Modified Gauss-Newton with sparse updates

Xiang Li, Aleksandr Aravkin, Tristan van Leeuwen, and Felix J.Herrmann*

Seismic Laboratory for Imaging and Modeling - The University of British Columbia

Abstract

Full-waveform inversion (FWI) is a data fitting procedure that relies on the collection of seismic data volumes and sophisticated computing to create high-resolution models. With the advent of FWI, the improvements in acquisition and inversion have been substantial, but these improvements come at a high cost because FWI involves extremely large multi-experiment data volumes. The main obstacle is the 'curse of dimensionality' exemplified by Nyquist's sampling criterion, which puts a disproportionate strain on current acquisition and processing systems as the size and desired resolution increases. In this paper, we address the 'curse of dimensionality' by using randomized dimensionality reduction of the FWI problem, coupled with a modified Gauss-Newton (GN) method designed to promote curvelet-domain sparsity of model updates. We solve for these updates using the spectral projected gradient method, implemented in the SPGℓ_1 software package. Our approach is successful because it reduces the size of seismic data volumes without loss of information. With this reduction, we can compute Gauss-Newton updates with the reduced data volume at the cost of roughly one gradient update for the fully sampled wavefield.

Introduction

As we reported in earlier work (Li and Herrmann, 2010), the cost of computing gradient and Newton updates is one of the major impediments preventing the successful application of FWI to industry-size data volumes. The cost of computing the gradient depends on the size of the data and on the discretization of the Helmholz operator, while Newton updates are difficult because the Hessian of FWI is dense and possibly indefinite (negative eigenvalues). Finally, FWI is both overdetermined (there are more datums than unknowns), and underdetermined, (there is a finite aperture and hence 'shadow zones'), so FWI requires prior information (Symes, 2008).

To address these issues, the classic nonlinear least squares FWI formulation (Tarantola, 1984; Pratt et al., 1998) is regularized by an attractor term, which penalizes the two-norm difference between an initial model guess for the model and the estimated model, or by total variation, which penalizes fluctionals to bring out discontinuities (Akcelik et al. (2002); Virieux and Operto (2009)). In contrast to these methods, we do not append any extra terms to the FWI objective, but instead we regularize the model updates to be sparse in the curvelet frame (Herrmann et al., 2008), which is known to provide compressible representations of models that are smooth except for discontinuities/wavefronts along piece-wise smooth curves. By promoting curvelet-domain sparsity for the updates, we exploit theoretically optimal decay for the magnitude-sorted curvelet coefficients for velocity distributions that contain singularities (modeled by zero-, first, and fractional-order discontinuities (Herrmann, 2003; Herrmann et al., 2001)) with conflicting dips. These updates are also compressible in the curvelet domain because they are the result of multidimensional correlations between the source and residual wavevields and curvelets are known to represent wavefields sparsely (Candes and Demanet, 2004).

To use these two properties, we modify the standard Gauss-Newton subproblem (Nocedal and Wright, 1999) of FWI by adding a sparsity-promoting ℓ_1-constraint. These sparsity-promoting subproblems are solved using the SPGℓ_1 program (SPGℓ_1 - Berg and Friedlander, 2008), which is implemented in a matrix-free manner specifically designed for large-scale problems. Aside from serving as a powerful prior, our sparsity-promoting modification also connects to recent insights from Compressive Sensing (CS in short throughout the paper, Candès et al., 2006; Donoho, 2006), where it is shown that compressible signals can be recovered from severely sub-Nyquist sampling rates by solving a sparsity promoting (ℓ_1) program. The key contribution of CS is that it provides precise and concrete design principles and theoretical performance estimates for one-norm based recovery, which have let to major paradigm shift in signal/image processing (see e.g. , the IEEE special issue on CS) or (Herrmann, 2010). As shown in Herrmann et al. (2009b), simultaneous/continuous acquisition (Beasley, 2008; Berkhout, 2008), can be seen as instances of CS. Apart from having an impact on seismic acquisition, we use this identification to make computations more efficient in the context of wavefield simulation (Neelamani et al., 2008; Herrmann et al., 2009b), imaging (Romero et al., 2000; Herrmann and Li, 2011), and FWI (Krebs et al., 2009; Herrmann et al., 2009a). In these approaches, conventional sequential impulsive sources are replaced by a limited number of simultaneous 'phase-encoded' sources, which reduces the number of right-hand sides, and hence the computational complexity of wavefield simulations and on-the-fly computations of the gradient and reduced Hessian of FWI.

In random phase-encoding, where all sources fire simultaneously with random weights, coherent crosstalk turns into Gaussian noise, as in CS. This opens the possibility to approximate Gauss-Newton updates by solving a sparsity-promoting program on a reduced system. We gain provided costs of solving the sparsity-promoting

program are smaller then the cost of computing GN updates for the complete system (Li and Herrmann, 2010). In this way, we overcome the problem of prohibitive costs of computing GN updates by using simultaneous sources as a dimensionality reduction technique, randomly resampling the sources at each linearization. By drawing new simultaneous sources after each linearization, we remove possible bias and thus increase the accuracy of our solution while still working with a dimensionality reduced data volume at every iteration. This feature is essential for large-scale geophysical problems.

Dimensionality reduction by compressive sampling

Full-waveform inversion (FWI) involves the solution of the following multi-experiment unconstrained optimization problem:

$$\min_{\mathbf{m}} \frac{1}{2K} \sum_{i=1}^{K} \|\mathbf{b}_i - \mathscr{F}_i[\mathbf{m}, \mathbf{q}_i]\|_2^2, \tag{1}$$

with $K = n_f \times n_s$ the batch size, given by the total number of monochromatic sources $\mathbf{q}_i$ (n_f, n_s are the total number of frequencies and source experiments). The vectors $\mathbf{b}_i$ represent the corresponding vectorized monochromatic shot records and the $\mathscr{F}_i[\mathbf{m}, \mathbf{q}_i] = \mathbf{P}\mathbf{H}^{-1}[\mathbf{m}]\mathbf{q}_i$ represents the monochromatic forward operator for the i^{th} source, with $\mathbf{P}$ the detection operator, resticting data to the receiver positions. For simplicity, we assume a fixed receiver array. We also neglect surface-related multiples by using an absorbing boundary condition at the surface.

Unfortunately, the solution of the above minimization problem with Newton, Gauss-Newton, or Quasi-Newton techniques is extremely costly because each update requires multiple iterations involving the solution of the forward and time-reversed (adjoint) Helmholtz system for all n_f frequencies and all n_s sources. We address this problem by combining dimensionality-reduction strategies with recovery based on sparsity promotion. More specifically, we reduce the number of sources, and hence the number of the Helmholtz solves by replacing Eq. 1 with

$$\min_{\mathbf{m}} \left\{ \frac{1}{2} \|\mathbf{R}\mathbf{M}(\mathbf{b} - \mathscr{F}[\mathbf{m}, \mathbf{q}])\|_2^2 = \frac{1}{2} \|\underline{\mathbf{b}} - \underline{\mathscr{F}}[\mathbf{m}, \underline{\mathbf{q}}]\|_2^2 \right\}. \tag{2}$$

In this expression (the $\min$ runs over the expressions within the brackets), the vectors $\mathbf{b}, \mathbf{q}$ contain all monochromatic sources and shot records, and $\mathscr{F}[\mathbf{m}, \mathbf{q}]$ is the corresponding modeling operator for all monochromatic shots. We calculate the dimensionality-reduced counterparts of these quantities (denoted by the underbar), via $\underline{\mathbf{b}} = \mathbf{R}\mathbf{M}\mathbf{b}$, $\underline{\mathbf{q}} = \mathbf{R}\mathbf{M}\mathbf{q}$, and $\underline{\mathscr{F}}[\mathbf{m}, \underline{\mathbf{q}}]$. After applying this operator, n_s sequential monochromatic sources are combined into $n_s' \ll n_s$ supershots. Each supershot is given by a random superposition of all shots. A different random subset of $n_f' \ll n_f$ frequencies is used for each supershot.

Mathematically, this means that each simultaneous-source experiment in the collection of supershots is given by a different restriction. The i^{th} block of this restriction matrix $\mathbf{R}$ is given by the Kronecker product: $\mathbf{R}_i := \mathbf{R}_i^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{R}_i^{\Omega}$ for $i = 1 \cdots n_s'$. The restriction selects one supershot and the subset of frequencies. The measurement matrix $\mathbf{M}$ is given by the Kronecker product $\mathbf{M} := \mathbf{M}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{I}$, where $\mathbf{M}^{\Sigma}$, acting along the source coordinate uses Romberg (2009)'s phase encoding as in Lin and Herrmann (2007).

The identity in Eq. 2 follows from linearity of randomized subsampling by $\mathbf{R}\mathbf{M}$ and from linearity of forward modeling operator w.r.t. the sources. As a consequence, the number of PDE solves required in Eq. 2 is reduced by a factor of K'/K with $K' = n_s' \times n_f'$ (see also Herrmann et al., 2009b, for details). However, this speed up comes at the expense of creating artifacts that are related to source crosstalk and the key question is to find a solver that mitigates these artifacts.

Replacing Gauss-Newton subproblem with sparsity promoting formulation: A common approach to solving FWI (or in our case, the subsampled FWI problem) is the Gauss-Newton method, where at each iteration an update $\delta\mathbf{m}$ is obtained by solving the following least-squares subproblem

$$\min_{\delta\mathbf{m}} \|\underline{\delta\mathbf{b}} - \underline{\nabla\mathscr{F}}[\mathbf{m}, \underline{\mathbf{q}}]\delta\mathbf{m}\|_2^2, \tag{3}$$

with $\underline{\delta\mathbf{b}} = \underline{\mathbf{b}} - \underline{\mathscr{F}}[\mathbf{m}, \underline{\mathbf{q}}]$ the dimensionality reduced data residue and $\underline{\nabla\mathscr{F}}$ the reduced Jacobian—i.e., the linearized Born scattering operator for simultaneous sources. As in any Gauss-Newton method for FWI, the reduced Hessian in this formulation can be interpreted as ignoring internal multiple reflections present in the true Hessian (Pratt et al., 1998). To avoid additional complications, we assume our data to be surface-multiple free.

To mitigate source crosstalk and ill conditioning of the reduced Hessian, we add sparsity-promoting constraints to recover sparse updates for the linearized subsampled FWI problem Eq. (3). Specifically, given a sparsifying transform $\mathbf{S}$ with adjoint $\mathbf{S}^H$ (curvelets), we compute the update $\delta\mathbf{m} = \mathbf{S}^H\delta\mathbf{x}$ (with ^{H} denoting the adjoint). We obtain curvelet coefficient vector for the updates $\delta\mathbf{x}$ by solving the so-called LASSO problem

$$\min_{\delta\mathbf{x}} \quad \|\underline{\delta\mathbf{b}} - \underline{\nabla\mathscr{F}}[\mathbf{m}, \underline{\mathbf{q}}]\mathbf{S}^H\delta\mathbf{x}\|_2^2 \quad \text{s.t. } \|\delta\mathbf{x}\|_1 \le \tau, \tag{4}$$

with τ chosen automatically by the algorithm. The LASSO problem is more difficult to solve than the least square subproblem (cf. Eq. (3)), but we can afford to do this on the dimensionality-reduced problem. The key idea is that the constraint $\|\delta\mathbf{x}\|_1 \le \tau$ favors updates $\delta\mathbf{m}$ that are sparse in the curvelet domain. Because GN updates are sparse in the curvelet-domain, solving the LASSO problem removes the crosstalk and restores the amplitudes because it selects the largest curvelet coefficients that lie above the crosstalk noise level.

1: initialize $\mathbf{m}$, $k \leftarrow 0$.
2: **for** $j = 1$: number of frequency bands **do**
3: **while** not converged **do**
4: Randomly subsample to obtain $\underline{\delta\mathbf{b}}^k, \underline{\mathbf{q}}^k$.
5: $\delta\widetilde{\mathbf{x}} \leftarrow \left\{ \begin{array}{ll} \arg\min_{\delta\mathbf{x}} & \|\underline{\delta\mathbf{b}}^k - \underline{\nabla\mathscr{F}}[\mathbf{m}^k, \underline{\mathbf{q}}^k]\mathbf{S}^H\delta\mathbf{x}\|_2^2 \\ \text{s.t.} & \|\delta\mathbf{x}\|_1 \le \tau_k, \end{array} \right\}$
6: $\mathbf{m}^{k+1} \leftarrow \mathbf{m}^k + \mathbf{S}^H\delta\widetilde{\mathbf{x}}$
7: $k \leftarrow k + 1$
8: **end while**
9: **end for**

Algorithm 1: Modified GN-method for FWI with sparse updates

Solving the LASSO Subproblem: In order to solve (4), we use the spectral projected gradient (SPG) algorithm implemented in SPGℓ_1 (Berg and Friedlander, 2008). At each iteration, the algorithm computes the gradient of the least squares objective in (4), and then obtains a modified

direction by projecting this gradient onto the set $\{\delta\mathbf{x} : \|\delta\mathbf{x}\| \leq \tau\}$. We use the SPG method as our computational kernel to compute the updates (see Algorithm 1).

Pareto Curve: In Algorithm 1, we solve a series of LASSO problems for different linearizations. It is also essential to draw new collections of supershots after each LASSO subproblem has been solved (Li and J.Herrmann, 2010; Li et al., 2011; Herrmann and Li, 2011). Each subproblem also requires a parameter τ_k as the constraint on the ℓ_1-norm of the update. The series of τ_k parameters for each k^{th} subproblem are picked automatically by the SPGℓ_1 algorithm using the tradeoff curve between the optimal value of the misfit and the sparsity level τ. Specifically, consider the optimal value function

$$v_k(\tau) = \min_{\delta\mathbf{x}} \left\{ \|\underline{\delta\mathbf{b}^k} - \underline{\nabla\mathscr{F}}[\mathbf{m}^k, \underline{\mathbf{q}^k}]\mathbf{S}^H\delta\mathbf{x}\|_2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \leq \tau \right\} .$$

Given τ and data for the k-th iterate, the value $v_k(\tau)$ is immediately obtained as the square root of the minimum value of the LASSO problem with the τ constraint. The curve traced out by $v_k(\tau)$ as τ varies is shown to be differentiable w.r.t. τ (Berg and Friedlander, 2008), and dubbed the Pareto curve (see figure 1). Steepness of the Pareto curve reflects the ability to make progress in the optimization for small increases in τ. The SPGℓ_1 algorithm therefore sets

$$\tau_k = \frac{\|\delta\mathbf{b}^k\|_2 - \sigma}{-v_k'(0)} ,$$

the ratio of the current misfit to the slope of the Pareto curve at the origin. The parameter σ in the above expression is a noise level, which serves as a minimum value for the linearlized misfit that we would ever want to aim for. In the experiments presented below, we set $\sigma = 0$. The derivative of the origin has a closed form expression:

$$v_k'(0) = -\|\nabla\mathscr{F}^H\underline{\delta\mathbf{b}^k}\|_\infty ,$$

where $\nabla\mathscr{F}^H$ is the adjoint of the Jacobian at the current iterate, and $\|\mathbf{w}\|_\infty$ is the element-wise maximum of the absolute values of $\mathbf{w}$. For a more thorough treatment of the ideas related to the Pareto curve see (Berg and Friedlander, 2008) and (Hennenfent et al., 2008).

Example

To demonstrate the performance of our algorithm, we run a series of experiments on the 2-D BG model plotted in Fig. 2(a). All simulations are carried out with 350 shot positions sampled at a 20m interval and 701 receiver positions sampled at a 10m interval, yielding an maximum offset of 7km. We used a Ricker wavelet with a central frequency of 12Hz. The time record has a duration of $2.4\,\text{s}$ and is sampled with a sample interval of 16ms. To improve convergence, the inversions are carried out sequentially in 10 overlapping frequency bands on the interval $2.9 - 22.5\text{Hz}$ (Bunks et al., 1995), each using 7 different simultaneous shots and 10 selected frequencies—i.e., $K' = 70$. For each subproblem, we use roughly 20 iterations of SPG. Hence, we obtain LASSO updates at a cost roughly equivalent to one tenth of the cost of a gradient calculation with all of the sources ($K = 17500$). As a starting model, we use a velocity profile obtained by smoothing the original model (plotted in Fig. 2(b), followed by horizontal averaging.
The result after ten LASSO iterations for each frequency band is depicted in Fig. 2(c). Since we do not use a line search, we never evaluate the misfit for all sources. The cost of ten LASSO iterations is then roughly equivalent to one evaluation of the full misfit. This gives us an order of magnitude speed up.
As we can see from Fig. 2(c), the inversion result is able to capture all the discontinuities in the model up to a resolution commensurate the frequency range over which we carried out the inversion. Our results benefit significantly from supershot renewals after solving each LASSO subproblem, since these renewals remove the bias that would occur if we used a fixed collection of supershots during the entire procedure. Because of space limitations, we can not include our inversion result without renewals. We are also not able to show the sparsity-promoting result for all data because this is computationally infeasible. However, the experiments shown here demonstrate that our results are competitive with quasi-Newton methods applied to the entire data.

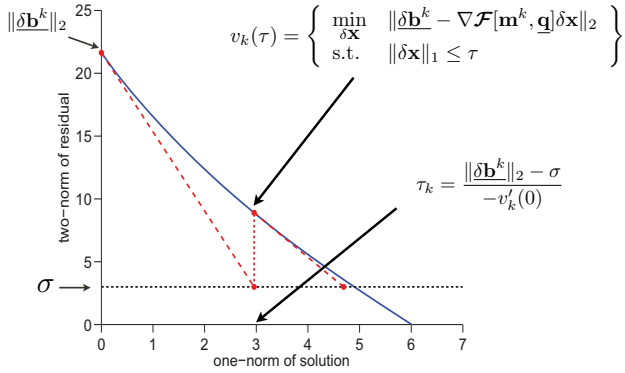

Figure 1: Pareto curve describes the tradeoff between the norm of the misfit and the sparsity parameter τ. A given value of τ determines the value function $v_1(\tau)$ as the solution to the corresponding LASSO problem. The curve is used to automatically select the value of τ_k at the k-th iteration.

Conclusions

We introduced an efficient algorithm to solve FWI, incorporating randomized dimensionality reduction into a modified Gauss-Newton method with sparse updates. Our method also uses techniques from stochastic optimization (Haber et al., 2010). In effect, we turn 'overdetermined' Gauss-Newton subproblems into underdetermined dimensionality-reduced subproblems with sparsity promotion. The resulting method is computationally efficient, as it works on a reduced data volume with fewer monochromatic source experiments. Our method also exploits natural sparsity of the updates in the curvelet domain, which acts as a powerful prior regularizing the inversion of the linearized subproblems.

Acknowledgments

We would like to thank the sponsors of the SINBAD project. This work was partly supported by a NSERC CRD Grant DNOISE II (375142-08). We also would like to thank Charles Jones for providing us with the BG model and the makers of CurveLab.

References

Akcelik, V., G. Biros, and O. Ghattas, 2002, Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation: Supercomputing, ACM/IEEE 2002 Conference, 41–41.

Beasley, C. J., 2008, A new look at marine simultaneous sources: The Leading Edge, **27**, 914–917.

Berg, E. v., and M. P. Friedlander, 2008, Probing the Pareto frontier for basis pursuit solutions: Technical Report 2, Department of Computer Science, University of British Columbia.

Berkhout, A. J., 2008, Changing the mindset in seismic data acquisition: The Leading Edge, **27**, 924–938.

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

Candès, E., J. Romberg, and T. Tao, 2006, Stable signal recovery from incomplete and inaccurate measurements: Comm. Pure Appl. Math., **59**, 1207–1223.

Candes, E. J., and L. Demanet, 2004, The curvelet representation of wave propagators is optimally sparse: Communications on Pure and Applied Mathematics, **58**, 1472–1528.

Donoho, D. L., 2006, Compressed sensing: IEEE Trans. Inform. Theory, **52**, 1289–1306.

Haber, E., M. Chung, and F. J. Herrmann, 2010, An effective method for parameter estimation with pde constraints with multiple right hand sides: Technical Report TR-2010-4, UBC-Earth and Ocean Sciences Department.

Hennenfent, G., E. van den Berg, M. P. Friedlander, and F. J. Herrmann, 2008, New insights into one-norm solvers from the Pareto curve: Geophysics, **73**, no. 4.

Herrmann, F. J., 2003, Multifractional splines: application to seismic imaging: Proceedings of SPIE Technical Conference on Wavelets: Applications in Signal and Image Processing X, SPIE, 240–258.

———, 2010, Randomized sampling and sparsity: Getting more information from fewer samples: Geophysics, **75**, WB173–WB187.

Herrmann, F. J., Y. A. Erlangga, and T. Lin, 2009a, Compressive sensing applied to full-waveform inversion: Presented at the , EAGE, EAGE.

———, 2009b, Compressive simultaneous full-waveform simulation: Geophysics, **74**, A35.

Herrmann, F. J., and X. Li, 2011, Efficient least-squares migration with sparsity promotion: Presented at the , EAGE, EAGE Technical Program Expanded Abstracts.

Herrmann, F. J., W. Lyons, and C. Stark, 2001, Seismic facies characterization by monoscale analysis: Geoph. Res. Lett., **28**, 3781–3784.

Herrmann, F. J., P. P. Moghaddam, and C. C. Stolk, 2008, Sparsity- and continuity-promoting seismic imaging with curvelet frames: Journal of Applied and Computational Harmonic Analysis, **24**, 150–173. (doi:10.1016/j.acha.2007.06.007).

Krebs, J. R., J. E. Anderson, D. Hinkley, A. Baumstein, S. Lee, R. Neelamani, and M.-D. Lacasse, 2009, Fast full wave seismic inversion using source encoding: , SEG, 2273–2277.

Li, X., A. Aravkin, T. van Leeuwen, and F. Herrmann, 2011, Full-waveform inversion with randomized l1 recovery for the model updates: Presented at the , EAGE, EAGE Technical Program Expanded Abstracts.

Li, X., and F. J. Herrmann, 2010, Full-waveform inversion from compressively recovered model updates: , SEG, 1029–1033.

Li, X., and F. J.Herrmann, 2010, Full-waveform inversion from compressively recovered updates: Presented at the , SEG, SEG Technical Program Expanded Abstracts.

Lin, T. T. Y., and F. J. Herrmann, 2007, Compressed wavefield extrapolation: Geophysics, **72**, SM77–SM93.

Neelamani, N., C. Krohn, J. Krebs, M. Deffenbaugh, and J. Romberg, 2008, Efficient seismic forward modeling using simultaneous random sources and sparsity: SEG International Exposition and 78th Annual Meeting, 2107–2110.

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

Pratt, R., C. Shin, and G. Hicks, 1998, Gauss-Newton and full Newton methods in frequency-space waveform inversion: Geoph. J. Int., **133**, 341–362.

Romberg, J., 2009, Compressive sensing by random convolution: SIAM Journal on Imaging Sciences, **2**, 1098–1128.

Romero, L. A., D. C. Ghiglia, C. C. Ober, and S. A. Morton, 2000, Phase encoding of shot records in prestack migration: Geophysics, **65**, no. 2, 426–436.

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

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

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

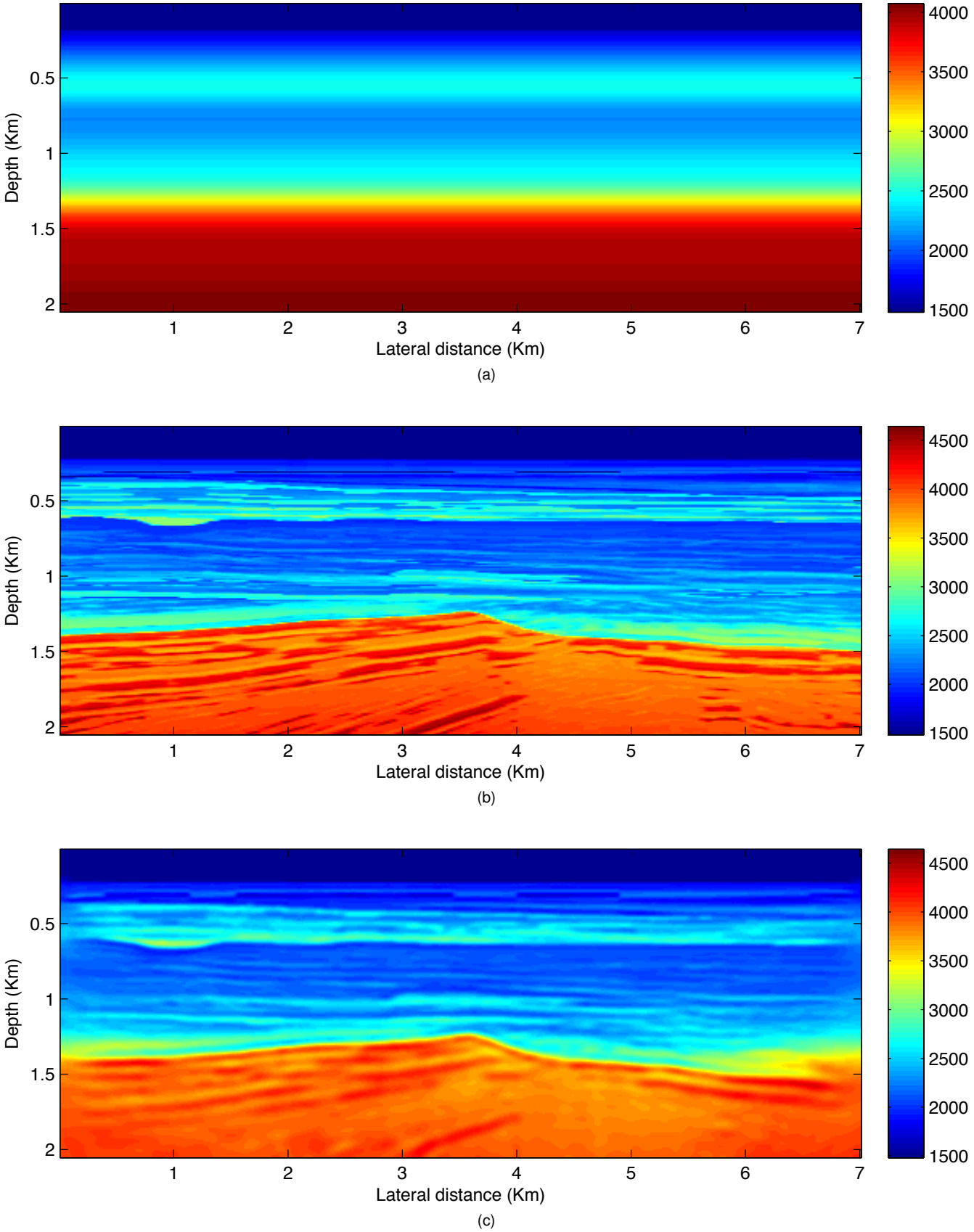

Figure 2: Full-waveform inversion result. **(a)** Initial model. **(b)** True model. **(c)** Inverted result starting from 2.9Hz with 7 simultaneous shots and 10 frequencies.