion results

Table 1

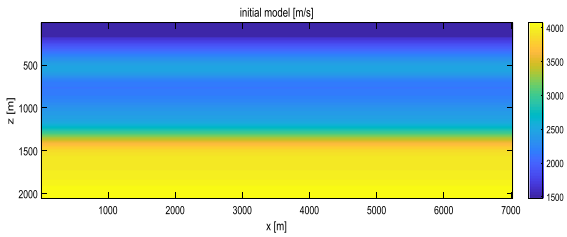
Quantitative performance results of the algorithms applied to various models

Models	Methods	Initialization			Final results			
		PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	Run times(s)
Marmousi model	TV	19.26	0.63	0.11	21.73	0.75	0.08	**4802.80**
	ASRI				**23.90**	**0.80**	**0.06**	4980.20
BG model	TV	12.73	0.26	0.23	15.46	0.55	0.17	**3069.68**
	ASRI				**19.14**	**0.69**	**0.11**	3103.63
SEG/EAGE salt model	TV	7.53	0.46	0.42	11.61	0.51	0.28	**203.21**
	ASRI				**12.02**	**0.54**	**0.25**	215.34

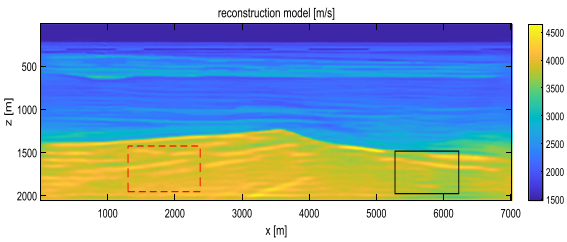
The bold values represent the best value among the comparison results



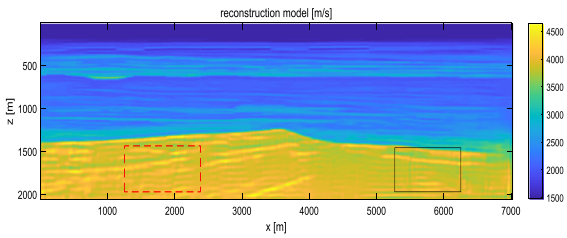
(a) The true velocity model



(b) The initial velocity model



(c) Inversion result using TV regularization



(d) Inversion result using ASRI mathod

Figure 4
Inversion results of the BG Compass model

is shown in Fig. 7. The quantitative results are shown in Table 1. In addition, we would like to mention that the proposed method cannot promise to overcome the local minimum problem, but it can be largely alleviated. Due to the presence of the salt body, the FWI results tend to have undesirable artifacts even

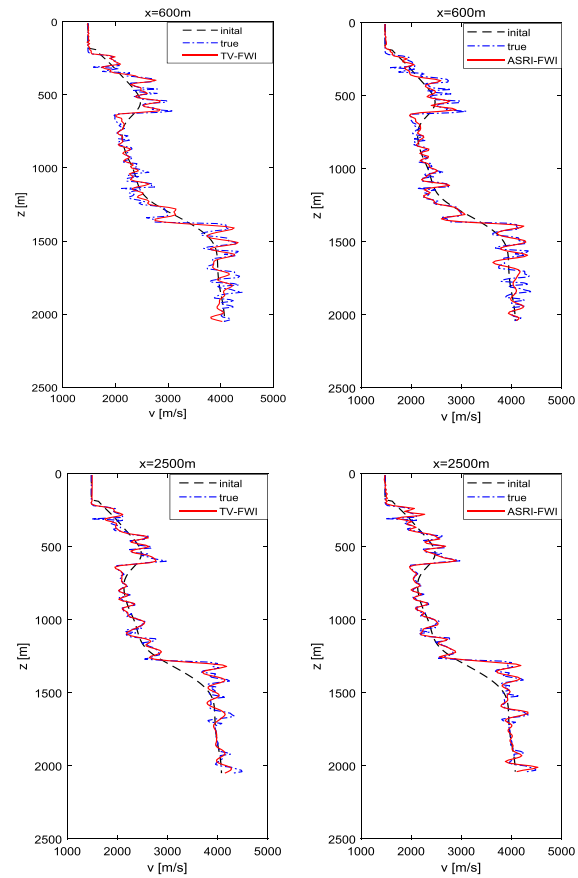
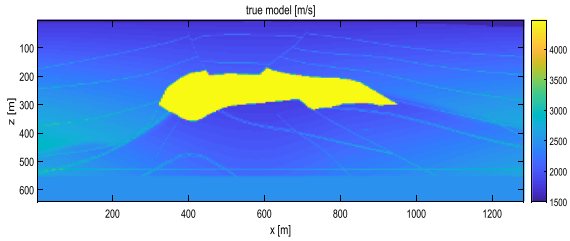
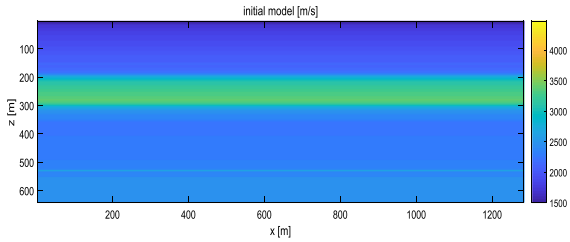


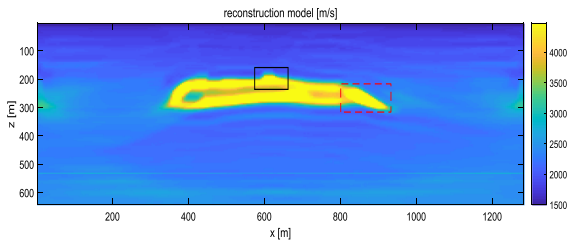
Figure 5
The vertical velocity profiles extracted from the BG Compass model



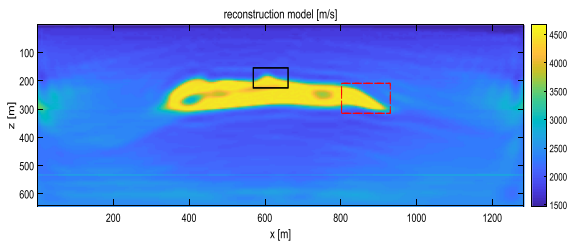
(a) The true velocity model



(b) The initial velocity model



(c) Inversion result using TV regularization



(d) Inversion result using ASRI mathod

Figure 6
Inversion results of the SEG/EAGE salt model

when the initial velocity is good. In particular, the deeper part of the estimated model tends to have incorrect results. As a regularization technique, our proposed approach can be used together with other FWI strategies to further improve the accuracy and the convergence speed.

Further, in order to investigate the sensitivity of the patch size ($\sqrt{m} \times \sqrt{m}$) and sliding distance S, multiple experiments are performed. The 2D SEG/

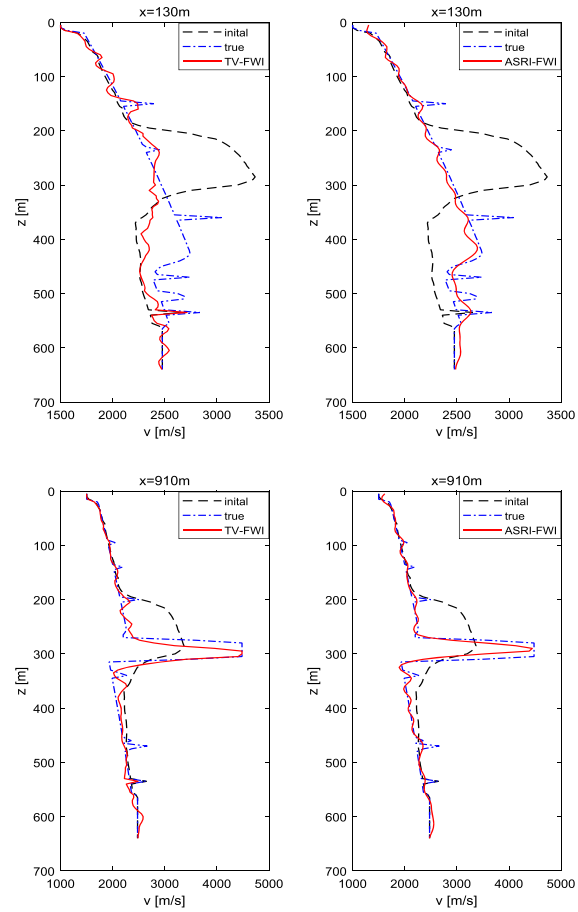


Figure 7
The vertical velocity graphs extracted from the SEG/EAGE salt model

EAGE salt model is selected as the test model. We note that the inversion results are less sensitive to these parameters as long as the learned dictionary is overcomplete.

5. *Conclusion*

Despite the rapid development of waveform inversion technique, the ill-posedness inherent of FWI is still a major problem in this field. In this paper, we propose a novel adaptive dictionary learning-based sparsity-promoting regularization for FWI called the ASRI-FWI method. In the inversion procedure, we use the estimated result (as training patches) from L-BFGS iterations to teach the dictionary and employ the learned dictionary to guide

the inversion in the current iteration. Compared with the traditional TV regularization method, our method effectively reduces the degrees of freedom in velocity parameters to be inverted and eliminates undesirable artifacts and preserves significant details and structural information of model parameters. The introduction of the overcomplete dictionary learning process may also alleviate the problem of local minima in FWI to some extent, but not completely. Yet another important problem is the computational cost. The computational cost of our ASRI-FWI method is higher than the traditional TV-FWI method, mainly because of the nested loop steps and the extra operations related to adaptive dictionary learning: extraction of patch, construction of the partition tree, determination of the dictionary from the partition tree, and sparse approximation calculation. Further work is required to study a more efficient choice of optimal parameters, using parallel computing techniques to accelerate the computation, and the most important is to test this approach for the elastic multi-parameter FWI problems, 3D FWI, and real geophysical inversion problems.

Acknowledgements

The authors thank the editor and the anonymous referees for their valuable comments, suggestions, and support. This work is partially supported by the National Natural Science Foundation of China (Grant No. 41474102).

References

Aghamiry, H. S., Gholami, A., & Operto, S. (2020a). Compound regularization of full-waveform inversion for imaging piecewise media. *IEEE Transactions on Geoscience and Remote Sensing,58*(2), 1192–1204.

Aghamiry, H. S., Gholami, A., & Operto, S. (2020). Full waveform inversion by proximal newton method using adaptive regularization. *Geophysical Journal International,224*(1), 169–80.

Aharon, M., Elad, M., & Bruckstein, A. (2006). The K-SVD: An algorithm for designing of ever complete dictionaries for sparse representation. *IEEE Transactions on Signal Processing,54*, 4311–4322.

Bao, C., Cai, J. F., & Ji, H. (2013). Fast sparsity-based orthogonal dictionary learning for image restoration. *Proceedings of the 2013 IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2013.420.

Bao, P., Zhou, J., & Zhang, Y. (2018). Few-view CT reconstruction with group-sparsity regularization: GSR-SART. *International Journal for Numerical Methods in Biomedical Engineering*. https://doi.org/10.1002/cnm.3101.

Beckouche, S., & Ma, J. (2014). Simultaneous dictionary learning and denoising for seismic data. *Geophysics,79*(3), A27–A31.

Bruckstein, A. M., Donoho, D. L., & Elad, M. (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review,51*(1), 34–81.

Byrd, R. H., Lu, P., & Nocedal, J. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing,16*(5), 1190–1208.

Chan, T., Marquina, A., & Mulet, P. (2000). High-order total variation-based image restoration. *SIAM Journal on Imaging Sciences,22*(2), 503–516.

Chen, Y. (2017). Fast dictionary learning for noise attenuation of multidimensional seismic data. *Geophysical Journal International,209*(1), 21–31.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics,59*, 797–829.

Donoho, D. L., Tsaig, Y., Drori, I., & Starck, J. (2012). Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Transactions on Information Theory,58*(2), 1094–1121.

Elad, M. (2010). *Sparse and redundant representations: From theory to applications in signal and image processing*. New York: Springer.

Esser, E., Guasch, L., Leeuwen, T. V., Aravkin, A. Y., & Herrmann, F. J. (2018). Total variation regularization strategies in full-waveform inversion. *SIAM Journal on Imaging Sciences,11*(1), 376–406.

Fu, H. S., Ma, M. Y., & Han, B. (2020). An accelerated proximal gradient algorithm for source-independent waveform inversion. *Journal of Applied Geophysics,177*, 104030.

Gao, K., & Huang, L. (2019). Acoustic-and elastic-waveform inversion with total generalized p-variation regularization. *Geophysical Journal International,218*(2), 933–957.

Huang, X., Liu, Y., & Wang, F. (2019). A robust full waveform inversion using dictionary learning. *Seg Technical Program Expanded Abstracts*. https://doi.org/10.1190/segam2019-3215989.1.

Huang, X., Eikrem, K. S., Jakobsen, M., & Nvdal, G. (2020). Bayesian full-waveform inversion in anisotropic elastic media using the iterated extended Kalman filter. *Geophysics,85*(4), C125–C139.

Kumar, R., Willemsen, B., Herrmann, F. J., et al. (2019). Enabling numerically exact local solver for waveform inversion low-rank approach. *Computing Geosciences, 23*, 829–847.

Lewicki, M. S., & Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation,12*(2), 337–365.

Li, D., & Harris, J. M. (2018). Full waveform inversion with nonlocal similarity and model-derivative domain adaptive

sparsity-promoting regularization. *Geophysical Journal International,34*(4), 1841–1864.

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

Liu, L., Plonka, G., & Ma, J. (2017). Seismic data interpolation and denoising by learning a tensor tight frame. *Inverse Problems,33*(10), 105011.

Liu, L., Ma, J., & Plonka, G. (2018). Sparse graph-regularized dictionary learning for suppressing random seismic noise. *Geophysics,83*(3), V215–V231.

Loris, I., Nolet, G., Daubechies, I., & Dahlen, F. A. (2007). Tomographic inversion using L1-norm regularization of wavelet coefficients. *Geophysical Journal International,170*(1), 359–370.

Mairal, J., Bach, F., & Ponce, J. (2014). Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision,8*(2), 85–283.

Mallat, S., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transaction Signal Processing,41*(12), 3397–3415.

Matharu, G., & Sacchi, M. (2019). A subsampled truncated-Newton method for multiparameter full-waveform inversion. *Geophysics,84*(3), R333–R340.

Nocedal, J., & Wright, S. (2006). *Numerical optimization* (2nd ed.). Berlin: Springer Science & Business Media.

Oh, J. W., Shin, Y., Alkhalifah, T., & Min, D. J. (2020). Multistage elastic full-waveform inversion for tilted transverse isotropic media. *Geophysical Journal International,223*(1), 57–76.

Plessix, R. E. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International,167*(2), 495–503.

Pratt, R. G., Shin, C., & Hick, G. J. (1998). Gauss–Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International,133*(2), 341–362.

She, B., Wang, Y., Zhang, J., Wang, J., & Hu, G. (2019). AVO inversion with high-order total variation regularization. *Journal of Applied Geophysics,161*, 167–181.

Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics,49*(8), 1259–1266.

Xue, Z., Zhu, H., & Fomel, S. (2017). Full-waveform inversion using seislet regularization. *Geophysics,82*(5), A43–A49.

Yong, P., Liao, W., Huang, J., & Li, Z. (2018). Total variation regularization for seismic waveform inversion using an adaptive primal dual hybrid gradient method. *Inverse Problems,34*(4), 045006.

Zeng, X., Bian, W., Liu, W., Shen, J., & Tao, D. (2015). Dictionary pair learning on Grassmann manifolds for image denoising. *IEEE Transactions on Image Processing,24*(11), 4556–4569.

Zhou, Y., Zhao, H. M., Shang, L., & Liu, T. (2014). Immune K-SVD algorithm for dictionary learning in speech denoising. *Neurocomputing,137*, 223–233.

Zhu, L., Liu, E., & Mcclellan, J. H. (2017). Sparse-promoting full-waveform inversion based on online orthonormal dictionary learning. *Geophysics, 82*(2), R87–R107.

(Received July 20, 2020, revised December 15, 2020, accepted January 13, 2021, Published online January 25, 2021)