Geophysical Journal International

Geophys. J. Int. (2022) **229,** 35–55
Advance Access publication 2021 October 27
GJI General Geophysical Methods

https://doi.org/10.1093/gji/ggab441

Full waveform inversion based on the non-parametric estimate of the probability distribution of the residuals

P.T.C. Carvalho [1], S.L.E.F. da Silva [2,*] E.F. Duarte,[2] R. Brossier [3] G. Corso[1,4] and J.M. de Araújo[1,2]

[1] *Postgraduate Program in Science and Petroleum Engineering, Federal University of Rio Grande do Norte, Natal, RN, Brazil. E-mail: pedro.c.carvalho@tecnico.ulisboa.pt*
[2] *Department of Theoretical and Experimental Physics, Federal University of Rio Grande do Norte, Natal, RN, Brazil*
[3] *ISTerre, University Grenoble Alpes, Grenoble, France*
[4] *Department of Biophysics and Pharmacology, Federal University of Rio Grande do Norte, Natal, RN, Brazil*

Accepted 2021 October 25. Received 2021 September 5; in original form 2021 May 7

SUMMARY

In an attempt to overcome the difficulties of the full waveform inversion (FWI), several alternative objective functions have been proposed over the last few years. Many of them are based on the assumption that the residuals (differences between modelled and observed seismic data) follow specific probability distributions when, in fact, the true probability distribution is unknown. This leads FWI to converge to an incorrect probability distribution if the assumed probability distribution is different from the real one and, consequently it may lead the FWI to achieve biased models of the subsurface. In this work, we propose an objective function which does not force the residuals to follow a specific probability distribution. Instead, we propose to use the non-parametric kernel density estimation technique (KDE) (which imposes the least possible assumptions about the residuals) to explore the probability distribution that may be more suitable. As evidenced by the results obtained in a synthetic model and in a typical *P*-wave velocity model of the Brazilian pre-salt fields, the proposed FWI reveals a greater potential to overcome more adverse situations (such as cycle-skipping) and also a lower sensitivity to noise in the observed data than conventional L^2- and L^1-norm objective functions and thus making it possible to obtain more accurate models of the subsurface. This greater potential is also illustrated by the smoother and less sinuous shape of the proposed objective function with fewer local minima compared with the conventional objective functions.

Key words: Inverse theory; Probability distributions; Statistical methods; Waveform inversion; Controlled source seismology; Wave propagation.

1 INTRODUCTION

Full waveform inversion (FWI) is a powerful geophysical technique that makes it possible to obtain accurate and sharp models of the subsurface from seismic data. This technique consists of an optimization problem that aims to find the physical properties of the subsurface that lead to the best fit of the modelled data to the observed seismic data (Lailly 1983; Tarantola 1984; Fichtner 2011; for a review see for example Virieux & Operto 2009 and Virieux *et al.* 2014).

The original FWI formulation proposed by Lailly (1983) and Tarantola (1984), and still the most applied, is based on minimizing the square of the L^2-norm of the residuals (differences between modelled and observed seismic data), which stems from the assumption that these residuals follow a Gaussian probability distribution (Tarantola 2005). However, this assumption is not always the most appropriate since (1) FWI is a nonlinear problem (Amundsen 1991; Fichtner 2011), (2) the amount and quality of information regarding each variable (physical properties of the subsurface) contained in the observed data is quite different (with predominant information regarding the physical properties of the better illuminated regions and little or no information regarding the physical properties of the poorly illuminated regions), (3) the subsurface models are in general quite heterogeneous, (4) the observed data may contain noise that is not random as, for example, coherent noise, and (5) errors from incorrect modelling may be present, for instance, when using the acoustic approximation in situations where elastic modelling would be closer to the actual propagation of the waves (Aravkin *et al.* 2011), among other factors. And, thus, the assumption that the residuals follow a Gaussian probability distribution may lead the

*Now at: Seismic Inversion and Imaging Group, Federal Fluminense Univ., Niterói, RJ, Brazil

FWI to achieve biased models of the subsurface far from the reality (Constable 1988).

In the literature, over the last few years, several alternative objective functions have been proposed for FWI, based on the assumption that the residuals follow well-defined probability distributions different than the Gaussian probability distribution, when in fact the true probability distribution is not known. For example, it is known that FWI based on the L^1-norm of the residuals, which comes from the assumption that the residuals follow a Laplace probability distribution (or double exponential, Tarantola 1987) is less sensitive to large errors and outliers than the L^2-norm. Therefore, it is possible to provide more reliable subsurface models using the L^1-norm when the observed data is noisy and even with outliers (Guitton & Symes 2003; Brossier *et al.* 2010). However, the L^1-norm has a singularity for null residuals. Guitton & Symes (2003) sought to combine the qualities of L^2- and L^1-norm and eliminate their drawbacks in a single objective function: the Huber (1973) norm, to obtain an objective function less sensitive to large residuals and differentiable everywhere. However, although this proposal, as well as other hybrids L^1/L^2 (e.g. Bube & Langan 1997), have considerable advantages over the conventional L^2-norm, they depend on a threshold (which defines where the transition between L^1 and L^2 occurs) that requires 'tedious trial-and-error investigations' to find the best one (Brossier *et al.* 2010). Also, Crase *et al.* (1990) investigated objective functions that are based on the assumption that the residuals follow Cauchy and hyperbolic secant probability distributions, demonstrating that these are also more robust than the L^2-norm.

Later, Aravkin *et al.* (2011) proposed an objective function that derives from the assumption that the residuals follow a Student's *t*-probability distribution demonstrating that it is more robust than the conventional ones and that it is particularly suitable for situations in which data have very poor quality (e.g. with large outliers) or situations where the modelling is poor or far from the real data generating process. Also Yuan & Wang (2013) assuming that the residuals follow a Student's *t*-probability distribution together with non-smooth regularization in a Bayesian framework show that it is possible to retrieve blocky structures and edges of geology body even in the presence of large errors. More recently da Silva *et al.* (2020a, b) proposed objective functions based on generalizations of the Gaussian probability distribution: the κ- and q-generalized Gaussian probability distributions associated with the Kaniadakis (2001) and Tsallis (1988) statistics, respectively. Both proposals provide more accurate models of the subsurface than the conventional L^2-norm, especially when the observed seismic data are very noisy and with outliers.

However, although these proposals may seem promising, they all impose specific probability distributions to the residuals, when in fact, the true probability distribution is unknown. Consequently, imposing probability distributions that may be not be the most suitable FWI may lead to biased models of the subsurface. In addition to the aforementioned objective functions, a wide range of other objective functions has also been proposed over the past few years that do not directly derive from the assumption of a specific probability distribution for the residuals, which also prove to be robust (Tejero *et al.* 2015; Métivier *et al.* 2016) but they are not focused on exploring the true probability distribution of the residuals.

In this work, we do not assume any specific probability distribution for the residuals, but instead, we propose to use the non-parametric kernel density estimation (KDE) technique (Rosenblatt 1956; Parzen 1962; Silverman 1986; Scott 1992) to explore the most suitable probability distribution. This technique aims to estimate the probability density function (PDF) of a random variable (in our problem, the residuals) exclusively from observations of the variable, imposing the least of assumptions on the random variable, other than some degree of smoothing of the probability distribution (Hart 1997; Fan & Yao 2003) and, thus, 'allowing the data speak for themselves' (e.g., Fan & Yao 2003). Therefore, our proposal makes it possible to approach the true probability distribution of the residuals, which can have any shape. For example, it may be asymmetric regarding the null residuals, contrary to what all proposals in the literature have considered so far, which have assumed symmetric probability distributions. It should further be pointed out that there are other techniques in the literature that seek to approximate probability distributions, however, they are mostly parametric techniques, that is, they are based on specific probability distributions, such as, for instance, the well-known Gaussian Mixture Models (GMM) which is based on the assumption that the true probability distribution is composed of a weighted sum of a certain number of Gaussian probability distributions (e.g. Bishop 2006). Conversely, in the KDE technique, each observation contributes with a kernel function (which can be Gaussian or any other, but which in practice has no significant influence on the final estimated probability distribution—Chen 2017) to estimate the probability of each value of the variable, resulting in the estimation of the probability distribution exclusively from the data and therefore in a completely non-parametric way. Note also that Xue *et al.* (2016) have used smoothing kernels in FWI, however, based on a different idea. Xue *et al.* (2016) proposed using smoothing kernels to smooth the residuals in seismic traces (over time). They proposed to start FWI with high smoothing parameters (similar to the bandwidths in our proposal), which corresponds to oversmoothing and consequently giving greater relevance to the lower frequencies and then, in subsequent steps, they proposed to decrease the bandwidths in a way to enable the progressive inclusion of information regarding the highest frequencies in the inversion. In contrast, in the present work, we propose to use the KDE technique to estimate the probability distribution of residuals in a completely non-parametric way. Also in contrast to the Xue *et al.*'s (2016) proposal, our proposal does not require the assumption of bandwidths throughout the FWI, but instead, the (optimal) bandwidths are estimated automatically and exclusively from the residuals themselves.

It is also noteworthy that most of the objective functions proposed in the literature are restricted to considering that the residuals corresponding to each instant of time (in temporal discretization) are completely independent of each other and, therefore, without taking into account possible relations between residuals at different instants of time. Thus these proposals do not have a 'global view' of all the differences between each modelled and observed seismic trace and, as a consequence, they tend to be more susceptible to incurring problems such as, for instance, cycle-skipping problems. This latter problem is one of the major problems that original FWI as well as many of the FWI proposals in the literature face. The cycle-skipping problems stem from the fact that the observed seismic data, in general, does not contain sufficient low-frequency information (usually below 3 Hz at exploration scales, Li & Demanet 2016) to prevent the FWI from being conducted to an incorrect fit during the optimization process when the model is far from the real subsurface. In conventional FWI, the cycle-skipping problems are overcome if one starts the optimization process from an initial model close enough to the true subsurface so that the corresponding modelled data match within half a period associated with the shortest wavelength of the observed data (Bunks *et al.* 1995; Virieux & Operto 2009). However, in most situations, there is no reasonable prior knowledge about the subsurface of the region

under study, making it difficult to know whether the initial model is close to the real subsurface or not. And, as a consequence, the FWI ends up being held hostage by other techniques capable of finding an initial model of the subsurface closest to the real one (Métivier *et al.* 2016). In contrast to most of the proposals in the literature, our proposal makes it possible to assess the relationship between residuals of different instants of time, enabling a global view of all differences between each modelled and observed seismic trace. And, therefore, it is expected that our proposal to have the potential to overcome more adverse situations (such as, e.g. some cycle-skipping situations) and consequently to achieve closer to the real and more accurate models of the subsurface.

The remainder of the paper is organized as follows. In the next section, we briefly review the conventional FWI formulation and then our proposal and the corresponding gradient are presented. In Section 3, three different numerical experiments are presented: in the first, a similar experiment to Mulder & Plessix (2008) is performed to examine the shape of the proposed objective function, then, in the second, our proposal is applied to the inversion of a synthetic model similar to the 'Camembert' model and, finally, our proposal is also applied to a more realistic velocity model which represents a typical *P*-wave velocity model of the Brazilian pre-salt field and the results obtained are compared with those provided by the conventional FWI (L^2-norm and L^1-norm of the residuals). Finally, in Section 4, the main conclusions are summarized.

2 METHODOLOGY

2.1 Conventional full waveform inversion

FWI is a technique formulated as an optimization problem that aims to find the physical properties of the subsurface that lead to the best fit of the computationally modelled data to the observed data. Its original formulation (1) is based on minimizing the square of the L^2-norm of the differences between modelled and observed seismic data (Lailly 1983; Tarantola 1984), which comes from the assumption that these differences (also known as residuals) follow a Gaussian probability distribution (Tarantola 2005):

$$\min S_{L2}(\mathbf{m}) = \frac{1}{2}\sum_{s=1}^{n_s}\sum_{r=1}^{n_r}\int_0^{t_{\max}} \left(d_{r,s}^{\mathrm{mod}}(\mathbf{m},t) - d_{r,s}^{\mathrm{obs}}(t)\right)^2 dt, \tag{1}$$

where $d_{r,s}^{\mathrm{mod}}(\mathbf{m},t)$ and $d_{r,s}^{\mathrm{obs}}(t)$ are the modelled and observed data at the receiver r and at time t, respectively, due to the triggering of the source s. $\mathbf{m}$ represents the model parameters (i.e. the physical properties of the subsurface), $t_{\max}$ is the acquisition time and, n_s and n_r are the number of sources and receivers, respectively.

Assuming that the residuals ($\Delta\mathbf{d} = \mathbf{d}^{\mathrm{mod}} - \mathbf{d}^{\mathrm{obs}}$) are independent and identically distributed (i.i.d.) according to a Gaussian probability distribution with zero mean and unit variance, applying the maximum-likelihood method:

$$\max_{\mathbf{m}} \mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\Delta d_i(\mathbf{m})^2\right), \tag{2}$$

where n is the number of samples of the traces recorded at the receivers. Taking the logarithm of the likelihood function:

$$\max_{\mathbf{m}} \log \mathcal{L} = \sum_{i=1}^{n}\left(\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\Delta d_i(\mathbf{m})^2\right), \tag{3}$$

which is equivalent to:

$$\min_{\mathbf{m}} -\log \mathcal{L} = n\log\left(\sqrt{2\pi}\right) + \sum_{i=1}^{n}\frac{1}{2}\Delta d_i(\mathbf{m})^2, \tag{4}$$

and disregarding the first term (since it does not depend on the model parameters), a discretized version of the objective function of conventional FWI (1) is reached.

In this work, the acoustic approximation is assumed and, therefore, the seismic data correspond to acoustic pressures obtained from the acoustic wave eq. (5) and the model parameters $\mathbf{m}$ are the *P*-wave velocities.

$$\frac{1}{\kappa(x)}\frac{\partial^2 p(x,t)}{\partial t^2} - \nabla.\left(\frac{1}{\rho(x)}\nabla p(x,t)\right) = f(t)\delta(x - x_s), \tag{5}$$

where $\kappa(x)$ is the bulk modulus of the medium at spatial position x of the model and defined as $\kappa(x) = \rho(x)v(x)^2$ (where $\rho(x)$ and $v(x)$ are the density and *P*-wave velocity of the medium, respectively), $p(x,t)$ is the acoustic pressure field and the term $f(t)\delta(x - x_s)$ represents the source $f(t)$ applied at the position x_s.

FWI is summarized in the search for the minimum of the objective function from a method based on the Newton's method: a quasi-Newton method. In this work, the L-BFGS-B quasi-Newton method (limited memory Broyden–Fletcher–Goldfarb–Shanno with boundaries, Liu & Nocedal 1989; Byrd *et al.* 1995; Zhu *et al.* 1997) was used, which has been shown to be one of the most effective methods (Brossier *et al.* 2009; Fei *et al.* 2014). This method estimates an approximation of the inverse of the Hessian matrix from a few gradients of the previous iterations, not requiring too much amounts of memory and, therefore, making the optimization process more efficient. Thus, the solution of FWI is obtained iteratively according to:

$$\mathbf{m}_{l+1} = \mathbf{m}_l + \alpha_l\, \tilde{\mathbf{H}}_S^{-1}(\mathbf{m}_l)\nabla S(\mathbf{m}_l), \tag{6}$$

where $\mathbf{m}_l$ are the model parameters of the lth iteration, $\tilde{\mathbf{H}}_S^{-1}$ is an approximation of the inverse of the Hessian matrix, ∇S is the gradient of the objective function with respect to model parameters and $\alpha_l > 0$ is the step length obtained by line search.

2.1.1 *Gradient computation*

Deriving the objective function (1) with respect to each model parameter m_k, the components of the gradient are obtained:

$$\frac{\partial S_{L2}}{\partial m_k} = \sum_{s=1}^{n_s}\sum_{r=1}^{n_r}\int_0^{t_{\max}} \Delta d_{r,s}(\mathbf{m},t)\frac{\partial d_{r,s}^{\mathrm{mod}}(\mathbf{m},t)}{\partial m_k}dt, \tag{7}$$

where $\Delta d_{r,s}(\mathbf{m},t) = d_{r,s}^{\mathrm{mod}}(\mathbf{m},t) - d_{r,s}^{\mathrm{obs}}(t)$ are the residuals and $\frac{\partial d_{r,s}^{\mathrm{mod}}}{\partial m_k}$ is the Jacobian or the Fréchet derivative of the modelled data with respect to the model parameter m_k.

Assuming, for instance, that the model parameters are the *P*-wave velocities, the Fréchet derivative of the modelled data with respect to the model parameter k corresponds to a wavefield resulting from a disturbance $\delta_k(t)$ of the model parameter (at position x_k of the model) (8). Therefore, for the computation of all components of the gradient, from the explicit computation of the Fréchet derivative, it would be necessary to solve one wave equation for each model parameter, which would be impractical in terms of computational cost. Thus, an adjoint formulation (Tarantola 1984; Plessix 2006; Yang *et al.* 2015) is used, which makes it possible to estimate all components of the gradient simultaneously and efficiently from just the solution of only two wave equations: from a propagation

forward in time to compute the wavefield p_o and from a propagation backwards in time to compute the wavefield $p_{\Delta d}^{\text{back}}$, according to:

$$\frac{\partial S_{L2}}{\partial v_k} = \sum_{s=1}^{n_s} \sum_{r=1}^{n_r} \int_0^{t_{\max}} \delta_k(t)\ p_{\Delta d}^{\text{back}}(x_k, t)\, dt, \tag{8}$$

where:

$$\delta_k(t) = \frac{2}{\kappa(x_k) v(x_k)} \frac{\partial^2 p_o(x_k, t)}{\partial t^2},$$

where $p_{\Delta d}^{\text{back}}$ is the wavefield resulting from the propagation backwards in time of the residuals $\Delta\mathbf{d}$ from the corresponding receivers and p_o is the wavefield obtained from the propagation of the source $f(t)$ (from eq. 5).

2.2 The KDE technique and the proposed FWI

Bearing in mind that the true probability distribution of the residuals is usually unknown, in this work a specific probability distribution for the residuals is not assumed, but instead, we seek to explore a probability distribution that best suits the residuals using the non-parametric KDE technique.

2.2.1 The KDE method

The KDE is a non-parametric method that aims to estimate the PDF $f(x)$ of a random variable X from a sample of observations $x_1, x_2, \ldots, x_n$ (in our case, the residuals), using the following expression (see e.g. Chen 2017; Hansen 2009; or Li & Racine 2007):

$$\tilde{f}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right), \tag{9}$$

where $K(.)$ is a kernel function and $h > 0$ is the bandwidth (or smoothing parameter).

The kernel function aims to define the contribution of the observations to the estimation of the probability of a given value of the variable X and the bandwidth governs the window of observations that contribute to the estimation of that probability. There are several kinds of kernel functions, however, the most commonly used are the Gaussian and Epanechnikov. In this work, the Gaussian kernel was used, which is defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right). \tag{10}$$

Whereas the choice of the kernel function usually does not have a significant effect on the estimation of the PDF, the bandwidth plays a crucial role (Chen 2017). If h is too high, the PDF becomes too smooth with few details and, therefore, important structures can be obscured due to the enormous amount of smoothing (i.e. oversmoothing occurs). On the other hand, if h is too small it results in a PDF with many structures that might derive just randomness (i.e. occurring undersmoothing, see, for instance, fig. 3 in Chen 2017). There are several methods to estimate the optimal bandwidth, such as cross-validation methods, plug-in methods among others (Chen 2017), however, a simpler way is to consider the rule-of-thumb suggested by Silverman (1986): $h = 0.9 \min(\sigma_n, \text{IQR}/1.34)\, n^{-1/5}$ for Gaussian kernels (where σ_n is the standard deviation of the sample data and IQR is the interquartile range).

On the other hand, adopting a single fixed (or global) bandwidth may be detrimental, since it may lead to an oversmoothing where the data (in our case, the residuals) are denser in the probability distribution and to an undersmoothing where the data are sparser (Van Kerm 2003). For this reason, in the present work, the Adaptive KDE method (Abramson 1982) is used, which enables the use of bandwidths that vary locally in the probability distribution. In this method, local bandwidths h_j are estimated for the region around each observation x_j and which are mainly a function of the local concentration of data around x_j (denoted by $\tilde{f}(x_j)$) and the global bandwidth h (Van Kerm 2003):

$$h_j = \lambda_j\, h, \tag{11}$$

where λ_j is:

$$\lambda_j(x_j) = \sqrt{\frac{G}{\tilde{f}(x_j)}},$$

and where $\tilde{f}(x_j)$ is the density estimate at x_j assuming a global bandwidth h obtained from eq. (9) and G is the geometric mean over all j of the density estimate $\tilde{f}(x)$:

$$G = \left(\prod_{j=1}^{n} \tilde{f}(x_j)\right)^{1/n} = \exp\left\{\frac{1}{n}\sum_{j=1}^{n} \log\left(\tilde{f}(x_j)\right)\right\}.$$

Hence, the greater the data concentration around x_j (denoted by a high $\tilde{f}(x_j)$) the lower will be λ_j and consequently the lower the local bandwidth h_j and the opposite where the data are more sparse.

And lastly, the PDF estimate in Adaptive KDE method is obtained from:

$$\widehat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_j} K\left(\frac{x - x_j}{h_j}\right). \tag{12}$$

This means that the estimate of the probability distribution at x corresponds to the sum of the contribution of each observation x_j with a kernel function weighted by the inverse of the bandwidth parameter h_j. The closer the value of observation x_j is to x the greater its contribution to the estimate of the probability distribution at x. Furthermore, if there are a large number of observations with values close to x_j, the bandwidth associated with x_j (h_j) will tend to be small, thus corresponding to a steep kernel function (approaching a Dirac delta function in the case of a Gaussian kernel) and, therefore, observation x_j will only actually contribute to the estimate of the probability distribution at x if it has a value very close to x. Conversely, if there are few observations with values close to x_j (as for instance in the case of outliers), h_j will tend to be high, corresponding to x_j a flatter kernel and, therefore, the observation x_j will contribute to the estimate of the probability distribution at x even if it is relatively far away from the x value.

2.2.2 Proposed FWI

Assuming that the residuals (Δd) corresponding to each seismic trace are i.i.d. and that their PDF can be estimated non-parametrically from eq. (12) and considering the Gaussian kernel (10), their PDF will be defined by:

$$\widehat{f}(\Delta d) = \frac{1}{n\sqrt{2\pi}} \sum_{j=1}^{n} \frac{1}{h_j} \exp\left(-\frac{1}{2}\left(\frac{\Delta d - \Delta d_j}{h_j}\right)^2\right). \tag{13}$$

Applying the maximum-likelihood method:

$$\max \mathcal{L} = \prod_{i=1}^{n} \text{probability}\left(\Delta d_i\right), \tag{14}$$

$$\max_{\mathbf{m}} \mathcal{L} = \prod_{i=1}^{n} \frac{1}{n\sqrt{2\pi}} \sum_{j=1}^{n} \frac{1}{h_j(\mathbf{m})} \times \exp\left(-\frac{1}{2}\left(\frac{\Delta d_i(\mathbf{m}) - \Delta d_j(\mathbf{m})}{h_j(\mathbf{m})}\right)^2\right), \quad (15)$$

and taking the logarithm of the likelihood function:

$$\max_{\mathbf{m}} \log \mathcal{L} = \sum_{i=1}^{n} \log\left[\frac{1}{n\sqrt{2\pi}} \sum_{j=1}^{n} \frac{1}{h_j} \times \exp\left(-\frac{1}{2}\left(\frac{\Delta d_i - \Delta d_j}{h_j}\right)^2\right)\right], \quad (16)$$

which is equivalent to:

$$\min_{\mathbf{m}} -\log \mathcal{L} = -n \log\left(\frac{1}{n\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \log\left[\sum_{j=1}^{n} \frac{1}{h_j} \exp\left(-\frac{1}{2}\left(\frac{\Delta d_i - \Delta d_j}{h_j}\right)^2\right)\right], \quad (17)$$

the proposed objective function is obtained (for all sources and receivers):

$$\min_{\mathbf{m}} S_p(\mathbf{m}) = n_s\, n_r\, n \log\left(n\sqrt{2\pi}\right) - \sum_{s=1}^{n_s}\sum_{r=1}^{n_r}\sum_{i=1}^{n} \log\left[\sum_{j=1}^{n} \frac{1}{h_j^{r,s}(\mathbf{m})} \exp\left(-\frac{1}{2} u_{i,j}^2(\mathbf{m})\right)\right], \quad (18)$$

where:

$$u_{i,j}(\mathbf{m}) = \frac{\Delta d_i^{r,s}(\mathbf{m}) - \Delta d_j^{r,s}(\mathbf{m})}{h_j^{r,s}(\mathbf{m})},$$

and $\Delta d_i^{r,s}(\mathbf{m}) = d_i^{\text{mod}}(\mathbf{m}) - d_i^{\text{obs}}$ and $\Delta d_j^{r,s}(\mathbf{m}) = d_j^{\text{mod}}(\mathbf{m}) - d_j^{\text{obs}}$ are the residuals at the instant of time i and j, respectively, $h_j^{r,s}(\mathbf{m})$ is the bandwidth, n is the number of samples of the trace recorded at the receiver position (i.e. the number of time steps in time discretization) and n_s and n_r are the number of sources and receivers, respectively.

A close look at the proposed objective function shows that it not only seeks to minimize the residuals of each instant of time but also it seeks to minimize the differences between the residuals at different instants of time, in particular the residuals of the closest value which also are usually in neighbouring instants of time. This last characteristic comes from the only relevant imposition by the KDE technique, which is the imposition of smoothness in the probability distribution of the residuals and which in the case of the proposed objective function it also leads to that the differences between the modelled and observed data are not abrupt over time. Note also that low (local) bandwidths $h_j^{r,s}$ lead to evaluations of differences only between similar residual values, which are often (but not only) found at neighbouring time instants in the seismic trace, whereas high bandwidths are associated with an evaluation of differences between a wider range of residual values and, therefore, possibly related to the evaluation of the relationship with more neighbours in the seismic trace.

It should also be noted that if we assume that in each seismic trace there is no relationship between the residuals corresponding to different instants of time and that they are completely independent of each other, the Δd_j term in the eq. (18) can be disregarded and, furthermore, if unit local bandwidths ($h_j = 1$) are assumed and constant terms are ignored, the proposed objective function becomes the conventional L^2-norm FWI.

2.2.3 *Computation of the proposed objective function gradient*

Taking the derivative of the proposed objective function (18) with respect to each model parameter m_k, one obtains the components of the gradient (for only one source and one receiver):

$$\frac{\partial S_p}{\partial m_k} = -\sum_{i=1}^{n} \frac{1}{\beta_i} \sum_{j=1}^{n} \frac{\gamma_{i,j}}{h_j} \times \left[-\frac{\partial h_j}{\partial m_k} + u_{i,j}\left(u_{i,j}\frac{\partial h_j}{\partial m_k} - \frac{\partial\left(\Delta d_i - \Delta d_j\right)}{\partial m_k}\right)\right], \quad (19)$$

where:

$$\gamma_{i,j} = \frac{\exp\left(-\frac{1}{2}u_{i,j}^2\right)}{h_j},$$

$$\beta_i = \sum_{l=1}^{n} \gamma_{i,l},$$

and, therefore:

$$\frac{\partial S_p}{\partial m_k} = \sum_{i=1}^{n} \frac{1}{\beta_i} \sum_{j=1}^{n} \frac{\gamma_{i,j}}{h_j} \times \left[\frac{\partial h_j}{\partial m_k}\left(1 - u_{i,j}^2\right) + u_{i,j}\frac{\partial\left(\Delta d_i - \Delta d_j\right)}{\partial m_k}\right]. \quad (20)$$

For the sake of simplicity, assuming that the factor λ_j of the local bandwidths does not depend on the model parameters and that the global bandwidth depends only on the standard deviation, the $\frac{\partial h_j}{\partial m_k}$ is:

$$\frac{\partial h_j}{\partial m_k} = \frac{\partial\left(\lambda_j\, h\right)}{\partial m_k} = \frac{\partial\left(\lambda_j\, 0.9\, \sigma_n\, n^{-1/5}\right)}{\partial m_k} = 0.9\, n^{-1/5}\lambda_j \frac{\partial \sigma_n}{\partial m_k} = \frac{h_j}{\sigma_n}\frac{\partial \sigma_n}{\partial m_k}. \quad (21)$$

And considering that the standard deviation of the residuals σ_n is:

$$\sigma_n = \sqrt{\frac{1}{n-1}\sum_{l=1}^{n}\left(\Delta d_l - \overline{\Delta d}\right)^2}, \quad (22)$$

which can also be written as:

$$\sigma_n = \sqrt{\frac{1}{n-1}\left(\sum_{l=1}^{n}\Delta d_l^2 - \frac{1}{n}\left(\sum_{l=1}^{n}\Delta d_l\right)^2\right)}, \quad (23)$$

where $\overline{\Delta d}$ is the mean of the residuals: $\overline{\Delta d} = \frac{1}{n}\sum_{l=1}^{n}\Delta d_l$. The derivative of the standard deviation of the residuals with respect to model parameter m_k is:

$$\frac{\partial \sigma_n}{\partial m_k} = \frac{1}{(n-1)\,\sigma_n}\left(\sum_{l=1}^{n}\Delta d_l\left(\frac{\partial d^{\text{mod}}}{\partial m_k}\right)_l - \overline{\Delta d}\sum_{l=1}^{n}\left(\frac{\partial d^{\text{mod}}}{\partial m_k}\right)_l\right), \quad (24)$$

or:

$$\frac{\partial \sigma_n}{\partial m_k} = \sum_{l=1}^{n}\frac{\Delta d_l - \overline{\Delta d}}{(n-1)\,\sigma_n}\left(\frac{\partial d^{\text{mod}}}{\partial m_k}\right)_l, \quad (25)$$

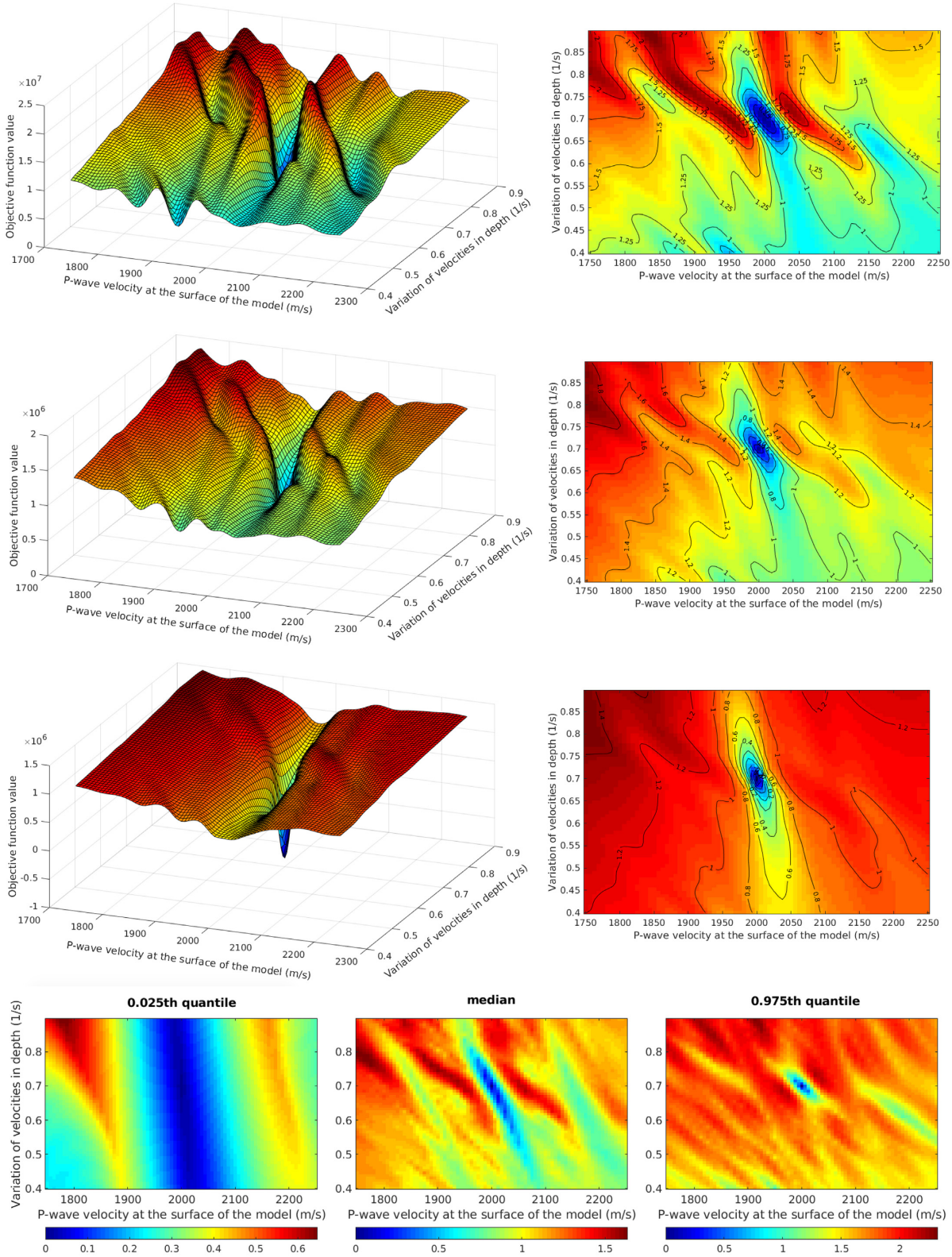


Figure 1. Variation of the value of the objective functions of conventional FWI (L^2-norm of the residuals) (above), of the objective function corresponding to the L^1-norm of the residuals (middle) and of the proposed objective function (below) with the parameters that define the velocity models: the *P*-wave velocity at the top of the model (v_i) and the velocity gradient in depth (φ). And statistics of the global bandwidths (median and limits of the 95 per cent interval) associated with each *P*-wave velocity model for the proposed objective function case (below).

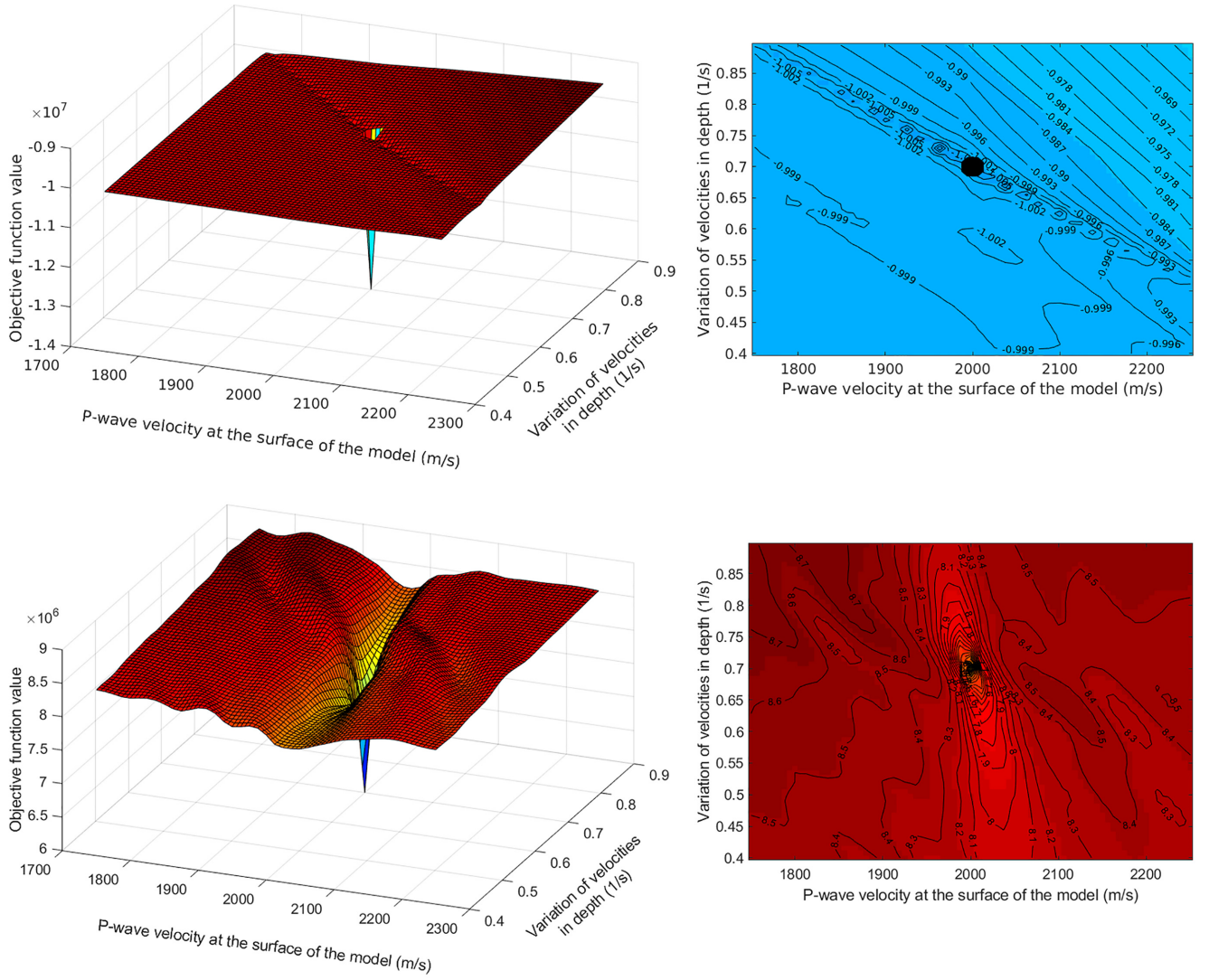


Figure 2. Variation of the value of the proposed objective function with the parameters that define the velocity models [*P*-wave velocity at the top of the model (v_i) and the velocity gradient in depth (φ)] for case of too small (above) and too high (below) global bandwidths ($h = 10^{-8}$ and 10^4, respectively).

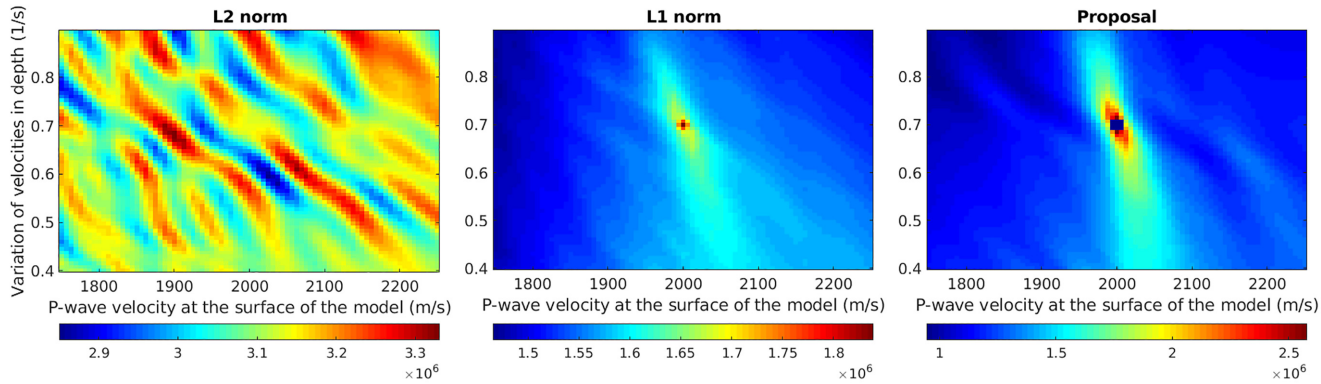


Figure 3. Differences between the value of the objective functions corresponding to the situation in which the 'observed' data contain Gaussian noise and the situation in which the 'observed' data have no noise, for the various objective functions.

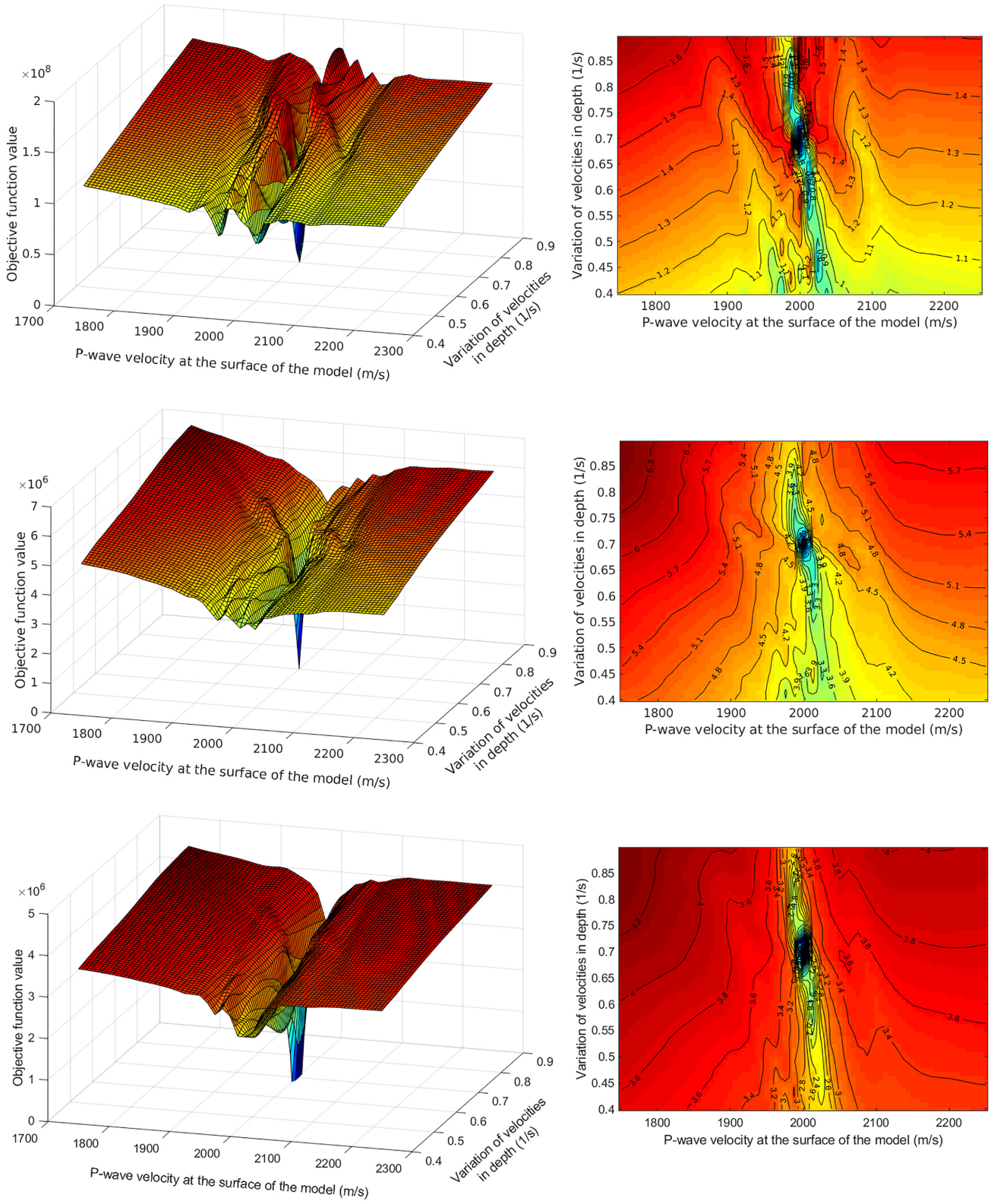


Figure 4. Variation of the value of the objective functions of conventional FWI (L^2-norm) (above), L^1-norm (middle) and of the proposed objective function (below) with the parameters that define the velocity models [the *P*-wave velocity at the top of the model (v_i) and the velocity gradient in depth (φ)] for the case where the 'observed' data contains a higher frequencies content (between about 3 and 60 Hz).

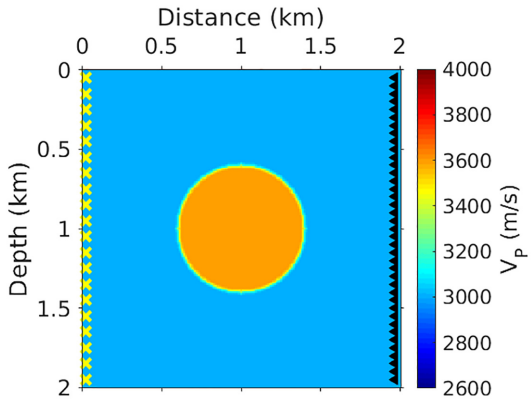


Figure 5. True *P*-wave velocity model.

and, therefore, the derivative of the proposed objective function with respect to model parameter m_k becomes:

$$\frac{\partial S_p}{\partial m_k} = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\gamma_{i,j}}{\beta_i}\left[\frac{1}{\sigma_n}\frac{\partial \sigma_n}{\partial m_k}\left(1-u_{i,j}^2\right) + \frac{u_{i,j}}{h_j}\left(\left(\frac{\partial d^{\mathrm{mod}}}{\partial m_k}\right)_i - \left(\frac{\partial d^{\mathrm{mod}}}{\partial m_k}\right)_j\right)\right], \tag{26}$$

and, finally, leading to (for the several sources and receivers):

$$\frac{\partial S_p}{\partial m_k} = \sum_{s=1}^{n_s}\sum_{r=1}^{n_r}\sum_{i=1}^{n}\Theta_i\left(\frac{\partial d^{\mathrm{mod}}}{\partial m_k}\right)_i, \tag{27}$$

where:

$$\Theta_i = \frac{\Delta d_i^{r,s} - \overline{\Delta d^{r,s}}}{(n-1)\left(\sigma_n^{r,s}\right)^2}\psi^{r,s} + \sum_{j=1}^{n}\left(\Delta d_i^{r,s} - \Delta d_j^{r,s}\right)\left(\frac{\gamma_{i,j}^{r,s}}{\beta_i^{r,s}\left(h_j^{r,s}\right)^2} + \frac{\gamma_{j,i}^{r,s}}{\beta_j^{r,s}\left(h_i^{r,s}\right)^2}\right), \tag{28}$$

and

$$\psi^{r,s} = \sum_{l=1}^{n}\left(\frac{\sum_{o=1}^{n}\gamma_{l,o}^{r,s}\left(1-u_{l,o}^2\right)}{\beta_l^{r,s}}\right),$$

$$u_{i,j} = \frac{\Delta d_i^{r,s} - \Delta d_j^{r,s}}{h_j^{r,s}},$$

$$u_{j,i} = \frac{\Delta d_j^{r,s} - \Delta d_i^{r,s}}{h_i^{r,s}},$$

and $\Delta d_i^{r,s}(\mathbf{m}) = d_i^{\mathrm{mod}}(\mathbf{m}) - d_i^{\mathrm{obs}}$ and $\Delta d_j^{r,s}(\mathbf{m}) = d_j^{\mathrm{mod}}(\mathbf{m}) - d_j^{\mathrm{obs}}$ are the residuals at the instant of time i and j, respectively, $h_i^{r,s}(\mathbf{m})$ and $h_j^{r,s}(\mathbf{m})$ are the bandwidths, $\left(\sigma_n^{r,s}\right)^2$ is the variance of the residuals corresponding to the receiver r and the source s, n is the number of samples of the trace recorded at the receiver position and n_s and n_r are the number of sources and receivers, respectively.

Comparing the gradient of the proposed objective function (27) and the gradient of the conventional objective function (7) and bearing in mind that, in discrete terms, the integral in eq. (7) represents the inner product between the residuals Δd and the Fréchet derivative, it can be seen that the residuals in the gradient of the conventional objective function were replaced by a vector with components Θ_i (corresponding to time instant i) in the gradient of the proposed objective function. Therefore, the efficient gradient of the proposed objective function consists of the propagation of a signal with components Θ_i (28) backward in time instead of simply the residuals as in gradient of conventional FWI (8).

Note also that the gradient will tend to zero when all residuals (in each trace) are as close as possible to each other and they approach to the mean of the residuals, which also means that, in contrast to most of the objective functions proposed in the literature, the proposed objective function can deal with situations in which the mean of the residuals is not exactly zero.

3 NUMERICAL EXPERIMENTS

In order to demonstrate the potential of our proposal, its application is presented below in several experiments together with the conventional objective functions L^2- and L^1-norms of the residuals. As a reminder, the conventional L^1-norm seeks to minimize the absolute value of the residuals instead minimizing the square of the residuals as in L^2-norm (1) and in the gradient computation the sign of the residuals is propagated backwards in time instead of propagating the residuals themselves as in L^2-norm (8).

For the purpose of examining the shape of the proposed objective function and its possible behaviour in FWI, firstly, a similar experiment to Mulder & Plessix's (2008) is carried out, where one investigates how the objective function varies with the variation of a velocity model defined by only two parameters. Then, the proposed objective function is applied to the inversion of a simple model similar to the Camembert model (Gauthier *et al.* 1986) and the obtained velocity models are compared with the velocity models obtained by conventional objective functions. And finally, the proposed objective function is applied to a more realistic velocity model which represents a typical *P*-wave velocity model of the Brazilian pre-salt field.

3.1 Shape of the objective functions for a velocity model defined by only two parameters

In order to get an idea of the shape of the proposed objective function and its behaviour in FWI, a similar experiment to Mulder & Plessix's (2008) was carried out. This experiment consists of evaluating how the value of the objective function varies with the variation of a simple *P*-wave velocity model with linearly increasing velocities in depth (and constant in the horizontal direction) defined by only two parameters: the *P*-wave velocity at the surface of the model (i.e. at zero depth, v_i in $\mathrm{m\,s^{-1}}$) and the velocity gradient in depth (φ in $(\mathrm{m\,s^{-1}})\,\mathrm{m^{-1}}$ or $\mathrm{s^{-1}}$), according to the following expression:

$$\text{velocities}(z,x) = v_i + \varphi\, z, \tag{29}$$

where z and x are the depth and the distance in the horizontal direction, respectively (both in metres).

The velocities at the top of the model were assumed to vary between 1750 and 2250 $\mathrm{m\,s^{-1}}$ (discretized at every 12.5 $\mathrm{m\,s^{-1}}$) and the velocity gradients to vary between 0.4 and 0.9 (discretized at every 0.015 $\mathrm{s^{-1}}$). The velocity models are 17 km long and 3.5 km deep.

The 'observed' seismic data were obtained from a model defined by a velocity at the top of the model of $v_i = 2000\ \mathrm{m\,s^{-1}}$ and by a velocity gradient of $\varphi = 0.7\ \mathrm{s^{-1}}$, which was assumed to be the true model. The acquisition geometry consisted of only one source located at the top centre of the model at a distance of $x = 8.5$ km and 60 m deep and 169 receivers located at the top of the model at 60 m deep between the distances $x = 100$ and 16900 m, equally spaced every 100 m. The seismic source was assumed to be a Ricker wavelet with a peak frequency of 5.5 Hz and in order to simulate a situation

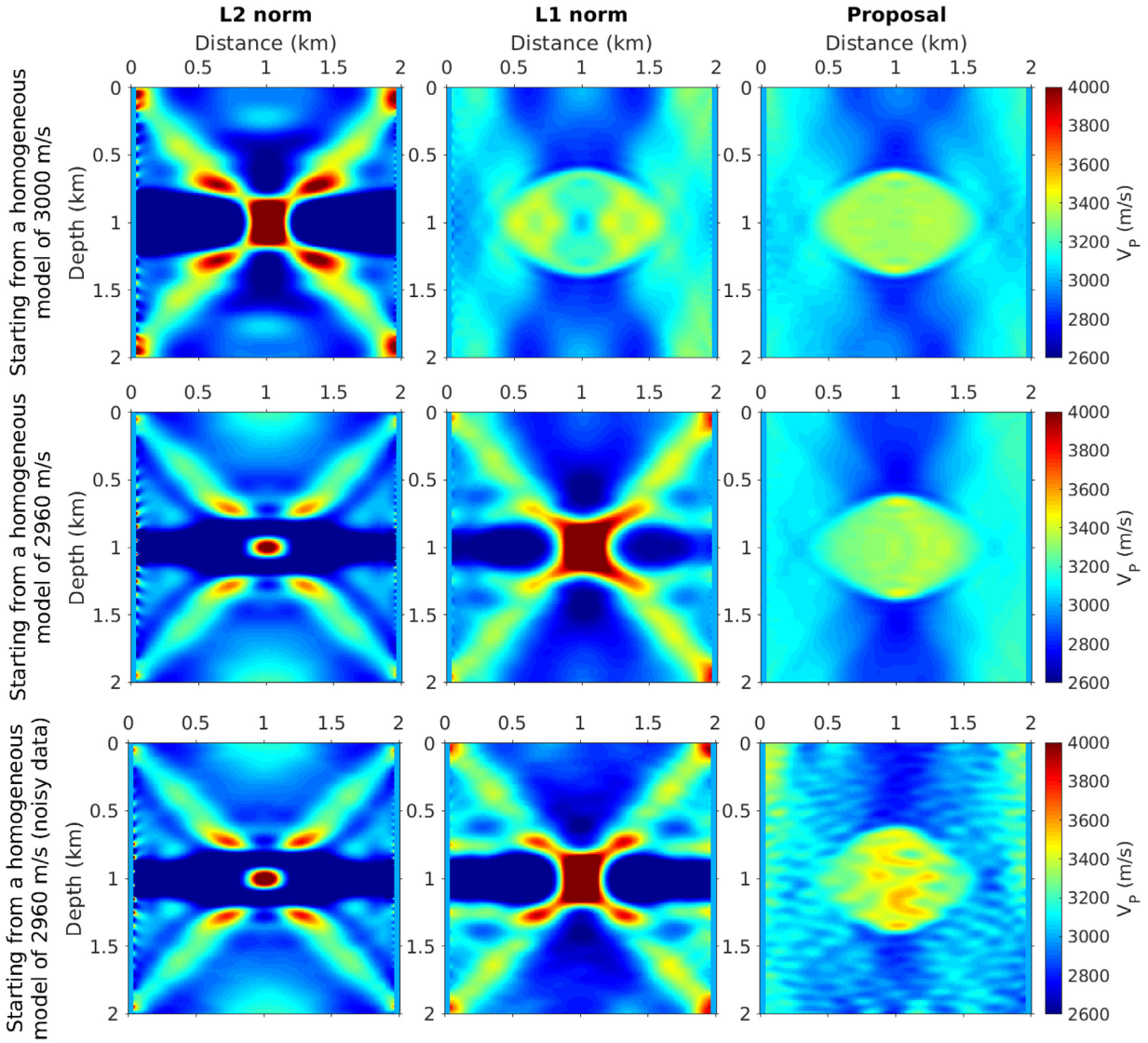


Figure 6. FWI results obtained by the conventional L^2- and L^1-norms and proposed objective functions when starting from the homogeneous model of velocity 3000 $\mathrm{m\,s^{-1}}$ (above) and from the homogeneous model of velocity 2960 $\mathrm{m\,s^{-1}}$ (middle). And FWI results when starting from the homogeneous model of velocity 2960 $\mathrm{m\,s^{-1}}$ and where the 'observed' data contain Gaussian noise (SNR of about 7 dB) (below).

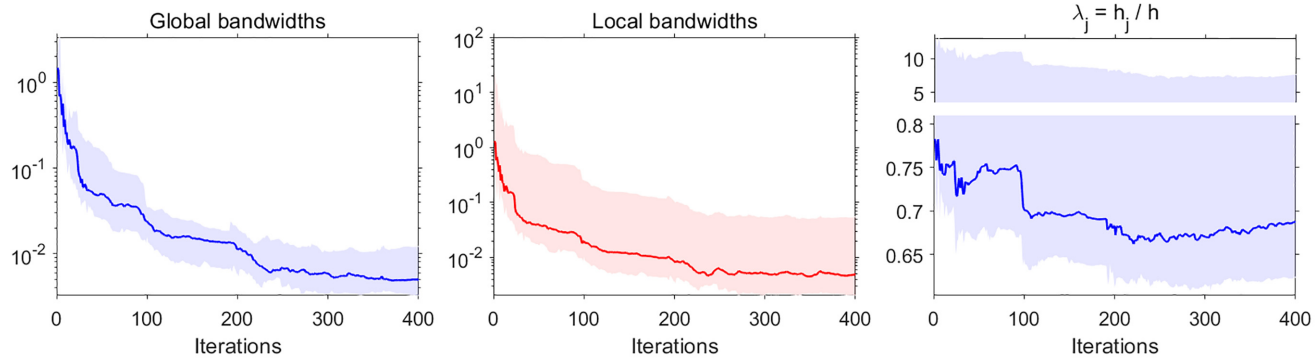


Figure 7. Statistics (median and limits of the 95 per cent interval) of global and local bandwidths and lambdas over the FWI corresponding to the source at 950 m deep in the Camembert model when one starts from a homogeneous model of velocity 3000 $\mathrm{m\,s^{-1}}$.

closer to the real one, frequencies below 3 Hz were removed. The total recording time was 5 s. In the modelling of the wavefields, the acoustic approximation is used and as boundary conditions the unsplit convolutional perfectly matched layers (C-PML) absorbing boundary condition (Komatitsch & Martin 2007) is used in all the surroundings of the models to avoid the reflection of the waves at the limits of the models, except at the top of the models where a free-surface condition was assumed.

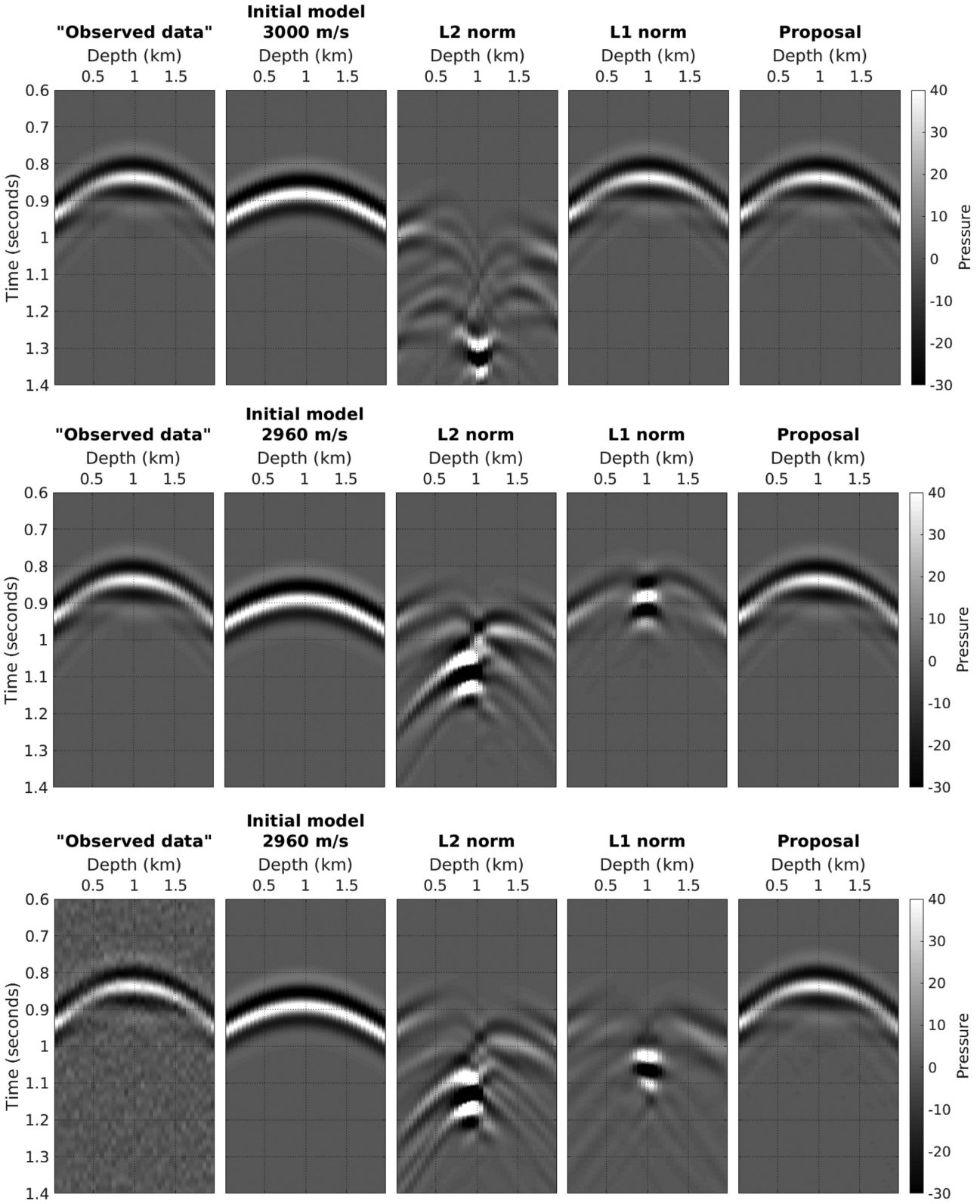
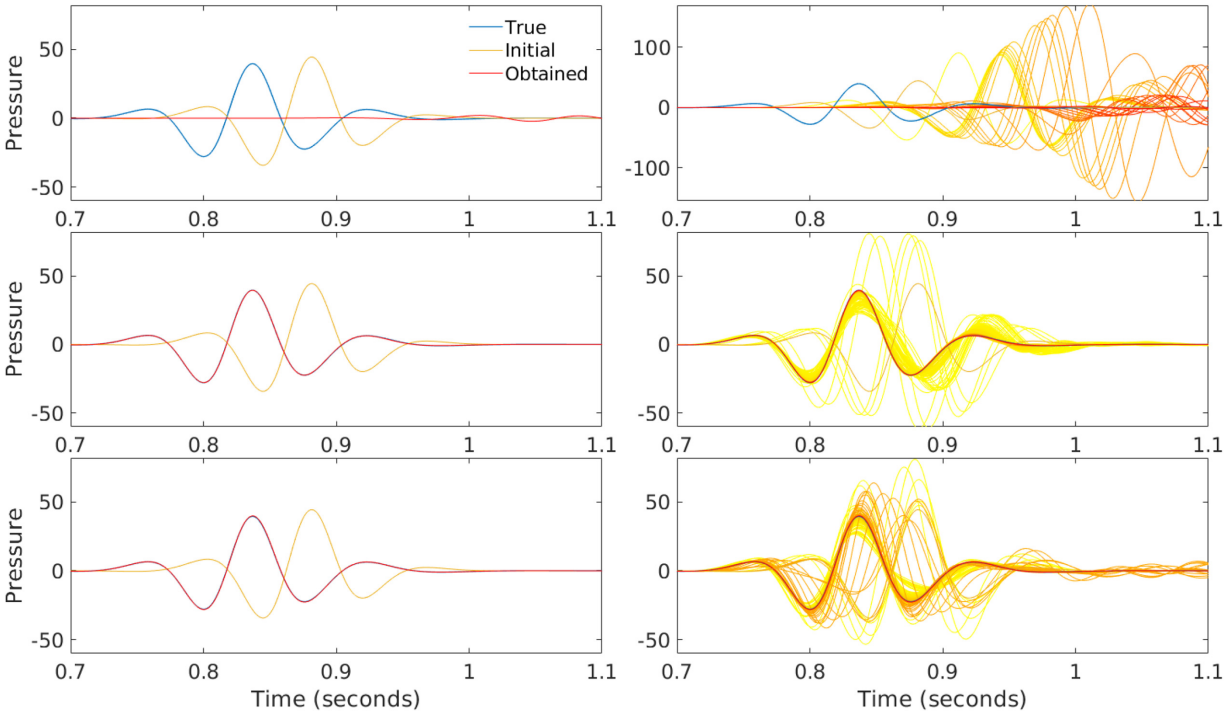
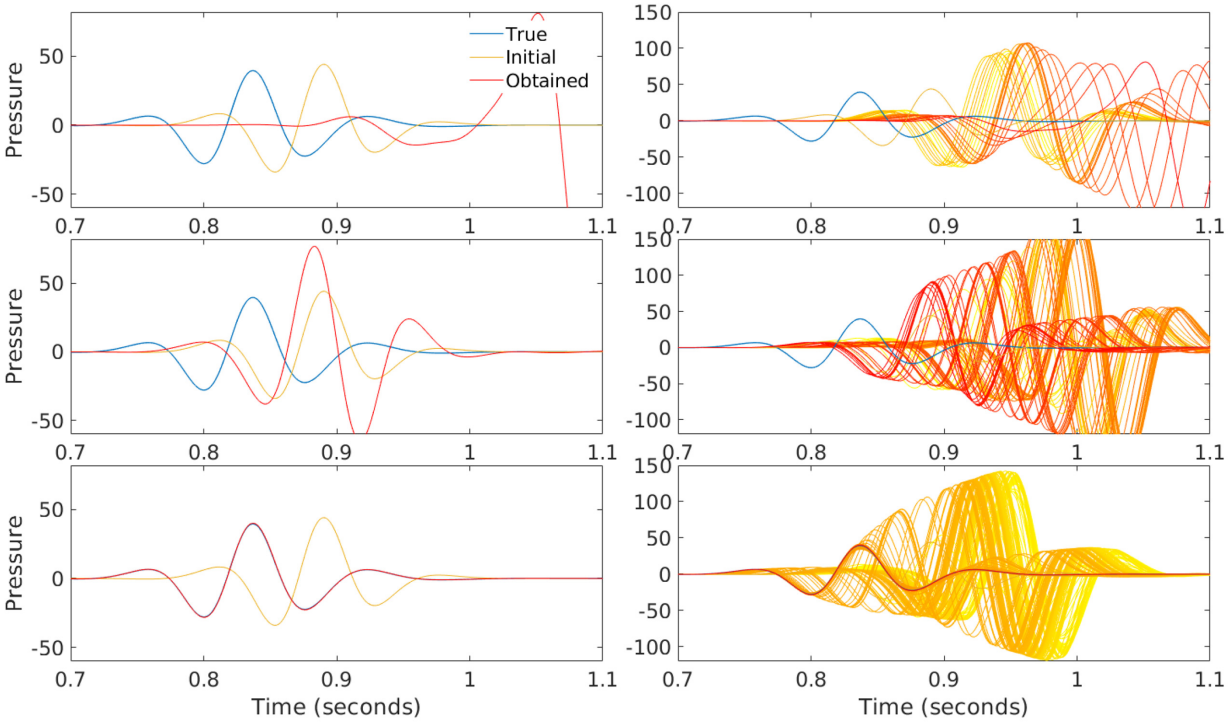


Figure 8. Common shot gathers corresponding to the source fired at 950 m deep obtained in the true, initial velocity models and in the velocity models obtained by the conventional L^2- and L^1-norms and proposed objective functions when starting from the homogeneous model of velocity 3000 $\mathrm{m\,s^{-1}}$ (above), from the homogeneous model of velocity 2960 $\mathrm{m\,s^{-1}}$ (middle) and for the case in which the 'observed' data contain noise (SNR of about 7 dB) (below) in the Camembert model case.



(a) For the case when starting from a homogeneous model of velocity 3000 m/s.



(b) For the case when starting from a homogeneous model of velocity 2960 m/s.

Figure 9. Traces recorded in the receiver at 950 m depth (due to the source fired at the same depth) obtained in the true, initial velocity models and in the velocity models obtained by the conventional L^2 (above) and L^1-norm (middle) and proposed (below) objective functions (left) and its evolution along the FWI's iterations (on the right) in the Camembert model case (the light yellow traces correspond to the first iteration, becoming more reddish throughout the iterations until reaching the red colour in the last iteration).

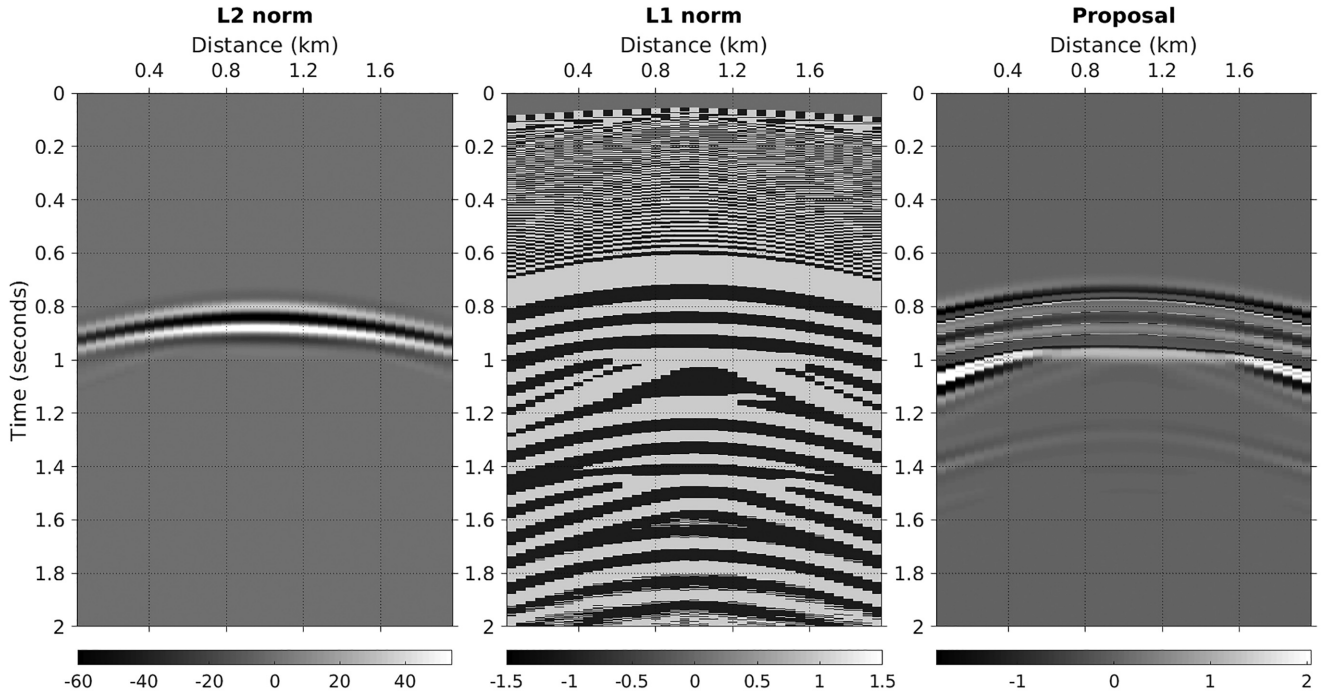


Figure 10. Adjoint sources of conventional L^2- and L^1-norms and proposed objective function for Camembert model when starting from the homogeneous model of velocity 3000 $\mathrm{m\,s^{-1}}$ for the source fired at 950 m deep.

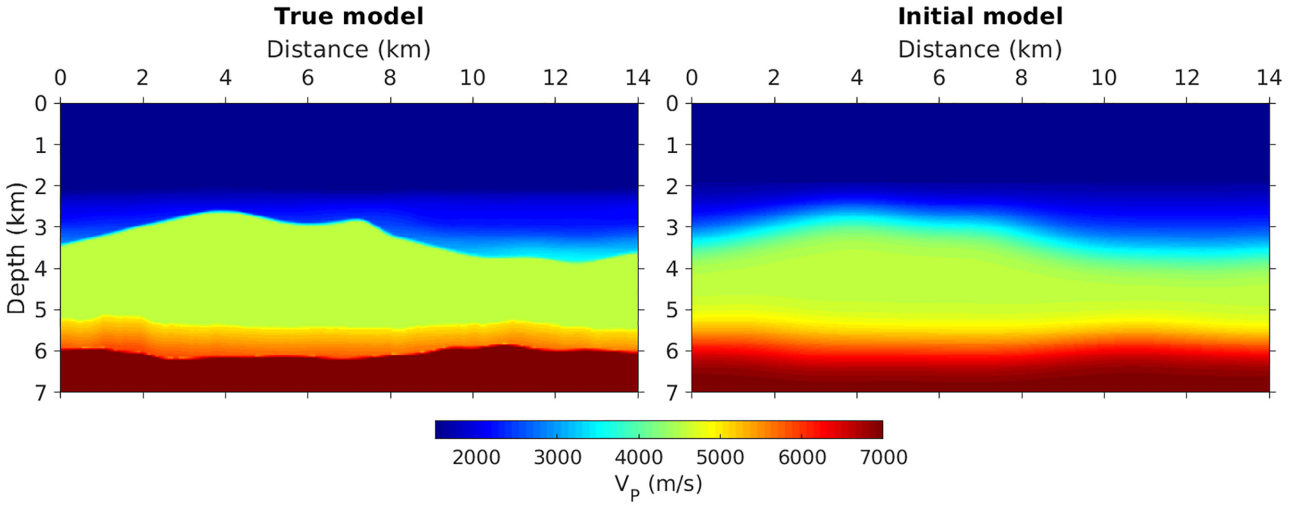


Figure 11. Left: true velocity model (typical *P*-wave velocity model of the Brazilian pre-salt field) and right: initial velocity model.

In Fig. 1, the shape of the conventional objective functions L^2- and L^1-norms of the residuals and of the proposed objective function (and corresponding statistics of the global bandwidths) are presented. As can be seen, the proposed objective function has a much smoother and less sinuous shape with fewer and much less pronounced local minima and also a well defined and more pronounced global minimum than conventional objective functions. This global minimum, in fact, corresponds to a high negative value of the objective function since as one moves towards the true velocity model the local bandwidths h_j tend to zero (around the zero residual) leading the proposed objective function to tend to $-\infty$ (eq. 18). Note that, contrary to what happens in conventional objective functions, in the proposed objective function the global minimum is not surrounded by steep hills difficult to cross, although further away from the global minimum some hills and local minima still persist, however, they are quite more smoothed or even non-existent and therefore much easier to overcome. These facts lead us to believe that our proposal has a better performance in FWI than the conventional objective functions.

Furthermore, it is also observed that if constant and too small global bandwidths are assumed [instead of local bandwidths, i.e. if using eq. (9) with constant global bandwidth instead of eq. (12) for the PDF estimation], the proposed objective function tends to become almost flat and with tenuous local minima (Fig. 2). And, on the contrary, if constant and too high global bandwidths are assumed, the proposed objective function becomes more sinuous and consequently with more pronounced local minima (Fig. 2). These facts, corresponding to limit situations, demonstrate the importance

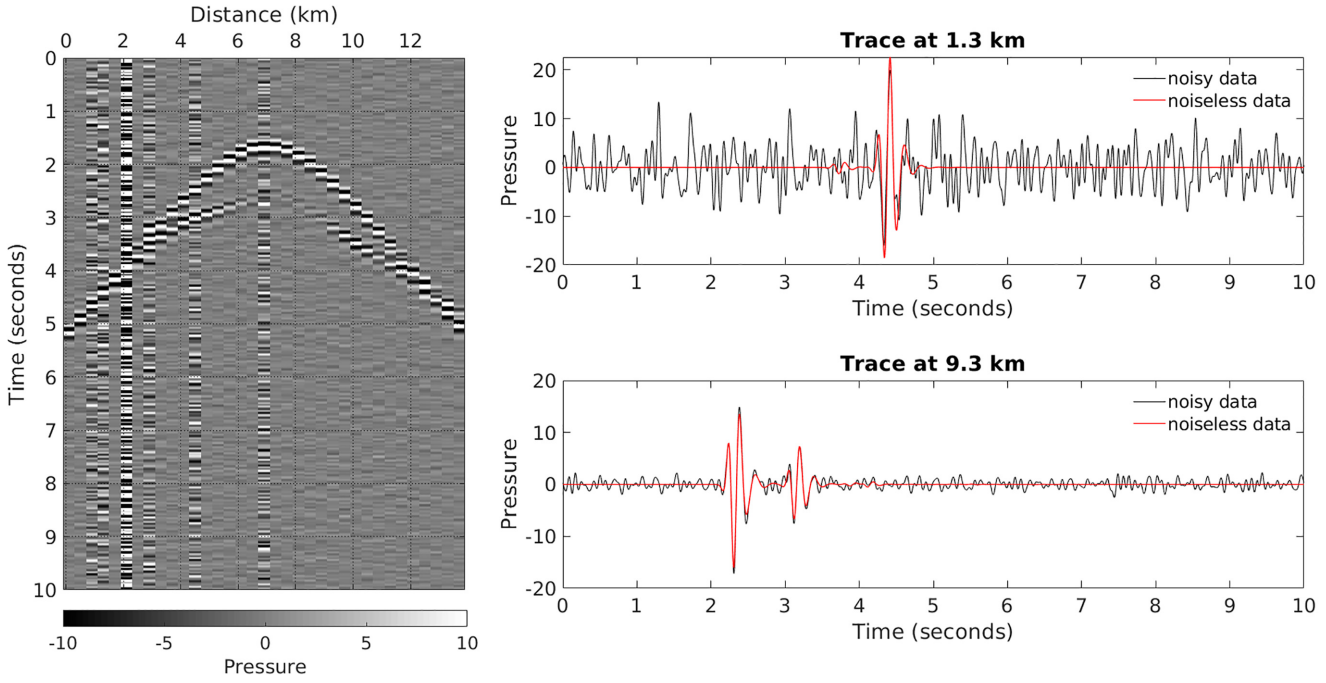


Figure 12. 'Observed' seismogram corresponding to the source fired at distance of 7 km (left) and examples of corresponding seismic traces with and without noise (right).

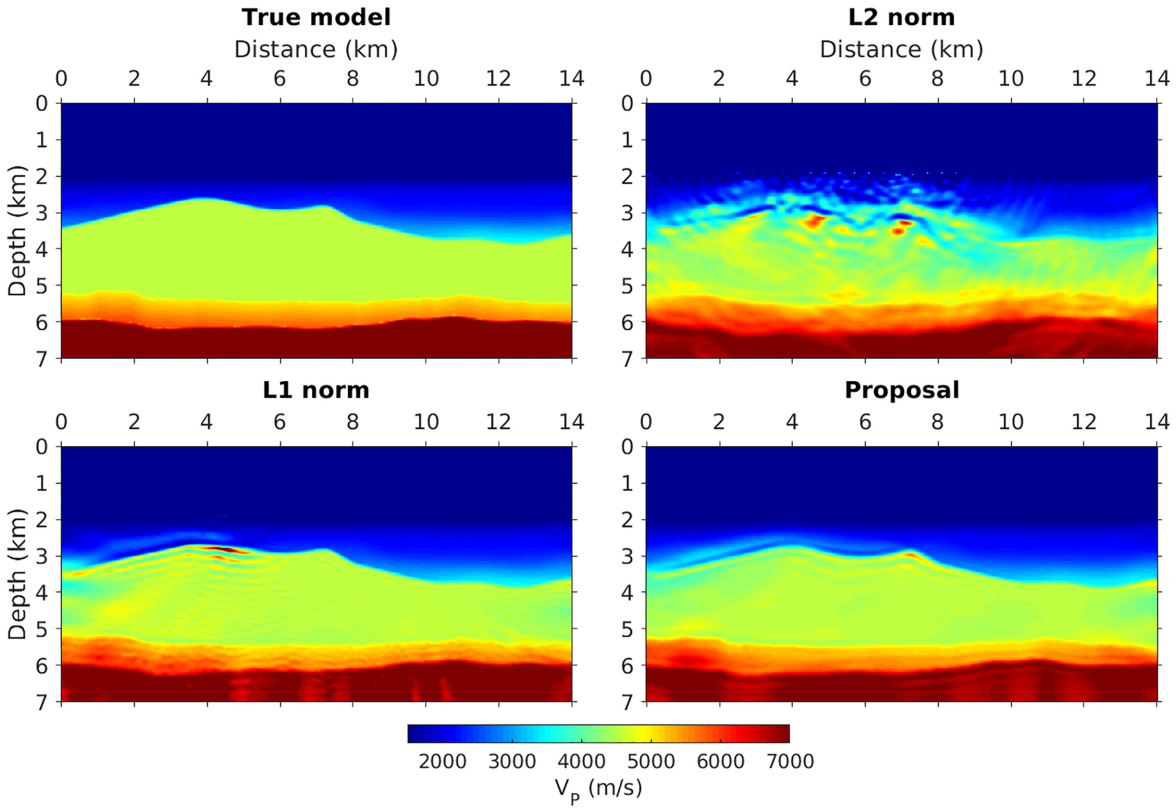


Figure 13. True velocity model and FWI results obtained by the conventional L^2- and L^1-norms and the proposed FWI for the case where the 'observed' data have no noise.

of using 'optimal' bandwidths, in particular the use of local bandwidths (which enable a more adequate estimation of the residuals probability distribution), so that the proposed objective function is neither too smooth leading to slow convergences to the global minimum nor too sinuous with more and more pronounced local minima.

With a view to investigate the influence of noise in the 'observed' data on the shape of the objective functions, we repeated the previous

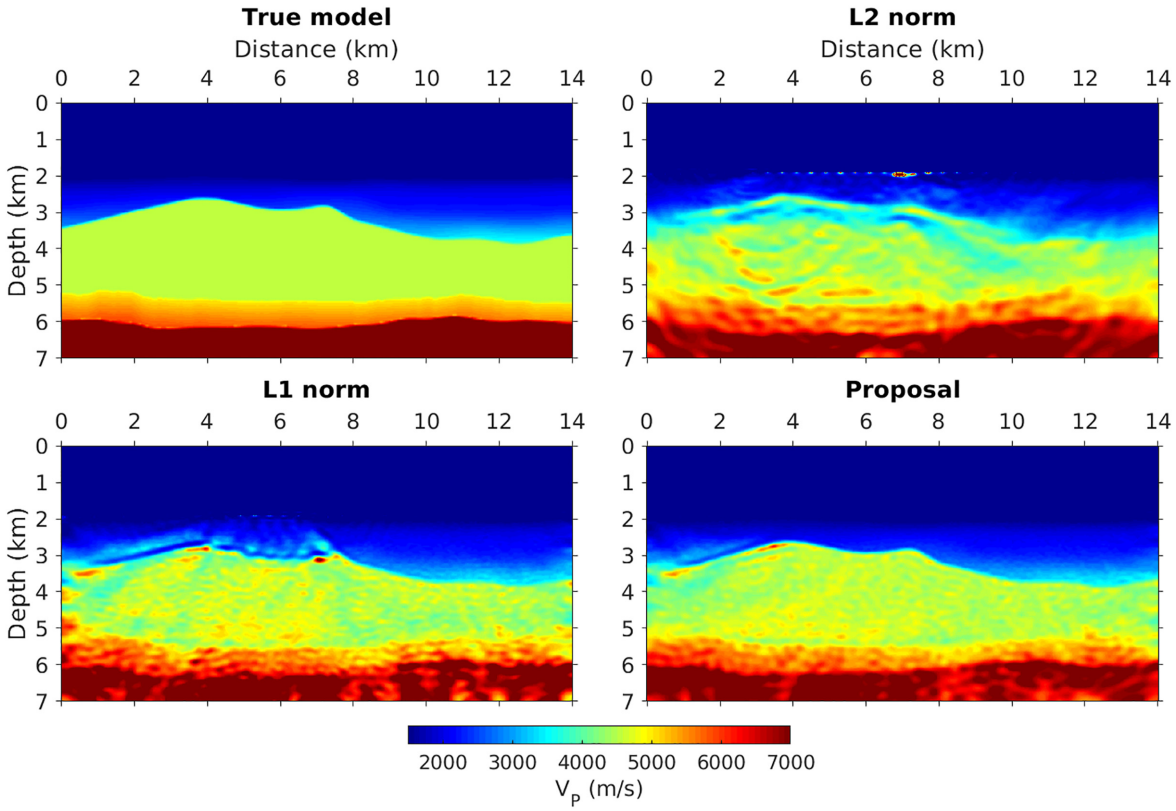


Figure 14. True velocity model and FWI results obtained by the conventional L^2- and L^1-norms and the proposed FWI for the noisy data case (SNR of about 7 dB).

experiment but this time with the 'observed' data contaminated with Gaussian noise [with an SNR (signal-to-noise ratio) of 6 dB]. As expected, the value of the objective functions increased (since now the residuals have an additional portion which corresponds to the noise), however, we found that the conventional objective functions undergo a smaller and more irregular increase (in particular the L^2-norm) than the proposed objective function (Fig. 3). The proposed objective function becomes smoother (particularly in the vicinity of the global minimum) both compared to the conventional objective functions and compared to the situation where the data have no noise. This reveals that the proposed objective function may not be as impaired as the conventional objective functions when the 'observed' data are noisy, as it tends to be smoother, which may also mean a smaller number of local minima.

Additionally, to evaluate the influence of a higher frequencies content on the shape of the objective functions, the same experiment was performed again but now a seismic source with a higher frequency content was assumed, that is, the seismic source was assumed to be a Ricker wavelet with a peak frequency of 20 Hz. The results show that, in general, as the content of higher frequencies increases, the objective functions become steeper, the valleys become narrower and deeper and consequently the local minima become more difficult to overcome (see particularly the conventional L^2-norm, Fig. 4), however, this is not so intense in the proposed objective function, which remains smoother than the conventional objective functions. These facts explain the increase in difficulties in conventional FWI when the content of higher frequencies increases and also indicating a possible better behaviour of the proposed objective function than conventional ones under these conditions.

3.2 Camembert model

Also with the aim of evaluating the behaviour of the proposed objective function in FWI, inversions were performed on a simple model similar to the Camembert model (Gauthier *et al.* 1986), a homogeneous model of 3000 m s^{-1} with a central circular anomaly of velocity 3600 m s^{-1} (Fig. 5). In order to the transition between the two velocities not to be so abrupt, a slight smoothing was carried out with a Gaussian filter with a standard deviation of 10 m.

In this experiment, the sources and receivers were placed in two lateral boreholes. The sources (20 sources) were placed in the well on the left at distance of $x = 20$ m and the receivers (39 receivers) were placed in the well on the right at distance of $x = 1970$ m. The sources and receivers were placed from the depth of 50 m to the depth of 1950 m, however, the sources are spaced every 100 m and the receivers are spaced 50 m apart. The source was assumed to be a Ricker wavelet with a peak frequency of 10 Hz and in order to simulate a situation closer to the real one, the frequencies below 2 Hz were filtered. The acquisition time was 2.0 s.

In the modelling of the wavefields, the acoustic approximation was used, a 10 m spatial discretization in both vertical and horizontal direction was considered and the C-PML boundary condition is used in all the surroundings of the models. In the FWI process, the quasi-Newton L-BFGS-B method was used as optimization algorithm.

Starting from homogeneous models of velocity 3000 and 2960 m s^{-1} and using the proposed objective function as well as the conventional L^2- and L^1-norms of the residuals, the velocity models presented in Fig. 6 were obtained. As can be seen in the various situations, the proposed objective function was able to reach reliable velocities close to the real velocities, particularly at the centre of the anomaly. The objective function based on the L^1-norm

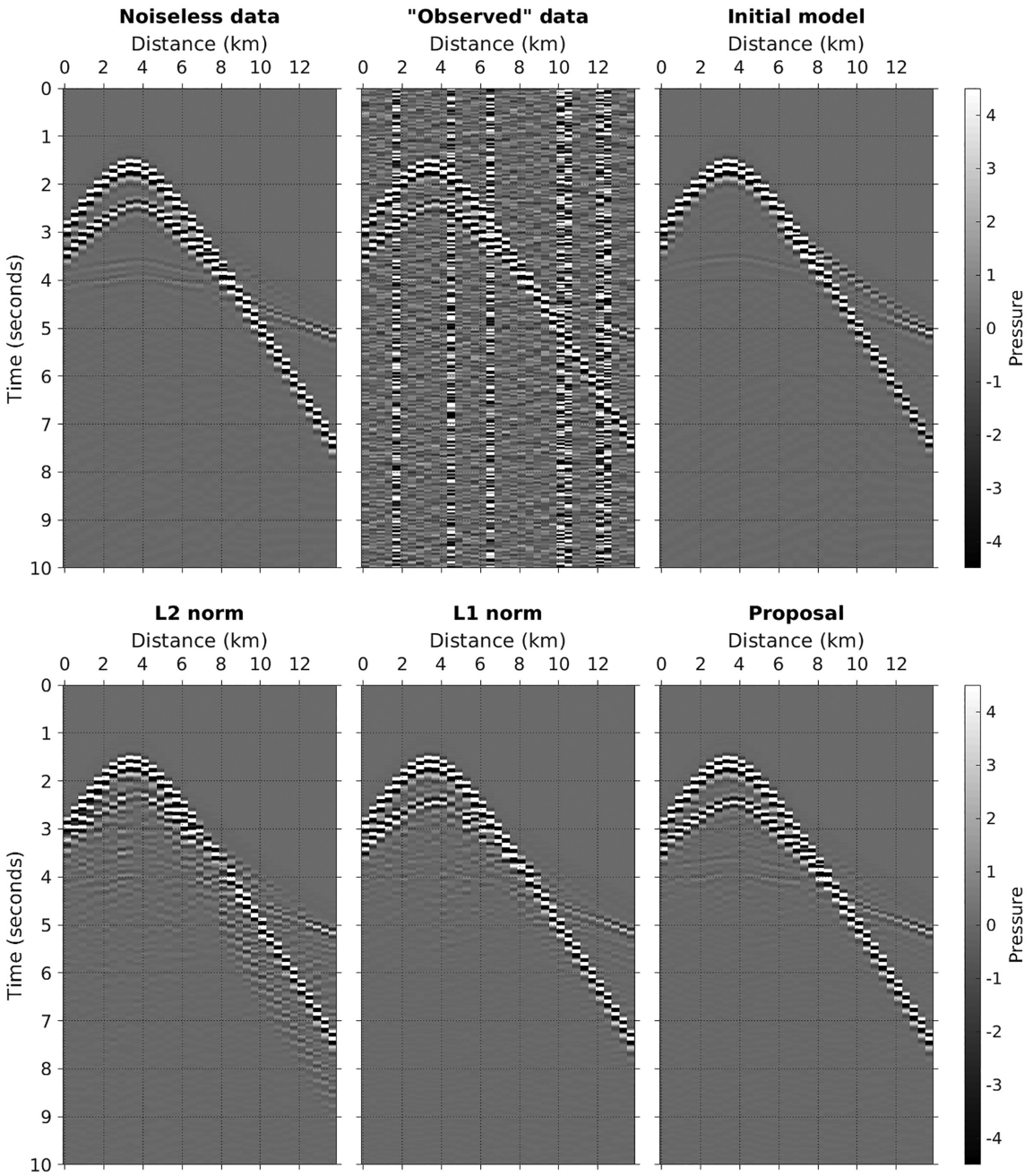


Figure 15. Seismograms corresponding to the source fired at distance of 3.4 km obtained in the true model (noiseless data), in the initial model and in the velocity models obtained by the conventional L^2- and L^1-norms and proposed FWI.

of the residuals still can identify the circular anomaly when one starts from an initial velocity model closer to the true model (from a homogeneous model of velocity 3000 $\mathrm{m\,s^{-1}}$), but the proposed objective function can achieve a velocity model closer to the real. However, when starting from a model further away from the true model, the conventional L^1-norm of the residuals can no longer able to reach an acceptable velocity model (Fig. 6). Adding some Gaussian noise to the 'observed' data, the proposed objective function continues to be able to identify the anomaly even if starting from a velocity model that is farther from the true one. This proves that the our proposal has the potential to overcome more adverse situations than conventional FWI.

In Fig. 7, as an example, are presented statistics of bandwidths over the FWI iterations for the source fired at 950 m deep when one starts from a homogeneous model of velocity 3000 $\mathrm{m\,s^{-1}}$. As can be seen, the bandwidths show a decreasing trend over the FWI iterations, as they are directly proportional to the standard deviation of the residuals and, therefore, as the FWI converges towards the true model, the residuals tend to become smaller as well as their standard deviations, and consequently also the bandwidths.

Analysing the data, it is noted that, in fact, cycle-skipping problems occur (Fig. 8). The initial models lead to arrival delays of more than half a period associated with the shortest wavelength

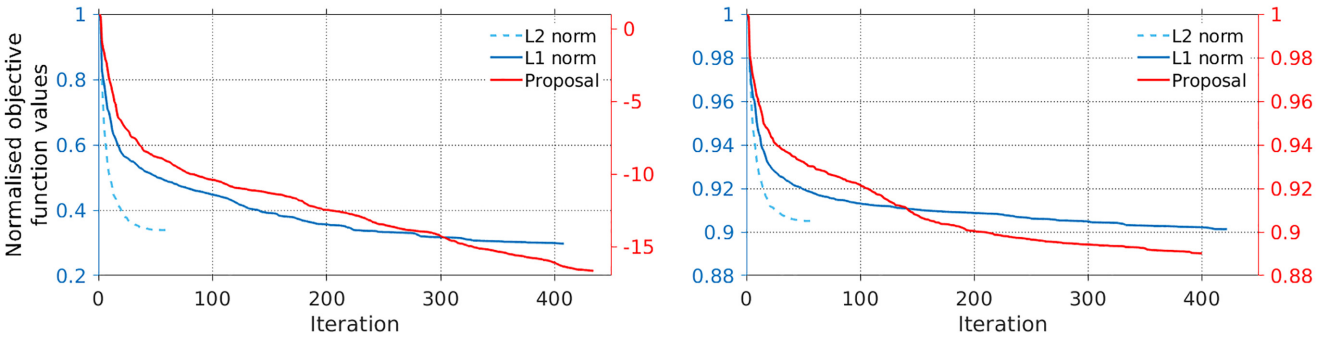


Figure 16. Evolution of the objective functions value (normalized by its maximum value) over the FWI iterations for the case where the 'observed' data have no noise (on the left) and for the noisy data case (on the right) (the blue vertical axis on the left corresponds to the conventional L^2- and L^1-norms and the red vertical axis on the right corresponds to the proposed FWI).

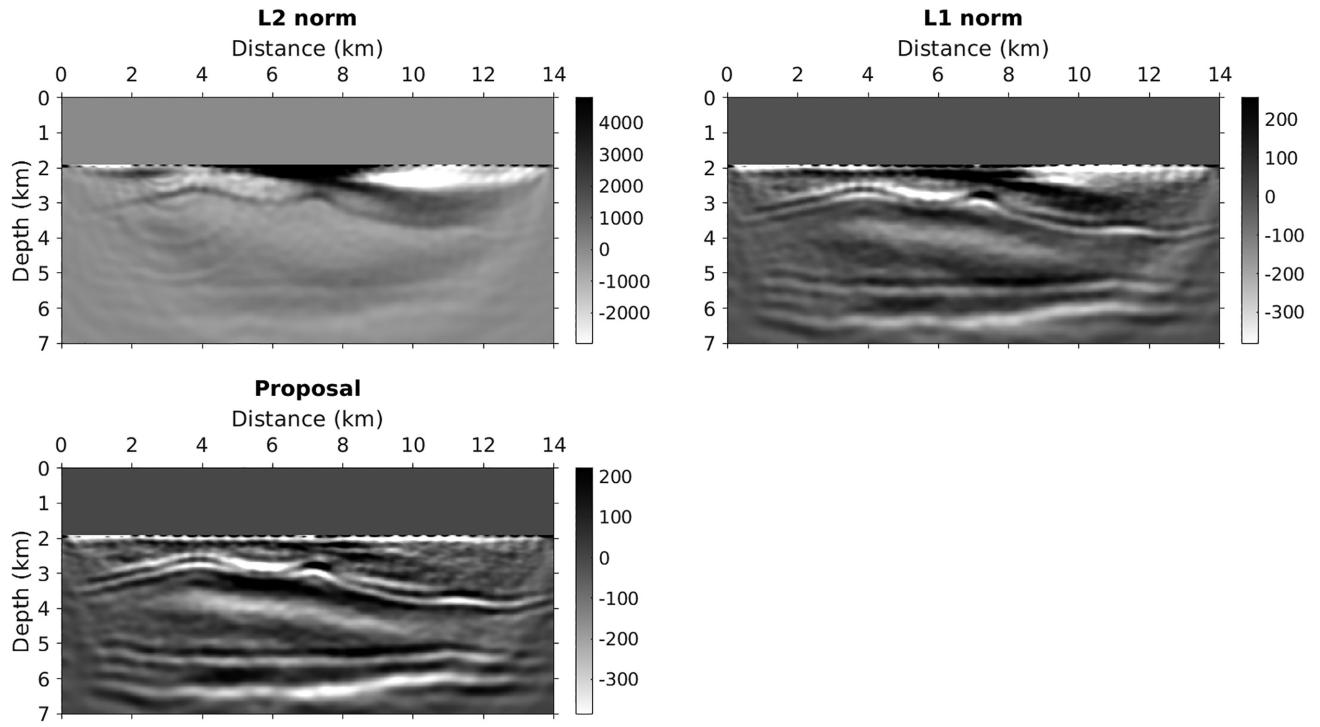


Figure 17. First gradients of the conventional L^2- and L^1-norms and proposed FWI for the case where the 'observed' data are noisy (the minimum and maximum of the colour scales correspond to the 0.01th and 0.99th quantiles of the components of the gradients, respectively).

(of the 'observed' data) in relation to the correct arrival, making the conventional objective functions, in particular the L^2-norm, find it difficult to converge to acceptable velocities. Note, for instance, the evolution of the modelled trace along the FWI iterations in the receiver at 950 m depth for the case of the L^2-norm objective function in Fig. 9 where instead of the velocities increase so that the modelled trace would get closer to the true trace, the velocities decreased. Conversely, the proposed objective function demonstrates in both situations the ability to overcome this difficulty and to converge towards the true trace (and the correct velocities) and furthermore with a more coherent and gradual evolution over the FWI iterations than conventional objective functions (Fig. 9).

Also analysing the adjoint sources, it is found that the adjoint sources corresponding to the proposed objective function have similarities with the L^1-norm adjoint sources, however, they are more regular (continuous), they are not limited to only values of −1 and 1 and have a higher content of lower frequencies, which are important to overcome cycle-skipping situations (Fig. 10).

3.3 Typical *P*-wave velocity model of the Brazilian pre-salt field

In order to demonstrate the potentialities of our proposal in a more realistic case, we applied it to a typical *P*-wave velocity model of the Brazilian pre-salt field (Fig. 11). The model consists of a layer of water of about 2 km followed by post-salt marine shales, salt, pre-salt reservoir and bed rock. The acquisition geometry considered was an ocean bottom node (OBN) where the receivers (35 receivers) are at the bottom of the sea between the depths of 1920 and 1950 m and between the horizontal distances 100 m and 13.7 km, equally

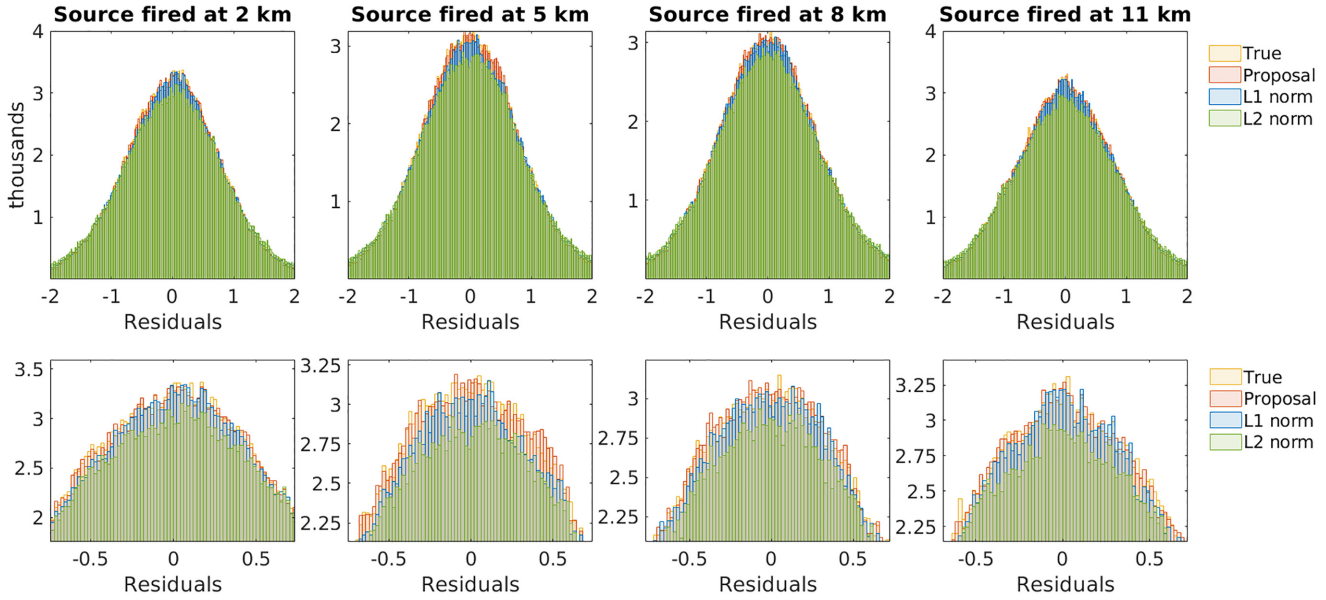


Figure 18. Histograms of the residuals corresponding to the velocity models obtained by the conventional L^2- and L^1-norm and proposed FWI (above) and its magnification (below) for the sources fired at 2, 5, 8 and 11 km. (The true residuals in this synthetic case study correspond to the difference between the data obtained in the true model (noiseless data) and the noisy data: $\Delta d = d^{\text{obs}}_{\text{noiseless}} - d^{\text{obs}}_{\text{noisy}}$).

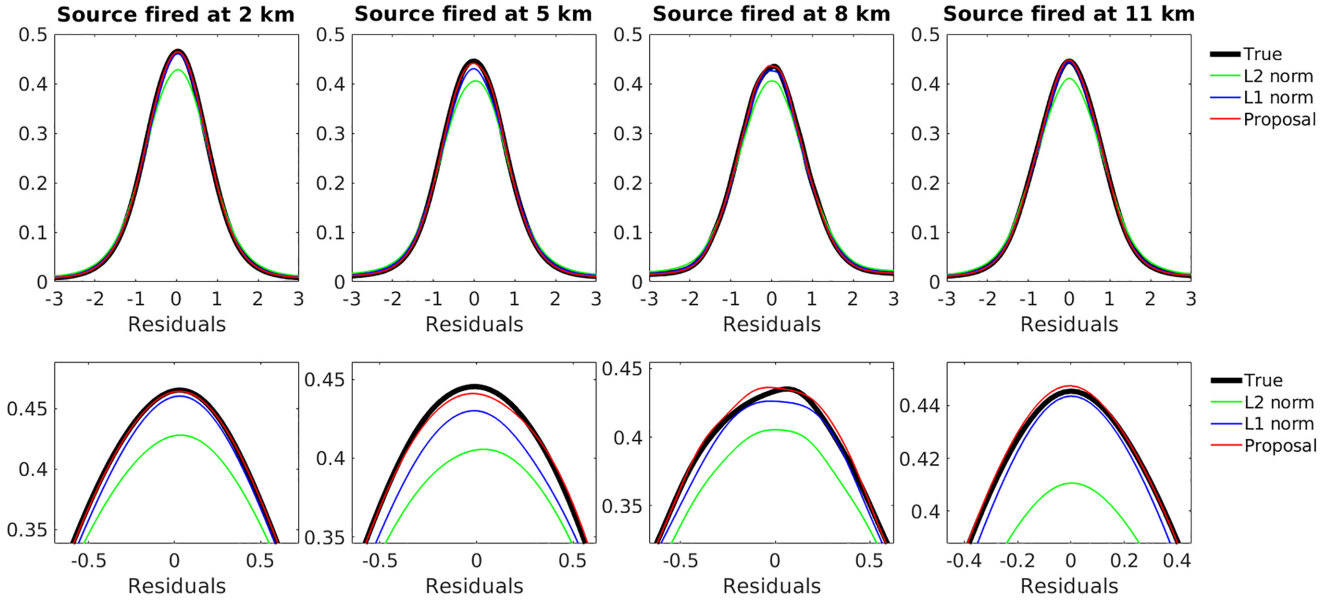


Figure 19. Probability distributions of the residuals (estimated by the KDE technique) corresponding to the velocity models obtained by the conventional L^2- and L^1-norms and proposed FWI (above) and its magnification (below) for the sources fired at 2, 5, 8 and 11 km.

spaced every 400 m. 69 sources were fired at a depth of 10 m between the distances 200 m and 13.8 km every 200 m. The seismic source was assumed to be a Ricker wavelet with a peak frequency of 5 Hz and in order to simulate a situation closer to the real, low frequencies below 1.5 Hz were filtered out. The total recording time is set to 10 s.

The 'observed' data have been generated synthetically from the true velocity model shown in Fig. 11 and were subsequently contaminated with Gaussian noise, and with the noise being amplified 5 times in 20 per cent of the seismic traces (randomly selected according to a uniform distribution) to simulate the presence of spikes (outliers), resulting in an 'observed' data with a median SNR of about 7 dB. Fig. 12 shows an example of a shot gather and seismic traces. In the generation of the 'observed' data as well as in the FWI process, the acoustic approximation was used. The velocity models were spatially discretized every 20 m, both in the vertical and horizontal directions, and as boundary conditions the C-PML absorbing boundary condition was used in all the surroundings of the models (since, in this case, no free-surface is assumed).

Starting from the initial model shown in Fig. 11 (at the right) (which corresponds to a smoothing of the true model by a Gaussian filter of a standard deviation of 400 m in the vertical direction and 800 m in the horizontal direction), the velocity models in Figs 13

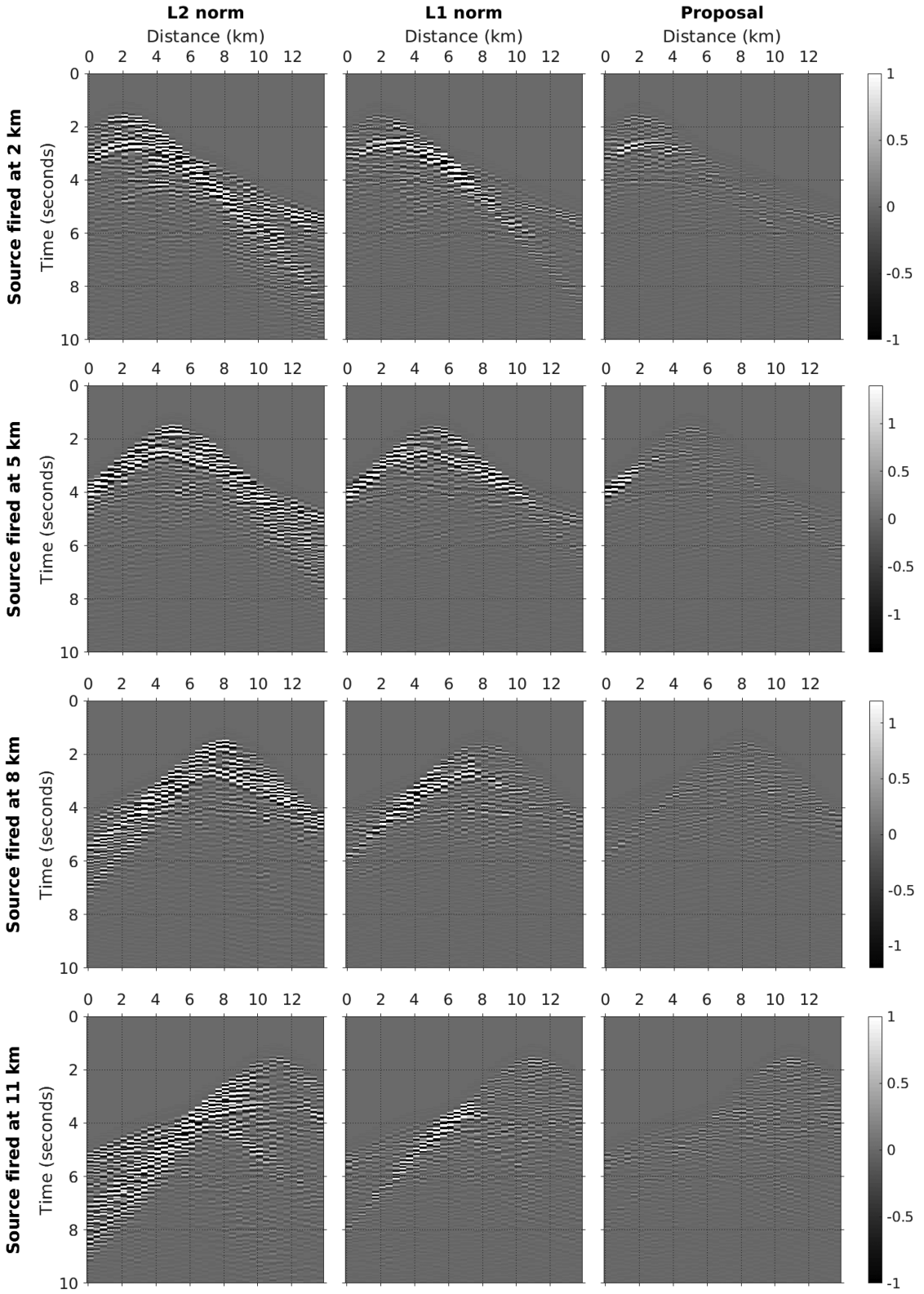


Figure 20. Differences between the seismograms corresponding to the velocity models obtained by the conventional and proposed objective functions and the seismograms corresponding to the true model (noiseless data) for the sources fired at 2, 5, 8 and 11 km.

and 14 were obtained by applying the proposed and conventional objective functions L^2- and L^1-norms of the residuals for the case where the 'observed' data have no noise and for the noisy data case, respectively.

In the FWI, the true velocities up to the depth of 1800 m were assumed to be known, where the water layer with a *P*-wave velocity of about 1500 m s^{-1} is located. The Laplace smoothing filter (Trinh *et al.* 2017) was also applied to the gradients of the objective

functions as a form of regularization. Regarding the optimization algorithm, the quasi-Newton L-BFGS-B method was used.

As can be seen, our proposal was able to achieve a velocity model closer to the real subsurface than the conventional objective functions (Figs 13 and 14). Apart from the fact that our proposal being able to identify the main interfaces between the different geological structures, it was also able to find velocities of the geological layers closer to the real ones. In the case where the 'observed' data are noisy, the velocity model reconstructed by the proposed FWI has, however, some artefacts (Fig. 14) but these are the result of the high level of noise in the 'observed' data. Note, for instance, the shot gather corresponding to the source fired at distance 3.4 km in Fig. 15, the proposed objective function was able to more reliably recover the reflections that occurred at the top of the salt structure as well as the diving waves.

Regarding the evolution of the objective function value over the FWI iterations, it is found that the proposed objective function presents a more irregular behaviour than the conventional L^2- and L^1-norm objective functions. Although, like conventional objective functions, the proposed objective function also presents a first stretch of faster convergence, unlike the conventional ones, it later presents several other stretches with different convergence rates (Fig. 16). It should also be noted that the proposed objective function initially (in the first stretch) tends to present a slower convergence than the conventional ones, although in the following stretches it acquires higher convergence rates and can reach and surpass conventional ones (Fig. 16, on the right). This is in agreement with the shape of the proposed objective function obtained in the previous experiment in Section 3.1, which has a smaller slope in the regions farthest from the global minimum, and only in the vicinity of the global minimum it become steeper (Fig. 1), which mean, therefore, a slower convergence when the model is still far from the true model. Note also, for instance, the case of the Camembert model when starting from the homogeneous model of velocity 3000 m s^{-1}, where the conventional L^1-norm showed a faster convergence than the proposed objective function, quickly approaching in relatively few iterations of the true seismic trace (Fig. 9a). These facts may thus indicate that the proposed objective function may exhibit a slower initial convergence than the conventional ones, in particular if the initial model is still far from the real one. It should also be noted that in this synthetic example of the Brazilian pre-salt field in which the observed data does not contain noise, the behaviour of the proposed objective function was already quite different (Fig. 16, on the left). Although it also had a faster convergence in the first stretch and a slower convergence in the following stretches, it always had a much faster convergence than the conventional objective functions (Fig. 16 on the left, note that the right and left vertical axes have different scales) and furthermore, unlike the conventional ones, it did not show a tendency to stabilize its value. However, it must be stressed that these cases in which the observed data are noise-free and the modelling corresponds exactly to the real wave propagation in the medium are actually quite rare.

In Fig. 17, the first gradients of the conventional and the proposed FWI are also presented. As can be seen, the main interfaces are much sharper and better defined in the first gradient of the proposed FWI than in the first gradients of the conventional FWI. The L^2 gradient is essentially dominated by updates closer to the surface and where the deeper layers are almost imperceptible (Fig. 17, at top left). In the L^1 gradient, although the geological layers are already much more visible, this gradient continues to be dominated by shallow updates as the L^2 gradient (Fig. 17, at top right). In contrast, in the proposed FWI gradient, the magnitude of the gradient components is already more uniform throughout the model, also leading to a greater emphasis on deeper geological structures (Fig. 17, below).

Finally, it should be noted that our proposal, as expected, makes it possible to reach probability distribution of the residuals closest to the true probability distributions (which in the present case are known because it is a synthetic case, Figs 18 and 19), which means, therefore, that our proposal was able to achieve the modelled data closest to the 'true observed' data (noiseless data) than conventional objective functions L^2- and L^1-norms of the residuals (as can be seen from Fig. 20).

It is also noteworthy that, in terms of computational cost, the proposed objective function does not represent a significant computational cost when compared to the conventional FWI. For instance, in this more realistic example, where the velocity model has a spatial discretization of 701 points in the horizontal direction and 351 points in depth and temporal discretization of 10001 points per seismic trace, in our implementation where we performed a resampling according to Nyquist sampling theorem for a more efficient computation of the gradient (e.g. Yang *et al.* 2016), the proposed objective function only has a computational cost of about 5 per cent higher than the conventional FWI (L^2-norm).

4 CONCLUSIONS

In contrast to previous works in the literature, in this work, a specific probability distribution for the residuals is not imposed, but instead, it is proposed to use the non-parametric KDE technique to explore the probability distribution that may be most suitable. And, thus, to avoid that FWI is forced to converge to an incorrect probability distribution different from the true probability distribution of the residuals and consequently preventing FWI from reaching biased subsurface models.

As the results obtained in our experiments demonstrate, our proposal has a greater potential to overcome adverse situations, such as, for instance, situations where the initial model is far away from the real subsurface and the observed data contain a relevant noise level and, consequently, to achieve subsurface models closer to the real ones. This greater potential can also be justified by its smoother shape and with fewer and less pronounced local minima than conventional objective functions.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from Shell Brasil through the New Methods for Full-Waveform Inversion project at Federal University of Rio Grande do Norte (UFRN) and the strategic importance of the support given by Brazilian National Agency for Petroleum, Natural Gas and Biofuels (ANP) through the R&D levy regulation. The authors are also thankful to the High-Performance Computing Center at UFRN (NPAD/UFRN) for making computer resources available. The authors also would like to thank Jorge L. Lopez from Shell Brasil for careful reading, comments and suggestions which helped to improve this article.

DATA AVAILABILITY

The data underlying this paper cannot be shared publicly due to confidential reasons. The data maybe shared on reasonable request to the corresponding author.

REFERENCES

Abramson, I.S., 1982. On bandwidth variation in kernel estimates—a square root law, *Ann. Stat.,* **10**(4), 1217–1223.

Amundsen, L., 1991. Comparison of the least-squares criterion and the cauchy criterion in frequency-wavenumber inversion, *Geophysics,* **56**(12), 2027–2035.

Aravkin, A., van Leeuwen, T. & Herrmann, F., 2011. Robust full-waveform inversion using the student's t-distribution, in *SEG Technical Program Expanded Abstracts 2011,* pp. 2669–2673. Society of Exploration Geophysicists.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning,* Information Science and Statistics.

Brossier, R., Operto, S. & Virieux, J., 2009. Seismic imaging of complex onshore structures by 2d elastic frequency-domain full-waveform inversion, *Geophysics,* **74**(6), WCC105–WCC118.

Brossier, R., Operto, S. & Virieux, J., 2010. Which data residual norm for robust elastic frequency domain full waveform inversion? *Geophysics,* **75**(3), R37–R46.

Bube, K.P. & Langan, R.T., 1997. Hybrid L1 / L2 minimization with applications to tomography, *Geophysics,* **62,** 1183–1195.

Bunks, C., Saleck, F.M., Zaleski, S. & Chavent, G., 1995. Multiscale seismic waveform inversion, *Geophysics,* **60,** 1457–1473.

Byrd, R., Lu, P., Nocedal, J. & Zhu, C., 1995. A limited memory algorithm for bound constrained optimization, *SIAM J. Sci. Comput.,* **16**(5), 1190–1208.

Chen, Y.-C., 2017. A tutorial on kernel density estimation and recent advances, *Biostat. Epidemiol.,* **1**(1), 161–187.

Constable, C.G., 1988. Parameter estimation in non-gaussian noise, *Geophys. J. Int.,* **94**(1), 131–142.

Crase, E., Pica, A., Noble, M., McDonald, J. & Tarantola, A., 1990. Robust elastic nonlinear waveform inversion: application to real data, *Geophysics,* **55,** 527–538.

da Silva, S.L.E.F., Carvalho, P.T.C., de Araújo, J.M. & Corso, G., 2020a. Full-waveform inversion based on kaniadakis statistics, *Phys. Rev. E,* **101**(5), 053311. https://doi.org/10.1103/PhysRevE.101.053311.

da Silva, S.L.E.F., Costa, C.A.N., Carvalho, P.T.C., de Araújo, J.M., Lucena, L. & Corso, G., 2020b. Robust full-waveform inversion using q-statistics, *Phys. A,* **548,** 124473. https://doi.org/10.1016/j.physa.2020.124473.

Fan, J. & Yao, Q., 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods,* Springer, New York.

Fei, Y., Rong, G., Wang, B. & Wang, W., 2014. Parallel L-BFGS-B algorithm on GPU, *Comput. Graph.,* **40,** 1–9.

Fichtner, A., 2011. *Full Seismic Waveform Modelling and Inversion,* Springer-Verlag.

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

Guitton, A. & Symes, W.W., 2003. Robust inversion of seismic data using the huber norm, *Geophysics,* **68,** 1310–1319.

Hansen, B.E., 2009. *Lecture Notes on Nonparametrics,* University of Wisconsin.

Hart, J., 1997. *Nonparametric Smoothing and Lack-of-Fit Tests,* Springer-Verlag, New York.

Huber, P.J., 1973. Robust regression: asymptotics, conjectures, and monte carlo, *Ann. Stat.,* **1,** 799–821.

Kaniadakis, G., 2001. Non-linear kinetics underlying generalized statistics, *Phys. A,* **296,** 405–425.

Komatitsch, D. & Martin, R., 2007. An unsplit convolutional perfectly matched layer improved at grazing incidence for the seismic wave equation, *Geophysics,* **72**(5), 155–167.

Lailly, P., 1983. The seismic inverse problem as a sequence of before stack migration, in Bednar, J.B., ed. *Conference on Inverse Scattering: Theory and Application,* pp. 206–220, SIAM.

Li, Q. & Racine, J., 2007. *Nonparametric Econometrics: Theory and Practice,* Princeton University Press, Princeton, NJ.

Li, Y.E. & Demanet, L., 2016. Full-waveform inversion with extrapolated low-frequency data, *Geophysics,* **81**(6), R339–R348.

Liu, D.C. & Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization, *Math Program: Ser. A and B,* **45,** 503–528.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. & Virieux, J., 2016. Measuring the misfit between seismogramsusing an optimal transport distance: application to full waveform inversion, *Geophys. J. Int.,* **205**(1), 345–377.

Mulder, W. & Plessix, R.E., 2008. Exploring some issues in acoustic full waveform inversion, *Geophys. Prospect.,* **56,** 827–841.

Parzen, E., 1962. On estimation of a probability density function and mode, *Ann. Math. Stat.,* **33**(3), 1065–1076.

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

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function, *Ann. Math. Stat.,* **27**(13), 832–837.

Scott, D.W., 1992. *Multivariate Density Estimation : Theory, Practice, and Visualization,* Wiley, New York.

Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis,* Chapman & Hall, London.

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

Tarantola, A., 1987. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation,* Elsevier Scientific Publ. Co., Inc.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* SIAM.

Tejero, C.E.J., Dagnino, D., Sallarès, V. & Ranero, C.R., 2015. Comparative study of objective functions to overcome noise and bandwidth limitations in full waveform inversion, *Geophys. J. Int.,* **203**(1), 632–645.

Trinh, P.T., Brossier, R., Métivier, L., Virieux, J. & Wellington, P., 2017. Bessel smoothing filter for spectral-element mesh, *Geophys. J. Int.,* **209**(3), 1489–1512.

Tsallis, C., 1988. Possible generalization of Boltzmann-Gibbs statistics, *J. Stat. Phys.,* **52,** 479–487.

Van Kerm, P., 2003. Adaptive kernel density estimation, *Stata J.,* **3,** 148–156.

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

Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A. & Zhou, W., 2014. 6. an introduction to full waveform inversion, in *Encyclopedia of Exploration Geophysics,* pp. R1–1-R1-40, Society of Exploration Geophysicists.

Xue, Z., Alger, N. & Fomel, S., 2016. Full-waveform inversion using smoothing kernels, in *SEG Technical Program Expanded Abstracts 2016,* pp. 1358–1363. Society of Exploration Geophysicists.

Yang, P., Gao, J. & Wang, B., 2015. A graphics processing unit implementation of time-domain full-waveform inversion, *Geophysics,* **80**(3), F31–F39.

Yang, P., Brossier, R. & Virieux, J., 2016. Wavefield reconstruction by interpolating significantly decimated boundaries, *Geophysics,* **81**(5), T197–T209.

Yuan, S. & Wang, S., 2013. Full waveform inversion using non-smooth data fidelity and non-smooth regularization, *Can. J. Explor. Geophys.,* **38**(1), 4–11.

Zhu, C., Byrd, R., Lu, P. & Nocedal, J., 1997. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization, *ACM T. Math. Softw.,* **23**(4), 550–560.