Efficient least-squares imaging with sparsity promotion and compressive sensing

Felix J. Herrmann[1] *and Xiang Li*[1]

ABSTRACT

Seismic imaging is a linearized inversion problem relying on the minimization of a least-squares misfit functional as a function of the medium perturbation. The success of this procedure hinges on our ability to handle large systems of equations—whose size grows exponentially with the demand for higher resolution images in more and more complicated areas—and our ability to invert these systems given a limited amount of computational resources. To overcome this "curse of dimensionality" in problem size and computational complexity, we propose a combination of randomized dimensionality-reduction and divide-and-conquer techniques. This approach allows us to take advantage of sophisticated sparsity-promoting solvers that work on a series of smaller subproblems each involving a small randomized subset of data. These subsets correspond to artificial simultaneous-source experiments made of random superpositions of sequential-source experiments. By changing these subsets after each subproblem is solved, we are able to attain an inversion quality that is competitive while requiring fewer computational, and possibly, fewer acquisition resources. Application of this concept to a controlled series of experiments showed the validity of our approach and the relationship between its efficiency—by reducing the number of sources and hence the number of wave-equation solves—and the image quality. Application of our dimensionality-reduction methodology with sparsity promotion to a complicated synthetic with well-constrained structure also yields excellent results underlining the importance of sparsity promotion.

INTRODUCTION

Modern-day seismic imaging technology depends increasingly on computationally and data-intensive "wave-equation" migration, which relies on full acquisition and high-fidelity wavefield simulations (see e.g. Rickett, 2003; Guitton, 2004; Plessix and Mulder, 2004). These challenges are compounded by a lack of available direct solvers for the time-harmonic Helmholtz equation in 3D. This is problematic because each source requires a separate PDE solve for indirect methods and this leads to simulation costs

[1]Seismic Laboratory for Imaging and Modeling, Department of Earth and Ocean Sciences, the University of British Columbia, 6339 Stores Road, Vancouver, V6T 1Z4, BC, Canada. Email: fherrmann@eos.ubc.ca

that increase linearly with the number of source experiments. This explains current-day interest in dimensionality-reduction techniques that aim to reduce exponentially growing data volumes acquired with exceedingly many sources.

Motivated by early work of Morton and Ober (1998); Romero et al. (2000) and more recently by Ayeni (2010); Fei et al. (2010), we overcome the challenge of the "curse of dimensionality" by decreasing the number of source experiments. As a result, we lower the computational burden of imaging significantly. To accomplish this goal, we extend the randomized dimensionality-reduction ideas presented by (Herrmann et al., 2009b; Neelamani et al., 2010) to the imaging problem.

Seismic imaging entails the inversion of an extremely large, but in the absence of noise, consistent "overdetermined" systems of equations. Even though there are generally more equations than unknowns, imaging is plagued by finite aperture and shadow zones, which make this system ill conditioned (Symes, 2008). Ill conditioning, in conjunction with extreme high costs of applying imaging operators, challenge iterative solution methods for least-squares imaging problems. To address this issue, we combine ideas from stochastic optimization (Bertsekas and Tsitsiklis, 1996; Shapiro et al., 2009; Nemirovski et al., 2009; Haber et al., 2010) and compressive sensing (CS—in short throughout this paper, Candès et al., 2006; Donoho, 2006; Mallat, 2009), yielding a formulation where we invert the large linearized system by solving a sequence of much smaller subproblems that act on source-encoded "supershots" (LI and Herrmann, 2010).

The presented approach differs from deterministic approaches, which include preconditioning, based on approximations of the wave-equation Hessian; data-dependent source syntheses, based on singular-value decompositions of the data matrix (Habashy et al., 2010); or the replacement of the Frobenius-norm (ℓ_2) by the matrix norm on the data residue (Symes, 2010). Instead, our method proposes to reduce the problem size. But contrary to (Sirgue and Pratt, 2004; Mulder and Plessix, 2004), who select deterministic subsets of angular frequencies in their imaging, we are motivated by recent ideas from Krebs et al. (2009); Haber et al. (2010) who utilize source-encoding to reduce the dimensionality of full-waveform inversion.

The outline of our paper is as follows. First, we motivate how stochastic optimization and CS are related to dimensionality reduction by randomized phase encoding in seismic imaging. Next, we introduce mini batches as collections of supershots, obtained by randomized sampling along the source and frequency axes. Inspired by stochastic optimization, we propose a solution to large-scale imaging problems via sequences of smaller dimensionality-reduced least-squares subproblems with or without sparsity constraints. Next, we identify these constrained subproblems as relaxed sparsity-promoting problems employed by large-scale one-norm solvers. We show that this leads to an efficient algorithm, which we subsequently analyze by performing a series of controlled imaging experiments. We conclude by applying the proposed method to a seismic imaging problem with well-constrained complexity.

MOTIVATION

After discretization, seismic imaging involves inversion of the linearized (time-harmonic) acoustic Born-scattering matrix linking data, collected in the vector $\mathbf{b} \in \mathbb{C}^{N_f N_r N_s}$ with N_f, N_r, and N_s the number of angular frequencies, receiver, and source positions, to perturbations in the medium parameters, collected in the vector $\mathbf{x} \in \mathbb{R}^M$, with M the number of gridpoints of the model. Without loss of generality, we will keep the density of mass fixed.

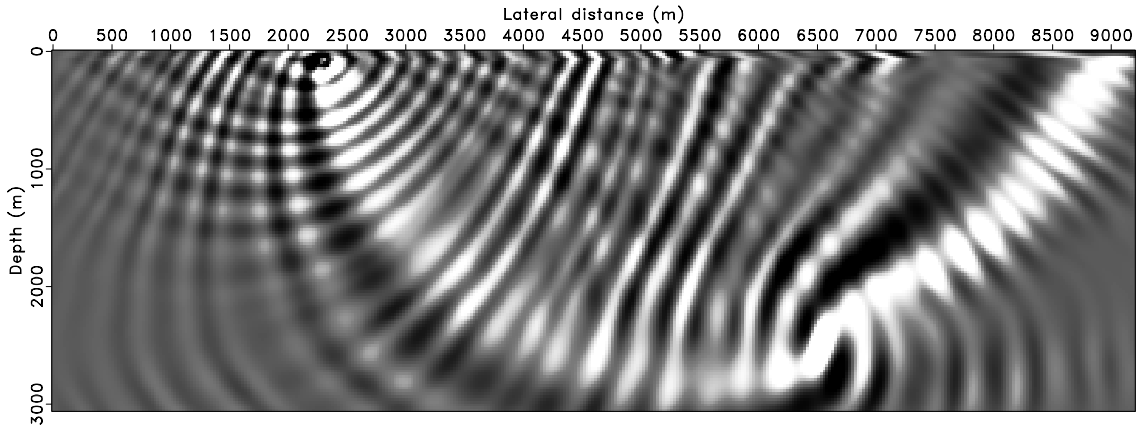
Because angular frequencies and sequential sources can be treated independently, the linearized inversion has the following separable form:

$$\underset{\mathbf{x}}{\text{minimize}} \frac{1}{2K}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \frac{1}{2K}\sum_{i=1}^{K}\|\mathbf{b}_i - \mathbf{A}_i\mathbf{x}\|_2^2, \tag{1}$$

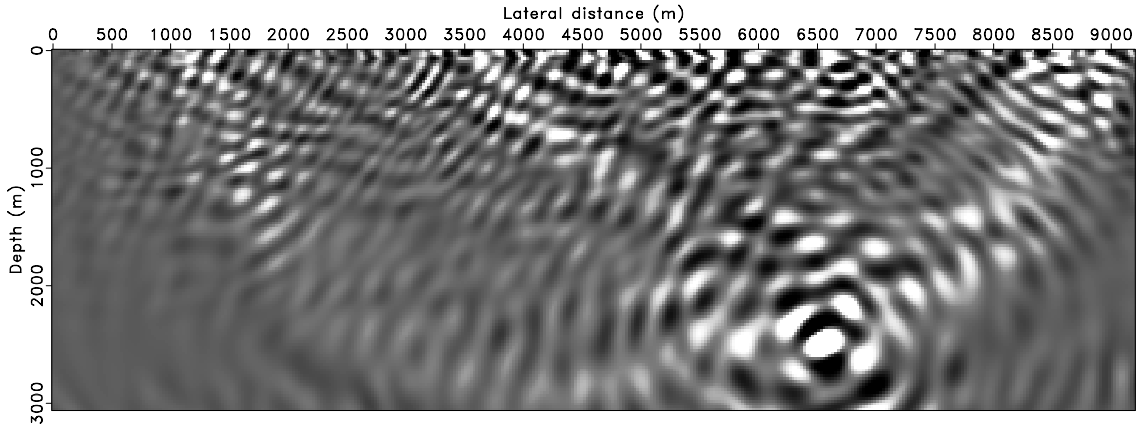
with $K = N_f N_s$ the batch size, given by the total number of monochromatic sources. The vectors $\mathbf{b}_i \in \mathbb{C}^{N_r}$ represent the corresponding vectorized monochromatic pre-processed (free of surface-multiples and direct waves) shot records. The matrix $\mathbf{A}_i$ represents the monochromatic linearized scattering matrix for the i^{th} source.

Unfortunately, solving this problem is problematic because each iteration requires $4K$ PDE solves: two to compute the action of $\mathbf{A}_i$ and two for the action of its adjoint $\mathbf{A}_i^H$. Both actions involve solutions of the forward source and reverse-time residual wavefields (Plessix and Mulder, 2004). Thus, the inversion costs grow linearly with the number of monochromatic experiments, multiplied by the number of matrix-vector multiplies required by the solver. (Because of the size of the problem, the matrices $\mathbf{A}_i$, $i = 1 \cdots K$ can not be formed explicitly and we have to rely on iterative methods to solve equation 1.)

While preconditioning techniques improve convergence of Lanczos methods (Herrmann et al., 2009a), these iterative techniques require multiple evaluations of the scattering operator and its adjoint. Unfortunately, these multiple passes through the complete data are computationally intractable. To overcome this difficulty, we use a dimensionality-reduction approach where sequential sources are replaced by a reduced number of simultaneous sources made of randomized superpositions. In this way, we not only explore linearity of the wave equation with respect to the sources but we also exploit the fact that randomized sources have a richer wavenumber content, which improves the image quality albeit the resulting image can be extremely noisy. Juxtapose Figure 1(a), obtained with a single sequential shot, with Figure 1(b) obtained from a single simultaneous shot. The key contribution of this paper is to mitigate this noisy source cross talk while still benefiting from computational gains related to the reduction of the number of required PDE solves. Before we outline the details of our randomized imaging algorithm, let us first briefly discuss recent developments in optimization and theoretical signal analysis that provide insights and justifications for our method.



(a)



(b)

Figure 1: Migration of a single spike positioned in the target zone of the Marmousi model with a single sequential or simultaneous source. **(a)** Migrated spike for one sequential shot. **(b)** The same but now one single simultaneous shot. Notice the improved image of the randomized simultaneous source due to the increased wavenumber diversity.

Stochastic optimization

In machine learning, separable optimization problems (cf. Equation 1) can be solved efficiently by either the stochastic-average approximation (SAA) or by the stochastic approximation (SA, Bertsekas and Tsitsiklis, 1996; Nemirovski et al., 2009; Haber et al., 2010). With SAA, Equation 1 is approximated by selecting subsets of frequencies and shots (possibly simultaneous shots), yielding $K' = n_f n_s \ll K$ for the reduced batch size with $n_f \ll N_f$ and $n_s \ll N_s$. For our imaging problem, this approach corresponds to carrying out least-squares migration with a randomized subset of shots. As shown by van Leeuwen et al. (2010b), the resulting error in the migrated image decay only slowly with increasing batchsize K'. This can be understood because SAA is essentially a Monte-Carlo sampling method, where errors decays only slowly ($\mathcal{O}(\sqrt{K'})$) with increasing batch size. Despite this disadvantage, SAA is popular because of its relative simplicity that offers flexibility with respect to the choice of solvers for the separable optimization in Equation 1. This flexibility allows us to use generic solvers such as LSQR (Paige and Saunders, 1982).

To address the relative slow convergence of SAA, SA directly intervenes in first-order optimization by computing gradients on randomized subsets of data. This turns deterministic gradient descent into stochastic-gradient descent (Bertsekas and Tsitsiklis, 1996; Nemirovski et al., 2009; Haber et al., 2010), which in our problem corresponds to randomly selecting a different monochromatic source—e.g., a different $\mathbf{A}_i$ with $i \subset [1 \cdots K]$—for each gradient update pertaining to Equation 1. For linear problems, this approach is reminiscent of randomized "block Kaczmarz" (Strohmer and Vershynin, 2009), which was used successfully in the deterministic case by Natterer (2001) in tomography. Like SAA, SA extends to nonlinear inversion problems (see e.g., Nemirovski et al., 2009), and was recently introduced by Haber et al. (2010) in the context of parameter-estimation problems with PDE's. SA also justifies recent work by Krebs et al. (2009) and provides a theoretical explanation for observed lack of convergence— SA only converges with $\mathcal{O}(1/k)$ with k the number of iterations (Nemirovski et al., 2009)—and instabilities with respect to noise (van Leeuwen et al., 2010a).

Compressive sensing

Randomized-dimensionality reduction also underlies recent advances in sampling theory for signals that exhibit structure, which translates into transform-domain sparsity. As opposed to stochastic optimization, where randomization is used to reduce variance (see e.g., Hutchinson, 1990; Avron and Toledo, 2010), compressive sensing uses randomization to turn coherent subsampling-related interferences—such as aliasing and shot "cross talk"—into relatively "harmless" Gaussian noise. According to CS, the noise level depends on the degree of subsampling and transform-domain sparsity. Consequently, sampling is no longer fully determined by Nyquist, but by transform-domain sparsity (see e.g., Herrmann, 2010, for an overview of the application of CS

in exploration seismology). Neelamani et al. (2010) and Herrmann et al. (2009b) both took advantage of this finding in wavefield simulations with time stepping or by inverting the time-harmonic Helmholtz system. In both cases, fully-sampled wavefields are recovered with curvelet-domain sparsity promotion from small subsets of (monochromatic) simultaneous-source experiments. This procedure leads to efficient simulations because the computational overhead of the recovery is small compared to computational gain from subsampling.

However, solving Equation 1 differs fundamentally from standard CS because the system in Equation 1 is "overdetermined"; there are more equations then unknowns. In addition, the scattering matrix is ill conditioned due to limitations in aperture that may lead to shadow zones. However, for reflectors that are in the range of the scattering operator the wave-equation Hessian $\mathbf{A}^H\mathbf{A}$ is near unitary, and curvelets are nearly invariant under the action of the Hessian (Herrmann et al., 2008). This property—in conjunction with the optimality of curvelets on images with reflectors that may include conflicting dips—motivates us to use curvelet-domain sparse recovery to mitigate the source crosstalk caused by the dimensionality reduction.

METHODOLOGY

To solve Equation 1 efficiently, we combine recent ideas from stochastic optimization and compressive sensing. For this purpose, let us first mathematically define seismic mini batches consisting of "supershots". Next, we present a pragmatic optimization strategy where we cast the original imaging problem into a series of much smaller subproblems that work on different subsets of random source-encoded supershots. For linearized inversion (Herrmann and Li, 2011), this approach corresponds to drawing a collection of supershots, followed by imaging, and using this image as a warm start for a new linearized inversion with a new independently drawn collection of supershots. We compare the performance of least-squares on these subproblems with and without sparsity constraints. Depending on these two choices, the formulation leads either to a Monte-Carlo type of algorithm, which relies on averaging to reduce the subsampling errors, or to a compressive-sensing type of algorithm, which relies on sparsity promotion to remove the source cross talk. In both cases, the error depends on the subsampling ratio K/K'.

The seismic mini batch: a collection of supershots

We base our algorithm on forming compressive seismic experiments—or to use the language of online machine learning mini batches—that consist of collections of small numbers of supershots. These supershots are made of randomized superpositions of sequential sources.

Mathematically, imaging experiments for mini batches with $K' \ll K$ monochro-

matic supershots, require the solution of the reduced system

$$\mathbb{P}_{\ell_2}(\mathbf{RM}): \qquad \underset{\mathbf{x}}{\text{minimize}} \left\{ \frac{1}{2} \|\mathbf{RM}\,(\mathbf{b}-\mathbf{Ax})\,\|_2^2 = \frac{1}{2}\|\underline{\mathbf{b}}-\underline{\mathbf{A}}\mathbf{x}\|_2^2 \right\}, \tag{2}$$

In this expression, we dimensionality reduced Equation 1 with the subsampling matrix $\mathbf{RM}$. This subsampling matrix reduces the tall expensive to compute system $\mathbf{A} \in \mathbb{C}^{(KN_r)\times M}$ to $\underline{\mathbf{A}} \stackrel{\text{def}}{=} \mathbf{RMA} \in \mathbb{C}^{(K'N_r)\times M}$ and the data to $\underline{\mathbf{b}} \stackrel{\text{def}}{=} \mathbf{RMb}$. The dimensionality-reduction matrix itself is factored into a restriction and mixing matrix. The restriction matrix $\mathbf{R}$ is defined by the Kronecker product: $\mathbf{R} \stackrel{\text{def}}{=} \mathbf{R}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{R}^{\Omega} \in \mathbb{R}^{(K'N_r)\times(KN_r)}$ with $\mathbf{R}^{\Sigma} \in \mathbb{R}^{n_s'\times n_s}$ selecting $n_s' \ll n_s$ rows uniform randomly amongst $[1\cdots n_s]$ and $\mathbf{R}^{\Omega} \in \mathbb{R}^{n_f'\times n_f}$ selecting $n_f' \ll n_f$ frequencies from the seismic frequency band. The matrix $\mathbf{I}$ represents the identity matrix. The mixing matrix $\mathbf{M} \in \mathbb{R}^{(KN_r)\times(KN_r)}$ is given by the Kronecker product $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{M}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{I}$. As in Lin and Herrmann (2007), we follow Romberg (2009) to phase encode sequential shots via

$$\mathbf{M}^{\Sigma} \stackrel{\text{def}}{=} \text{sign}(\boldsymbol{\eta}) \odot \mathbf{F}_{\Sigma}^{H}\, \text{diag}\left(e^{j\boldsymbol{\theta}}\right) \mathbf{F}_{\Sigma}, \tag{3}$$

with $\boldsymbol{\theta} = \text{Uniform}([0, 2\pi])$ a random phase rotation, and $\mathbf{F}_{\Sigma}$ the Fourier transform along the source coordinate. The vector $\boldsymbol{\eta} \in N(0,1)$ is used to define a random-sign pattern by which the phase-encoded vector is premultiplied (the symbol $\odot$ represents element-wise product). This definition for the matrix $\mathbf{M}^{\Sigma}$ is fast ($\mathcal{O}(n_s \log n_s)$) and mimics the action of a matrix with Gaussian *i.i.d.* entries.

Using linearity of randomized subsampling by $\mathbf{RM}$, in combination with linearity with respect to monochromatic sources and the separability of the imaging problem (cf. Equation 1), it is easy to show that the number of PDE solves required for each iteration of the solution of Equation 2 is slimmed down by a factor of K'/K (see also Herrmann et al., 2009b, for details). Essentially, the action of $\mathbf{RM}$ "commutes". Note that recent work by Haber et al. (2010) uses the same principle. Unfortunately, speed ups by randomized source supersposition and subsampling go typically at the expense of leaking energy from imaged reflectors to incoherent artifacts. Hence, the key question is to find a solver that mitigates these artifacts and restores the amplitudes at an overhead small compared to the speed up.

Stochastic-average approximations with warm starts

To address the slow decay of the error of SAA and the delicacy of the SA, we combine ideas underlining these methods by casting the original imaging problem into a series of much smaller subproblems that work on different independent subsets of random source-encoded supershots (Herrmann and Li, 2011). For linearized inversion, this approach corresponds to drawing a collection of supershots, followed by least-squares imaging, and using these images as warm starts for a new linearized inversion with a new independently drawn collection of supershots. This process is repeated until no longer progress is made towards the solution. In Algorithm 1, we outline this procedure for a generic subproblem solver $\mathbb{P}(\mathbf{RM}; \mathbf{x}_0)$ that uses warm starts $\mathbf{x}_0$.

Two subproblem solvers

Because algorithm 1 gives us flexibility regarding the subproblem solver, we propose to compare two solvers, namely $\mathbb{P}_{\ell_2}(\mathbf{RM})$—solved by a limited number of iterations of LSQR (Paige and Saunders, 1982)—and

$$\mathbb{P}_{\ell_1}(\mathbf{RM},\tau): \qquad \underset{\mathbf{x}}{\text{minimize}}\, \frac{1}{2}\|\underline{\mathbf{b}}-\underline{\mathbf{A}}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \tau. \tag{4}$$

We solve the latter problem, known as the Least Absolute Shrinkage and Selection Operator (LASSO Tibshirani, 1997) problem, by a spectral-projected gradient method (see e.g., Berg and Friedlander, 2008, for details).

With the LSQR solver, we regularize the inverse of the wave-equation Hessian by limiting the number of iterations of LSQR (Hansen, 1997). This is necessary because otherwise we may create imaging artifacts related to the null space of the (dimensionality-reduced) Hessian. The total number of PDE solves required by LSQR is proportional to $N_{\ell_2}K'$, with N_{ℓ_2} the number of iterations required by the ℓ_2-norm solver. Conversely, we control the null space with LASSO by sparsity promotion. To take full advantage of sparsity promotion as a regularization, we include the curvelet synthesis matrix (Candès et al., 2006) in the definition of the dimensionality-reduced Born scattering operator $\underline{\mathbf{A}}$. The migrated image is then calculated by applying curvelet synthesis on the $\mathbf{x}$ that solves Equation 4. The total number of PDE solves required by this algorithm is proporional to $N_{\ell_1}K'$ with N_{ℓ_1} the number of iterations required by the ℓ_1-norm solver. In these computations, the computational overhead of the curvelet transform is negligible compared to the cost of solving PDE's and is therefore ignored.

Leveraging the Pareto curve

In the noise-free case, sparsity-promoting imaging involves the solution of the following optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}}\, \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{b}=\mathbf{A}\mathbf{x}. \tag{5}$$

Efficient ℓ_1 solvers for this problem are typically based on solutions of a series of relaxed subproblems, where components are allowed to enter into the solutions controllably. It is widely known that these approaches lead to a reduction in the number of iterations to reach the solution. The spectral projected-gradient algorithm (SPGℓ_1 Berg and Friedlander, 2008) uses this principle by solving a series of LASSO problems where the τ's are increased intelligently. In this method, the Pareto boundary—the trade-off curve delineating feasible and infeasible solutions as a function the ℓ_2-norm of the data misfit and the model's ℓ_1-norm—is exploited to compute the relaxations by root finding that uses convexity and smoothness of the Pareto curve. See Figures 2(a) and 2(b), which illustrate this principle, and the corresponding solution path. As we can see this approach uses a limited number of matrix-vector multiplies. Because

the cost of the solver is determined by this number of multiplies, this approach is particularly suitable for large-scale geophysical problems (Hennenfent et al., 2008).

Unfortunately, the degree of randomized dimensionality reduction determines the amount of cross-talk that results from the inversion, and hence we can not reduce the problem size too much. Therefore, we improve convergence by drawing new mini batches whenever a LASSO subproblem is solved. Because the solution is maximally sparse at that point, it is natural to select the new set of supershots and continue with a warm start of the algorithm for the next subproblem. We calculate the τ's with SPGℓ_1's root finding. This principle is illustrated in Figure 2(b) where the system of equations now changes after solving each subproblem. Of course, this approach is only justified as long as K' is not too small such that Pareto curves remain similar for different realization of the **RM**'s. To verify the assumption of similarity amongst different Pareto curves, we plotted four realizations of these curves for $K' = 12$ (four simultaneous shots and three frequencies) in Figure 2(c). These curves clearly make the case that we should be able to continue using SPGℓ_1's root finding.

EMPIRICAL PERFORMANCE STUDY

To compare the proposed algorithm, we conduct a series of synthetic imaging experiments based on the smoothed Marmousi model (Bourgeois et al., 1991) plotted in Figure 3(a). This model, which also defines the medium perturbation (Figure 3(b)), is used as the background-velocity model for migration. With this model, we generate time-harmonic input data by subtracting solutions of the Helmholtz equation for the true and smoothed velocity. We use a nine-point stencil (Jo et al., 1996) and absorbing boundary conditions on a 143×384 grid with a grid size of 24 m. To mimic real applications, we solve the Helmholtz systems on the fly during the inversion. The length of the time record is 2.4 s and we use 192 shot locations, with a shot spacing of 48 m, and 384 receiver positions sampled with a receiver spacing of 24 m.

To establish a baseline for comparison, we first compute a least-squares reference image by solving Equation 1 for all 192 shots and 10 randomly selected frequencies and 10 iterations of LSQR. This baseline image corresponds to a batchsize of $K = 1920$. The result of this exercise is included in Figure 4 for a Ricker wavelet with a central frequency of 12 Hz.

We also solve a series of dimensionality reduced subproblems with 8 supershots and 3 frequencies ($K' = 24$) for 10 subproblems with LSQR ($\mathbb{P}_{\ell_2}(\mathbf{RM})$) and SPG$\ell_1$ ($\mathbb{P}_{\ell_1}(\mathbf{RM})$) each with and without independent renewals of **RM**. The results of these experiments are summarized in Figure 5. From these experiments, we can make the following observations. First, redrawing the supershots after solving each subproblem improves the performance of both solvers. This can be understood because these renewals remove possible correlations between **RM** and the current estimate for the (curvelet-domain) velocity perturbation. Second, the images obtained by sparsity promotion are clearly superior in quality compared to the least-squares results. This

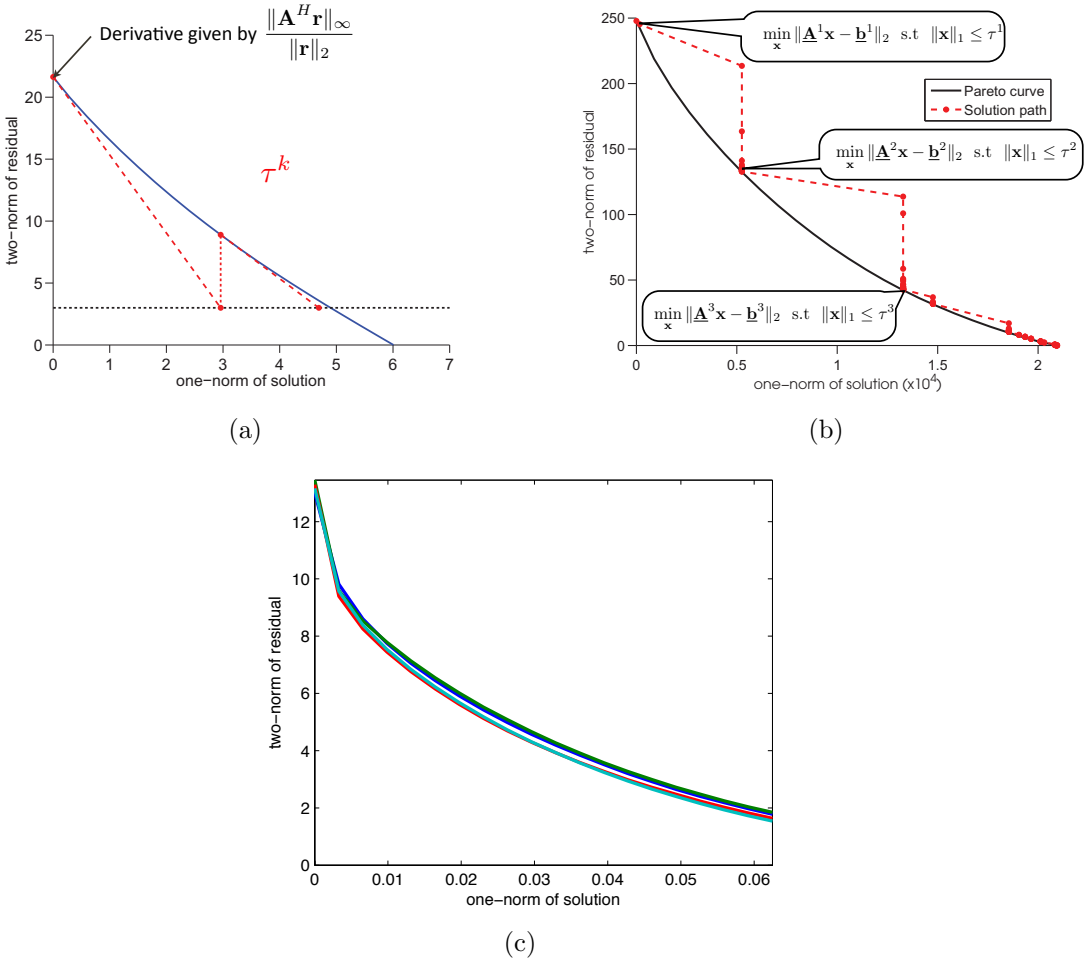
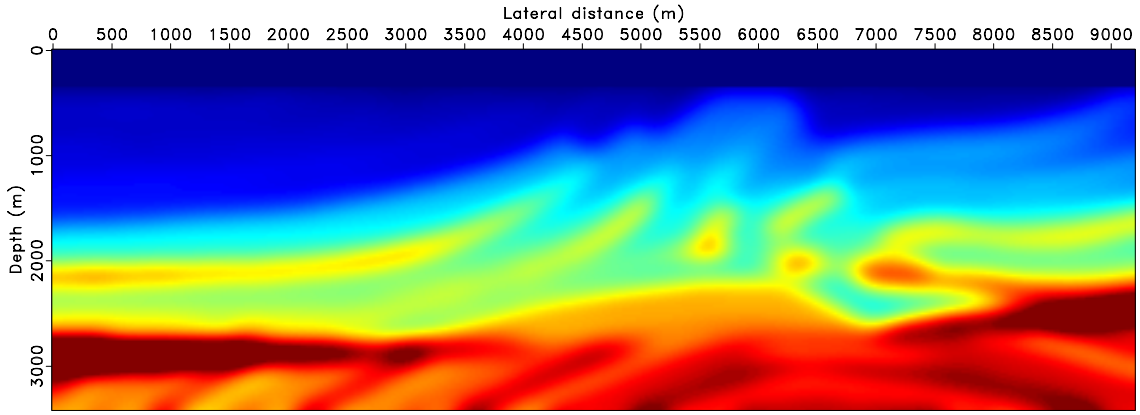
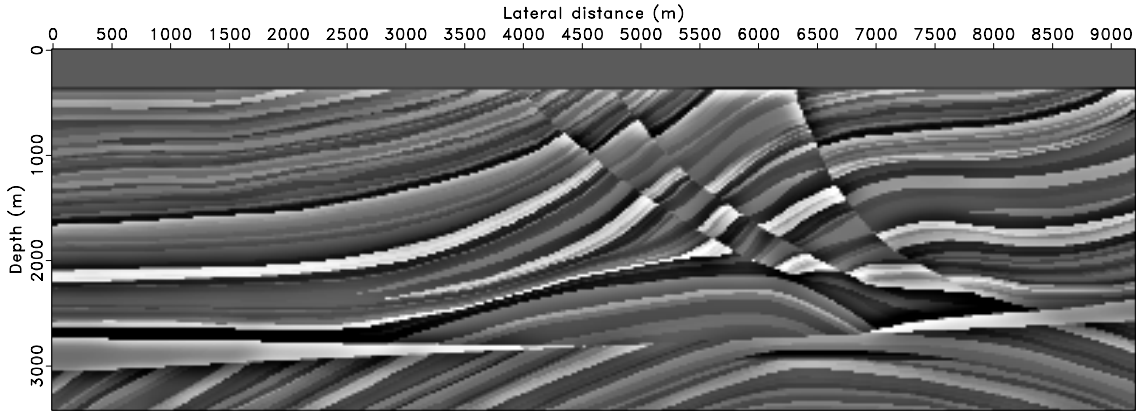


Figure 2: SPGℓ_1 and batching. **(a)** Newton root finding using the convexity and smoothness of the Pareto curve (adapted from Berg and Friedlander (2008)). **(b)** Series of LASSO subproblems with renewals for the collections of supershots (adapted from Berg and Friedlander (2008)). **(c)** Pareto curves for different independent realizations of the dimensionality-reduction ($K' = 12$) operator **RM**.

can be explained by CS. Third, the sparsifying result with renewals, albeit noisy, compares favorably to the baseline image in certain areas; e.g. it has higher resolution and better resolved amplitudes at depth. In summary, we obtained a remarkably good result with a significantly reduced computational cost. We attribute this performance to curvelet-domain compressibility, which serves as a strong prior that mitigates source crosstalk and regularizes the inversion.



(a)



(b)

Figure 3: The Marmousi model. **(a)** Smoothed background-velocity model. **(b)** Velocity perturbation, defined by the difference between the true and smoothed background-velocity models.

Convergence as a function of the number of PDE solves

The possible gains in computation speed of our solvers hinges on the interplay between the mini batch size and the number of matrix-vector multiplies required by the solver to bring down both the data residue and to recover the artifact-free image. The product of these two factors determines the number of PDE solves. To measure the

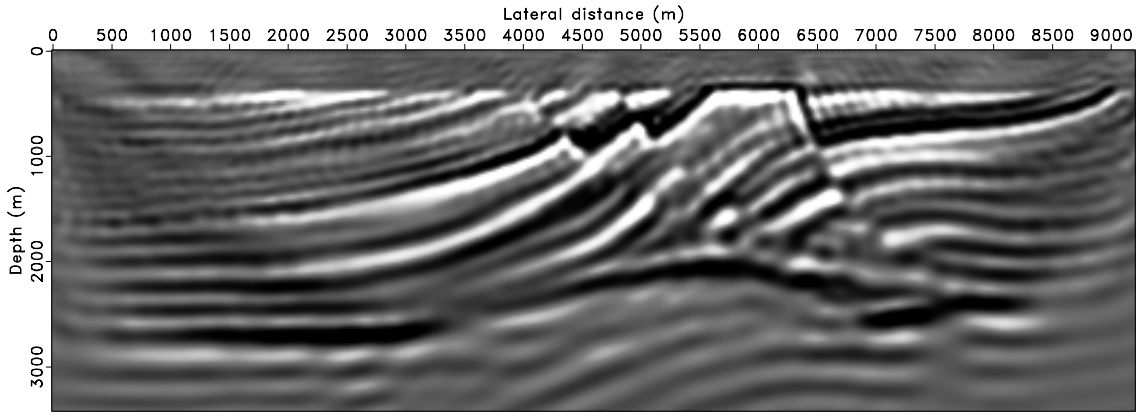
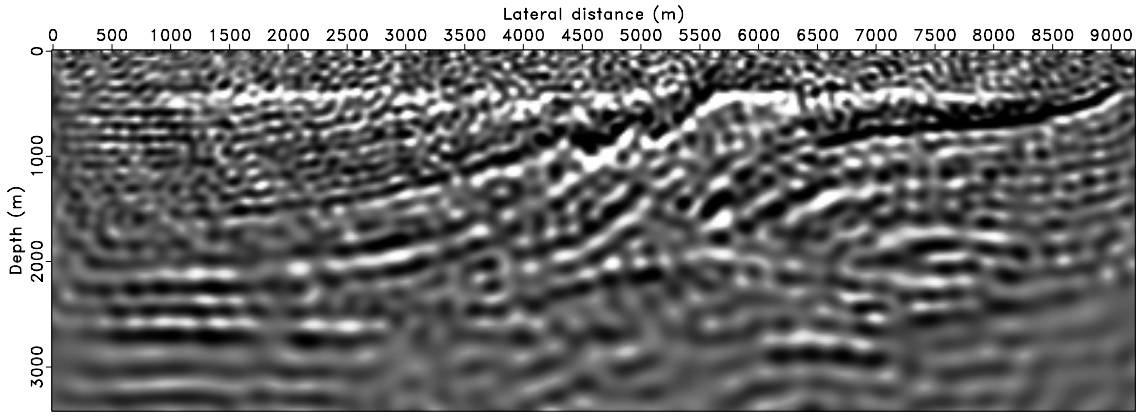
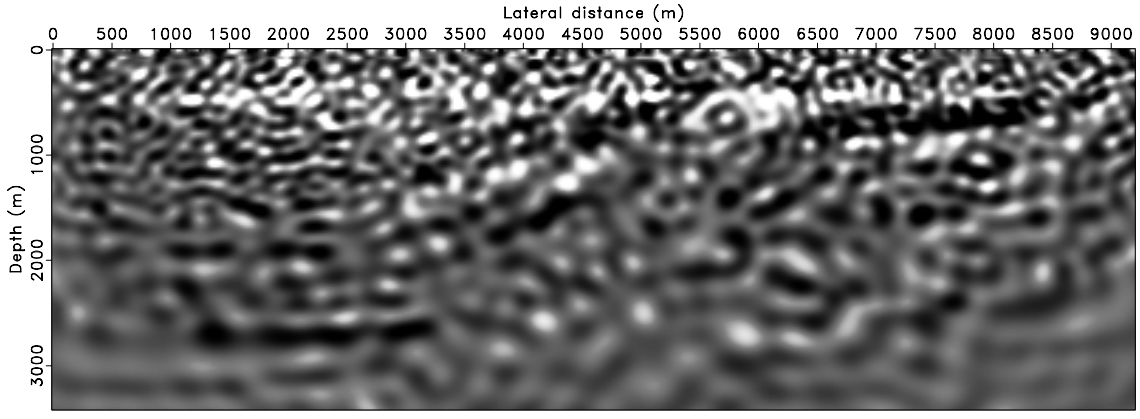


Figure 4: Baseline image calculated for all data (192 sources and 10 frequencies) with 10 iterations of LSQR.

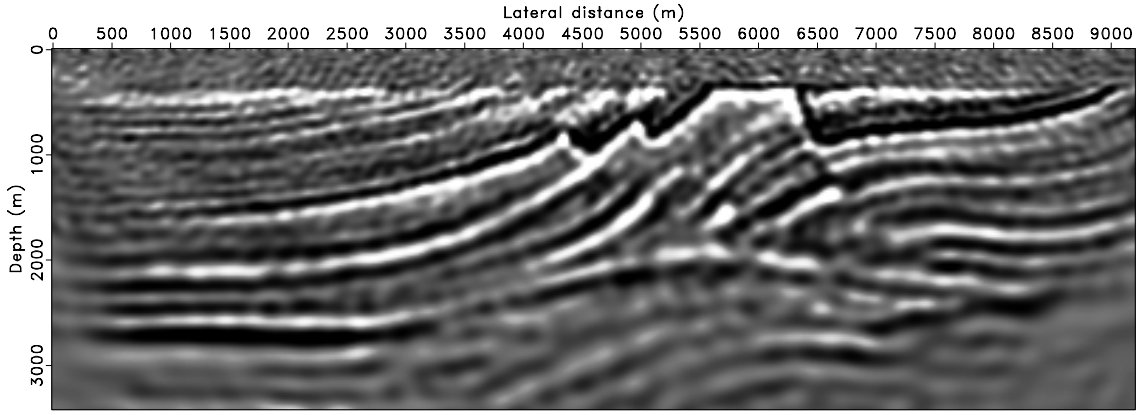


(a)

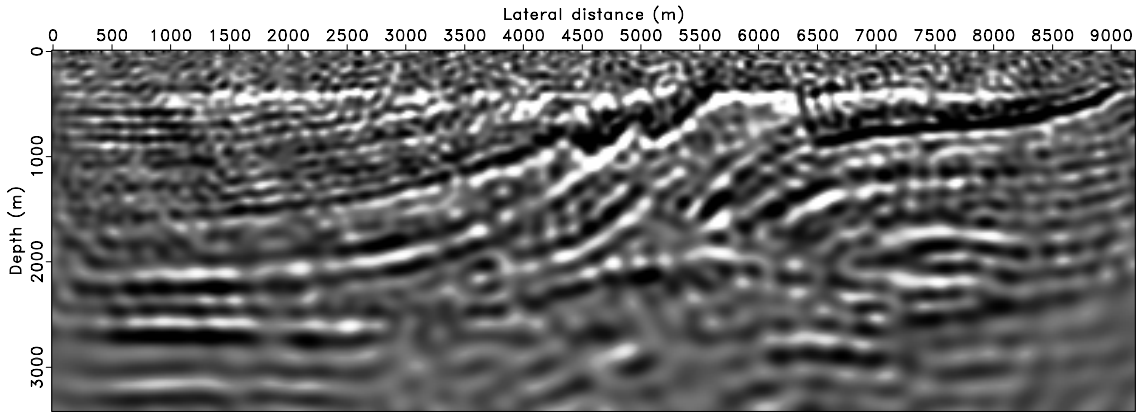


(b)

Figure 5: Stochastic-average approximation with LSQR. **(a)** Image obtained by Algorithm 1 with $\mathbb{P}_{\ell_2}(\mathbf{RM})$ with 10 independent redraws for $\mathbf{RM}$. **(b)** The same but with the same $\mathbf{RM}$.



(a)



(b)

Figure 6: Stochastic-average approximation with SPGℓ_1. **(a)** Image obtained by Algorithm 1 with $\mathbb{P}_{\ell_1}(\mathbf{RM})$ for approximately the same number of PDE solves.**(b)** The same but with the same **RM**.

performance of the proposed curvelet-based stochastic-average approximation with warm starts, we plot this number versus the model-error energy for a fixed ratio of $K/K' = 80$ and $n'_s/n'_f = 8/3$. The results of this exercise with and without redraws are plotted in Figure 7. This plot clearly demonstrates a more rapid decay for the model-space error (difference between true and estimated model) in case of independent redraws of the **RM**'s. Compared to the baseline problem with "all" data, we obtain approximately a fourfold speedup if we factor in the number of iterations required by the solver. On first glance, this speedup may not be significant. However, we expect a larger uplift in 3D where there are many more sources. In addition, with our dimensionality reduction we are able to approximately solve the BP problem (cf. Equation 5). Without the dimensionality reduction this would not have been possible because of the number of iterations required by one-norm solvers.

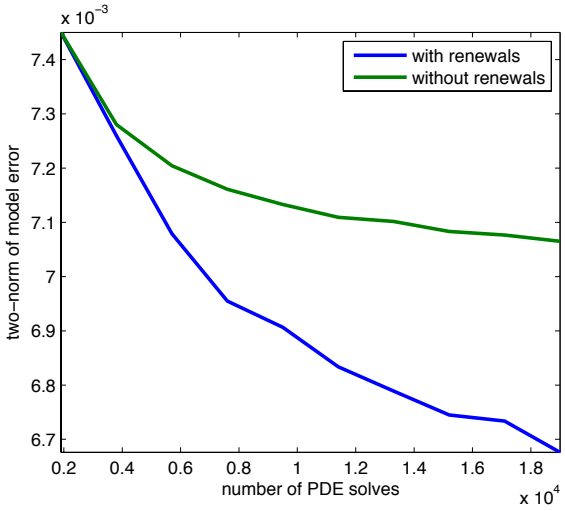


Figure 7: Two-norm error between the true and recovered medium perturbation as a function of the number of PDE solves (for $\mathbb{P}_{\ell_1}(\mathbf{RM})$). Convergence is clearly improved by drawing new randomized collections of supershots after each subproblem is solved.

Recovery quality as a function of batchsize

As we have seen from the previous example, computational gains can be made using the proposed stochastic-average approximation with relaxed LASSO's. To get a better understanding of the relationship between the recovery error and minibatch size, we also conduct a series of experiments where we vary the subsampling ratio's and the ratio supershots-over-frequencies while approximately fixing the number of PDE solves. The results of this exercise are summarized in Table 1. As expected, the numbers in Table 1 generally confirm increasing recovery errors for increasing subsampling ratios albeit there is not a very strong relationship between the recovery quality and the subsampling ratio compared to results presented in the literature (LI and Herrmann, 2010). The fact that our results were generated with renewals offers an explanation for the weaker dependence on the subsampling ratios.

CASE STUDY: THE BG COMPASS MODEL

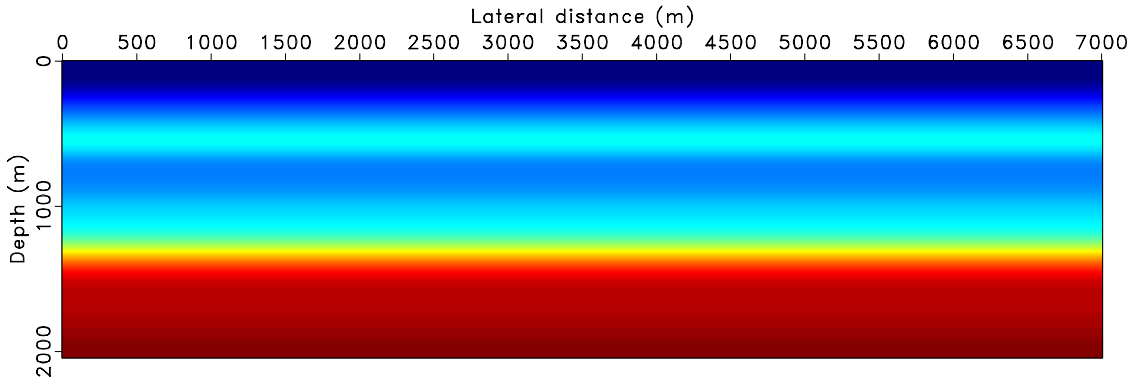
To test our imaging algorithm in a more realistic setting, we consider a synthetic velocity model with a large degree of well-constrained variability. To build the background-velocity model, we employ our modified Gauss-Newton method with sparse updates in 10 overlapping frequency bands on the interval $2.9 - 22.5$Hz and with initial model plotted in Figure 8(a). (Note that this approach reported in Li et al. (2011) is based on a similar dimensionality reduction technique as presented in this paper.) The output of this procedure is plotted in Figure 8(b) and is used as the background-velocity model for our imaging algorithm.

We parametrize the velocity perturbation on a 409×1401 grid with a gridsize of 5 m. Again, we use the Helmholtz solver to generate data from 350 source positions sampled at an interval of 20 m and 701 receivers sampled with an interval of 10 m. We use 10 random frequencies in our simulations selected from the interval $20 - 50$ Hz and scaled by the spectrum of a 30 Hz Ricker wavelet. The input data is given by the difference between simulations with the true and initial velocity models (Figure 8(b)). As before, we solve 10 subproblems $\mathbb{P}_{\ell_1}(\mathbf{RM})$ with and without independent redraws of $\mathbf{RM}$. The result of this exercise is summarized in Figure 9 and clearly show significant improvements from the redraws. Not only is the crosstalk removed more efficiently but the reflectors are also better imaged in particular at the deeper parts of the model where recovery without redraws is not able to image the events.

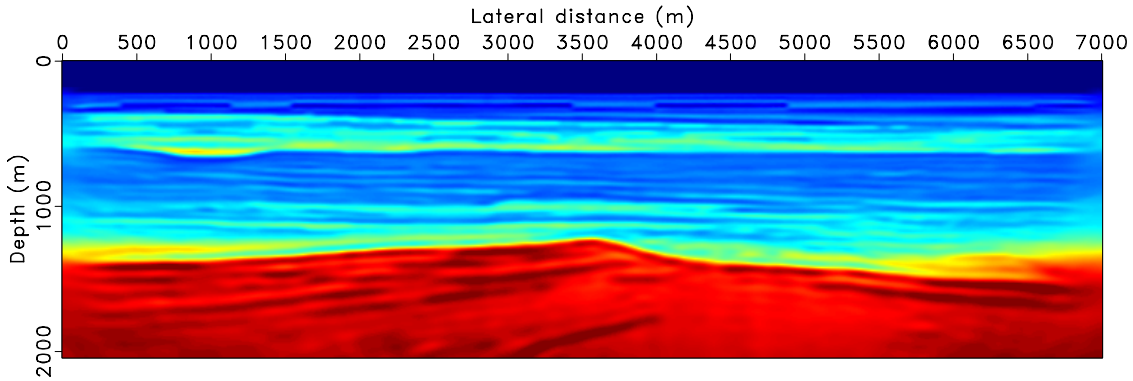
DISCUSSION

Efforts to speed up computational costs of linearized imaging roughly fall into two categories. First, there are methods that aim to "nearly diagonalize" Green's functions (Douma and de Hoop, 2007) or wave-equation Hessians (Herrmann et al., 2008) using transform-domain techniques such as curvelets. These methods exploit the property that curvelets remain near invariant under wave propagation, which in principle, leads to fast algorithms. Unfortunately, the engineering of concrete and explicit implementations of these approximations is involved, and may carry a significant overhead (Andersson et al., 2008). Conversely, randomized dimensionality reduction is simpler because it utilizes the notion of curvelet-invariance implicitly via transform-domain sparsity. (Propagated wavefields remain compressible in curvelet frames. See Smith, 1998; Demanet and Peyré, 2011, who rigorously prove this property.) This feature explains the success of our proposed method, which benefits from the availability of wave simulators and the ability of curvelets to sparsely represent seismic images. By promoting sparsity, we are able to exploit continuity along the reflectors without requiring a data-adaptive step that requires prior information on the dip field of the reflectors (Guitton et al., 2010).

By considering dimensionality-reduced subproblems as CS-like sparse recovery problems—where the originally "overdetermined" system is turned deliberately into an underdetermined system—we remove the crosstalk artifacts and restore the ampli-

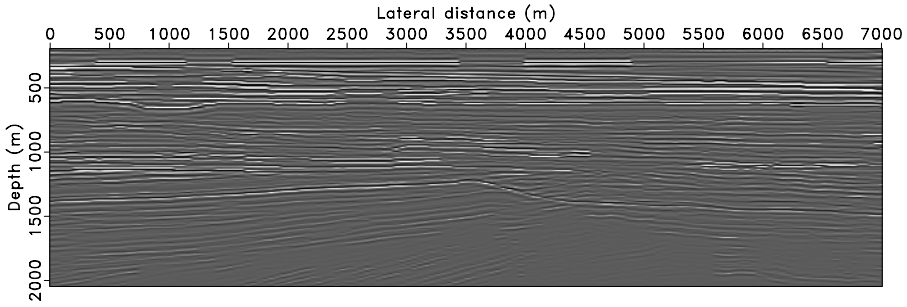


(a)

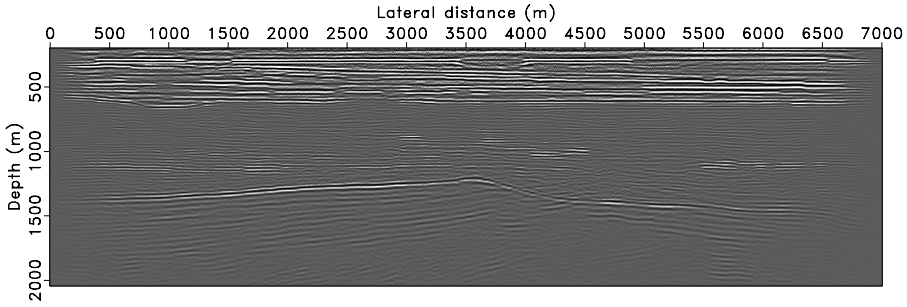


(b)

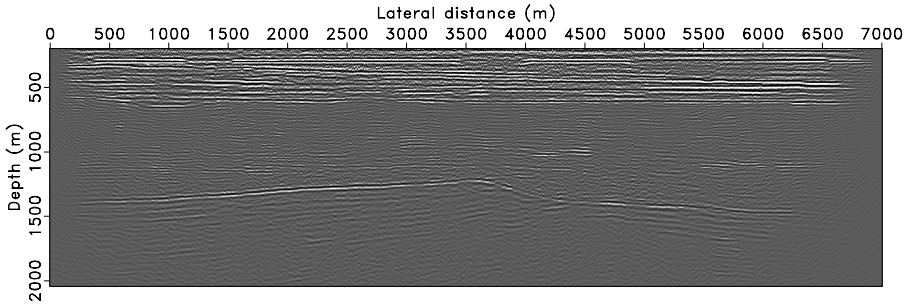
Figure 8: Full-waveform inversion result. **(a)** Initial model. **(b)** Inverted result starting from 2.9Hz with 7 simultaneous shots and 10 frequencies in each of the 10 frequency bands.



(a)



(b)



(c)

Figure 9: Dimensionality-reduced sparsity-promoting imaging from random subsets of 17 simultaneous shots and 10 frequencies. We used the background velocity-model plotted in Figure 8(b) **(a)** True perturbation given by the difference between the true velocity model and the FWI result plotted in Figure 8(b). **(b)** Imaging result with redraws for the supershots. **(c)** The same but without redraws. Notice the significant improvement in image quality when renewing collection of supershots after solving each LASSO subproblem.

tudes by iterating on highly dimensionality reduced subproblems. Not withstanding an ill-conditioned Hessian, spectral-projected gradients makes good progress towards the solution in relatively few matrix-vector multiplies. A possible explanation for this phenomenon and the benefit from redraws is that compressive sampling the sources does not make the dimensionality-reduced Hessian significantly more ill conditioned. This is after all the premise of compressive sensing where the condition number of sampling matrices is being controlled by design. In addition, recent work by Montanari (2010) has shown that redraws help to remove possible correlations between the solution vector and the source encoding and this increases the convergence of solutions based on iterative soft thresholding. We can argue that we are observing this effect empirically. Finally, our approach is also reminiscent of randomized "block Kaczmarz" (Strohmer and Vershynin, 2009) and to recent work by Friedlander and Schmidt (2011); Tristan van Leeuwen and Herrmann. (2011) who also propose sampling strategies. However, finding the best sampling strategy itself is still an open research problem; see, for example, Friedlander and Schmidt (2011); Tristan van Leeuwen and Herrmann. (2011).

Because our formulation includes contributions form the wave-equation Hessian more theoretical work will be necessary to (i) compensate for the "coloring" by this operator, e.g. by solving weighted ℓ_1-norm problems and (ii) analyze the coherence and restricted-isometry properties of the dimensionality reduced Hessian, using practical techniques recently developed by Mansour et al. (2011). These latter results are particularly exciting because they allow for

- an efficient imaging technology with a controllable error. As in CS, this error depends on the subsampling ratio and on the compressibility of the model in the sparsifying domain. This means that our sparse recovery algorithm returns images that can be considered as images we would have obtained by keeping only a small fraction of the largest transform-domain (curvelet) coefficients. The larger the batch size, the larger this fraction, and the better the recovery by virtue of the transform-domain compressibility. Clearly, this property differs fundamentally from Monte-Carlo techniques where the error decays slowly with the batch size.

- an integration of dimensionality reduction with acquisition. For instance, we could envisage "online" acquisition during which simultaneously acquired data is continuously inverted with a procedure reminiscent of the approach outlined in this paper.

CONCLUSIONS

We introduced an efficient algorithm to solve the linearized imaging problem. Our method combines recent findings from the fields of stochastic optimization and compressive sensing and turns the originally "overdetermined" seismic imaging problem into a series of underdetermined dimensionality-reduced subproblems. By considering

these subproblems as sparse-recovery problems, we were able to create high-fidelity images at a fraction of the computational cost. The final image can be considered as the result of curvelet-domain sparsity-promoting migration.

We found this curvelet-domain sparsity promotion essential because it controls the null space of the Hessian and removes source crosstalk due to the randomized dimensionality reduction via supershots. Solvers of the reduced problems that solely rely on least-squares are not able to accomplish this.

In summary, our approach can be seen as an instance of a new randomized dimensionality-reduction paradigm where the costs of computations are no longer dominated by the discretization but by transform domain sparsity of the model. In this new paradigm of randomized inversion, dimensionality reductions allow us to solve (linearized) inversion problems in ways, which previously, would have been computationally infeasible. The examples presented in this paper support this observation and show highly competitive results on synthetic model with realistic complexity.

ACKNOWLEDGMENTS

We are grateful to Charles Jones from BG for providing us with the BG Compass velocity model. We would also like to thank the authors of SPOT (`http://www.cs.ubc.ca/labs/scl/spot/`), SPGℓ_1 (`http://www.cs.ubc.ca/labs/scl/spgl1/`), and CurveLab (`curvelet.org`). This paper was prepared with Madagascar (`rsf.sf.net`) and was in part financially supported by by the CRD Grant DNOISE 334810-05. The industrial sponsors of the Seismic Laboratory for Imaging and Modelling (SLIM) BG Group, BP, Chevron, ConocoPhillips, Petrobras, Total SA., and WesternGeco are also gratefully acknowledged.

REFERENCES

Andersson, F., M. V. de Hoop, H. F. Smith, and G. Uhlmann, 2008, A multi-scale approach to hyperbolic evolution equations with limited smoothness: Communications in Partial Differential Equations, **33**, 988–1017.

Avron, H., and S. Toledo, 2010, Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix.

Ayeni, G., 2010, Seismic reservoir monitoring with permanent encoded seismic arrays: , SEG, 4221–4226.

Berg, E. v., and M. P. Friedlander, 2008, Probing the Pareto frontier for basis pursuit solutions: SIAM Journal on Scientific Computing, **31**, 890–912.

Bertsekas, D., and J. Tsitsiklis, 1996, Neuro-Dynamic Programming (Belmont, MA: Athena Scientific).

Bourgeois, A., M. Bourget, P. Lailly, M. Poulet, P. Ricarte, and R. Versteeg, 1991, Marmousi data and model, *in* The Marmousi experience: EAGE, **5-9**.

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

Candès, E. J., L. Demanet, D. L. Donoho, and L. Ying, 2006, Fast discrete curvelet transforms: Multiscale Modeling and Simulation, **5**, no. 3, 861–899.

Demanet, L., and G. Peyré, 2011, Compressive wave computation: Found. Comput. Math., **11**, no. 3, 257–303.

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

Douma, H., and M. V. de Hoop, 2007, Leading-order seismic imaging using curvelets: Geophysics, **72**, S231–S248.

Fei, T. W., Y. Luo, S. Aramco, and G. T. Schuster, 2010, De-blending reverse-time migration: , SEG, 3130–3134.

Friedlander, M. P., and M. Schmidt, 2011, Hybrid deterministic-stochastic methods for data fitting: Tech. rep., Department of Computer Science, University of British Columbia, Vancouver.

Guitton, A., 2004, Adaptive subtraction of multiples using the l^1-norm: Geophys Prospect, **52**, 27–27.

Guitton, A., G. Ayeni, and G. Gonzales, 2010, A preconditioning scheme for full waveform inversion: SEG, Expanded Abstracts, **29**, 1008–1012.

Habashy, T. M., A. Abubakar, G. Pan, and A. Belani, 2010, Full-waveform seismic inversion using the source-receiver compression approach: , SEG, 1023–1028.

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.

Hansen, P. C., 1997, Rank-Deficient and Discrete Ill-posed Problems: SIAM.

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

Herrmann, F. J., C. R. Brown, Y. A. Erlangga, and P. P. Moghaddam, 2009a, Curvelet-based migration preconditioning and scaling: Geophysics, **74**, A41.

Herrmann, F. J., Y. A. Erlangga, and T. Lin, 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., 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).

Hutchinson, M., 1990, A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines: J. Commun. Statist. Simul, **19**.

Jo, C. H., C. Shin, and J. H. Suh, 1996, An optimal 9-point, finite-difference, frequency-space, 2-D scalar wave extrapolator: Geophysics, **61**, 529–537.

Krebs, J. R., J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M.-D. Lacasse, 2009, Fast full-wavefield seismic inversion using encoded sources: Geophysics, **74**, WCC177–WCC188.

Li, X., A. A. an Tristan van Leeuwen, and F. J. Herrmann, 2011, Modified gauss-newton with sparse updates: Presented at the , SBGF.

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

Lin, T. T. Y., and F. J. Herrmann, 2007, Compressed wavefield extrapolation: Geophysics, **72**, no. 5, SM77–SM93.

Mallat, S. G., 2009, A Wavelet Tour of Signal Processing: the Sparse Way: Academic Press.

Mansour, H., H. Wason, T. T. Lin, and F. J. Herrmann, 2011, Simultaneous-source marine acquisition with compressive sampling matrices: Technical Report TR-2011-02, UBC-Earth and Ocean Sciences Department. (submittedTR).

Montanari, A., 2010, Graphical models concepts in compressed sensing: CoRR, **abs/1011.4328**.

Morton, S. A., and C. C. Ober, 1998, Faster shot-record depth migrations using phase encoding: SEG Technical Program Expanded Abstracts, SEG, 1131–1134.

Mulder, W., and R. Plessix, 2004, How to choose a subset of frequencies in frequency-domain finite-difference migration: Geoph. J. Int., **158**, 801–812.

Natterer, F., 2001, The mathematics of computerized tomography: Society for Industrial Mathematics.

Neelamani, R. N., C. E. Krohn, J. R. Krebs, J. K. Romberg, M. Deffenbaugh, and J. E. Anderson, 2010, Efficient seismic forward modeling using simultaneous random sources and sparsity: Geophysics, **75**, WB15–WB27.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro, 2009, Robust stochastic approximation approach to stochastic programming: SIAM Journal on Optimization, **19**, 1574–1609.

Paige, C. C., and M. A. Saunders, 1982, Lsqr: An algorithm for sparse linear equations and sparse least squares: ACM TOMS, **8**, 43–71.

Plessix, R., and W. Mulder, 2004, Frequency-domain finite difference amplitude-preserving migration: Geoph. J. Int., **157**, 975–987.

Rickett, J. E., 2003, Illumination-based normalization for wave-equation depth migration: Geophysics, **68**, no. 4.

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.

Shapiro, A., D. Dentcheva, and D. Ruszczynski, 2009, Lectures on stochastic programming: Modeling and theory: SIAM, Philadelphia.

Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248.

Smith, H. F., 1998, A Hardy space for Fourier integral operators.: J. Geom. Anal., **8**, 629–653.

Strohmer, T., and R. Vershynin, 2009, A randomized Kaczmarz algorithm with exponential convergence: Journal of Fourier Analysis and Applications, **15**, 262–278.

Symes, W., 2010, Source synthesis for waveform inversion: , SEG, 1018–1022.

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

Prospecting, **56**.

Tibshirani, R., 1997, Regression shrinkage and selection via the lasso: J. Royal. Statist. Soc B., **58**, 267–288.

Tristan van Leeuwen, Mark Schmidt, M. F., and F. Herrmann., 2011, A hybrid stochastic-deterministic optimization method for waveform inversion: Presented at the , EAGE, EAGE Technical Program Expanded Abstracts.

van Leeuwen, T., A. Aravkin, and F. J. Herrmann, 2010a, Seismic waveform inversion by stochastic optimization: Technical Report TR-2010-5, UBC-Earth and Ocean Sciences Department.

van Leeuwen, T., M. Schmidt, M. Friedlander, and F. J. Herrmann, 2010b, A hybrid stochastic-deterministic optimization method for waveform inversion: Technical Report TR-2010-6, UBC-Earth and Ocean Sciences Department.

Result: Estimate for the model $\widetilde{\mathbf{x}}$
$\mathbf{x}_0 \longleftarrow \mathbf{0}$; `// initial model`
$\mathbf{k} \longleftarrow \mathbf{0}$; `// initial counter`
while $\|\mathbf{x}_0 - \widetilde{\mathbf{x}}\|_2 \geq \epsilon$ **do**
  $k \longleftarrow k+1$; `// increase counter`
  $\widetilde{\mathbf{x}} \longleftarrow \mathbf{x}_0$; `// update warm start`
  $\mathbf{RM} \longleftarrow$ Draw($\mathbf{RM}$); `// draw new subsampler`
  $\mathbf{x}_0 \longleftarrow$ Solve($\mathbb{P}(\mathbf{RM}); \widetilde{\mathbf{x}}$); `// solve the subproblem`
end

Algorithm 1: Stochastic-average approximation with warm restarts

Subsample ratio	0.0006	0.0013	0.0026	0.0033
n'_f/n'_s	Signal-noise ratio (dB)			
2	**1.60**	**1.63**	**1.86**	**1.95**
1	**1.62**	**1.75**	**1.87**	**1.99**
0.5	**1.63**	**1.77**	**1.98**	**2.06**
Speed up (×)	1536	768	384	307

Table 1: Signal-to-noise ratios, SNR $= -20\log_{10}(\frac{\|\mathbf{x}-\mathbf{x}_0\|_2}{\|\mathbf{x}\|_2})$ for sparse curvelet-based recovery for different subsample and frequency-to-shot ratios. The vector $\mathbf{x}$ is the inverted perturbation and $\mathbf{x}_0$ is the true perturbation given by the difference between the true and smooth background model.