Solving nonlinear inverse problems by linearized model extension - a survey of possible methods

Biondo Biondi, Rahul Sarkar, and Joseph Jennings

ABSTRACT

Nonlinear inverse problems are challenging because gradient-based inverse algorithms may converge toward local minima instead of the desired global minimum. Different methods can be used to solve nonlinear inverse problems based on a linearized model extension; these methods differ in their global-convergence characteristics and in their convergence rate. We present and analyze, both analytically and numerically, three types of such methods. All three methods show attractive global convergence properties. However, our analysis is both incomplete and based on a simple 1D wave-propagation problem where the medium is characterized by a single slowness value. We discuss the convergence rate of the three types of solution we proposed, but, at the current stage of our research, we cannot reach any definitive conclusions on their convergence rate.

INTRODUCTION

Full waveform inversion (FWI) has a well-known convergence problem when the starting velocity model is far from the correct one and low frequency are not present in the data, or they are too noisy. One of the most promising direction of research for overcoming this problem is based on an extension of the velocity model. The first proposed solutions were based on extension of the reflectivity model (e.g. migrated image); we will refer to all of these methods, somewhat inappropriately, as wave-equation migration velocity analysis (WEMVA) methods (Symes and Carazzone, 1991; Biondi and Sava, 1999; Sava and Biondi, 2004; Shen and Symes, 2008; Zhang and Biondi, 2013). More recently, Symes (2008) and Biondi and Almomin (2014) have proposed extensions of the whole velocity model; that is, of long wavelength as well as short wavelength. These methods have been successful to converge to good models when applied to field data; however, their convergence might be too slow for being directly applicable to large scale problems (Almomin and Biondi, 2014). With another rough, but convenient, generalization we will refer two these methods as tomographic waveform inversion (TFWI) methods.

The main goals of our project are: 1) to develop faster-converging algorithms to apply to the extended FWI methods (e.g. TFWI), and 2) to explore the applicability of the idea of a linearized extension to the solution of other important nonlinear inverse problems in geophysics. In the first section of this report we formalize the idea of applying a linearized model extension to a generic non-quadratic optimization problem.

The second section presents the modeling equation of a simple 1D waveform inversion problem. We assume the data to be a single trace recorded from a transmission experiment between one source and one receiver in an homogeneous medium, with the slowness of the medium being the only parameter to estimate. We also introduce a useful linear extension of the nonlinear modeling operator that will be used to test inverse methods.

In the last sections of this report we present and analyze three ways of solving the nonlinear inverse problem by using a linearized model extension. The first way is based on the optimization of a two-term objective function, and it is related to the TFWI methods proposed by Symes (2008) and Biondi and Almomin (2014); The first term measures data fitting whereas the second one is a regularization (model focusing) term.

The second approach is based on alternating between the optimization of two objective functions. It is conceptually related to WEMVA methods as presented by Biondi and Sava (1999); Sava and Biondi (2004); Zhang and Biondi (2013) (e.g. migration and velocity model updating), where the velocity updating is driven by the matching of the current image to a better focused image.

The third proposed methods minimizes an objective function with a single term. This single term depends on the slowness model through two modeling operators: the original nonlinear operator and the extended linear operator. Our analysis shows that this one-term objective function has some characteristics of both the FWI and the WEMVA objective functions; we will refer to it as the FWI-WEMVA objective function. A version of this objective function was first presented by Symes (2008). In that paper this objective function was used as the basis for the formalization of the velocity analysis problem as a constrained optimization problem, rather than to be directly minimized.

SOLVING NON-QUADRATIC OPTIMIZATION PROBLEMS BY LINEARIZED MODEL EXTENSION

As mentioned above, one of our goals is to explore the applicability of the idea of a linearized extension to the solution of other important nonlinear inverse problems in seismology. Therefore, we first formalize the inversion problem in general terms, and then we study several inversion approaches by analyzing their application to a specific 1D waveform inversion problem.

We want to estimate the vector of model parameters, $\mathbf{m}$, from the recorded data vector, $\mathbf{d}_r$, recorded as the output of an operator $\mathcal{L}$ that is non-linear with respect to the model parameters; that is,

$$\mathbf{d}_r = \mathcal{L}\left(\overline{\mathbf{m}}\right), \tag{1}$$

where $\overline{\mathbf{m}}$ is the "true" value of the parameter vector, and it is the ideal solution of the estimation problem. We can set up the estimation as the least-squares problem of minimizing

$$J(\mathbf{m}) = \frac{1}{2} \left\| \mathcal{L}(\mathbf{m}) - \mathbf{d}_r \right\|_2^2 . \tag{2}$$

Because of the non-linear dependency between the modeled data, $\mathbf{d} = \mathcal{L}(\mathbf{m})$ and the parameter vector $\mathbf{m}$, the objective function in equation 2 is not quadratic, and in general, is not even convex, and presents many local minima. Therefore, when we apply gradient-based methods to solve the optimization problem in 2 we are likely to converge towards a local minimum, instead of the desired global one ($\overline{\mathbf{m}}$).

We are interested in improving the convergence towards the global minimum by solving a different optimization problem that shares the global minimum with the one expressed in 2, but does not have local minima, or at least, can be "safely driven" to converge to the global minimum starting from an arbitrary starting solution $\mathbf{m}_0$. We start by extending the non-linear operator by adding to the modeled data ($\mathbf{d}$) the output, ${}_{\mathrm{L}}\mathbf{d}$, of an appropriately defined linear operator, $\tilde{\mathbf{L}}$, applied to an additional parameter vector ${}_{\mathrm{L}}\mathbf{m}$. The"total" modeled data vector is thus expressed as,

$${}_{\mathrm{T}}\mathbf{d} = \tilde{\mathcal{L}}(\mathbf{m}, {}_{\mathrm{L}}\mathbf{m}) = \mathcal{L}(\mathbf{m}) + \tilde{\mathbf{L}}(\mathbf{m})\, {}_{\mathrm{L}}\mathbf{m} = \mathbf{d} + {}_{\mathrm{L}}\mathbf{d}, \tag{3}$$

where $\tilde{\mathcal{L}}$ is the "extended" modeling operator that is function of both the original parameter vector, $\mathbf{m}$, and the *extended model* parameter vector, ${}_{\mathrm{L}}\mathbf{m}$. Notice that the linear operator $\tilde{\mathbf{L}}$ is itself a non-linear function of the parameter vector $\mathbf{m}$.

The optimization problem in 2 can be modified into the following

$$J_{\mathrm{E}}(\mathbf{m}, {}_{\mathrm{L}}\mathbf{m}) = \frac{1}{2} \left\| \tilde{\mathcal{L}}(\mathbf{m}, {}_{\mathrm{L}}\mathbf{m}) - \mathbf{d}_r \right\|_2^2 , \tag{4}$$

that obviously has a global minimum for $(\mathbf{m} = \overline{\mathbf{m}}, {}_{\mathrm{L}}\mathbf{m} = 0)$. In a more compact notation, if we combine the two model vectors into one, we can write

$$J_{\mathrm{E}}({}_{\mathrm{T}}\mathbf{m}) = \frac{1}{2} \left\| \tilde{\mathcal{L}}({}_{\mathrm{T}}\mathbf{m}) - \mathbf{d}_r \right\|_2^2 , \tag{5}$$

where ${}_{\mathrm{T}}\mathbf{m} = (\mathbf{m}, {}_{\mathrm{L}}\mathbf{m})$ and ${}_{\mathrm{T}}\overline{\mathbf{m}} = (\mathbf{m} = \overline{\mathbf{m}}, {}_{\mathrm{L}}\mathbf{m} = 0)$.

The extension operator $\tilde{\mathbf{L}}$ plays an important role in the method, and it should be defined according to the specific non-linear behavior of $\mathcal{L}(\mathbf{m})$ that prevents convergence to the global minimum of J in the practical problems that we want to address. Ideally, $\frac{d\tilde{\mathcal{L}}}{d_{\mathrm{T}}\mathbf{m}}$ is close to be a unitary operator, or at least the following approximation is valid:

$$\begin{aligned} \frac{d\tilde{\mathcal{L}}}{d_{\mathrm{T}}\mathbf{m}}(\mathbf{m}) \frac{d\tilde{\mathcal{L}}}{d_{\mathrm{T}}\mathbf{m}}'(\mathbf{m}) \left[\mathcal{L}(\mathbf{m}) - \mathbf{d}_r\right] &= \\ \frac{d\tilde{\mathcal{L}}}{d_{\mathrm{T}}\mathbf{m}}(\mathbf{m}) \frac{d\tilde{\mathcal{L}}}{d_{\mathrm{T}}\mathbf{m}}'(\mathbf{m}) \left[\mathcal{L}(\mathbf{m}) - \mathcal{L}(\overline{\mathbf{m}})\right] &\approx \left[\mathcal{L}(\mathbf{m}) - \mathcal{L}(\overline{\mathbf{m}})\right] \ \forall\, \overline{\mathbf{m}} \text{ and } \mathbf{m}. \end{aligned} \tag{6}$$

In practice, there is not a single linear extension that fulfills the requirement expressed in 6 for all possible events present in typical reflection-seismic data, and for any possible error in starting velocity model. For example, Biondi and Almomin (2014) showed that a time-lag extension of the slowness model is capable of modeling the kinematics of the data residuals caused by long-wavelength errors in the velocity model. However, the same linear extension is not as effective when the nonlinear wave phenomena that hamper global convergence are related to multiple-scattering by discrete interfaces; e.g. multiples.

Focusing operator

The optimization problem in 4 (and 5) is under-constrained. For any choice of the extended model vector $\mathbf{m}$, it is likely that there is a corresponding value of ${}_{\mathrm{L}}\mathbf{m}$ that minimizes J_{E}. Furthermore, the extension of the modeling operator is only an end to achieve the goal of robustly converging toward $\overline{\mathbf{m}}$. At convergence, or even in a neighborhood of the global minimum, we would would like to have the contributions of the linearized extension to be negligible. To achieve this goal, we introduce a "focusing" $\mathcal{F}$ operator.

An essential property of the focusing operator is that its output is more focused than its input. If we define an optimally focused model vector $\overline{{}_{\mathrm{L}}\mathbf{m}}$ when all the energy is focused in at one particular model coordinate $\overline{\mathbf{x}_f}$, we can define a measure of the defocusing as

$$D\left({}_{\mathrm{L}}\mathbf{m}\right) = \left\|\mathbf{X}_f\,{}_{\mathrm{L}}\mathbf{m}\right\|_2^2, \tag{7}$$

where $\mathbf{X}_f$ is a diagonal matrix with the absolute value of the distance from $\overline{\mathbf{x}_f}$; that is, $\mathbf{X}_f = \mathbf{diag}(|\mathbf{x_f} - \overline{\mathbf{x_f}}|)$. By construction $D\left(\overline{{}_{\mathrm{L}}\mathbf{m}}\right) = 0$.

The focusing operator must have at least the property that

$$\begin{cases} \mathcal{F}\left({}_{\mathrm{L}}\mathbf{m}\right) = \overline{{}_{\mathrm{L}}\mathbf{m}} & \text{if } {}_{\mathrm{L}}\mathbf{m} = \overline{{}_{\mathrm{L}}\mathbf{m}} \\ D\left(\mathcal{F}\left({}_{\mathrm{L}}\mathbf{m}\right)\right) < D\left({}_{\mathrm{L}}\mathbf{m}\right) & \text{if } {}_{\mathrm{L}}\mathbf{m} \neq \overline{{}_{\mathrm{L}}\mathbf{m}}. \end{cases} \tag{8}$$

In the following sections we discuss how the focusing operator can be introduced in a regularization term to be added the data-fitting term of the objective function, or how we can directly introduce it into the data-fitting term of the objective function. Depending on the way that the focusing operator is introduced in the objective function(s), and on the choice of $\tilde{\mathbf{L}}$, other properties, in addition to 8, of $\mathcal{F}$ are required for defining an estimation method that robustly converges towards the global minimum.

SIMPLE 1D WAVEFORM INVERSION PROBLEM

In the following sections we present several approaches to solve the estimation problem. We analyze the behavior of these methods using a simple 1D "wavelet-shift" modeling operator. In a homogeneous medium with slowness $\bar{s}$, the wavefield generated by one source and recorded by one receiver at a distance l is equal, when we ignore an amplitude scaling, to the source wavelet time-shifted by $\Delta t = ls$. The recorded-data vector $\mathbf{d}_r \in \Re^{N_t}$ is expressed as follows:

$$\mathbf{d}_r = \mathcal{L}\left(\bar{s}\right) = \mathbf{S}\left(l\bar{s}\right)\mathbf{g}, \tag{9}$$

where $\mathbf{g} \in \Re^{N_g}$ is the source vector, and $\mathbf{S}\left(\Delta t\right) \in \Re^{N_t \mathrm{x} N_g}$ is a time-shift operator that shifts an input vector by the time interval Δt.

We define the linearized extension of modeling as the time convolution of the shifted wavelet with a filter c; that is as,

$$_{\mathrm{L}}d\left(t\right) = g\left(t - ls\right) *_\tau c\left(\tau\right), \tag{10}$$

where with the symbol $*_\tau$ we denote convolution with the filter $c\left(\tau\right)$ along the time-lag axis τ. For matter of convenience, we fix the value of c at the origin to be equal to zero; that is, $c\left(\tau = 0\right) = 0$.

The data vector, $_{\mathrm{L}}\mathbf{d}$, produced by the linearized extension can be expressed by the following matrix-vector product

$$_{\mathrm{L}}\mathbf{d} = \tilde{\mathbf{L}}\left(s\right)\mathbf{c}, \tag{11}$$

where $\tilde{\mathbf{L}}\left(s\right) \in \Re^{N_t \mathrm{x} N_\tau}$ is a matrix appropriately constructed with the elements of the shifted source vector $\mathbf{S}\left(l\bar{s}\right)\mathbf{g}$, and $\mathbf{c}$ is a vector of length N_τ representing the discretization of convolutional filter c. The total modeled data $_{\mathrm{T}}\mathbf{d}$ can thus be expressed as follows:

$$_{\mathrm{T}}\mathbf{d} = \tilde{\mathcal{L}}\left(s, \mathbf{c}\right) = \mathcal{L}\left(s\right) + \tilde{\mathbf{L}}\left(s\right)\mathbf{c} = \mathbf{S}\left(ls\right)\mathbf{g} + \tilde{\mathbf{L}}\left(s\right)\mathbf{c}. \tag{12}$$

As mentioned before, we set the zero-lag coefficient of $\mathbf{c}$ to zero; that is, $c_0 = 0$. For convenience, we also assume that $\mathbf{c}$ is centered around the origin; that is, $c_i \neq 0$ for $N_\tau/2 \leq i \leq N_\tau/2$ with N_τ even.

Notice that our choice of $\tilde{\mathbf{L}}$ is, in purpose, not the most obvious one. The most natural choice would have been to define $\tilde{\mathbf{L}}\mathbf{c}$ as the convolution of $\mathbf{c}$ with the first time derivative of the shifted source wavelet; that is, to build $\tilde{\mathbf{L}}$ with elements of $-l\mathbf{S}\left(l\bar{s}\right)\dot{\mathbf{g}}$, instead of $\mathbf{S}\left(l\bar{s}\right)\mathbf{g}$ (Biondi and Almomin, 2014). In practice, there is a little differences between these two choices. The main reason we made this choice is to emphasize that the definition of $\tilde{\mathbf{L}}$ does not need to be based on physical arguments. The main criterion driving the definition of $\tilde{\mathbf{L}}$ should be to fulfill the approximation expressed in 6 as accurately as possible.

Linearization of modeling operators

When computing gradients of the proposed objective functions, we need to linearize the modeling operators with respect to perturbations of the model parameters. In this section we derive the linearization of the simple 1D modeling operator $\tilde{\mathcal{L}}(s, \mathbf{c})$ with respect to slowness s and the coefficients of the convolutional filter $\mathbf{c}$. Taking advantage of the fact that slowness is a scalar, the operators derived from these linearizations can also be expressed as vector and matrices.

The derivative of the operator $\mathcal{L}(s)$ with respect to slowness, that we will denote as the vector $\mathbf{L} \in \Re^{N_t}$, is given by the following:

$$\frac{\partial \mathcal{L}(s)}{\partial s} = \mathbf{L}(s) = -l\mathbf{S}(ls)\,\dot{\mathbf{g}}, \tag{13}$$

where $\dot{\mathbf{g}}$ is the discretization of the first time derivative of the source function $g(t)$.

The derivative of the extension operator $\tilde{\mathbf{L}}(s, \mathbf{c})$ with respect to slowness is given by the matrix-vector product of the matrix $\dot{\tilde{\mathbf{L}}}\in \Re^{N_t \mathrm{x} N_\tau}$ with the vector $\mathbf{c}$ as follows:

$$\frac{\partial \tilde{\mathbf{L}}}{\partial s}(s, \mathbf{c}) = \dot{\tilde{\mathbf{L}}}(s)\,\mathbf{c}. \tag{14}$$

As for the matrix $\tilde{\mathbf{L}}$ (equation 11), also the matrix $\dot{\tilde{\mathbf{L}}}$ can be appropriately constructed with the elements of the first time derivative of the shifted source vector $\mathbf{S}(ls)\,\dot{\mathbf{g}}$ scaled by $-l$.

Since ${}_{\mathrm{L}}\mathbf{d}$ is linear with respect to the convolutional filter $\mathbf{c}$, the linearization of the extension operator $\tilde{\mathbf{L}}(s, \mathbf{c})$ with respect to $\mathbf{c}$ is simply the matrix $\tilde{\mathbf{L}}$; that is,

$$\frac{\partial \tilde{\mathbf{L}}}{\partial \mathbf{c}}(s, \mathbf{c}) = \tilde{\mathbf{L}}. \tag{15}$$

Numerical example

We illustrate the properties of the operators defined above by using a numerical example. Figure 1 shows the results of this numerical example.

Figure 1a shows the wavelet that we used for this example. It was derived by taking the first time derivative of a Ricker wavelet with fundamental frequency of 7 Hz. For convenience, in the proceedings we will refer to it as the *Ricker-derived wavelet.* Figure 1b shows the data residuals $(\mathcal{L}(s_o) - \mathcal{L}(\bar{s}))$ for a range of starting slowness values (s_o), and a true slowness $\bar{s} = 1$ s/km. The source-receiver distance l is 4 km.

Figures 1c and 1d show the *back projection* of the data residuals into the model space. These back projections are an important component of any gradient-based estimation algorithm. Figure 1c shows the application of the adjoint of $\mathbf{L}(s_o)$ to the corresponding data residuals shown in Figure 1b. Similarly, Figure 1d shows the application of the adjoint of $\tilde{\mathbf{L}}(s_o)$ to the same data residuals shown in Figure 1b.

As discussed, above, an important characteristic of the extended modeling operator is that its linearization should be close to be unitary, in the sense defined by

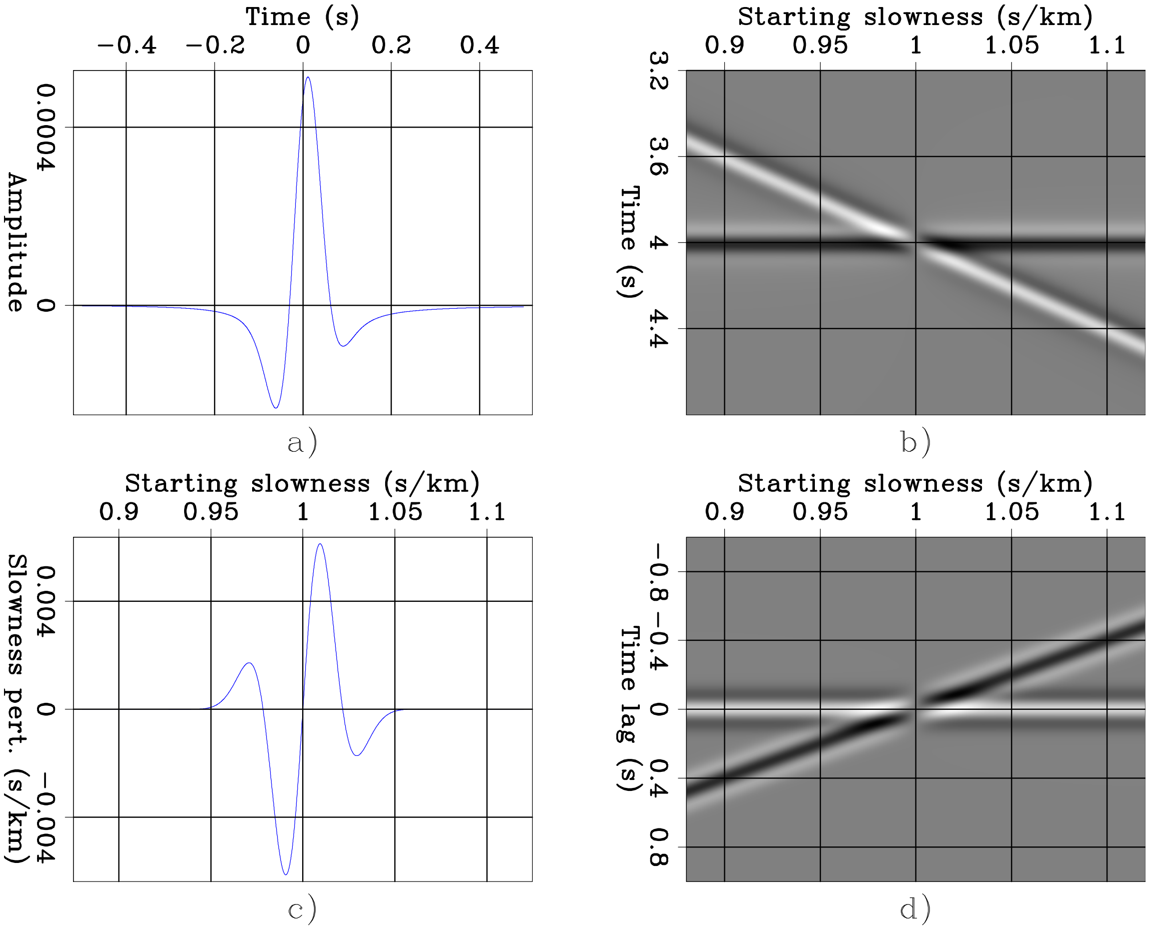


Figure 1: a) Ricker-derived wavelet used for the numerical examples. b) data residuals for a range of starting slowness values. c) back projection of the data residuals into slowness space for the same range of starting slowness values. d) back projection of the data residuals into the convolutional filter space. [**ER**]

the relation in equation 6. Figure 2 shows that the original (non extended) linearized operator is far from fulfilling that condition, whereas the linearized extension fulfills that condition, at least with regards to the kinematics of the reconstructed residuals.

Figure 2b shows the application of $\mathbf{L}\mathbf{L}'$ to the data residuals shown in Figure 1b, which for convenience of the reader are also displayed in Figure 2a. Figure 2b can also be described as the *forward projection* into the data space of the slowness perturbations shown in Figure 1c. The data residuals are well reconstructed only in a small interval of starting slowness centered around the true slowness ($\bar{s}$). Furthermore, at both ends of this range, the reconstructed residuals have the wrong polarity. In contrast, Figure 2c shows the application of $\tilde{\mathbf{L}}\tilde{\mathbf{L}}'$ to the data residuals. It can also be described as the forward projection into the data space of the convolutional filter perturbations shown in Figure 1d. The kinematics of the events in the data residuals are well reconstructed. The only noticeable difference between panel 2a and panel 2c is that the reconstructed events are convolved with the square of the source wavelet.

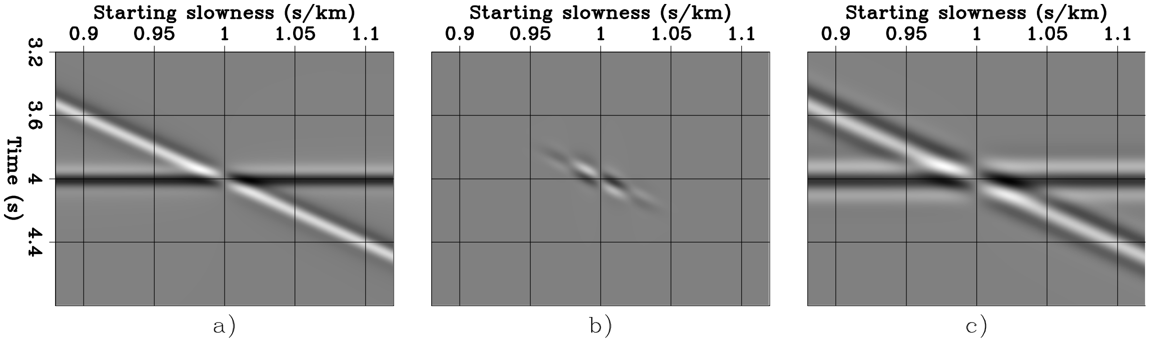


Figure 2: a) Data residuals $[\mathcal{L}(s_o) - \mathcal{L}(\bar{s})]$. b) $\mathbf{L}\mathbf{L}'[\mathcal{L}(s_o) - \mathcal{L}(\bar{s})]$. c) $\tilde{\mathbf{L}}\tilde{\mathbf{L}}'[\mathcal{L}(s_o) - \mathcal{L}(\bar{s})]$. **[ER]**

Focusing operators

We tested different focusing operators; all of them are linear with respect to the convolutional filter $\mathbf{c}$. Therefore, we will denote as $\mathbf{F}$ and write $\mathcal{F}(\mathbf{c}) = \mathbf{F}\mathbf{c}$. There are two families of focusing operators that can be useful. The operators belonging to the first family scale the filter coefficients as a function of the distance from the origin; that is, as a function of the time lag τ. We refer to the focusing operators belonging to this family as *amplitude* focusing operators because they simply scale the amplitudes of the input filter. The operators belonging to the second family "shift" the filter coefficients towards the origin. Therefore, we refer to these operators as *phase* focusing operators because they actually change the phase of the convolutional filter.

DSO (Differential Semblance Operator)

The first operator we analyze is derived from the classical DSO operator (Symes and Carazzone, 1991) and defined as $\mathbf{F}_D = \mathbf{diag}(\tau_f)$ with

$$\tau_f = \begin{cases} \frac{\tau_{\mathrm{w}}\tau_{\mathrm{max}}-|\tau|}{\tau_{\mathrm{w}}\tau_{\mathrm{max}}} & \text{if } |\tau| < \tau_{\mathrm{w}}\tau_{\mathrm{max}} \\ 0 & \text{if } |\tau| \geq \tau_{\mathrm{w}}\tau_{\mathrm{max}}, \end{cases} \tag{16}$$

where τ_{max} is the maximum τ represented in $\mathbf{c}$, and τ_{w} (with $0 \leq \tau_{\mathrm{w}} \leq 1$) is an adimensional parameter that determines the width of the triangular window; this parameter may change with iterations. Notice that the classical DSO operator is equal to $\mathbf{I} - \mathbf{F}_D$ with $\tau_{\mathrm{w}} = 1$.

Gaussian window

The second operator we analyze is also a simple amplitude focusing operator; it is defined as the truncated Gaussian window $\mathbf{F}_G = \mathbf{diag}(\tau_f)$, with

$$\tau_f = \begin{cases} e^{-5\frac{\tau^2}{(\tau_{\mathrm{w}}\tau_{\mathrm{max}})^2}} & \text{if } |\tau| < \tau_{\mathrm{W}}\tau_{\mathrm{max}} \\ 0 & \text{if } |\tau| \geq \tau_{\mathrm{W}}\tau_{\mathrm{max}}, \end{cases} \tag{17}$$

where τ_{max} and τ_{W} have the same meaning as in equation 16.

Shift

The third operator we analyze is the simplest phase focusing operator (Almomin, personal communication). The operator $\mathbf{F}_S$ shifts by one sample the coefficients of $\mathbf{c}$ towards the origin. The i-th coefficient $\hat{c}_i$ of the output filter are computed from the coefficients c_i of the input filter as:

$$\hat{c}_i = \begin{cases} c_{i+1} & \text{if } i > 0 \\ c_{i-1} & \text{if } i < 0. \\ 0 & \text{if } |i| = N_\tau/2 \text{ or } i = 0. \end{cases} \tag{18}$$

Shrink

The fourth, and last, operator we analyze is also a phase focusing operator; it scales the τ axis of its input filter by a factor α, that is $\hat{c}(\tau) = c(\alpha\tau)$, with $\alpha \geq 1$. The operator $\mathbf{F}_\alpha$ is the discrete implementation of this axis-shrinking operator that employs a sinc interpolator.

Examples of application of focusing operators

Figure 3 shows the outputs of these four focusing operators as applied to the convolutional filters shown in Figure 1d. Figures 3a, 3b, 3c, and 3d show the results of applying $\mathbf{F}_D$, $\mathbf{F}_G$, $\mathbf{F}_S$, and $\mathbf{F}_\alpha$, respectively. The output of $\mathbf{F}_S$ is indistinguishable from its input because $\mathbf{F}_S$ shifts its input by one sample only.

Figure 4 shows the plots of the defocusing measure D defined in equation 7 to the input and output $\mathbf{c}$ vectors shown in Figure 1d and Figure 3. It demonstrates with a numerical example that all four focusing operators fulfill the condition introduced in 8. For this example, we set $\tau_{\mathrm{W}} = 1.0$ for both $\mathbf{F}_D$ and $\mathbf{F}_G$, and $\alpha = 1.111$ for $\mathbf{F}_\alpha$. The functions are obviously symmetric around $s_o = 1$ s/km, and thus for clarity we plotted them only for $s_o \geq 1$ s/km. The values of $D\left(\mathbf{F}_S\mathbf{c}\right)$ are uniformly smaller than the values of $D\left(\mathbf{c}\right)$; however, their respective plots shown in Figure 4 almost perfectly overlap because their difference is tiny.

TFWI OBJECTIVE FUNCTIONS

Model-space regularization

As previously discussed, there are three different ways in which we constrain problem 4 using the focusing operator $\mathcal{F}$. One of these approaches is to use it in a model regularization term in addition to the data-fitting term defined in problem 4. For the simple 1D problem described above, we can write this objective function as:

$$J_{\mathrm{M}}\left(s,\mathbf{c}\right) = \frac{1}{2}\left\|\tilde{\mathcal{L}}\left(s,\mathbf{c}\right) - \mathbf{d}_r\right\|_2^2 + \frac{\epsilon}{2}\left\|\left(\mathbf{I} - \mathbf{F}\right)\mathbf{c}\right\|_2^2, \tag{19}$$

where $\mathbf{F}$ is the linear focusing operator. Through our numerical experiments, we have found that choosing $\mathbf{F}$ as the DSO operator provides the best convergence properties. This observation is related to the fact that the phase-focusing operators are nearly unitary and therefore, the result of their forward and adjoint applications in the computation of the gradients does not provide the correct focusing of the extended model space.

The results of running 7000 iterations of a non-linear conjugate gradient inversion are shown in Figure 5. For this inversion, the starting physical slowness (s_0) was 1.12 s/km and the starting extended slowness ($\mathbf{c}_0$) was $\mathbf{0}$. For the regularization, we set $\epsilon = 100$ and $\mathbf{F}$ to be the DSO operator.

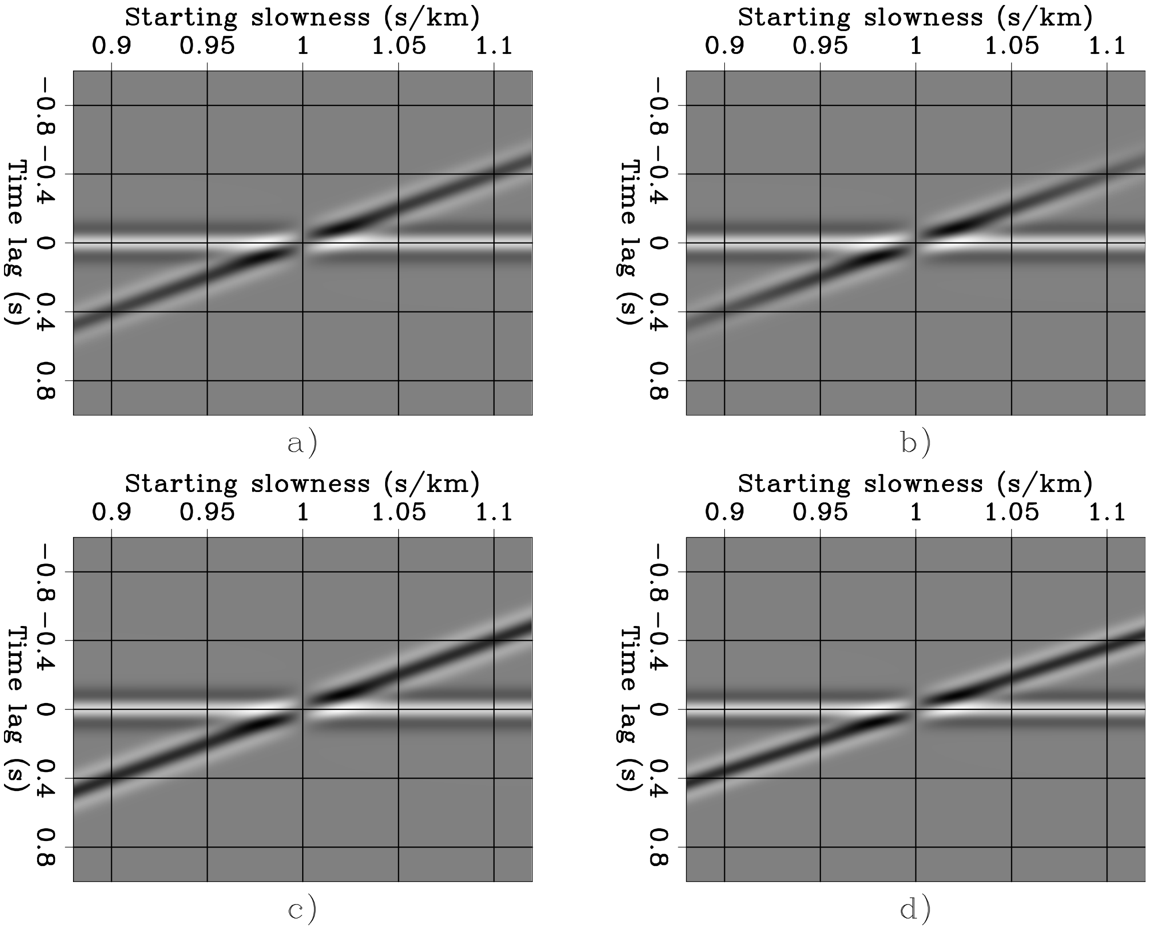


Figure 3: Results of applying the four focusing operators to the convolutional filters **c** shown in Figure 1d: a) $\mathbf{F}_D\mathbf{c}$, b) $\mathbf{F}_G\mathbf{c}$, c) $\mathbf{F}_S\mathbf{c}$, and d) $\mathbf{F}_\alpha\mathbf{c}$. For both $\mathbf{F}_D$ and $\mathbf{F}_G$, $\tau_{\rm W} = 1.$, and $\alpha = 1.111$ for $\mathbf{F}_\alpha$. **[ER]**

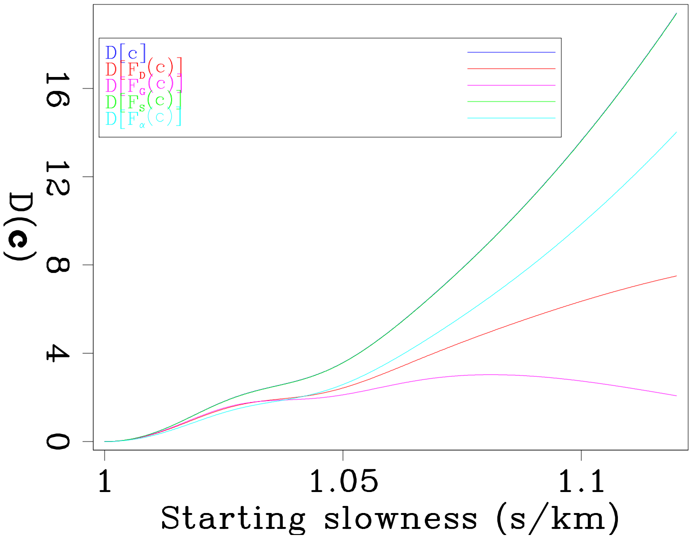


Figure 4: Plots of: $D(\mathbf{c})$, $D(\mathbf{F}_D\mathbf{c})$, $D(\mathbf{F}_G\mathbf{c})$, $D(\mathbf{F}_S\mathbf{c})$, and $D(\mathbf{F}_\alpha\mathbf{c})$. For both $\mathbf{F}_D$ and $\mathbf{F}_G$, $\tau_{\rm W} = 1.$, and $\alpha = 1.111$ for $\mathbf{F}_\alpha$. Notice that the plot of $D(\mathbf{F}_S\mathbf{c})$ almost perfectly overlaps the plot of $D(\mathbf{c})$; however, it is below it for all values of s_o. **[ER]**

Data-space regularization

Another approach of regularization is to add a data-space regularization term. One way to write this objective is as:

$$J_{\mathrm{D}}\left(s,\mathbf{c}\right)=\frac{1}{2}\left\|\tilde{\mathcal{L}}\left(s,\mathbf{c}\right)-\mathbf{d}_r\right\|_2^2+\frac{\epsilon}{2}\left\|\tilde{\mathcal{L}}\left(s,\mathbf{c}\right)-\tilde{\mathcal{L}}\left(s,\mathbf{F}\mathbf{c}\right)\right\|_2^2. \tag{20}$$

As was discussed with the model-space regularization, choosing $\mathbf{F}$ to be the DSO operator provides the best convergence properties for optimizing this objective function. The results of running this inversion for 10000 iterations and a starting model of $s_0 = 1.12$ s/km and $\mathbf{c}_0 = \mathbf{0}$ with $\epsilon = 10$ and the DSO operator as the focusing operator $\mathbf{F}$ are shown in Figure 5.

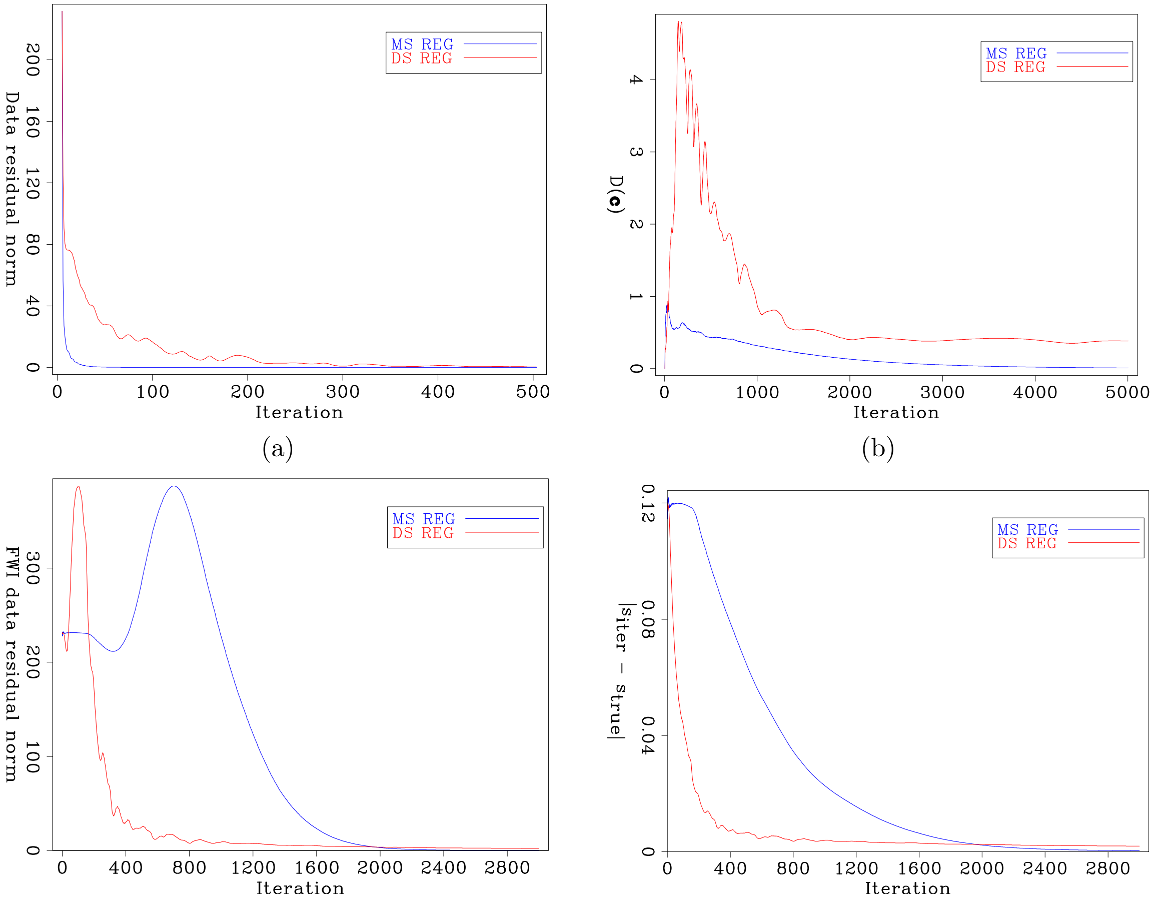
Comparison

Comparing both model and data space regularization results (blue and red curves respectively) in each of the panels in Figure 5 we observe from the data residual norms (panel (a)) that the model space regularization reduces the extended FWI objective function ($||\tilde{\mathcal{L}}\left(s,\mathbf{c}\right)-\mathbf{d}_r||_2^2$) faster than does the data space. This is due to the fact that with the model space regularization, we are directly focusing the model therefore providing better focusing of the extended model ($\mathbf{c}$). This is clear in panel (b) which shows the focusing measure $D(\mathbf{c})$ with iteration. However, the data space regularization does update the model much faster than the model space regularization. This is evident in panels (c) and (d) where the FWI data residual norm ($\frac{1}{2}||\mathcal{L}\left(s\right)-\mathbf{d}_r||_2^2$) and the model residual are shown respectively. From these figures it is clear that in only 600 iterations the updated physical slowness is within 0.45% of the true slowness. In contrast, the model space regularization needs more than 1500 iterations before it reaches that point in the inversion. At later iterations, it is evident in both panels (c) and (d) that the model-space regularization does reduce the objective function more than the data-space regularization. This only occurs when the residuals are quite small.

These results suggest that a data-space regularization with frequent restarts; that is, by resetting set $\mathbf{c}$ to zero might be the best approach within this class of methods. Biondi and Almomin (2014) presented a nested optimization algorithm that included a restart at each outer iteration. At the limit, if the algorithm is restarted by resetting set $\mathbf{c}$ to zero after each update of the slowness model, we will implement an inversion scheme close the the alternating algorithms presented in the next section.

A CLASS OF ALTERNATING ALGORITHMS

In this section, we discuss a class of algorithms to solve the extended FWI problem. We begin by first introducing an algorithm based on a simple intuitive idea, which



(a) (b)

(c) (d)

Figure 5: Results of inversions with model space (blue curve) and data space (red curve) regularization. Both inversions ran for 7000 iterations and plots are windowed here for display. The inversions were run with starting a model of $s_0 = 1.12$ s/km and $\mathbf{c}_0 = 0$. (a) Data residual norms of extended FWI objective function $(\frac{1}{2}||\tilde{\mathcal{L}}\,(s, \mathbf{c}) - \mathbf{d}_r||_2^2)$, (b) focusing measure $D(\mathbf{c})$ as defined in equation 7, (c) data residual norms of FWI objective function $(\frac{1}{2}||\mathcal{L}\,(s) - \mathbf{d}_r||_2^2)$, and (d), model residual ($|s_{\text{iter}} - \bar{s}|$ where s_{iter} is the physical slowness at the current iteration). [**CR**]

we will refer to as the "alternating algorithm". Guided by computational evidence that the algorithm converges to the true slowness $\bar{s}$ for any incorrect starting slowness s_0, we attempt to understand the underlying cause of its global convergence. Mathematical analysis reveals the key property governing its convergence properties. We discuss some of these findings briefly.

Although we are not able to mathematically prove global convergence of the alternating algorithm, we are able to build on the analysis to develop a simpler modified algorithm that retains the global convergence properties. We will refer to it as the "modified alternating algorithm". It should be noted that the modified alternating algorithm is under investigation and it converges faster, but this property may be specific to the 1D problem analyzed in this paper. Whether these properties hold for the general case is not known at this time.

Throughout this section we do not enforce the requirement that $c_0 = 0$. It is not necessary for the class of algorithms discussed in this section for the specified 1D problem. In addition, we also assume that the length of the convolution filter is given by $N_\tau = 2N_t - 1$.

Alternating algorithm

We motivate the first algorithm by considering the residual of the total modeled data using the extended modeling operator measured in the l^2 norm. For a given starting slowness s_0 and convolution filter $\mathbf{c}$, we denote this quantity as $J_{\mathbf{c}}(s_0, \mathbf{c})$. It is defined below as:

$$J_{\mathbf{c}}(s_0, \mathbf{c}) = \frac{1}{2} || \widetilde{\mathcal{L}}(s_0, \mathbf{c}) - \mathbf{d}_r ||_2^2. \tag{21}$$

It is important to remember that we had originally introduced the convolution filter $\mathbf{c}$ to represent an extended set of model parameters that we could change as we like to model the recorded data, for any starting slowness s_0. A natural way to achieve this goal is to fix s_0 in equation 21, and then perform gradient descent to determine a suitable $\mathbf{c}$ that minimizes $J_{\mathbf{c}}(s_0, \mathbf{c})$. In fact for fixed s_0, the function $J_{\mathbf{c}}(s_0, \mathbf{c})$ is a semidefinite quadratic, and thus doing a sequence of steepest descent iterations over $\mathbf{c}$ will converge to a global minimum, which in absence of noise in the data happens to be zero. However, solving such a subproblem to convergence involves repeated iterations, irrespective of the optimization algorithm employed.

We look at a cheaper alternative which is to only look at the negative gradient at the first iteration, still for fixed s_0 but with $\mathbf{c} = \mathbf{0}$. We denote this quantity $\hat{\mathbf{c}}(s_0)$ and define it below,

$$\hat{\mathbf{c}}(s_0) = -\frac{\partial J_{\mathbf{c}}(s, \mathbf{c})}{\partial \mathbf{c}} \bigg|_{s=s_0, \mathbf{c}=\mathbf{0}} = \tilde{\mathbf{L}}'(s_o) \left[\mathbf{d}_r - \widetilde{\mathcal{L}}(s_0, \mathbf{0}) \right] = \tilde{\mathbf{L}}'(s_o) \left[\mathbf{d}_r - \mathcal{L}(s_0) \right]. \tag{22}$$

Note that $-\hat{\mathbf{c}}(s_0)$ is the same quantity that was plotted before in Figure 1d, i.e the application of $\tilde{\mathbf{L}}'(s_o)$ to the data residual $\mathcal{L}(s_0) - \mathbf{d}_r$. This figure can be understood as a superposition of a central positive band that is invariant with respect to s_0, and a diagonal negative band that depends on s_0. The positive band is exactly the term $\tilde{\mathbf{L}}'(s_o)\,\mathcal{L}(s_0)$, while the negative band is the term $-\tilde{\mathbf{L}}'(s_o)\,\mathbf{d}_r$ appearing in the expression for $\hat{\mathbf{c}}(s_0)$ in equation 22.

It can be mathematically proved that the positive band is invariant with respect to s_0. It is also clear from the figure that the closer s_0 is to $\bar{s}$, the diagonal negative band is closer to the zero lag coefficient of $\hat{\mathbf{c}}(s_0)$. This observation is key in understanding this algorithm and motivates the following idea : given any $\hat{\mathbf{c}}(s_0)$, we can try to apply a focusing operator $\mathbf{F}$ to transform $\hat{\mathbf{c}}(s_0)$ to an approximation of $\hat{\mathbf{c}}(s_0+\Delta s)$, where Δs is a slowness perturbation towards $\bar{s}$. Mathematically, this idea is expressed below:

$$\mathbf{F}\hat{\mathbf{c}}(s_0) \approx \hat{\mathbf{c}}(s_0 + \Delta s). \tag{23}$$

Assuming that we have carried out the above transformation, the only thing remaining to do is to find a way to recover Δs from $\mathbf{F}\hat{\mathbf{c}}(s_0)$. An intuitive idea would be to match the quantities $\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}(s_0))$ and $\widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}}(s_0))$ in the least squares sense, where s is close to s_0. For fixed s_0, this leads to the following objective function:

$$J_s(s) = \frac{1}{2}||\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}(s_0)) - \widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}}(s_0))||_2^2. \tag{24}$$

The quantity $\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}(s_0))$ represents the total modeled data using the extended modeling operator for slowness s_0 and convolution filter $\hat{\mathbf{c}}(s_0)$, while $\widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}}(s_0))$ represents the total modeled data for any slowness value s close to s_0 and the focused convolution filter $\mathbf{F}\hat{\mathbf{c}}(s_0)$. The expectation is that minimizing the objective function $J_s(s)$ will yield a slowness perturbation Δs towards $\bar{s}$. Putting all these ideas together, we have the following algorithm:

1. For any starting slowness s_0, compute $\hat{\mathbf{c}}(s_0) = \tilde{\mathbf{L}}'(s_o)\left[\mathbf{d}_r - \mathcal{L}(s_0)\right].$
2. Start from $s = s_0$, and solve the following local optimization problem:

$$s_* = \underset{s}{\operatorname{argmin}}\ J_s(s) \tag{25}$$

3. Set $s_0 = s_*$ and iterate 1-3 till convergence.

In the above discussion $\mathbf{F}$ is a general focusing operator. However in this section, we restrict ourselves to the particular case of the shift focusing operator, which was introduced earlier in equation 18. Note that the choice of the shift focusing operator is inherently restrictive, because we only shift the coefficients in the extended model vector by one sample towards the zero lag coefficient. Doing that automatically enforces small changes in Δs. It is indeed possible to incorporate bigger shifts up to a limit into $\mathbf{F}_S$, and we would obtain similar results to what we present next. The choice of studying the one sample shift focusing operator captures all the effects that would be true with bigger shifts. The subsequent analysis of the alternating algorithm also holds in this regime in a slightly modified form. In fact, with bigger shifts the rate of convergence is much faster at each iteration when s_0 is far away from $\bar{s}$. However, if the shifts are too large we lose the property that Δs is a slowness perturbation towards $\bar{s}$ for all starting slowness s_0.

Such ideas of incorporating bigger shifts and speeding up convergence can also be incorporated with the use of other types of focusing operators like the shrink focusing operator $\mathbf{F}_\alpha$ introduced earlier.

Numerical results

We provide computational evidence that the alternating algorithm converges to $\bar{s}$ for any starting slowness s_0. To illustrate this we start by plotting the objective function $J_s(s)$ for different values of s_0 in Figure 6. It is clearly seen from each panel that minimizing $J_s(s)$ starting from s_0 will yield a slowness update Δs towards $\bar{s}$. Therefore, if this process is repeated at every iteration, we will reach $\bar{s}$. The iterates eventually stop changing when $s_0 = \bar{s}$ as the gradient of $J_s(s)$ with respect to s at $s = \bar{s}$ becomes zero. This test shows that the alternating algorithm converges to the true slowness $\bar{s}$ starting from any s_0.

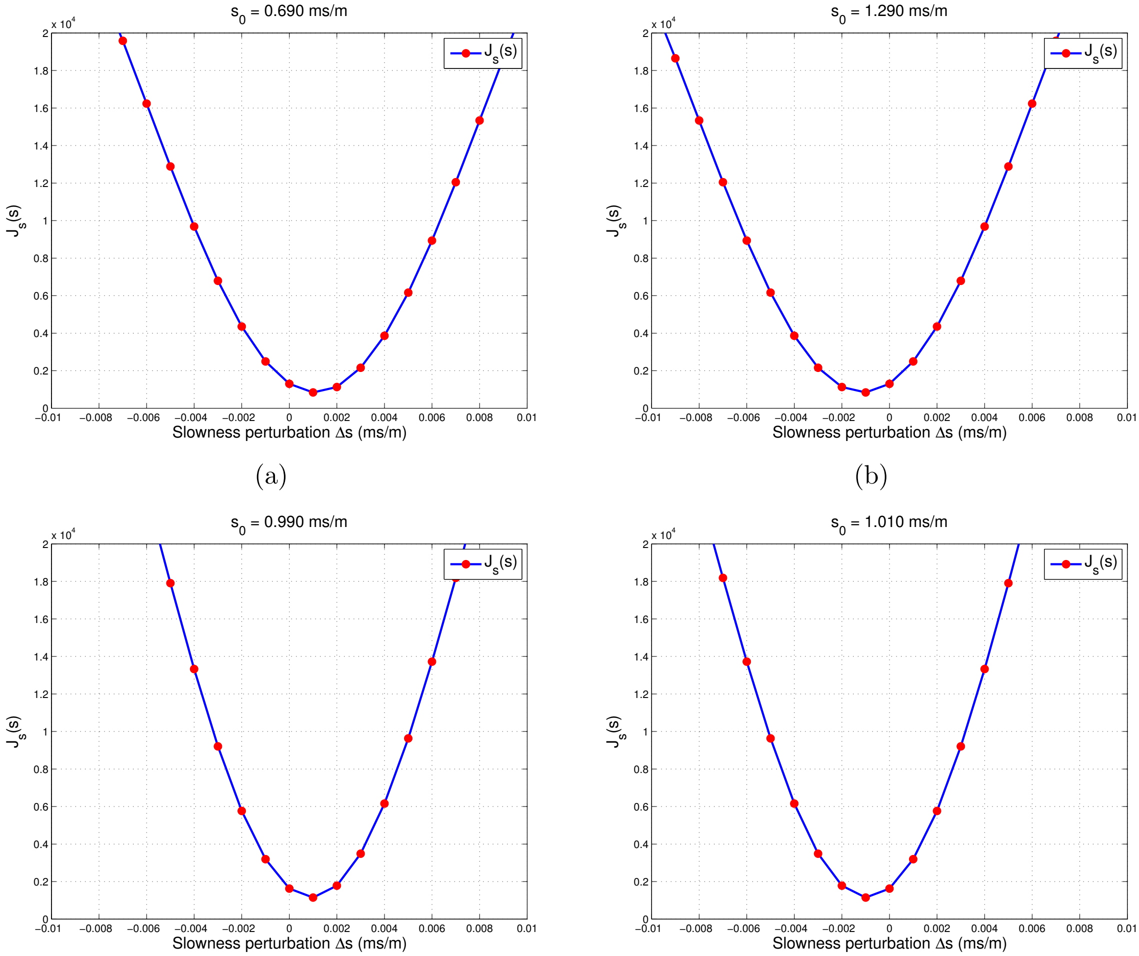
Another way of seeing the global convergence property is to evaluate the gradient of the objective function $J_s(s)$ for each s_0 at $\Delta s = 0$, i.e $\frac{\partial J_s(s_0)}{\partial s}$. We first calculate the gradient of $J_s(s)$ below:

$$
\begin{aligned}
\frac{\partial J_s(s)}{\partial s} &= -\left[\frac{\partial \mathcal{L}(s)}{\partial s} + \frac{\partial \tilde{\mathbf{L}}\,(s)}{\partial s}\mathbf{F}\hat{\mathbf{c}}(s_0))\right]' \left[\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}(s_0)) - \widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}}(s_0))\right] \\
&= -\left[\mathbf{L}(s) + \dot{\tilde{\mathbf{L}}}\,(s)\,\mathbf{F}\hat{\mathbf{c}}(s_0))\right]' \left[\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}(s_0)) - \widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}}(s_0))\right] \qquad (26) \\
&= -\left[\mathbf{L}'(s) + \hat{\mathbf{c}}'(s_0)\mathbf{F}'\,\dot{\tilde{\mathbf{L}}}'\,(s)\right] \left[\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}(s_0)) - \widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}}(s_0))\right] \quad .
\end{aligned}
$$

Evaluating this quantity at $s = s_0$ gives:

$$
\frac{\partial J_s(s_0)}{\partial s} = -\left[\mathbf{L}'(s_0) + \hat{\mathbf{c}}'(s_0)\mathbf{F}'\,\dot{\tilde{\mathbf{L}}}'\,(s_o)\right] \tilde{\mathbf{L}}\,(s_o)\,[\mathbf{I} - \mathbf{F}]\hat{\mathbf{c}}(s_0) \quad . \qquad (27)
$$

We have plotted the quantity $\frac{\partial J_s(s_0)}{\partial s}$ as a function of different starting slowness values s_0 in Figure 7. The figure tells us that for $s_0 > \bar{s}$, the gradient is always positive, while for $s_0 < \bar{s}$, the gradient is always negative. Thus, the search direction which is the negative of the gradient always points in the correct direction. Finally when $s_0 = \bar{s}$, the gradient is zero, which means that the algorithm will terminate when $s_0 = \bar{s}$.



(a) (b)

(c) (d)

Figure 6: Plot of the function $J_s(s)$ for different values of s_0 : (a) $s_0 = 0.69$ ms/m represents the case when s_0 is slow compared to $\bar{s}$, (b) $s_0 = 1.29$ ms/m represents the case when s_0 is fast compared to $\bar{s}$, (c) $s_0 = 0.99$ ms/m represents the case when s_0 is almost close to $\bar{s}$ on the slower side, and (d) $s_0 = 1.01$ ms/m represents the case when s_0 is almost close to $\bar{s}$ on the faster side. [**ER**]

Analysis of the alternating algorithm

The computational evidence of global convergence to $\bar{s}$ naturally leads us to take a closer look at the alternating algorithm. Here the main question of interest is *"why is the algorithm converging ?"*. We want to isolate the pieces of the objective function that is responsible for its convergence.

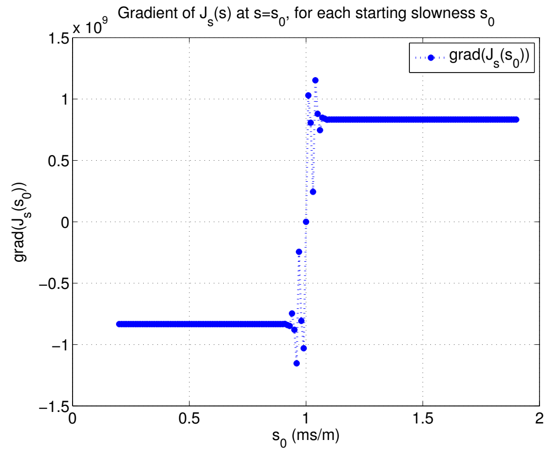


Figure 7: Plot of the quantity $\frac{\partial J_s(s_0)}{\partial s}$ in equation 27 as a function of different starting slowness values s_0. [**ER**]

We first focus our attention to the behavior of $J_s(s)$ for some fixed s_0. We will also henceforth use $\hat{\mathbf{c}}$ to denote $\hat{\mathbf{c}}(s_0)$, unless stated otherwise to avoid excessive notation. We start by expanding the expression for $J_s(s)$ and group similar terms to get the following:

$$\begin{aligned} J_s(s) &= \frac{1}{2}||\widetilde{\mathcal{L}}(s_0, \hat{\mathbf{c}}) - \widetilde{\mathcal{L}}(s, \mathbf{F}\hat{\mathbf{c}})||_2^2 = \frac{1}{2}||\left[\mathcal{L}(s_0) - \mathcal{L}(s)\right] + \left[\tilde{\mathbf{L}}\left(s_o\right)\hat{\mathbf{c}} - \tilde{\mathbf{L}}\left(s\right)\mathbf{F}\hat{\mathbf{c}}\right]||_2^2 \\ &= \frac{1}{2}||\left[\mathcal{L}(s_0) - \mathcal{L}(s)\right] - \left[\tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right)\mathcal{L}(s_0) - \tilde{\mathbf{L}}\left(s\right)\mathbf{F}\tilde{\mathbf{L}}'\left(s_o\right)\mathcal{L}(s_0)\right] \\ &+ \left[\tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r - \tilde{\mathbf{L}}\left(s\right)\mathbf{F}\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r\right]||_2^2 \\ &= \frac{1}{2}||\mathbf{u}+\mathbf{v}+\mathbf{w}||_2^2 = \frac{1}{2}\left[||\mathbf{u}||_2^2 + ||\mathbf{v}||_2^2 + ||\mathbf{w}||_2^2\right] + \left[\mathbf{u}'\mathbf{v} + \mathbf{v}'\mathbf{w} + \mathbf{w}'\mathbf{u}\right] \quad , \end{aligned} \tag{28}$$

where we have denoted,

$$\begin{aligned} \mathbf{u} &:= \mathbf{u}(s) = \mathcal{L}(s_0) - \mathcal{L}(s) \\ \mathbf{v} &:= \mathbf{v}(s) = -[\tilde{\mathbf{L}}\left(s_o\right) - \tilde{\mathbf{L}}\left(s\right)\mathbf{F}]\tilde{\mathbf{L}}'\left(s_o\right)\mathcal{L}(s_0) \\ \mathbf{w} &:= \mathbf{w}(s) = [\tilde{\mathbf{L}}\left(s_o\right) - \tilde{\mathbf{L}}\left(s\right)\mathbf{F}]\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r \quad . \end{aligned} \tag{29}$$

The different terms involving u,v and w

It turns out that the most interesting term is the one involving only $||\mathbf{w}||_2^2$. This term is the key in getting the correct update when s_0 is far away from $\bar{s}$, and so the modeled data $\mathcal{L}(s_0)$ and the recorded data $\mathbf{d}_r$ don't interfere with each other. When this is the case, two key mathematical properties hold for the 1D problem being studied, as stated below:

$$\mathbf{F}\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r = \tilde{\mathbf{L}}'\left(s_0+\Delta s\right)\mathbf{d}_r + \mathcal{O}((\delta t)^2) \quad , \tag{30a}$$

$$\tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right) = \tilde{\mathbf{L}}\left(s_0+\Delta s\right)\tilde{\mathbf{L}}'\left(s_0+\Delta s\right) + \mathcal{O}((\delta t)^2) \quad . \tag{30b}$$

Demonstration of equations 30 is straightforward but lengthy; therefore, we decided to omit it from the text.

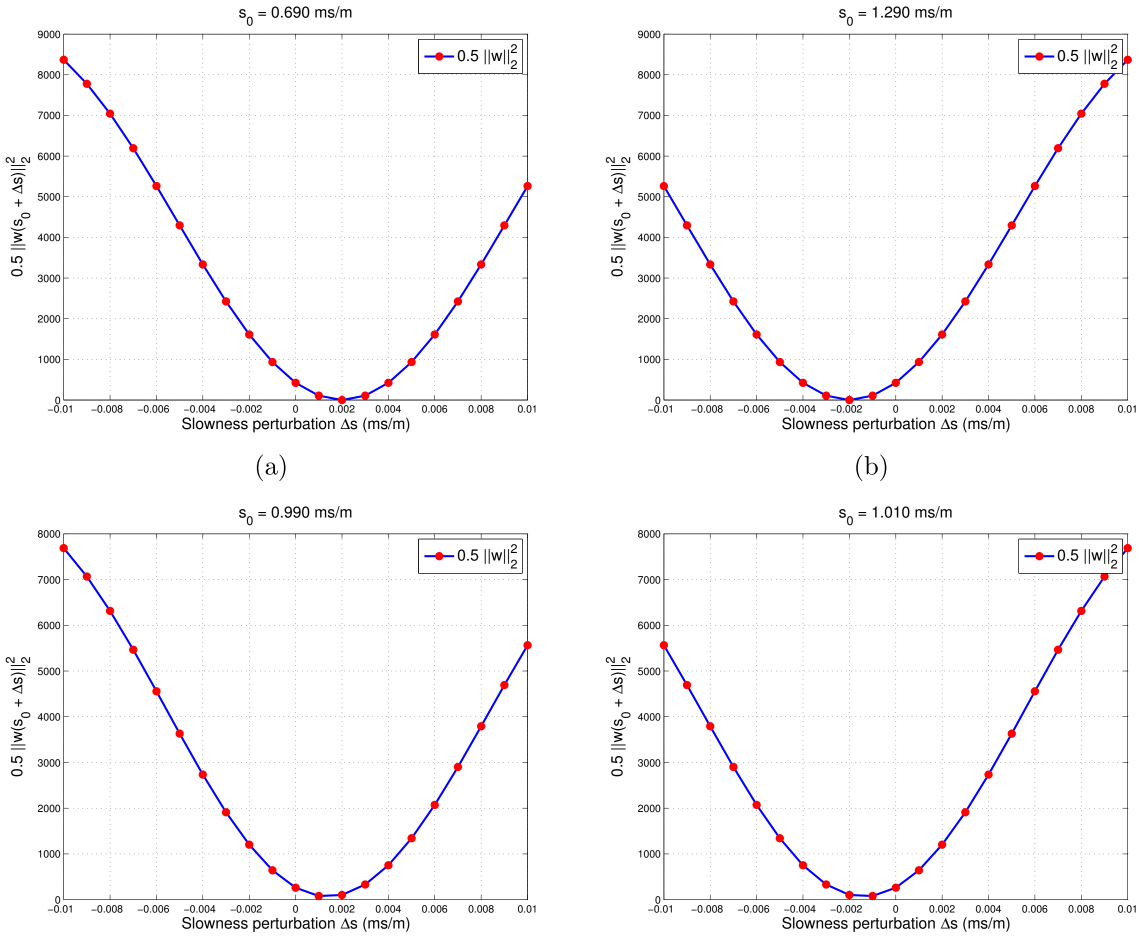
In equation 30, Δs is a slowness perturbation in the correct direction (towards $\bar{s}$). The immediate consequence of these relations is that the quantity $\frac{1}{2}||\mathbf{w}||_2^2$ vanishes at $s_0+\Delta s$. This is illustrated below in equation 31, where the second line follows from 30(a) and the third line follows from 30(b).

$$\begin{aligned}\frac{1}{2}||\mathbf{w}||_2^2\Big|_{s_0+\Delta s} &= \frac{1}{2}||\tilde{\mathbf{L}}\left(s_0+\Delta s\right)\mathbf{F}\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r - \tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r||_2^2 \\ &= \frac{1}{2}||\tilde{\mathbf{L}}\left(s_0+\Delta s\right)\tilde{\mathbf{L}}'\left(s_0+\Delta s\right)\mathbf{d_r} - \tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r + \mathcal{O}((\delta t)^2)||_2^2 \\ &= \frac{1}{2}||\tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r - \tilde{\mathbf{L}}\left(s_o\right)\tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r + \mathcal{O}((\delta t)^2)||_2^2 \\ &= \mathcal{O}((\delta t)^4) \approx 0 \quad .\end{aligned} \tag{31}$$

Numerical studies also suggest that the function $\frac{1}{2}||\mathbf{w}||_2^2$ is increasing in the interval $[\min(s_0, s_0+\Delta s), \max(s_0, s_0+\Delta s)]$. Thus if we carry out local minimization of the $\frac{1}{2}||\mathbf{w}||_2^2$ term starting from the initial slowness s_0, we would obtain the new optimal point $s_0+\Delta s$, which is always closer to the true slowness $\bar{s}$. It is to be noted that the above argument only holds in the regime when s_0 is far away from $\bar{s}$. When this is not the case, the modeled data $\mathcal{L}(s_0)$ and the recorded data $\mathbf{d}_r$ begin to interfere and equation 31 does not hold. However, numerical studies seem to indicate that one can still get the correct update direction by minimizing $\frac{1}{2}||\mathbf{w}||_2^2$.

These aspects are illustrated in Figure 8, where we have plotted the quantity $\frac{1}{2}||\mathbf{w}(s_0+\Delta s)||_2^2$ as a function of the slowness perturbation Δs around s_0, for different values of s_0. Figures 8(a) and 8(b) represent cases when s_0 and $\bar{s}$ are sufficiently far apart so that there is no interference between $\mathcal{L}(s_0)$ and $\mathbf{d}_r$. As can be seen on both the figures, the function goes to zero on the "correct" side, i.e if we start from $\Delta s = 0$ and try to minimize $\frac{1}{2}||\mathbf{w}(s_0+\Delta s)||_2^2$, the optimal Δs would represent a step towards $\bar{s}$. The same fact is true also for Figures 8(c) and 8(d), which represent cases where s_0 is so close to $\bar{s}$ that there is interference between $\mathcal{L}(s_0)$ and $\mathbf{d}_r$. In both of these cases, one can see that the minima of $\frac{1}{2}||\mathbf{w}(s_0+\Delta s)||_2^2$ is still in the right direction, but the function does not become zero at the minima. This is precisely connected to the fact that equation 31 is losing accuracy in this regime.

It seems from the analysis of the $\frac{1}{2}||\mathbf{w}||_2^2$ term that if one ignored all the other terms in the expression for $J_s(s)$, we would still have global convergence. In fact



(a)

(b)

(c)

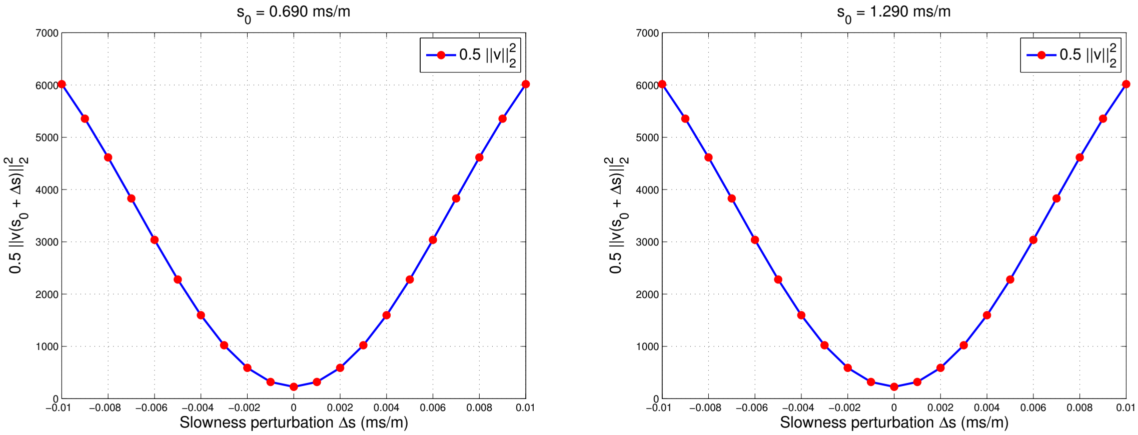
(d)

Figure 8: Plot of the function $\frac{1}{2}||\mathbf{w}(s_0 + \Delta s)||_2^2$ for different values of s_0 : (a) $s_0 = 0.69$ ms/m represents the case when s_0 is slow compared to $\bar{s}$, (b) $s_0 = 1.29$ ms/m represents the case when s_0 is fast compared to $\bar{s}$, (c) $s_0 = 0.99$ ms/m represents the case when s_0 is almost close to $\bar{s}$ on the slower side, and (d) $s_0 = 1.01$ ms/m represents the case when s_0 is almost close to $\bar{s}$ on the faster side. [**ER**]

this is actually the case, and would later motivate the development of the "modified alternating algorithm". However, we still need to consider the behavior of the remaining terms in the expression for $J_s(s)$ and understand if they help or impede the convergence. We thus turn to the $\frac{1}{2}||\mathbf{u}||_2^2$ and $\frac{1}{2}||\mathbf{v}||_2^2$ terms next.

The $\frac{1}{2}||\mathbf{u}||_2^2$ term is not useful. Using the definition of $\mathbf{u}$ in equation 29, we have $\frac{1}{2}||\mathbf{u}||_2^2 = \frac{1}{2}||\mathcal{L}(s) - \mathcal{L}(s_0)||_2^2$. This looks exactly like the FWI objective function $\frac{1}{2}||\mathcal{L}(s) - \mathbf{d}_r||_2^2$, but where the recorded data $\mathbf{d}_r = \mathcal{L}(\bar{s})$ has been replaced by $\mathcal{L}(s_0)$, i.e data modeled using an incorrect slowness s_0. We know from the properties of the FWI objective function that it is a convex function in a sufficiently close neighborhood of the true slowness. Therefore, this property also carries over to the $\frac{1}{2}||\mathbf{u}||_2^2$ term, and we conclude that the $\frac{1}{2}||\mathbf{u}||_2^2$ is locally convex around s_0. This means that if we start sufficiently close to s_0, the gradient of s with respect to the $\frac{1}{2}||\mathbf{u}||_2^2$ term will always point towards s_0. This observation is important, as it says that while the presence of the $\frac{1}{2}||\mathbf{w}||_2^2$ term in $J_s(s)$ will provide the correct update direction, including the $\frac{1}{2}||\mathbf{u}||_2^2$ term in $J_s(s)$ will oppose the correct slowness update. Clearly, this is not a desirable property and we thus conclude that we should omit the $\frac{1}{2}||\mathbf{u}||_2^2$ term in the definition of $J_s(s)$.

Finally we consider the $\frac{1}{2}||\mathbf{v}||_2^2$ term. It turns out that this term is also convex locally around s_0. For this particular 1D case, an interesting fact about this term is that its profile does not vary with s_0. This is illustrated in Figure 9. As clearly seen, the curves in both the figures are exactly same. In fact, this is also true for all possible values of s_0 irrespective of whether s_0 is close or far from $\bar{s}$, but this is most likely only true for the specific 1D problem we are studying. However, we think that the local convexity property of $\frac{1}{2}||\mathbf{v}||_2^2$ close to s_0 may be more general. Thus, presence of this term will also impede convergence as its gradient will oppose any change from s_0.



(a) (b)

Figure 9: Plot of the function $\frac{1}{2}||\mathbf{v}(s_0 + \Delta s)||_2^2$ for different values of s_0 : (a) $s_0 = 0.69$ ms/m corresponds to the case when s_0 is slower compared to $\bar{s}$, and (b) $s_0 = 1.29$ ms/m corresponds to the case when s_0 is faster compared to $\bar{s}$. [**ER**]

The cross terms $\mathbf{u}'\mathbf{v} + \mathbf{v}'\mathbf{w} + \mathbf{w}'\mathbf{u}$ are quite complicated and evades mathematical analysis at the moment. Preliminary analysis has not revealed any clear understanding of how they may be affecting convergence. In general, they produce gradients which are sometimes in the correct direction and at other times not. So it is quite a miracle that the full expression for $J_s(s)$ produces the correct update direction, as evidenced in the numerical results shown earlier.

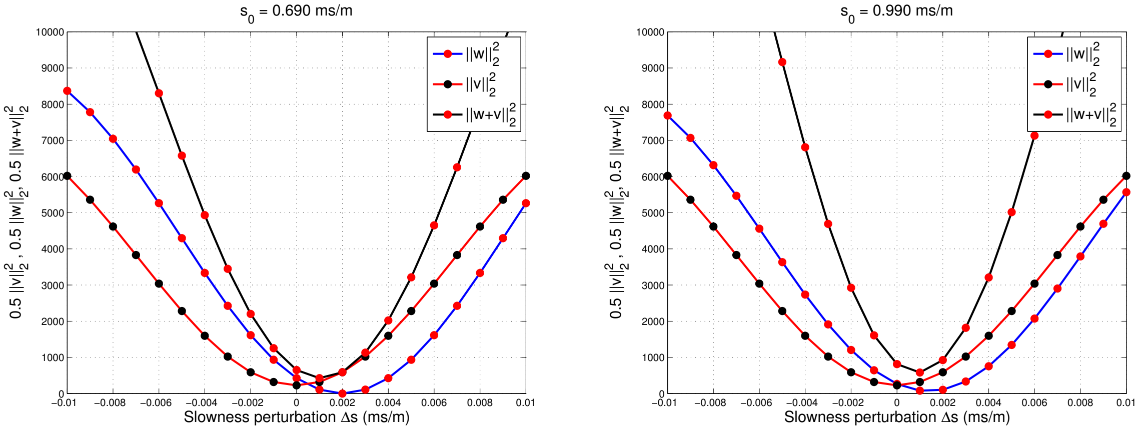
Slowdown of convergence

It is quite clear that for this 1D problem, the effect of the $\frac{1}{2}||\mathbf{w}||_2^2$ term dominates that of the others in the full expression for $J_s(s)$. So as a net, we obtain the correct slowness perturbation Δs by minimizing $J_s(s)$ at every step of the iterative algorithm. Starting from any slowness s_0, one step of the algorithm takes us closer to the true slowness $\bar{s}$, and this process iterated enough times finally takes us to $\bar{s}$. The process converges to $\bar{s}$ because as we get closer and closer to $\bar{s}$, the updates Δs get smaller and smaller. In fact for $s_0 = \bar{s}$, we have $\frac{\partial J_s(\bar{s})}{\partial s} = 0$, and so we do not get any more update.

However from the point of view of algorithmic efficiency and rate of convergence, we would really like to avoid the ill-effects associated with the terms involving $\mathbf{u}$ and $\mathbf{v}$. Based on the analysis done so far, we would like to only use the $\frac{1}{2}||\mathbf{w}||_2^2$ term for solving the local minimization problem at each iteration. But first, we provide some numerical evidence of how the convergence is affected. As an example, we have chosen to drop all terms involving $\mathbf{u}$ from the expression of $J_s(s)$, and analyze the behavior of the remaining terms given by $J_s(s) = \frac{1}{2}||\mathbf{w}||_2^2 + \frac{1}{2}||\mathbf{v}||_2^2 + \mathbf{v}'\mathbf{w} = \frac{1}{2}||\mathbf{w}+\mathbf{v}||_2^2$. This situation is plotted in Figures 10(a),10(b) for slowness values $s_0 = 0.69$ ms/m, 0.99 ms/m. In each figure, we have displayed the quantities $\frac{1}{2}||\mathbf{v}||_2^2$, $\frac{1}{2}||\mathbf{w}||_2^2$ and $\frac{1}{2}||\mathbf{w}+\mathbf{v}||_2^2$. As we can clearly see in both cases, minimizing $\frac{1}{2}||\mathbf{w}+\mathbf{v}||_2^2$ or $\frac{1}{2}||\mathbf{w}||_2^2$ starting from s_0 will yield a slowness perturbation towards $\bar{s}$. But in the latter case, the step will be larger compared to the former. It should be mentioned that similar conclusions are also obtained for the cases when s_0 is faster than $\bar{s}$. We thus conclude that minimizing the complete expression for $J_s(s)$ will take longer to converge to $\bar{s}$ in the presence of the $\mathbf{u}$ and $\mathbf{v}$ terms.

Modified alternating algorithm

To address the convergence issue, we now discuss a modified alternating algorithm that is based on the idea of replacing the original expression for $J_s(s)$ in equation 24 by $J_s(s) = \frac{1}{2}||\mathbf{w}||_2^2 = \frac{1}{2}||[\tilde{\mathbf{L}}\,(s_o) - \tilde{\mathbf{L}}\,(s)\,\mathbf{F}]\tilde{\mathbf{L}}'\,(s_o)\,\mathbf{d}_r||_2^2$. As we have shown previously, local minimization of the $\frac{1}{2}||\mathbf{w}||_2^2$ term produces the correct update direction at every iteration. We had also argued previously based on Figures 8(c) and 8(d) that minimizing the $\frac{1}{2}||\mathbf{w}||_2^2$ term gives the correct update direction, even when s_0 is close to $\bar{s}$. This is indeed the case for the choice of the modeling parameters like wavelet



(a)

(b)

Figure 10: Plot of the different terms $\frac{1}{2}||\mathbf{v}(s_0+\Delta s)||_2^2$, $\frac{1}{2}||\mathbf{w}(s_0+\Delta s)||_2^2$ and $\frac{1}{2}||\mathbf{w}(s_0+\Delta s)+\mathbf{v}(s_0+\Delta s)||_2^2$ for different values of s_0, indicated by the red, blue and black curves respectively : (a) $s_0 = 0.69$ ms/m represents the case when s_0 is slow compared to $\bar{s}$, and (b) $s_0 = 0.99$ ms/m represents the case when s_0 is almost equal $\bar{s}$, but slightly slower. [**ER**]

and sampling interval used to create those plots. However, it has also been observed that for other choices of modeling parameters, the $\frac{1}{2}||\mathbf{w}||_2^2$ term starts to have the minima exactly at s_0, when s_0 starts to get close to $\bar{s}$ and the algorithm terminates before reaching $\bar{s}$. It is quite difficult to characterize this situation analytically due to the interference between $\mathcal{L}(s_0)$ and $\mathbf{d}_r$, but we suspect that the effect is related to the shape and frequency of the wavelet and also numerical inaccuracies stemming from the choice of the modeling parameters. We discuss a robust alternative below that was found to not suffer from this issue by modifying the objective function in such a manner that when s_0 is far from $\bar{s}$, it is exactly equal to $\frac{1}{2}||\mathbf{w}||_2^2$, and the only differences are when s_0 is close to $\bar{s}$.

We first define the unit shift operators in the positive and negative directions $\mathbf{S}_+$ and $\mathbf{S}_-$. $\mathbf{S}_+$ defines a linear map $\Re^{N_\tau} \rightarrow \Re^{N_\tau}$, where each sample is shifted down by one sample. $\mathbf{S}_-$ defines a linear map $\Re^{N_\tau} \rightarrow \Re^{N_\tau}$, where each sample is shifted up by one sample. These operators have the explicit matrix form as defined below:

$$\mathbf{S}_+ = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & 1 & 0 & \\ & & & 1 & 0 \end{bmatrix}_{N_\tau \times N_\tau} , \quad \mathbf{S}_- = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}_{N_\tau \times N_\tau} . \tag{32}$$

We next introduce the masking operator $\mathbf{M}_i$. It also defines a linear map $\Re^{N_\tau} \rightarrow$

$\Re^{N_\tau}$, and its action on a vector $\mathbf{c} \in \Re^{N_\tau}$ is defined as follows:

$$\mathbf{M}_i \mathbf{c} = \tilde{\mathbf{c}} \quad \text{, where,} \quad \tilde{c}_j = \begin{cases} c_j, & \text{if } j \leq i \\ 0, & \text{if } j > i \end{cases} \quad . \tag{33}$$

The above quantities $\mathbf{S}_+, \mathbf{S}_-, \mathbf{M}_i$ form the building blocks of the modified algorithm. For the remainder of this section we will redefine the quantity $\hat{\mathbf{c}}$ to be $\hat{\mathbf{c}} = \tilde{\mathbf{L}}'(s_o)\,\mathbf{d}_r$. We also need to define two more quantities $\mathbf{r}_+(s)$ and $\mathbf{r}_-(s)$, which can be interpreted as generalized residuals and defined below:

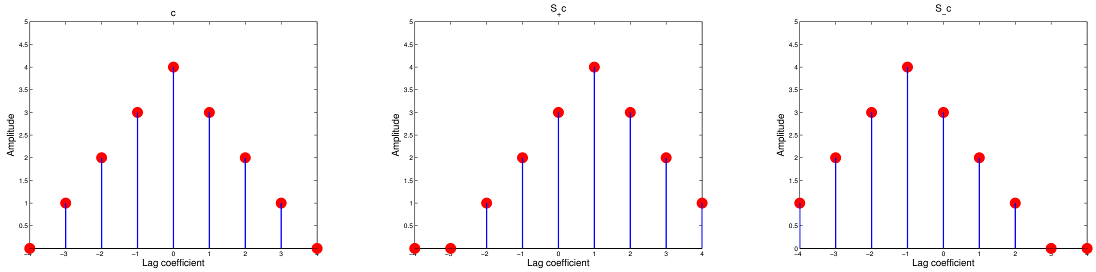
$$\begin{aligned} \mathbf{r}_+ &:= \mathbf{r}_+(s) = \tilde{\mathbf{L}}(s)\,\mathbf{M}_0 \mathbf{S}_+ \hat{\mathbf{c}} - \tilde{\mathbf{L}}(s_o)\,\mathbf{M}_{-1} \hat{\mathbf{c}} \\ \mathbf{r}_- &:= \mathbf{r}_-(s) = \tilde{\mathbf{L}}(s)\,[\mathbf{I} - \mathbf{M}_{-1}]\mathbf{S}_- \hat{\mathbf{c}} - \tilde{\mathbf{L}}(s_o)\,[\mathbf{I} - \mathbf{M}_0]\hat{\mathbf{c}} \quad . \end{aligned} \tag{34}$$

It is instructive to look at what the combination of terms appearing in equation 34 involving $\mathbf{S}_+, \mathbf{S}_-$ and $\mathbf{M}_0, \mathbf{M}_{-1}$ look like when applied to a vector $\mathbf{c}$. To clarify this point, we choose a dummy vector $\mathbf{c}$ corresponding to the choice of parameters $N_t = 5$, $N_\tau = 2N_t - 1 = 9$ as shown in Figure 11(a). We then applied the unit positive and negative shifts and plotted the results in Figures 11(b) and 11(c), respectively. In Figures 12(a), 12(b), 12(c), and 12(d) we plot the terms $\mathbf{M}_0\mathbf{S}_+\mathbf{c}$, $\mathbf{M}_{-1}\mathbf{c}$, $[\mathbf{I} - \mathbf{M}_{-1}]\mathbf{S}_-\mathbf{c}$ and $[\mathbf{I} - \mathbf{M}_0]\mathbf{c}$, respectively. As seen from these figures, the terms $\mathbf{M}_0\mathbf{S}_+\mathbf{c}$ and $\mathbf{M}_{-1}\mathbf{c}$ are exactly shifted copies of each other with the property that all positive lag coefficients are zero. Similarly, the terms $[\mathbf{I} - \mathbf{M}_{-1}]\mathbf{S}_-\mathbf{c}$ and $[\mathbf{I} - \mathbf{M}_0]\mathbf{c}$ are also shifted copies of each other with the property that all negative lag coefficients are zero.

Thus $\mathbf{r}_+$ term can be interpreted as the difference between the data produced by the linearized extension using the starting slowness s_0 and convolution filter $\mathbf{M}_{-1}\hat{\mathbf{c}}$ (which is the $\tilde{\mathbf{L}}(s_o)\,\mathbf{M}_{-1}\hat{\mathbf{c}}$ term), and the linearized extension using a slowness s close to s_0 and convolution filter $\mathbf{M}_0\mathbf{S}_+\hat{\mathbf{c}}$ (which is the $\tilde{\mathbf{L}}(s)\,\mathbf{M}_0\mathbf{S}_+\hat{\mathbf{c}}$ term). A similar interpretation also holds for $\mathbf{r}_-(s)$. It is clear from the definition of the mask operator that when the support of $\hat{\mathbf{c}}$ is strictly negative, $\mathbf{r}_- = \mathbf{0}$ holds identically, and when the support of $\hat{\mathbf{c}}$ is strictly positive, $\mathbf{r}_+ = \mathbf{0}$ holds identically. These situations correspond to the cases $s_0 >> \bar{s}$ and $s_0 << \bar{s}$ respectively and hence the modeled data $\mathcal{L}(s_0)$ and recorded data $\mathbf{d}_r$ do not interfere with each other in both these cases.

The preceding observations allow us to define the modified objective function $J_M(s)$ in terms of the squared l^2 norms of $\mathbf{r}_+$ and $\mathbf{r}_-$ which we define to be $J_+(s)$ and $J_-(s)$ respectively, as below:

$$\begin{aligned} J_+(s) &= \frac{1}{2}||\mathbf{r}_+||_2^2 = \frac{1}{2}||\tilde{\mathbf{L}}(s)\,\mathbf{M}_0\mathbf{S}_+\hat{\mathbf{c}} - \tilde{\mathbf{L}}(s_o)\,\mathbf{M}_{-1}\hat{\mathbf{c}}||_2^2 \\ J_-(s) &= \frac{1}{2}||\mathbf{r}_-||_2^2 = \frac{1}{2}||\tilde{\mathbf{L}}(s)\,[\mathbf{I} - \mathbf{M}_{-1}]\mathbf{S}_-\hat{\mathbf{c}} - \tilde{\mathbf{L}}(s_o)\,[\mathbf{I} - \mathbf{M}_0]\hat{\mathbf{c}}||_2^2 \\ J_M(s) &= J_+(s) + J_-(s) \quad . \end{aligned} \tag{35}$$



(a) (b) (c)

Figure 11: Plot of the result of application of the shift operators to a vector corresponding to the choice of parameters $N_t = 5$, $N_\tau = 9$. (a) This is the vector $\mathbf{c}$. (b) Result of application of unit positive shift $\mathbf{S}_+\mathbf{c}$. (c) Result of application of unit negative shift $\mathbf{S}_-\mathbf{c}$. [**ER**]

We can finally write the modified alternating algorithm:

1. $\hat{\mathbf{c}} = \tilde{\mathbf{L}}'\left(s_o\right)\mathbf{d}_r$.

2. Start from $s = s_0$, and solve the local minimization problem:

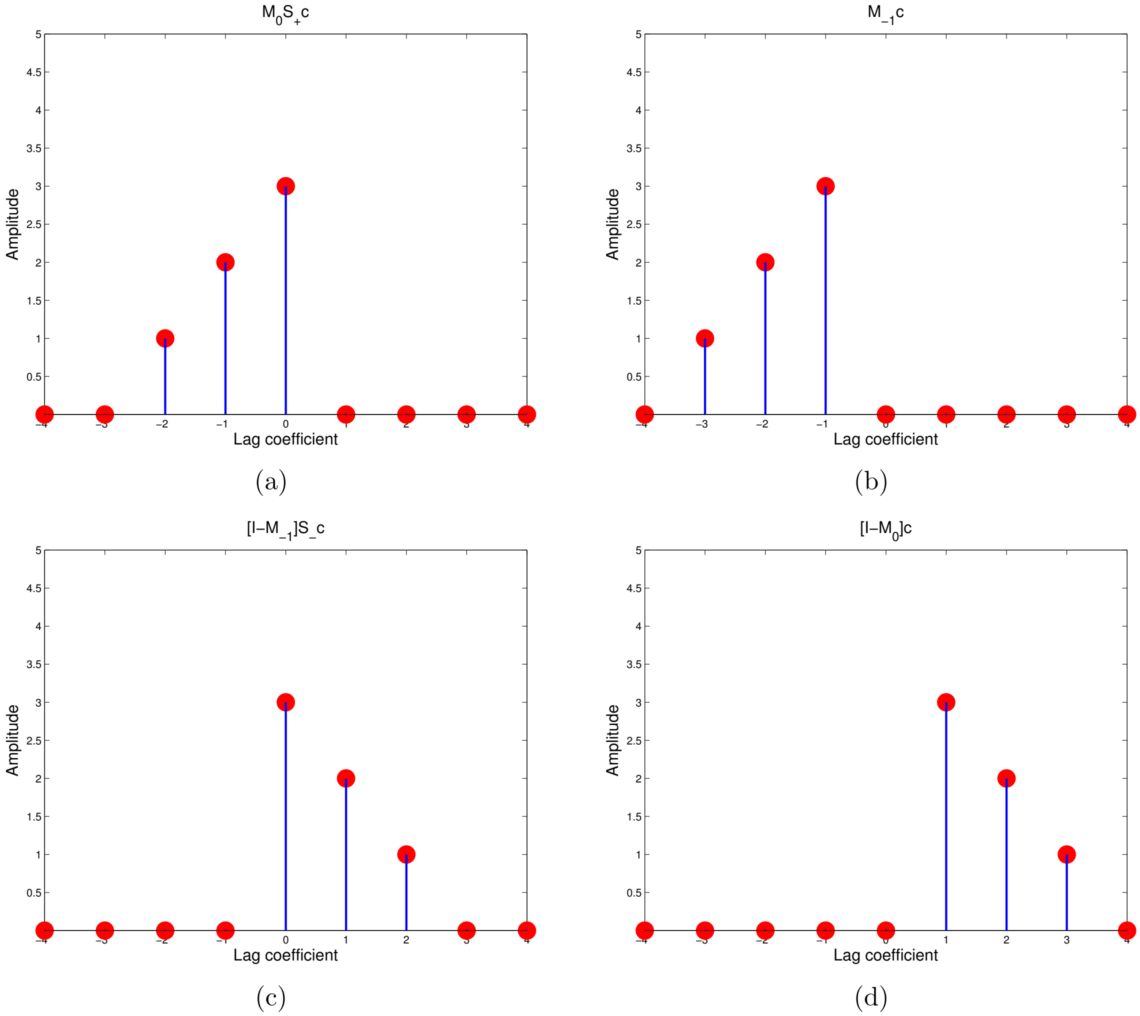
$$s_* = \underset{s}{\operatorname{argmin}} \;\; J_M(s) \tag{36}$$

3. Set $s_0 = s_*$ and iterate 1-3 till convergence.

The objective function $J_M(s)$ has two interesting properties. The first one is that when s_0 is far away from $\bar{s}$, $J_M(s) = \frac{1}{2}||\mathbf{w}||_2^2$. The second property is that when $s_0 >> \bar{s}$, $J_-(s) = 0$ and hence $J_M(s) = J_+(s)$, and similarly when $s_0 << \bar{s}$, $J_+(s) = 0$ and hence $J_M(s) = J_-(s)$. Both $J_+(s)$ and $J_-(s)$ are non-zero only when $\mathcal{L}(s_0)$ and $\mathbf{d}_r$ are interfering, which happen when s_0 is close to $\bar{s}$. We have plotted the behavior of the functions $J_M(s)$, $J_+(s)$, $J_-(s)$ in Figure 13. It is clear from these plots that the modified alternating algorithm will converge to the true slowness $\bar{s}$ from any starting slowness s_0, as minimizing $J_M(s)$ yields a step Δs towards $\bar{s}$.

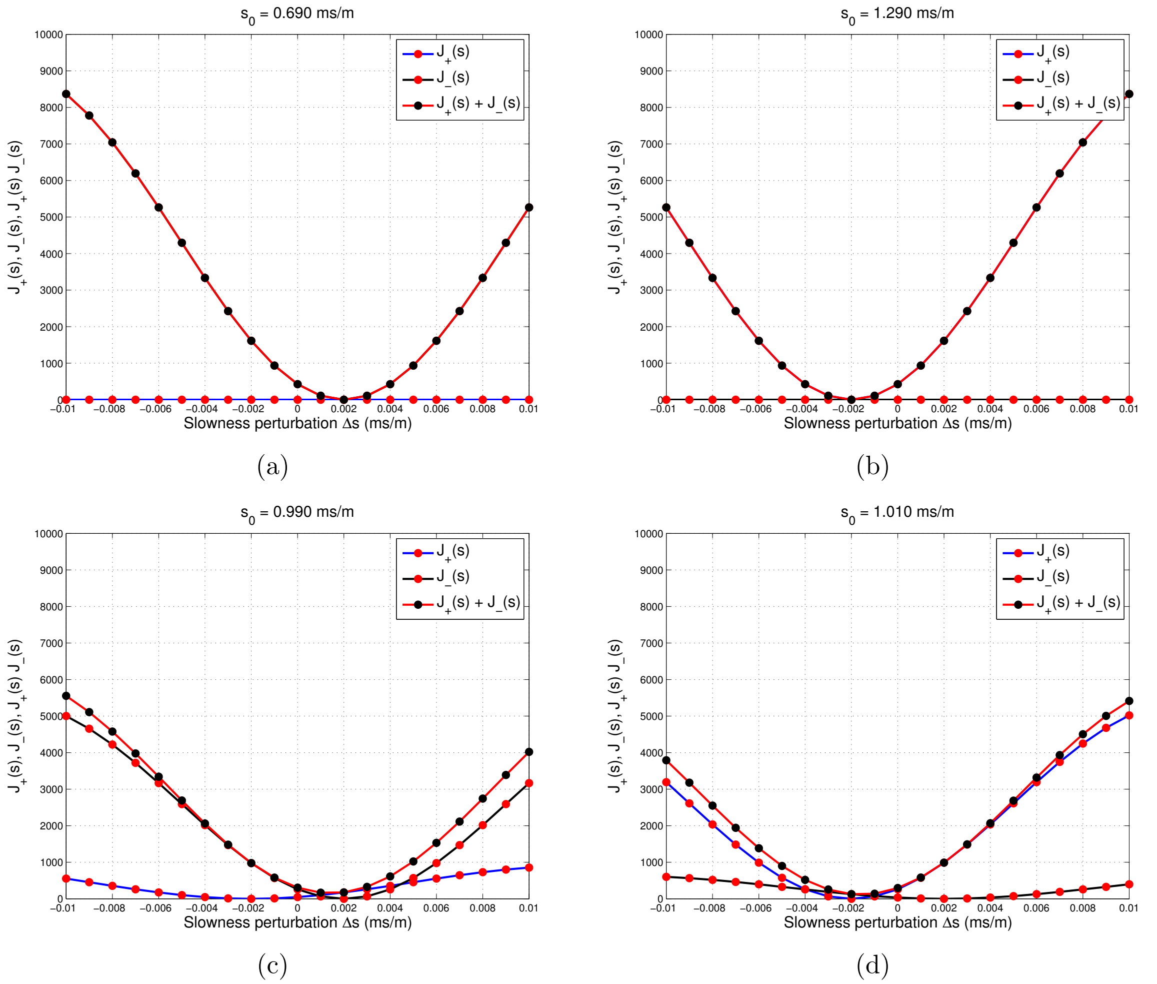
We finally note that the modified alternating algorithm converges faster than the alternating algorithm. This can be realized by noting the fact already mentioned that for s_0 sufficiently far away from $\bar{s}$, $J_M(s) = \frac{1}{2}||\mathbf{w}||_2^2$. This observation can be seen by comparing the profiles of $J_M(s)$ in Figures 13(a) and 13(b), with the profiles of $\frac{1}{2}||\mathbf{w}||_2^2$ in Figures 8(a) and 8(b). Thus in the regime when s_0 is far away from $\bar{s}$, the convergence is controlled by the behavior of the $\frac{1}{2}||\mathbf{w}||_2^2$ term. But we know that the $\frac{1}{2}||\mathbf{w}||_2^2$ term has better convergence than $J_s(s)$, and thus so does the modified alternating algorithm.

It must be mentioned here that just like it was discussed that it is possible to incorporate the idea of bigger shifts in the shift focusing operator used in the alternating algorithm, it is possible to do the same thing also with the modified alternating



(a) (b) (c) (d)

Figure 12: Plot of the composite terms appearing in equation 34, involving the mask and shift operators, with vector $\mathbf{c}$ in Figure 11(a). (a) $\mathbf{M}_0\mathbf{S}_+\mathbf{c}$ - The result is the zeroing of all samples of $\mathbf{S}_+\mathbf{c}$ after the sample at index 0. (b) $\mathbf{M}_{-1}\mathbf{c}$ - The result is the zeroing of all samples of $\mathbf{c}$ after the sample at index -1. (c) $[\mathbf{I}-\mathbf{M}_{-1}]\mathbf{S}_-\mathbf{c}$ - The result is the zeroing of all samples of $\mathbf{S}_-\mathbf{c}$ before the sample at index 0. (d) $[\mathbf{I}-\mathbf{M}_0]\mathbf{c}$ - The result is the zeroing of all samples of $\mathbf{c}$ before the sample at index 1. [**ER**]



(a)

(b)

(c)

(d)

Figure 13: Plot of the functions $J_M(s), J_+(s), J_-(s)$ for different value of starting slowness s_0 : (a) $s_0 = 0.69$ ms/m represents the case when s_0 is slow compared to $\bar{s}$, and hence $J_+(s) = 0$, (b) $s_0 = 1.29$ ms/m represents the case when s_0 is fast compared to $\bar{s}$, and hence $J_-(s) = 0$, (c) $s_0 = 0.99$ ms/m represents the case when s_0 is almost close to $\bar{s}$ on the slower side, and (d) $s_0 = 1.01$ ms/m represents the case when s_0 is almost close to $\bar{s}$ on the faster side. [**ER**]

algorithm. Doing so will increase the magnitude of the update Δs obtained at each iteration leading to faster convergence, but this will only work up to some maximum shift beyond which the global convergence property will be lost. However, the essential features of running the alternating algorithm with bigger shifts are similar to what have been presented here with the unit shift operators.

FWI-WEMVA OBJECTIVE FUNCTION

Finally we analyze the behavior of the following objective function that incorporates the focusing operator directly in the data fitting term:

$$\begin{aligned} J_{\mathrm{FW}}(s) &= \frac{1}{2}\left\|(\mathbf{I}-\mathbf{F})\,\tilde{\mathbf{L}}'(s)\left[\tilde{\mathcal{L}}(s,\mathbf{c}=0)-\mathbf{d}_r\right]\right\|_2^2 \\ &= \frac{1}{2}\left\|(\mathbf{I}-\mathbf{F})\,\tilde{\mathbf{L}}'(s)\left[\mathcal{L}(s)-\mathbf{d}_r\right]\right\|_2^2. \qquad (37) \end{aligned}$$

This objective function is related to the one presented by Symes (2008) in equation 14. One of its attractive properties is that it depends on slowness through both the data residuals $(\mathcal{L}(s)-\mathbf{d}_r)$ and the focusing of the backprojection of these residuals into the space of $\mathbf{c}$ by the operator $\tilde{\mathbf{L}}'(s)$. Our conjecture is that during the inversion process, the gradient component corresponding to the direct dependency on the data residuals introduces short wavelengths into the slowness model, whereas the gradient component corresponding to the focusing of $\mathbf{c}$ introduces long wavelengths into the slowness model. If that were the case, optimizing this objective function would have the potential of achieving simultaneous inversion of all model scales. Unfortunately, this conjecture cannot be fully tested using our simple 1D model because we assumed the slowness to be a scalar, and obviously cannot be decomposed into different scales.

Another attractive properties of the objective function in equation 37 is that, when the amplitude focusing operators $\mathbf{F}_D$ and $\mathbf{F}_G$ are used, its behavior substantially changes according to the value of the parameter τ_{W}. These changes in behavior of J_{FW} can be easily understood by analyzing the gradient of the objective function with respect to slowness. This gradient has two terms because both $\mathcal{L}$ and $\tilde{\mathbf{L}}'$ are function of the slowness. The term deriving from the dependency of $\mathcal{L}$ from s is a FWI-like gradient, whereas the one deriving from the dependency of $\tilde{\mathbf{L}}'$ from s is a WEMVA-like gradient. The total gradient can be expressed as follows:

$$\begin{aligned} \nabla J_{\mathrm{FW}} &= \underbrace{\mathbf{L}'(s)\,\tilde{\mathbf{L}}(s)\,(\mathbf{I}-\mathbf{F})'\,(\mathbf{I}-\mathbf{F})\,\tilde{\mathbf{L}}'(s)\left[\mathcal{L}(s)-\mathbf{d}_r\right]}_{\mathrm{FWI-like\ \ gradient}} \qquad (38) \\ &+ \underbrace{\left[\mathcal{L}(s)-\mathbf{d}_r\right]'\,\dot{\tilde{\mathbf{L}}}(s)\,(\mathbf{I}-\mathbf{F})'\,(\mathbf{I}-\mathbf{F})\,\tilde{\mathbf{L}}'(s)\left[\mathcal{L}(s)-\mathbf{d}_r\right]}_{\mathrm{WEMVA-like\ \ gradient}}. \qquad (39) \\ &= \nabla J_{\mathbf{F}\mathrm{W}}+\nabla J_{\mathrm{F}\mathbf{W}} \qquad (40) \end{aligned}$$

When $\tau_{\mathrm{W}} \approx 0$ the first term in the gradient ($\nabla J_{\mathbf{F}\mathrm{W}}$ in equation 40) is close to the conventional FWI gradient because $\tilde{\mathbf{L}}\tilde{\mathbf{L}}'(\mathcal{L}(s)-\mathbf{d}_r) \approx (\mathcal{L}(s)-\mathbf{d}_r)$ by virtue of the approximation in equation 6. It also dominates the gradient, because the second term ($\nabla J_{\mathbf{FW}}$ in equation 40) is small (it would be actually zero if we had not imposed the constraint of $c_0 = 0$).

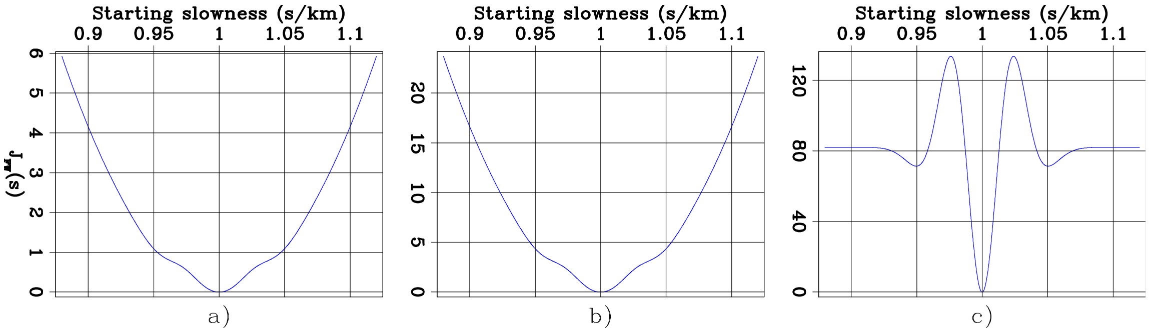
In contrast, when $\tau_{\mathrm{W}} \approx 1$, the first term in the gradient, $\nabla J_{\mathbf{F}\mathrm{W}}$, is small because the application of $(\mathbf{I}-\mathbf{F})'(\mathbf{I}-\mathbf{F})$ strongly attenuates the time lags in $\mathbf{c}$ that contribute the most to the backprojection of the residuals (see Figure 2b). Consequently the WEMVA-like term, $\nabla J_{\mathbf{FW}}$, dominates the gradient, and ensures convergence towards the global minimum.

This behavior of the objective function in equation 37 is illustrated by Figure 14 through Figure 17. Figures 14 and 15 show $J_{\mathrm{FW}}(s_o)$ for $\mathbf{F} = \mathbf{F}_D$ and $\mathbf{F} = \mathbf{F}_G$, respectively. In each figure, the three panels correspond to different values of τ_{W}: for panels a) $\tau_{\mathrm{W}} = 1.0$, for panels b) $\tau_{\mathrm{W}} = 0.5$, and for panels c) $\tau_{\mathrm{W}} = 0.0$. The objective functions in the leftmost panels are convex. A gradient-based method would have no problems to converge towards the global minimum; the convergence, however, would be slow. In the middle panels, the objective functions are "tighter" but still convex. In contrast, the objective functions plotted in the rightmost panels are oscillatory and may cause similar convergence problems as experienced in the minimization of conventional FWI objective function. On the other hand, the high sensitivity of these oscillating objective functions to small changes in slowness may also enable the inversion to achieve high resolution, once we are close enough to the correct slowness. These observations suggest the application of an iterative inversion process that starts with wide focusing operators ($\tau_{\mathrm{W}} = 1.0$) and that slowly tightens the focusing operators toward $\tau_{\mathrm{W}} = 0.0$ as the data kinematics are fitted. Such an algorithm has the potential of achieving both robust global convergence from arbitrary starting model and fast local convergence close to the desired global minimum.

Figures 16 and 17 show the FWI-like gradient term for $\mathbf{F} = \mathbf{F}_D$ and $\mathbf{F} = \mathbf{F}_G$, whereas Figures 18 and 19 show the WEMVA-like gradient term for $\mathbf{F} = \mathbf{F}_D$ and $\mathbf{F} = \mathbf{F}_G$. Notice that the FWI-like gradient term is strongly oscillatory for all values of τ_{W}, but also that its amplitude is higher than the amplitude of the WEMVA-like term only for $\tau_{\mathrm{W}} = 0$ (panels c) in the figures. In contrast, the WEMVA-like term of the gradient that is shown in panels a) and b) is well-behaved for both choices of focusing operator. However, close to convergence; that is for s_o close to $\bar{s}$, the gradient is small. If we had to rely only on this gradient-component, the resolution of the inversion would be likely to suffer. In panels c) the WEMVA-like term becomes oscillatory and has "wrong" sign even close to convergence, but its amplitude insignificant compared to the amplitude of the corresponding FWI-like gradient terms.

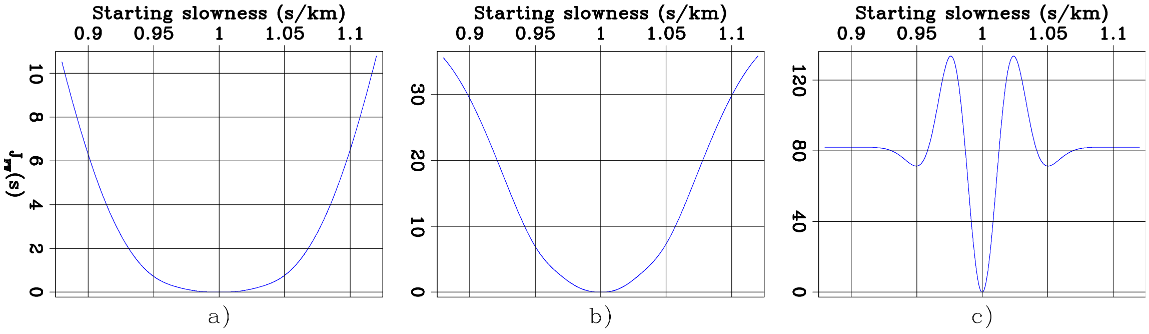
We can also observe the WEMVA-like gradients in panels a) and b) are smoother when $\mathbf{F} = \mathbf{F}_G$ (Figure 19) than when $\mathbf{F} = \mathbf{F}_D$ (Figure 18). This difference may be indicative of a difference in robustness between the two focusing operators. To test this hypothesis we conducted a a similar test, but with a zero-phase wavelet in place of the Ricker-derived wavelet shown Figure 1a. The zero-phase wavelet has the same central frequency as the Ricker-derived wavelet, but it is more ringing. Consequently the objective functions and gradients are more oscillatory than the ones shown in previous figures.

Figure 20a shows the objective function computed using the highly-ringing zero-phase wavelet when $\mathbf{F} = \mathbf{F}_D$ and $\tau_{\mathrm{W}} = 1.0$; it corresponds to the objective function computed using the Ricker-derived wavelet and shown in Figure 14a. With this new wavelet, the objective function is not any more convex, and the total (FWI-like plus WEMVA-like term) gradient (Figure 20b) has two zero-crossing on each side of $\bar{s}$. On the contrary, the objective function computed with $\mathbf{F} = \mathbf{F}_G$ (Figure 21a) is still convex; its total gradient (Figure 21b) gets close to the horizontal axis, but it does not cross it, except at the expected zero-crossing at $s_o = \bar{s}$. This difference in behavior can be explained by comparing the WEMVA-like gradient terms (Figure 20c and Figure 21c). The one computed with $\mathbf{F} = \mathbf{F}_G$ is smoother than the one computed with $\mathbf{F} = \mathbf{F}_D$.



a) b) c)

Figure 14: $J_{\mathrm{FW}}(s_o)$ computed with $\mathbf{F} = \mathbf{F}_D$ and with: a) $\tau_{\mathrm{W}} = 1.0$, b) $\tau_{\mathrm{W}} = 0.5$, and c) $\tau_{\mathrm{W}} = 0.0$. [**ER**]



a) b) c)

Figure 15: $J_{\mathrm{FW}}(s_o)$ computed with $\mathbf{F} = \mathbf{F}_G$ and with: a) $\tau_{\mathrm{W}} = 1.0$, b) $\tau_{\mathrm{W}} = 0.5$, and c) $\tau_{\mathrm{W}} = 0.0$. [**ER**]

DISCUSSIONS

All the three approaches that we presented to solve the extended inverse problem show promises to lead to inversion algorithms with robust global convergence. However,

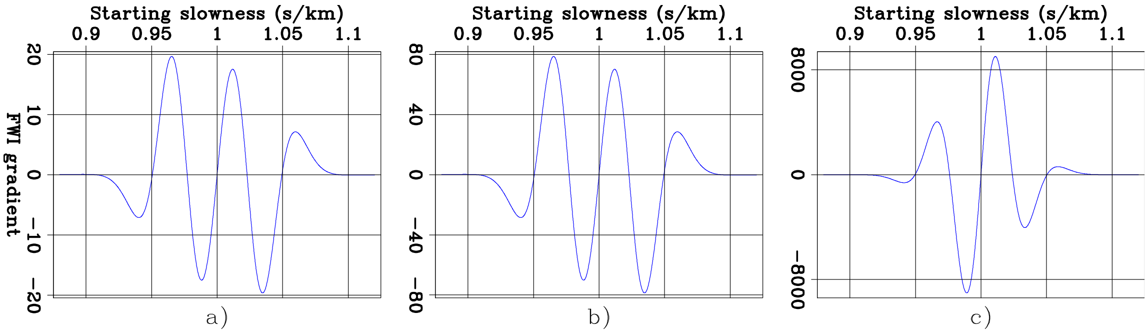


Figure 16: $\nabla J_{\mathbf{F}\mathrm{W}}(s_o)$ computed with $\mathbf{F} = \mathbf{F}_D$ and with: a) $\tau_{\mathrm{W}} = 1.0$, b) $\tau_{\mathrm{W}} = 0.5$, and c) $\tau_{\mathrm{W}} = 0.0$. [**ER**]

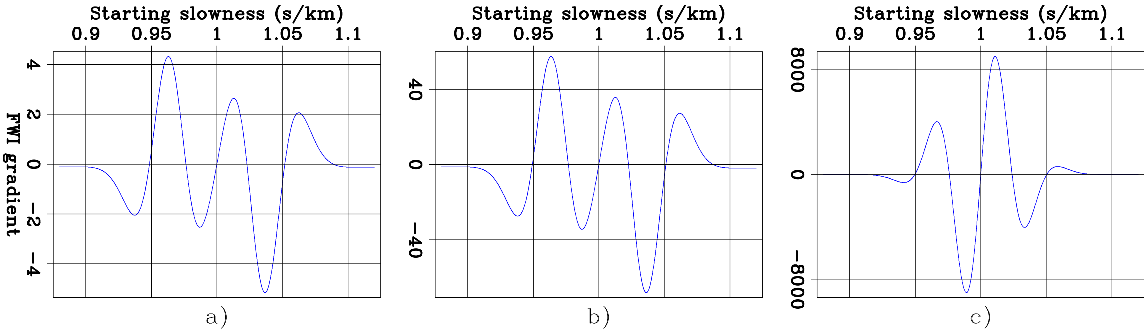


Figure 17: $\nabla J_{\mathbf{F}\mathrm{W}}(s_o)$ computed with $\mathbf{F} = \mathbf{F}_G$ and with: a) $\tau_{\mathrm{W}} = 1.0$, b) $\tau_{\mathrm{W}} = 0.5$, and c) $\tau_{\mathrm{W}} = 0.0$. [**ER**]

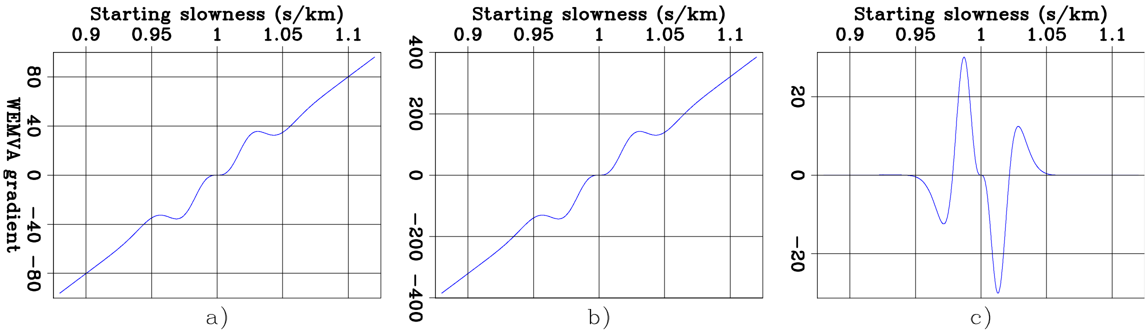
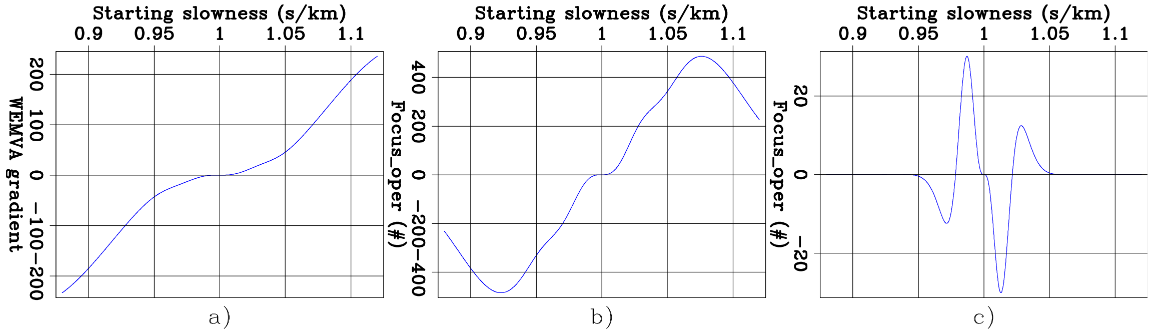


Figure 18: $\nabla J_{\mathrm{F}\mathbf{W}}(s_o)$ computed with $\mathbf{F} = \mathbf{F}_D$ and with: a) $\tau_{\mathrm{W}} = 1.0$, b) $\tau_{\mathrm{W}} = 0.5$, and c) $\tau_{\mathrm{W}} = 0.0$. [**ER**]

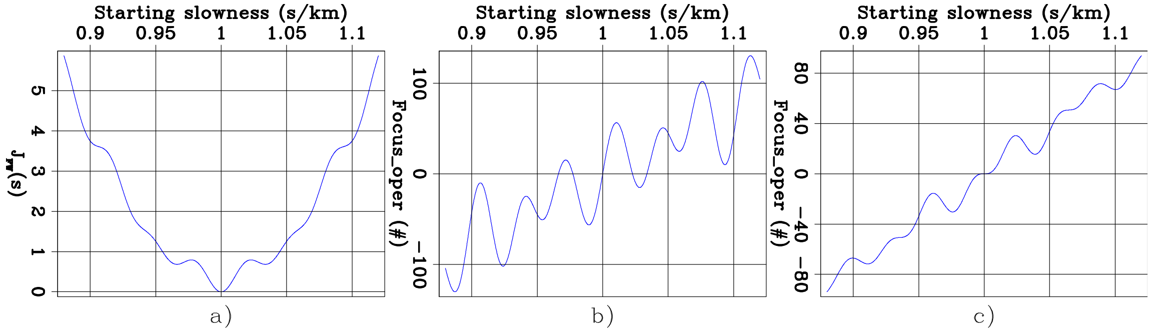


a)

b)

c)

Figure 19: $\nabla J_{\mathrm{F}\mathbf{W}}\left(s_o\right)$ computed with $\mathbf{F} = \mathbf{F}_G$ and with: a) $\tau_{\mathrm{W}} = 1.0$, b) $\tau_{\mathrm{W}} = 0.5$, and c) $\tau_{\mathrm{W}} = 0.0$. [**ER**]

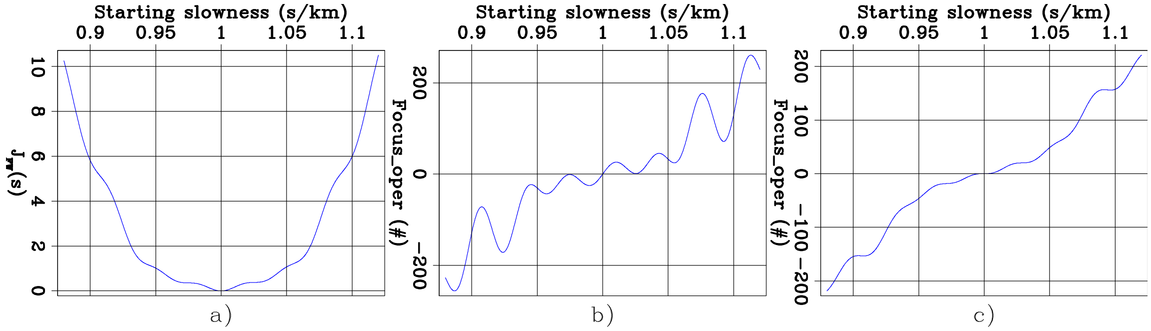


a)

b)

c)

Figure 20: a)$J_{\mathrm{FW}}\left(s_o\right)$ computed with a zero-phase wavelet that was more ringing than the Ricker-derived wavelet used for the previous figures. b) $\nabla J_{\mathrm{FW}}\left(s_o\right)$, and c) $\nabla J_{\mathrm{F}\mathbf{W}}\left(s_o\right)$. All these three curves were computed with $\mathbf{F} = \mathbf{F}_D$ and $\tau_{\mathrm{W}} = 1.0$. [**ER**]



a)

b)

c)

Figure 21: a)$J_{\mathrm{FW}}\left(s_o\right)$ computed with a zero-phase wavelet that was more ringing than the Ricker-derived wavelet used for the previous figures. b) $\nabla J_{\mathrm{FW}}\left(s_o\right)$, and c) $\nabla J_{\mathrm{F}\mathbf{W}}\left(s_o\right)$. All these three curves were computed with $\mathbf{F} = \mathbf{F}_G$ and $\tau_{\mathrm{W}} = 1.0$. [**ER**]

we have not sufficiently developed our analysis to draw firm conclusions on the rate of convergence of the proposed methods. Since the development of efficient inversion algorithms is one of our main goals, further work in this direction is needed.

The 1D wave-propagation problem we used to analyze the proposed inversion methods has two useful advantages: 1) it enables comprehensive analysis of global convergence because it is computationally fast and 2) objective functions and gradients can be analyzed as simple 1D plots. However, it has also two (related to each other) main shortcomings: 1) it models transmitted events but not reflected ones, and 2) its model space (a simple scalar) cannot be decomposed into different scales, and thus does not enable insights on how different model scales (long vs. short wavelengths) behave during the inversion process.

REFERENCES

Almomin, A. and B. Biondi, 2014, Preconditioned tomographic full waveform inversion by wavelength continuation: SEG Technical Program Expanded Abstracts, **33**, 944–948.

Biondi, B. and A. Almomin, 2014, Simultaneous inversion of full data bandwidth by tomographic full waveform inversion: Geophysics, **79**, WA129–WA140.

Biondi, B. and P. Sava, 1999, Wave-equation migration velocity analysis: SEG Technical Program Expanded Abstracts, **18**, 1723–1726.

Sava, P. and B. Biondi, 2004, Wave-equation migration velocity analysis-I: Theory: Geophysical Prospecting, **52**, 593–606.

Shen, P. and W. W. Symes, 2008, Automatic velocity analysis via shot profile migration: Geophysics, **73**, VE49–VE59.

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

Symes, W. W. and J. J. Carazzone, 1991, Velocity inversion by differential semblance optimization: Geophysics, **56**, 654–663.

Zhang, Y. and B. Biondi, 2013, Moveout-based wave-equation migration velocity analysis: Geophysics, **78**, U31–U39.