An optimal transport approach for seismic tomography: Application to 3D full waveform inversion

L. Métivier [1,2], R. Brossier[2], Q. Mérigot[3], E. Oudet[1], J. Virieux[2]

[1] Laboratoire Jean Kuntzmann (LJK), Univ. Grenoble Alpes, CNRS, France
[2] Institut des sciences de la Terre (ISTerre), Univ. Grenoble Alpes, France
[3] Laboratoire CEREMADE, Univ. Paris-Dauphine, CNRS, France

E-mail: `ludovic.metivier@ujf-grenoble.fr`

Abstract. The use of optimal transport distance has recently yielded significant progress in image processing for pattern recognition, shape identification, and histograms matching. In this study, the use of this distance is investigated for a seismic tomography problem exploiting the complete waveform; the full waveform inversion. In its conventional formulation, this high resolution seismic imaging method is based on the minimization of the L^2 distance between predicted and observed data. Application of this method is generally hampered by the local minima of the associated L^2 misfit function, which correspond to velocity models matching the data up to one or several phase shifts. Conversely, the optimal transport distance appears as a more suitable tool to compare the misfit between oscillatory signals, for its ability to detect shifted patterns. However, its application to full waveform inversion is not straightforward, as the mass conservation between the compared data can not be guaranteed, a crucial assumption for optimal transport. In this study, the use of a distance based on the Kantorovich-Rubinstein norm is introduced to overcome this difficulty. Its mathematical link with the optimal transport distance is made clear. An efficient numerical strategy for its computation, based on a proximal splitting technique, is introduced. We demonstrate that each iteration of the corresponding algorithm requires solving the Poisson equation, for which fast solvers can be used, relying either on the fast Fourier transform or on multigrid techniques. The development of this numerical method make possible applications to industrial scale data, involving tenths of millions of discrete unknowns. The results we obtain on such large scale synthetic data illustrate the potentialities of the optimal transport for seismic imaging. Starting from crude initial velocity models, optimal transport based inversion yields significantly better velocity reconstructions than those based on the L^2 distance, in 2D and 3D contexts.

Keywords: optimal transport, full waveform inversion, non-smooth convex optimization

Submitted to: *Inverse Problems*

1. Introduction

1.1. Optimal transport in imaging sciences

Optimal transport finds its roots in the work of the French engineer Gaspard Monge (1780), in an attempt to devise the most efficient strategy to move sand from a given location to a building site. In this sense, the problem is formulated as an optimal assignment problem, given a cost measuring the effort for moving mass units from one location to the other. Almost two hundred years later, [Kantorovich, 1942] introduced a relaxation of the optimal transport problem, where the mass is allowed to split, giving the first proofs of existence for the (relaxed) optimal transport problem. Considering two probability distributions μ and ν on two separable metric spaces X and Y, the optimal transport problem amounts to find, among all the mapping from X to Y transforming μ into ν, the one minimizing a cost function measuring the effort to achieve this mapping in terms of elementary displacement from X to Y. In the last fifteen years, this field of research has been put on the front scene through the work of many mathematicians, as testified by the books of [Villani, 2003, Villani, 2008, Ambrosio et al., 2008, Santambrogio, 2015]. In particular, the metric underlain by the optimal transport distance is used to establish new existence results of solution to nonlinear partial differential equations such as the Boltzmann equations. Applications of optimal transport in geometry processing and image processing have also been investigated in the last years. Contrast and color mappings are well known applications in image processing [Ferradans et al., 2014], but a more complete list would include pattern recognition and shape classification, texture synthesis and texture mixing, and image smoothing among others (see [Lellmann et al., 2014] and references therein for a detailed state-of-the-art in image processing). One of the interesting properties of optimal transport distances relies on the fact that they give the possibility to perform comparisons between images based on global properties of the images, and not only local, pixel-by-pixel comparisons, underlain by the use of conventional L^p distances. This property is the main reason for the interest of these distances for applications in pattern recognition and shape identification.

1.2. Full waveform inversion: a seismic imaging method

Full waveform inversion (FWI) is based on the minimization of the misfit between observed and predicted data [Lailly, 1983, Tarantola, 1984]. The observed data is a collection of seismic signals recorded at the surface, after the propagation of waves either generated by a controlled source at the exploration scale, with applications for oil and gas industry, or earthquakes, from regional to global scales, with applications in seismology. The predicted data is computed through the numerical solution of a wave propagation problem, from simple acoustic modeling to more realistic visco-elastic anisotropic modeling. The minimization of the misfit between predicted and observed data is performed on a selected set of discrete parameters, including wave propagation velocities, density, attenuation, and anisotropy parameters. A review of FWI and its application is provided in [Virieux and Operto, 2009].

FWI is now routinely used, both in the academy and the industry, mainly for the reconstruction of wave propagation velocities [Fichtner et al., 2010, Tape et al., 2010, Peter et al., 2011, Sirgue et al., 2010, Plessix and Perkins, 2010, Zhu et al., 2012, Warner et al., 2013, Vigh et al., 2014, Borisov and Singh, 2015, Operto et al., 2015]. Its advantage over conventional tomography methods, which are based on the interpretation of selected arrival times only, is its ability to recover higher resolution estimates. As the method involves millions (2D acoustic) to tens of billions (3D elastic) of discrete unknowns, it is based on a local minimization of the misfit, which iteratively updates an initial estimation of the wave velocity. In its conventional formulation, the misfit is computed as the L^2 distance between observed and predicted data. This yields a significant difficulty: the corresponding misfit function has multiple local minima. The initial guess of the solution should thus be sufficiently close from the global minimum for the method to converge to the desired estimation. The reason for the presence of these local minima is well understood. Smooth, macro-scale modifications of the velocity structure mainly impact the oscillatory seismic waveform through modifications of the travel-times, resulting in time shifted waveforms [Jannane et al., 1989]. However, the L^2 norm is not an appropriate tool for capturing these time-shifts. Seeing the waveform as a purely oscillatory signal, a local minimum is reached each time the velocity model matches the data with one or several phase shifts. This phenomenon is referred to as cycle skipping or phase ambiguity in the FWI community. The use of the L^2 norm thus requires to start from an initial velocity model kinematically compatible with the data, in the sense that the observed data is matched within half a phase, to prevent cycle skipping.

This strong requirement is of course a difficulty for the application of FWI as such an initial model may not be always available. This difficulty has prompted numerous investigations attempting to overcome this restriction. Multi-scale frequency approaches have been first introduced to mitigate the sensitivity to cycle skipping by enlarging the phase in the first steps of the method [Bunks et al., 1995, Pratt, 1999]. This strategy is limited by the lowest available frequency, which is most of the time not low enough to sufficiently constrain the model and avoid cycle skipping. Image domain techniques have also been proposed (see [Symes, 2008] and references therein for a review). The methods are based on improving the consistency of the velocity model through the re-focusing of migrated images along a dimension introduced artificially in the imaging condition (time lag, subsurface offset, illumination angle). These methods are able to generate smooth updates of the initial background models. However the high computation cost related to the repeated construction of image volumes through migration algorithms seems to have precluded their use in 3D configurations up to now. In addition, these methods rely exclusively on reflected waves as they are based on the construction of reflectivity images. Data-domain techniques represent a third type of strategies to mitigate cycle skipping. These methods are based on the modification of the misfit measurement: cross-correlation [Luo and Schuster, 1991], and later on warping techniques [Hale, 2013, Ma and Hale, 2013], have been proposed to access directly the time shifts between seismograms without travel-time picking. Misfit functions based on the instantaneous phase and envelope of the signals have been investigated in seismology [Fichtner et al., 2008, Bǒzdag et al., 2011]. Recently, deconvolution approaches [Luo and Sava, 2011, Warner and Guasch, 2014] have also been promoted for FWI, where Wiener filters are used to match observed and predicted data. All these strategies share the common purpose to produce more convex misfit functions, possibly at the expense of the high resolution expected from full waveform inversion. Another drawback of these strategies is related to the necessity to design a suitable penalization function to maximize the energy at zero time shift (or to minimize the energy away from zero time shift), which may require non-trivial parameter tuning. In addition, identification and windowing of the data may be required to robustly estimate the time shifts between traces, which is a difficult task for seismic data acquired at the exploration scale.

From an optimization point of view, the convergence to local minima results from the use of local optimization techniques. As a consequence, since the introduction of FWI, several attempts to apply global optimization schemes to FWI based on Monte Carlo techniques or genetic algorithms have been performed. The crucial issue for the success of these methods is the design of a suitable subsurface velocity parameterization allowing to drastically reduce the number of unknowns, which is not straightforward. Indeed, current high performance computing devices may provide the capability to perform such a global model space exploration for problems involving no more than several hundred discrete parameters, which is several order of magnitudes lower than the standard FWI problem size for realistic applications. In [Diouane et al., 2016], a recent example of such an attempt is proposed, where the model parameterization is based on a Discrete Cosine Transform and a step function transform, and the global optimization method is based on the CMA-ES algorithm [Hansen, 2006].

1.3. Contribution

In a recent study, the use of an optimal transport distance in the framework of FWI has been suggested as a novel data-domain technique [Engquist and Froese, 2014]. The main motivation relies on the fact that the optimal transport distance, as a global comparison tool, is particularly suited to capture time shifts between signals. This property is emphasized in [Engquist and Froese, 2014] on 1D time-shifted Ricker signals: the optimal transport distance is a convex, nearly quadratic function of the time shift.

Building up on this idea, we have proposed a methodology for using an optimal transport distance within FWI in realistic 2D configurations [Métivier et al., 2016]. In this context, the seismic data is interpreted as a collection of 2D images, one for each seismic source. The signal recorded by each receiver is gathered in a 2D panel depending on the physical location of the receiver at the surface. This organization of the data is used for years by geophysicists for analyzing the data as it allows to easily identify all the seismic events (direct propagation, refraction, pre-and-post critical reflection, conversion). However, this data representation is rarely accounted for in the inversion algorithm. The L^2 distance performs pixel-by-pixel comparisons, while cross-correlation and deconvolution approaches rely on a computation of time-shifts trace by trace. Nonetheless, macro-scale variations of the velocity structures impacts the data by shifting the seismic events not only along the time axis but also along the receiver (space) axis. Warping techniques [Hale, 2013] appear to us as a first attempt to account for these shifts considering the whole seismogram. Tracking these shifts using a 2D optimal transport distance should allow to account for these shifts more robustly, similarly as in pattern recognition applications. Applications of this strategy to 2D realistic case studies have confirmed its interest.

The methodology we have proposed [Métivier et al., 2016] is based on a modified dual Kantorovich problem. This formulation allows for the non conservation of the mass between the data, which is a crucial point: standard optimal transport distance relies on the assumption that the total mass is conserved between the compared images. For FWI applications, the concept of mass used in imaging science is to be understood as the intensity of the recorded signal at a given time by a given receiver. In this context, there is no reason for the total mass of the observed data to be equal to the total mass of the predicted data for seismic applications. For instance, if reflections are missing in the predicted data (due to the absence of the corresponding reflectors in the subsurface model), the predicted data will contain less mass than the observed one. Noise also contaminates the observed data in real applications, which in essence unpredictably contributes to its total mass.

In the present study, we propose to further analyze the mathematical foundation of this modified dual strategy. We show that the distance which is used is actually related to the Kantorovich-Rubinstein (KR) norm, which has strong connections with the dual Kantorovich problem, as well as with the L^1 distance, as pointed out in [Lellmann et al., 2014]. We also propose a novel numerical strategy for the computation of the Kantorovich-Rubinstein norm, making possible to apply this strategy to realistic size 3D data sets for the first time. The corresponding problem is formulated as a non-smooth convex optimization problem, solved with the Simultaneous Descent Method of Multipliers (SDMM), a proximal splitting strategy [Combettes and Pesquet, 2011]. We prove that the linear system to be solved at each iteration of SDMM corresponds to the solution of the Poisson's equation with homogeneous Neumann boundary conditions, which can be solved in linear complexity through multigrid techniques [Brandt, 1977]. This numerical strategy is implemented within the framework of time-domain acoustic FWI. An analysis of the KR distance is first performed for the comparison of time-signals to investigate the dependence of this distance with respect to time shifts, as it is performed in [Engquist and Froese, 2014]. The KR distance exhibits a single global minimum in this case. However, it appears not to be a convex function of the time-shift, as it seems non-differentiable at its minimum. This feature confirms the link between KR and L^1 distances. The shape of the L^2 and KR misfit function is then compared in a 2D more realistic context, for a bi-dimensional problem. Using the KR distance appears to level-up the secondary valleys of the misfit function, reducing the risk of being trapped in a local minimum following a gradient-based local optimization method. A 2D FWI application to the Marmousi model is presented, to emphasize the interest of the strategy, and its robustness to the presence of noise. Finally, a fully 3D experiment performed on the overthrust SEG/EAGE model is proposed. From these experiments, the use of the optimal transport distance appears as an interesting tool for FWI as it seems to mitigate efficiently the cycle skipping issue, while the numerical strategy which is proposed appears as feasible for realistic size 3D problems.

1.4. Structure of the paper

The study is divided in six sections. In Section 2, the mathematical background of optimal transport is reviewed quickly, as well as the link between the KR norm and optimal transport. In Section 3, the numerical strategy we set up to compute the KR norm for large scale problems is presented. In Section 4, the formulation of the FWI problem using the KR norm is presented. In Section 5, we present three different case studies, from 1D to 3D, emphasizing the main properties of FWI based on the KR distance. Concluding remarks and perspectives are given in Section 6.

1.5. Notations

In what follows, X and Y denote two metric spaces. The space of probability distribution on X and Y are denoted by $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ respectively. The push-forward distribution of $\mu \in \mathcal{P}(X)$ by the mapping T

$$\left\{ \begin{array}{ccc} X & \longrightarrow & Y \\ T : x & \longrightarrow & T(x), \end{array} \right. \tag{1}$$

is denoted by $T_{\#}\mu \in \mathcal{P}(Y)$, such that for any measurable set $A \subset Y$, we have

$$\left(T_{\#}\mu\right)(A) \equiv \mu\left(T^{-1}(A)\right) = \nu(A). \tag{2}$$

For the numerical computation of the KR norm, we work with a compact subset of $\mathbb{R}^3$ denoted by Ω, such that

$$\Omega = \prod_{s \in \{x,y,z\}} [a_s ; b_s]. \tag{3}$$

The compact Ω is discretized using a Cartesian mesh with constant step sizes in each direction $h_s,\ s \in \{x,y,z\}$ such that

$$h_s = \frac{b_s - a_s}{N_s}, \quad s \in \{x,y,z\}, \tag{4}$$

where N_s is the number of grid points in the direction $s \in \{x,y,z\}$. A point on the mesh is denoted by $(x_i, y_j, z_k) \in \mathbb{R}^3$ such that

$$\left\{ \begin{array}{l} x_i = a_x + (i-1)h_x \\ y_j = a_y + (j-1)h_y \\ z_k = a_z + (k-1)h_z. \end{array} \right. \tag{5}$$

We use the standard discrete notations such that, $\forall\ \varphi(x,y,z)$

$$\begin{array}{cccc} & \Omega & \longrightarrow & \mathbb{R} \\ \varphi : & (x,y,z) & \longrightarrow & \varphi(x,y,z), \end{array} \tag{6}$$

we have

$$\varphi(x_i, y_j, z_k) = \varphi_{ijk}, \tag{7}$$

and we introduce the set of indexes on the mesh

$$\mathcal{A} = \left\{(i,j,k) \in \mathbb{N}^3,\ \ 1 \leq i \leq N_x,\ \ 1 \leq j \leq N_y,\ \ 1 \leq k \leq N_z\right\}. \tag{8}$$

The total number of points in the mesh is $N = N_x \times N_y \times N_z$. We will also use the set of indexes $\mathcal{A}_x, \mathcal{A}_y, \mathcal{A}_z$, such that

$$\begin{array}{ll} \mathcal{A}_x & = \left\{(i,j,k) \in \mathbb{N}^3,\ \ 1 \leq i \leq N_x - 1,\ \ 1 \leq j \leq N_y,\ \ 1 \leq k \leq N_z\right\}, \\ \mathcal{A}_y & = \left\{(i,j,k) \in \mathbb{N}^3,\ \ 1 \leq i \leq N_x,\ \ 1 \leq j \leq N_y - 1,\ \ 1 \leq k \leq N_z\right\}, \\ \mathcal{A}_z & = \left\{(i,j,k) \in \mathbb{N}^3,\ \ 1 \leq i \leq N_x,\ \ 1 \leq j \leq N_y,\ \ 1 \leq k \leq N_z - 1\right\}. \end{array} \tag{9}$$

The integer $P \in \mathbb{N}$ is defined by

$$P = \mathrm{card}(\mathcal{A}_x) + \mathrm{card}(\mathcal{A}_y) + \mathrm{card}(\mathcal{A}_z) + N. \tag{10}$$

Finally, we define the space of real matrices with $n \in \mathbb{N}$ rows and $p \in \mathbb{N}$ columns by $\mathbb{M}_{n,p}(\mathbb{R})$. The Kronecker product between two matrices $A \in \mathbb{M}_{n,p}(\mathbb{R})$ and $B \in \mathbb{M}_{q,r}(\mathbb{R})$ is defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1p}B \\ \vdots & & \vdots \\ a_{n1}B & \dots & a_{np}B \end{pmatrix} \in \mathbb{M}_{nq,pr}(\mathbb{R}). \tag{11}$$

2. Optimal transport and the Kantorovich-Rubinstein norm

We start by recalling the standard Monge formulation for the optimal transport problem. Given two probability distributions $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, and a cost function $c(x, y)$

$$\left\{ \begin{array}{lll} X \times Y & \longrightarrow & \mathbb{R}_+ \\ c : (x, y) & \longrightarrow & c(x, y), \end{array} \right. \tag{12}$$

the optimal transport problem is defined as

$$\inf_T \left\{ \int c(x, T(x)) d\mu(x), \;\; T_\# \mu = \nu \right\}. \tag{13}$$

The constraint $T_\# \mu = \nu$ indicates that the push forward distribution $T_\# \mu$ of μ by the mapping T is equal to the distribution ν. The optimal transport problem can thus be interpreted as determining the mapping T which transports the distribution μ onto the distribution ν in the sense of the equation (2), which minimizes the cost defined in (13), for a given cost function $c(x, y)$.

The problem (13) is difficult to solve, in particular because of the constraint (2). A relaxation of this problem has been proposed by [Kantorovich, 1942], under the linear programming problem

$$\inf_\gamma \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y), \;\; s.c. \;\; \gamma \in \Pi(\mu, \nu) \right\}, \tag{14}$$

where the ensemble of transport plans $\Pi(\mu, \nu)$ is defined by

$$\Pi(\mu, \nu) = \left\{ \gamma \in \mathcal{P}(X \times Y), \;\; (\pi_X)_\# \gamma = \mu, \;\; (\pi_Y)_\# \gamma = \nu \right\}. \tag{15}$$

The operators π_X and π_Y are the projectors on X and Y respectively. The problem (14) generalizes (13) in the sense that, instead of considering a mapping T transporting each particle of the distribution μ to the distribution ν, it considers all pairs (x, y) of the space $X \times Y$ and for each pair defines how many particles of μ go from x to y. While in the context of the Monge formulation (13), each point of the space X has only one possible destination on Y, given by $T(x)$, in the context of the Kantorovich formulation (14), the particles at point x can have various destinations in Y, given by $\gamma(x, y)$ for $y \in Y$. The constraint (15) ensures that the distribution μ is transported onto the distribution ν. The relaxed problem (14) admits a solution under very mild hypothesis, unlike Monge's problem (13). In addition, when (14) admits a solution T, the measure $\pi = (I, T)_\# \mu$ is a solution of the relaxed problem (13) [Ambrosio, 2003, Pratelli, 2007].

The dual problem of (14) is formulated as

$$\left\{ \begin{array}{l} \displaystyle\max_{\varphi, \psi} \int_X \varphi d\mu + \int_Y \psi d\nu, \\ \varphi \in \mathcal{C}_b(X), \;\; \psi \in \mathcal{C}_b(Y), \;\; \forall (x, y) \in X \times Y, \;\; \varphi(x) + \psi(y) \leq c(x, y), \end{array} \right. \tag{16}$$

where the ensemble of continuous and bounded functions on X (resp. Y) is denoted by $\mathcal{C}_b(X)$ (resp. $\mathcal{C}_b(Y)$). The standard duality result shows that the solution of (14) is equal to the solution of (16) (see for instance [Villani, 2003] or [Santambrogio, 2015] for a complete proof).

In the particular case where $Y = X$ and the cost function $c(x, y)$ is a distance (which will be the case in the remainder of this study), the dual problem (16) can be simplified into

$$\max_\varphi \int_X \varphi d\left(\mu - \nu\right), \;\; \varphi \in \mathrm{Lip}_{1,c}, \tag{17}$$

where $\mathrm{Lip}_{1,c}$ denotes the space of 1 Lipschitz function for the distance $c(x, y)$, *i.e.*

$$\mathrm{Lip}_{1,c} = \left\{ \varphi : x \in X \longrightarrow \mathbb{R}, \;\; \forall (x, x') \; \in X \times X, \;\; |\varphi(x) - \varphi(x')| \leq c(x, x') \right\}. \tag{18}$$

The duality result in this particular case is known as the Kantorovich-Rubinstein theorem [Santambrogio, 2015, chap. 3.2.1]. The modification of the initial Monge problem into the dual form of the Kantorovich problem thus leads to the solution of the linear problem with linear constraints (17).

While, under its primal form (14), the optimal transport problem is defined for probability measures, the dual form is defined for general measures provided they have the same total mass, *i.e* the mass is conserved from the mass distribution μ to the mass distribution ν. An illustration of this requirement follows. Consider the case where μ and ν have a different mass

$$\int_X d(\mu - \nu) \neq 0. \tag{19}$$

Consider the constant function $\varphi_\alpha(x) = \alpha$ for $\alpha \in \mathbb{R}^*$. We have $\varphi_\alpha \in Lip_{1,c}$ and

$$\int_X \varphi_\alpha d\left(\mu - \nu\right) = \alpha \int_X d(\mu - \nu) \neq 0. \tag{20}$$

Thus, the problem (17) has no solution as for any $\varphi \in Lip_{1,c}$, one can find $\varphi_\alpha \in Lip_{1,c}$ such that, for α sufficiently large,

$$\int_X \varphi_\alpha d\left(\mu - \nu\right) > \int_X \varphi d\left(\mu - \nu\right). \tag{21}$$

This simple example suggests a straightforward generalization of the dual Kantorovich problem which remains well posed when the total mass between μ and ν is not conserved. This generalization consists in complementing the 1-Lipschitz constraint with a bound constraint. The problem (17) thus becomes

$$\max_{\varphi} \int_X \varphi d\left(\mu - \nu\right), \quad s.c. \;\; \varphi \in \mathrm{Lip}_{1,c}, \;\; \|\varphi\|_\infty \leq \lambda. \tag{22}$$

In this study, we focus on the particular case for which the distance function $c(x, y)$ is the distance associated with the ℓ_1 norm on $\mathbb{R}^d$

$$c(x, y) = |x - y| = \sum_{i=1}^{d} |x_i - y_i|. \tag{23}$$

Interestingly, with this choice of cost function $c(x, y)$, the generalization (22) corresponds to the definition of the KR norm [Bogachev, 2007]. This norm is defined in the space of Radon measures on Ω, which is the dual space, for the $\|.\|_\infty$ norm, of the space of real valued continuous functions defined on Ω which are zero at infinity, denoted by $(\mathcal{C}_0\left(\Omega, \mathbb{R}\right), \|.\|_\infty))$. Besides the link with optimal transport, the KR norm can also be interpreted as a generalization of the L^1 norm (in a similar sense that the generalization from Total Variation to Total Generalized Variation norms), and shares some properties with the Meyer's G-norm. These similarities are studied in detail in [Lellmann et al., 2014], where the use of the KR norm is proposed as an alternative to the L^1 norm in a TV denoising problem. In the remainder of the paper, the space of 1-Lipschitz functions for the distance induced by the ℓ_1 norm on $\mathbb{R}^d$ and with infinity norm bounded by λ will be denoted by $\mathrm{BLip}_{1,\lambda}$.

3. An efficient numerical strategy for computing the Kantorovich-Rubinstein norm

3.1. Discretization and reduction of the number of constraints

In this section we assume that the dimension d is set to 3. Using the notations introduced in Section 1.5, the problem (22) is discretized as

$$\begin{array}{l} \displaystyle\max_{\varphi_{ijk}} \sum_{ijk} \varphi_{ijk}\left(\mu_{ijk} - \nu_{ijk}\right), \quad s.c. \\ \left\{ \begin{array}{ll} \forall\, (i, j, k), (l, m, n) \in \mathcal{A}^2, & |\varphi_{ijk} - \varphi_{lmn}| < |x_i - x_l| + |y_j - y_m| + |z_k - z_n|, \\ \forall\, (i, j, k) \in \mathcal{A}, & |\varphi_{ijk}| \leq \lambda. \end{array} \right. \end{array} \tag{24}$$

Computing a numerical approximation of the solution of (24) requires the solution of a convex optimization problem involving $O(N^2)$ linear constraints. This would lead to an unacceptable computational time for the large scale problems induced by FWI applications. However, a property of the ℓ_1 norm on $\mathbb{R}^d$ can be used to reduce the number of constraints from $O(N^2)$ to $O(N)$. Note that the following proposition can be rephrased in terms of graph theory by saying that the restriction of the ℓ^1 distance to the grid Z^d is a geodesic distance on the graph over Z^d where two nodes $a, b \in Z^d$ are joined by an edge if and only if $|a - b|_{\ell^1} \leq 1$ (Fig.1).

Proposition.
The two following assertions are equivalent

$$(A1) \quad \forall\,(i,j,k),(l,m,n) \in \mathcal{A}^2,\ |\varphi_{ijk} - \varphi_{lmn}| < |x_i - x_l| + |y_j - y_m| + |z_k - z_n|,$$

$$(A2) \quad \left\{ \begin{array}{ll} \forall\,(i,j,k) \in \mathcal{A}_x, & |\varphi_{i+1,j,k} - \varphi_{ijk}| < |x_{i+1} - x_i|, \\ \forall\,(i,j,k) \in \mathcal{A}_y, & |\varphi_{i,j+1,k} - \varphi_{ijk}| < |y_{j+1} - y_j|, \\ \forall\,(i,j,k) \in \mathcal{A}_z, & |\varphi_{i,j,k+1} - \varphi_{ijk}| < |z_{k+1} - z_k|. \end{array} \right. \tag{25}$$

Proof.
$(A1)$ obviously implies $(A2)$. To prove the reciprocal implication, consider a pair of points on the mesh denoted by u and v, such that

$$u = (x_i, y_j, z_k), \quad v = (x_l, y_m, z_n). \tag{26}$$

A sequence of points $w_q = (x_{i_q}, y_{j_q}, z_{k_q}),\ q = 1, \ldots, M$ can be selected to form a path on the mesh from u to v, such that $w_1 = u$, $w_M = v$, and w_q are all adjacent on the grid, with monotonically varying coordinates. The key is to see that, for such a sequence of points, the ℓ_1 distance on $\mathbb{R}^d$ ensures that

$$||v - u||_1 = \sum_{q=1}^{M} ||w_{q+1} - w_q||_1. \tag{27}$$

(see Fig. 1 for an illustration).

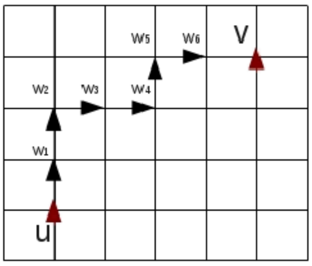

Figure 1. Illustration of the property (27) for a 2D mesh. Considering two points u and v, a sequence of adjacent points $w_1, \ldots, w_6$ with monotonically varying coordinates is found to connect them. Such a sequence always exists and is non-unique.

Now, consider a function φ satisfying $(A2)$. The triangle inequality yields

$$||\varphi(v) - \varphi(u)||_1 \leq \sum_{q=1}^{M} ||\varphi(w_{q+1}) - \varphi(w_q)||_1. \tag{28}$$

As the points w_q are adjacent, the local inequalities described by $(A2)$, satisfied by φ, yield

$$\sum_{q=1}^{M} ||\varphi(w_{q+1}) - \varphi(w_q)||_1 \leq \sum_{q=1}^{M} ||w_{q+1} - w_q||_1. \tag{29}$$

Putting together equations (28), (29) and (27) yields

$$||\varphi(v) - \varphi(u)||_1 \leq ||v - u||_1, \tag{30}$$

or

$$|\varphi_{ijk} - \varphi_{lmn}| < |x_i - x_l| + |y_j - y_m| + |z_k - z_n|, \tag{31}$$

which proves the proposition.

□

Using the equivalence (25), the problem (24) can be rewritten in its equivalent form

$$\max_{\varphi_{ijk}} \sum_{ijk} \varphi_{ijk} \left(\mu_{ijk} - \nu_{ijk}\right), \quad s.c.$$
$$\begin{cases} \forall\,(i,j,k) \in \mathcal{A}_x, & |\varphi_{,i+1,jk} - \varphi_{ijk}| < |x_{i+1} - x_i| = h_x, \\ \forall\,(i,j,k) \in \mathcal{A}_y, & |\varphi_{i,j+1,k} - \varphi_{ijk}| < |y_{j+1} - y_j| = h_y, \\ \forall\,(i,j,k) \in \mathcal{A}_z, & |\varphi_{i,j,k+1} - \varphi_{ijk}| < |z_{k+1} - z_k| = h_z, \\ \forall\,(i,j,k) \in \mathcal{A}, & |\varphi_{ijk}| < \lambda. \end{cases} \tag{32}$$

The problem (32) is equivalent to (24) and only involves $2P = O(N)$ constraints. This reduction of the order of the number of constraints gives the possibility to design an efficient numerical strategy to compute the KR norm.

3.2. Proximal splitting technique for the solution of (32)

3.2.1. The SDMM method

The problem (32) is reformulated as the convex non-smooth problem

$$\max_{\varphi} f_1(\varphi) + f_2(A\varphi), \tag{33}$$

where

$$f_1(\varphi) = \sum_{i,j,k \in \mathcal{A}} \varphi_{ijk} \left(\mu_{ijk} - \nu_{ijk}\right), \quad f_2(\varphi) = i_K\left(\varphi\right), \tag{34}$$

with K the unit hypercube

$$K = \left\{x \in \mathbb{R}^P, \ |x_i| \leq 1, \ i = 1, \ldots P\right\}, \tag{35}$$

i_K the indicator function of K

$$i_K(x) = \left| \begin{array}{lll} 0 & \text{if} & x \in K \\ +\infty & \text{if} & x \notin K, \end{array} \right. \tag{36}$$

and $A \in \mathbb{M}_{P,N}(\mathbb{R})$ a rectangular real matrix with P rows and N columns such that

$$A = \left[D_x \;\; D_y \;\; D_z \;\; \frac{1}{\lambda} I_N\right]^T, \tag{37}$$

where I_N is the real identity matrix of size N and D_x, D_y, D_z are the forward finite differences operators

$$\begin{cases} \left(D_x\varphi\right)_{ijk} = \dfrac{\varphi_{i+1,j,k} - \varphi_{ijk}}{h_x}, \\ \left(D_y\varphi\right)_{ijk} = \dfrac{\varphi_{i,j+1,k} - \varphi_{ijk}}{h_y}, \\ \left(D_z\varphi\right)_{ijk} = \dfrac{\varphi_{i,j,k+1} - \varphi_{ijk}}{h_z}. \end{cases} \tag{38}$$

Convex optimization problems of type (33) involving at least one non differentiable functions, here $f_2(\varphi)$, can be efficiently solved through proximal splitting techniques, such as forward-backward algorithms, Douglas-Rashford splitting, or alternating direction method of multipliers (ADMM) [Combettes and Pesquet, 2011]. From our numerical experiments, among this class of proximal splitting techniques, the simultaneous-direction method of multipliers (SDMM) was found to achieve the fastest convergence. The method can be briefly sketched as follows

Proximal splitting strategies rely on a splitting of the problem in terms of the functions $f_1(\varphi)$ and $f_2(\varphi)$ and the computation of the proximity operators of these two functions (scaled by a positive factor γ). Proximity operators can be seen as a generalization of convex projection operators. They are defined as

$$\text{prox}_f(x) = \arg\min_y f(y) + \frac{1}{2}\|x - y\|_2^2, \tag{39}$$

where the standard Euclidean distance on $\mathbb{R}^d$ is denoted by $\|.\|_2$. For the particular case of the function f_1 and f_2, closed-form formulations can be found such that

$$\text{prox}_{\gamma f_1}(\varphi) = \varphi - \gamma(\mu + \nu), \tag{40}$$

```
$\gamma > 0,\ y_1^0 = 0,\ y_2^0 = 0,\ z_1^0 = 0,\ z_2^0 = 0$;
for $n = 0, 1, \dots$ do
    $\varphi^n = \left(I_N + A^T A\right)^{-1} \left[(y_1^n - z_1^n) + A^T\left(y_2^n - z_2^n\right)\right]$;
    $y_1^{n+1} = \mathrm{prox}_{\gamma f_1}\left(\varphi^n + z_1^n\right)$ ;
    $z_1^{n+1} = z_1^n + \varphi^n - y_1^{n+1}$ ;
    $y_2^{n+1} = \mathrm{prox}_{\gamma i_K}\left(A\varphi^n + z_2^n\right)$ ;
    $z_2^{n+1} = z_2^n + A\varphi^n - y_2^{n+1}$ ;
end
```

Algorithm 1: SDMM method for the solution of the problem (33).

$$\forall i = 1, \dots, P,\ \ \left(\mathrm{prox}_{\gamma f_2}(x)\right)_i = \left(\mathrm{prox}_{i_K}(x)\right)_i = \left| \begin{array}{ll} x_i & \text{if } \ -1 \leq x_i \leq 1 \\ 1 & \text{if } \ x_i > 1 \\ -1 & \text{if } \ x_i < -1. \end{array} \right. \tag{41}$$

Note that the scaling γ only acts on the proximity operator of γf_1 as $\gamma i_K = i_K$. However, the choice of γ should be done with care. Small values for γ could slow down the convergence of the algorithm while too large values can yield numerical instabilities and hamper the convergence. In practice, we choose $\gamma = 0.9$. The closed-form formulations (40) and (41) are inexpensive to compute with an overall complexity in $O(N)$ operations. However, the SDMM algorithm requires the solution of a linear system involving the matrix $I + A^T A$, which is the most time-consuming part of the algorithm.

3.2.2. A Laplacian operator with homogeneous Neumann boundary conditions

Let us inspect in more details what is the form of the matrix $A^T A$. We assume the following ordering of the discrete vector of $\mathbb{R}^N$ using the mapping

$$\begin{array}{lll} i, j, k & \longrightarrow & l = i + (j-1) \times N_x + (k-1) \times N_x \times N_y \\ \mathcal{A} & \longrightarrow & \{1, \dots, N\}. \end{array} \tag{42}$$

With this ordering, the matrices D_x, D_y and D_z may be defined by

$$D_x = I_{N_z} \otimes I_{N_y} \otimes F_{N_x},\ \ D_y = I_{N_z} \otimes F_{N_y} \otimes I_{N_x},\ \ D_z = F_{N_z} \otimes I_{N_y} \otimes I_{N_x}, \tag{43}$$

where the definition of the Kronecker product $\otimes$ defined in Section 1.5 is used, and, for $N_s \in \mathbb{N}, s \in \{x, y, z\}$, the matrix F_{N_s} is defined by

$$F_{N_s} = \frac{1}{h_s} \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{M}_{N_s - 1, N_s}(\mathbb{R}),\ \ s \in \{x, y, z\}. \tag{44}$$

The matrix $A^T A \in \mathbb{M}_{N,N}(\mathbb{R})$ is given by

$$A^T A = \Delta + \frac{1}{\lambda^2} I,\ \ \Delta = D_x^T D_x + D_y^T D_y + D_z^T D_z. \tag{45}$$

The following theorem proves that the matrix Δ actually corresponds to the second-order finite differences discretization of the 3D Laplacian operator defined on Ω with homogeneous Neumann boundary conditions.

Theorem. *The theorem is stated for an arbitrary number of dimension $d \in \mathbb{N}$. In this context Ω is a compact subset of $\mathbb{R}^d$ such that*

$$\Omega = \prod_{i=1}^{d} [a_i, b_i]. \tag{46}$$

The integer $N_i \in \mathbb{N}$, $i = 1, \dots, d$, is the number of discretization points in the direction i, and the discretization step h_i is defined by

$$h_i = \frac{b_i - a_i}{N_i},\ i = 1, \dots, d. \tag{47}$$

Let D_i be the matrix such that

$$D_i = I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes F_{N_i} \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1}. \tag{48}$$

Then the matrix $\Delta = \sum_{i=1}^{d} D_i^T D_i$ corresponds to the second-order finite differences discretization of the Laplacian operator defined on Ω with homogeneous Neumann boundary conditions.

Proof. We start with the case $d = 1$. In this case the second-order finite differences discretization of the Laplacian operator with homogeneous Neumann boundary conditions is the matrix of size N_1

$$L_{N_1} = \frac{1}{h_1^2} \begin{pmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{pmatrix}. \tag{49}$$

In the case $d = 1$, we also have $D_1 = F_{N_1}$, therefore $D_1^T D_1 = F_{N_1}^T F_{N_1}$. A simple matrix product allows us to verify that $F_{N_1}^T F_{N_1} = L_{N_1}$ which proves the theorem for $d = 1$.

For $d > 1$, the second-order finite differences discretization of the Laplacian operator with homogeneous Neumann boundary conditions on the Cartesian mesh $\mathcal{A}$ can be expressed as

$$\Delta = \sum_{i=1}^{d} B_i, \quad B_i = I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes L_{N_i} \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1}. \tag{50}$$

We have

$$D_i^T D_i = \left(I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes F_i \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1} \right)^T I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes F_i \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1}, \tag{51}$$

which is equivalent to

$$D_i^T D_i = \left(I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes F_i^T \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1} \right) I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes F_i \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1}, \tag{52}$$

as the transpose operation is distributive for the Kronecker product. This yields

$$D_i^T D_i = I_{N_d} \otimes \ldots \otimes I_{N_{i+1}} \otimes F_i^T F_i \otimes I_{N_{i-1}} \otimes \ldots \otimes I_{N_1}, \tag{53}$$

using the mixed-product property of the Kronecker product. However, $F_i^T F_i = L_{N_i}$, therefore $B_i = D_i^T D_i$ which completes the proof.

□

The linear system which has to be solved at each iteration of the SDMM algorithm thus corresponds to a second-order finite-differences discretization of the Poisson's problem

$$-(\Delta + 2I)\varphi^n = f, \tag{54}$$

where Δ is a Laplacian operator with homogeneous Neumann boundary conditions. The best numerical strategies for the solution of such problems appears to rely either on Fast Fourier Transform algorithm with $O(N \log N)$ complexity [Swarztrauber, 1974] or multigrid solvers with $O(N)$ complexity [Brandt, 1977]. In this study, we use the multigrid method implemented in the MUDPACK library [Adams, 1989].

The combination of the reduction of the number of constraints using the property of the ℓ_1 distance and the observation that the matrix appearing in the SDMM strategy actually corresponds to the discretization of the Poisson's equation offers the possibility to design an efficient numerical method to compute the KR norm for large scale problems.

4. Application to full waveform inversion

4.1. Formulation of the FWI problem

The observed data corresponding to N_s sources and N_r receivers is denoted by

$$d_{obs,s}(x_r, t), \;\; t \in [0; T], \;\; x_r \in \mathbb{R}^d, \; r = 1, \dots, N_r, \; s = 1, \dots, N_s, \tag{55}$$

where x_r denotes the position of the receivers. In this study, we restrict to the acoustic approximation with constant density, and the predicted data is computed as the solution of

$$\frac{1}{v_P^2} \partial_{tt} p(x, t) - \Delta p(x, t) = f_s(x, t), \tag{56}$$

where T is the recording time, $v_P(x)$ denotes the pressure wave (P-wave) velocity, $p(x, t)$ is the pressure wavefield and $f_s(x, t)$ an explosive seismic source which we assume to be known in this study. For a given P-wave velocity $v_P(x)$ and a given source $f_s(x, t)$, the predicted data is denoted by

$$d_{pred,s}[v_P](t) = \left[p(x_1, t), \dots, p(x_R, t)\right], \tag{57}$$

where $p(x, t)$ is the solution of (56).

Conventional FWI is formulated as the minimization over the set of pressure wave velocity functions $v_P(x)$ of the L^2 distance between $d_{pred}[v_P](x_r, t)$ and $d_{obs}(x_r, t)$ expressed as

$$\min_{v_P} f_{L^2}(v_P) = \frac{1}{2} \sum_{s=1}^{N_s} \|d_{pred,s}[v_P] - d_{obs,s}\|_2^2 \overset{def}{=} \frac{1}{2} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \int_0^T |d_{pred,s}(x_r, t) - d_{obs,s}(x_r, t)|^2 dt. \tag{58}$$

In this study, we investigate the effect of comparing the data using the distance associated with the KR norm, reformulating the FWI problem as

$$\min_{v_P} f_{KR}(v_P) = \sum_{s=1}^{N_s} \|d_{pred,s}[v_P] - d_{obs,s}\|_{KR}. \tag{59}$$

4.2. Gradient computation

The numerical solution to the FWI problem (59) is computed through a local minimization technique. The quasi-Newton l-BFGS algorithm introduced by [Nocedal, 1980] is used to this purpose. This requires the ability to compute the gradient of the misfit function $f_{KR}(v_P)$. The adjoint-state technique provides an efficient strategy to compute this quantity [Lions, 1968, Chavent, 1974, Plessix, 2006] which we present quickly here. For the sake of simplicity, a single seismic source is considered in what follows ($N_s = 1$), the generalization to $N_s > 1$ sources being straightforward by summation. The modeling equation (56) and the relation between the predicted data and the pressure wavefield (57) are rewritten in compact form as

$$F(v_P, p) = 0, \;\; d_{pred}[v_P] = Rp, \tag{60}$$

where

$$F(v_P, p) = \frac{1}{v_P^2} \partial_{tt} p(x, t) - \Delta p(x, t) - f_s(x, t), \tag{61}$$

and R is the extraction operator at the receivers location such that

$$R : p(x, t) \longrightarrow \left[p(x_1, t), \dots, p(x_R, t)\right]. \tag{62}$$

Consider the Lagrangian function

$$L(v_P, p, \lambda) = g(Rp, d_{obs}) + \left(F(v_P, p), \lambda\right)_{\mathcal{W}}, \tag{63}$$

where the L^2 scalar product in the wavefield space is denoted by $(., .)_{\mathcal{W}}$ and g is a general distance function measuring the discrepancy between Rp and d_{obs}. For $\overline{p}(v_P)$ solution of the modeling equation (56), the Lagrangian function is

$$L(v_P, \overline{p}(v_P), \lambda) = g(R\overline{p}(v_P), d_{obs}) \equiv f_g(v_P), \tag{64}$$

where $f_g(v_P)$ denotes a general misfit function associated with the distance g. For the sake of simplicity, the dependence of $\overline{p}$ with v_P is not written explicitly in the sequel. Taking the derivatives of eq. (64) with respect to v_P yields

$$\frac{\partial L(v_P, \overline{p}, \lambda)}{\partial p}\frac{\partial \overline{p}}{v_P} + \left(\frac{\partial F(v_P, \overline{p})}{\partial v_P}, \lambda\right)_{\mathcal{W}} = \frac{\partial g(R\overline{p}, d_{obs})}{\partial v_P} = \nabla f_g(v_P). \tag{65}$$

The adjoint state variable $\overline{\lambda}$ is chosen to cancel the first term of (65), such that the gradient of the misfit function $f(v_P)$ can be computed as the scalar product in the wavefield space

$$\nabla f_g(v_P) = \left(\frac{\partial F(v_P, \overline{p})}{\partial v_P}, \overline{\lambda}\right)_{\mathcal{W}}, \tag{66}$$

avoiding the computation of $\partial \overline{p}/\partial v_P$, the Jacobian operator of the pressure wavefield. Computing this matrix is prohibitively expensive for large-scale applications. The problem of computing efficiently the gradient is thus brought back to the ability to compute $\overline{\lambda}$. Cancelling the first term of (65) gives the adjoint equation

$$\frac{\partial F(v_P, \overline{p})^T}{\partial p}\overline{\lambda} = -\frac{\partial g(R\overline{p}, d_{obs})}{\partial p}. \tag{67}$$

In the context of the wave equation, the operator $F(v_P, p)$ is linear with respect to p and self-adjoint. For this reason, the adjoint wavefield $\overline{\lambda}$ is the solution of the acoustic wave equation (56) backward in time, with a source term depending on the distance function g. Interestingly, this source term is the only quantity involved in the gradient computation which is impacted by the choice of the distance function g (as already noticed for instance in [Brossier et al., 2010, Luo and Sava, 2011]). In the case where the L^2 norm is used, this source term is simply

$$-\frac{\partial \|R\overline{p} - d_{obs}\|^2_{L^2}}{\partial p} = -2R^T(R\overline{p} - d_{obs}). \tag{68}$$

If the KR norm is used, the source term becomes

$$-\frac{\partial \|R\overline{p} - d_{obs}\|_{KR}}{\partial p} = -\frac{\partial\left(\max_{\varphi \in \mathrm{BLip}_{1,\lambda}} \int_X \varphi\left(R\overline{p} - d_{obs}\right)\right)}{\partial p}. \tag{69}$$

We denote by $\overline{\varphi}$ the solution of (17), such that

$$\overline{\varphi} = \arg \max_{\varphi \in \mathrm{BLip}_{1,\lambda}} \int_X \varphi\left(R\overline{p} - d_{obs}\right). \tag{70}$$

Using the almost-everywhere differentiability of concave functions, we thus have for a.e. $\bar{p}$

$$-\frac{\partial \|R\overline{p} - d_{obs}\|_{KR}}{\partial p} = -R^T\overline{\varphi}. \tag{71}$$

Using the KR norm to compare the observed and predicted data within a FWI framework thus only requires to modify the definition of the source term of the adjoint equation from (68) to (71). In terms of implementation, this means that the maximization problem (17) has to be solved only once per source and per iteration of the FWI loop. The value of the criterion reached at the maximum yields the misfit function value, while the function $\overline{\varphi}$ achieving this maximal value is the source term of the adjoint equation. As a consequence, introducing the KR norm within an existing FWI code has only a limited impact on the code structure. With a slight abuse of language, $\overline{\varphi}$ is referred to as the KR residuals in the following.

5. Numerical experiments

5.1. Normalization

In the following experiments, the seismic data are considered to be defined on

$$\Omega = \prod_{s=1}^{d} [0, 1]. \tag{72}$$

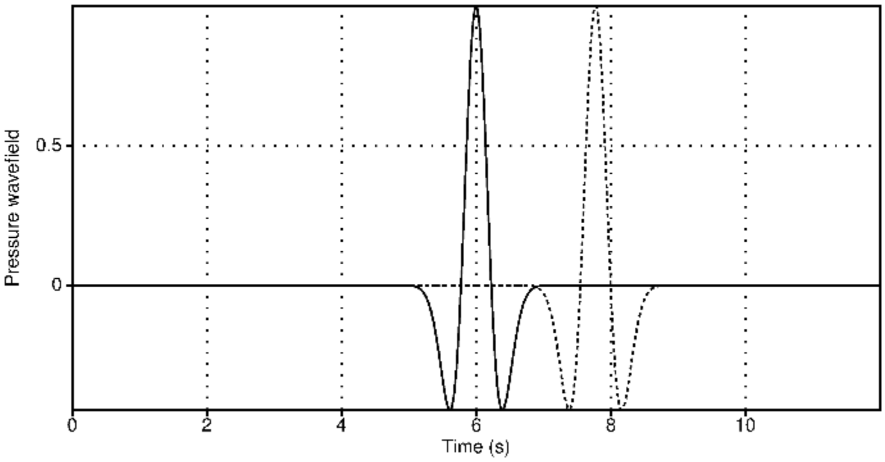

Figure 2. Ricker signal playing the role of observed data (thick line). Example of shifted in time Ricker (dash line).

The discretization steps $h_s,\ s \in \{x, y, z\}$ are thus defined as

$$h_s = \frac{1}{N_s} \tag{73}$$

This amounts to treat the seismic data as square images with no preferred dimensions to perform the comparison. In addition, the parameter λ which controls the infinity norm of the solution to 32 is set to 1. This is a pragmatical choice. In turns, it requires to scale appropriately the initial residuals $d_{pred}[v_P] - d_{cal}$ so that their infinity norm is not too far from 1.

5.2. 1D case study: sensitivity to time shift

We start the numerical investigation with a schematic 1D experiment similar to the one proposed in [Engquist and Froese, 2014]. A Ricker time signal serves as observed data, and the predicted data corresponds to this same Ricker signal, shifted in time (Fig. 2). The L^2 and KR distances, as functions of the time shift, are compared. In this particular case, as the energy of the signal is conserved by the time-shift, the bound constraint on the dual variable φ is relaxed by setting λ to an arbitrary high value. The misfit profiles using the L^2 and KR distances are presented in Figure 3. The misfit based on the L^2 distance presents two local minima, typical of cycle skipping, apart the global minimum. The misfit based on the KR distance presents a single minimum, which indicates a better robustness to the time shift. However, compared to the 2-Wasserstein distance used in [Engquist and Froese, 2014, Fig. 3] which yields a convex misfit function, the misfit function here appears as not differentiable at its minimum and concave. The non-differentiability at the minimum is reminiscent of the L_1 norm, which might not be surprising according to the strong relation between the KR norm and this norm [Lellmann et al., 2014]. Note that, compared with the study of [Engquist and Froese, 2014], the KR norm does not require to separate the signal in its positive and negative part.

A further insight on the KR distance is given in Figure 4 where the L^2 residuals and the KR residuals are compared for the comparison of the two signals presented in the Figure 2. While the L^2 residuals only correspond to the difference between these two signals, the KR residuals appears as an envelope of the L^2 residuals, with positive and negative DC components at the beginning and the end of the signal respectively. In addition, the extrema of the KR residuals present an angular shape typical of the L^1 norm. This particular shape may be related again to the Lipschitz constraint. The similarities between the KR norm and the L^1 norm may question the use of standard quasi-Newton solvers dedicated to the minimization of smooth functions such as the l-BFGS algorithm. However the numerical experiments presented in the next section demonstrate that, for 2D and 3D FWI applications, this property does not preclude the use of these solvers to minimize the KR misfit function, as was already observed previously for the L^1 norm [Brossier et al., 2010].

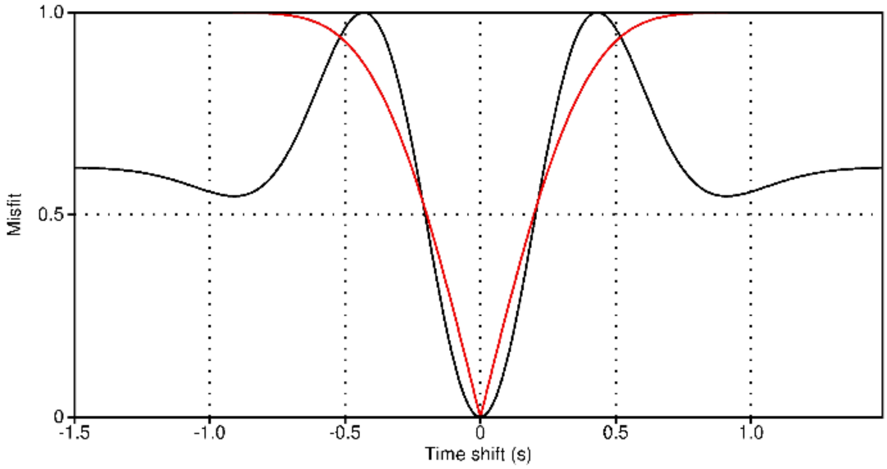

Figure 3. Misfit function depending on the time shift of the Ricker signal, using the L^2 distance (black) and the KR distance (red).

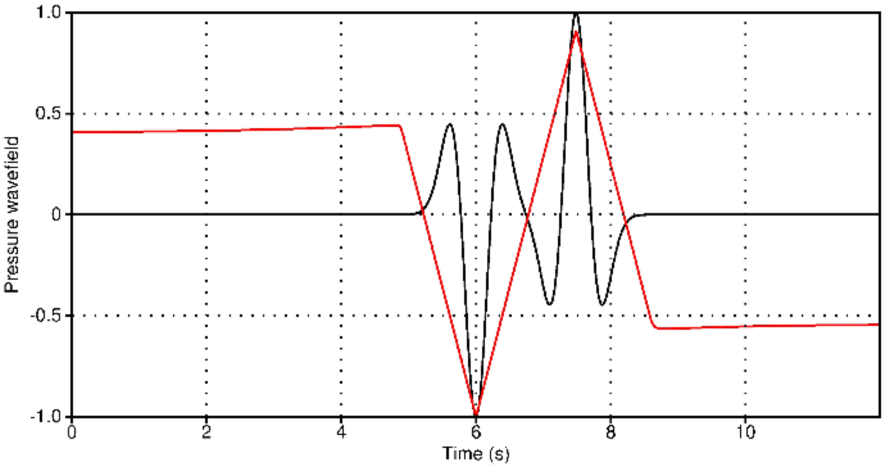

Figure 4. Comparison between the L^2 (black) and KR (red) residuals for the two shifted Ricker signals presented in Figure 2.

5.3. 2D case study: misfit function comparison in a more realistic configuration

In this experiment, a 2D configuration is considered. A fixed-spread surface acquisition is used, constituted of 168 receivers equally spaced each 100 m, at 50 m depth, from $x = 0$ km to $x = 16.85$ km. A single source is located at $x = 8.45$ km. Similar to the experiment presented in [Mulder and Plessix, 2008] to emphasize the local minima of the L^2 misfit function, the velocity model is assumed to vary linearly in depth such that

$$v_P(x, z) = v_{P,0} + \alpha z. \tag{74}$$

The P-wave velocity is thus parameterized by the velocity at the origin $v_{P,0}$ and the velocity vertical gradient α. The reference velocity model is chosen so that $v_{P,0} = 2000$ m s^{-1} and $\alpha = 0.7$ s^{-1}. A reference data set is computed in this model through the solution of the wave equation (56). The L^2 and KR misfit functions are then evaluated on a grid of 41×41 points such that

$$v_{P,0} \in [1750, 2250], \quad \alpha \in [0.4, 0.9], \tag{75}$$

with discretization steps $\Delta v_{P,0} = 12.5$ m s^{-1} and $\Delta\alpha = 0.015$ s^{-1}. For each couple of parameters on this grid, the misfit functions f_{L^2} and f_{KR} are evaluated. The results are presented in Figure 5. Interestingly, the L^2 misfit function presents two narrow valleys of attraction, on both sides of

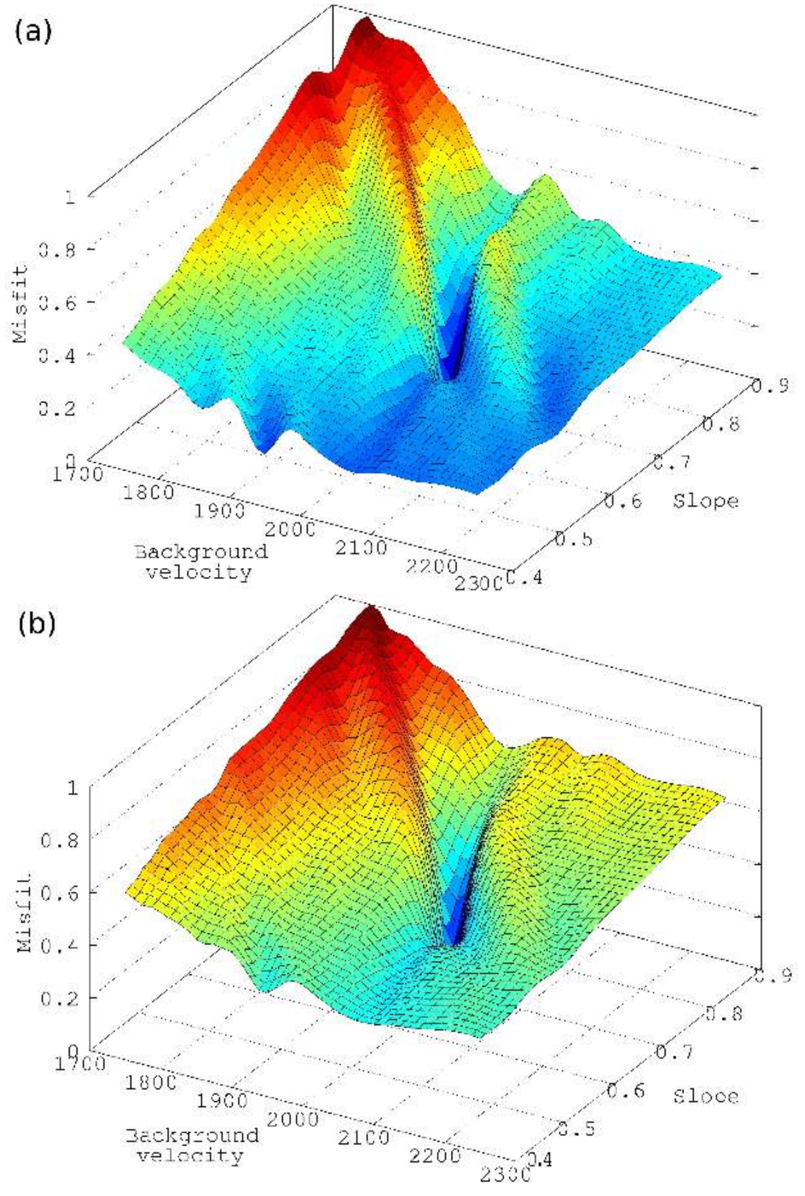

Figure 5. L^2 misfit function (a) and KR misfit function (b) depending on the background velocity $v_{P,0}$ and the slope α.

the valley where the global minimum is located. Finding this global minimum with local descent methods thus requires to start in the correct valley. Conversely, even if the KR misfit function still possesses local minima, the two narrow valleys of attraction on both sides of the central valley containing the global minima have been lifted up. The valley on the left is not anymore an obstacle to converge to the central valley. The valley on the right still plays the role of a barrier, however the height of the barrier has been significantly reduced compared to the L^2 case. This is an indication of a better behavior of the KR misfit function compared to the L^2 misfit function.

Contrary to the previous 1D experiment, the misfit function appears as more regular. A better insight on the shape of the global minimum is provided in Figure 6 where misfit function profiles along the velocity gradient α and the background velocity $v_{P,0}$ respectively are presented. These profiles well illustrate how the secondary valley of attraction are lifted up for the KR distance. They also show that even if the KR misfit function is smoother, in the vicinity of the global minimum, the misfit function exhibits similarities with the L_1 norm and appears as not differentiable at the global minimum. In terms of resolution power, it should be noted that the width of the global valley of attraction is almost the same for the L^2 and KR misfit function. This is different from what is observed when cross-correlation or deconvolution approaches are used to reduce the sensitivity to the initial model and cycle skipping: in this case the valley of attraction is strongly widen. This reflects a resolution loss of the imaging method as, near the solution, models possibly quite different yield approximately the same misfit. The KR misfit function is thus expected to keep the same resolution of the L^2 misfit function while relaxing the constraint on the choice of the initial model.

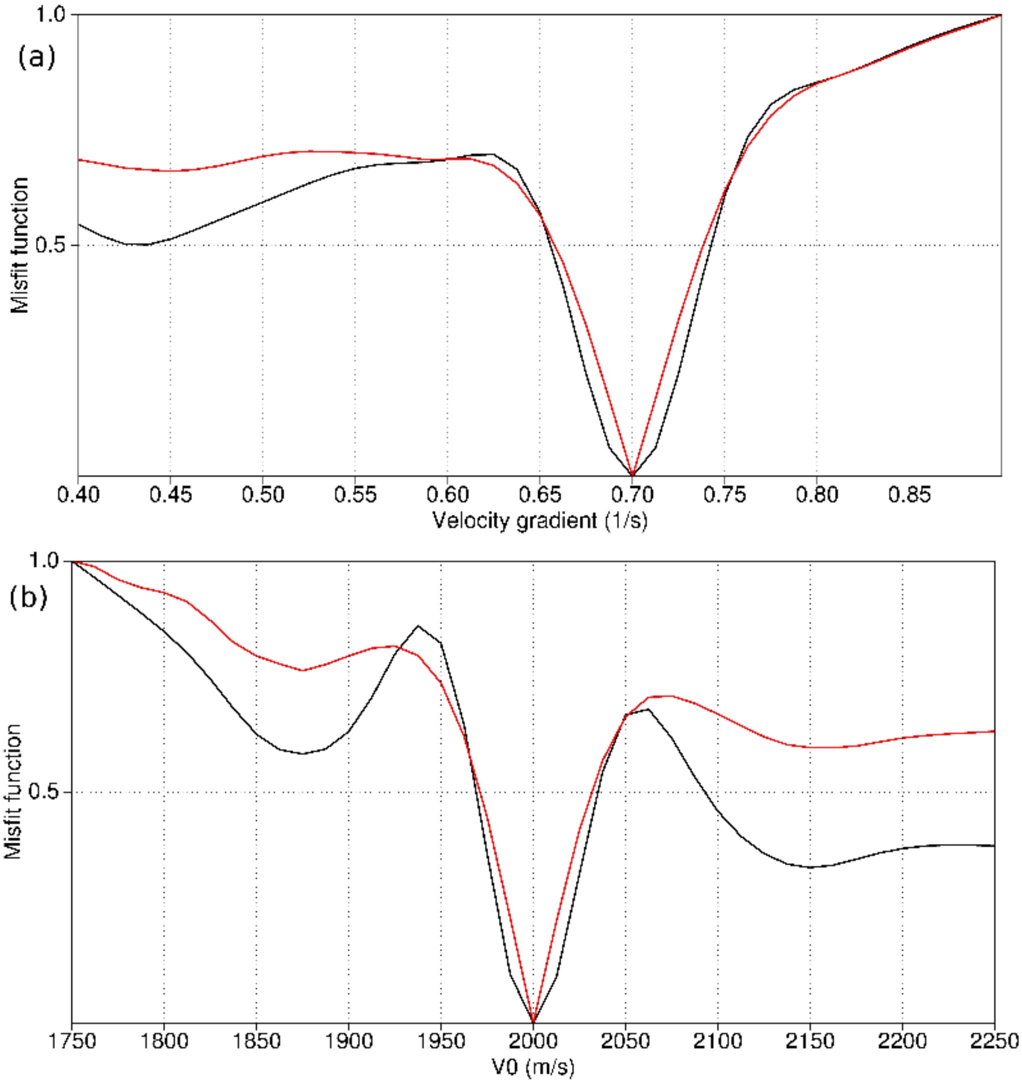

Figure 6. L^2 (black) and KR (red) misfit function depending on the velocity gradient α for a constant $v_{P,0} = 2000$ m.s^{-1} (a). L^2 (black) and KR (red) misfit function depending on $v_{P,0}$ for a constant velocity gradient $\alpha = 0.7$ (b).

In Figure 7, the L^2 and KR residuals are presented for the velocity model corresponding to $v_{P,0} = 2250$ m s^{-1}, $\alpha = 0.9$ s^{-1}. Cycle skipped diving and direct arrivals can be identified on the L^2 residuals for far offset receivers, between $t = 3$ s and $t = 4$ s. The corresponding KR residuals appear in this 2D case as a smooth version of the L^2 residuals, with a re-balancing of the energy between the mismatched direct and diving waves. Low frequency components seem also to be introduced in the KR residuals (black and white diffuse energy from left to the right of Figure 7 (c)). In the following experiments, this tendency to smooth and weight approximately equally the different mismatched seismic events is confirmed. The smoothing effect is not surprising: actually it is directly related to the repeated application of an approximate inverse of the Laplacian operator within the SDMM algorithm (solution of the linear system (54)).

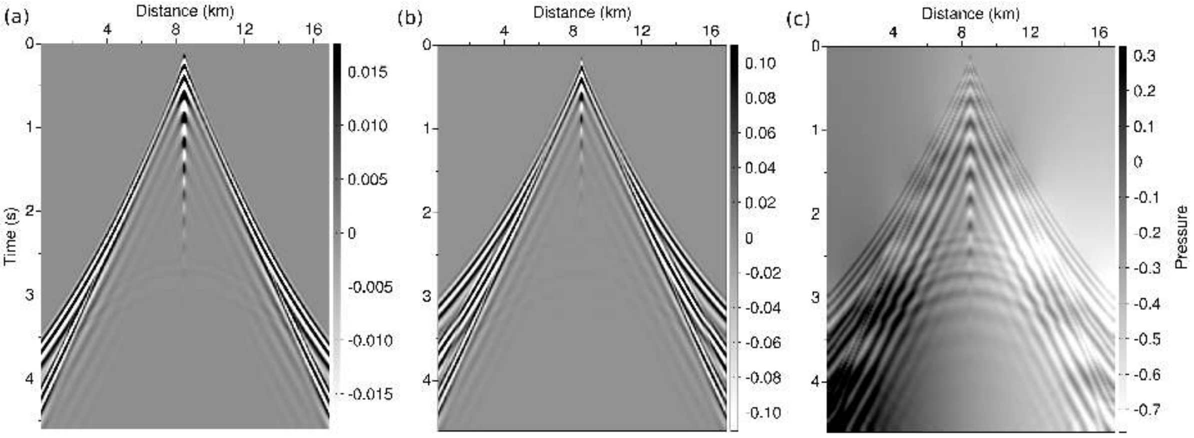

Figure 7. Observed data for the reference model (a). L^2 (b) and KR (c) residuals for the model $v_{P,0} = 2250$ m s^{-1}, $\alpha = 0.9$ s^{-1}.

5.4. Application to 2D and 3D realistic configurations

In the two following experiments, we consider FWI of synthetic offshore data in 2D (Marmousi model) and 3D (overthrust SEG/EAGE model) configurations. In both cases, the misfit functions $f_{L^2}(v_P)$ and $f_{KR}(v_P)$ are minimized using the quasi-Newton l-BFGS method [Nocedal, 1980] with the memory parameter l set to 20. The FWI implementation which is used interfaces the l-BFGS method provided in the SEISCOPE optimization toolbox [Métivier and Brossier, 2016]. A regularization strategy based on a gradient smoothing is used, designed as a Gaussian filter with a correlation length equal to a fraction of the local wavelength [Operto et al., 2006]. As a surface acquisition is considered in both experiments, a preconditioning of the gradient simply based on a linear (in 2D) or quadratic (in 3D) amplification in depth is also applied.

The KR distance is computed with a fixed number of SDMM iterations. This number is selected such that the criterion which is maximized during the solution of the problem (33) only marginally evolves after this number of iterations. For the 2D Marmousi problem, this number is set to 50, while it is set to 100 for the 3D overthrust problem. The computational overhead associated with the use of the KR distance is related to the increase of the gradient computational time. For the 2D Marmousi case, this overhead is equal to 3.8 s for a gradient computational time of 20.6 s in the conventional L^2 distance FWI formulation. This represents almost an 19% increase of the computational time. For the 3D overthrust model, the overhead is equal to 180 s for a gradient computational time of 240 s in the conventional L^2 distance FWI formulation. This represents approximately a 75% increase of the computational time.

5.4.1. 2D Marmousi case study, sensitivity to noise

The P-wave velocity of the Marmousi 2 benchmark model is presented in Figure 8a. A fixed-spread surface acquisition with 128 sources each 125 m and 168 receivers each 100 m, at 50 m depth is considered. The synthetic data is generated using a Ricker source function centered on 5 Hz. The frequency content of the source is low-pass filtered below 3 Hz to mimic realistic seismic data for which this frequency band is contaminated by noise and therefore unavailable for inversion. Two initial models are considered: the first contains the main features of the exact model, only with smoother interfaces. The second is a strongly smoothed version of the exact model with very weak lateral variation and underestimated growth of the velocity in depth. In the following experiments, the misfit functions $f_{L^2}(v_P)$ and $f_{KR}(v_P)$ are minimized using the quasi-Newton l-BFGS method [Nocedal, 1980].

The results obtained using the L^2 and KR distances are presented in Figure 8(d-g). The convergence to a correct estimation of the P-wave velocity model is reached using both the L^2 and KR distances starting from the first initial model. The models presented in Figure 8d and 8e have been obtained after 100 l-BFGS iterations. A slightly better estimation is obtained using the KR norm in the low velocity zone at $x = 11$ km, $z = 2.5$ km, where a high velocity artifact can be seen for the L^2 estimation. Starting from the second initial model, only the results obtained using the KR distance are meaningful (Fig.8g). The poor initial approximation of the P-wave velocity is responsible for cycle skipping and the L^2 estimation converges towards a local minimum (Fig.8f). The estimation obtained with the KR norm is significantly closer from the true model, despite low velocity artifacts in the shallow part at $x = 1.5$ km, $z = 1$ km and in depth at $x = 12$ km, $z = 3.4$ km. The results obtained with the KR distance are obtained after 439 l-BFGS iterations. The minimization of the L^2 misfit function with the same strategy fails after 83 iterations. From this experiment, the KR distance appears as an appropriate tool to mitigate cycle skipping in FWI.

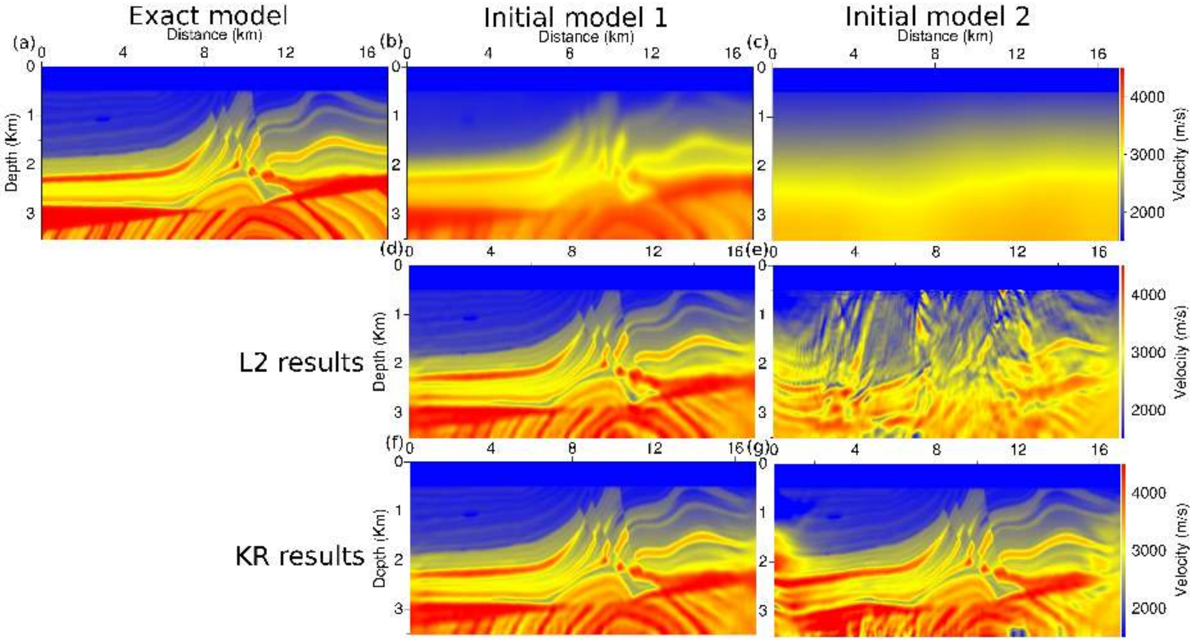

Figure 8. Marmousi model case study. Exact model (a), initial model 1 (b), initial model 2 (c), results obtained with the L^2 distance starting from model 1 (d), from model 2 (e), results obtained with the KR distance starting from model 1 (f), from model 2 (g).

In practice, low frequency seismic data required to initiate multi-scale FWI cycles are significantly contaminated by noise. Therefore, it is important to design stable strategies with respect to the interpretation of noisy data. For this reason, an additional inversion is performed using the KR distance, starting from the second initial model, with noisy data. A white Gaussian noise is added to the synthetic data. This white noise is band-pass filtered such that it belongs to the frequency-band of the data *i.e* between $3 - 12.5$ Hz. The signal over noise ratio (SNR) is taken equal to 5 (the power of the original data is 5 times higher than the power of the noise in the frequency), which is representative of the SNR observed on real data for this frequency band. The result of this experiment is presented in Figure 9. The P-wave velocity estimation is only slightly degraded by the presence of noise. Not surprisingly, the less illuminated zone on the borders of the model are more affected. This is understandable as these are the zones where the redundancy of the information is the weakest. A high frequency oscillation is also introduced in the estimation, which could be removed in a second stage through smoothing/denoising strategies. Compared to the first case where no noise is introduced, the convergence is observed here after 502 l-BFGS iterations, corresponding to a limited increase of 63 iterations. The method thus seems stable with respect to the introduction of noise.

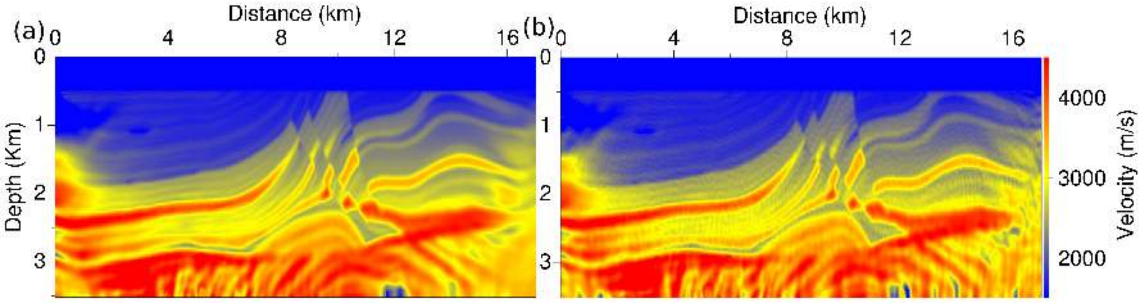

Figure 9. Results obtained on the Marmousi model using the KR distance with clean data (a), noisy data (b).

A better insight on how the method reacts to the addition of noise can be obtained by observing the L^2 and KR residuals in the second initial model, with and without noise (Fig. 10). As already observed, the KR residuals appear smoother than the L^2 residuals, with and without noise. This smoothing effect may mitigate the impact of noise on the KR results. More interestingly, the comparison between KR and L^2 residuals without noise shows that the KR distance enhance the weighting of weaker amplitude seismic events, such as those observed after 3 s for the receivers in the vicinity of the source (located at $x = 8$ km). This different weighting seems to be preserved when noise is introduced for the KR distance. Conversely, these weakly energetic events are in the noise limit for the L^2 residuals. The ability to keep these different weighting may explain the weak impact of the noise on the P-wave velocity estimation, when the KR norm is employed.

5.4.2. 3D application on the SEG/EAGE overthrust model

In this experiment, the shallow left part of the 3D SEG/EAGE overthrust model is considered (Fig. 11). A 250 m deep water layer is added on top. This model covers a surface of 10×10 km^2 and is 2.5 km deep. A fixed spread surface acquisition is used, with $8 \times 8 = 64$ sources (respectively $97 \times 97 = 9409$ receivers) regularly located each 1.2 km (respectively 100 m) in both x and y directions, and at 50 m depth. The synthetic dataset is generated using a Ricker source band-pass filtered between 3 Hz and 7.5 Hz (Fig. 12). The spatial discretization leads to a representation of the P-wave velocity model with $201 \times 201 \times 51$ discrete points with a discretization step h equal to 50 m. The time step is chosen equal to 0.004 s to respect the CFL condition. The recording time for one seismogram is fixed to 4 s (1000 discrete time steps). Each seismogram thus corresponds to a data cube of $97 \times 97 \times 1000 \simeq 10^7$ discrete points.

The purpose of this experiment is to focus on cycle skipping problems in a 3D context and compare the results obtained with the L^2 distance and the KR distance. Cycle skipping is mostly observed on diving waves, which sample the shallowest part of the model. For this reason, the initial model is chosen to poorly represent the exact model, especially in its shallow part. Slices of the exact and initial models at constant $y = 5$ km and constant depth $z = 1.5$ km, $z = 2$ km are presented in Figure 13. The initial model is an almost constant velocity model around 3000 m.s^{-1}, while the velocity of the exact model reaches 3500 m.s^{-1} already at $z = 1$ km depth. For this reason, the kinematic of the diving waves is not correctly predicted by the initial model. This can be observed in Figure 14, where the data associated with the source located at $x = 4.8$ km, $y = 4.8$ km, computed in the exact and the initial model are presented. For each model, three data panels are presented, corresponding to a slice in the data volume at constant $x = 5$ km, constant $y = 5$ km, and constant $t = 3$ s. The data is dominated by the direct arrival propagating in the water layer and the strong reflection coming from the interface between the water layer and the see bottom. The relative complexity of the signal is related to the source signature: the Butterworth filters applied to the Ricker wavelet yield a complicated wavelet with a large time support (Fig. 12). As the source and the water layer are considered to be known, the initial model correctly

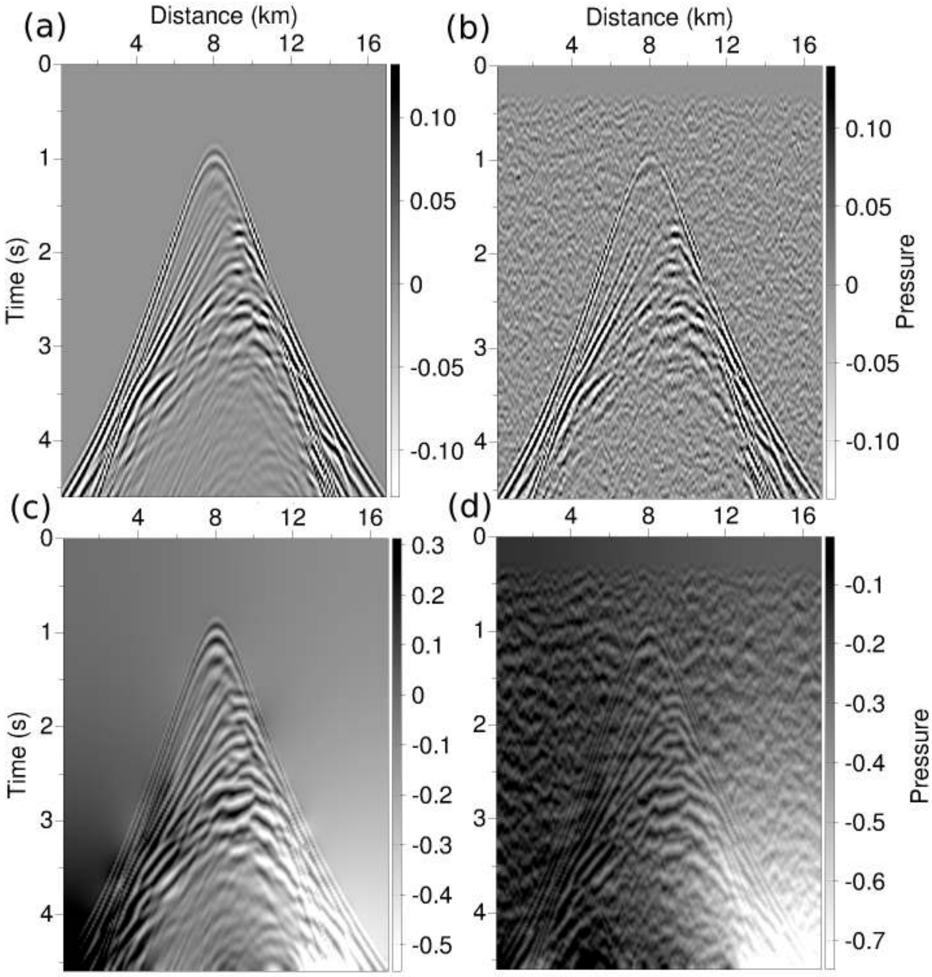

Figure 10. L^2 residuals in the second initial model with clean data (a), noisy data (b). KR residuals in the second initial model with clean data (c), noisy data (d).

reproduces the direct arrivals. However, a time shift of at least 0.3 s can be observed for the diving waves recorded by the farthest receivers. Conventional FWI using the L^2 distance is thus likely to produce inaccurate results in this configuration.

In Figure 15, the L^2 and KR residuals in the initial model are presented. The three panels correspond to slices in the residual volume at constant $x = 5$ km, constant $y = 5$ km, and constant $t = 3$ s., similarly as for the data in Figure 14. As it has been noticed in the previous 2D experiments, the energy of the seismic events in the KR residuals is balanced so that the amplitude of each event is comparable, while the L^2 residuals are dominated by short-to-intermediate offset missing events. Interestingly, this balance can be observed in the three panels, which testifies that the solution of the optimal transport problem is performed in the 3D volume without privileging one dimension over the two others.

Two distinct workflows for comparing the L^2 and the KR distance are followed. The first one simply consists in performing the inversion giving the freedom to the l-BFGS optimizer to perform as many iterations as possible. As for the Marmousi experiment, a Gaussian smoothing of the gradient is used, the correlation length being a fraction of the local wavelength [Operto et al., 2006]. Following this workflow, the inversion using the L^2 distance terminates after 61 iterations, while 229 iterations are performed with the KR distance. Both terminations are related to a linesearch failure: the optimizer is not able to minimize further the misfit function. The second workflow is based on a restarting process. At each stage, 100 l-BFGS iterations are allowed. The initial model for each stage corresponds to the final model of the previous stage. After the first termination on

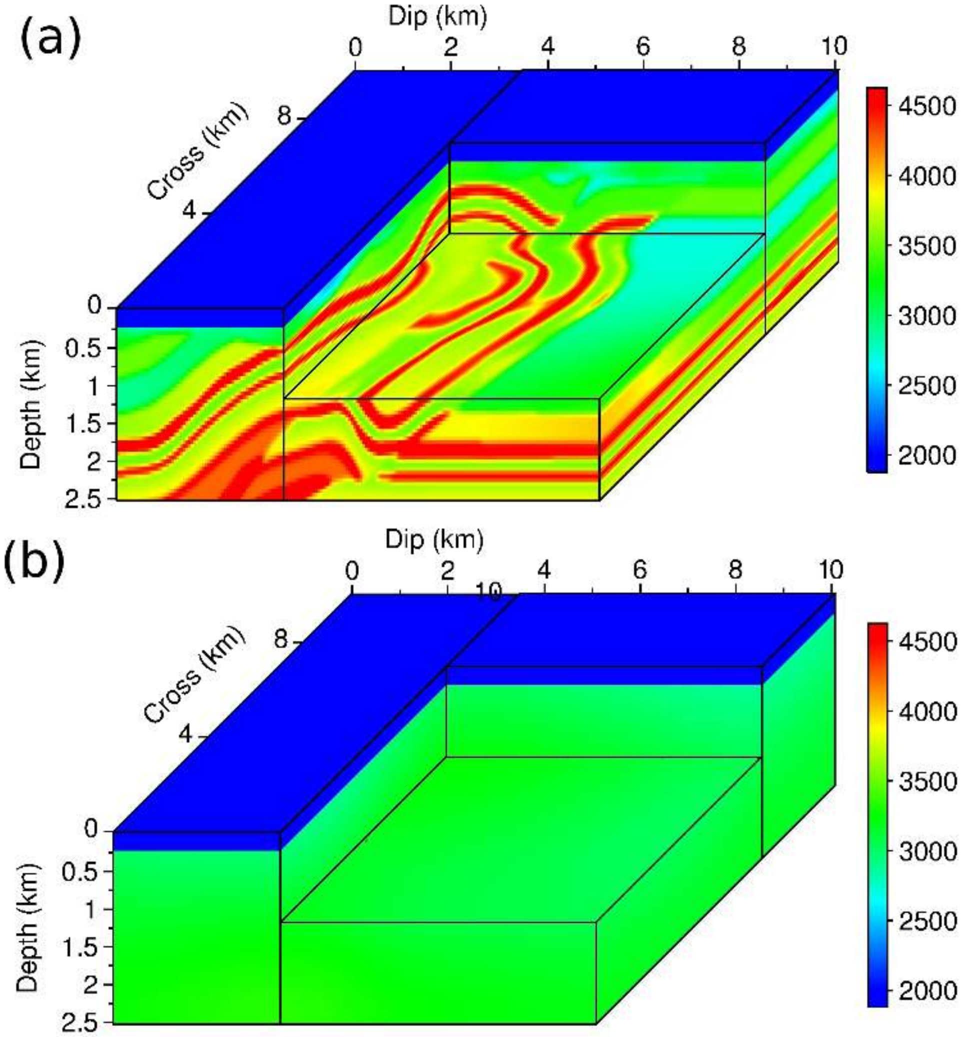

Figure 11. Exact (a) and initial (b) models used for the 3D SEG/EAGE overthrust case study.

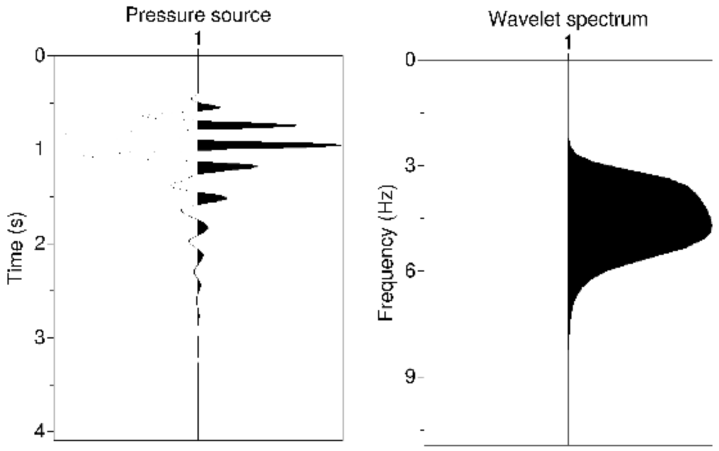

Figure 12. Source wavelet profile and spectrum for the 3D overthrust experiment.

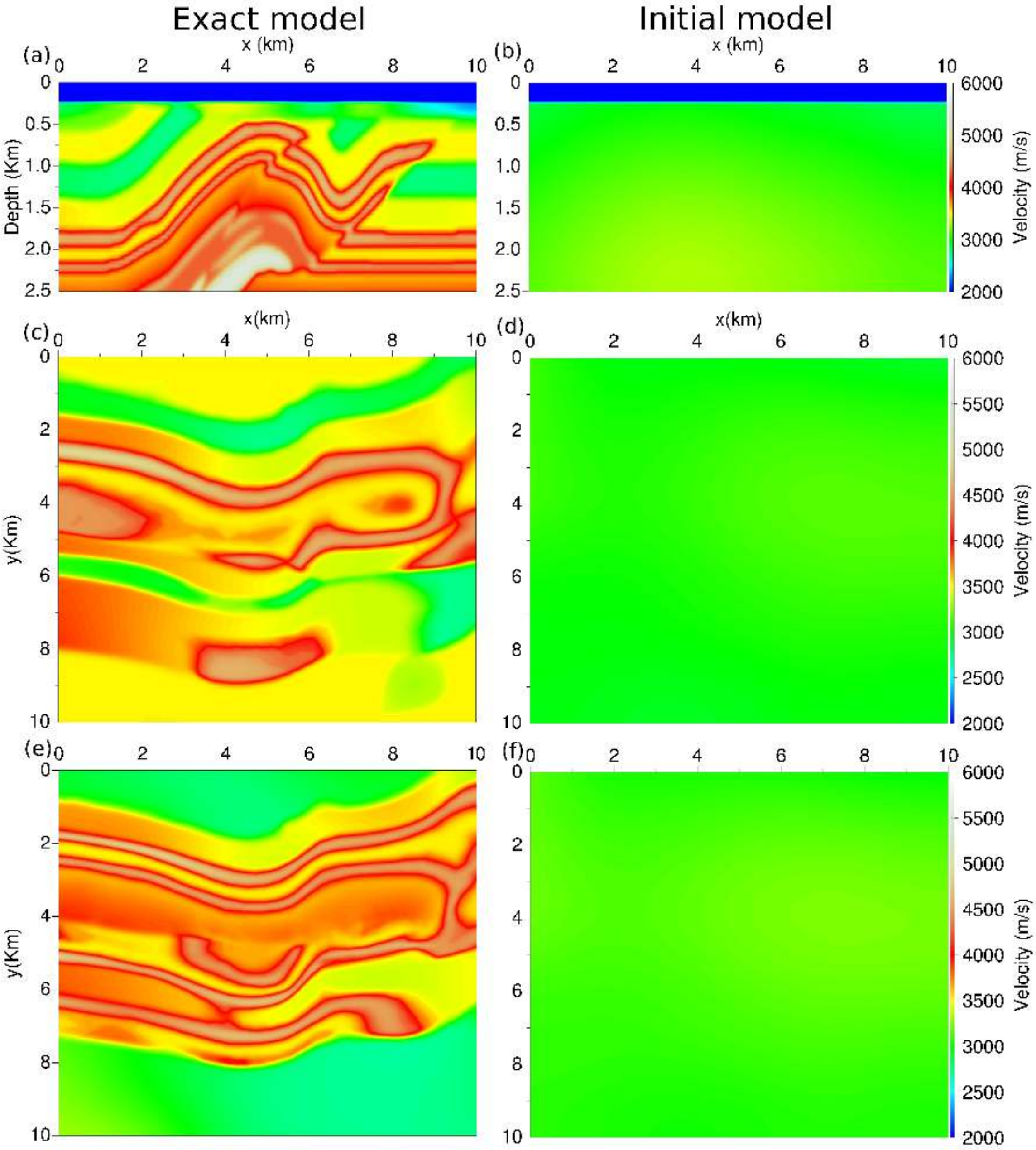

Figure 13. Exact and initial model overthrust model cross-sections. Cross-section at constant $y = 5$ km for the exact model (a), initial model (b). Cross-section at constant $z = 1.5$ km for the exact model (c), initial model (d). Cross-section at constant $z = 2$ km for the exact model (e), initial model (f).

a linesearch failure (instead of meeting the maximum 100 iterations criterion), the iterations are restarted from the previous model, however the regularization is cancelled. The process ends when the second linesearch failure is detected. Following this second workflow, the inversion using the L^2 distance terminates the first stage after 61 iteration, and 517 additional iterations are performed (with a restart of l-BFGS each 100 iterations), for a total of 578 iterations. The inversion using the KR distance terminates the first stage after 361 iterations, and 239 additional iterations with no regularization are performed, for a total of 600 iterations. The reason for this second workflow is that we observed that the l-BFGS optimizer can be sensitive to the gradient smoothing. This rather pragmatical modification of the descent direction is not accounted for in the misfit function. Therefore, errors can be introduced in the estimation of the descent direction using the l-BFGS

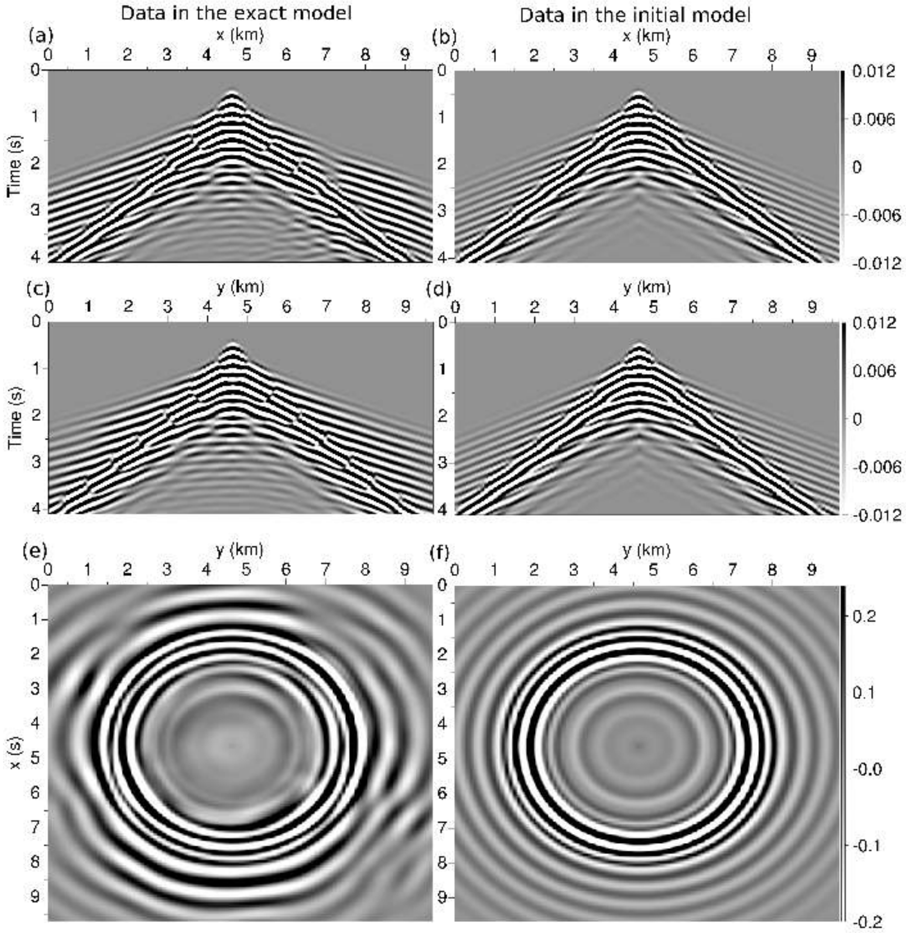

Figure 14. Seismograms in the exact (left column) and initial (right column) models. Seismogram cross-section at constant $y = 5$ km in the exact (a) and initial (b) models. Seismogram cross-section at constant $x = 5$ km in the exact (a) and initial (b) models. Seismogram cross-section at constant $t = 3$ s in the exact (e) and initial (f) models.

algorithm. This workflow also mimics hierarchical strategies which are often used for real data applications: in this sense it is a more realistic comparison than performing a single optimization as in workflow 1.

The results obtained following workflow 1 are presented in Figure 16. Obvious signs of cycle skipping are visible in the estimation obtained with the L^2 distance following workflow 1. In the constant y sections (Fig. 16a), low velocity artifacts can be observed at 1 km depth, in zones where the velocity update should be positive. Conversely, the KR distance provides a more reliable result in the shallow part of the model, until $z > 1.5$ km (Fig. 16b). This difference between the L^2 and KR distance is emphasized by the constant z cross-sections presented in Figure 16c-f. At $z = 1.5$ km, the cross-section of the KR estimation presents the main structures of the exact model (Fig. 16d). Conversely, the L^2 estimation does not exhibit these structures and present low velocity artifacts caused by cycle skipping (Fig. 16c). At $z = 2$ km, the KR estimation still provides some relevant information on the exact model, for instance in the zone 6 km $< y <$ 8 km, 2 km $< x <$ 8 km (Fig. 16f). Conversely, the L^2 estimation at this depth is completely cycle skipped exhibiting low velocity structures at the location where high velocity updates would be expected (Fig. 16g).

An additional illustration of the cycle skipping in the shallow part of the model using the L^2 distance is provided in Figure 17. The slices at constant $y = 5$ km of the estimated models are compared with the exact one and the zone below $z = 1.25$ km is shaded. As can be seen, the

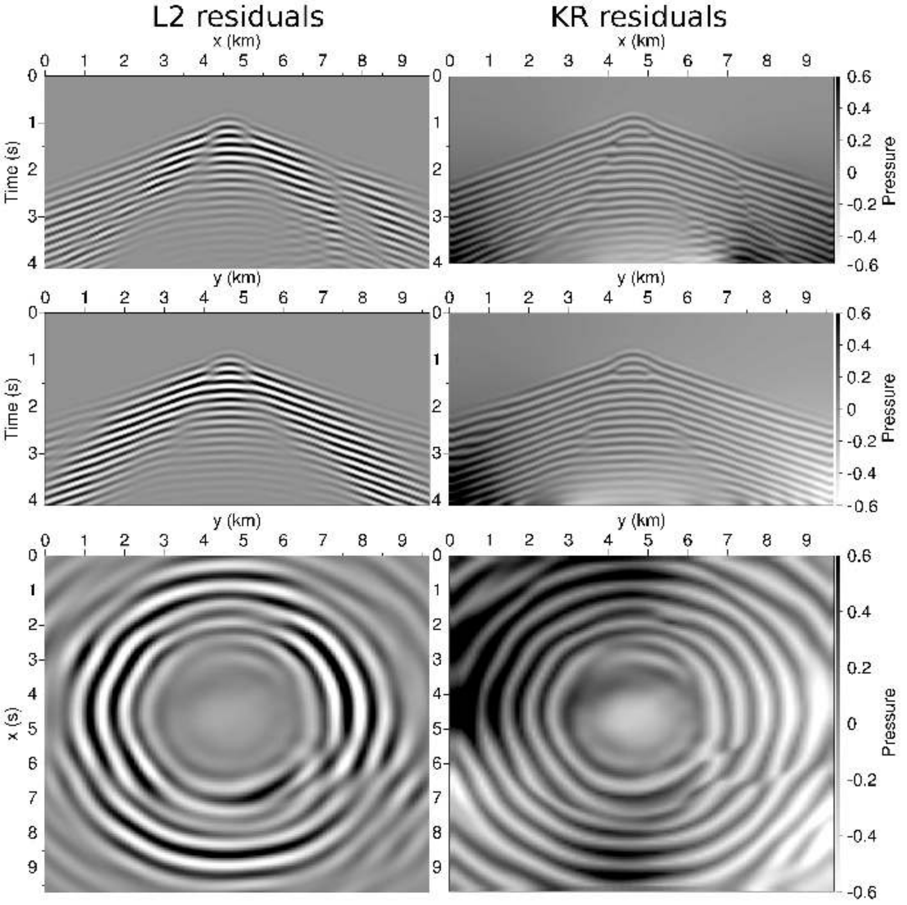

Figure 15. Residuals in the initial model for the L^2 distance (first row), for the KR distance (second row). Cross-section for constant $y = 5$ km (a,b). Cross-section for constant $x = 5$ km (c,d). Cross-section for constant $t = 3$ s (e,f).

curved reflector at 1 km depth is replaced with a low velocity anomaly in the L^2 estimation, while its structure is arising in the KR estimation.

The results obtained following workflow 2 are presented in Figure 18. The restarting procedure yields better estimation than the results obtained with a single optimization (workflow 1) both for the L^2 and KR misfit functions. For both distances, the shallow part until $z = 1.5$ km is approximately correctly recovered. However, stronger differences can be seen in depth. The slice at constant $z = 2$ km reveals that the L^2 estimation still suffers from cycle skipping, with a low velocity anomaly located near $x =$ 4km, $y = 4$ km. Conversely, the KR estimation at this depth seems in better accordance with the exact model. This low velocity artifact is replaced with the correct high-velocity structure.

An analysis of the residuals in the final estimations is provided in Figures 19 and 20. For workflow 1, as can be seen in constant x and constant y panels (Fig. 19b,e) , the residuals corresponding to the diving waves (arrival between $t = 2$ s and $t = 2.5$ for farthest offset receivers) are strong in the L^2 final model. Comparatively, these residuals are strongly attenuated in the KR final model (Fig. 19c,f). This observation is confirmed by the residual panel at constant $t = 3$ s, where the fringes associated to mismatched event are considerably reduced for receivers between $x =$ 2km, $y = 2$ km and $x = 8$ km, $y = 8$ km (Fig. 19h,i). For workflow 2, the difference between the residuals obtained with the L^2 distance and the KR distance is less obvious (Fig. 20). For both estimations, these residuals are considerably reduced compared to those obtained in the final estimations following workflow 1. The uninterpreted part of the data corresponds to strong reflections coming from the deepest part of the model, which is consistent with the inaccurate reconstruction observed in depth. However, one can note again lower amplitude residuals associated

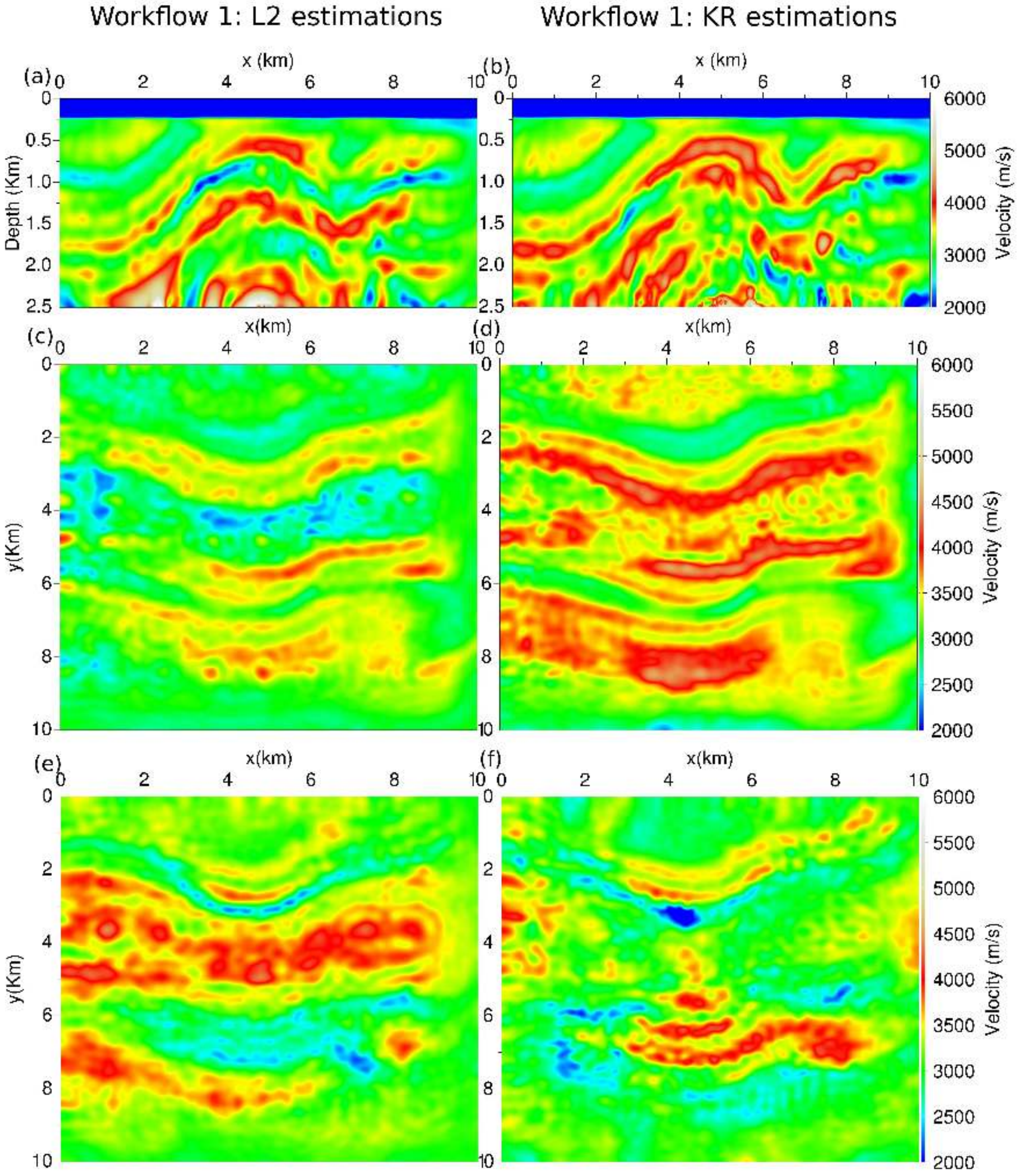

Figure 16. Cross-sections of the L^2 and KR estimations following workflow 1. Cross-section at constant $y = 5$ km for the L^2 estimation (a), KR estimation (b). Cross-section at constant $z = 1.5$ km for the L^2 estimation (c), KR estimation (d). Cross-section at constant $z = 2$ km for the L^2 estimation(e), KR estimation (f).

with shallow diving waves using the KR distance, showing that in this case the kinematic of these waves is better reconstructed despite the cycle skipping observed in the initial model.

As a final remark on this 3D experiment, one shall have in mind that the initial model which has been chosen is particularly inaccurate: in practice better estimations can be obtained without too much effort, accounting for instance for the velocity increase in depth. In this sense, this 3D experiment may not represent a realistic application of FWI, and one could question the relative inaccuracy of the results obtained using both L^2 and KR distance, especially in depth. Better results with both strategies could of course be obtained starting from a more accurate initial

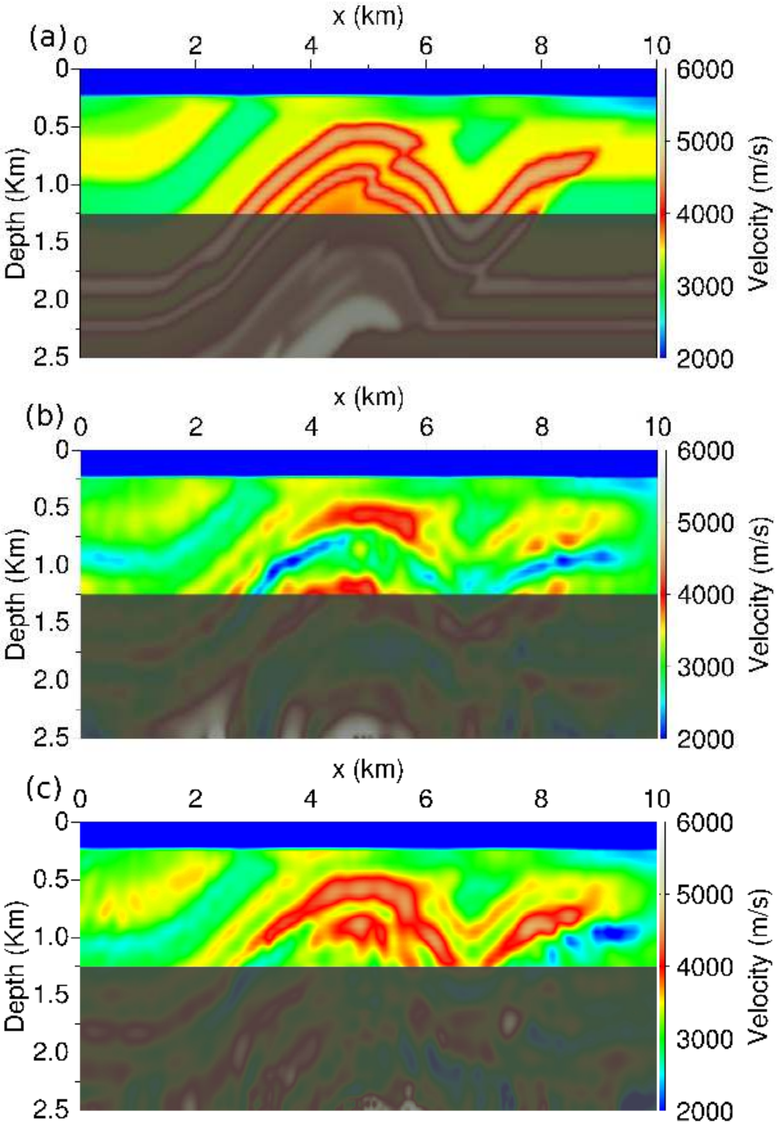

Figure 17. Focus on the shallow reconstruction following workflow 1. Constant $y = 5$ km cross-section for the exact model (a), L^2 estimation (b), KR estimation (c).

model. However, the purpose of this experiment is rather to push standard FWI strategy based on the L^2 distance to its limit, and observe what could be brought by the change to the KR distance. In this respect, the KR distance based FWI appears as a more robust tool for mitigating cycle skipping issues, feasible for realistic 3D applications.

6. Conclusion and perspectives

Large scale, smooth perturbations of the velocity are mainly responsible for delaying or accelerating the waves propagating within the subsurface, resulting in shifted events in the predicted and observed seismograms. Contrary to the L^2 distance, optimal transport distances have the capability to detect the shifts of recognizable patterns between images. For this reason, they can provide an interesting alternative for the reconstruction of velocity model.

In this study, the optimal transport distance which is proposed is based on the Kantorovich-Rubinstein norm. The computation of this norm is recast as a non-smooth optimization problem,

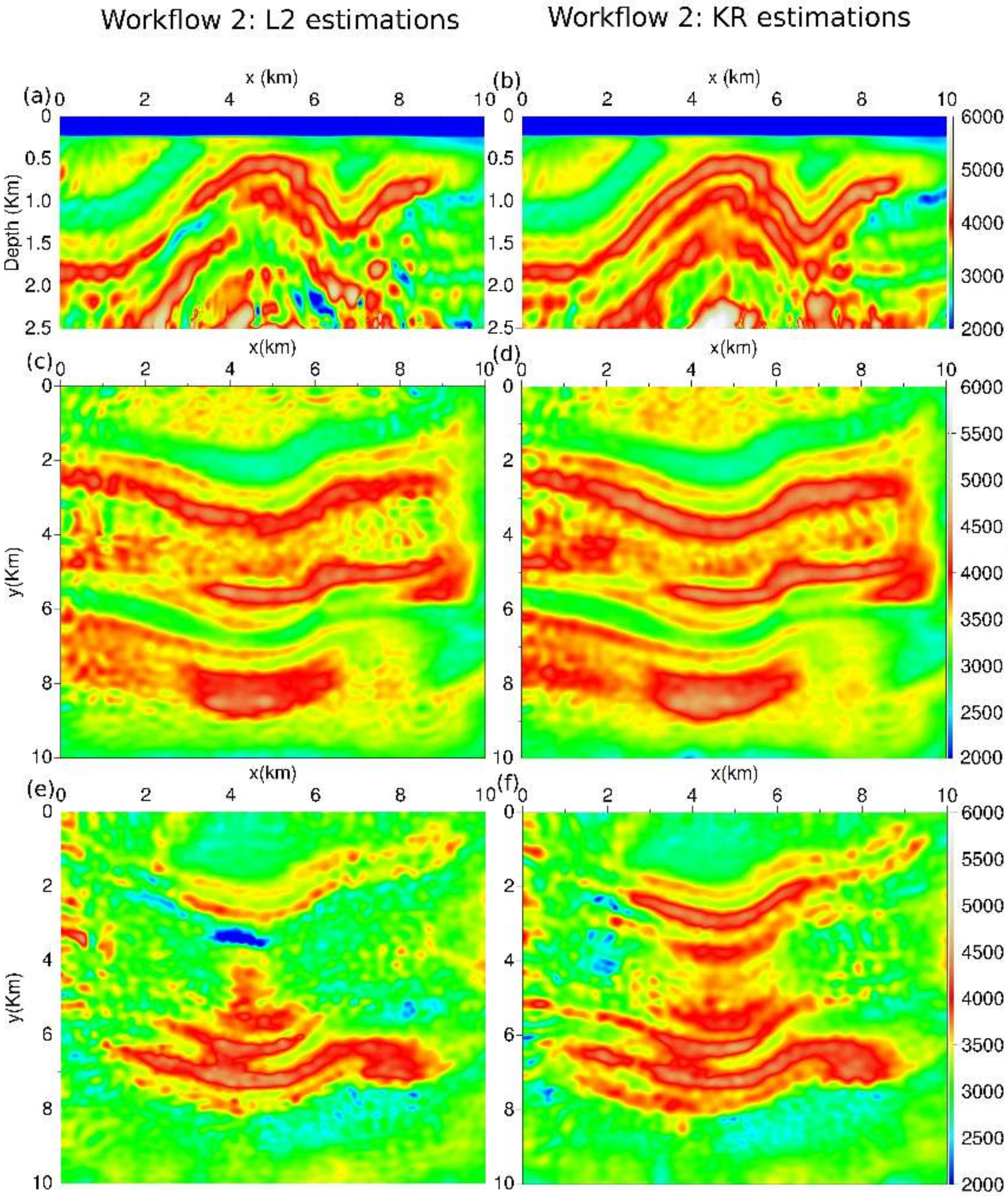

Figure 18. Cross-sections of the L^2 and KR estimations following workflow 2. Cross-section at constant $y = 5$ km for the L^2 estimation (a), KR estimation (b). Cross-section at constant $z = 1.5$ km for the L^2 estimation (c), KR estimation (d). Cross-section at constant $z = 2$ km for the L^2 estimation(e), KR estimation (f).

which is solved efficiently following a proximal splitting technique, namely the SDMM algorithm. Each iteration of this algorithm requires the solution of a Poisson's equation with homogeneous Neumann boundary conditions, which is efficiently performed using a multigrid algorithm.

The synthetic experiments which are performed confirm the interest of this approach for FWI application. The KR distance between time-shifted 1D signals presents a single minimum as a function of the time-shift. In a 2D time-domain framework with a surface acquisition, the misfit maps using the L^2 distance and KR distance are computed, for a velocity model represented as

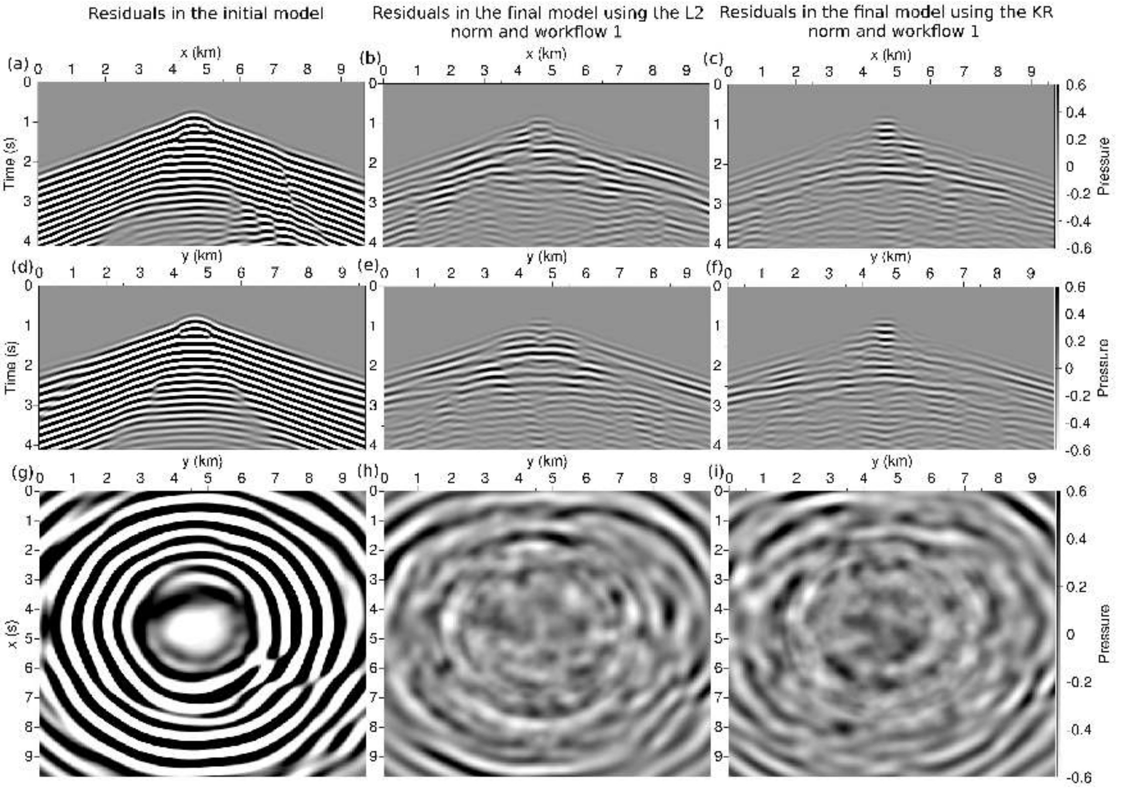

Figure 19. L^2 residuals in the initial model (left column), in the final model obtained using the L^2 (middle column) and KR (right column) distances following workflow 1. Cross-section for constant $y = 5$ km in the initial model (a), in the final model obtained using the L^2 (b) and KR (c) distances following workflow 1. Cross-section for constant $x = 5$ km in the initial model (d), in the final model obtained using the L^2 (e) and KR (f) distances following workflow 1. Cross-section for constant $t = 3$ s in the initial model (g), in the final model obtained using the L^2 (h) and KR (i) distances following workflow 1.

a linear function of the depth. The effect of the introduction of the KR distance shows that the secondary valleys are lifted up, reducing the probability to converge towards a local minimum when using a gradient-based method. A 2D time-domain FWI experiment on the Marmousi model shows a better robustness of the KR distance with respect to the accuracy of the initial model. Interestingly, the method appears robust to the addition of noise to the data. Finally, the 3D experiment on the overthrust SEG/EAGE model emphasizes the possibility of using the KR norm even in this large scale configuration. The method appears again as a reliable tool to mitigate cycle skipping issues: starting from a poor initial model, FWI based on the KR distance is able to reconstruct more accurately the part of the model sampled by diving waves, which are heavily cycle skipped in the initial model.

Future work will include application of this method to 2D and 3D real data in the FWI context, as well as a combination of this method with reflection FWI strategies [Chavent et al., 1994, Plessix et al., 1999, Xu et al., 2012, Brossier et al., 2015, Zhou et al., 2015]. These methods seek to improve the FWI reconstruction of the velocity in depth through the alternate reconstruction of the subsurface reflectivity and the smooth velocity background. This allows to account for transmission kernels between reflectors and the surface acquisition. The use of an optimal transport in this context may improve further the velocity reconstruction.

Beyond FWI, this work yields interesting perspectives for tomography methods in general. As soon as a model parameter influencing the kinematic of the propagation is reconstructed from the matching of shifted measurements, the use of optimal transport distance could be advantageously considered. The KR distance is flexible enough to offer the possibility to compare non-positive

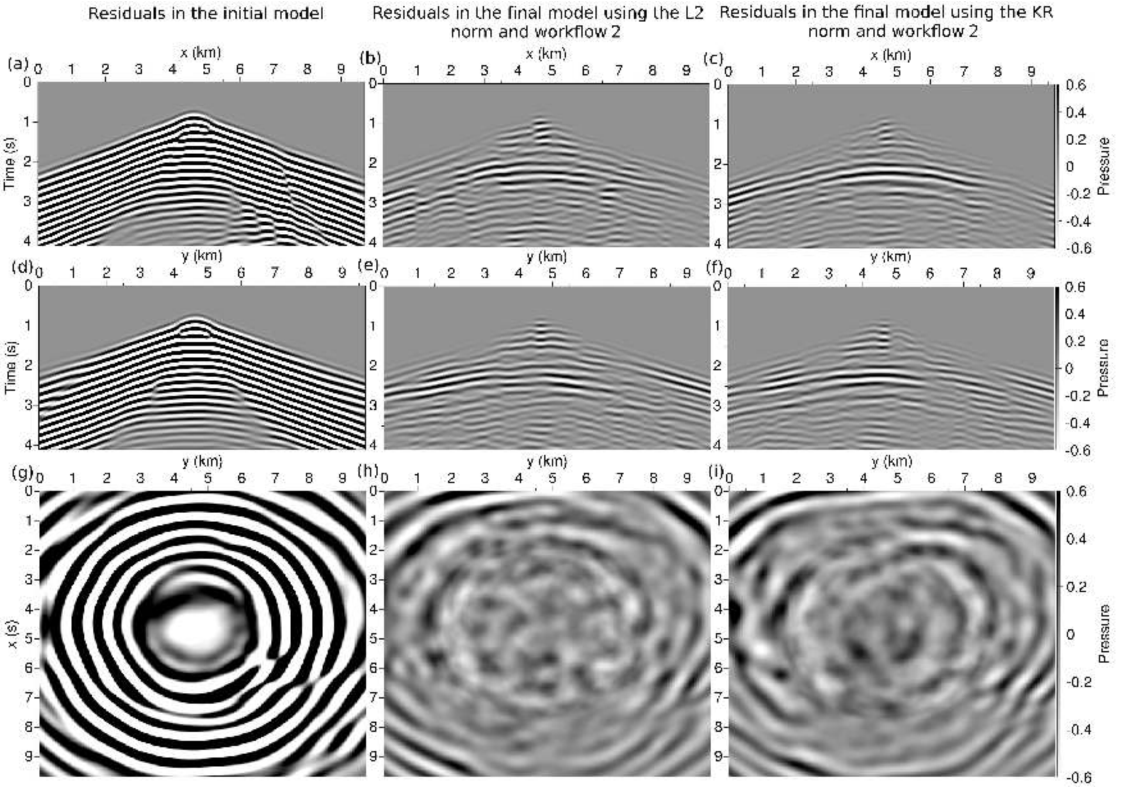

Figure 20. L^2 residuals in the initial model (left column), in the final model obtained using the L^2 (middle column) and KR (right column) distances following workflow 2. Cross-section for constant $y = 5$ km in the initial model (a), in the final model obtained using the L^2 (b) and KR (c) distances following workflow 2. Cross-section for constant $x = 5$ km in the initial model (d), in the final model obtained using the L^2 (e) and KR (f) distances following workflow 2. Cross-section for constant $t = 3$ s in the initial model (g), in the final model obtained using the L^2 (h) and KR (i) distances following workflow 1.

signals in the sense of optimal transport, without requiring the mass conservation between the signals. The numerical strategy presented in this study also gives the possibility to consider large scale application: in this study, the data volume for the 3D application reaches $O(10^7)$ discrete samples. As the method scales in linear complexity, larger-scale applications should be considered in the future.

References

[Adams, 1989] Adams, J. C. (1989). MUDPACK: Multigrid portable FORTRAN software for the efficient solution of linear elliptic partial differential equations. *Applied Mathematics and Computation*, 34(2):113–146.

[Ambrosio, 2003] Ambrosio, L. (2003). Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces*, volume 1812 of *Lecture Notes in Mathematics*, pages 1–52. Springer Berlin Heidelberg.

[Ambrosio et al., 2008] Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

[Bogachev, 2007] Bogachev, V. I. (2007). *Measure Theory*. Number vol. I,II in Measure Theory. Springer Berlin Heidelberg.

[Borisov and Singh, 2015] Borisov, D. and Singh, S. C. (2015). Three-dimensional elastic full waveform inversion in a marine environment using multicomponent ocean-bottom cables: a synthetic study. *Geophysical Journal International*, 201:1215–1234.

[Brandt, 1977] Brandt, A. (1977). Multi-level adaptive solutions to boundary-value problems. *Mathematics of Computation*, 31:333–390.

[Brossier et al., 2010] Brossier, R., Operto, S., and Virieux, J. (2010). Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3):R37–R46.

[Brossier et al., 2015] Brossier, R., Operto, S., and Virieux, J. (2015). Velocity model building from seismic reflection data by full waveform inversion. *Geophysical Prospecting*, 63:354–367.

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

[Bŏzdag et al., 2011] Bŏzdag, E., Trampert, J., and Tromp, J. (2011). Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870.

[Chavent, 1974] Chavent, G. (1974). Identification of parameter distributed systems. In Goodson, R. and Polis, M., editors, *Identification of function parameters in partial differential equations*, pages 31–48. American Society of Mechanical Engineers, New York.

[Chavent et al., 1994] Chavent, G., Clément, F., and Gòmez, S. (1994). Automatic determination of velocities via migration-based traveltime waveform inversion: A synthetic data example. *SEG Technical Program Expanded Abstracts 1994*, pages 1179–1182.

[Combettes and Pesquet, 2011] Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., and Wolkowicz, H., editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optimization and Its Applications*, pages 185–212. Springer New York.

[Diouane et al., 2016] Diouane, Y., Gratton, S., Vasseur, X., Vicente, L. N. ., and Calandra, H. (2016). A parallel evolution strategy for an Earth imaging problem in geophysics. *Optimization and Engineering*, 17(1):3–26.

[Engquist and Froese, 2014] Engquist, B. and Froese, B. D. (2014). Application of the wasserstein metric to seismic signals. *Communications in Mathematical Science*, 12(5):979–988.

[Ferradans et al., 2014] Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014). Regularized Discrete Optimal Transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882.

[Fichtner et al., 2008] Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H. P. (2008). Theoretical background for continental- and global-scale full-waveform inversion in the time-frequency domain. *Geophysical Journal International*, 175:665–685.

[Fichtner et al., 2010] Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H. P. (2010). Full waveform tomography for radially anisotropic structure: New insights into present and past states of the Australasian upper mantle. *Earth and Planetary Science Lettters*, 290(3-4):270–280.

[Hale, 2013] Hale, D. (2013). Dynamic warping of seismic images. *Geophysics*, 78(2):S105–S115.

[Hansen, 2006] Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In Lozano, J. A., Larrañaga, P., Inza, I., and Bengoetxea, E., editors, *Towards a New Evolutionary Computation*, volume 192 of *Studies in Fuzziness and Soft Computing*, pages 75–102. Springer Berlin Heidelberg.

[Jannane et al., 1989] Jannane, M., Beydoun, W., Crase, E., Cao, D., Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Roeth, G., Singh, S., Snieder, R., Tarantola, A., and Trezeguet, D. (1989). Wavelengths of Earth structures that can be resolved from seismic reflection data. *Geophysics*, 54(7):906–910.

[Kantorovich, 1942] Kantorovich, L. (1942). On the transfer of masses. *Dokl. Acad. Nauk. USSR*, 37:7–8.

[Lailly, 1983] Lailly, P. (1983). The seismic problem as a sequence of before-stack migrations. In Bednar, J., editor, *Conference on Inverse Scattering: Theory and Applications*. SIAM, Philadelphia.

[Lellmann et al., 2014] Lellmann, J., Lorenz, D., Schönlieb, C., and Valkonen, T. (2014). Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859.

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

[Luo and Sava, 2011] Luo, S. and Sava, P. (2011). A deconvolution-based objective function for wave-equation inversion. *SEG Technical Program Expanded Abstracts*, 30(1):2788–2792.

[Luo and Schuster, 1991] Luo, Y. and Schuster, G. T. (1991). Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653.

[Ma and Hale, 2013] Ma, Y. and Hale, D. (2013). Wave-equation reflection traveltime inversion with dynamic warping and full waveform inversion. *Geophysics*, 78(6):R223–R233.

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

[Métivier et al., 2016] Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016). Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Journal International*, 205:345–377.

[Mulder and Plessix, 2008] Mulder, W. and Plessix, R. E. (2008). Exploring some issues in acoustic full waveform inversion. *Geophysical Prospecting*, 56(6):827–841.

[Nocedal, 1980] Nocedal, J. (1980). Updating Quasi-Newton Matrices With Limited Storage. *Mathematics of Computation*, 35(151):773–782.

[Operto et al., 2015] Operto, S., Miniussi, A., Brossier, R., Combe, L., Métivier, L., Monteiller, V., Ribodetti, A., and Virieux, J. (2015). Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to Valhall in the visco-acoustic vertical transverse isotropic approximation. *Geophysical Journal International*, 202(2):1362–1391.

[Operto et al., 2006] 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.

[Peter et al., 2011] Peter, D., Komatitsch, D., Luo, Y., Martin, R., Le Goff, N., Casarotti, E., Le Loher, P., Magnoni, F., Liu, Q., Blitz, C., Nissen-Meyer, T., Basini, P., and Tromp, J. (2011). Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes. *Geophysical Journal International*, 186(2):721–739.

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

[Plessix et al., 1999] Plessix, R. E., Chavent, G., and Roeck, Y.-H. D. (1999). Waveform inversion of reflection seismic data for kinematic parameters by local inversion. *SIAM Journal of Scientific Computing*, 20:1033–1052.

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

[Pratelli, 2007] Pratelli, A. (2007). On the equality between monge's infimum and kantorovich's minimum in optimal mass transportation. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 43(1):1 – 13.

[Pratt, 1999] Pratt, R. G. (1999). Seismic waveform inversion in the frequency domain, part I : theory and verification in a physical scale model. *Geophysics*, 64:888–901.

[Santambrogio, 2015] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.

[Sirgue et al., 2010] Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., and Kommedal, J. H. (2010). Full waveform inversion: the next leap forward in imaging at Valhall. *First Break*, 28:65–70.

[Swarztrauber, 1974] Swarztrauber, P. N. (1974). A Direct Method for the Discrete Solution of Separable Elliptic Equations. *SIAM Journal on Numerical Analysis*, 11(6):1136–1150.

[Symes, 2008] Symes, W. W. (2008). Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, 56:765–790.

[Tape et al., 2010] Tape, C., Liu, Q., Maggi, A., and Tromp, J. (2010). Seismic tomography of the southern California crust based on spectral-element and adjoint methods. *Geophysical Journal International*, 180:433–462.

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

[Vigh et al., 2014] Vigh, D., Jiao, K., Watts, D., and Sun, D. (2014). Elastic full-waveform inversion application using multicomponent measurements of seismic data collection. *Geophysics*, 79(2):R63–R77.

[Villani, 2003] Villani, C. (2003). *Topics in optimal transportation*. Graduate Studies In Mathematics, Vol. 50, AMS.

[Villani, 2008] Villani, C. (2008). *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin.

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

[Warner and Guasch, 2014] Warner, M. and Guasch, L. (2014). Adaptative waveform inversion - FWI without cycle skipping - theory. In *76th EAGE Conference and Exhibition 2014*, page We E106 13.

[Warner et al., 2013] Warner, M., Ratcliffe, A., Nangoo, T., Morgan, J., Umpleby, A., Shah, N., Vinje, V., Stekl, I., Guasch, L., Win, C., Conroy, G., and Bertrand, A. (2013). Anisotropic 3D full-waveform inversion. *Geophysics*, 78(2):R59–R80.

[Xu et al., 2012] Xu, S., Wang, D., Chen, F., Lambaré, G., and Zhang, Y. (2012). Inversion on reflected seismic wave. *SEG Technical Program Expanded Abstracts 2012*, pages 1–7.

[Zhou et al., 2015] Zhou, W., Brossier, R., Operto, S., and Virieux, J. (2015). Full waveform inversion of diving & reflected waves for velocity model building with impedance inversion based on scale separation. *Geophysical Journal International*, 202(3):1535–1554.

[Zhu et al., 2012] Zhu, H., Bŏzdag, E., Peter, D., and Tromp, J. (2012). Structure of the european upper mantle revealed by adjoint tomography. *Nature Geoscience*, 5:493–498.