Algorithmic Strategies for Full Waveform Inversion: 1D Experiments

Carsten Burstedde and Omar Ghattas, Jackson School of Geosciences and Institute for Computational Engineering and Sciences, The University of Texas at Austin*

SUMMARY

The full waveform, output least squares seismic inversion method is in principle capable of automated operation and high resolution. However, a number of difficulties arise when it is applied to highly-resolved model parameterizations and high frequency sources. First, the least squares functional suffers from spurious local minima, which necessitates an accurate initial guess of the low-wavenumber background. Second, total variation regularization methods that are used to resolve sharp interfaces create significant numerical difficulties. Third, inclusion of bound constraints on continuous model parameters presents considerable difficulties for commonly-used active set or interior point methods. Finally, common gradient-based optimization methods have difficulties scaling to the large-scale parameter spaces that result when the continuous fields are discretized. Here we introduce an optimization strategy that incorporates a number of techniques addressing these four difficulties, including objective reformulation and time-window continuation, primal-dual methods for treatment of inequalities and regularization, and inexact matrix-free Newton-Krylov optimization. Examples demonstrate good performance for a "100 wavelength" 1D inverse problem

INTRODUCTION

Seismic inversion via least-squares optimization using the full acoustic or elastic wave equation remains a challenge, particularly for high frequency problems. The difficulties include:

Multiple minima. The least squares misfit objective function is oscillatory in directions associated with wavenumber components of the model that are longer than seismic wavelengths (i.e. smooth components) (Symes and Carazzone (1991)). The common practice of grid-and-frequency continuation (Bunks et al. (1995)) can be effective at avoiding local minima, but only when the observations contain sufficient low frequency content to initialize the continuation process.

Ill-posedness. The objective function is relatively insensitive to wavenumber components of the model that are shorter than seismic wavelengths (i.e. rough components). Tikhonov regularization addresses this ill-posedness, but blurs interfaces within layered media. Total variation (TV) regularization (Acar and Vogel (1994)) are very effective at resolving sharp interfaces, but creates substantial numerical difficulties.

Variational inequalities. When infinite-dimensional inequalities associated with bound constraints on continuous model parameters are discretized on fine grids, very large systems of inequalities result. Active set methods (Nocedal and Wright, 1999, Ch.16) can suffer from combinatorial complexity and have difficulty scaling as the grid is refined; interior point methods (Conn et al. (2000)) often have better scaling properties, but good preconditioners for solution of the resulting optimality systems remain elusive.

Scalable optimization algorithms. The severe nonlinearities make the least squares optimization problem difficult to solve. The use of gradient methods such as nonlinear conjugate gradients, in combination with adjoint computation of the gradient, results in a method that is cheap to apply at each iteration (Tarantola (1986)) yet is only linearly convergent (Nocedal and Wright, 1999, Ch.5). On the other hand, Newton's method is quadratically convergent (Nocedal and Wright, 1999, Ch.6) but requires solution of a dense system of equations at each iteration (Pratt et al. (1998)), which can be intractable for fine discretizations of continuous model parameters.

Here, we incorporate a number of techniques that address one of the difficulties above. To help mitigate multiple minima, we augment the least squares misfit objective with a new convexifying term that involves moments in time, in combination with a continuation on length of the time window. We appeal to recent work from image processing on primal-dual TV regularization (Chan et al. (1999)) and from optimal control theory on primal-dual active set strategies for bound inequalities (Hintermüller et al. (2003)), both of which demonstrate superior scalability relative to conventional methods. Finally, we employ an inexact matrix-free Newton-Krylov method, which combines the curvature-exploiting properties of Newton's method with a computational structure resembling a gradient-only method; these methods have demonstrated grid-independent convergence for 2D and 3D highly-resolved inverse wave propagation problems (Akcelik et al. (2002, 2003)). Experiments on 1D problems illustrate that the combination of these ideas results in a fast, scalable, and robust method that performs well on a "100 wavelength" inverse problem.

THEORY AND METHOD

Problem Setting

Let $u(\mathbf{x},t)$ be the wavefield in space $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ and time $t \in (0,T)$, and $f(\mathbf{x},t)$ the force corresponding to one or more seismic sources, related by the (partial differential) forward operator $\mathscr{L}u = f$. The discretizations of u and f shall be denoted by boldface $\mathbf{u} \in \mathbb{R}^{KN_w}$ and $\mathbf{f} \in \mathbb{R}^{KN_w}$, where N_w is generally a multiple of the number of space grid points and K the number of discrete time values. The particular (finite difference or finite element) discretization used should of course be suited to the forward modeling operator, which takes the form $\mathbf{Lu} = \mathbf{f}$. The matrix $\mathbf{L} \equiv \mathbf{L}_\mu$ is assumed to depend (nonlinearly) on the model parameters $\mu \in \mathbb{R}^{N_p}$. These can be any combination of density, seismic velocities, elasticity, or attenuation parameters.

We choose to describe the optimization formulation and algo-

rithms at the finite-dimensional level, the implementation of which guarantees consistency of gradient computations (Gunzburger (2003)). The optimization problem is:

$$\min_{\mathbf{u},\mu} J(\mathbf{u},\mu) := F(\mathbf{u}) + \alpha R(\mu) \quad \text{s.t.} \quad \mathbf{L}_\mu \mathbf{u} = \mathbf{f}, \qquad (1)$$

and additional bound constraints on model parameters. The first term $F(\mathbf{u})$ defines the measure of similarity between the synthetics and the data u^*. For output-least squares, is a numerical quadrature formula approximating the integral

$$F_{\text{ls}}(\mathbf{u}) \approx \frac{1}{2}\sum_r \int_0^T \|u^*(\mathbf{x}_r,t) - u(\mathbf{x}_r,t)\|^2 \mathrm{d}t,$$

where $\mathbf{x}_r$ are the receiver locations. This choice can lead to highly oscillatory function behavior, and we use several ideas to address this difficulty. We propose a variation of this term as follows,

$$F_{\text{mo},p}(\mathbf{u}) \approx \frac{1}{2}\sum_r \left[\int_0^T \big(u^*(\mathbf{x}_r,t) - u(\mathbf{x}_r,t)\big)(t-t_0)^p \mathrm{d}t\right]^2,$$

which realizes moments of order p in time about the instant t_0. As here the square is outside the integral, this is weaker than the least-squares approach, so we generally combine the two up to some order $p_{\max}$,

$$F = F_{\text{ls}} + \sum_{p=0}^{p_{\max}} \beta_p F_{\text{mo},p}. \qquad (2)$$

In (1), $R(\mu)$ is the regularization term. We examine two possibilities, Tikhonov and total variation (TV), which in the case of a scalar material parameter μ read

$$R_2(\mu) \approx \frac{1}{2}\int_\Omega \|\nabla\mu\|^2 \mathrm{d}x \qquad \text{for Tikhonov,}$$

$$R_{1,\varepsilon}(\mu) \approx \int_\Omega \sqrt{\|\nabla\mu\|^2 + \varepsilon}\, \mathrm{d}x \qquad \text{for TV.}$$

The top expression generally leads to a smoothing of the model, while the second allows steep jumps, yet is numerically more difficult to handle (Chan et al. (1999)). Again, we write the $\approx$ sign here since R_2 and $R_{1,\varepsilon}$ operate on the discrete representation of μ.

Optimization Method

To derive optimality conditions for the constrained minimization problem (1), we define a Lagrange multiplier, or adjoint wavefield, $\mathbf{p} \in \mathbb{R}^{KN_w}$ and form the Lagrangian function

$$L(\mathbf{p},\mathbf{u},\mu) := J(\mathbf{u},\mu) + \mathbf{p}^T(\mathbf{L}_\mu \mathbf{u} - \mathbf{f}).$$

Setting the gradient of the Lagrangian, $\nabla L_{\mathbf{p},\mathbf{u},\mu}$, to zero yields the state and adjoint equations,

$$\mathbf{L}\mathbf{u} = \mathbf{f}, \qquad \mathbf{L}^T\mathbf{p} = -\nabla F(\mathbf{u}), \qquad (3)$$

and the reduced gradient in μ,

$$\mathbf{g} := \nabla_\mu L = \mathbf{p}^T\left[\mathscr{D}_\mu \mathbf{L}(\cdot)\right]\mathbf{u} + \alpha\nabla R. \qquad (4)$$

Here we have introduced the differential as

$$\mathscr{D}_{\mathbf{x}} f(\Delta\mathbf{x}) := \lim_{h\to 0} \frac{f(\mathbf{x}+h\Delta\mathbf{x}) - f(\mathbf{x})}{h}.$$

Eliminating state and adjoint wavefields $(\mathbf{u},\mathbf{p})$ by solving the state and adjoint equations in (3) and inserting into (4) with $\alpha = 0$ gives the expression for the reduced gradient,

$$\mathbf{g} = -\nabla F^T \mathbf{L}^{-1}\left[\mathscr{D}_\mu \mathbf{L}(\cdot)\right]\mathbf{L}^{-1}\mathbf{f},$$

which is the familiar Born approximation Fréchet kernel. Note also that the transpose $\mathbf{L}^T$ in (3) naturally leads to a reverse-time propagation of the adjoint variable. Thus, the gradient can be computed by solving for $\mathbf{u}$ and $\mathbf{p}$ in one forward and one adjoint computation, and assembling it through the inner products in (4). Gradient-based optimization methods such as nonlinear CG and limited memory BFGS can be implemented at this point; see e.g. Nocedal and Wright (1999).

Unfortunately, the rate of convergence of gradient-based methods is only linear, and one requires Hessian information to improve the rate to (asymptotically) quadratic. The resulting Newton's method takes the form

$$\Delta\mu^{(k)} := \mu^{(k+1)} - \mu^{(k)} = -(\nabla^2_\mu L)^{-1}\nabla L_\mu\,\big|_{\mu^{(k)}}.$$

The reduced Hessian matrix $\mathbf{H} := \nabla^2_\mu L$ is obtained by differentiating (4) and using the chain rule on the derivatives of $\mathbf{p}$ and $\mathbf{u}$ with respect to μ computed from (3). The full expression reads

$$\begin{aligned}\mathbf{H} = \quad & \mathbf{E}_{\mathbf{u}}^T\mathbf{L}^{-T}(\nabla^2 F)\mathbf{L}^{-1}\mathbf{E}_{\mathbf{u}} + \alpha\nabla^2 R \\ & + \mathbf{E}_2 - \mathbf{E}_{\mathbf{u}}^T\mathbf{L}^{-T}\mathbf{E}_{\mathbf{p}} - \mathbf{E}_{\mathbf{p}}^T\mathbf{L}^{-1}\mathbf{E}_{\mathbf{u}}\end{aligned}$$

where the first line contains the so-called Gauss-Newton approximation, and the second line the remaining second-order terms. Here we have recast the differentials as matrices acting on $\Delta\mu$,

$$\begin{aligned}\mathbf{E}_{\mathbf{u}}\,\Delta\mu &:= \left[\mathscr{D}_\mu \mathbf{L}(\Delta\mu)\right]\mathbf{u}, \\ \mathbf{E}_{\mathbf{p}}\,\Delta\mu &:= \left[\mathscr{D}_\mu \mathbf{L}^T(\Delta\mu)\right]\mathbf{p}, \\ \Delta\mu^T\mathbf{E}_2\,\Delta\mu' &:= \mathbf{p}^T\left[\mathscr{D}^2_\mu \mathbf{L}(\Delta\mu,\Delta\mu')\right]\mathbf{u}.\end{aligned}$$

This form makes it clear that the Hessian $\mathbf{H}$ is a dense matrix that is far too costly to store and factor for fine model discretizations. The remedy here is to resort to iterative Krylov solvers for the Newton system

$$\mathbf{H}\Delta\mu = -\mathbf{g}, \qquad (5)$$

in particular (preconditioned) conjugate gradients (CG) and the Steihaug variant. In the inexact Newton-Krylov setting, CG is terminated early to (1) avoid directions of negative curvature, which guarantees a descent direction and thus global convergence with a suitable line search; (2) prevent "oversolving" when far from the minimum; and (3) remain inside the trust region (if this globalization variant is used). See Nocedal and Wright (1999) for details. Moreover, since CG requires only matrix-vector products, the method can be implemented in a matrix-free manner, and $\mathbf{H}$ is never stored. Instead, Hessian-vector products are formed on-the-fly at each CG iteration, each of which requires a total of three state or adjoint wave propagation solutions. An advantage of this approach is its generality: tuning of algorithmic parameters and stopping criteria alone can realize almost any tradeoff between cost per nonlinear iteration and convergence rate, creating a continuum

of methods from basic steepest descent (one CG iteration per Newton step) to full Newton (full CG convergence).

There remain several important parameters to choose, such as β_p, α, and ε, and tolerances hidden inside the algorithms. Also, upper and lower bounds on models parameters must be enforced (for simplicity we left the bounds out of the discussion of optimality conditions). As mentioned in the Introduction, we enforce bounds via an adaptation of a primal-dual active set strategy (Hintermüller et al. (2003)), which allows for faster identification of the correct active set, and can exhibit mesh-independent convergence.

EXAMPLES

Preliminaries

The basis of the computational experiments described here is a set of density and velocity data acquired from a borehole. The measurements cover a depth from $z_0 = 170$m to $z_1 = 1051$m, thus a depth range of $l = z_1 - z_0 = 881$m, with data points placed at 12.5cm intervals. See Figure 1 for plots of the density $\rho(z)$ and the elasticity modulus $\mu(z)$ inferred from the compressional wave speed.

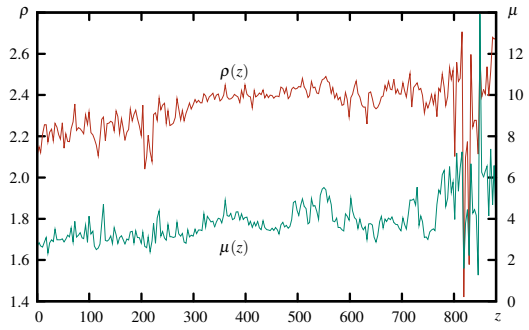

Figure 1: Density $\rho(z)$ and elasticity parameter $\mu(z)$ filtered to $\Delta z = 3.4$m intervals. The units are g/cm^3 for ρ, and 4.68×10^9 Pa for μ. The compressional wave speeds range from 2160m/s to 5610m/s.

We use the one-dimensional wave equation without attenuation in the time domain,

$$\mathscr{L}u = \rho \frac{\partial^2}{\partial t^2} u + \frac{\partial}{\partial z} \mu \frac{\partial}{\partial z} u.$$

The operator is discretized with piecewise linear finite elements on the interval $\Omega = (0, l)$, with free surface conditions on the top end and absorbing boundary conditions at $z = l$. The source and receiver are located at the top position $z = 0$, and the source time function is modeled by a Ricker wavelet whose central frequency will be stated for each experiment.

While the well data are real, we need to generate the receiver signal $u^*(z = 0, t)$ within our code as a preprocessing step. The more model detail is taken into account in this precomputation, the more faithful (and relevant) the whole experiment is to the physical reality. On the other hand, a detailed target model leads to a strongly varying receiver signal and hence to a large probability that local minima are encountered in the numerical inversion. In the following experiments, we have strived to push the level of realism to the convergence limits of our inversion method.

The key to the success of the inversions described here is a continuation principle, which successively refines the knowledge about the model and leads to a sequence of guesses. In all our experiments, this incremental process has been found to be more reliable in the presence of local minima than a single inversion from an initial guess. The continuation principle can be applied to various parameters, such as time horizon, regularization parameters and source frequency. We also switch the optimization algorithm while iterating, from a Gauss-Newton method initially to a full Newton's method when close to converging.

Experiment A

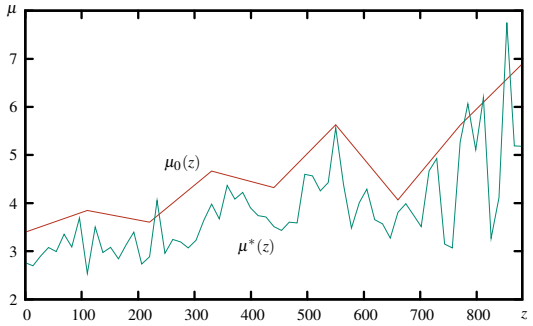

Figure 2: Initial guess and target for Experiment A. The target is resolved at $\Delta z = 13.8$m resolution. The initial guess is chosen very coarse and 20% too high (that is 10% too fast) on average.

The purpose of this first experiment is to demonstrate the continuation principle and to point out the differences between Tikhonov and total variation (TV) regularization. We use a source frequency of 25Hz and a receiver signal generated on a grid comprising 64 cells. The time horizon T is initially chosen such that the model is illuminated only in the top region, and then successively elongated. Once the depth is fully covered, the regularization parameter is decreased in several steps. The initial guess has been deliberately chosen very coarse and consistently too high by 20%, which makes for a difficult situation due to the large shifts of the arrival times. Target and initial guess are shown in Figure 2.

Figure 3 shows the results of the inversion for Tikhonov regularization. The values of T and α used were $T = .3$s, $T = .5$s, $T = .7$s all at $\alpha = 10^{-3}$, $T = 1$s at $\alpha = 10^{-4}$, and $T = 1.6$s at $\alpha = 10^{-4}$ to $\alpha = 10^{-7}$. It can be seen that the model is refined from top to bottom, the final iterate being in excellent agreement with the target. The analogous plot for TV is given in Figure 4. The main difference is the blocky step structure of the results, which is characteristic for this type of regularization, and may be preferable depending on the application. The convergence of this method can be improved by using time moment observation (2) with $p_{\max} = 3$. This increased the tolerance for erroneous initial guesses from 10% to 15%, which is then close to Tikhonov regularization (20%).

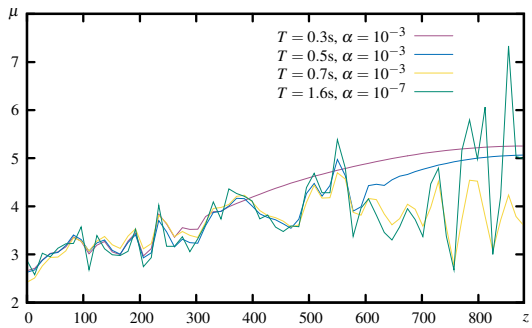

Figure 3: Selected stages of the inversion in Experiment A at 25Hz with Tikhonov regularization. The relative L_2 error of the final iterate (green) against the target in Figure 2 is 2% .

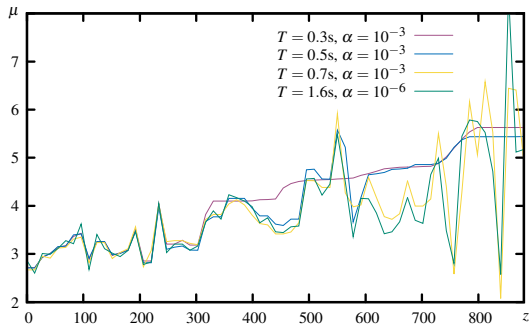

Figure 4: Selected stages of the inversion in Experiment A with TV regularization, obtained with time moment observation. The L_2 error of the final iterate (green) is about 3%.

Experiment B

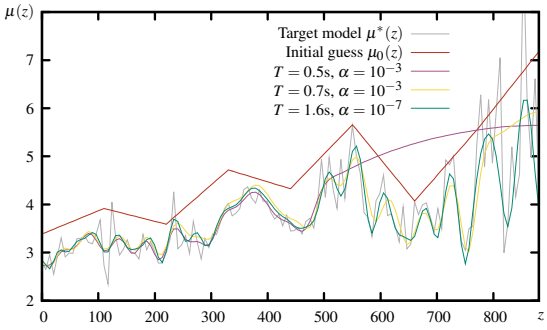

Figure 5: The first stage of Experiment B uses a grid resolution of $\Delta z = 6.9$m, time moment observation and an initial frequency of 15Hz. Hence, the result (green) is smoother than that in Figure 3.

In our second experiment, we push the limit of resolution in a two-stage process. We begin with a target which is resoled twice as high as before. The highest frequency to converge here was 10Hz for least-squares observation, and could be improved to 15Hz by activating time-moment observation. We then apply time horizon and Tikhonov regularization continuation as in Experiment A. Plots of the initial guess, the target model and some inversion iterates are provided in Figure 5.

Using the result of the first stage as initial guess, we increase both frequency and grid resolution from $\Delta z = 3.4\text{m}/f = 25$Hz over $\Delta z = 1.7\text{m}/f = 50$Hz to $\Delta z = 86\text{cm}/f = 100$Hz. The time-moment observation is switched off in the second stage. The results are shown in Figure 6. The final iterate indeed captures every detail of the target down < 10m.

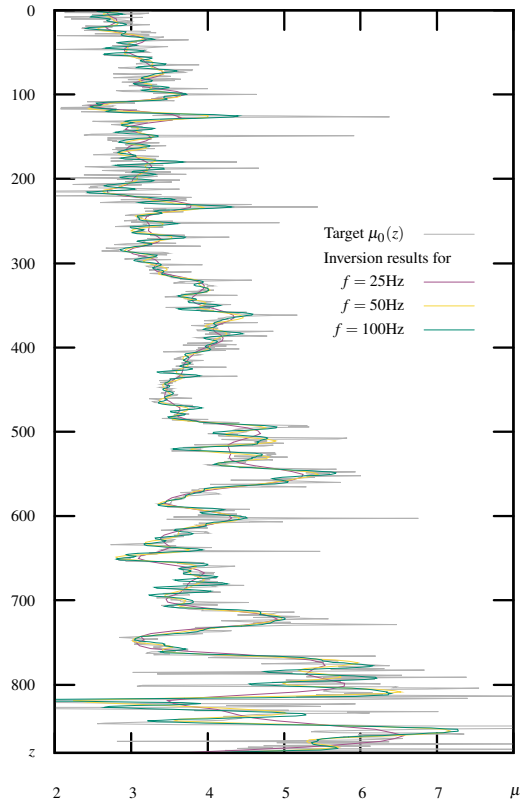

Figure 6: This high-detail result (green) is achieved through three frequency continuation steps, starting from the first-stage result in Figure 5. The central frequency of the final Ricker wavelet is 100Hz.

CONCLUSIONS

In this paper, we formulate a flexible and scalable strategy for waveform inversion. To mitigate multiple minima, a time-moment observation term is introduced, and the time window is elongated successively. Additional techniques are implemented to handle ill-posedness and inequality constraints. We employ an inexact Krylov-Newton optimization method for its scalability and grid-independent convergence. The method is then applied to 1D well data. Using pulse frequencies of up to 100Hz, we can recover the target model to high detail.

ACKNOWLEDGMENTS

The authors would like to thank Apache Corporation for their cooperation in providing the data and the permission to publish this work, and Sergey Fomel, Mark Tomasso and Georg Stadler at UT for extended discussion. NSF grant EAR-0326449 is gratefully acknowledged.

EDITED REFERENCES

Note: This reference list is a copy-edited version of the reference list submitted by the author. Reference lists for the 2007 SEG Technical Program Expanded Abstracts have been copy edited so that references provided with the online metadata for each paper will achieve a high degree of linking to cited sources that appear on the Web.

REFERENCES

Acar, R. and C. Vogel, 1994, Analysis of bounded variation penalty methods for ill-posed problems: Inverse Problems, 10, 1217–1229.

Akcelik, V., J. Bielak, G. Biros, I. Epanomeritakis, A. Fernandez, O. Ghattas, E. J. Kim, J. Lopez, D. O'Hallaron, T. Tu, and J. Urbanic, 2003, High-resolution forward and inverse earthquake modeling on terascale computers: Presented at the IEEE/ACM SC2003 Conference.

Akcelik, V., G. Biros, and O. Ghattas, 2002, Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation: Presented at the IEEE/ACM SC2002 Conference.

Bunks, C., F. Saleck, S. Zaleski, and G. Chavent, 1995, Multi-scale seismic waveform inversion: Geophysics, 50, 1457–1473.

Chan, T., G. Golub, and P. Mulet, 1999, A nonlinear primal-dual method for total variation-based image restoration: SIAM Journal on Scientific Computing, 20, 1964–1977.

Conn, A., N. Gould, and P. Toint, 2000, Trust-region methods: SIAM.

Gunzburger, M. D., 2003, Perspectives in flow control and optimization: SIAM.

Hintermuller, M., K. Ito, and K. Kunisch, 2003, The primal-dual active set strategy as a semi-smooth Newton method: SIAM Journal on Optimization, 13, 865–888.

Nocedal, J. and S. J. Wright, 1999, Numerical optimization: Springer Verlag.

Pratt, R. G., C. Shin, and G. J. Hicks, 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, 133, 341–362.

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

Tarantola, A., 1986, A strategy for nonlinear elastic inversion of seismic reflection data: Geophysics, 51, 1893–1903.