We A12 02

# FWI with Optimal Transport: a 3D Implementation and an Application on a Field Dataset

R. Poncet (CGG), J. Messud* (CGG), M. Bader (CGG), G. Lambaré (CGG), G. Viguier (CGG), C. Hidalgo (INEOS)

## Summary

We present the application to a 3D real dataset of full waveform inversion (FWI) with optimal transport (OT) using the Kantorovich-Rubinstein (KR) distance as proposed by Métivier et al. (2016). This approach involves an efficient numerical implementation for OT in time and space directions, allowing the lateral coherency of the traces to be taken into account; this has an important impact on the quality of the results. The approach also exhibits a slightly reduced sensitivity to local minima compared to least squares (LSQ) misfit. Moreover the iterative method used for the computation of the KR distance allows the production of a set of intermediary solutions that span progressively from LSQ to OT. We recall the main components of the approach and present its numerical implementation in 3D. We show the improvement of the results compared to conventional FWI on 2D synthetic and 3D real datasets for the same number of velocity update iterations.

## Introduction

Conventional full waveform inversion (FWI) is based on a least squares (LSQ) misfit function. It has proven to be effective for high-resolution velocity model building in areas investigated by diving and reflected waves. Due to cycle skipping, this cost function is however plagued by many local minima, and the local optimization process requires starting from a good initial model. Many alternate misfit functions have been proposed to mitigate this. Among those, Wasserstein distances, based on optimal transport (OT) theory, recently aroused attention in geophysics (Engquist and Froese, 2014; Métivier et al, 2016). This enthusiasm comes from their ability to deal with shifts between data (Yang et al, 2017), thus potentially reducing the local minima issue. The introduction of OT into FWI is however not straightforward: "vanilla" OT theory addresses the comparison of positive data with the same mass (integral), which is not the case with seismic data, and computational aspects are critical due to the size of seismic data.

Among the various formulations of OT FWI, Métivier et al (2016, 2016b) appears quite promising based on the results from synthetic 2D and 3D cases. Their approach is based on the use of a cost function involving a Wasserstein distance, and is directly applied on the seismic data. To remedy the issue of the conservation of the mass, the Wasserstein distance is slightly modified becoming the Kantorovich-Rubinstein (KR) distance (Lellmann et al, 2014). As nothing is done concerning data positivity, the method loses a part of its OT behaviour, i.e. has a reduced ability to sense shifts in the data. Interestingly, they use a multidimensional KR distance accounting for correlations between time samples, and traces within a shot, which has an important impact on the quality of the results. They propose a robust and efficient iterative numerical scheme to compute the KR distance, offering perspective for 3D industrial applications.

In this article, we present an application of OT FWI to 3D real data, following the approach proposed by Métivier et al. (2016). We compare the results to the LSQ approach, and show the improvements brought by OT FWI (structural coherency of the velocity and slightly more robust to cycle skipping). We also emphasize the connection between LSQ misfit and KR distance: the iterative method used for the computation of the KR distance produces a set of intermediary solutions that span progressively from LSQ to OT.

## Theory

For each shot, we denote by $d_{obs}(\boldsymbol{x})$ the observed data and $d[m](\boldsymbol{x})$ the data modelled using a subsurface model $m$. The data space $\boldsymbol{X}$ is parameterized by the time and the receiver positions, i.e. a vector $\boldsymbol{x} = (t_{\boldsymbol{x}}, \boldsymbol{r}_{\boldsymbol{x}})^{+}$. For a FWI misfit measure $\sum_{shots} J(d_{obs}, d[m])$, the data gradient $\delta J/\delta d[m]$ defines the adjoint-source, and thus the model gradient via the adjoint-state method (Plessix, 2006). $c_p(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_p$ denotes the L$_p$ distance between vectors in the $\boldsymbol{X}$ space. The LSQ misfit and data gradient are

$$J_{L_2}(d_{obs}, d[\boldsymbol{m}]\,) \;=\; \frac{1}{2} c_2^2(d_{obs}, d[m]) \qquad \text{and} \qquad \frac{\delta\, J_{L_2}}{\delta d[m]} = \Delta d[m] \qquad (1)$$

where $\Delta d[m] = d_{obs}(\boldsymbol{x}) - d[m]$ denotes the residual. The p-Wasserstein distance for two probability densities $d_1$ and $d_2$ in the data space is shown in Eq. (2), where s.t. denotes "subject to constraint".

$$J_{W_p}^{p}(d_1, d_2\,) = \inf_{\boldsymbol{T}} \int_{\boldsymbol{X}} c_p^p(\boldsymbol{x}, \boldsymbol{T}(\boldsymbol{x}))\, d_2(\boldsymbol{x}) d\boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{T} \in maps\ that\ rearrange\ d_2\ into\ d_1 \quad (2)$$

Eq. (2) is similar to the original Monge OT problem that seeks the minimum cost to transport mass from $d_2$ to $d_1$, from the $\boldsymbol{X}$ space cost $c_p^p$ point of view (Villani, 2008). As OT requires positive $d_1$ and $d_2$ with equal masses, it cannot be readily applied to seismic data. To overcome this limitation, Yang et al (2017) and Qiu et al (2017) proposed positive transformations of the seismic data followed by rescaling to the same mass. One of the drawbacks of this transformation is that noise or unpredicted data can largely influence the inversion (Métivier et al., 2016). They chose the *p=2* case, i.e. the squared 2-Wasserstein distance related to the LSQ cost $c_2^2$. As solving eq. (2) in the multi-dimensional data space is computationally demanding because it implies solving the Monge-Ampère equation (Qiu et al, 2017), most of their applications consider 1D data space, i.e. measure time direction differences. Then $\boldsymbol{x}$ represents the time direction, $x = t$, and eq. (2) is solved for each trace independently.

Métivier et al (2016) started from the 1-Wasserstein distance, i.e. *p=1* in eq. (2). Kantorovich and then Rubinstein showed that then eq. (2) can be reformulated as linear (dual) problems that can be computed using linear programming. But the use of data with different masses creates mathematical singularities. To overcome that, one can add a bounding constraint that leads to the so-called KR distance (Lellmann et al, 2014) where the seismic data can directly be used

$$J_{KR}(d_{obs}, d[m]) = \max_{\varphi} \int_{\boldsymbol{X}} \varphi(\boldsymbol{x})\Delta d[m](\boldsymbol{x})d\boldsymbol{x} \ \text{ s.t. } |\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| \le c_1(\boldsymbol{x}, \boldsymbol{y}) \ \& \ |\varphi(\boldsymbol{x})| \le K \quad (3)$$

$\varphi$ is the solution of the dual problem (3). The first constraint on $\varphi$ is called 1-Lipschitz for the metric $c_1$. It imposes that $\varphi$ is continuous and changes in $\varphi$ are sufficiently slow with respect to $c_1$. This produces low frequencies in $\varphi$, which is crucial for OT behaviour. The 2nd finite bound constraint on $\varphi$ allows us to overcome the mass conservation requirement and use seismic data with different masses without creating singularities. The solution of equation 3 denoted by $\varphi^{max}$ depends on the residual. In this case, the adjoint-source is defined by $\delta J_{KR}/\delta d[m] = \varphi^{max}$ (because we can show $\int_{\boldsymbol{X}} \frac{\delta \varphi^{max}(\boldsymbol{x})}{\delta d[m](\boldsymbol{y})} \Delta d[m](\boldsymbol{x})d\boldsymbol{x} \approx 0$), allowing us to use the adjoint-state method to compute the model gradient. The 1-Lipschitz constraint in eq. (3) can be reformulated as a local constraint which leads to a computationally tractable scheme to iteratively solve the discretized eq. (3) (Métivier et al, 2016b).

## Implementation and methodology improvement

We have implemented the Simultaneous Descent Method of Multipliers (SDMM) convex optimization method and the Laplace solver as proposed by Métivier et al (2016b) considering time and inline receiver dimensions in $\boldsymbol{x} = (t_{\boldsymbol{x}}, r_{\boldsymbol{x}}^{inline})^{+}$, i.e. resolving eq. (3) for each crossline (the crossline direction being sparser and more aliased than the inline one in the marine case).

*K* value: From $L_1$ to OT

Choosing a good *K* value is important for the success of the scheme. For large *K*, the KR distance (3) becomes equivalent to the 1-Wasserstein distance, i.e.pure OT, in the case of positive and equal mass data. Using seismic data requires sufficiently reducing *K* to avoid singularities. But if *K* is chosen too small, i.e. causing all the bounding constrainst in eq. (3) to saturate, it can be proven that the KR distance becomes then equivalent to the $L_1$ distance (up to a proportionality constant that is compensated by line search) (Lellmann et al, 2014). We were able to find the good balance for *K*.

Number of iterations: From LSQ to OT

*N* denotes the number of SDMM inner iterations to resolve eq. (3). In our implementation for *N=0,* KR FWI reduces to LSQ FWI, and that increasing *N* will “add” more and more OT. Our experience shows that for real data, full convergence can require quite large *N* values (up to 1000) which can be computationally demanding in 3D cases. We were able to define an *N* value that represents a good balance between sufficient OT behavior and full convergence.

The *K* and *N* parameters are interesting because they allow us to define a hybrid misfit that mixes OT with conventional misfits if desired, namely $L_1$ (from *K*) and LSQ (from *N*). We believe this flexibility and the controllable continuum between those 3 misfits is a strong point.

Parameterization of the metric: Crucial for success

Because $\boldsymbol{x}$ mixes different physical units, the $c_1$ metric that appears in the 1-Lipschitz constraint of eq. (3) must necessarily be a generalized $L_1$ distance. We parameterize it in a Mahalanobis-like fashion

$$c_1(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\sigma_r^{inline}} \{v|t_{\boldsymbol{x}} - t_{\boldsymbol{y}}| + |r_{\boldsymbol{x}}^{inline} - r_{\boldsymbol{y}}^{inline}|\}$$

where $\sigma_r^{inline}$ represents a variance in the inline (distance) direction and $\sigma_t = v/\sigma_r^{inline}$ a variance in the time direction. $v$ is a velocity. It can be demonstrated that for full convergence (i.e. sufficiently large *N*), any choice for $\sigma_r^{inline}$ is almost equivalent in eq. (3) when an optimal *K* is considered (they will almost only differ by a proportionality constant compensated by line search). But in practice, we can only run finite number of iterations. Hence, our objective is to find $\sigma_r^{inline}$ and $\sigma_t$ that can improve the rate of convergence. $v$ defines the average direction along which most correlations between traces occur, thus is closely related to the average move-out direction. We were able to define optimum $\sigma_r^{inline}$ and $v$ values.

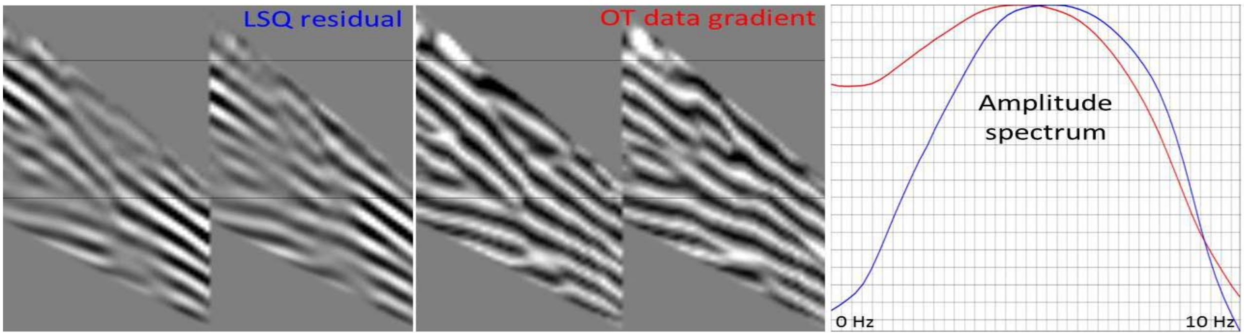

***Figure 1** 3D real dataset with a mute for a 6Hz inversion. Left: LSQ data gradient (or residual). Middle: KR data gradient for N=600. Right: Corresponding amplitude spectra comparison.*

Fig. 1 compares, for a real 3D shot-gather data (with a mute), a LSQ gradient (or residual) and a KR distance gradient for N=600. The latter has much more low frequencies and a skeleton-like texture in both time and inline directions due to the 1-Lipschitz constraint. The use of KR with multidimensional data space (time and inline directions) allowed us to recover nice continuity and amplitude balancing in the move-out direction.

## Application to synthetic and field dataset

We tested the scheme on the 2D Marmousi 2 dataset and on a real narrow-azimuth marine dataset. We used *N*≤600 SDMM inner iterations for each KR problem, and a preconditioned L-BFGS optimization scheme for the FWI optimization process, with a number of FWI iterations between 6 and 20.

For Marmousi, 20 FWI iterations were performed directly at up to 10Hz, keeping all the data (first break, reflections, multiples…) and starting with a smooth initial velocity model obtained by Gaussian filter smoothing of the true model. The top panels of Fig. 2 show those models. The synthetic data was created in the constant-density acoustic approximation with a Ricker source centred on 6 Hz. The bottom panels of Fig. 2 shows that the model estimated by KR FWI matches the true model better than LSQ FWI, especially in highlighted zones. KR better mitigates cycle-skipping and produces more continuity along structures.

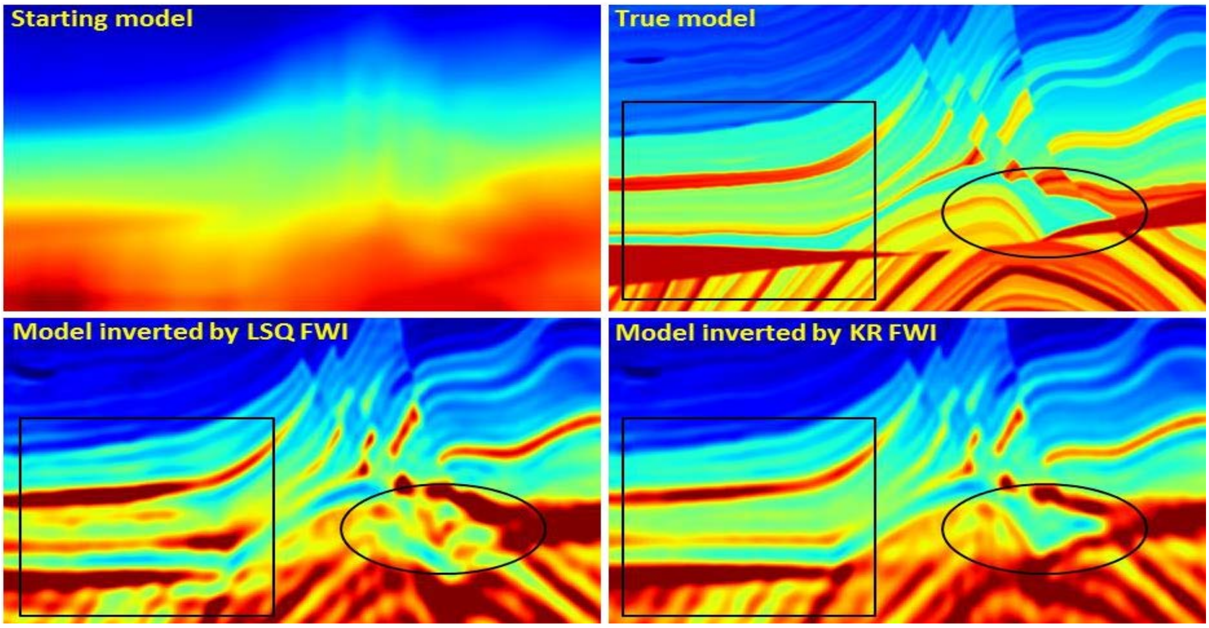

***Figure 2** Marmousi 2 model test. FWI velocity. 20 FWI iterations performed directly at up to 10Hz, keeping all the data (first break, reflections, multiples…) and using the same smooth starting model.*

For the 3D case, the field dataset was taken from the Siri field in the Danish North Sea. The survey consists of 6 towed streamers with a maximum offset of 3km. Time domain FWI was run at 6, 8 and 10 Hz consecutively with 6 iterations at each frequency block. Due to the relatively simple and flat geology and to the absence of long offsets (maximum penetration of diving waves < 800m) there was no sign of cycle skipping in the part resolvable with FWI even with a quite smoothed starting model. Nevertheless, KR FWI was able to deliver a better model, more coherent with geological structures compared to LSQ FWI (Fig. 3). It thus seems a final LSQ FWI pass is unnecessary. The accuracy of KR FWI is confirmed with migrated gather flatness. We expect a much larger uplift from KR FWI in a more complex geological setting, with longer offset acquisitions and lower frequency content data.

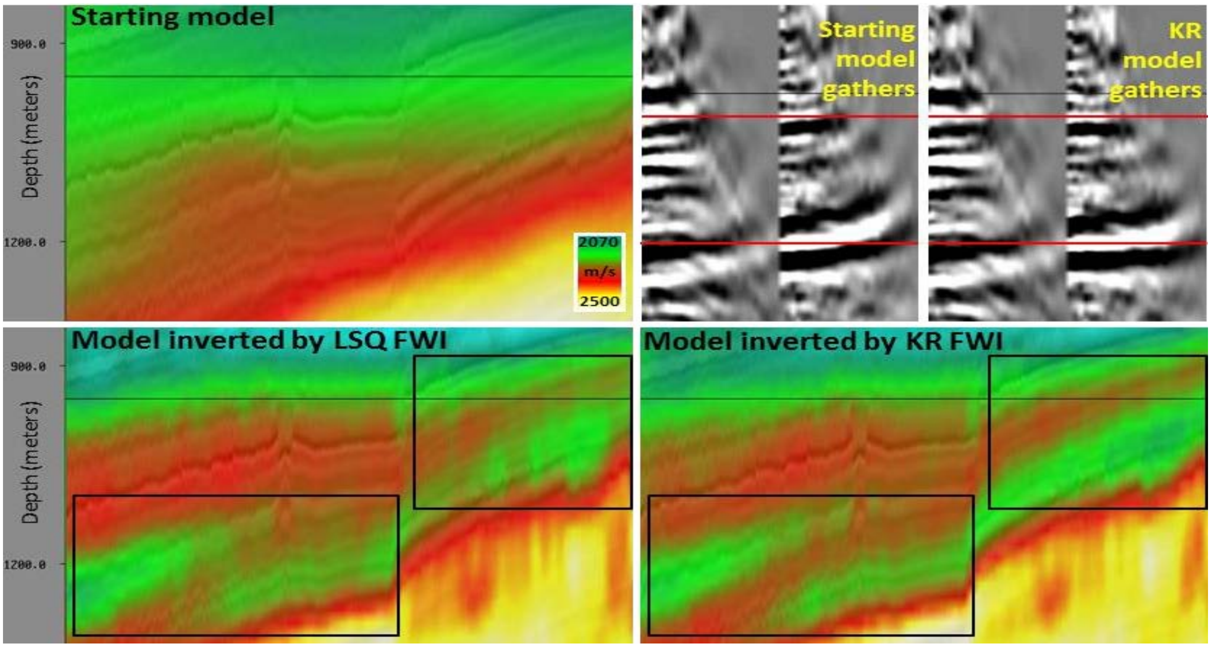

***Figure 3*** *3D real data test. Stack overlaid with FWI velocity, and migrated gathers (Data courtesy of WesternGeco Multiclient). LSQ and KR FWI are run with same configuration: 6 iterations at 6, 8 and 10Hz using same smooth starting model.*

## Conclusions

We have presented the application to a 3D real dataset of OT FWI using the KR distance as proposed by Métivier et al. (2016). It is based on an efficient numerical implementation allowing for time and space directions OT, with the benefits of an improved structural consistency and a reduced sensitivity to cycle skipping compared to LSQ FWI. In practice it offers an interesting extension to LSQ FWI, producing a set of intermediary solutions that span progressively from LSQ to OT.

## Acknowledgements

We thank WesternGeco, INEOS and CGG for their permission to publish this work. We gratefully thank Olivier Leblanc and Hervé Prigent for their involvement.

## References

Engquist, B. and Froese, B. D. [2014] Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences* **12**, 979 – 988.

Lellmann, J., Lorenz, D. A., Schönlieb, C. and Valkonen T. [2014] Imaging with Kantorovich--Rubinstein Discrepancy. *SIAM Journal on Imaging Sciences* **7**, 2833-2859.

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.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. and Virieux, J. [2016b] An optimal transport approach for seismic tomography: application to 3D full waveform inversion. *Inverse Problems* **32**, 115008.

Plessix, R. E. [2006] A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Communications in Mathematical Sciences* **12**, 979 – 988.

Qiu, L., Ramos-Martínez, J., Valenciano, A., Yang, Y. and Engquist, B. [2017] Full waveform inversion with an exponentially-encoded optimal transport norm. *87th SEG Expanded Abstracts*, p 1286.

Villani, C. [2008] Optimal transport: Old and new. *Springer, Berlin*.

Yang, Y., Engquist, B., Sun, J. and Froese, B. D. [2017] Application of Optimal Transport and the Quadratic Wasserstein Metric to Full-Waveform Inversion. *Accepted for publication in Geophysics*.