Full Waveform Inversion by Proximal Newton Method using Adaptive Regularization

H Aghamiry, A Gholami, S Operto

Geophys. J. Int. (XXX) **XXX**, 1–**??**

Full Waveform Inversion by Proximal Newton Method using Adaptive Regularization

H. S. Aghamiry.[1], A. Gholami[2], S. Operto[1]

[1] *Université Côte d'Azur, CNRS, Observatoire de la Côte d'Azur, IRD , Géoazur, Valbonne, France.*
E-mail: aghamiry@geoazur.unice.fr, operto@geoazur.unice.fr
[2] *University of Tehran, Institute of Geophysics, Tehran, Iran.*
E-mail: agholami@ut.ac.ir

Received XXX; in original form XXX

SUMMARY
Regularization is necessary for solving nonlinear ill-posed inverse problems arising in different fields of geosciences. The base of a suitable regularization is the prior expressed by the regularizer, which can be non-adaptive or adaptive (data-driven), smooth or non smooth, variational-based or not. Nevertheless, tailoring a suitable and easy-to-implement prior for describing geophysical models is a nontrivial task. In this paper, we propose two generic optimization algorithms to implement arbitrary regularization in nonlinear inverse problems such as full-waveform inversion (FWI), where the regularization task is recast as a denoising problem. We assess these optimization algorithms with the plug-and-play block-matching BM3D regularization algorithm, which determines *empirical* priors adaptively without any optimization formulation. The nonlinear inverse problem is solved with a proximal Newton method, which generalizes the traditional Newton step in such a way to involve the gradients/subgradients of a (possibly non-differentiable) regularization function through operator splitting and proximal mappings. Furthermore, it requires to account for the Hessian matrix in the regularized least-squares optimization problem. We propose two different splitting algorithms for this task. In the first, we compute the Newton search direction with an iterative method based upon the first-order generalized iterative shrinkage-thresholding algorithm (ISTA), and hence Newton-ISTA (NISTA). The iterations require only Hessian-vector products to compute the gradient step of the quadratic approximation of the nonlinear objective function. The second relies on the alternating direction method of multipliers (ADMM), and hence Newton-ADMM (NADMM), where the least-squares optimization subproblem and the regularization subproblem in the composite objective function are decoupled through auxiliary variable and solved in an alternating mode. The least-squares subproblem can be solved with exact, inexact, or quasi-Newton methods. We compare NISTA and NADMM numerically by solving FWI with BM3D regularization. The tests show promising results obtained by both algorithms. However, NADMM shows a faster convergence rate than Newton-ISTA when using L-BFGS to solve the Newton system.

Key words: Inverse theory; Waveform inversion; Controlled source seismology; Numerical modelling.

1 INTRODUCTION

Non-linear inverse problems frequently arise in different fields of geosciences (Tarantola 2005). Large-scale problems are typically solved with iterative local optimization (gradient-based) techniques such as Newton's method. Furthermore, such problems are inherently ill-posed and thus require regularization techniques to be implemented such that assumptions and priors about the unknown models are encoded in the optimization. At the heart of a suitable regularization is a priori information expressed by the regularizer or regularization function (Gholami & Siahkoohi 2010). A proper regularizer can be added to the objective function as penalty terms or constrains (Peters & Herrmann 2017; Peters et al. 2019), potentially renders the solution unique, increases its stability, and prevents data overfitting. It should be able to (mathematically) describe the solution while being easy to implement with iterative linearization methods. These specifications make tailoring a suitable regularizer nontrivial. A prior can be *adaptive* or *non-adaptive*, where by *adaptive* is meant the adaptation of the regularization function to the problem of interest. Traditional priors used to solve inverse problems such as smoothness, sparseness, blockiness are *non-adaptive* (Tikhonov & Arsenin 1977; Tarantola 2005). They are defined ac-

cording to the preliminary assumptions about the targeted model, which are independent of the input of the problem to be solved. In contrast, *adaptive* priors are solely derived from the input and tailored to the output model accordingly. Complex models require complex priors, which can be hard to derive. Different priors lead to different forms of regularization, ranging from smooth and convex single-parameter regularizers (Tikhonov & Arsenin 1977) to non-smooth and non-convex multi-parameter ones (Gholami & Hosseini 2011; Selesnick & Farshchian 2017).

Denoising as the simplest inverse problem (Section 2.1) has contributed to enormous progress in developing sophisticated adaptive and non-adaptive priors for complicated signal recovery from noisy signals (Milanfar 2012). Some recently proposed excellent denoising methods include nonlocal means filters (Milanfar 2012; Goyal et al. 2020) and block matching and 3D filtering (BM3D) (Dabov et al. 2007) and its variants (Goyal et al. 2020). These patch-based methods use both local and nonlocal redundancy of information in the input signal to preserve structures in the solution by yielding locally adaptive filters via similarity kernels. Specifying the kernel function in these methods is essentially equivalent to estimating a particular type of empirical prior from the input signal (Milanfar 2012). This somehow contrasts with the traditional non-adaptive regularization methods, for which the prior is fixed and independent from the input signal (Tarantola 2005). Such an adaptive regularization has been applied to linear inverse problems in, e.g., Danielyan et al. (2011) and Venkatakrishnan et al. (2013). We refer the reader to Appendix A for a more detailed review of the BM3D method that will be used in this study.

The focus of this paper is to present a generic optimization framework to implement arbitrary regularizer in nonlinear inverse problems such as Full Waveform Inversion (FWI) (Tarantola 1984; Pratt et al. 1998; Virieux & Operto 2009). This framework extends proximal Newton-type algorithms to deal with adaptive plug-and-play regularizer, which may not have an explicit optimization (or variational) formulation as BM3D (Dabov et al. 2007). Similar to the classical Newton-type methods, proximal Newton methods solve a nonlinear inverse problem iteratively as $\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \Delta\mathbf{m}_k$, where $\mathbf{m}_k$ is the model parameters at iteration k, $\Delta\mathbf{m}_k$ is the search direction and α_k is the step length. When a composite objective function includes a general (and possibly non-differentiable) regularization term, $\Delta\mathbf{m}_k$ must further involve the gradients/subgradients of the regularization function (Lee et al. 2014). Proximal Newton methods achieve this task by breaking down the original complex problem into simpler subproblems through operator splitting and proximal mappings. We propose two distinct algorithms to solve the regularized problem with proximal Newton methods. In the first, called NISTA, the Newton search direction $\Delta\mathbf{m}_k$ is computed at iteration k by minimizing a composite objective function given by the sum of the locally quadratic approximation of the nonlinear misfit function involving the Hessian and the regularization function. The minimum of this surrogate objective function is found iteratively in an inner loop inside iteration k with a proximal gradient method (Section 2.2) based upon the shrinkage-thresholding algorithm (ISTA) (Daubechies et al. 2004; Attouch & Peypouquet 2016). This method requires one Hessian-vector product per inner iteration to build the gradient of the surrogate linearized misfit function.

The second algorithm, called NADMM, relies on the alternating direction method of multipliers (ADMM) (Boyd et al. 2010; Aghamiry et al. 2019b). ADMM decouples the linearized least-squares objective function and the regularization term via an auxiliary variable and solves the two subproblems in alternating mode with the primal-dual method of multipliers. The first subproblem requires to solve at each iteration k a linear system involving the Hessian, just like classical Newton-type methods. This system can be solved exactly or approximately with inexact or quasi-Newton algorithms (Nocedal & Wright 2006; Métivier et al. 2017).

An important property of the proposed algorithms is that they only need the output of the regularizer without requiring any information about its functional form and statistical properties. This black-box implementation brings flexibility to the algorithms for implementing arbitrary adaptive and non-adaptive regularizations in the nonlinear inverse problem. Therefore, the main properties of the proposed regularization method can be summarized as follow: [1] It can be easily implemented for existing nonlinear optimization algorithms. [2] The regularizer is treated as black-box denoiser, and thus, regularization which has not an explicit optimization formulation such as plug-and-play denoiser can be used. [3] Irrespective of the differentiability of the regularizer, it can be implemented with iterative gradient-based solvers. [4] The computational overhead generated by the regularization is the computation of the proximal/denoising operator at each iteration, and hence is negligible in most cases.

We implement the proposed adaptive regularization to solve full-waveform inversion (FWI), an ill-posed PDE-constrained nonlinear optimization problem, in which the subsurface parameters and the wavefields are defined as the minimizers of the Euclidean distance between observed and calculated data (Tarantola 1984; Pratt et al. 1998). Among different methods to solve this constrained optimization problem, we consider a variable projection formulation leading to the classical FWI (Pratt et al. 1998) and a PDE-relaxation formulation implemented with ADMM (Aghamiry et al. 2019c). The ADMM formulation, which updates the parameters and the wavefields in alternating mode, is referred to iteratively refined wavefield reconstruction inversion (IR-WRI) (Aghamiry et al. 2019c,b). Numerical tests performed show outstanding performance of the adaptive regularization in building complicated velocity models by the above waveform inversion methods.

2 PRELIMINARIES

We start with a brief review of the concepts and formulas of the linear inverse theory used in this paper. This section is intended for background, and the readers who are familiar with linear inverse theory can skim over this section.

In linear inverse problems, the desired model, denoted by the column vector $\mathbf{m}$, needs to be estimated from measurements $\mathbf{d}$ that are related to $\mathbf{m}$ via a linear operator/matrix $\mathbf{A}$, i.e. $\mathbf{d} = \mathbf{A}\mathbf{m} + \mathbf{e}$ for some random noise $\mathbf{e}$. For a Gaussian distributed random noise, the estimation problem usually appears as determination of the minimizer of a suitably defined objective function

$$\arg\min_{\mathbf{m}} \frac{1}{2}\|\mathbf{d} - \mathbf{A}\mathbf{m}\|_2^2 + \lambda\mathcal{R}(\mathbf{m}), \tag{2.1}$$

where $\mathcal{R}$ is a regularizer or regularization function, which somehow prevents data overfitting and λ determines regularization weight. Different forms of $\mathcal{R}$ have been proposed, ranging from smooth and convex single-parameter functions (Tikhonov & Arsenin 1977) to non-smooth and non-convex multi-parameter ones (Gholami & Hosseini 2011; Selesnick & Farshchian 2017). In its simplest form, $\mathcal{R}(\mathbf{m}) = \|\mathbf{m} - \mathbf{m}^{prior}\|_2^2$ is a damping term that encourages $\mathbf{m}$ not to be very far from the prior model $\mathbf{m}^{prior}$ (Tarantola 2005).

2.1 Denoising and Proximal Operator

In denoising problem, $\mathbf{A} = \mathbf{I}$ (the identity matrix) and the estimate is simply defined as

$$\text{prox}_{\lambda\mathcal{R}}(\mathbf{d}) = \arg\min_{\mathbf{m}} \frac{1}{2}\|\mathbf{d}-\mathbf{m}\|_2^2 + \lambda\mathcal{R}(\mathbf{m}). \tag{2.2}$$

This is called the proximal operator of $\mathcal{R}$ (Combettes & Pesquet 2011; Parikh & Boyd 2013). Despite its simple definition, proximal operators are powerful tools in optimization because 1) the general optimization problem (2.1) can be solved by proximal algorithms, which merely require to evaluate the gradient of the misfit function, $\mathcal{M}(\mathbf{m}) = \frac{1}{2}\|\mathbf{d}-\mathbf{A}\mathbf{m}\|_2^2$, and the proximal operator (2.2). 2) Since a proximal operator involves the information about gradients/subgradients of $\mathcal{R}$, proximal algorithms handle both differentiable and nondifferentiable forms of $\mathcal{R}$. This contrasts with traditional algorithms, such as the Newton's algorithm, which requires the objective to be differentiable. Furthermore, the interpretation of the proximal operator as a denoising problem (Kamilov et al. 2017) allows us to solve optimization problem (2.1) with advanced regularizations embedded in sophisticated denoising algorithms such as non-local means (NLM), Block-matching and 3D filtering (BM3D) (Appendix A) or deep learning denoisers (Meinhardt et al. 2017).

2.2 The Proximal Gradient Method

The proximal-gradient method is an important tool for solving non-linear problems we describe in subsequent sections. In order to see how the proximal operator (2.2) helps to solve (2.1), we use the majorization-minimization (MM) approach (Lange 2016), which has a simple convergence proof. The governing idea of MM is to find the minimum of non-convex/convex function $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ via the iterative minimization of a simpler convex surrogate function $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ that majorizes $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ at step k (i.e. $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m}) \geq \mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$). Fig. 1 shows a schematic of the MM process. The non-convex function $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ is shown in blue while a few surrogate functions $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ for points $\mathbf{m}_k$, $k \in \{0,1,2\}$, are shown in orange, pink and red, respectively. This figure shows how the iterative MM algorithm approaches a local minimum of $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ through the minimum of easy to minimize surrogate functions $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$.

It is interesting to note that, for $\mathcal{M}(\mathbf{m}) = \frac{1}{2}\|\mathbf{d}-\mathbf{A}\mathbf{m}\|_2^2$ and $c \in (0, 1/\|\mathbf{A}\|^2)$ with $\|\mathbf{A}\|$ the largest singular value of $\mathbf{A}$, we have that

$$\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m}) \leq \widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m}), \tag{2.3}$$

with equality at $\mathbf{m} = \mathbf{m}_k$, where

$$\widetilde{\mathcal{M}_k}(\mathbf{m}) = \mathcal{M}(\mathbf{m}_k) + (\mathbf{m}-\mathbf{m}_k)^T\nabla\mathcal{M}(\mathbf{m}_k) + \frac{1}{2c}\|\mathbf{m}-\mathbf{m}_k\|_2^2, \tag{2.4}$$

in which $\mathbf{m}_k$ is a reference model (previous iterate) and $\nabla\mathcal{M}(\mathbf{m}_k)$ is the gradient vector. This approximation allows us to minimize (2.1) by iteratively minimizing a simpler problem

$$\mathbf{m}_{k+1} = \arg\min_{\mathbf{m}} \widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m}). \tag{2.5}$$

Simple algebra shows that equation (2.5) is equivalent to

$$\mathbf{m}_{k+1} = \text{prox}_{c\lambda\mathcal{R}}(\mathbf{m}_k - c\nabla\mathcal{M}_k(\mathbf{m}_k)). \tag{2.6}$$

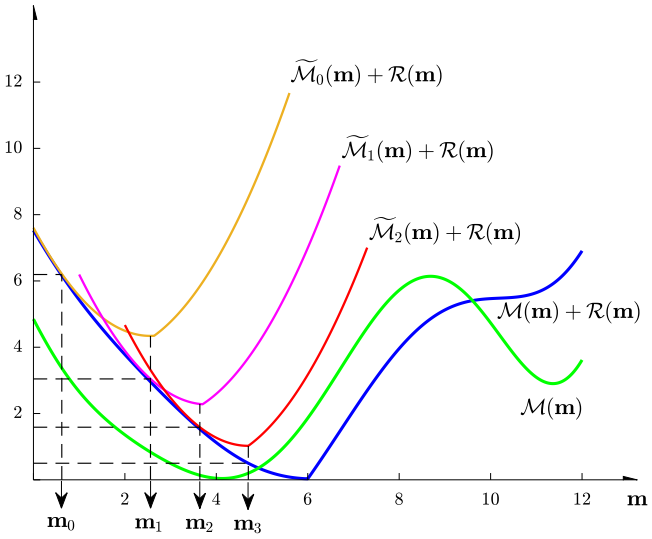

Figure 1. The function $\mathcal{M}(\mathbf{m})$ is shown in green where it has a global minimum at (4,0). Also $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ is shown in blue with a global minimum at (6,0) while a few surrogate functions $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$, $k \in \{0,1,2\}$ are shown in orange, pink and red, respectively. The MM algorithm seeks to find a local minimizer of $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$, blue curve, by iteratively minimizing easy-to-minimize functions $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$. Minimization of $\widetilde{\mathcal{M}_k}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ gives a new optimal point,$\mathbf{m}_1$, $\mathbf{m}_2$ and $\mathbf{m}_3$, and it reaches to the minimum of $\mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m})$ when k tends to infinity.

This is nothing but the famous iterative shrinkage-thresholding algorithm (ISTA) (Daubechies et al. 2004) (also known as forward-backward splitting algorithm and proximal gradient method). FISTA (Beck & Teboulle 2009) is an accelerated version that uses a particular linear combination of the two last iterates to perform the update. A simple acceleration is obtained by using the extrapolation method of Nesterov (Nesterov 1983), leading to the generalized form of ISTA (Attouch & Peypouquet 2016)

$$\begin{cases} \mathbf{m}_{k+1} = \text{prox}_{c\lambda\mathcal{R}}(\mathbf{p}_k - c\nabla\mathcal{M}_k(\mathbf{p}_k)) \\ \mathbf{p}_{k+1} = \mathbf{m}_{k+1} + \frac{k-1}{k+2}(\mathbf{m}_{k+1} - \mathbf{m}_k). \end{cases} \tag{2.7}$$

3 METHOD

A nonlinear inverse problem such as FWI with a general form of regularization can be written as

$$\min_{\mathbf{m}} \mathcal{M}(\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m}), \tag{3.1}$$

where $\mathbf{m}$ gathers the model parameters. In optimization problem (3.1), $\mathcal{M}(\mathbf{m})$ is the data misfit function. Its minimization ensures that the simulated data $F(\mathbf{m})$ are close to the measurements $\mathbf{d}$, where F is a nonlinear differentiable function. $\mathcal{R}(\mathbf{m})$ is the possibly non-differentiable regularization, which encodes the prior knowledge about the model parameters and prevents data overfitting. λ is the trade-off parameter that balances between the data misfit and regularization terms.

A Newton-type method majorizes the misfit term with a local quadratic function of form

$$\widetilde{\mathcal{M}_k}(\mathbf{m}) = \mathcal{M}(\mathbf{m}_k) + (\mathbf{m}-\mathbf{m}_k)^T\nabla\mathcal{M}(\mathbf{m}_k) + \frac{1}{2}(\mathbf{m}-\mathbf{m}_k)^T\mathbf{H}_k(\mathbf{m}-\mathbf{m}_k), \tag{3.2}$$

where $\mathbf{m}_k$ is the iterate at iteration k, $\nabla\mathcal{M}(\mathbf{m}_k)$ is the gradient vector, and $\mathbf{H}_k$ is the Hessian matrix $\nabla^2\mathcal{M}(\mathbf{m}_k)$ or an approximation of it.

Using the approximation in equation (3.2), proximal Newton-type methods solve problem (3.1) iteratively as

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \Delta\mathbf{m}_k, \tag{3.3}$$

where α_k is the step length, which can be determined by a line search method, and

$$\Delta\mathbf{m}_k = \arg\min_{\Delta\mathbf{m}} \widetilde{\mathcal{M}}_k(\mathbf{m}_k + \Delta\mathbf{m}) + \lambda\mathcal{R}(\mathbf{m}_k + \Delta\mathbf{m}) \tag{3.4}$$

is a search direction (Lee et al. 2014). Computation of the search direction $\Delta\mathbf{m}_k$ is the most computationally expensive part of this algorithm because it requires the minimization of a composite function given by the sum of a quadratic term involving the Hessian matrix, equation (3.2), and the regularization term $\mathcal{R}$. For $\lambda = 0$, the algorithm reduces to a classical Newton method, where an approximation of the Hessian can be employed, leading to quasi-Newton methods or gradient method if $\mathbf{H}_k$ reduces to a scaled version of the identity matrix. For $\lambda \neq 0$, however, determination of the search direction in problem (3.4) is more challenging. In the following, we propose two methods for this task.

3.1 Newton-ISTA (NISTA)

NISTA relies on the first-order iterative shrinkage-thresholding algorithm (ISTA). We refer the reader to Daubechies et al. (2004) and Section 2.2 for a more detailed review of this method to estimate iteratively the Newton search direction of problem (3.4). This requires to implement the following inner loop within the outer loop over k

$$\begin{cases} \Delta\mathbf{m}_k^{\ell+\frac{1}{2}} = \Delta\mathbf{p}^{\ell} - c_k[\mathbf{H}_k\Delta\mathbf{p}^{\ell} + \nabla\mathcal{M}(\mathbf{m}_k)], \\ \Delta\mathbf{m}_k^{\ell+1} = \text{prox}_{c_k\lambda\mathcal{R}}(\mathbf{m}_k + \Delta\mathbf{m}_k^{\ell+\frac{1}{2}}) - \mathbf{m}_k, \\ \Delta\mathbf{p}^{\ell+1} = \Delta\mathbf{m}_k^{\ell+1} + \frac{\ell-1}{\ell+2}(\Delta\mathbf{m}_k^{\ell+1} - \Delta\mathbf{m}_k^{\ell}), \end{cases} \tag{3.5}$$

where l is the inner iteration count, $\Delta\mathbf{p}^0 = 0$, and $c_k \in (0, 1/\|\mathbf{H}_k\|^2)$. The term in bracket in the first line of equation (3.5) is the gradient of the surrogate function $\widetilde{\mathcal{M}}_k(\mathbf{m})$, equation (3.2). Also, the operator in the second line of equation (3.5), prox(•), is the proximal operator of the regularization term $\mathcal{R}(\mathbf{m})$, which can be viewed as a denoiser (Section 2.1). For many choices of the regularizer $\mathcal{R}$, there can be a closed-form expression for the denoiser in the second subproblem of (3.5). The main property of this formulation is that it can be generalized to exploit multiple (even data-driven) priors by using different denoisers (e.g., BM3D (Dabov et al. 2007), the weighted nuclear norm (Kamilov et al. 2017)) instead of the prox operator. It is important that the denoiser function is treated as a black box, i.e., we only need access to the output of the denoiser for a given input, irrespective of its functional form.

The NISTA is summarized in Algorithm 1. The algorithm is started with $\Delta\mathbf{p}_0 = \mathbf{0}$. However, to reduce the number of iterations, we can perform a warm start of the inner loop by using the results of the previous iteration. Any approximation of the Hessian (diagonal, BFGS approximation) can be used to perform the Hessian-vector product in line 5 of Algorithm 1. Alternatively, a second-order adjoint-state method can be used to perform this product (Métivier et al. 2013).

Algorithm 1 Adaptive regularization by NISTA.

Require: starting point $\mathbf{m}_k$
1: set $\Delta\mathbf{p}^0 = \mathbf{0}$
2: **repeat**
3: Compute the step direction:
4: **for** $\ell = 1$ to $N-1$ **do**
5: $\Delta\mathbf{m}_k^{\ell+\frac{1}{2}} = \Delta\mathbf{p}^{\ell} - c_k(\mathbf{H}_k\Delta\mathbf{p}^{\ell} + \nabla\mathcal{M}(\mathbf{m}_k))$
6: $\Delta\mathbf{m}_k^{\ell+1} = \text{prox}_{c_k\lambda\mathcal{R}}(\mathbf{m}_k + \Delta\mathbf{m}_k^{\ell+\frac{1}{2}}) - \mathbf{m}_k$
7: $\Delta\mathbf{p}^{\ell+1} = \Delta\mathbf{m}_k^{\ell+1} + \frac{\ell-1}{\ell+2}(\Delta\mathbf{m}_k^{\ell+1} - \Delta\mathbf{m}_k^{\ell})$
8: **end for**
9: Select step length α_k with a line search algorithm.
10: Update: $\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k\Delta\mathbf{m}_k^N$.
11: **until** stopping conditions are satisfied.

3.2 Newton-ADMM (NADMM)

NADMM is obtained by solving optimization problem (3.4) via the alternating direction method of multipliers (ADMM) (Boyd et al. 2010). By introducing the auxiliary variable $\mathbf{p} = \mathbf{m}_k + \Delta\mathbf{m}$, we recast the minimization problem in (3.4) as the following constrained problem:

$$\min_{\Delta\mathbf{m},\mathbf{p}} \quad \widetilde{\mathcal{M}}_k(\mathbf{m}_k + \Delta\mathbf{m}) + \lambda\mathcal{R}(\mathbf{p}) \quad \text{subject to} \quad \mathbf{m}_k + \Delta\mathbf{m} = \mathbf{p}. \tag{3.6}$$

Solving problem (3.6) with an augmented Lagrangian method leads to the following saddle point problem

$$\min_{\Delta\mathbf{m},\mathbf{p}} \max_{\mathbf{q}} \widetilde{\mathcal{M}}_k(\mathbf{m}_k + \Delta\mathbf{m}) + \lambda\mathcal{R}(\mathbf{p}) + \langle\mathbf{q}, \mathbf{m}_k + \Delta\mathbf{m} - \mathbf{p}\rangle + \frac{1}{2c_k}\|\mathbf{m}_k + \Delta\mathbf{m} - \mathbf{p}\|_2^2, \tag{3.7}$$

where $\mathbf{q}$ is the Lagrange multiplier and $1/c_k$ serves as a penalty parameter. Applying the scaled form of ADMM (Boyd et al. 2010, section 3.1.1) to problem (3.7), when combined with equation (3.3), gives the iteration

$$\begin{cases} \Delta\mathbf{m}_k = \arg\min_{\Delta\mathbf{m}} \widetilde{\mathcal{M}}_k(\mathbf{m}_k + \Delta\mathbf{m}) \\ \qquad\qquad + \frac{1}{2c_k}\|\mathbf{m}_k + \Delta\mathbf{m} - \mathbf{p}_k - \mathbf{q}_k\|_2^2 \\ \mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k\Delta\mathbf{m}_k \\ \mathbf{p}_{k+1} = \text{prox}_{c_k\lambda\mathcal{R}}(\mathbf{m}_{k+1} - \mathbf{q}_k) \\ \mathbf{q}_{k+1} = \mathbf{q}_k + \mathbf{p}_{k+1} - \mathbf{m}_{k+1}, \end{cases} \tag{3.8}$$

where the primal and dual variables are updated in alternating mode. A key property of the ADMM algorithm, equation (3.8), is that we don't need to solve it exactly at each iteration. The errors which are made by inexact solving of equation (3.8) are compensated by the Lagrange multipliers. This statement was corroborated by numerical experiments that showed that one (inner) iteration of equation (3.8) generates solutions that are accurate enough to guarantee the fast convergence of the ADMM algorithm (Goldstein & Osher 2009; Boyd et al. 2010; Aghamiry et al. 2019b, 2020a).
With a change of variable $\mathbf{m}_k^{prior} = \mathbf{p}_k + \mathbf{q}_k$, the first subproblem in (3.8) requires us to solve

$$\min_{\Delta\mathbf{m}} \widetilde{\mathcal{M}}_k(\mathbf{m}_k + \Delta\mathbf{m}) + \frac{1}{2c_k}\|\mathbf{m}_k + \Delta\mathbf{m} - \mathbf{m}_k^{prior}\|_2^2, \tag{3.9}$$

which has a closed-form minimizer given by

$$\Delta\mathbf{m}_k = (c_k\mathbf{H}_k + \mathbf{I})^{-1}(-c_k\nabla\mathcal{M}(\mathbf{m}_k) + \Delta\mathbf{m}_k^{prior}), \tag{3.10}$$

where $\Delta\mathbf{m}_k^{prior} = \mathbf{m}_k^{prior} - \mathbf{m}_k$. This is a *generalized* gradient step because it implicitly includes the information carried out by

the gradient/subgradient of the possibly non-differentiable regularizer. It is seen that the prior information introduced by $\mathcal{R}(\mathbf{m})$ in the original problem (3.1), regardless of its mathematical form or its differentiability, is replaced by the a priori information that the (unknown) model at each iteration is a sample of a known Gaussian probability density whose mean is $\mathbf{m}_k^{prior}$ and whose covariance matrix is a scaled identity matrix. The regularization appeared as a damping term that encourages the model not to be very far from the dynamic prior/reference model $\mathbf{m}_k^{prior}$, unlike traditional Bayesian approach (Tarantola 2005) where the a priori model is static. The $(c_k\mathbf{H}_k + \mathbf{I})^{-1}$ in Newton system (3.10), line 4 of Algorithm 2, can be approximated with any quasi-Newton or inexact Newton methods such as L-BFGS (Nocedal & Wright 2006) or the truncated Newton method (Métivier et al. 2017). The proposed NADMM method is summarized in Algorithm 2.

Algorithm 2 Adaptive regularization by NADMM.

Require: starting point $\mathbf{m}_0$
1: set $\mathbf{p}_0 = \mathbf{q}_0 = \mathbf{0}$
2: **repeat**
3: Compute the Hessian $\mathbf{H}_k$ or an approximation to it.
4: Compute the step direction: $\Delta\mathbf{m}_k = (c_k\mathbf{H}_k + \mathbf{I})^{-1}(-c_k\nabla\mathcal{M}(\mathbf{m}_k) + \mathbf{p}_k + \mathbf{q}_k - \mathbf{m}_k)$.
5: Select step length α_k with a line search algorithm.
6: Update: $\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k\Delta\mathbf{m}_k$.
7: Update: $\mathbf{p}_{k+1} = \text{prox}_{c_k\lambda\mathcal{R}}(\mathbf{m}_{k+1} - \mathbf{q}_k)$.
8: Update: $\mathbf{q}_{k+1} = \mathbf{q}_k + \mathbf{p}_{k+1} - \mathbf{m}_{k+1}$
9: **until** stopping conditions are satisfied.

3.3 Application to Full Waveform Inversion

In the *Numerical example* section, we assess the algorithms 1 and 2 against seismic full-waveform inversion (FWI) (Virieux & Operto 2009) methods with a series of benchmark velocity models. FWI is a nonlinear multivariate PDE-constrained optimization problem, which can be formulated in the frequency domain as

$$\min_{\mathbf{m},\mathbf{u}} \quad \|\mathbf{P}\mathbf{u} - \mathbf{d}\|_2^2, \quad \text{subject to} \quad \mathbf{A}(\mathbf{m})\mathbf{u} = \mathbf{b}, \tag{3.11}$$

where $\mathbf{m} \in \mathbb{R}^{N\times 1}$ denotes the discrete subsurface model parameters, $\mathbf{b} \in \mathbb{C}^{N\times 1}$ the source term, $\mathbf{u} \in \mathbb{C}^{N\times 1}$ the modelled wavefield, $\mathbf{d} \in \mathbb{C}^{M\times 1}$ the recorded seismic data and $\mathbf{P} \in \mathbb{R}^{M\times N}$ a linear observation operator that samples $\mathbf{u}$ at the receiver positions. This objective function is for one frequency and one source. Extension to multiple frequencies and sources is simply implemented by summation over sources and frequencies in the objective function and by adding multiple right-hand sides in the constraint, equation (3.11) (Aghamiry et al. 2019a). The matrix $\mathbf{A}(\mathbf{m}) \in \mathbb{C}^{N\times N}$, whose coefficients depend on $\mathbf{m}$, represents the discretized PDE (Pratt et al. 1998; Plessix 2007). Here, we briefly review the two different formulations of FWI that will be used. The first classical one relies on variable projection to recast the nonlinear constrained problem as an unconstrained problem with a reduced search space (Plessix 2006). The second extends the linear regime of the waveform inversion by relaxation of the PDE constraint with ADMM (Aghamiry et al. 2019c) in the framework of the wavefield reconstruction inversion (WRI) method (van Leeuwen & Herrmann 2013, 2016) (see also Huang et al. (2018a) and Huang et al. (2018b) for some variants). This recasts the original nonlinear constrained problem as a biconvex problem according to the bilinearity of the wave equation in wavefield and model parameters.

3.3.1 Reduced-space FWI

Classical FWI (Pratt et al. 1998; Plessix 2006) strictly enforces the PDE constraint, $\mathbf{u} = \mathbf{A}(\mathbf{m})^{-1}\mathbf{b}$, in the misfit function at each iteration by projection of the full multivariate search space onto the parameter search space for sake of computational efficiency. This leads to the following monovariate misfit function

$$\mathcal{M}(\mathbf{m}) = \frac{1}{2}\|\mathbf{d} - F(\mathbf{m})\|_2^2, \tag{3.12}$$

where $F(\mathbf{m}) = \mathbf{P}\mathbf{A}^{-1}(\mathbf{m})\mathbf{b}$ is the calculated data. The gradient and the Hessian of equation (3.12), which are required for Proximal-Newton methods, are given by (Pratt et al. 1998)

$$\nabla\mathcal{M}(\mathbf{m}) = -\mathbf{J}^T\boldsymbol{\Delta}\mathbf{d}, \tag{3.13}$$

and

$$\nabla^2\mathcal{M}(\mathbf{m}) = \mathbf{J}^T\mathbf{J} + \frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}, [\boldsymbol{\Delta}\mathbf{d}|\cdots|\boldsymbol{\Delta}\mathbf{d}], \tag{3.14}$$

where $\boldsymbol{\Delta}\mathbf{d} = \mathbf{d} - F(\mathbf{m})$ and $\mathbf{J}$ is the sensitivity or the Fréchet derivative matrix, defined as

$$\mathbf{J}_{ij}(\mathbf{m}) = \frac{\partial[F(\mathbf{m})]_i}{\partial\mathbf{m}_j}. \tag{3.15}$$

In practice, the gradient is calculated with the adjoint-state method (Plessix 2006). The action of the Hessian can be taken into account approximately with preconditioned quasi-Newton method (l-BFGS) or truncated Newton methods (see Métivier & Brossier (2016) for an overview).

3.3.2 ADMM-based Wavefield Reconstruction Inversion (IR-WRI)

In classical FWI, the wave-equation $\mathbf{A}(\mathbf{m})\mathbf{u} = \mathbf{b}$ is solved exactly at each iteration to generate the reduced form of the objective function (3.12). In the wavefield reconstruction inversion (WRI) method (van Leeuwen & Herrmann 2013, 2016), the wave-equation is processed as a weak constraint (namely, it is satisfied approximately) with a penalty method such that the simulated wavefields match the observations to a great extent. Then, the parameters are updated from the wavefields by least-squares minimization of the wave equation errors (van Leeuwen & Herrmann 2013; Aghamiry et al. 2019c). Updating the wavefields and the subsurface parameters in alternating mode at iteration k leads to the following objective function for $\mathbf{m}$

$$\mathcal{M}(\mathbf{m}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}(\mathbf{m})\mathbf{u}_k\|_2^2, \tag{3.16}$$

where the so-called data-assimilated wavefield $\mathbf{u}_k$ (Aghamiry et al. 2020b) is the least-squares solution of the overdetermined system gathering the wave equation and the observation equation

$$\begin{pmatrix}\mathbf{A}(\mathbf{m}_{k-1})\\ \mu\mathbf{P}\end{pmatrix}\mathbf{u}_k = \begin{pmatrix}\mathbf{b}\\ \mu\mathbf{d}\end{pmatrix}, \tag{3.17}$$

where $\mu > 0$ is the penalty parameter. Note that equations (3.16) and (3.17) are provided assuming a single source experiment. For multiple sources, the objective function (3.16) is simply obtained by summation over sources, while one augmented system (3.17) per source needs to be solved (Aghamiry et al. 2019a). For (3.16), the gradient and Hessian are given by

$$\nabla\mathcal{M}(\mathbf{m}) = -\mathbf{L}^T(\mathbf{b} - \mathbf{A}(\mathbf{m})\mathbf{u}_k), \tag{3.18}$$

where

$$\nabla^2 \mathcal{M}(\mathbf{m}) = \mathbf{L}^T \mathbf{L}, \tag{3.19}$$

and

$$\mathbf{L}_{ij} = \frac{\partial [\mathbf{A}(\mathbf{m})\mathbf{u}_k]_i}{\partial \mathbf{m}_j}. \tag{3.20}$$

The columns of the matrix $\mathbf{L}$ contain the so-called virtual sources (Pratt et al. 1998), that makes this matrix diagonal when optimization parameters are squared slownesses.

4 NUMERICAL EXAMPLES

4.1 Wavefield Inversion of Inclusions Models

We first assess the NISTA and NADMM algorithms in the FWI/WRI framework using BM3D as denoiser. To highlight the adaptivity power of BM3D, we design four different velocity models, where inclusions of different shape are added to a 2 km $\times$ 2 km homogeneous background velocity model ($\mathbf{v} = 2$ km/s) (first column of Fig. 2). Also, to show the flexibility of the proposed adaptive regularization in managing different priors simultaneously, we put all the four inclusions together in the background model and reconstruct them jointly (Fig. 3). For all the tests of this section, data are generated by five sources at the surface (with 400 m spacing) and 50 m equally spaced receivers placed on all the boundaries except the surface. The forward modelling is performed with a 9-point stencil finite-difference method implemented with anti-lumped mass and PML absorbing boundary conditions to solve the Helmholtz equation, where the stencil coefficients are optimized to the frequency (Chen et al. 2013) (this scenario is considered for all wave-propagation examples in this paper). The source signature is a Ricker wavelet with a 10 Hz dominant frequency. We start the inversion from the homogeneous background model ($\mathbf{v} = 2$ km/s) and invert simultaneously four frequency components (5, 7, 10, and 12.5 Hz) of noiseless and noisy data with FWI and IR-WRI with and without the BM3D regularization.

We first apply FWI and IR-WRI via NADMM without and with BM3D regularization. It is worth to mention that, when a denoiser is used for regularization, the regularization weight λ is embedded in the standard deviation of noise (σ) of the denoiser, $\sigma = \sqrt{c_k \lambda}$ (see equation 2.6). We set σ of BM3D equal to 25 and 45 for noiseless and noisy data, respectively (Appendix A). We perform FWI with the L-BFGS quasi-Newton method with line search to perform NADMM. We perform the inversion with noiseless data and set the maximum number of iteration to 70 as stopping criterion for IR-WRI in both cases (without and with BM3D). For FWI, the stopping criteria is set to the model error (l_2-norm of the difference between true and estimated model) achieved by IR-WRI for a fair comparison between the two waveform inversion methods. Fig. 2 shows the results obtained by FWI and IR-WRI for all four models without and with regularization. It is clearly seen that, for both methods, regularization improved the results and successfully recovered the shape of the different anomalies, thanks to the adaptive nature of the BM3D. It is worth noting that IR-WRI performed better than FWI, although the focus of this paper is not to compare the two methods as this topic has been already addressed in Aghamiry et al. (2019c). Instead, we aim to show how adaptive regularization can improve the results of both methods when it is implemented with proximal Newton algorithms.

We continue by using a model that includes all four inclusions (Fig. 3a). Fig. 3 shows the velocity models estimated by FWI (Figs. 3b-c) and IR-WRI (Figs. 3d-e) with and without regularization. The different inclusions are jointly reconstructed with the same accuracy as in the case where they are reconstructed independently (Fig. 2). This highlights the local-adaptivity property of the BM3D regularization.

4.1.1 Sensitivity to the regularization parameter

The only free parameter of BM3D is σ (Appendix A) which is selected by trial and error in this paper. Here we want to asses the sensitivity of the final results against σ. We apply NADMM-based FWI and IR-WRI with BM3D as regularizer on noiseless data using different values of σ. For each σ, the stopping criteria is chosen as in the previous test (70 iterations for IR-WRI and model error achieved by IR-WRI for FWI). The velocity models estimated by FWI and IR-WRI with BM3D regularization for $\sigma = [160, 40, 5, 1]$ are shown in Fig. 4. Note that the results with $\sigma = 20$ are shown in Figs. 3c and 3e for FWI and IR-WRI, respectively. It is seen that both of FWI and IR-WRI can reconstruct an acceptable model for a wide range of values of σ, i.e. $\sigma = [5 - 40]$.

4.1.2 Robustness against noise

We continue by assessing the robustness of the proposed method against random noise. We apply NADMM-based FWI and IR-WRI without and with BM3D when the data are contaminated with different level of random noises. The relative root mean square error (RMSE) curves versus signal to noise ratio (SNR) is depicted in Fig. 5, where RMSE and SNR are defined as

$$\text{RMSE} = 100 \frac{\|\mathbf{m} - \mathbf{m}_*\|_2}{\|\mathbf{m}_*\|_2}, \tag{4.1}$$

in which $\mathbf{m}$ and $\mathbf{m}_*$ are the estimated and true models, respectively, and

$$\text{SNR} = 20 \log \left(\frac{Signal\ RMS\ amplitude}{Noise\ RMS\ amplitude} \right). \tag{4.2}$$

Fig. 5 shows the average value (over 20 runs) for each SNR. Furthermore, we use $\|\mathbf{P}\mathbf{u}_k - \mathbf{d}\|_2 = 1.01\varepsilon$ as the stopping criterion of iteration, where ε is the ℓ_2 norm of the noise. The velocity models estimated by FWI and IR-WRI without/with BM3D regularization for SNR=5db are shown in Fig. 6. In order to show how the data are fitted, the difference between the estimated and noiseless (10 Hz) data for the tests of Fig. 6 are shown in Fig. 7.

4.1.3 A comparison between NISTA and NADMM

Here, we use the BM3D regularized FWI with noiseless and noisy data to compare NISTA and NADMM. Both NISTA and NADMM algorithms are implemented with L-BFGS to account for the action of the Hessian (Line 5 and 4 of Algorithms 1 and 2, respectively). Fig. 8 shows the velocity models reconstructed by both algorithms and Fig. 9 shows the corresponding convergence history (the value of the objective function as a function of the iteration count). Although we perform approximately 100 inner iterations of proximal gradient to estimate the search direction of NISTA, NADMM still performs better. Furthermore, since we implement both algorithms with L-BFGS, the results show that in practice NADMM should be preferred to NISTA.

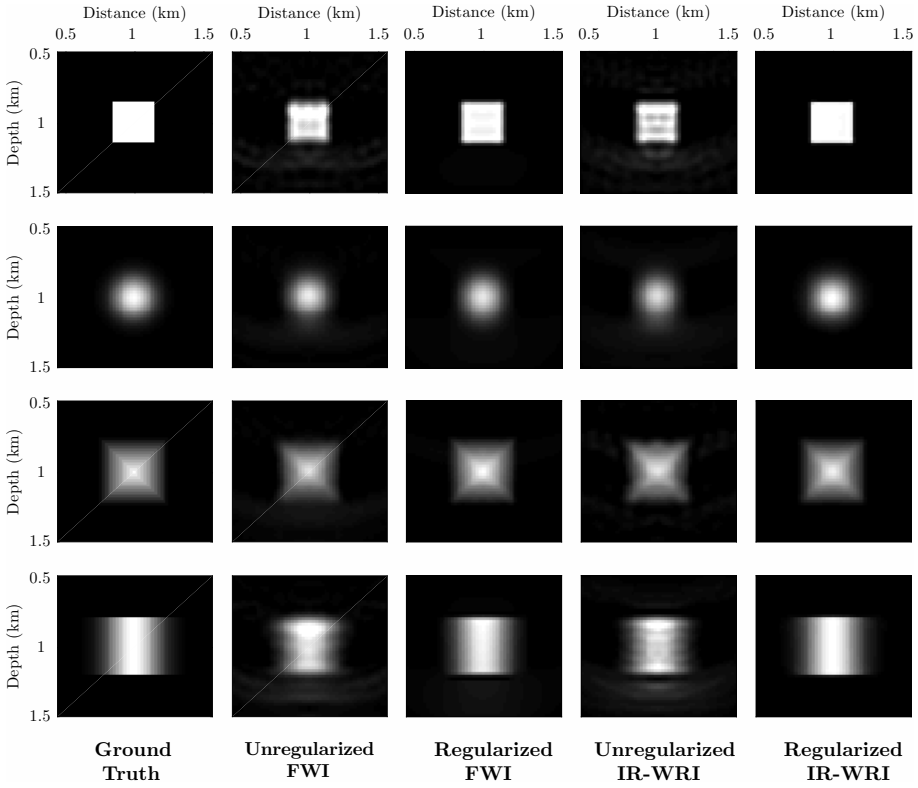

Figure 2. The performances of FWI and IR-WRI with adaptive NADMM for reconstruction of different velocity structures. The data were generated by five sources at the surface (with 400 m spacing) and 50 m equally spaced receivers positioned on all the boundaries except the surface. In all figures, the colour bar varies between 2 km/s and 2.5 km/s with low velocities in black and high velocities in white.

4.2 Performance on benchmark models

We continue by assessing the performance of the NADMM-based adaptive IR-WRI algorithm using well-documented 2D benchmark subsurface velocity models in exploration seismic, e.g. the Marmousi II (Martin et al. 2006), SEG/EAGE overthrust (Aminzadeh et al. 1997), SEG/EAGE salt (Aminzadeh et al. 1997), synthetic Valhall (Prieux et al. 2011) and 2004 BP salt (Billette & Brandsberg-Dahl 2004) benchmark velocity models. The selected target from these benchmark models are shown in the first column of Fig. 10, respectively. The fixed-spread acquisition with a few equally spaced sources at the sea bottom and a line of equally spaced receivers at the 25 m depth is used for all of the tests (see Table 1 for more technical details). Also, the models are discretized with 25 m spacing in horizontal and vertical directions. We compute the wavefields using Perfectly-Matched Layer (PML) absorbing boundary conditions along the bottom, right, and left sides of the model using 10 grid points in the PMLs and a free-surface boundary condition at the surface and a 10 Hz Ricker wavelet is used as the source signature. We design a multiscale inversion with a classical continuation frequency strategy in the selected frequency band by proceeding over small batches of two frequencies with a frequency interval of 0.5 Hz. We also perform three paths through the batches, where the starting and finishing frequencies of the paths and other technical details about the models are reported in Table 1. Also, we use 60, 40, and 25, respectively, for σ of BM3D in the first, second and third path. The initial velocity models are crude models, as shown in the second column of Fig. 10. Accordingly, we tackle these benchmarks with IR-WRI only since FWI would remain stuck in a local minimum due to cycle skipping. We set the number of IR-WRI iterations per frequency batch equal to 10 or ℓ_2-norm of source residuals equal to 1e-3 as the stopping criteria. The estimated models without and with BM3D regularization with NADMM (Algorithm 2) are shown in the third and fourth columns of Fig. 10, respectively. A direct comparison along vertical logs (as depicted with dashed lines in first column of Fig. 10) between the true velocity, the initial and the final velocity models without/with BM3D regularization are shown in Figs. 11a-d, for Marmousi II, SEG/EAGE salt, Synthetic Valhall and 2004 BP salt models, respectively. The results show that, although the different benchmark models are characterized by different kinds of structures, adaptive regularization combined with IR-WRI manages to reconstruct accurately each of them with a significant jump of quality compared to the case where BM3D is not used.

5 CONCLUSIONS

In this paper, we proposed a flexible optimization framework based upon proximal Newton methods to implement arbitrary regularizations (smooth, nonsmooth, adaptive) in nonlinear inverse problem such as FWI and its WRI variant. The regularization problem is recast as a denoising problem through proximal mapping such that the optimization algorithm only needs the output of the denoiser. This means that any denoiser can be easily implemented as a black box in the optimization algorithm. This open the door to leading-edge plug-and-play denoisers such as the block-matching BM3D method, which has no explicit optimization formulation (it can not be cast as a variational problem). Two different algorithms (NISTA

Table 1. Technical details of the benchmark models

	Size (km × km)	Source interval(m)	Receiver interval(m)	Inverted frequency band (Hz) [starting-finishing]	Outer iterations frequencies (Hz) [starting-finishing]
Marmousi II	4.25 × 11.5	250	50	[3-10]	[3-8], [4-9], [5-10]
SEG/EAGE salt model	2.1 × 7.8	100	25	[3-7]	[3-6], [3.5-7], [4-7]
Synthetic Valhall	5.25 × 16	500	100	[3-13]	[3-9], [4-11.5], [5-13]
2004 BP salt	5.8 × 16.25	250	50	[3-13]	[3-9.5], [3.5-11.5], [5-13]

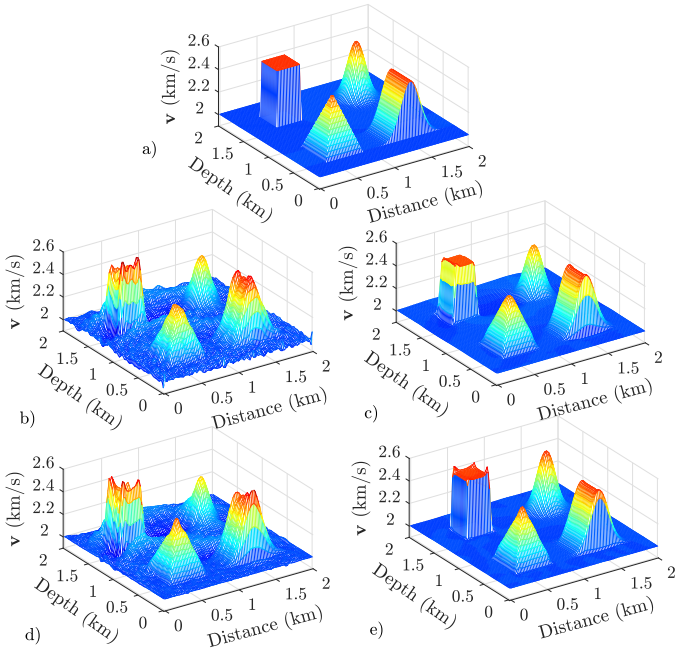

Figure 3. Inclusion test. (a) True velocity model. (b-e) Velocity models estimated by (b) FWI without regularization, (c) FWI with regularization, (d) IR-WRI without regularization, (e) IR-WRI with regularization.

and NADMM) based upon FISTA and ADMM are proposed to implement the proximal Newton method. Both of them requires to perform Hessian-vector products, which can be implemented with quasi-Newton or second-order adjoint-state method. Both algorithms are assessed against a series of well-documented subsurface models. In all cases, the results show that NADMM should be preferred at the expense to NISTA. Although the adaptive BM3D regularizer was used in this study, more conventional Tikhonov, Total Variation or a combination of the two can be easily implemented in NISTA-based and NADMM-based FWI and IR-WRI. To allow people to evaluate the NISTA and NADMM algorithm, we have made some codes available at https://gitlab.oca.eu/wind.

6 ACKNOWLEDGEMENTS

We would like to thank editors Frederik Simons, Fern Storey, and two anonymous reviewers for their comments which help improving the manuscript. This study was partially funded by the WIND consortium (*https://www.geoazur.fr/WIND*), sponsored by Chevron, Shell, and Total. This study was granted access to the HPC resources of SIGAMM infrastructure (http://crimson.oca.eu), hosted by Observatoire de la Côte d'Azur and which is supported by the Provence-Alpes Côte d'Azur region, and the HPC resources of CINES/IDRIS/TGCC under the allocation A0050410596 made by GENCI. Hossein S. Aghamiry would like to thank the IDEX UCA JEDI with the WIMAG project for their support. Ali Gholami would like to acknowledge the financial support of university of Tehran for this research under grant number 27711-1-06.

REFERENCES

Aghamiry, H., Gholami, A., & Operto, S., 2019a. ADMM-based multi-parameter wavefield reconstruction inversion in VTI acoustic media with TV regularization, *Geophysical Journal International*, **219**(2), 1316–1333.

Aghamiry, H., Gholami, A., & Operto, S., 2019b. Implementing bound constraints and total-variation regularization in extended full waveform inversion with the alternating direction method of multiplier: application to large contrast media, *Geophysical Journal International*, **218**(2), 855–872.

Aghamiry, H., Gholami, A., & Operto, S., 2019c. Improving full-waveform inversion by wavefield reconstruction with alternating direction method of multipliers, *Geophysics*, **84(1)**, R139–R162.

Aghamiry, H., 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., 2020b. Accurate and efficient data-assimilated wavefield reconstruction in the time domain, *Geophysics*, **85**(2), A7–A12.

Aminzadeh, F., Brac, J., & Kunz, T., 1997. *3-D Salt and Overthrust models*, SEG/EAGE 3-D Modeling Series No.1.

Attouch, H. & Peypouquet, J., 2016. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than 1/k^2, *SIAM Journal on Optimization*, **26**(3), 1824–1834.

Beck, A. & Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal Imaging Sciences*, **2(1)**, 183–202.

Billette, F. J. & Brandsberg-Dahl, S., 2004. The 2004 BP velocity benchmark, in *Extended Abstracts, 67th Annual EAGE Conference & Exhibition, Madrid, Spain*, p. B035.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J., 2010. Distributed optimization and statistical learning via the alternating direction of multipliers, *Foundations and trends in machine learning*, **3**(1), 1–122.

Chen, Z., Cheng, D., Feng, W., & Wu, T., 2013. An optimal 9-point finite difference scheme for the Helmholtz equation with PML, *International Journal of Numerical Analysis & Modeling*, **10**(2).

Combettes, P. L. & Pesquet, J.-C., 2011. Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, vol. 49 of **Springer Optimization and Its Applications**, pp. 185–212, eds Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., & Wolkowicz, H., Springer New York.

Dabov, K., Foi, A., & Egiazarian, K., 2007. Video denoising by sparse 3D transform-domain collaborative filtering, in *2007 15th European Signal Processing Conference*, pp. 145–149, IEEE.

Danielyan, A., Katkovnik, V., & Egiazarian, K., 2011. Bm3d frames and

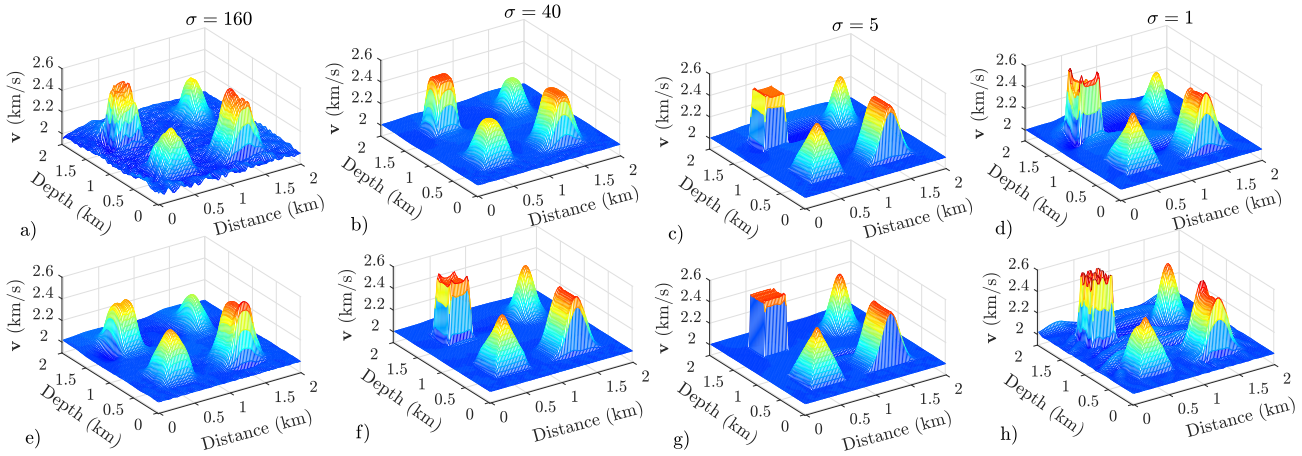

Figure 4. Inclusion test with different σ for BM3D. (a-d) Estimated velocity model using FWI with BM3D regularization with $\sigma = 160$ (a), 40 (b), 5 (c), and 1 (d). (e-h) Same as (a-d) but for IR-WRI.

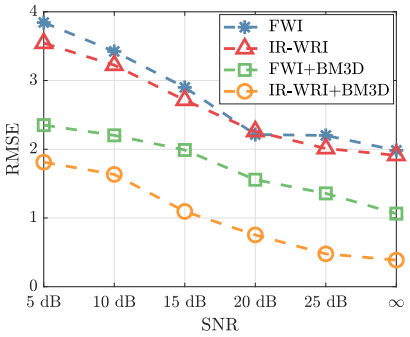

Figure 5. RMSE for FWI and IR-WRI without and with BM3D using NADMM when data are contaminated with different level of noises.

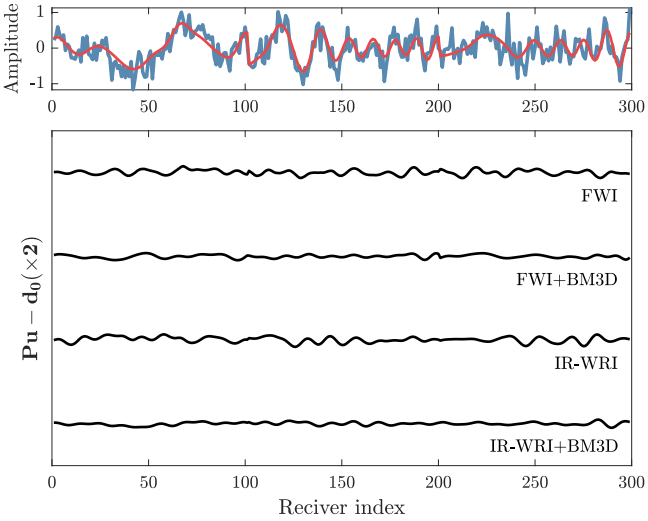

Figure 7. Real part of the 10 Hz data. (a) The noisy data with SNR=5db are shown in blue, while the noiseless data are shown in red. (b) The difference between predicted data ($\mathbf{Pu}_k$) and noiseless data ($\mathbf{d}_0$) at the final iteration of Fig. 6. The residual curves are scaled by factor 2.

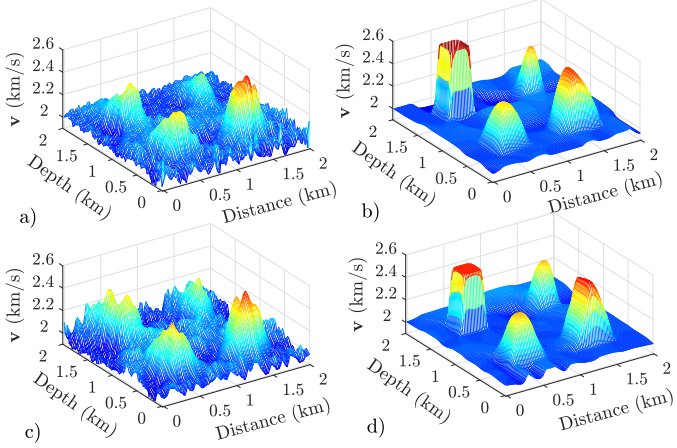

Figure 6. Inclusion test with SNR=5db. Estimated velocity model using (a) FWI without regularization, (b) FWI with regularization, (c) IR-WRI without regularization, (d) IR-WRI with regularization.

variational image deblurring, *IEEE Transactions on Image Processing*, **21**(4), 1715–1728.

Daubechies, I., Defrise, M., & De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, **57**(11), 1413–1457.

Gholami, A. & Hosseini, S. M., 2011. A general framework for sparsity-based denoising and inversion, *IEEE transactions on signal processing*, **59**(11), 5202–5211.

Gholami, A. & Siahkoohi, H., 2010. Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints, *Geophysical Journal International*, **180**(2), 871–882.

Goldstein, T. & Osher, S., 2009. The split Bregman method for L1-regularized problems, *SIAM Journal on Imaging Sciences*, **2**(2), 323–343.

Goyal, B., Dogra, A., Agrawal, S., Sohi, B., & Sharma, A., 2020. Image denoising review: From classical to state-of-the-art approaches, *Information Fusion*, **55**, 220–244.

Huang, G., Nammour, R., & Symes, W. W., 2018a. Source-independent extended waveform inversion based on space-time source extension: Frequency-domain implementation, *Geophysics*, **83**(5), R449–R461.

Huang, G., Nammour, R., & Symes, W. W., 2018b. Volume source-based extended waveform inversion, *Geophysics*, **83**(5), R369–387.

Kamilov, U. S., Mansour, H., & Wohlberg, B., 2017. A plug-and-play priors approach for solving nonlinear imaging inverse problems, *IEEE Signal Processing Letters*, **24**(12), 1872–1876.

Lange, K., 2016. *MM optimization algorithms*, vol. 147, SIAM.

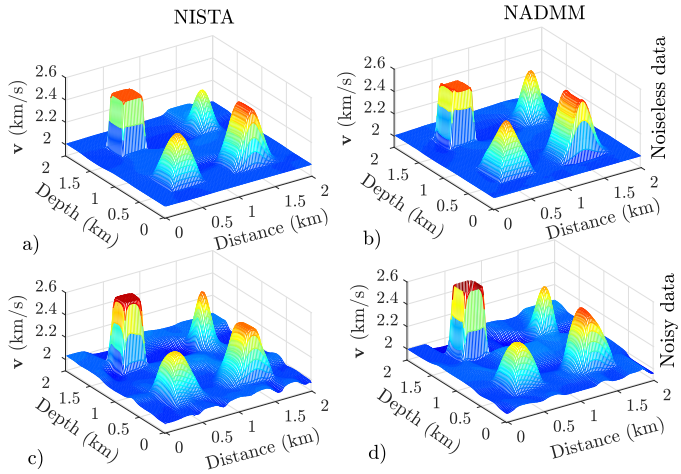

Figure 8. Inclusion test with BM3D regularized FWI using NISTA (a and c) and NADMM (b and d). (a-b) Noiseless data, (c-d) Noisy data with SNR=5db.

Lebrun, M., 2012. An analysis and implementation of the bm3d image denoising method, *Image Processing On Line*, **2**, 175–213.

Lee, J. D., Sun, Y., & Saunders, M. A., 2014. Proximal Newton-type methods for minimizing composite functions, *SIAM Journal on Optimization*, **24**(3), 1420–1443.

Martin, G. S., Wiley, R., & Marfurt, K. J., 2006. Marmousi2: An elastic upgrade for Marmousi, *The Leading Edge*, **25**(2), 156–166.

Meinhardt, T., Moller, M., Hazirbas, C., & Cremers, D., 2017. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1781–1790.

Métivier, L. & Brossier, R., 2016. The seiscope optimization toolbox: A large-scale nonlinear optimization library based on reverse communication, *Geophysics*, **81**(2), F11–F25.

Métivier, L., Brossier, R., Virieux, J., & Operto, S., 2013. Full Waveform Inversion and the truncated Newton method, *SIAM Journal On Scientific Computing*, **35(2)**, B401–B437.

Métivier, L., Brossier, R., Operto, S., & J., V., 2017. Full waveform inversion and the truncated Newton method, *SIAM Review*, **59**(1), 153–195.

Milanfar, P., 2012. A tour of modern image filtering: New insights and methods, both practical and theoretical, *IEEE signal processing magazine*, **30**(1), 106–128.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate $O(1/k^2)$, in *Dokl. Akad. Nauk SSSR*, vol. 269, pp. 543–547.

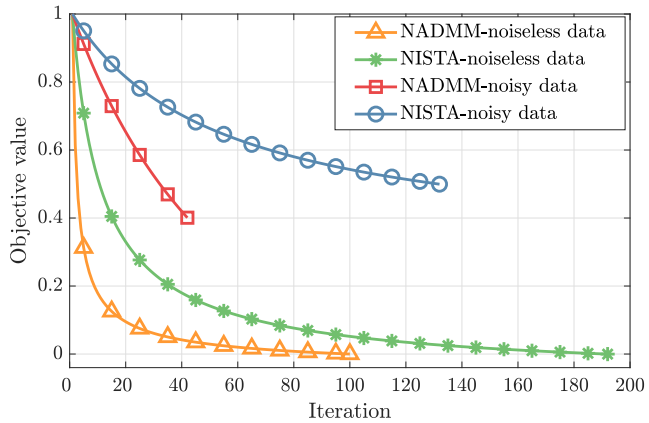

Figure 9. Evaluation of the objective function of four different tests of Fig. 8.

Nocedal, J. & Wright, S. J., 2006. *Numerical Optimization*, Springer, 2nd edn.

Parikh, N. & Boyd, S., 2013. Proximal algorithms, *Foundations and Trends in Optimization*, **1(3)**, 123–231.

Peters, B. & Herrmann, F. J., 2017. Constraints versus penalties for edge-preserving full-waveform inversion, *The Leading Edge*, **36**(1), 94–100.

Peters, B., Smithyman, B. R., & Herrmann, F. J., 2019. Projection methods and applications for seismic nonlinear inverse problems with multiple constraints, *Geophysics*, **84**(2), R251–R269.

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.

Plessix, R. E., 2007. A Helmholtz iterative solver for 3D seismic-imaging problems, *Geophysics*, **72**(5), SM185–SM194.

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

Prieux, V., Brossier, R., Gholami, Y., Operto, S., Virieux, J., Barkved, O., & Kommedal, J., 2011. On the footprint of anisotropy on isotropic full waveform inversion: the Valhall case study, *Geophysical Journal International*, **187**, 1495–1515.

Selesnick, I. & Farshchian, M., 2017. Sparse signal approximation via nonseparable regularization, *IEEE Transactions on Signal Processing*, **65**(10), 2561–2575.

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

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

Tikhonov, A. & Arsenin, V., 1977. *Solution of ill-posed problems*, Winston, Washington, DC.

van Leeuwen, T. & Herrmann, F., 2016. A penalty method for PDE-constrained optimization in inverse problems, *Inverse Problems*, **32(1)**, 1–26.

van Leeuwen, T. & Herrmann, F. J., 2013. Mitigating local minima in full-waveform inversion by expanding the search space, *Geophysical Journal International*, **195(1)**, 661–667.

Venkatakrishnan, S. V., Bouman, C. A., & Wohlberg, B., 2013. Plug-and-play priors for model based reconstruction, in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948, IEEE.

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

APPENDIX A: BLOCK-MATCHING AND 3D FILTERING (BM3D)

Most of the denoising techniques are based on shrinking small coefficients in a transformed domain where the coefficient distribution is sparse. BM3D (Dabov et al. 2007) is a novel image denoising technique, which is based on an enhanced sparse representation in a transformed domain where the self-similarities in the image are exploited. This method contains four steps (Lebrun 2012):

- Select an image patch with size $k \times k$ and find all the similar patches in the original image and group them in a 3D cube.
- Apply a 3D linear transform on the cube.
- Filter the transform coefficients by thresholding or Wiener filtering.
- Apply inverse 3D transform and return the filtered patches to their correct positions.

Redundancy between patches enables BM3D to reconstruct smooth and flat parts properly. Furthermore, it is capable to reconstruct fine details and sharp edges. Despite good performance of this technique, it has some drawback like artifacts in denoised images when

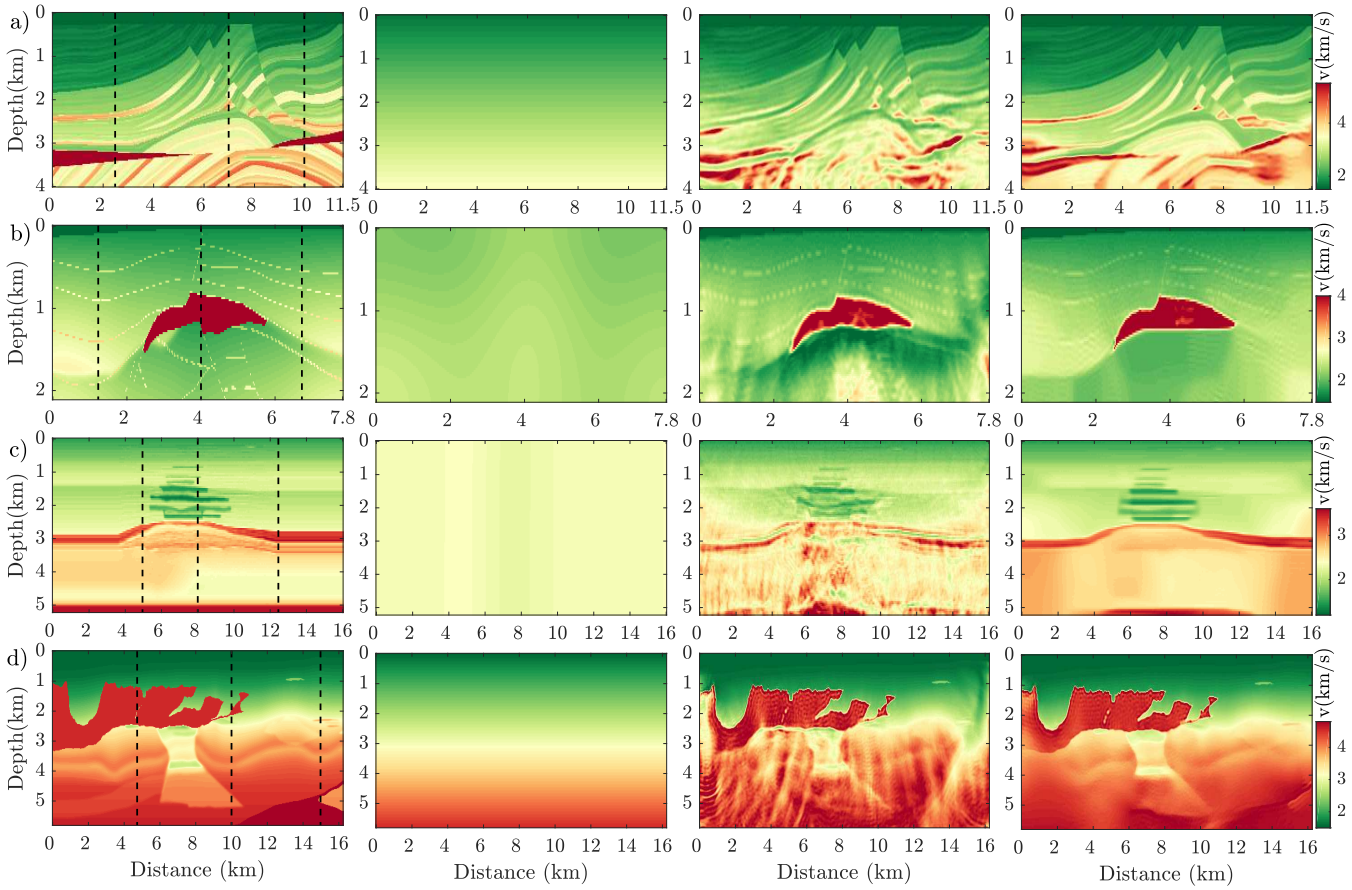

Figure 10. IR-WRI without/with BM3D regularization on benchmark models. (a) Marmousi II, (b) SEG/EAGE salt, (c) Synthetic Valhalla and (d) 2004 BP salt models. The columns of this figure are as follows: True velocity model, initial velocity model, IR-WRI and BM3D regularized IR-WRI with NADMM. The vertical dashed lines indicate the location of vertical logs of Fig. 11

the signal to noise ratio is poor. Also, the computational complexity is much higher than that of simple denoising methods. The most computational burden of this method is generated by the first step, which is grouping and comparing similar patches. For a $N \times N$ image with a $k \times k$ patch, it is in order of $O(N^4k^2)$. The burden of the remaining steps is related to the used sparsifying transform. In this paper, we use 3D discrete cosine transform (DCT) and the computational complexity can be expressed as $O(k^2R\,log(k^2R))$ where R is the number of similar blocks in each cube. Moreover, BM3D and all of the patch-based denoising methods have a large number of parameters that are difficult to adjust properly. In this paper, we used the official package of BM3D, just set the standard deviation of noise (σ) and used the default values for the rest of the parameters according to table I of Dabov et al. (2007). The σ is selected by trial and error in this paper.

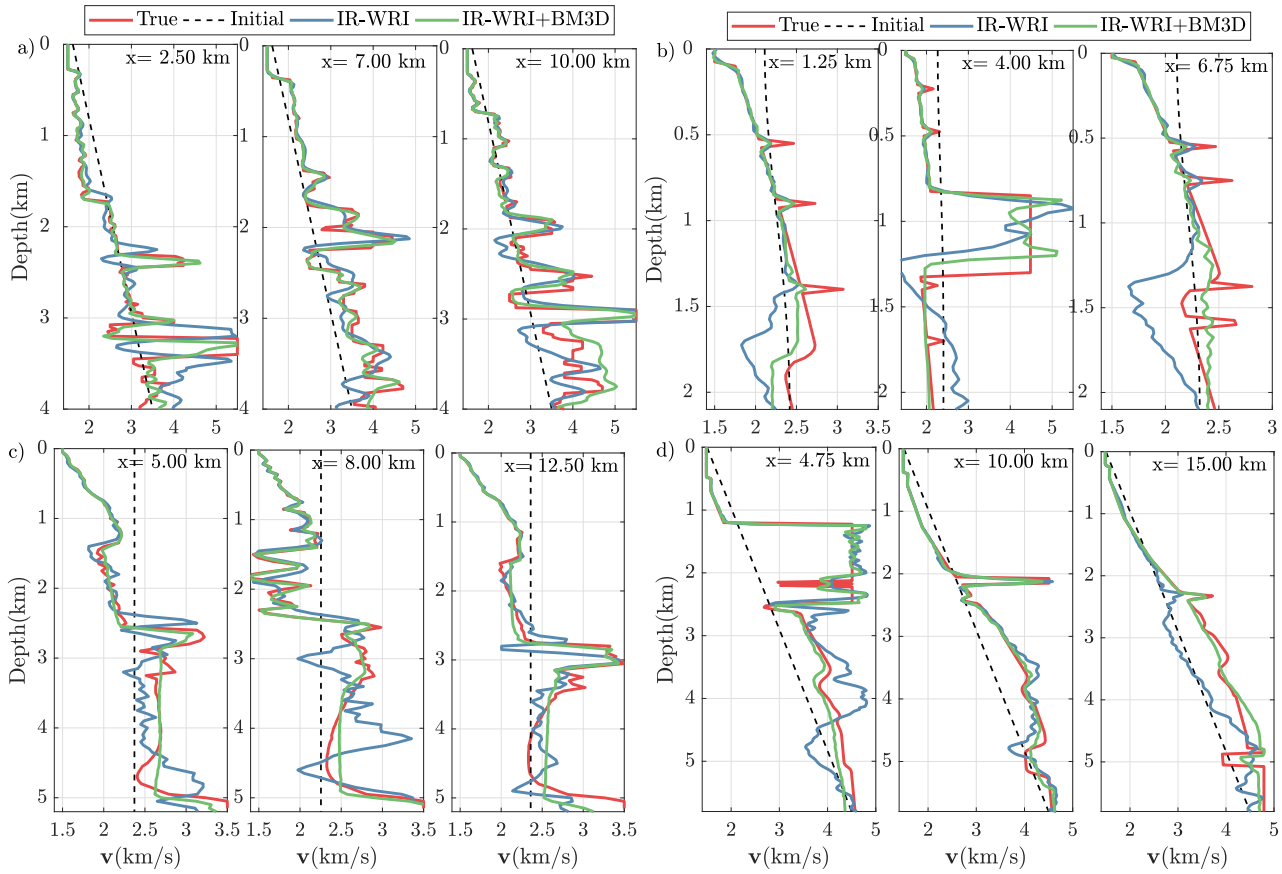

Figure 11. Direct comparisons between velocity models estimated by IR-WRI without/with BM3D regularization . (a) Marmousi II, (b) SEG/EAGE salt, (c) Synthetic Valhall and (d) 2004 BP salt models. In each panel, the true model is solid black, initial model is dashed black, the estimated model without regularization is blue and estimated model with regularization is orange. The horizontal location of each log is written in each panel and depicted with vertical lines in the first column of Fig. 10.