Full Waveform Inversion and the truncated Newton: quantitative imaging of complex subsurface structures

Ludovic Métivier, François Bretaudeau, Romain Brossier, Stéphane Operto, Jean Virieux

Geophysical Prospecting (200?) **XX,** 000–000

Full Waveform Inversion and the truncated Newton: quantitative imaging of complex subsurface structures

Ludovic Métivier[1,2], Romain Brossier[1], Stéphane Operto[3], and Jean Virieux[1]

[1] *ISTerre, Université de Grenoble I, BP 53, 38041 Grenoble Cedex 09 France*

[2] *LJK, CNRS, Université de Grenoble I, BP 53, 38041 Grenoble cedex 09, France*

[3] *Géoazur, Université de Nice Sophia-Antipolis, CNRS, IRD, Université Pierre et Marie-Curie, Villefranche-sur-mer, France*

Accepted 2008 ?? ??. Received 2008 June ??; in original form 2008 June ??

SUMMARY

Full Waveform Inversion appears to be a powerful tool for quantitative seismic imaging from wide-azimuth seismic data. The method is based on the minimization of the misfit between observed and simulated data. This amounts to the resolution of a large-scale nonlinear minimization problem. The analysis of the methods used to solve this problem emphasizes the crucial role played by the inverse Hessian operator in the reconstruction process. Accounting accurately for the effect of this operator within the minimization scheme should correct for illumination deficits, restore the amplitude of the subsurface parameters, and help to remove artifacts generated by energetic multiple reflections. As conventional methods, such as the preconditioned nonlinear conjugate gradient, only approximate the effect of this operator, we are interested in this study to another class of minimization methods, named as truncated Newton methods. Instead of approximating the inverse Hessian operator, the truncated Newton methods are based on the resolution of the linear system associated with the computation of the Newton descent direction using a matrix-free conjugate gradient solver. The aim of this study is to propose a feasible implementation of this method for the Full Waveform Inversion problem, and to

investigate in which context the truncated Newton method can outperform conventional methods. To this purpose, three case studies are presented: the first originates from a geotechnical application, the second is based on the Marmousi II model, the third on the benchmark BP 2004 model. These tests emphasize the interest of using the truncated Newton method, which better accounts for the inverse Hessian operator, for imaging complex subsurface models containing highly contrasted structures.

Key words: Full Waveform, Theory, Computing aspects, Numerical study, Imaging

1 INTRODUCTION

Full Waveform Inversion (FWI) is a powerful seismic imaging tool, dedicated to the estimation of subsurface parameters such as P-wave velocity, density, impedance and shear wave velocity. The method is based on the minimization of a misfit function that measures the distance between recorded seismic data and predicted data computed using a wave propagation engine. An initial subsurface model is iteratively updated to produce the final estimation.

The FWI method can be seen as an alternative to the standard two-steps workflow widely used in the industry which consists in first estimating a background velocity model from seismic reflection tomography, and then "migrate" the seismic data to generate a map of the subsurface reflectivity (Claerbout 1971, 1976). The interpretation of the complete wavefield within the FWI method yields the theoretical background to unify these two steps. Large offset data sets should provide enough information to recover large-scale structures of the subsurface while short-offset data sets may give access to finer-scale variations. As a result, while standard method generates qualitative images of the subsurface discontinuities, a *high resolution quantitative* estimation of the subsurface parameters is expected from FWI. The theoretical resolution reaches $\lambda/2$, where λ is the local wavelength of the recorded seismic waves (Wu and Toksöz 1987).

In addition, the presence of complex structures such as salt domes, sub-basalt targets, thrustbelts, or foothills is known to prevent this standard two-steps workflow to provide good quality subsurface images, or at the expense of intensive human interventions. Due

to the presence of these structures, reflection travel-time tomography is unable to provide automatically correct estimations of the wave velocity background. As a consequence, the migration of the seismic data fails to correctly focus the energy on the subsurface discontinuities and the quality of the migrated images that are obtained is often poor. FWI is expected to overcome this difficulty.

The formalism of the FWI method has been introduced by Lailly (1983) and Tarantola (1984), based on a time domain discretization of the wave equation. Its first application to 2D synthetic data in the acoustic approximation was performed by Gauthier *et al.* (1986). Later on, a hierarchical frequency domain approach has been introduced by Pratt for cross-hole tomography (Pratt and Worthington 1990; Pratt 1990). During the past ten years, the simultaneous advances in acquisition systems (development of wide-azimuth seismic surveys for instance) and high performance computing facilities have made possible the successful application of FWI to surface data, both in the 2D acoustic or 2D elastic approximation (Operto *et al.* 2004, 2005; Ravaut *et al.* 2004; Gao *et al.* 2006; Brossier *et al.* 2009; Prieux *et al.* 2011; Plessix *et al.* 2012). More recently, application of FWI to real surface data in the 3D acoustic approximation has also been performed (Sirgue *et al.* 2008; Plessix and Perkins 2010; Vigh *et al.* 2010). For a detailed overview of the FWI method and its application, the reader is referred to the review paper of Virieux and Operto (2009).

Despite these promising results, FWI is subject to one important limitation: the method requires a sufficiently accurate initial model to converge to the desired model. The misfit function which is minimized often contains several local minima, and the optimization schemes which are employed only guarantee the convergence to the nearest local minimum from the starting point. This is the expression of the nonlinear and undetermined nature of the seismic imaging inverse problem. Indeed, even if the forward problem is linear for the wavefield estimation, it depends nonlinearly on the subsurface parameters. For this reason, the seismic inverse problem is strongly nonlinear. In addition, the illumination of the subsurface is incomplete, both in terms of the acquisition system geometry (surface, streamer,

cross-hole) and frequency content (lack of low and high frequencies). This is responsible for the indetermination of the FWI problem.

In the context of wave velocity estimation by FWI, the convergence to local minima can be interpreted physically as the cycle skipping problem: the kinematic information contained in the starting model is not valid and the recorded data is accounted for up to one or several phases. Hierarchical approaches are defined to mitigate this difficulty: the data can be interpreted from the lowest available frequencies to the highest (Bunks *et al.* 1995; Sirgue and Pratt 2004) in addition to time-windowing techniques, which account the data from early seismic events to later ones (Brenders and Pratt 2007; Brossier *et al.* 2009).

The strong nonlinearity of the FWI problem claims for a careful understanding of the optimization routines which are employed to minimize the misfit function. The state-of-the-art methods used in the context of FWI are gradient-based algorithms, such as the steepest descent or the nonlinear conjugate-gradient method. From a given initial model, the sequence of updates yielding the final model is defined by the gradient of the misfit function. However, it is well known that these methods have poor convergence properties; in practice they may even fail to converge to the desired subsurface parameter estimation, depending on the complexity of the application case.

Conversely, Newton-based methods possess better convergence properties (superlinear to quadratic convergence rate). These methods are based on a descent direction given by the gradient of the misfit function multiplied by the inverse Hessian operator *. This operator provides useful information on the local curvature of the misfit function and can be used to generate a more accurate descent direction.

Pratt *et al.* (1998) emphasizes the importance of this operator in the context of FWI. The inverse Hessian operator acts as a deconvolution operator that accounts for the limited bandwidth of the seismic data and corrects for the loss of amplitude of poorly illuminated subsurface parameters. In addition, it helps to remove artifacts that the second order reflected waves may generate on the gradient descent direction. This analysis suggests that it

*The Hessian is the matrix of the second-order derivatives of the misfit function.

should be extremely crucial to account accurately for the inverse Hessian operator within the minimization schemes employed to solve the FWI problem.

However, because of the large-scale aspect of the FWI problem, which easily involves hundreds of thousands discrete unknowns in 2D up to billions discrete unknowns in 3D, explicit computation of the inverse Hessian operator is beyond current computational capabilities. As a consequence, several attempts for accounting for the inverse Hessian operator have been developed in the recent years.

A first possibility consists in approximating the diagonal of the Hessian and using its inverse as a preconditioner for the steepest-descent method or the nonlinear conjugate gradient method. For instance, Operto *et al.* (2006) compute the diagonal terms of the Gauss-Newton approximation of the Hessian, which requires some extra-computation. A cheaper strategy based on the so-called pseudo-Hessian operator is also proposed by Shin *et al.* (2001).

A second possibility consists in approximating the inverse Hessian operator by finite differences of previous values of the misfit gradient. Among this class of methods, known as quasi-Newton methods, the l-BFGS method is the most popular (Nocedal and Wright 2006; Byrd *et al.* 1995). Instead of approximating only the diagonal elements, a positive definite approximation of the full inverse Hessian is computed. In the framework of 2D elastic FWI, Brossier *et al.* (2009) implemented a l-BFGS optimization using a diagonal pseudo-Hessian as initial guess, showing good convergence properties compared to preconditioned non-linear conjugate-gradient.

An alternative to these approximation methods, consists in computing the descent direction as the solution of the linear system associated with the definition of the Newton descent direction, following a truncated Newton approach (Nash 2000). A "truncation" is performed in the sense that the linear system is not solved exactly: an approximation of the solution is rather computed using an iterative linear solver. Implemented in a "matrix-free" fashion, this iterative solver only requires to compute Hessian-vector products. The Hessian operator has not to be formed explicitly.

Although this class of methods is well known in the numerical optimization community,

the application of truncated Newton method in the FWI context has still not been fully investigated. Given the importance of the inverse Hessian operator in the FWI reconstruction scheme, we believe that this method could, in some specific cases, benefit from a better approximation of the inverse Hessian effect, and provides more accurate subsurface parameter estimations than standard optimization schemes. Therefore, the ambition of this study is to focus on the two following points:

- Designing a feasible implementation of the truncated Newton method for FWI in terms of computational time;
- Investigateing in which situations the truncated Newton method better accounts for the inverse Hessian operator and enhances the subsurface model estimation, compared to conventional optimization schemes.

In Section 2, we present a general overview of the standard minimization algorithms that are used to solve the FWI problem. We also precise the role of the inverse Hessian operator in the FWI scheme and we present the principle of the truncated Newton method. In Section 3, we describe how this method can be efficiently implemented in the context of FWI. In Section 4, we compare the efficiency of the truncated Newton method with gradient-based and quasi-Newton methods on three case studies. These tests are performed in the context of the 2D acoustic frequency domain FWI. The first one derives from a near-surface geotechnical imaging problem. The second one is the standard Marmousi II benchmark model. The third one is based on the 2004 BP benchmark model, partly inspired from the deep water Mexico Gulf subsurface geology. A conclusion and perspectives are given in Section 5.

2 NONLINEAR MINIMIZATION SCHEMES FOR SOLVING THE FWI PROBLEM

For the sake of clarity, the mathematical results presented in this section are formulated for a number of sources equal to 1. The extension to a multi-source context is straightforward.

2.1 Problem settings

We consider the frequency-domain forward problem

$$S(p)u = \varphi \tag{1}$$

where

- $p \in \mathcal{M}$ denotes the subsurface parameters;
- $u \in \mathcal{W}$ is the complex-valued seismic wavefield;
- φ is a source term;
- $S(p)$ is a partial differential operator related to the wave equation (from the acoustic case to the visco-elastic anisotropic dynamics).

The FWI problem is defined as the minimization over the parameter space of a distance between the data predicted by the forward problem and the recorded data.

$$\min_{p \in \mathcal{M}} f(p) = \frac{1}{2}\|Ru(p) - d\|^2, \tag{2}$$

where

- $u(p)$ is the solution of the forward problem (1) for the source term φ and the subsurface parameter p;
- R is a mapping of the wavefield to the receivers locations;
- d is the data set associated to the source φ;
- $\|.\|$ is a norm in the data space $\mathcal{D}$.

For practical reasons, the use of the L^2 norm is common. However, more general L^p norm could be also selected (Tarantola 2005). The L^1 norm is, for instance, a good choice when high-amplitude noise (outliers) corrupts the data (Brossier *et al.* 2010). More complex measurements of the distance between data sets can also be proposed to mitigate the sensitivity of FWI to the initial model. This is, however, beyond the scope of the work presented here.

2.2 Standard minimization methods

From a numerical point of view, FWI is a large-scale nonlinear minimization problem. The high number of discrete parameters prevents from using global or semi-global optimization techniques to solve this problem. Therefore, we focus on local optimization methods, which are based on the following recurrence: from an initial guess p_0, a sequence p_k is computed such that

$$p_{k+1} = p_k + \gamma_k \Delta p_k, \tag{3}$$

where Δp_k is the descent direction and γ_k is a scalar parameter computed through a line-search or a trust-region procedure (Bonnans *et al.* 2006; Nocedal and Wright 2006). The minimization schemes we consider differ from the formula used to compute Δp_k.

2.2.1 Gradient based-methods

The simplest local optimization method is the steepest-descent, for which Δp_k is computed as the opposite of the gradient of the misfit function

$$\Delta p_k^{SD} = -\nabla f(p_k). \tag{4}$$

This method is, however, known to converge only linearly to the nearest local minima. The nonlinear conjugate gradient is usually preferred, since it generally presents better convergence properties. It is based on the definition of Δp_k such that

$$\begin{cases} \Delta p_0^{CG} = -\nabla f(p_0), \\ \Delta p_k^{CG} = -\nabla f(p_k) + \beta_k \Delta p_{k-1}^{CG}. \end{cases} \tag{5}$$

The quantity β_k depends on $\nabla f(p_k)$ and $\nabla f(p_{k-1})$. Different formulas can be chosen to compute β_k, each defining a particular nonlinear conjugate gradient method: the Fletcher-Reeves and the Polak-Ribiere methods are the most common choices, although other formulations have been proposed more recently (Nocedal and Wright 2006).

Both the steepest-descent and the nonlinear conjugate gradient methods are gradient-based methods: the role played by the inverse Hessian operator in the reconstruction of the subsurface parameters is neglected. It is, however, possible to incorporate this information

into these two optimization schemes by using a preconditioning matrix P. Formulas (4) and (5) then become respectively

$$\Delta p_k^{PSD} = -P\nabla f(p_k), \tag{6}$$

and

$$\begin{cases} \Delta p_0^{PCG} = -P\nabla f(p_0) \\ \Delta p_k^{PCG} = -P\nabla f(p_k) + \beta_k \Delta p_{k-1}^{PCG} \end{cases} \tag{7}$$

In the context of FWI, an interesting preconditioning technique is introduced by Shin *et al.* (2001). This method consists in setting P as the inverse diagonal of the so-called pseudo-Hessian operator. This approximation of the Hessian operator can be evaluated using only the diffraction matrices $\partial_{p_j} S(p)$ related to the subsurface parameters and the wavefield $u(p)$. Since the computation of the gradient requires to compute $u(p)$, the construction of the pseudo-Hessian matrix does not induce any extra-computation cost once the gradient is known. This defines a first strategy to account for the influence of the inverse Hessian operator in the inversion scheme, at a reasonable computation cost. As this introduction of an approximation of the inverse hessian as a preconditioner may appear slightly artificial, we would like to emphasize that this strategy is actually intimately related to the Newton algorithm.

2.2.2 Quasi-Newton methods

Within the framework of Newton algorithms, the increment Δp_k is defined by

$$H(p_k)\Delta p_k^N = -\nabla f(p_k), \tag{8}$$

where $H(p_k)$ is the Hessian operator. However, the explicit computation of the inverse Hessian operator $H(p_k)^{-1}$ is far beyond the current computational capabilities, even for 2D applications. On this basis, the class of quasi-Newton methods is defined. These algorithms are based on an approximation Q_k of the inverse Hessian operator, and the descent direction Δp_k is computed from

$$\Delta p_k^{QN} = -Q_k \nabla f(p_k). \tag{9}$$

We recognize in formula (9) the definition of the model update Δp_k^{PSD} (6) associated with the preconditioned steepest-descent method. The matrix Q_k plays the role of the preconditioning matrix P: the introduction of an approximation of the inverse Hessian as a preconditioner for gradient-based methods originates naturally from the Newton method.

The method employed to compute the approximation matrix Q_k defines a variety of quasi-Newton methods. Among several possibilities, the l-BFGS method has become one of the most popular. It proposes a systematic approximation of $H(p_k)^{-1}$ through finite differences of l previous values of the misfit gradient $\nabla f(p_{k-l+1}), \dots, \nabla f(p_k)$ (Byrd *et al.* 1995). While the approximation proposed by Shin *et al.* (2001) is only applicable in the context of surface-data FWI and comes from physical interpretation, the l-BFGS formula is general and depends only on stored values of the misfit gradient.

The l-BFGS method has proven very useful in numerous large-scale applications (S. G. Nash 1991). This method can also be seen as a generalization of the nonlinear conjugate gradient method. Indeed, the nonlinear conjugate gradient algorithm computes the descent direction as a linear combination of the current and the previous gradient values $\nabla f(p_k)$ and $\nabla f(p_{k-1})$, while the l-BFGS methods approximates $H(p_k)^{-1}$ using l previous values of the gradients. The link between the l-BFGS method and the nonlinear conjugate gradient is made explicit in Nocedal and Wright (2006).

We have seen that both preconditioned gradient-based methods and quasi-Newton methods use approximation of the inverse Hessian. From a numerical optimization point of view, this operator gathers information on the local curvature of the misfit function, and helps to produce a more accurate descent direction than the one given by the opposite of the gradient. Within the context of FWI, a physical interpretation of the inverse Hessian operator can also be given.

2.3 The inverse Hessian operator

From the definition of the misfit function $f(p)$, we can write $H(p)$ as

$$H(p) = B(p) + C(p), \tag{10}$$

where

$$\begin{cases} B(p) = \mathcal{R}\left(J^\dagger(p) R^\dagger R J(p)\right), \\ C(p) = \mathcal{R}\left(R^\dagger \left(Ru(p) - d\right) \partial_{pp} u(p)\right). \end{cases} \tag{11}$$

In this expression, $J(p) = \partial_p u(p)$ is the Jacobian matrix of the wavefield $u(p)$, $\partial_{pp} u(p)$ represent the second-order derivatives of the wavefield $u(p)$, the symbol $\dagger$ denotes the conjugate transpose operator, and $\mathcal{R}$ denotes the real part operator. The matrix $B(p)$ is known as the Gauss-Newton approximation of the Hessian.

Assuming the system is discretized, denoting by p_j the discretized parameters, N_r the total number of receivers, and x_r their locations, a coefficient ij of the matrix $B(p)$ is

$$B_{ij} = \mathcal{R}\left(\sum_{r=1}^{N_r} \partial_{p_i} u(x_r) \partial_{p_j} u(x_r)\right). \tag{12}$$

The coefficient B_{ij} appears as the zero-lag correlation of $\partial_{p_i} u(x_r)$ and and $\partial_{p_j} u(x_r)$, the first-order derivatives with respect to p_i (respectively p_j) of the wavefield $u(p)$, recorded at the receiver locations. As a consequence, B_{ij} decreases with the distance which separates the parameters p_i and p_j and reaches its maximum for the autocorrelation of the two derived wavefields ($i = j$). In the high-frequency approximation, the zero-lag correlation of the derivatives of the wavefield with respect to two different parameters would be zero and $B(p)$ would be diagonal. However, the frequency content of the seismic data is limited and the matrix $B(p)$ is banded. In addition , the values of the first-order derivatives wavefields $\partial_{p_i} u$ and $\partial_{p_j} u$ decrease for parameters for which the wavefield is less sensitive. Therefore, filtering the gradient $\nabla f(p)$ with the matrix $B(p)^{-1}$ acts as a refocusing filter. The amplitude of poorly illuminated parameters is compensated in the model update given by the gradient perturbation.

In the same way, in discrete form, a coefficient ij of $C(p)$ is given by

$$C_{ij} = \mathcal{R}\left(\sum_{r=1}^{N_r} \left(u(x_r) - d(x_r)\right) \partial_{p_i p_j} u(x_r)\right). \tag{13}$$

The expression (13) shows that C_{ij} is the zero-lag correlation of the differences between the predicted and the recorded data at the receivers (also named residuals) and the second-order derivatives of the wavefield recorded at the receivers, with respect to the parameters

p_i and p_j. This second-order wavefield is also called double scattered wavefield: it represents a recorded signal that has been scattered twice, at the locations of the parameter p_i and p_j. In the presence of high amplitude multi-reflected waves, this double scattered wavefield generates strong artifacts on the gradient $\nabla f(p_k)$ for the optimization procedure, which only accounts for single scattered waves. Therefore, applying the matrix $C(p)^{-1}$ to the gradient may allow to compensate the artifacts generated by double scattered waves for the inversion.

This short analysis summarizes the results given by Pratt *et al.* (1998). It emphasizes the crucial importance of the inverse Hessian operator in the FWI reconstruction scheme: the descent direction given by the gradient should be scaled so as to restore the correct amplitude to the elements of the model update; the artifacts generated by high amplitude second-order reflections should be removed. As a consequence, we are interested in applying another kind of optimization to the FWI problem, named truncated Newton methods, that could possibly account more accurately for these effects.

2.4 The truncated Newton method

The truncated Newton method is an alternative to the minimization schemes presented in section 2.2. Instead of building an approximation Q_k of $H(p_k)^{-1}$, the linear system (8) is solved using a matrix-free version of the conjugate gradient algorithm (Saad 2003). This only requires the capability of computing Hessian-vector products $H(p_k)v$ where v is an arbitrary vector in the model space $\mathcal{M}$. The truncated Newton method thus results in a two nested loops algorithm:

- the external loop consists in the iterative update of the initial subsurface parameter estimation, following equation (3);
- the internal loop consists in the iterative resolution of the linear system (8), in order to compute the model update Δp_k.

Compared to the quasi-Newton methods, the truncated Newton method involves a computational expense associated with the resolution of the linear system (8). However, this additional expense should be balanced by an improvement of the convergence speed of the

external loop. Therefore, as mentioned by Nash (2000), an efficient implementation of the truncated Newton method relies on the reduction of this extra cost. This can be achieved in the context of FWI by

- defining second-order adjoint formulas for the efficient computation of Hessian vector products $H(p_k)v$ for any $v \in \mathcal{M}$;
- defining an adapted stopping criterion for the resolution of the linear system (8) to limit as much as possible the number of iteration of the conjugate gradient algorithm at each step of the external loop;
- using an appropriate preconditioner to accelerate the convergence of the resolution of the linear system (8).

The truncated Newton strategy can also be implemented using the Gauss-Newton $B(p)$ approximation of the Hessian operator $H(p)$. The Gauss-Newton optimization is a well known method designed for large-scale nonlinear least-squares problems. Within this framework, the linear system (8) is replaced by the linear system

$$B(p_k)\Delta p_k = -\nabla f(p_k), \tag{14}$$

where $B(p)$ is defined by the equation (11). This method may seem appealing for the following reasons:

- $B(p)$ is a positive definite matrix by construction, therefore the conjugate gradient algorithm is well adapted for the resolution of the linear system (14);
- close to the solution, the matrix $B(p)$ should be a good approximation of the full Hessian matrix $H(p)$, since the matrix $C(p)$ depends on the amplitude of the residuals.

From the implementation point of view, the truncated Gauss-Newton procedure differs only in the computation of matrix-vector products $B(p)v$ instead of $H(p)v$ from the truncated Newton method.

While preconditioned gradient-based methods and quasi-Newton methods use rough or sparse approximations of the inverse Hessian operator, truncated Newton and Gauss-Newton method may account more accurately for the inverse Hessian operator even if it is not formed

explicitly. Instead, the approximation is transferred to the truncation strategy, which is related to the definition of a particular stopping criterion.

In the next section, we investigate how the truncated Newton and Gauss-Newton methods can be efficiently implemented in the context of FWI.

3 IMPLEMENTATION OF THE TRUNCATED NEWTON METHOD

3.1 Linear systems resolution

The resolution of the linear systems (8) and (14) first requires to compute the right-hand side $-\nabla f(p_k)$. The computation of $\nabla f(p_k)$ is efficiently achieved through the well known first-order adjoint-state method, introduced by Lions (1968). For the sake of simplicity, we assume in the sequel that the subsurface parameter p is discretized in $\mathbb{R}^m$, $m \in \mathbb{N}$.

3.1.1 Computation of the gradient and first-order adjoint method

In the context of FWI, the first-order adjoint method amounts to compute $\nabla f(p_k)$ as the zero-lag cross-correlation of the incident wavefield and the adjoint wavefield. The adjoint wavefield, denoted by $\lambda \in \mathcal{W}$, is computed through the backpropagation of the residuals. It is defined as the solution of

$$S(p)^\dagger \lambda = R^\dagger (d - Ru(p)), \tag{15}$$

where $u(p)$ is the solution of (1). Based on the definition of the adjoint state λ, the component i of the gradient can be computed from the equation

$$\nabla f(p)_i = \mathcal{R}(\partial_{p_i} S(p) u(p), \lambda(p))_{\mathcal{W}}, \quad i = 1, \dots, M, \tag{16}$$

where $(.,.)_{\mathcal{W}}$ is the scalar product in $\mathcal{W}$ associated with the choice of the norm $\|.\|$. Using this method, the computation cost of the gradient amounts to the resolution of two wave propagation problems: one forward problem (1) and one adjoint problem (15). A survey of the application of the first order adjoint-state method to seismic imaging is proposed by Plessix (2006).

Since the right hand side of (8) and (14) can be efficiently computed through this method,

the use of a matrix-free conjugate gradient for the resolution of these systems only requires Hessian-vector products $H(p)v$ (respectively $B(p)v$) to be computed. This is achieved through second-order adjoint methods.

3.1.2 Computation of $H(p)v$ through second-order adjoint method

The computation of Hessian-vector products through second-order adjoint methods is a topic that has already been investigated in the field of data assimilation and weather forecasting (Le Dimet *et al.* 2002). However, the control variable in data assimilation is an initial condition for the system, whereas in seismic imaging, the control variable is a coefficient of the partial differential equation that describes the system. A formula for the computation of Hessian-vector products have been given by Pratt *et al.* (1998) in the seismic imaging context, for the Gauss-Newton approximation in the discrete frequency domain. Fichtner and Trampert (2011) propose more general formulas for the computation of Hessian kernels. Epanomeritakis *et al.* (2008) give the formulas corresponding to the elastic case.

We propose here a general framework in the frequency domain for deriving these formulas, with no assumption on the discretization and the kind of partial differential equations that are used for the wave propagation description. The method can be straightforwardly adapted to the time-domain formulation by adding proper initial and final conditions, and boundary conditions. Based on the definition of a new Lagrangian function, we derive an algorithm to compute, for given $(v, p) \in \mathcal{M}^2$, the matrix-vector product $H(p)v$. We first define the functional $g_v(p)$ as

$$g_v(p) = (\nabla f(p), v)_{\mathcal{M}}, \tag{17}$$

where $(.,.)_{\mathcal{M}}$ denotes the scalar product on the parameter space $\mathcal{M}$. By definition, the gradient of the functional $g_v(p)$ is

$$\nabla g_v(p) = H(p)v. \tag{18}$$

A direct derivation of the misfit function $f(p)$ gives the following formula for $\nabla f(p)$ as

$$\nabla f(p) = \mathcal{R}\left(J^\dagger(p) R^\dagger \left(Ru(p) - d\right)\right). \tag{19}$$

Note that the computation of the misfit gradient through formula (19) involves the explicit computation of the Jacobian matrix $J(p)$ which is not adapted to the large-scale aspect of the problem. The first-order adjoint state method avoids this computational burden.

Using (19), $g_v(p)$ can be rewritten as

$$g_v(p) = (R^\dagger(Ru(p) - d), J(p)v)_{\mathcal{W}}. \tag{20}$$

Deriving the forward problem with respect to the parameters p_j in the directions v_j yields

$$\left(\partial_{p_j} S(p).v_j\right) u + S(p)\left(\partial_{p_j} u.v_j\right) = 0, \quad j = 1, \dots, m. \tag{21}$$

Summing on j gives

$$\sum_{j=1}^{m} \left(\partial_{p_j} S(p).v_j\right) u + \sum_{j=1}^{m} S(p)\left(\partial_{p_j} u.v_j\right) = 0, \tag{22}$$

which is equivalent to

$$S(p)(J(p)v) = \Phi_v(p, u), \tag{23}$$

where

$$\Phi_v(u) = -\sum_{j=1}^{m} (\partial_{p_j} S(p).v_j)u. \tag{24}$$

The expression $J(p)v \in \mathcal{W}$ is denoted by $\overline{\alpha}_v(p)$ and is the solution of the forward problem (1) for the source term $\Phi_v(p, u)$. Thus, we may consider the constrained minimization problem

$$\min_p (R^\dagger(Ru - d), \alpha)_{\mathcal{W}}, \quad s.\ t.\ \ S(p)u = \varphi, \quad S(p)\alpha = \Phi_v. \tag{25}$$

Introducing the adjoint variables $(\lambda, \mu) \in \mathcal{W}^2$, the Lagrangian function associated with the problem (25) is

$$\begin{aligned} L_v(p, u, \alpha, \lambda, \mu) \;=\; & \mathcal{R}((R^\dagger(Ru - d), \alpha)_{\mathcal{W}}) + \\ & \mathcal{R}\left((S(p)u - \varphi, \mu)_{\mathcal{W}}\right) + \\ & \mathcal{R}\left((S(p)\alpha - \Phi_v, \lambda)_{\mathcal{W}}\right). \end{aligned} \tag{26}$$

Thus, for $\overline{u}(p)$ solution of (1) and $\overline{\alpha}_v(p)$ solution of (23), we have

$$L_v(p, \overline{u}(p), \overline{\alpha}_v(p), \lambda, \mu) = g_v(p), \tag{27}$$

and

$$\partial_p L_v(p, \overline{u}(p), \overline{\alpha}_v(p), \lambda, \mu) = \nabla g_v(p). \tag{28}$$

If we define the adjoint states $\overline{\lambda}(p)$ and $\overline{\mu}(p)$ such that

$$\begin{cases} \partial_u L_v(p, \overline{u}(p), \overline{\alpha}_v(p), \overline{\lambda}(p), \overline{\mu}(p)) = 0 \\ \partial_\alpha L_v(p, \overline{u}(p), \overline{\alpha}_v(p), \overline{\lambda}(p), \overline{\mu}(p)) = 0, \end{cases} \tag{29}$$

which is equivalent to

$$\begin{cases} S(p)^\dagger \overline{\mu} = -R^\dagger R \overline{\alpha}_v - \sum_{j=1}^{m} (\partial_{p_j} S(p).v_j)^\dagger \overline{\lambda} & (30) \\ S(p)^\dagger \overline{\lambda} = R^T (d - R\overline{u}(p)), & (31) \end{cases}$$

we obtain the three-terms Hessian-vector product formula

$$\begin{aligned} H(p)v \;\; = \;\; & \mathcal{R}\left((\partial_p S(p)\overline{u}(p), \overline{\mu}(p))_{\mathcal{W}}\right) + \\ & \mathcal{R}\left((\partial_p S(p)\overline{\alpha}_v(p), \overline{\lambda}(p))_{\mathcal{W}}\right) + \\ & \mathcal{R}\left(\sum_{j=1}^{m} v_j \left(\left(\partial_{p_j}\partial_p S(p)\right)\overline{u}(p), \overline{\lambda}(p)\right)\right). \end{aligned} \tag{32}$$

For a given $v \in \mathcal{M}$, and a given subsurface model $p \in \mathcal{M}$, the computation of the Hessian-vector product $H(p)v$ requires to solve four wave propagation problems: two forward problems for the computation of $\overline{u}(p)$ and $\overline{\alpha}_v(p)$, two adjoint problems for the computation of $\overline{\lambda}(p)$ and $\overline{\mu}(p)$. During the process, the Hessian matrix $H(p)$ is never computed explicitly. Note the identity between the adjoint state $\overline{\lambda}(p)$ derived here and the one which is derived for the computation of the misfit gradient through the first-order adjoint state formula.

3.1.3 Computation of $B(p)v$

The product $B(p)v$ can be written as

$$B(p)v = J(p)^\dagger R^T R J(p) v. \tag{33}$$

We consider the constrained minimization problem

$$\min_p g_w(p) = (u(p), w)_{\mathcal{W}} \;\; \textit{subject to} \;\; S(p)u = \varphi, \tag{34}$$

for an arbitrary $w \in \mathcal{W}$. Note that

$$\nabla g_w(p) = J(p)^\dagger w. \tag{35}$$

The Lagrangian function associated with this problem is

$$L_w(p, u, \nu) = (u, w) + \mathcal{R}(S(p)u - \varphi, \nu), \tag{36}$$

where $\nu \in \mathcal{W}$ is a new adjoint variable. From the first-order adjoint state method, we can state that

$$\nabla g_w(p) = \mathcal{R}(\partial_p S(p)\overline{u}(p), \overline{\nu}(p)), \tag{37}$$

where

$$S(p)^T \overline{\nu} = -w. \tag{38}$$

The formula (37) gives $J^\dagger(p)w$ for any $w \in \mathcal{M}$. Replacing w by $R^T R J(p)v$ in equation (38) thus yields the formula for the computation of $B(p)v$. Define $\overline{\xi}(p) \in \mathcal{W}$ such that

$$S(p)^T \overline{\xi} = -R^T R \overline{\alpha}_v(p), \tag{39}$$

where $\overline{\alpha}_v(p)$ is defined by (23), we have

$$B(p)v = \mathcal{R}\left(\partial_p S(p)\overline{u}(p), \overline{\xi}(p)\right). \tag{40}$$

As a consequence, for a given $v \in \mathcal{M}$ and a given subsurface model $p \in \mathcal{M}$, the computation of the matrix-vector product $B(p)v$ amounts to solving three wave propagation problems: two forward problems for the computation of $\overline{u}(p)$ and $\overline{\alpha}_v(p)$ and one adjoint problem for the computation of $\overline{\xi}(p)$. Note that

- the computation of $\overline{\xi}(p)$ amounts to setting $\overline{\lambda}(p)$ to 0 in the equation (30);
- the computation of $B(p)v$ amounts to setting $\overline{\lambda}(p)$ to 0 in the equation (32).

In view of these results, it could appear that the Gauss-Newton approximation requires less computational efforts, as only one adjoint problem has to be solved instead of two in the exact Newton case. Nonetheless, in the context of the truncated Newton method, *the two methods are equivalent in terms of computation cost*, as it is demonstrated in the next paragraph.

3.1.4 Computation cost

Consider the resolution of the linear system (8) for the computation of the descent direction Δp_k. The right hand side in (8) is the opposite gradient $-\nabla f(p)$. It is computed using the first-order adjoint state method, which requires to solve one forward problem for $\overline{u}(p)$ and one adjoint problem for $\overline{\lambda}(p)$. Provided these wavefields can be stored, the computation of $H(p)v$ or $B(p)v$ thus only requires to solve two additional problems: one forward problem for the computation of $\overline{\alpha}_v(p)$, and one adjoint problem, either for the computation of $\overline{\mu}(p)$ for $H(p)v$ or $\overline{\xi}(p)$ for $B(p)v$. Even if the computation of $\overline{\lambda}(p)$ is not required for the computation of $B(p)v$, it is imposed by the computation of the right-hand side of the linear system (8). The computation cost of the action of the Hessian operator or its Gauss-Newton approximation on an arbitrary vector is thus *the same* in terms of number of wave equations to be solved.

The overall computation cost is thus given by

$$C = N_{ext}(2 + 2 \times N_{int,k}) \tag{41}$$

where N_{ext} is the total number of iterations of the external loop and $N_{int,k}$ is the number of conjugate gradient iterations performed at the k^{th} iteration of the external loop. The choice of an appropriate stopping criterion for the conjugate gradient helps to reduce the quantities $N_{k,iter}$. Note that this computational cost depends on the ability of storing $\overline{u}(p)$ and $\overline{\lambda}(p)$, which is a reasonable assumption for 2D applications. This is a more complex issue in a 3D context, in which memory and I/O management methods should be required.

3.2 Definition of an adapted stopping criterion

At this stage, it is important to recall that the Newton method is an iterative minimization of local quadratic expansions of the misfit function. Indeed, the resolution of the system (8) amounts to the minimization of the quadratic form

$$q_k(\Delta p) = f(p_k) + (\nabla f(p_k), \Delta p) + \frac{1}{2}(H(p_k)\Delta p, \Delta p). \tag{42}$$

The definition of the stopping criterion is related to the accuracy of these quadratic expansions. If the local quadratic approximation of the misfit function is accurate, a precise

solution of the system (8) should be computed. Conversely, if this approximation is not accurate, computing an exact solution of the system (8) amounts to explore an unwanted zone in the subsurface parameter space.

This idea is exploited by Eisenstat and Walker (1994). They consider a stopping criterion for the CG iterations of the form

$$\|H(p_k)\Delta p_k + \nabla f(p_k)\| \leq \eta_k \|\nabla f(p_k)\|. \tag{43}$$

where η_k is called the forcing term. The role devoted to this forcing term is to account for the accuracy of the local quadratic approximation. When this accuracy increases, η_k should decrease, so as to require to solve the linear system (8) more accurately. Conversely, when the accuracy decreases, η_k should increase, so as to allow a less precise resolution of the linear system (8). This is achieved by defining η_k as the measure of the distance between the first order Taylor expansion of the gradient at the iteration $k-1$ and the gradient at iteration k:

$$\eta_k = \frac{\|\nabla f(p_k) - \nabla f(p_{k-1}) - \gamma_{k-1} H(p_{k-1})\Delta p_{k-1}\|}{\|\nabla f(p_{k-1})\|}. \tag{44}$$

The definition of the stopping criterion is complemented with an appropriate strategy to deal with the detection of negative eigenvalues of the Hessian operator. The conjugate gradient algorithm is designed for the resolution of symmetric definite positive systems. However, far from the solution, the full Hessian operator $H(p)$ may be indefinite. Therefore, during the resolution of the linear system (8) with the conjugate gradient method, the probability of encountering a curvature associated with a negative eigenvalue of the operator $H(p)$ is not negligible. In this case, the linear iterations are stopped and the last value of the descent direction Δp_k which is computed is returned. If this negative curvature is met at the very first linear iteration, the steepest-descent direction is returned. This strategy, proposed by Eisenstat and Walker (1994), ensures superlinear convergence properties far from the solution, and quadratic convergence when entering the attraction basin of the minimum.

3.3 Preconditioning

In order to speed-up the convergence of the resolution of the linear system (8), it is natural to introduce a preconditioning matrix. Different types of preconditioning techniques can be considered, with the following restriction: it is not possible to access to the coefficients of the Hessian operator, since this operator is not computed explicitly. This prevents from using incomplete LU or incomplete Cholesky factorization.

A feasible approach consists in using a l-BFGS approximation of the inverse Hessian operator. This approximation can be updated at each external iteration: in this case the truncated Newton method can be seen as a direct extension of the l-BFGS method, since performing no inner conjugate gradient iterations amounts to selecting the l-BFGS descent direction.

The l-BFGS approximation can also be built during each cycle of inner conjugate gradient iterations. As mentioned in section 3.2, these inner conjugate gradient iterations minimize the quadratic form $q_k(\Delta p)$ defined by the equation (42). Let us remind that the l-BFGS approximation of the Hessian operator associated with a particular function is defined from l estimations of the gradient of this function. In addition, note that the Hessian operator associated with $q_k(\Delta p)$ as a function of Δp is precisely $H(p_k)$. As a consequence, the l-BFGS approximation of $H(p_k)$ can be constructed from l values of the gradient of the quadratic approximation $\nabla q_k(\Delta p)$, which is given by

$$\nabla q_k(\Delta p) = H(p_k)\Delta p + \nabla f(p_k). \tag{45}$$

From an initial guess $\Delta p_0 = 0$, the inner conjugate gradient build a sequence Δp_r that converges to a solution Δp of the linear system (8). At each iteration, the algorithm performs the computation of the residuals associated with Δp_r which are given by

$$H(p_k)\Delta p_r + \nabla f(p_k) = \nabla q_k(\Delta p_r). \tag{46}$$

Therefore, for no extra-cost, the l-BFGS approximation of $H(p_k)$ can be constructed at iteration k from the successive evaluations of $\nabla q_k(\Delta p_r)$. Assuming that the Hessian matrix $H(p_{k+1})$ is "close" in some sense from $H(p_k)$, this approximation is used as a preconditioner

of the linear system (8) at iteration $k+1$. Based on these two approaches, it is even possible to define an interlaced algorithm that alternate between cycles of l-BFGS iterations and cycles of truncated Newton iterations. This strategy is proposed by Morales and Nocedal (2000). Nonetheless, the implementation of this procedure requires to define an efficient criterion for switching from one l-BFGS cycle to a truncated Newton cycle, which can be uneasy in practice.

In this study, we do not use this method and we focus on the special preconditioner related to the FWI problem proposed by Shin *et al.* (2001). The diagonal elements of the Gauss-Newton part of the Hessian $B(p)$ are approximated using the pseudo-Hessian approach, which appears to be relevant for surface seismic. Let us denote α_j the column j of the Jacobian matrix. From equation (21), we see that α_j is the solution of the forward problem

$$S(p)\alpha_j = -\partial_{p_j} S(p)u. \tag{47}$$

An exact computation of the entire Jacobian matrix $J(p)$ would thus require to solve m forward problems, which is intractable from a computational cost point of view. Nonetheless, a cheap approximation can be built by approximating the forward problem operator $S(p)$ as the identity matrix I in the left-hand side of equation (47). This leads to the definition of the pseudo-Hessian matrix entries:

$$\widetilde{H}_{ij}(p) = \left(\left[\partial_{p_i} S(p)u(p)\right]^T \left[\partial_{p_j} S(p)u(p)\right] \right), \quad i,j = 1,\dots,M. \tag{48}$$

The preconditioner used by Shin *et al.* (2001) is defined as

$$P_k = diag\left(\frac{1}{\widetilde{H}_{ii}(p_k)} \right) \quad i = 1,\dots,M. \tag{49}$$

However, because of the fast decrease of the wavefield with depth, very small values appear on the diagonal entries of $\widetilde{H}(p)$ corresponding to deep subsurface parameters. Therefore, using directly P_k as a preconditioner may yield numerical instabilities. We thus introduce a threshold parameter $\theta \in \mathbb{R}$, the constant $C_k \in \mathbb{R}$ such that

$$C_k = \max_j \widetilde{H}_{jj}(p_k), \tag{50}$$

and we define the matrix P_k^θ such that

$$P_k^\theta = diag\left(\frac{1}{\widetilde{H}_{ii}(p_k) + \theta C_k}\right), \quad i = 1, \ldots, M. \tag{51}$$

Finally the norm of the misfit gradient $\nabla f(p)$ should be preserve by the preconditioner*. Therefore, we introduce

$$P_k^{\nu,\theta} = \nu P_k^\theta \tag{52}$$

such that

$$\nu = \frac{\|\nabla f(p_k)\|}{\|P_k^\theta \nabla f(p_k)\|} \tag{53}$$

We use $P_k^{\nu,\theta}$ as a preconditioner of the linear system (8) at each iteration of the external loop. As it is demonstrated in the next section, $P_k^{\nu,\theta}$ can also be used as a preconditioner of the nonlinear conjugate gradient method.

4 CASE STUDIES

4.1 Numerical framework

The numerical tests we present are performed in the 2D frequency domain, using an acoustic modeling of the wave propagation. The density is assumed to be constant, and the subsurface is described by the P-wave velocity v_p. In this context, the forward problem (1) associated with a given angular frequency ω is equivalent to the Helmholtz equation

$$\omega^2 u + \frac{1}{v_p^2}\Delta u = \varphi \tag{54}$$

The equation (54) is discretized with a fourth-order finite difference scheme with a compact stencil (Hustedt *et al.* 2004). Perfectly Matched Layers (PML) (Berenger 1994; Métivier 2009) are used to avoid fictitious reflections on the boundaries of the computation domain. The resolution of (54) reduces to the resolution of a sparse linear system. This is performed

*Preserving the norm of the gradient is actually useful for the comparison of preconditioned and non-preconditioned truncated Newton methods using the same linesearch procedure. In addition, since the stopping criterion for the linear system (8) is based on the reduction of the linear residuals with respect to the norm of the gradient, if appears natural that the preconditioner keeps the gradient norm unchanged.

through a parallel LU factorization using the MUMPS algorithm (Amestoy *et al.* 2000). In addition, the LU factorization of the stiffness matrix associated with the discretization of (54) is reused to solve the adjoint problems. This interesting feature is one of the reason for working in the frequency domain: provided the LU factorization of the stiffness matrix system can be stored, this approach largely reduces the computational costs, compared to the time domain approach. This is especially important when the number of data sets is large: the same LU factorization is used to solve the forward and adjoint problems associated with each data set.

In the following three tests which are presented, based on this wave propagation modeling, an estimation of the P-wave velocity is computed using a FWI scheme. For each of these three tests, we compare the performances of seven minimization strategies, which may be grouped in three classes:

- nonlinear conjugate gradient methods;
- quasi-Newton l-BFGS methods;
- truncated Newton methods;

Only one method belongs to the first class: this is the preconditioned nonlinear conjugate gradient, using preconditioning matrices $P_k^{\nu,\theta}$. This method may be considered as the state-of-the-art method in the FWI context.

The second class contains two methods. The first is the standard l-BFGS method, which uses an approximation of the inverse Hessian matrix through finite differences of l values of the misfit gradient $\nabla f(p)$. This approximation is based on a prior estimation of the inverse Hessian, chosen as the identity matrix. In the following, we shall refer to this method as the standard l-BFGS method. A more accurate prior approximation may improve the performance of the l-BFGS method. As mentioned by Nocedal and Wright (2006), the l-BFGS method provides the required framework to choose a particular prior estimation at each nonlinear iteration. Therefore, the preconditioning matrices $P_k^{\nu,\theta}$ can be incorporated in the l-BFGS procedure as prior estimations of the inverse Hessian. This is implemented

in the second method of this class. We shall refer to this method as the improved l-BFGS method in the following.

The third family contains four methods. The two first correspond to the truncated Newton method using the Gauss-Newton approximation or the full Hessian matrix without any preconditioning. The two last methods correspond to the preconditioned versions of these two algorithms, using the matrices $P_k^{\nu,\theta}$.

For each case study, a trial-and-error approach is used to determine the optimal values for θ. Note that the choice of θ is also related to the depth of investigation of each case study. Pragmatical values range from 10^{-1} to 10^{-5}. In the applications presented here, θ varies between 10^{-2} (first case study) and 10^{-5} (Marmousi II case study).

For the comparison of all the minimization methods to make sense, we implement our own version of each, *using the same linesearch method*. This specific part of the algorithms is dedicated to the computation of the coefficient γ_k in equation (3). The linesearch algorithm we implement satisfies the Wolfe conditions to ensure the global convergence of our algorithms toward a local minimum (Bonnans *et al.* 2006). This implies to use a particular form of the nonlinear conjugate gradient method. Indeed, as mentioned by Nocedal and Wright (2006), Fletcher-Reeves and Polak-Ribiere versions of the nonlinear conjugate gradient algorithm requires to implement a linesearch algorithm satisfying the *strong* Wolfe conditions to guarantee global convergence toward a local minimum. Therefore we select the version of Dai and Yuan (1999), which is compatible with a linesearch procedure that only enforces the Wolfe condition. This amounts to select β_k in formula (7) as

$$\beta_k = \frac{\nabla f(p_k)^T P_k^{\nu,\theta} \nabla f(p_k)}{\left(\nabla f(p_k) - \nabla f(p_{k-1}\right)^T \Delta p_{k-1}}. \tag{55}$$

The computational efficiency of the inversion schemes is compared in terms of the required number of forward problem resolutions per sources. This gives a better insight of the computational cost associated with the six methods than the comparison of either

- the number of nonlinear iterations: methods based on the truncated Newton approach

perform significantly less nonlinear iterations, but each of these requires a far more important computation effort;

- the overall computation time: the implementation of the minimization algorithms have not been optimized in the same way, and the computation time depends on these implementation details.

For the three case studies, we implement the following stopping criterion: the iterations end as soon as

$$f(p_k)/f(p_0) < \epsilon. \tag{56}$$

The quantity ϵ is set to 10^{-4} for the two first experiment and 10^{-3} for the third. In addition, if no acceptable steplength is found after 20 linesearch iterations, the minimization is stopped and an error flag is returned. This situation occurs when a local minimum is reached.

In our implementation of the truncated Newton algorithm, we complement the Eisenstat stopping criterion by setting to 30 the maximum number of inner iterations that can be performed. If the conjugate gradient has not converged, the descent direction computed at the 30^{th} iteration is returned.

Finally, note that for the 3 tests that follow, the memory parameter l for the two l-BFGS methods, which corresponds to the number of gradient stored to compute the sparse approximation of the inverse Hessian, is set to $l = 20$. This rather large value is chosen to produce a reasonably accurate l-BFGS approximation of the Hessian operator. However, performing the same tests with $l = 5$ or $l = 40$ has led us to the same conclusions.

4.2 A geotechnical application

4.2.1 Presentation

The first case study we consider comes from a geotechnical problem: two concrete foundations are buried in the subsurface at few meters depth. We aim at detecting and correctly imaging these structures. The exact P-wave velocity model is presented in figure 1. Note that the ratio between vertical and horizontal distances is not respected to improve the readability of

the model figures. The exact model is composed of a homogeneous background (300 m.s^{-1}) and two superposed concrete structures (4000 m.s^{-1}). In addition, a layer at 500 m.s^{-1} is located at the bottom of the model. The depth of investigation is limited to 3 m, and the width of the exact model is 30 m. We use a discretization step $h = 0.15$ m, which amounts to 5025 discrete unknown parameters. A 10 points width PML layer surrounds the model.

The acquisition system is composed of one line of sources/receivers located at the surface, and two lines of sources/receivers respectively located in two wells on each of the lateral side of the domain. The sources and receivers are placed each 0.15 m, excepted at the surface: when the acquisition line crosses the concrete structures, the sources and receivers are removed.

The initial model is composed of the exact homogeneous background model at 300 m.s^{-1} and the bottom layer at 500 m.s^{-1} (Fig. 1).

Using the exact model, we compute nine data sets, from 100 Hz to 300 Hz, with a 25 Hz sampling. We invert simultaneously these nine data sets: the misfit function is computed as the sum of the misfit between the predicted and recorded data. An incremental strategy may be also used. This would consist in inverting each data set one by one starting by the lowest frequency data, and using the inversion result of one data set as a starting model for inverting the next data set. This hierarchical approach can also be applied to subgroups of data sets. Such a strategy can be particularly useful in some cases to mitigate strong nonlinearities of the misfit function (Sirgue and Pratt 2004). In this particular case, we have experimented that performing a hierarchical inversion of monofrequency data does not provide satisfactory results: several data sets are required to provide enough information for the method to converge. Conversely, we have not noticed any differences between the incremental inversion of groups of data sets and the simultaneous inversion of the data sets. This is the reason why we perform a simultaneous inversion of the data.

In figure 2, two common-shot gathers in the time domain are presented. These shot-gathers give a better insight of the complexity of the data associated with this subsurface configuration. The first shot-gather is associated with the exact model. The second one is

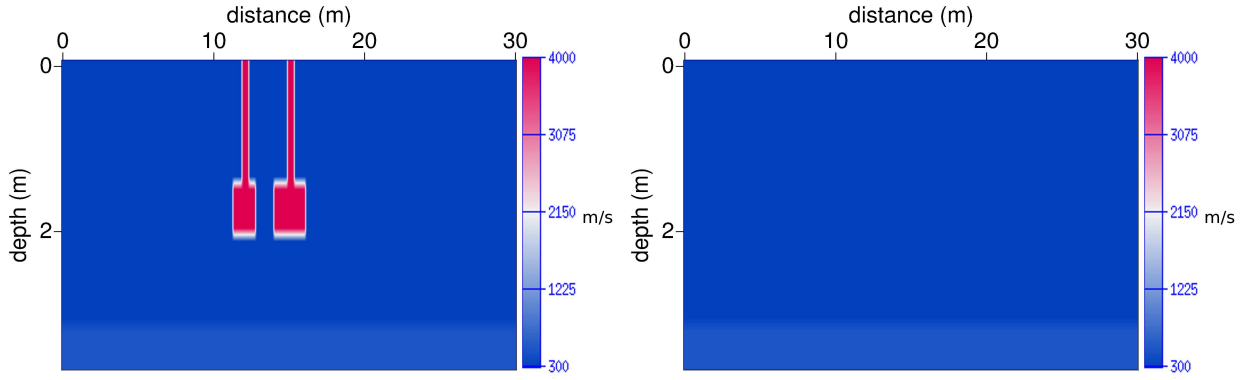

Figure 1. Exact P-wave velocity model (top) and its corresponding initial guess (bottom)

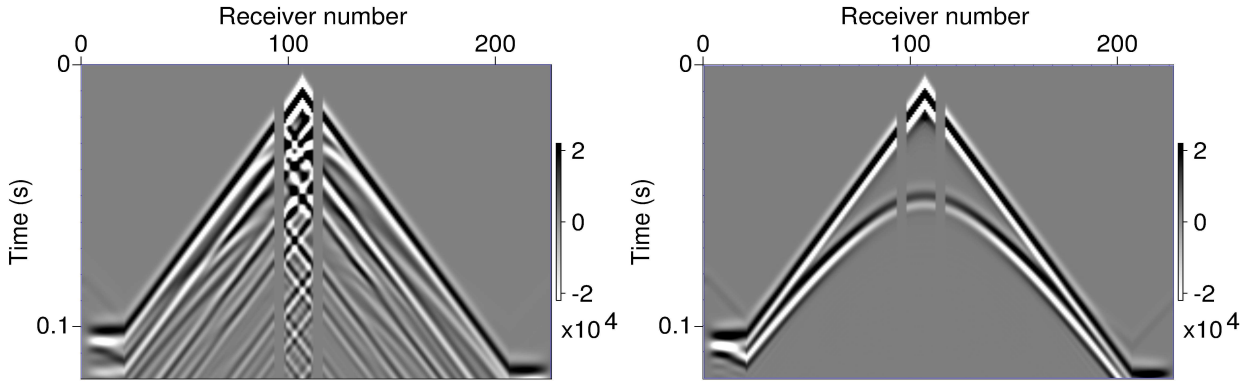

Figure 2. Common Shot Gather for the exact P-wave velocity domain (top) and the initial model (bottom). Please note the two vertical acquisition at both sides of the figures.

associated with the initial model. These time-domain data are generated using a Ricker type source, with a frequency bandwidth centered around 100 Hz, and located at the surface between the two concrete structures. The short distance of propagation inside the model due to the near-surface configuration makes difficult to discriminate between the different type of waves recorded in the seismograms. The very high velocity contrast between the background (300 m.s^{-1}) and the concrete foundations (4000 m.s^{-1}) generates high-amplitude reflections. In addition, the close distance between the two structures is responsible for important multiple scattering. The initial model only predicts correctly the main arrival and the reflection on the bottom layer. The challenge for the reconstruction algorithms is the interpretation of the multi-scattered waves.

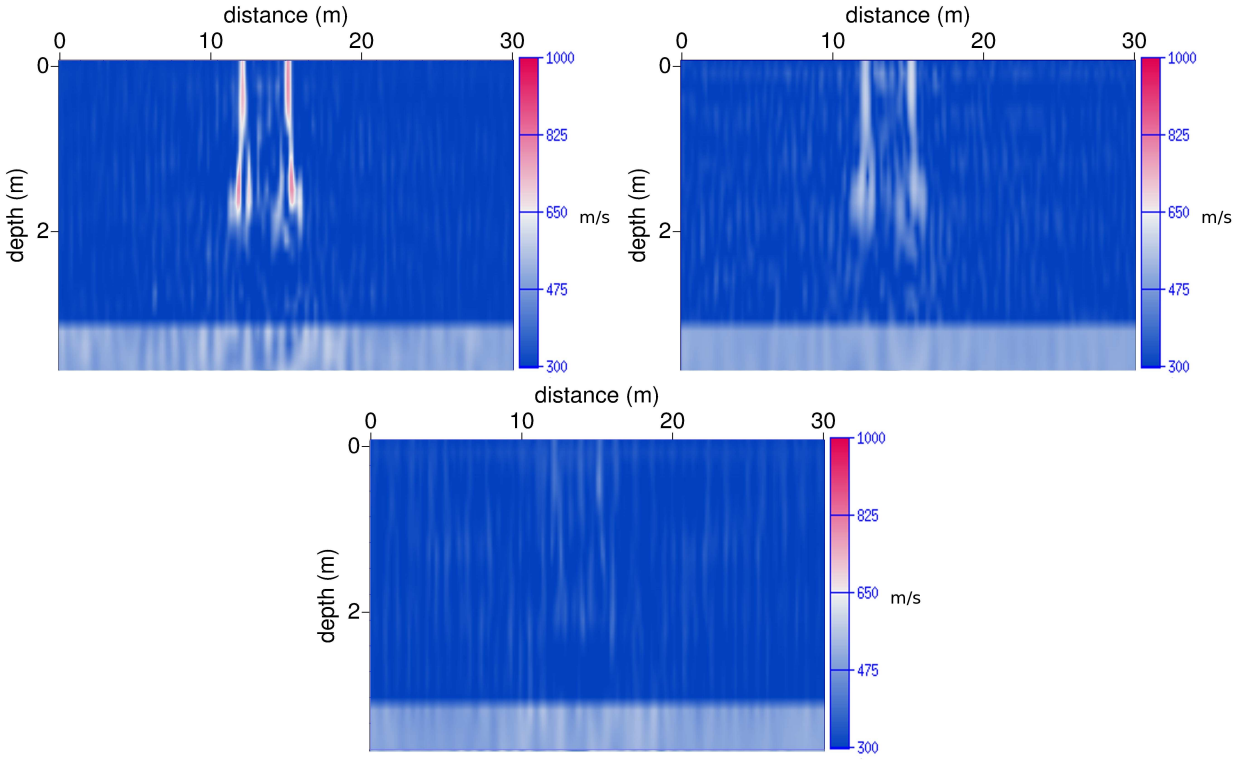

Figure 3. P-wave velocity estimations for the first case study. From top to bottom: preconditioned nonlinear conjugate gradient method, standard l-BFGS method, improved l-BFGS method.

4.2.2 Estimated models

The final estimations provided by the seven methods are presented in figures 3 and 4. Note that the color scale is different from the one used to present the exact model: the maximum estimated wave velocity amplitude only reaches 1000 m.s^{-1} while the exact amplitude of the concrete structures is 4000 m.s^{-1}.

The presence of the concrete structures is detected by the preconditioned conjugate gradient method, the standard l-BFGS method, and the truncated Gauss-Newton method with and without preconditioning. However, the shape of the structure is significantly distorted, a strong blurring affects the image, and the amplitude of the P-wave velocity is underestimated. The improved l-BFGS method is almost unable to detect the structures.

Conversely, the truncated Newton method is able to restore the correct shape of two structures, but the P-wave velocity remains underestimated, since it only reaches 600 m.s^{-1}.

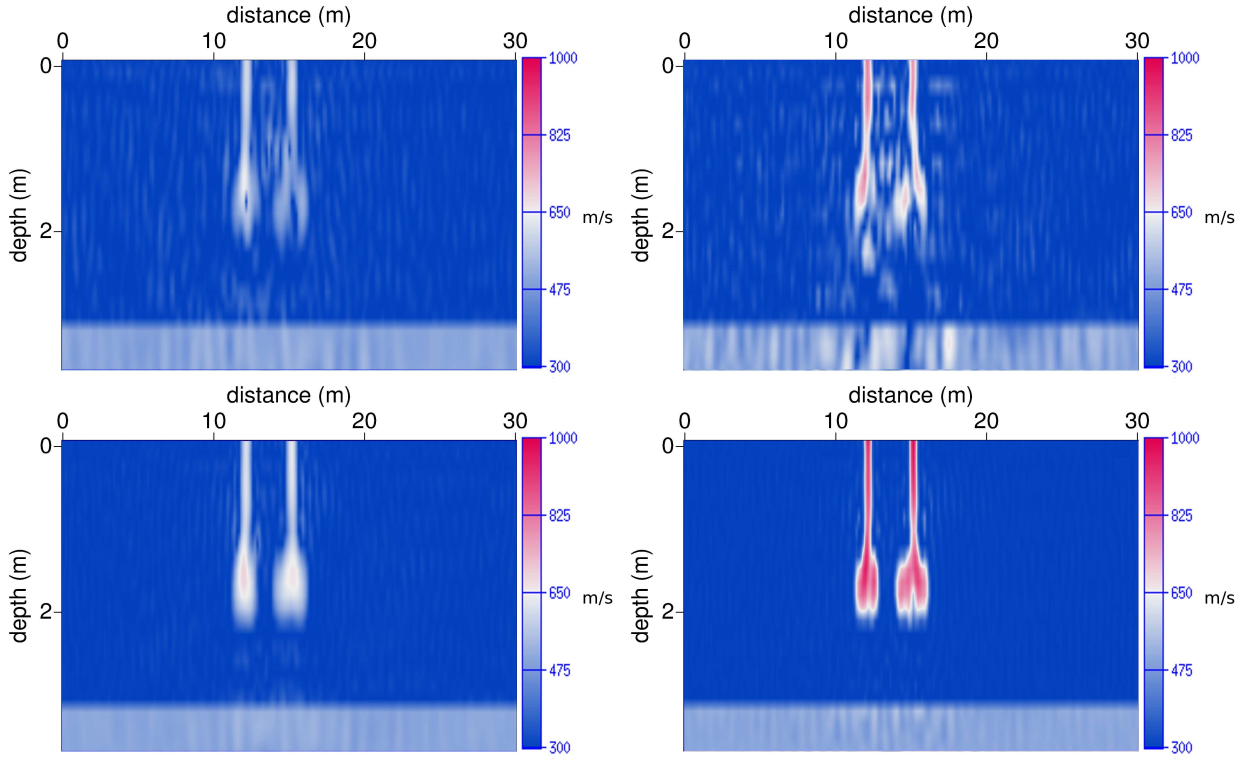

Figure 4. P-wave velocity estimations for the first case study. From top to bottom: truncated Gauss-Newton, preconditioned truncated Gauss-Newton, truncated Newton, preconditioned truncated Newton.

The best estimation is provided by the preconditioned truncated Newton method: the structures are well delineated, and the maximum recovered amplitude reaches 1000 m.s^{-1}.

4.2.3 Final residuals

We compute the time-domain residuals associated with each of the seven final estimated models. We use the same configuration as the one which is used for the computation of the seismograms presented in figure 2. These residuals are presented in figure 5. As expected, the amplitude of the residuals associated with the models provided by the preconditioned nonlinear conjugate gradient, the two l-BFGS methods, and the truncated Gauss-Newton methods (with and without preconditioning), is large. In particular, the multiple reflections between the two concrete structures are not interpreted by these algorithms.

Conversely, the amplitude of the residuals associated with the truncated Newton method and the preconditioned truncated Newton method is significantly smaller. The residuals

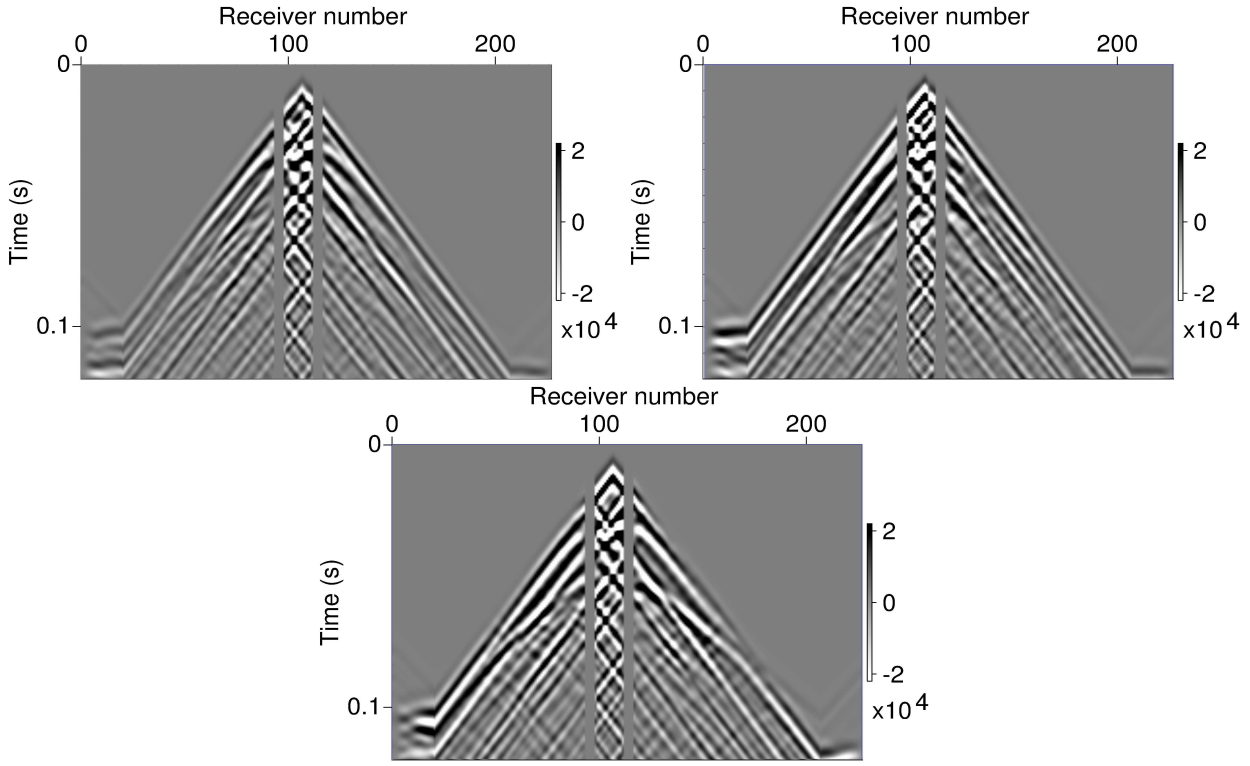

Figure 5. Final residuals for the first case study. From top to bottom: preconditioned nonlinear conjugate gradient method, standard l-BFGS method, improved l-BFGS method.

corresponding to the multi-scattered waves between the two structures significantly decrease. Analyzing the differences between the residuals associated with these two methods shows that the preconditioned truncated Newton residuals are smaller. Only these two methods are able to interpret correctly the multi-scattered wavefield between the two structures.

4.2.4 Convergence analysis

The convergence profiles of the truncated Newton method and its preconditioned version are presented in figure 7. The two methods stop on a linesearch failure: none is able to minimize the misfit function to 0.01% of the initial misfit. The truncated Newton method achieves however a significant reduction of the cost function and reaches 1% of the initial misfit. This, however, requires the resolution of more than 6000 thousand wave propagation problems per source. Conversely, the truncated Newton method solves only 1100 wave propagation problems per source to converge to 1% of the initial misfit, and reaches 0.06% of the initial

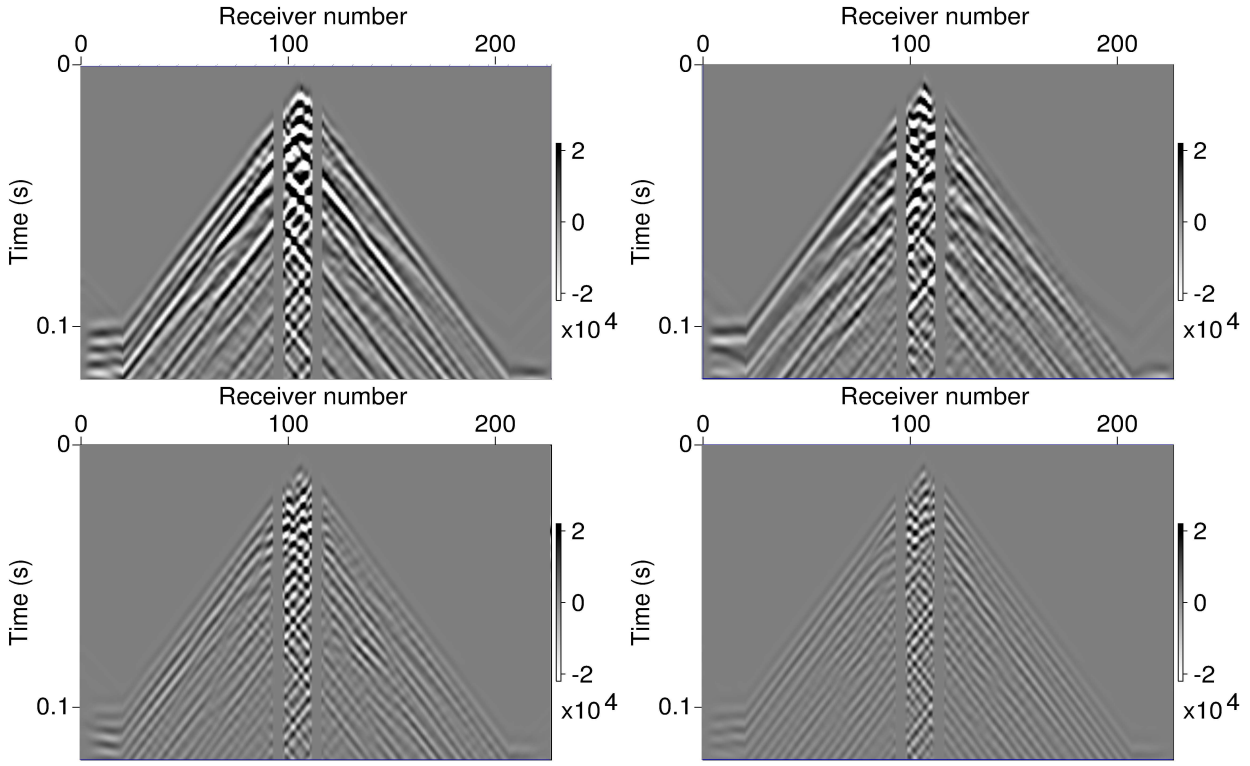

Figure 6. Final residuals for the first case study. From top to bottom: truncated Gauss-Newton, preconditioned truncated Gauss-Newton, truncated Newton, preconditioned truncated Newton.

misfit after the resolution of less than 4300 wave propagation problems per source. This emphasizes the efficiency of the preconditioning method introduced in section 3.3.

4.2.5 Interpretation

The interpretation of these results should be made in view of the Hessian definition given in equations (10) and (11). The second-order part of the Hessian $C(p)$ is related to double-scattered waves. The high velocity contrast between the background model and the concrete structures generates high amplitude multi-scattered waves. As a consequence, in this particular case, this second-order part of the Hessian is non-negligible. As mentioned by Pratt *et al.* (1998), high amplitude multi-scattered waves generate strong artifacts on the gradient descent direction. The preconditioned nonlinear conjugate gradient method is unable to correct these artifacts, as the preconditioning matrix which is used only approximates the

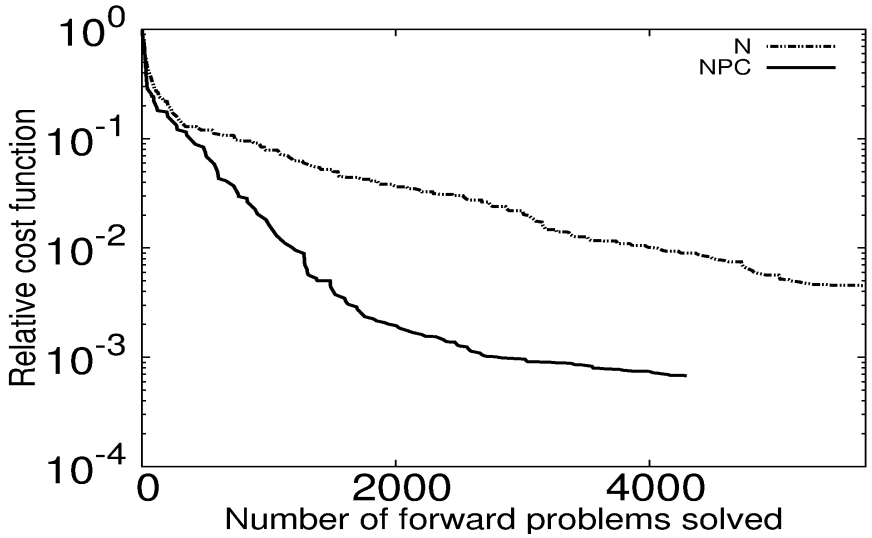

Figure 7. Convergence curves for the first case study: effect of the preconditioner. Comparison of the truncated Newton method (N) and the preconditioned truncated Newton method (NPC).

Gauss-Newton part of the Hessian matrix. For the same reason, the two truncated Gauss-Newton based methods are unable to interpret the data.

The case of the l-BFGS methods is slightly more complex. Indeed, the l-BFGS approximation should approximate the entire Hessian operator, and not only its Gauss-Newton part. Nonetheless, the l-BFGS method generates a symmetric definite positive approximation of the Hessian. However, in this case, because of the second order part $C(p)$, the true Hessian operator is indefinite. Therefore, for the first iterations, the l-BFGS approximation is inaccurate and prevents the l-BFGS algorithm from minimizing efficiently the misfit function $f(p)$. Since the amplitude of the residuals remains large, the second order part $C(p)$ remains important and the true Hessian matrix stays indefinite, while the successive l-BFGS approximations are positive definite. This discrepancy between the actual Hessian and its l-BFGS approximation may be responsible for the failure of the l-BFGS inversion scheme. Accounting for the preconditioning matrices $P_k^{\nu,\theta}$ as prior estimation of the l-BFGS approximation is also inefficient: only the Gauss-Newton part $B(p)$ of the Hessian is approximated by these matrices, and no information of the second order part $C(p)$ is added.

The differences between the estimations provided by the truncated Newton method and its preconditioned version demonstrates the scaling effect of the Gauss-Newton part of the

Hessian operator. The approximation of the diagonal of $B(p)$ used as a preconditioner corrects for the amplitude of the subsurface parameters. The preconditioned truncated Newton method thus combines the advantages of using the full inverse Hessian operator and correcting the amplitude of the subsurface parameter estimation through the preconditioner.

As a conclusion, this near-surface application case gives a good insight of the importance of the Hessian operator within the FWI reconstruction scheme, in particular in presence of complex multi-scattered wavefield. In the two following case studies, we investigate the efficiency of the truncated Newton approach for more conventional seismic applications, starting with the Marmousi II test case.

4.3 The Marmousi II test case

4.3.1 Presentation

The first Marmousi model is a synthetic P-wave velocity model designed at IFPEN (Institut Français du Pétrole Energies Nouvelles) in 1988 and proposed as a benchmark for testing seismic imaging method (Lailly *et al.* 1991). The geometry of this model is based on a realistic profile. In 2006, an upgrade toward the so-called Marmousi II model is proposed (Martin *et al.* 2006). The initial model is enlarged, and associated shear wave velocity and density models are added for elastic wave propagation modeling. In this study, we restrict ourselves to the acoustic approximation, and we only use the Marmousi II P-wave velocity model, presented in figure 8. As in the previous case, the ratio between vertical and horizontal distances is not respected to improve the readability of the model figures.

The Marmousi II model is 16 km wide and 3.5 km deep, with a 400 m deep water layer at the top of the model. In the following experiment, this layer is kept constant. This results in a reduction of the parameter space and a stabilization of the problem: since these parameters are close from the sources and the receivers, the wavefield is very sensitive to small variations of these parameters.

We use a discretization step of 25 m, which corresponds to approximately 10^5 pressure wave velocity discrete parameters. As for the previous test case, the PML size is set to 10

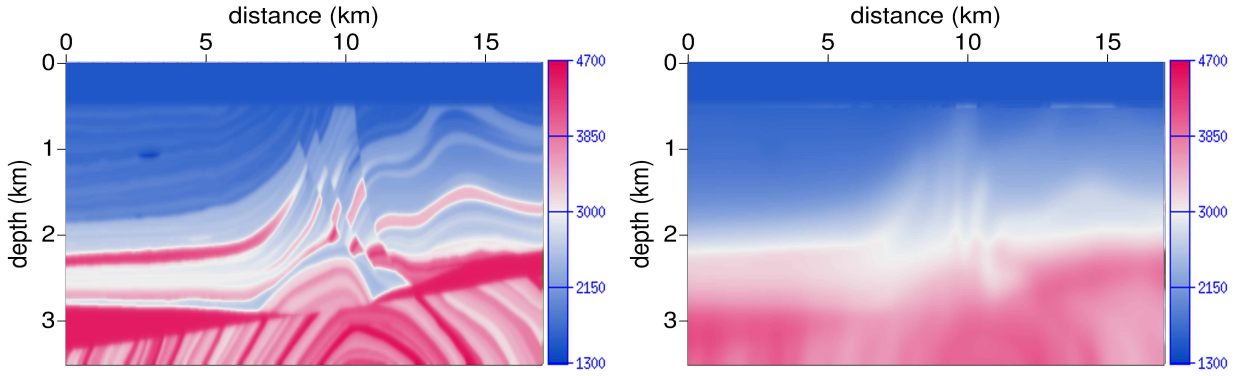

Figure 8. The Marmousi II pressure wave velocity model (top), initial guess (bottom).

discretization points on each side of the domain. No free-surface condition is implemented: this amounts to inverting pre-processed seismic data from which the free-surface multiples have been filtered. We use a marine acquisition system composed of 144 sources and 660 receivers located at the top of the model (25 m depth) in the water layer. The receivers are disposed from $x = 0.05$ km to $x = 16,5$ km each 25 m. The sources are located from $x = 0.05$ km to $x = 14,35$ km with a spatial sampling of 100 m. This generates a slight deficit of illumination of the right part of the model.

We compute 4 synthetic data sets, for the frequencies 3 Hz, 5 Hz, 8 Hz, 12 Hz. The corresponding number of discrete data is approximately $380,000$. We use a smoothed version of the exact Marmousi II model as initial guess (Fig. 8). This smoothing is performed using the Seismic Unix `smooth2` function, which performs a quadratic interpolation of the model, with smoothing parameters $r_1 = r_2 = 20$. This model could be for instance the result of a travel-time tomography of the data. As in the previous case, we invert simultaneously the 4 data sets.

We compute two shot-gathers, corresponding respectively to the exact and the initial model (Fig. 9). The source is a Ricker type signal, with a frequency bandwidth centered around 5 Hz. It is located at the surface at $x = 7$ km. The data computed using the exact model is clearly less complex than in the previous case. The first arrival dominates the data, the amplitude of reflected waves decreases rapidly and the different seismic events can be

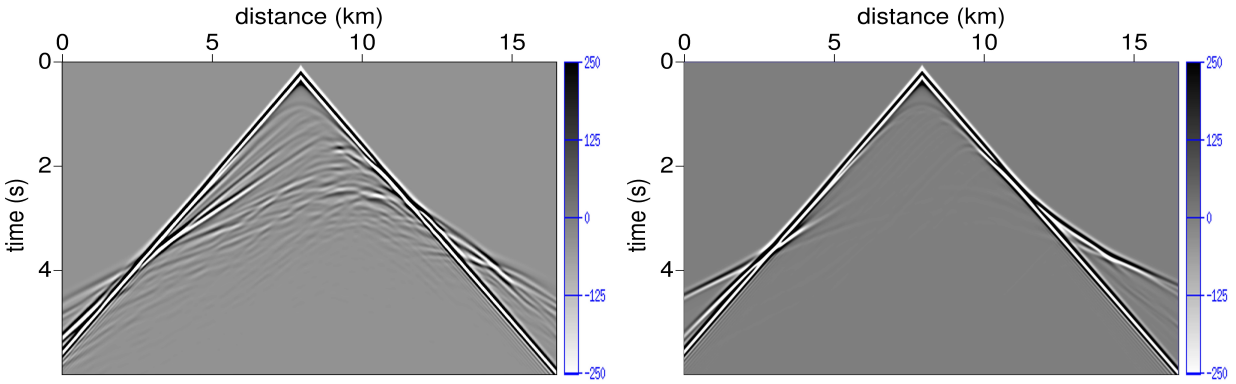

Figure 9. Time-domain data associated with the exact model (top), with the initial model (bottom).

easily separated. This is not surprising, as the Marmousi II P-wave velocity model does not contain highly contrasted interfaces.

4.3.2 Convergence analysis

The stopping criterion ϵ is set to 10^{-4}: the iterations are stopped when the misfit function reaches 0.01% of its initial value. The convergence profiles of the seven methods are presented in figure 10.

Among the seven minimization methods, the truncated Newton and truncated Gauss-Newton methods present the slowest convergence profiles: the two methods require respectively 2196 and 1502 wave propagation problem resolutions per source to satisfy the convergence criterion. The preconditioned nonlinear conjugate gradient and the standard l-BFGS methods are significantly more efficient: they require to solve only 824 and 786 wave propagation problems per source respectively to reach the same level of accuracy. The preconditioned truncated Newton and preconditioned truncated Gauss-Newton method provide even a faster convergence since the convergence is achieved with the resolution of respectively 682 and 324 wave propagation problems per source. Finally, for the Marmousi II case study, the fastest method to converge is the improved l-BFGS method, which reaches the required level of accuracy after the resolution of only 200 wave propagation problems per source.

The analysis of these convergence profiles underlines the effect of the preconditioner for

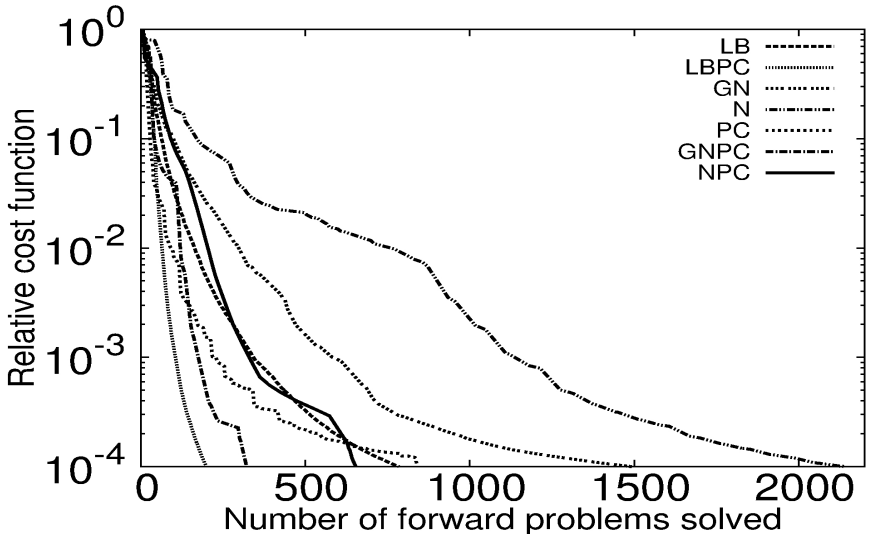

Figure 10. Misfit function decrease for the Marmousi test case. Comparison of the preconditioned nonlinear conjugate gradient method (PC), the standard l-BFGS method (LB), the improved l-BFGS method (LBPC), the truncated Gauss-Newton method (GN), the truncated Newton method (N), the preconditioned truncated Gauss-Newton method (GNPC) and the preconditioned truncated Newton method (NPC).

the truncated Newton method. Using no preconditioning, the truncated Gauss-Newton and truncated Newton methods converge slower than conventional methods such as the preconditioned nonlinear conjugate gradient or the standard l-BFGS method. Using the pseudo-Hessian preconditioner, the truncated Newton strategy outperforms these approaches. This preconditioning thus yields a significant improvement of the truncated Newton method in terms of convergence speed. However in this case, the incorporation of the diagonal of the pseudo-Hessian matrix within the l-BFGS framework turns out to be the most efficient strategy, since the best convergence profile is provided by the improved l-BFGS method.

4.3.3 Estimated models

The velocity models estimated with the improved l-BFGS method and the preconditioned truncated Gauss-Newton methods methods are presented in figure 11. From a low resolution approximation of the solution, the FWI method provides a high resolution quantitative estimation of the solution. We present in figures 12 and 13 the differences between the exact model and the estimations provided by the six minimization algorithms.

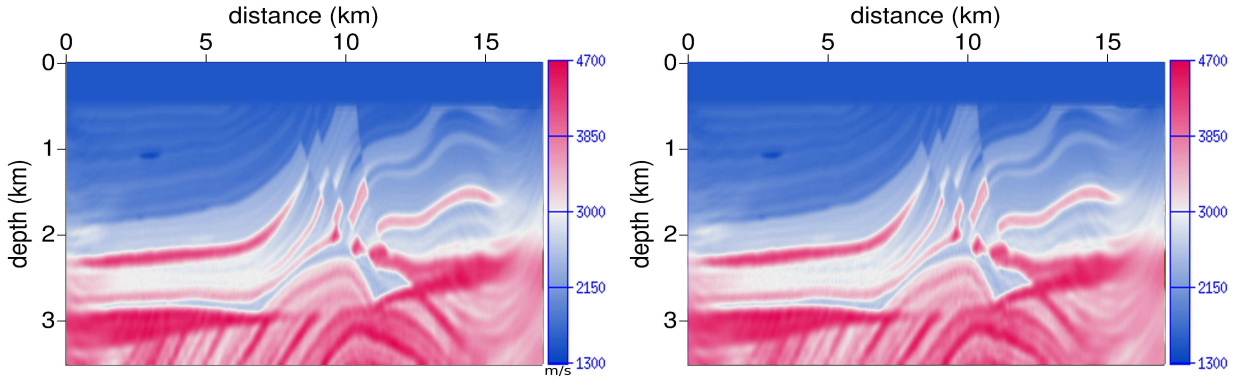

Figure 11. Improved l-BFGS (top) and preconditioned truncated Gauss-Newton (bottom) FWI results for the Marmousi II test case.

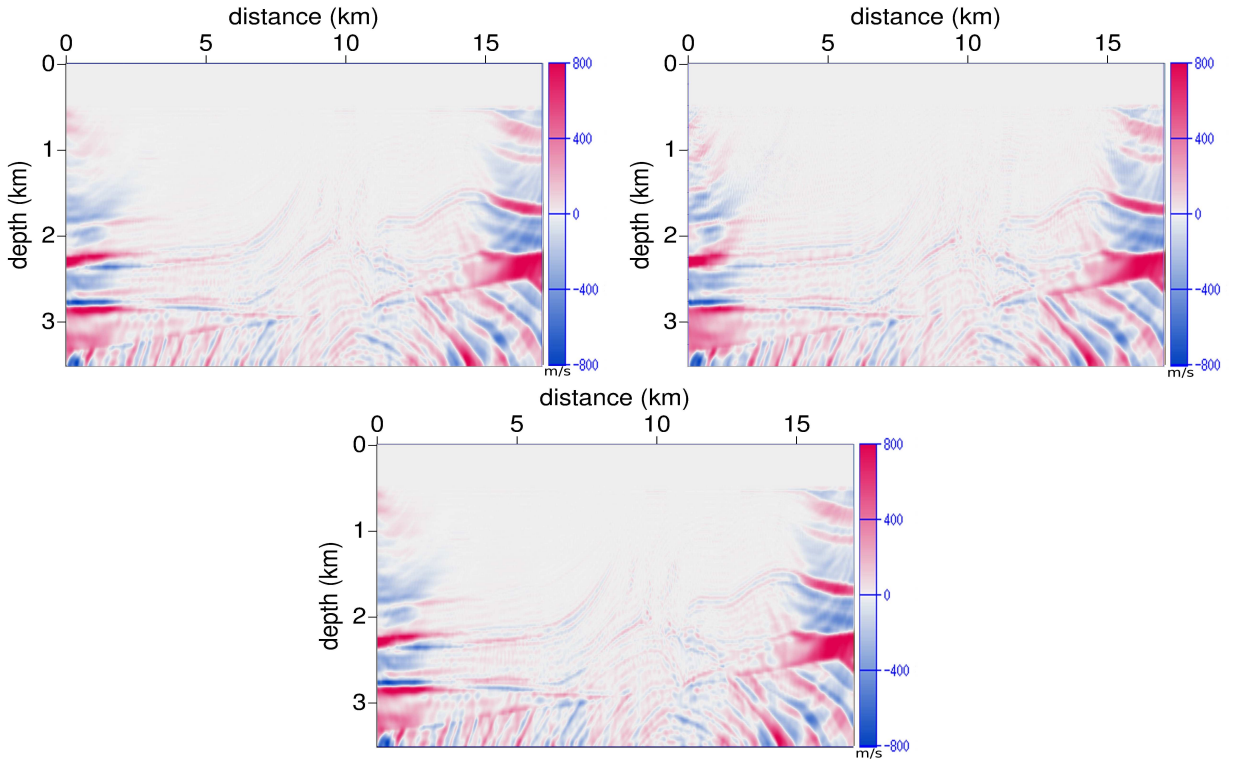

Figure 12. Differences between the exact and the estimated models for the Marmousi II case study. From top to bottom: preconditioned nonlinear conjugate-gradient method, standard l-BFGS method, improved l-BFGS method.

The seven estimations are very similar. Their differences with the exact model are almost the same. It is however possible to make some distinctions. The estimated model provided by the l-BFGS method presents rapid oscillations in the upper part of the model, which is not case for the other estimated models. However, in the deeper part, the differences between the

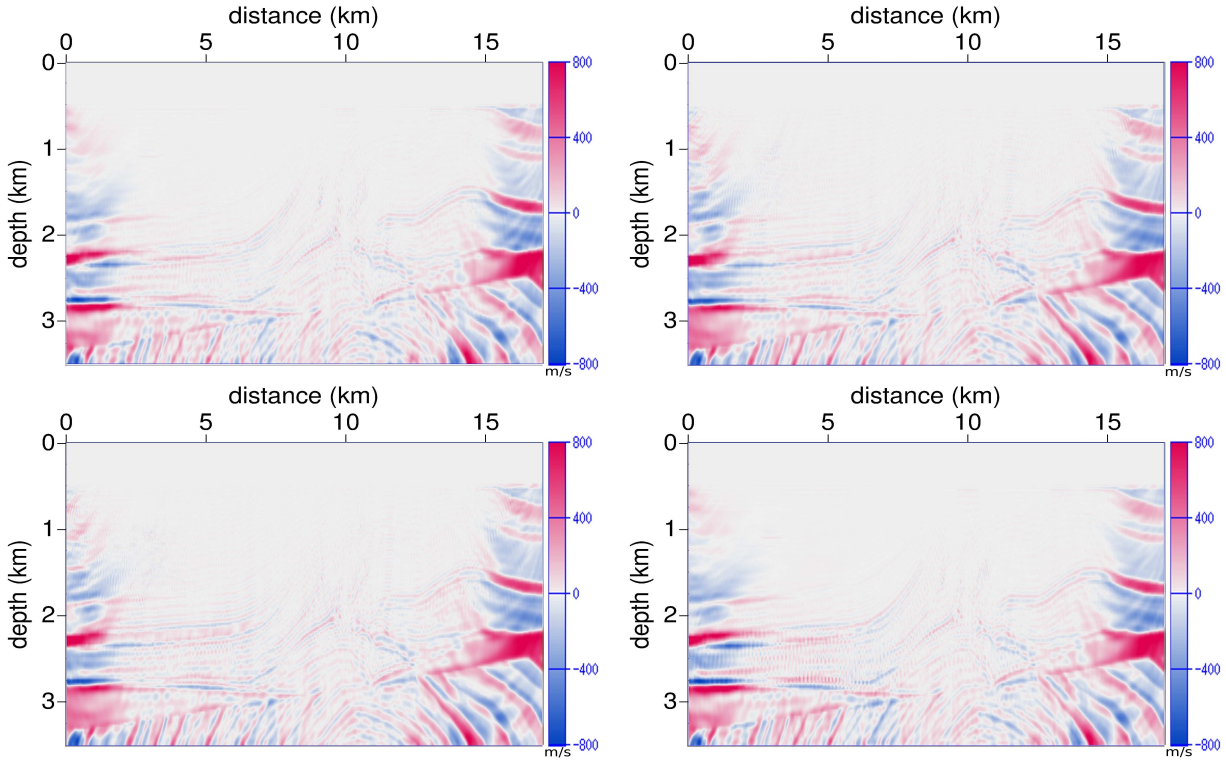

Figure 13. Difference between the exact and the estimated models for the Marmousi II case study. From top to bottom: truncated Gauss-Newton, preconditioned truncated Gauss-Newton, truncated Newton, preconditioned truncated Newton

model provided by the preconditioned nonlinear conjugate gradient method and the exact model are slightly larger than the differences associated with the improved l-BFGS model or the preconditioned truncated Gauss-Newton method.

4.3.4 Interpretation

Compared to the previous experiment, the seismic data which is interpreted here is less complex, as visible in figure 9. From a smooth and low resolution initial model, all the seven minimization schemes provide an accurate estimation of the exact subsurface model. The differences between the seven methods can be stated in terms of convergence speed. The truncated Newton and truncated Gauss-Newton methods demand the resolution of numerous wave propagation problems. Only using the preconditioning matrices $P_k^{\nu,\theta}$ these two methods outperforms the standard optimization schemes. The most efficient method turns out in this case to be the improved l-BFGS method, which converges at very fast rate.

We can infer from this experiment that, for reasonably complex subsurface models, the second-order part of the inverse Hessian operator can be neglected. In this configuration, the Gauss-Newton approximation of the Hessian operator is accurate enough to provide high convergence speed. This is underlined by the improved efficiency of the Gauss-Newton approach with respect to the full Newton approach (with or without preconditioning). The combination of the l-BFGS approximation and an accurate prior estimation of the Hessian through the matrices $P_k^{\nu,\theta}$ within the improved l-BFGS method provides here the best optimization strategy.

In the last case study, we investigate another imaging problem at the seismic exploration scale, that exhibits the same kind of complexity as presented in the first case study in terms of multi-reflected waves.

4.4 The 2004 BP model

4.4.1 Presentation

The 2004 BP model has been originally designed as a benchmark model for testing wave velocity estimation methods in presence of complex subsurface structures (Billette and Brandsberg-Dahl 2004). The whole model, presented in figure 14, is 67 km long and 12 km deep, and defined on a 6.25 m $\times$6.25 m grid. This represents more than 10×10^6 discrete unknown parameters. We reduce the size of the model by decimation, and we focus on the left part of the original model. This part presents a complex rugose salt body, and sub-salt slow velocity anomalies that represent over-pressured zones. This intends to mimic the geology that can be found in the Gulf of Mexico. The main challenges in this area are related to obtaining a precise delineation of the salt and recovering information on the sub-salt velocity variations.

The P-wave velocity in the salt reaches 4900 m.s^{-1}, while it is equal to 1486 m.s^{-1} in the water. Even if the contrast is not as important as in the first case study (section 4.2), the discrepancy between these two values is enough for the salt structures to generate high amplitude reflections. In addition these reflected waves are reflected back at the top of the

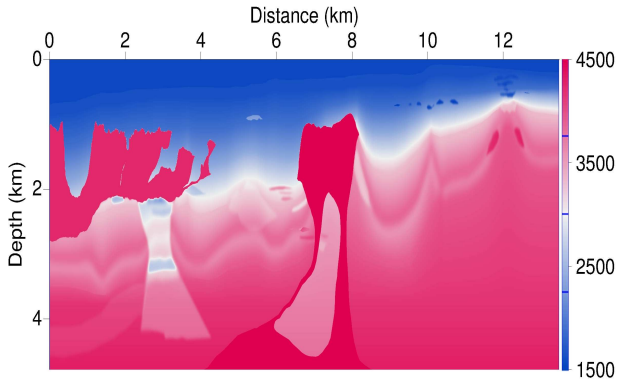

Figure 14. Exact whole BP 2004 model.

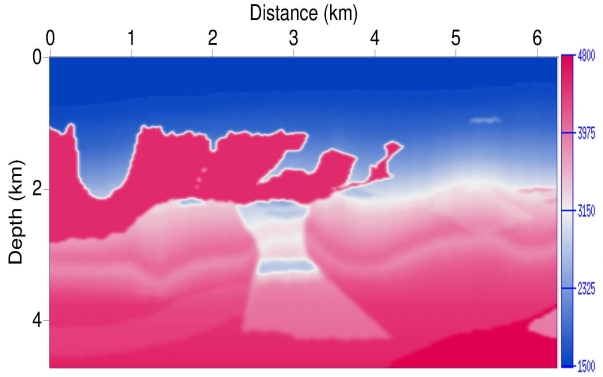

Figure 15. Exact reduced BP 2004 model.

water layer. We recognize the same pattern as in the first case study: the proximity of two reflectors generates multiple scattering.

We perform the decimation of the original model taking one parameter value each ten grid points, and we choose a 25 m discretization grid. We end up with a 6.2 km wide and 4.2 km deep reduced model, described by approximately 5×10^4 discrete parameters. A quadratic smoothing filter with a 50 m characteristic length is applied to smooth out the discontinuities generated by the decimation. The resulting model is presented in figure 15. As for the Marmousi II test case, the water layer is kept constant throughout the iterations, so as to stabilize the problem. The bathymetry of the sea-bottom is respected.

We use a surface acquisition system with 62 sources and 248 receivers, from $x = 50$ m to $x = 6225$ m at 25 m below the sea-level. The spatial sampling of the receivers and the sources is set up to 25 m and 100 m respectively. Contrary to the two previous cases,

we use here a free-surface condition at the top of the model, to account for the surface multiples, as our ambition is to evaluate the efficiency of the seven minimization methods for the interpretation of complex seismic data.

A 10 grid-points PML layer is added to the other sides of the domain. We generate 19 data sets, from 2 Hz frequency to 20 Hz with 1 Hz sampling. This represents approximately 2×10^6 discrete data. We apply here a hierarchical strategy in frequency. The complexity of the model requires to use such a technique to prevent from cycle skipping effect. Therefore, the 19 data sets are gathered into 6 overlapping subgroups, each of them containing 4 data sets.

- Group 1 : 2 Hz, 3 Hz, 4 Hz, 5 Hz
- Group 2 : 5 Hz, 6 Hz, 7 Hz, 8 Hz
- Group 3 : 8 Hz, 9 Hz, 10 Hz, 11 Hz
- Group 4 : 11 Hz, 12 Hz, 13 Hz, 14 Hz
- Group 5 : 14 Hz, 15 Hz, 16 Hz, 17 Hz
- Group 6 : 17 Hz, 18 Hz, 19 Hz, 20 Hz

We apply the FWI method iteratively to each group of data sets: the estimation computed from Group i is used as an initial model for applying the FWI algorithm to Group $i + 1$.

The cycle of FWI starts with an initial model computed as a smooth version of the exact model. As for the Marmousi II test case, this smoothing is performed using the Seismic Unix `smooth2` function. We use here smoothing parameters such that $r_1 = r_2 = 15$. The initial model is presented in figure 16.

As in the previous cases, we compute two shot-gathers, corresponding respectively to the exact and the initial model. We use the same source as in the Marmousi II case, located at $x = 3.1$ km, at 25 m below the sea-level. The two gathers are presented in figure 17.

As expected, the data computed in the exact domain exhibits the same high amplitude multi-reflected waves, below the first arrival. These waves are generated by the salt structures and their reflections on the free-surface at the sea-level. The data computed in the initial model contains only the first-arrival and small amplitude events related to the structure of

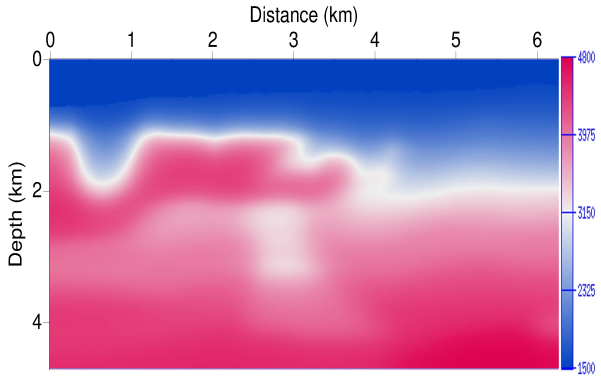

Figure 16. Initial model for the BP 2004 case study.

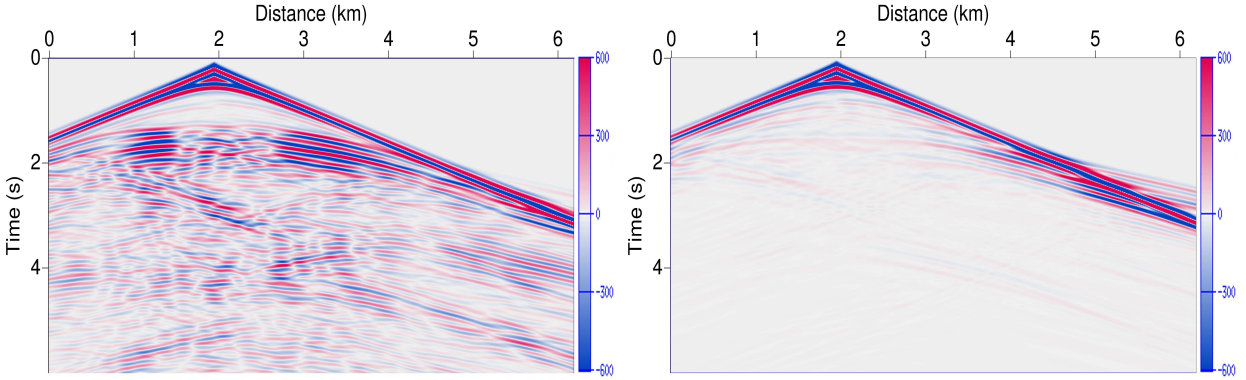

Figure 17. Time-domain data for the BP case study. Exact model (top), initial model (bottom)

the initial model. Even if the multi-reflection phenomenon is not as significant as in the first case study, the complexity of the data make this test case challenging for FWI.

4.4.2 Estimated models

The models estimated by the seven minimization methods are presented in figures 18 and 19. These estimations are obtained at the end of the cycle of inversions.

The two *l*-BFGS methods and the non-preconditioned versions of the truncated Newton and Gauss-Newton methods fail to correctly interpret the data. These methods are trapped in a local minima during the inversion of the first group of data sets. The inversion of the other groups does not yield significant improvements. As it can be expected in FWI, the interpretation of the lowest frequencies is crucial. The truncated Newton and Gauss-Newton methods generate a model very close from the initial model, failing to minimize the residuals,

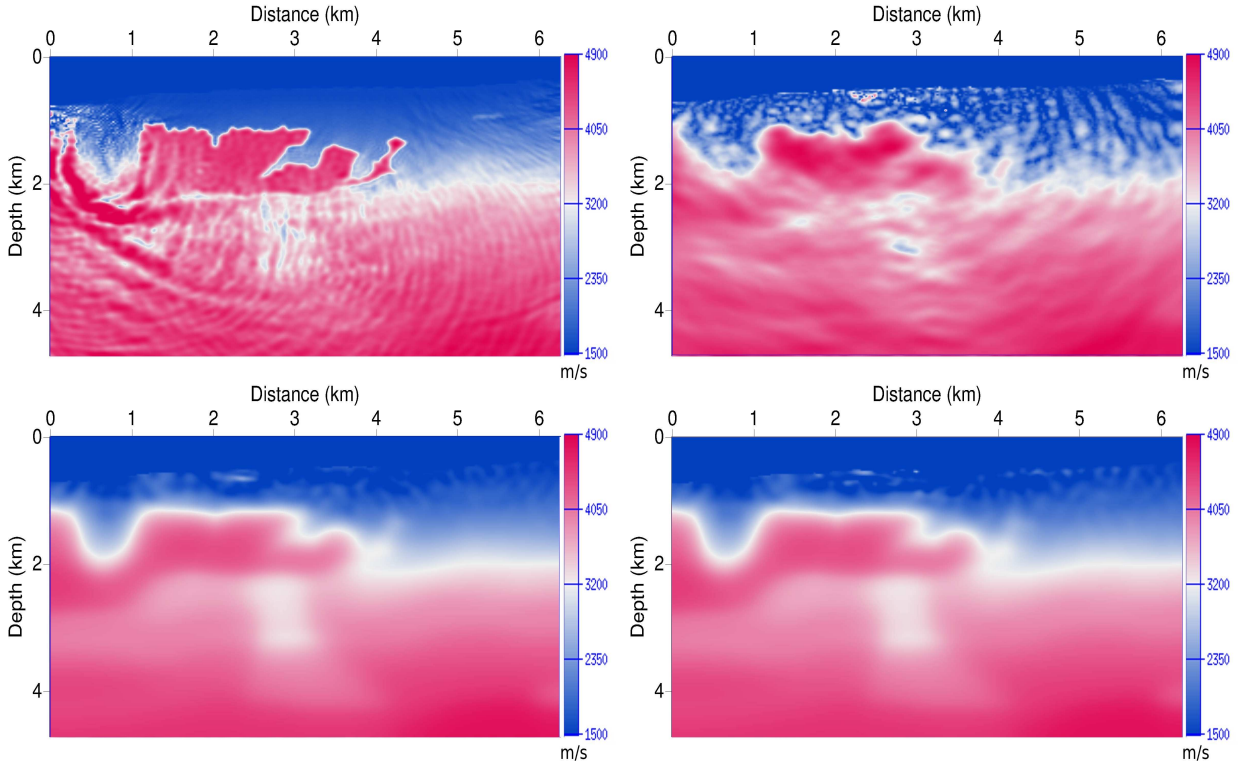

Figure 18. Estimated models for the BP case study. From top to bottom: standard l-BFGS method, improved l-BFGS method, truncated Newton method, truncated Gauss-Newton method.

while the two l-BFGS methods provide highly perturbed P-wave velocity estimations. These perturbations are mainly located in the upper left part of the model, where the cavity in the salt-structure may be responsible for the generation of high amplitude multi-scattered waves.

Conversely, the preconditioned nonlinear conjugate gradient method, the preconditioned truncated Gauss-Newton method and the preconditioned truncated Newton method are able to provide correct estimations of the P-wave velocity. In particular, the top of the salt dome is well delineated. However, the estimation computed by the preconditioned nonlinear conjugate gradient method is still subject to instabilities. In particular, an erroneous salt zone of large size is created in the salt cavity in the top left part of the model. These erroneous salt structures also appear in the estimation provided by the preconditioned truncated Gauss-Newton method. They are almost totally removed in the preconditioned truncated Newton estimation.

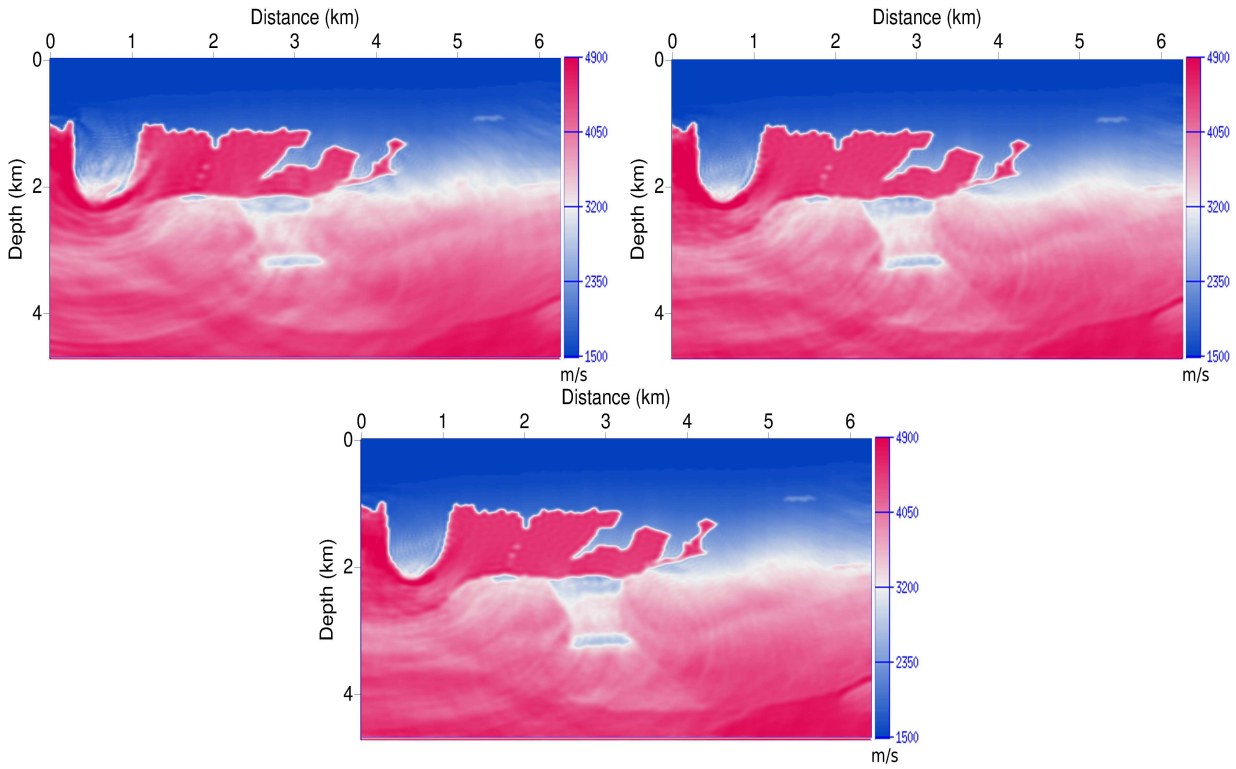

Figure 19. Estimated models for the BP case study. From top to bottom: preconditioned nonlinear conjugate gradient method, preconditioned truncated Gauss-Newton method, preconditioned truncated Newton method.

In figure 20, the differences between the estimations provided by the 3 latter methods and the exact model are presented. Not surprisingly, the preconditioned nonlinear conjugate gradient estimation appears to be the less correct, as the amplitude of the differences with the exact model in the upper left part is larger than for the two other estimations. The preconditioned truncated Gauss-Newton estimation also appears not to be as good as the one provided by the preconditioned truncated Newton method especially in the upper left part of the model which has been identified as the most sensitive part of the model.

4.4.3 Residuals

In figure 21, the residuals in the time domain associated with the model estimated by the preconditioned nonlinear conjugate gradient method, the preconditioned truncated Gauss-Newton method, and the preconditioned truncated Newton method are presented. These residuals are computed using the same acquisition system as the one which is used to compute

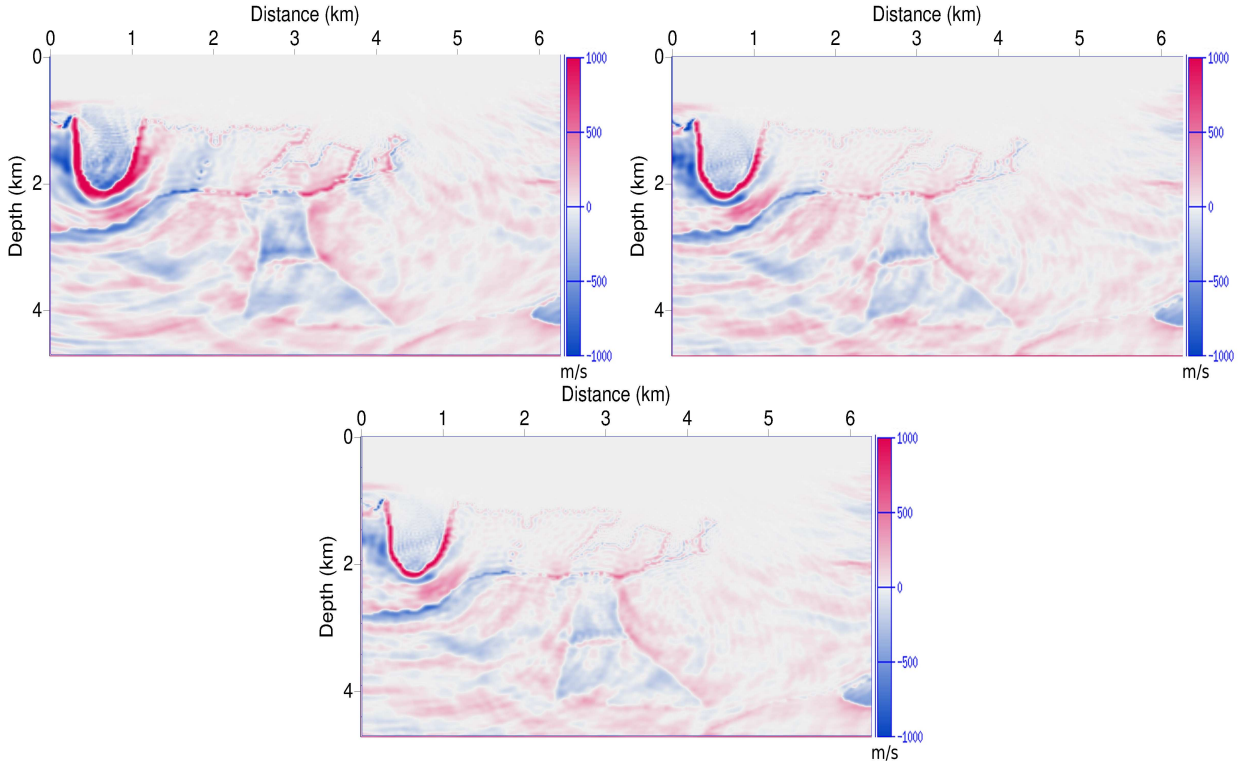

Figure 20. Differences between the estimation and the exact model for the BP study. From top to bottom: preconditioned nonlinear conjugate gradient method, preconditioned truncated Gauss-Newton method, preconditioned truncated Newton method.

the data in the time for the exact and initial models in figure 17. As expected, the amplitude of the residuals corresponding to the preconditioned nonlinear conjugate gradient estimation is larger than for the two other methods. The residuals associated with the preconditioned truncated Gauss-Newton models are also slightly larger than for the full Newton version of this method.

4.4.4 Convergence analysis

The convergence profiles of the three methods that provide a correct estimation of the P-wave velocity are presented for each group of data sets in figures 22 to 27. For each group of frequency, the decrease achieved by the preconditioned nonlinear conjugate gradient method is less important than the one provided by the two others. Excepted for the first frequency group, this method is unable to minimize the misfit function to the required accuracy level $\epsilon = 10^{-3}$. A linesearch failure is detected instead.

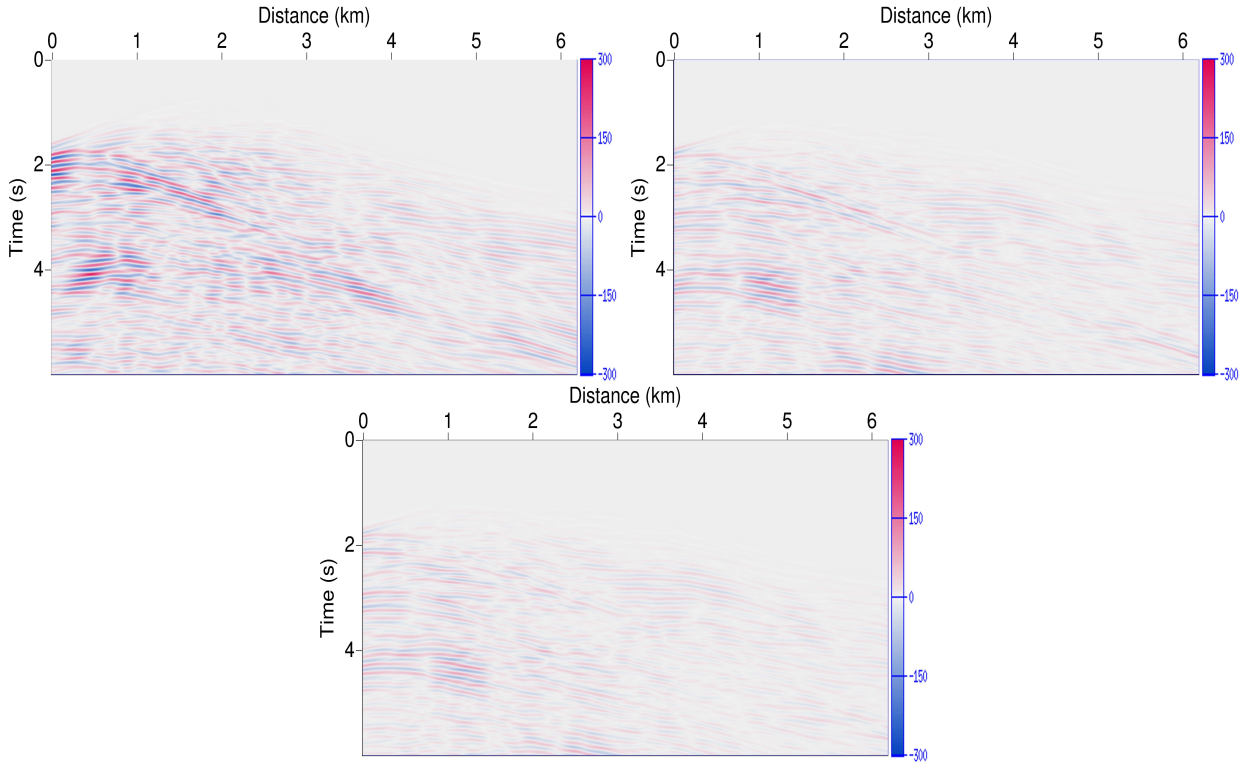

Figure 21. Residuals in the time domain for the BP case study. From top to bottom: preconditioned nonlinear conjugate gradient method, preconditioned truncated Gauss-Newton method, preconditioned truncated Newton method.

Conversely, the preconditioned truncated Newton and Gauss-Newton methods converges to this level of accuracy for all the frequency groups, excepted the preconditioned truncated Gauss-Newton method which stops on a linesearch failure at 0.2% of the misfit function on the last frequency group. For the five first frequency groups, this method is also the most efficient: the Gauss-Newton approximation allows to converge faster than the full Newton method. However, for the last frequency group, the truncated Newton method converges slightly faster.

4.4.5 Interpretation

The presence of largely contrasted structures generates the propagation of complex wavefields, with the presence of high amplitude reflected waves which are difficult to discriminate. As a consequence, standard minimization schemes fail to converge and interpret this data. The *l*-BFGS method is unable to provide a correct estimation, with or without us-

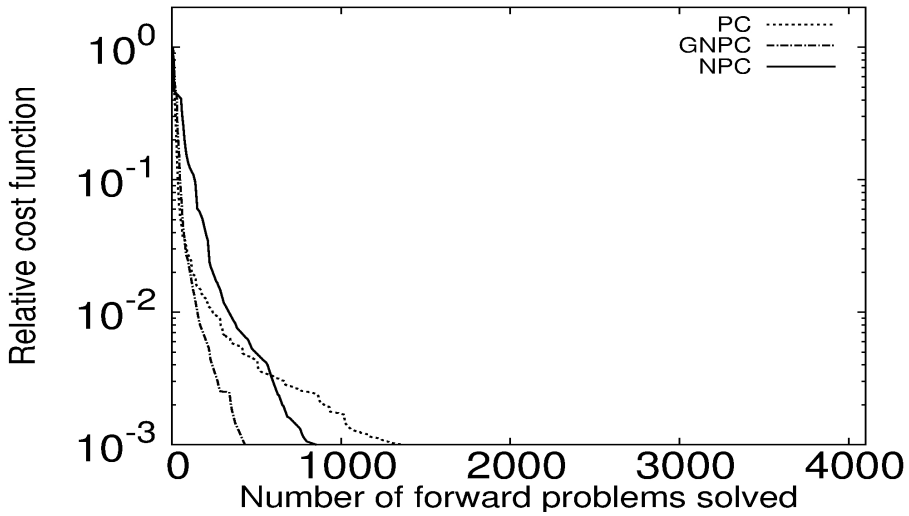

Figure 22. Misfit function decrease for the BP 2004 case study: 1st group of datasets. PC: preconditioned nonlinear conjugate gradient, GNPC: preconditioned truncated Gauss-Newton method, NPC: preconditioned truncated Newton method.

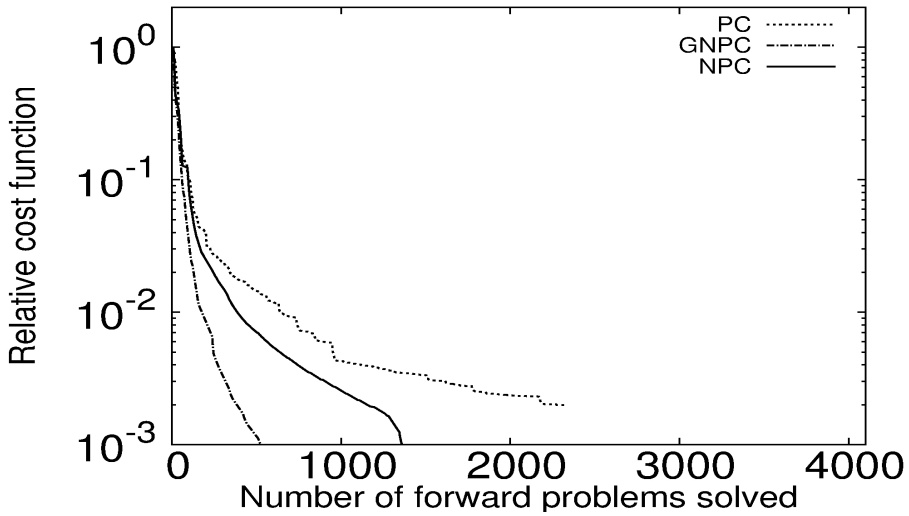

Figure 23. Misfit function decrease for the BP 2004 case study: 2d group of datasets. PC: preconditioned nonlinear conjugate gradient, GNPC: preconditioned truncated Gauss-Newton method, NPC: preconditioned truncated Newton method.

ing the rescaling information embedded in the diagonal of the pseudo-Hessian operator. The truncated Newton method is also unable to converge, using the full Hessian, or in the Gauss-Newton approximation.

From the first experiment, one can infer that the presence of high amplitude reflected

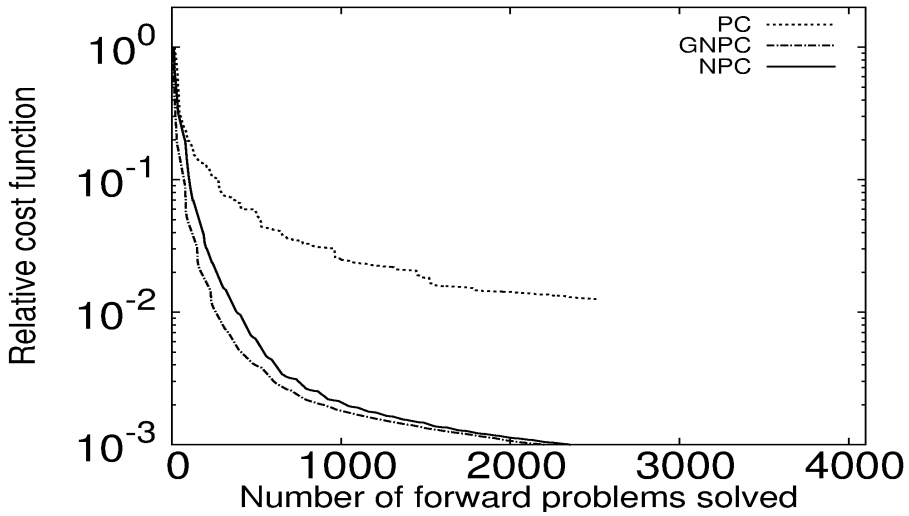

Figure 24. Misfit function decrease for the BP 2004 case study: 3rd group of datasets. PC: preconditioned nonlinear conjugate gradient, GNPC: preconditioned truncated Gauss-Newton method, NPC: preconditioned truncated Newton method.

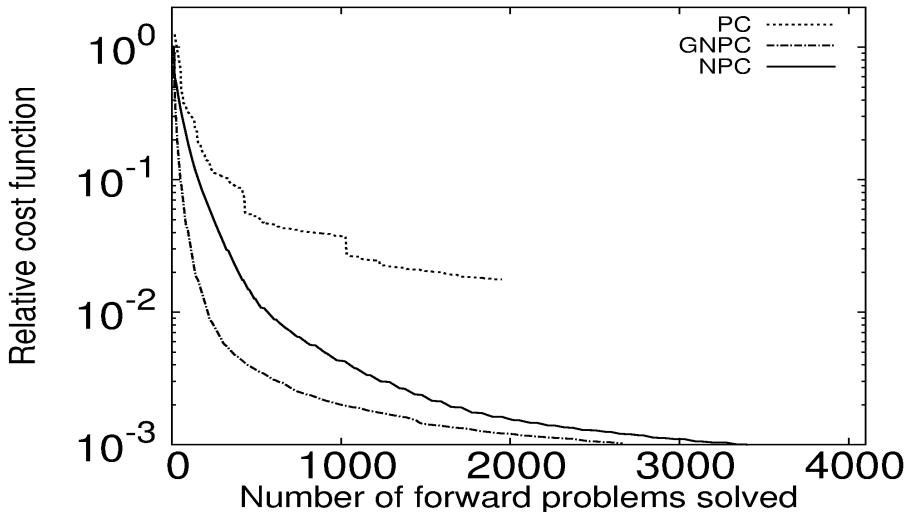

Figure 25. Misfit function decrease for the BP 2004 case study: 4th group of datasets. PC: preconditioned nonlinear conjugate gradient, GNPC: preconditioned truncated Gauss-Newton method, NPC: preconditioned truncated Newton method.

waves should be again responsible for the second-order part of the Hessian operator $C(p)$ to be non-negligible. As a consequence, this could be the cause of the failure of the truncated Gauss-Newton method and the two l-BFGS methods, which assume that the Hessian operator is positive definite, while the second-order part $C(p)$ makes it indefinite.

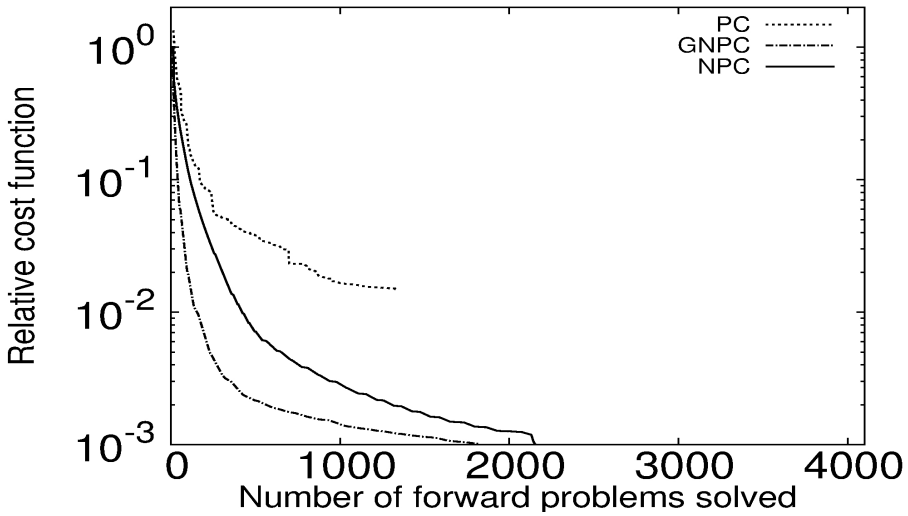

Figure 26. Misfit function decrease for the BP 2004 case study: 5th group of datasets. PC: preconditioned nonlinear conjugate gradient, GNPC: preconditioned truncated Gauss-Newton method, NPC: preconditioned truncated Newton method.

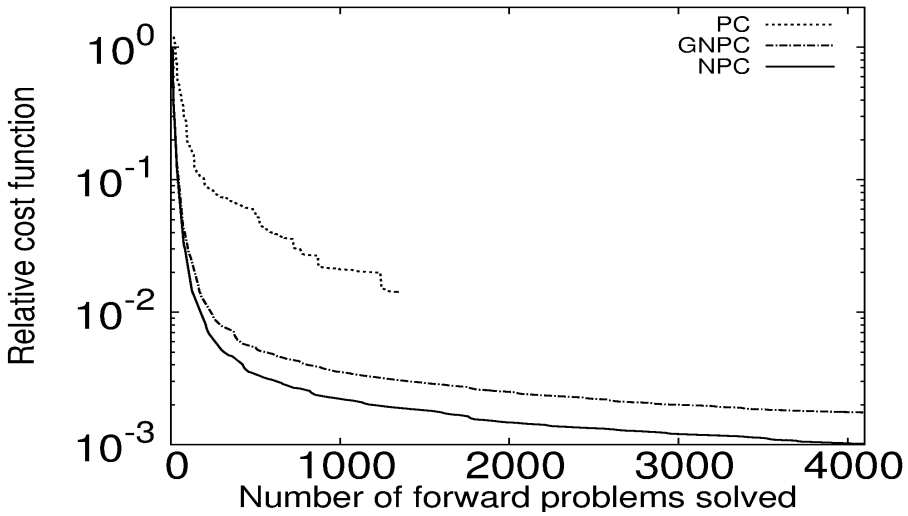

Figure 27. Misfit function decrease for the BP 2004 case study: 6th group of datasets. PC: preconditioned nonlinear conjugate gradient, GNPC: preconditioned truncated Gauss-Newton method, NPC: preconditioned truncated Newton method.

This explanation is, however, incompatible with the fact that the preconditioned truncated Gauss-Newton and the preconditioned nonlinear conjugate gradient methods are able to converge, while the truncated Newton method fails to converge. Indeed, these two first

methods do not account for the full Hessian within the inversion scheme, while the truncated Newton method does.

This experiment actually emphasizes the crucial importance of the scaling effect of the preconditioner based on the pseudo-Hessian operator. Only the methods using this preconditioner are able to minimize the misfit function. In addition, in this experiment, accounting for this preconditioner within the l-BFGS scheme seems not to be effective, as the improved l-BFGS method fails to converge.

The effect of the high amplitude multi-reflected waves may, however, be detected within the inversion of the last frequency group. Only for this group the preconditioned truncated Newton method outperforms the Gauss-Newton version of this method. This is in accordance with the fact that the presence of reflected waves increases with the frequency. In this context, accounting for the full Hessian operator may become more important, as it is suggested by the convergence profiles.

Finally, this experiment also emphasizes that the preconditioned truncated Newton or Gauss-Newton methods should be preferred in this context to the more simple preconditioned nonlinear conjugate gradient method. Even if the result provided by this method is already satisfactory, the two latter methods present better convergence properties, and provide a better estimation of the P-wave velocity model.

In the next section, we summarize the results presented in this study, and propose some perspectives of future work.

5 CONCLUSION AND PERSPECTIVES

In this study, we investigate the application of a new type of Newton method to the FWI problem, named as truncated Newton methods. Instead of computing directly an approximation of the inverse Hessian operator, the linear system associated with the computation of the Newton descent direction is solved through a matrix-free linear conjugate gradient solver. This first requires to be able to compute Hessian-vector products. The second-order adjoint method yields an efficient algorithm for computing these quantities: only two ad-

ditive wave propagation problems are needed. Second, based on the work of Eisenstat and Walker (1994), an adaptive stopping criterion is defined to control the accuracy required for the resolution of the inner linear systems. A forcing term related to the accuracy of the local quadratic approximation of the misfit function is computed. Third, a preconditioner can be used to improve the convergence speed of each linear system resolution through the conjugate gradient method.

The numerical experiments presented in this study aims at emphasizing the important role of the inverse Hessian operator, and investigating if the truncated Newton method can help to better account for it. Therefore we compare seven minimization methods, namely the preconditioned nonlinear conjugate gradient method, the l-BFGS method, the truncated Newton method in the Gauss-Newton approximation, the truncated Newton method using the full Hessian operator, and the preconditioned version of the two latter algorithms. As a preconditioner, we use the pseudo-Hessian diagonal matrix introduced by Shin *et al.* (2001). The same preconditioner is applied for the nonlinear conjugate gradient method and the two preconditioned truncated Newton methods.

The first experiment we consider is a canonical case study for which high amplitude multi-reflected waves dominate the data. Two structures with large amplitude P-wave velocity are buried at few meters depth in a slow homogeneous background. The contrast between the structures and the background and the proximity of the two structures generate high amplitude reflected waves. In this context, conventional minimization methods such as the preconditioned nonlinear conjugate gradient or the l-BFGS method fail to correctly reconstruct the structure. Only the truncated Newton method, using the full Hessian operator, is able to recover the shape and the location of the two structures. The use of the pseudo-Hessian-based diagonal preconditioner improves the convergence speed of the truncated Newton method and helps to recover the amplitude of the P-wave velocity in the structures. This test clearly indicates that the interpretation of the multi-reflected waves is only correctly performed when using the information of the full Hessian operator through the truncated Newton method strategy.

The second experiment which is proposed is the Marmousi II test case (Martin *et al.* 2006). This conventional seismic exploration test case exhibits less complexity than in the previous case. Consequently, the seven optimization methods generate very similar estimations of the exact model. The differences between the minimization schemes come from the convergence speed. In this case, using no preconditioning, the truncated Newton method, either in the Gauss-Newton approximation or in the full Hessian context, is outperformed by the conventional methods (preconditioned nonlinear conjugate gradient, l-BFGS method). The use of the pseudo-Hessian diagonal as a preconditioning matrix greatly improves the convergence of the truncated Newton methods, which in this case outperform the l-BFGS and the preconditioned nonlinear conjugate gradient methods. In the context of this experiment, the most efficient method seems however to include the information embedded in the preconditioner within the l-BFGS iterations, as proposed by Nocedal and Wright (2006).

The third case study which is presented originates from the BP 2004 model (Billette and Brandsberg-Dahl 2004). We focus on the left part of this model, which is inspired from the geology of the Mexico Gulf. Complex salt structures with large P-wave velocity generates high amplitude reflected waves, that are reflected back by the free-surface at the top of the model. This seismic exploration case study thus exhibits the same type of complex patterns that have been identified in the first canonical experiment. In this case, the l-BFGS method fails to converge, as well as non-preconditioned versions of the truncated Newton method. Only using the diagonal of the pseudo-Hessian matrix as a preconditioner makes possible to minimize the misfit function. The integration of the preconditioning matrix within the l-BFGS method is not enough in this case to make the method converge. Therefore, only the preconditioned nonlinear conjugate gradient method and the preconditioned truncated Newton and Gauss-Newton methods provide reliable P-wave model estimations. For highest frequencies, the increasing importance of multi-reflected waves make the full Hessian operator necessary to reach the convergence.

We conclude from this study that for the interpretation of reasonably complex data, the use of a sophisticated approximation of the inverse Hessian is not crucial. In this case,

the truncated Newton method, properly preconditioned, yields only a slight improvement in terms of convergence speed of the minimization process compared to standard methods. However, when the complexity of the data to be inverted increases, for instance when high amplitude reflected waves are recorded, the truncated Newton method seems to improve much notably the subsurface parameter estimation. In this case, the full inverse Hessian operator should be accounted for, since both its first-order and second-order part have non-negligible amplitude. This can only be done through the use of the truncated Newton method.

This first analysis should be confirmed by applying the truncated Newton method to real data, for instance to the 2D acoustic Valhall dataset (Prieux *et al.* 2011). The application of the method to elastic FWI and multi-parameter inversion should also be carried out. The intrinsic scaling of the model update related to inverse Hessian operator is indeed crucial for multi-parameter inversion of attenuation, anisotropic parameters, density, and shear-wave velocity. The truncated Newton method could help to mitigate the coupling effect between parameters. Extension to 3D FWI requires a careful implementation to manage memory issues related to the matrix-free Hessian-vector products formalism. Because of the computational expense of the resolution of a 3D wave propagation problem, the benefits related to the potential fast convergence of the truncated Newton minimization scheme could be substantial.

From a methodological point of view, the integration of regularization terms should be investigated. It is well known that conventional regularization method such as the Tikhonov regularization shift the smallest eigenvalues of the Hessian operator and tends to limit the influence of its indefinite part. In this context, the Gauss-Newton approximation of the Hessian could be more accurate, as the l-BFGS approximation and the pseudo-Hessian approximation. The benefits of the full Hessian approach could therefore be minored in this context.

The three experiments presented in section 4 also demonstrates the importance of the preconditioner for the truncated Newton method to converge rapidly. The construction or

improvement of the Shin preconditioner should thus be devised. The application of the new pseudo-Hessian matrix proposed by Choi and Shin (2008) could for instance be investigated. It is also possible to introduce more sophisticated linear solvers to replace the conjugate gradient method in charge of solving the inner linear system within the truncated Newton algorithm: deflated conjugate gradient or Lanczos iterative methods can be considered. Finally, the implementation of the truncated Newton method which is presented here is based on a linesearch globalization technique. The use of trust-regions method, as proposed by Steihaug (1983) could also be considered to improve the method.

ACKNOWLEDGMENTS

This study was funded by the SEISCOPE consortium ($http://seiscope.oca.eu$), sponsored by BP, CGG-VERITAS, ENI, EXXON-MOBIL, PETROBRAS, SAUDI ARAMCO, SHELL, STATOIL and TOTAL. This study was granted access to the HPC facilities of CIMENT (Université Joseph Fourier Grenoble), and of GENCI-CINES under Grant 2011-046091 of GENCI (Grand Equipement National de Calcul Intensif). The authors would also like to thank S. Bellavia and S. Gratton for useful discussions and advices

REFERENCES

Amestoy, P., Duff, I. S., and L'Excellent, J. Y., 2000. Multifrontal parallel distributed symmetric and unsymmetric solvers, *Computer Methods in Applied Mechanics and Engineering*, **184**(2-4), 501–520.

Berenger, J-P, 1994. A perfectly matched layer for absorption of electromagnetic waves, *Journal of Computational Physics*, **114**, 185–200.

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

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A., 2006. *Numerical Optimization, Theoretical and Practical Aspects*, Springer series, Universitext.

Brenders, A. J. and Pratt, R. G., 2007. Full waveform tomography for lithospheric imaging: results from a blind test in a realistic crustal model, *Geophysical Journal International*, **168**, 133–151.

Brossier, R., Operto, S., and Virieux, J., 2009. Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion, *Geophysics*, **74**(6), WCC63–WCC76.

Brossier, Romain, Operto, Stéphane, and Virieux, Jean, 2010. Which data residual norm for robust elastic frequency-domain full waveform inversion?, *Geophysics*, **75**(3), R37–R46.

Bunks, C., Salek, F. M., Zaleski, S., and Chavent, G., 1995. Multiscale seismic waveform inversion, *Geophysics*, **60**(5), 1457–1473.

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

Choi, Y. and Shin, C., 2008. Frequency-Domain Elastic Full Waveform Inversion Using the New Pseudo-Hessian Matrix: Experience Of Elastic Marmousi 2 Synthetic Data, *Bulletin of the Seismological Society of America*, **98**(5), 2402–2415.

Claerbout, J.F., 1971. Towards a unified theory of reflector mapping, *Geophysics*, **36**, 467–481.

Claerbout, J. F., 1976. *Fundamentals of Geophysical Data Processing*, McGraw-Hill Book Co.

Dai, Y. and Yuan, Y., 1999. A nonlinear conjugate gradient method with a strong global convergence property, *SIAM Journal on Optimization*, **10**, 177–182.

Eisenstat, S. C. and Walker, H. F., 1994. Choosing the forcing terms in an inexact Newton method, *SIAM Journal on Scientific Computing*, **17**, 16–32.

Epanomeritakis, I., Akçelik, V., Ghattas, O., and Bielak, J., 2008. A Newton-CG method for large-scale three-dimensional elastic full waveform seismic inversion, *Inverse Problems*, **24**, 1–26.

Fichtner, A. and Trampert, J., 2011. Hessian kernels of seismic data functionals based upon adjoint techniques, *Geophysical Journal International*, **185**(2), 775–798.

Gao, F., Levander, A. R., Pratt, R. G., Zelt, C. A., and Fradelizio, G. L., 2006. Waveform tomography at a groundwater contamination site: surface reflection data, *Geophysics*, **72(5)**, G45–G55.

Gauthier, O., Virieux, J., and Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: numerical results, *Geophysics*, **51**(7), 1387–1403.

Hustedt, B., Operto, S., and Virieux, J., 2004. Mixed-grid and staggered-grid finite difference methods for frequency domain acoustic wave modelling, *Geophysical Journal International*, **157**, 1269–1296.

Lailly, P., 1983. The seismic inverse problem as a sequence of before stack migrations, in *Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia*, edited by R. Bednar and Weglein, pp. 206–220.

Lailly, P., Rocca, F., and Versteeg, R., 1991. Synthesis of the Marmousi workshop, in *The Marmousi Experience*, pp. 169–194.

Le Dimet, F. X., Navon, I. M., and Daescu, D. N., 2002. Second-order information in data assim-

ilation, *American Meteorological Society*, **184**(2-4), 501–520.

Lions, J. L., 1968. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris.

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

Métivier, L., 2009. Utilisation des équations euler-pml en milieu hétérogène borné pour la résolution d'un problème inverse en géophysique., *ESAIM: Proc.*, **27**, 156–170.

Morales, J. L. and Nocedal, J., 2000. Enriched methods for large-scale unconstrained optimization, *Computational Optimization and Applications*, **21**, 143–154.

Nash, S. G., 2000. A survey of truncated Newton methods, *Journal of Computational and Applied Mathematics*, **124**, 45–59.

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

Operto, S., Ravaut, C., Improta, L., Virieux, J., Herrero, A., and Dell'Aversana, P., 2004. Quantitative imaging of complex structures from dense wide-aperture seismic data by multiscale traveltime and waveform inversions: a case study, *Geophysical Prospecting*, **52**, 625–651.

Operto, S., Virieux, J., and Dessa, J. X., 2005. High-resolution crustal seismic imaging from OBS data by full-waveform inversion: application to the eastern-Nankai trough, in *EOS Trans. AGU*, vol. 86, American Geophysical Union.

Operto, S., Virieux, J., Dessa, J. X., and Pascal, G., 2006. Crustal imaging from multifold ocean bottom seismometers data by frequency-domain full-waveform tomography: application to the eastern Nankai trough, *Journal of Geophysical Research*, **111**(B09306), doi:10.1029/2005JB003835.

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

Plessix, R. E. and Perkins, C., 2010. Full waveform inversion of a deep water ocean bottom seismometer dataset, *First Break*, **28**, 71–78.

Plessix, R.-E., Baeten, G., de Maag, J. Willem, and ten Kroode, F., 2012. Full waveform inversion and distance separated simultaneous sweeping: a study with a land seismic data set, *Geophysical Prospecting*, **60**, 733 – 747.

Pratt, R. G., 1990. Inverse theory applied to multi-source cross-hole tomography. part II : elastic wave-equation method, *Geophysical Prospecting*, **38**, 311–330.

Pratt, R. G. and Worthington, M. H., 1990. Inverse theory applied to multi-source cross-hole tomography. Part I: acoustic wave-equation method, *Geophysical Prospecting*, **38**, 287–310.

Pratt, R. G., Shin, C., and 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.I., and Kommedal, J.H., 2011. On the footprint of anisotropy on isotropic full waveform inversion: the Valhall case study, *Geophysical Journal International*, **187**, 1495–1515.

Ravaut, C., Operto, S., Improta, L., Virieux, J., Herrero, A., and dell'Aversana, P., 2004. Multi-scale imaging of complex structures from multi-fold wide-aperture seismic data by frequency-domain full-wavefield inversions: application to a thrust belt, *Geophysical Journal International*, **159**, 1032–1056.

S. G. Nash, J. Nocedal, 1991. A numerical study of the limited memory bfgs method and truncated newton method for large scale optimization, *Siam Journal on Optimization*, **1**, 358–372.

Saad, Y., 2003. *Iterative methods for sparse linear systems*, SIAM, Philadelphia.

Shin, C., Jang, S., and Min, D. J., 2001. Improved amplitude preservation for prestack depth migration by inverse scattering theory, *Geophysical Prospecting*, **49**, 592–606.

Sirgue, L. and Pratt, R. G., 2004. Efficient waveform inversion and imaging : a strategy for selecting temporal frequencies, *Geophysics*, **69**(1), 231–248.

Sirgue, L., Etgen, J. T., and Albertin, U., 2008. 3D Frequency Domain Waveform Inversion using Time Domain Finite Difference Methods, in *Proceedings 70th EAGE, Conference and Exhibition, Roma, Italy*, p. F022.

Steihaug, T., 1983. The conjugate gradient method and trust regions in large scale optimization, *Siam Journal on Numerical Analysis*, **20**(3).

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.

Vigh, Denes, Starr, Bill, Kapoor, Jerry, and Li, Hongyan, 2010. 3d full waveform inversion on a gulf of mexico waz data set, *SEG Technical Program Expanded Abstracts*, **29**(1), 957–961.

Virieux, J. and Operto, S., 2009. An overview of full waveform inversion in exploration geophysics, *Geophysics*, **74**(6), WCC127–WCC152.

Wu, R. S. and Toksöz, M. N., 1987. Diffraction tomography and multisource holography applied to seismic imaging, *Geophysics*, **52**, 11–25.