Fighting the curse of dimensionality: compressive sensing in exploration seismology

Felix J. Herrmann, Michael P. Friedlander, Özgür Yılmaz
January 21, 2012

***Abstract*—Many seismic exploration techniques rely on the collection of massive data volumes that are mined for information during processing. This approach has been extremely successful, but current efforts toward higher-resolution images in increasingly complicated regions of the Earth continue to reveal fundamental shortcomings in our typical workflows. The "curse of dimensionality" is the main roadblock, and is exemplified by Nyquist's sampling criterion, which disproportionately strains current acquisition and processing systems as the size and desired resolution of our survey areas continues to increase.**

We offer an alternative sampling strategy that leverages recent insights from compressive sensing towards seismic acquisition and processing for data that are traditionally considered to be undersampled. The main outcome of this approach is a new technology where acquisition and processing related costs are no longer determined by overly stringent sampling criteria.

Compressive sensing is a novel nonlinear sampling paradigm, effective for acquiring signals that have a sparse representation in some transform domain. We review basic facts about this new sampling paradigm that revolutionized various areas of signal processing, and illustrate how it can be successfully exploited in various problems in seismic exploration to effectively fight the curse of dimensionality.

***Index Terms*—Compressive sensing, curvelet transform, sparsity promotion, exploration seismology, seismic acqusition, seismic imaging, seismic inversion, and convex optimization**

I. The curse of dimensionality in seismic exploration

Modern-day seismic-data processing, imaging, and inversion rely increasingly on computationally and data-intensive techniques to meet society's continued demand for hydrocarbons. This approach is problematic because it leads to exponentially increasing costs as the size of the area of interest increases. Motivated by recent findings from compressive sensing (CS) and earlier work in seismic data regularization [1] and phase encoding [2], we confront the challenge of the "curse of dimensionality" with a randomized dimensionality-reduction approach that decreases the cost of acquisition and subsequent processing significantly. Before we discuss possible solutions to the curse of dimensionality in exploration seismology, we first discuss how sampling is typically conducted in exploration seismology.

Felix J. Herrmann is with the Department of Earth and Ocean Sciences, The University of British Columbia, Vancouver, Canada.

Michael P. Friedlander is with the Department of Computer Science, The University of British Columbia, Vancouver, Canada.

Özgür Yılmaz is with the Department of Mathematics, The University of British Columbia, Vancouver, Canada.

A. Classical approaches

During seismic data acquisition, data volumes are collected that represent dicretizations of analog finite-energy wavefields in up to five dimensions including time. So, we are concerned with the acquisition of an analog spatio-temporal wavefield $\bar{f}(t,x) \in L^2((0,T] \times [-X,X])$ with time T in the order of seconds and length X in the order of kilometers. The sampling intervals are of the order of milliseconds and of meters.

After proper discretization and analog-to-digital conversion, it is convenient to organize these high-resolution samples into a vector $f := \big\{f[q]\big\}_{q=0,\ldots,N-1} \in \mathbb{R}^N$. Note that in practice often we have missing samples, i.e., instead of f, the acquired data is $b = Rf$ where R is a $n \times N$ *restriction matrix* that consists of n rows of the $N \times N$ identity matrix.

B. Bottlenecks and compromises

Unfortunately, pressures for increased resolution make complete sampling ($n = N$) economically and physically infeasible and therefore the data is sampled at a rate below Nyquist, i.e., $n \ll N$. For the spatial coordinates, this typically corresponds to periodic subsampling of the sources/receivers while the total acquisition time is reduced by reducing the time between sequential single-source experiments. Unfortunately, these subsamplings can lead to serious artifacts and a lot of research has recently been devoted to come up with improved sampling schemes that randomize spatial locations of sources and receivers or that randomize the sources, e.g., by random dithering of marine sources or by source encodings on land.

C. Dimensionality reduction by Compressive Sensing

While recent proposals to expedite seismic acquisition or computations through simultaneous sourcing have proven successful, the proposed methods miss a rigorous framework that would allow for the design of rigorous workflows. By recognizing these clever new sampling schemes as instances of CS, we are able to make a start towards sampling and computation strategies that employ structure in seismic data, which translates into transform-domain sparsity. This attribute allows us to come up with sub-Nyquist sampling schemes whose sampling is proportional to the sparsity rather than to the dimensionality of the problem. The success of these techniques hinges on subsamplings that break periodicity of conventional samplings. To demonstrate how this works, we first give a brief introduction to the theory CS, followed by its application to problems in exploration seismology. Recovery

from the subsamplings depends on solving large-scale convex optimization problems, described in §III.

II. Compressive sampling as a dimension reduction method

Various classes of signals such us audio, images, and seismic signals admit *sparse approximations*, i.e., they can be well-approximated by a linear superposition of a few atoms of an appropriate basis. Compressed sensing (or compressive sampling)—championed by Candès, Romberg, and Tao [3] and Donoho [4]—has emerged as a novel paradigm for sensing such signals more efficiently as compared to the classical approach based on Shannon-Nyquist sampling theory. Signals that admit sparse approximations can be acquired from significantly fewer measurements than their ambient dimension using nonlinear recovery algorithms, e.g., ℓ_1 minimization or greedy algorithms However, these greedy algorithms are not suitable for large-scale problems because they only bring a single component into the solution per iteration.

In CS is the number of samples required to achieve a certain accuracy scales logarithmically with the ambient dimension, which, in the case of spatial sampling, is the sampling grid size. Thus, in problems where the sheer number of the measurements to be obtained is prohibitively large, CS is invaluable.

Next, we introduce the mathematical framework of CS and discuss the challenges we face in exploration seismology.

A. Compressive acquisition of sparse signals

The main signal model of CS is nonlinear: the signals are *sparse* (only a few of the entries are non-zero) or *compressible* (can be well-approximated by a sparse signal), either in the canonical basis or in some transform domain. Formally, consider a high-dimensional signal $x \in \mathbb{R}^N$. We first make the naive assumption that x is k-sparse, i.e., $\|x\|_0 \leq k$, where $\|x\|_0$ denotes the number of non-zero entries of the vector x. (We later relax the sparsity assumption to make way for more realistic signal ensembles including seismic.) The goal in CS is to obtain x (or an approximation) from non-adaptive linear measurements $y = \Psi x$, where Ψ is an appropriate full rank $n \times N$ *measurement matrix* with $n \ll N$.

Note that since $n < N$, i.e., the number of measurements is less than the ambient dimension, the system $\Psi z = b$ has infinitely many solutions, rendering it generally impossible to recover x from y. In CS, we aim to recover x by utilizing the prior information that x is sparse (or compressible): find the solution x^* of $\Psi z = b$ with the smallest number of non-zero entries. This is the *sparse recovery problem*. Unfortunately, this problem is NP-hard [5] and sensitive to the sparsity assumption and additive noise, thus not useful in practice. The major breakthrough in CS has been to specify explicit conditions under which the sparse recovery problem is equivalent to

$$\underset{z}{\text{minimize}} \quad \|z\|_1 \quad \text{subject to} \quad \Psi z = b, \tag{II.1}$$

which is computationally tractable. Specifically, these conditions [3], [4] determine what measurement matrices Ψ can be used so that (II.1) is guaranteed to recover all k-sparse x in $\mathbb{R}^N$ from n measurements given by b. In words, the main requirement is that Ψ nearly preserves the length of all sparse vectors.

Various random matrix ensembles have been shown to be effective compressive measurement matrices, e.g., Gaussian and Bernoulli matrices, and Fourier matrices with randomly selected rows. An important question is how the number of measurements required for exact recovery scales with the sparsity level k, the number of measurements n, and the ambient dimension N. (In the classical sampling theory, k is analogous to "bandwidth", n is analogous to sampling frequency, and N is analogous to the size of the sampling grid.) The following theorem, adapted from [3], summarizes one answer to this question.

Theorem 1. *Suppose Ψ is an $n \times N$ Gaussian random matrix. If $n \gtrsim k \log(N/n)$, with overwhelming probability* (II.1) *recovers all k-sparse x from the measurements $y = \Psi x$.*

In words, if the measurement matrix is chosen appropriately, the number of measurements scales only logarithmically with the ambient dimension N—a tremendous improvement over linear scaling of the classical sampling theory. While one-norm minimization is most commonly used, recent work by [6] generalizes to p-norm minimization with $0 < p < 1$.

B. Compressible signals and robustness to noise

For CS to be practicable, two major modifications to the setting described in § II-A need to be considered. First, it is naive to expect signals in practice to be exactly sparse, and a more realistic model is that the magnitude-sorted coefficients decay rapidly, leaving us with a vector with few large entries and many small ones. Such signals can be well approximated by sparse signals and are said to be *compressible*. Second, practical applications typically have measurements contaminated by noise, and it is again crucial that the CS approach is robust in this regard [3].

Theorem 2. *Let $x \in \mathbb{R}^N$ be arbitrary, and let Ψ be an appropriate $n \times N$ measurement matrix (e.g., a Gaussian matrix). Suppose that the noisy measurements are given by $b = \Psi x + e$ where e is additive noise with $\|e\|_2 \leq \epsilon$. Denote by x^* the solution of the following convex program:*

$$\underset{z}{\text{minimize}} \ \|z\|_1 \text{ subject to } \|\Psi z - b\| \leq \epsilon. \tag{II.2}$$

Then for absolute constants C_1 and C_2,

$$\|x - x^*\|_2 \leq C_1 \epsilon + C_2 k^{-1/2} \sigma_k(x)_{\ell_1},$$

whenever $n = \mathcal{O}(k \log[N/n])$ and where $\sigma_k(x)_{\ell_1}$ is the best k-term approximation error.

In words, the recovered approximation is within the noise level and nearly as accurate as the approximation we would obtain by measuring directly the largest k entries of x.

This theorem can play a pivotal role in exploration seismology. Its first main consequence is that sparse recovery from noise-free compressively sampled data gives an error that has, up to a logarithmic factor, the same decay as the

best nonlinear approximation error. This represents a major improvement over linear approximations that may have a much slower decay and hence a much lower quality for the recovery. Second, empirical evidence report by [7] shows that compressive sensing yields small recovery errors, even with low "oversampling ratios". This is underpinned by Theorem 2, which establishes that the recovery error is proportional to that of the best k-term nonlinear approximation error. It also shows that the recovered approximation is within the noise level.

Of course, many of these ideas are related to pioneering work. For example, Claerbout [8] explored sparsity promoting norms; "spiky deconvolution' 'has been analyzed by mathematicians, and randomized acquisition [9] is a precursor. Compressive sensing can be considered as a unifying framework—and more—for all of these approaches. We are motivated by these findings to apply and adapt this framework to solve problems in exploration seismology. See [10], [11] for other applications of compressive sensing in the geosciences.

C. Extensions

So far, the signals of interest were simply sparse in the canonical basis, which, of course, is rarely the case in practice. Images, for example, are sparse with respect to wavelet bases, and seismic signals admit sparse approximations in terms of curvelets [12]–[14]. Formally, consider signals $f \in \mathbb{R}^N$ that are sparse with respect to a basis or frame S, i.e., $f = S^H x$, x sparse. Here, S is a $P \times N$ matrix, with $P \geq N$, that admits a left-inverse; the superscript H denotes the adjoint.

In the next sections we discuss how and to what extent the CS paradigm can be used when S is either an orthonormal basis or a redundant frame.

1) S is an orthonormal basis: In the case when S is an orthonormal basis (i.e., $S^{-1} = S^H$), CS theory applies essentially unchanged. Specifically, the compressive samples of the signal f is given by $b = \Psi f = \Psi S^H x$ where x is sparse (or compressible). In turn, the effective measurement matrix is ΨS^H and if this matrix satisfies the requirements of Theorems 1 and 2, the conclusions of these theorems remain valid.

The main challenge is choosing Ψ for a given S so that ΨS^H is still a good *measurement matrix*. One possible way of constructing a good Ψ tailored to a given sparsity basis S is to first choose an appropriate *measurement basis* M that is *incoherent* with S. The coherence of two bases S and M is reflected by the largest-in-magnitude entry of the matrix MS^H. Once we choose an incoherent M, we discard all but n rows from M and use the resulting $n \times N$ matrix as our measurement matrix. More precisely, we set $\Psi = RM$ where R is an $n \times N$ restriction matrix (consisting of n rows of the $N \times N$ identity matrix). It can be shown that such a Ψ can be used for CS. For example, if the signal u is sparse in Fourier domain, i.e., S is the DFT matrix, then an optimally incoherent measurement basis is given by the $N \times N$ identity basis $M = I_N$. This leads to the following convex program:

$$\underset{z}{\text{minimize}} \ \|z\|_1 \text{ subject to } \|RMS^H z - y\| \leq \epsilon.$$

Note that there is a universal strategy for choosing Ψ that does not require prior knowledge of the sparsity basis S: if we choose Ψ to be an appropriate random measurement matrix then ΨS^H is guaranteed to be also a good measurement matrix independent of the orthonormal basis S.

2) S is a redundant: The problem becomes significantly more challenging if the sparsifying dictionary is overcomplete. This means that the signal f can be decomposed as $f = S^H x$, where S is $P \times N$ with $P > N$ and S admits a left inverse. For example, seismic signals are compressible with respect to curvelets, which are overcomplete. There are some differences in this setup, as compared to the orthonormal bases: (i) the expansion coefficients are not unique, and there are (infinitely) many x that explain the same signal f, and (ii) the columns of S^H must be correlated. Accordingly, the approaches used in the orthonormal case do not readily generalize immediately to this case. Empirically, the CS paradigm has been observed to be effective for acquisition of signals that are compressible with respect to redundant transforms. —see Section IV-A for an empirical study of compressive seismic-data acquisition using curvelet frames as the sparsifying transform.

D. Challenges in seismic

According to CS, successful dimensionality reduction hinges on an incoherent sampling strategy where coherent aliases are turned into relatively harmless white Gaussian noise. The challenges of adapting this approach to exploration seismology are twofold. First, seismic data acquisition is subject to physical constraints on the placement, type, and number of (possibly simultaneous) sources, and numbers of receivers. These constraints in conjunction with the extreme large size of seismic data call for seismic problem-specific solutions. Second, while CS offers significant opportunities for dimensionality reduction, there remain still challenges in adapting the scientific-computing workflow to this new approach, and again, CS offers an opportunity to make computation more efficient.

III. Solving the sparse-optimization problem

The main computational challenge in the dimensionality-reduction techniques that we describe can be ascribed to solving the convex optimization problem

$$\text{BP}_\sigma: \quad \underset{x}{\text{minimize}} \ \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \sigma,$$

where σ is an estimate of the required data misfit, often related to the noise level and model mismatch. (The value $\sigma = 0$ yields the *basis pursuit* problem [15].) The nonsmoothness of the objective is the essential difficulty. If these problems were relatively small, then many of the current workhorse algorithms for convex optimization (e.g., simplex and interior-methods) could be used off-the-shelf. However, these methods typically rely on explicit matrix representations of A. Thus there is now significant effort devoted to developing matrix-free algorithms tailored to these problems, which are typically characterized by large dense matrices. The terrific problem sizes in the seismic context is yet a further challenge: a "small" seismic problem can easily have 2^{21} variables.

One of our aims here is to give a broad view of the main approaches, and to describe the approach used by the

SPGL1 software package [16], [17], which we use routinely for tackling seismic sparse recovery problems.

A. Main approaches

Most approaches for solving BP_σ are based on its "Lagrangian" reformulation

$$\text{QP}_\lambda: \quad \underset{x}{\text{minimize}} \ \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1.$$

The positive parameter λ is related to the Lagrange multiplier of the constraint in BP_σ, and it balances the tradeoff between the two norm of the data misfit and the one norm of the solution, which promotes sparsity. For an appropriate choice of λ, this formulation has the same solution to BP_σ, and thus in some sense these two problems are equivalent. However, except for very special cases, the value of λ that induces the equivalence cannot be determined without first solving BP_σ. The typical approach is thus based on solving a sequence of problems QP_λ defined by a decreasing sequence of parameters λ [18]. This gradually decreases the data misfit, which usually allows more nonzeroes into the solution. The overall process terminates when the data mismatch reaches a prescribed accuracy. As we illustrate later, this can be an inefficient approach that requires the solution of too many subproblems.

Many algorithms are available for solving QP_λ or closely related variations, including the Lasso [19] variation

$$\text{LS}_\tau: \quad \underset{x}{\text{minimize}} \ \tfrac{1}{2}\|Ax-b\|^2 \text{ subject to } \|x\|_1 \le \tau.$$

B. Pareto curve

While great progress has been made addressing the non-smoothness component and selection of appropriate sequences of step lengths, a fundamental problem remains: even if we do have an effective algorithm for QP_λ (or LS_τ), how do we best choose a parameter λ (or τ) that yields a required data misfit? The Pareto curve, which traces the optimal trade-off between the two-norm of the residual $r = b - Ax$ and the one-norm of the solution x, is a helpful tool for visualizing the effect of regularization. Fig. 1 gives a schematic illustration of a the curve and some of its features. Points below the curve are not attainable. Any point on the curve, which is uniquely defined by a given A and b, gives the corresponding values σ (vertical axis) and τ (horizontal axis) that cause BP_σ and LS_τ to have the same solution. The negative of the slope at that point gives the corresponding value of λ that causes QP_λ to have the same solution; e.g., see point ①. Point ② coincides with the solution of BP_σ with $\sigma = \|b\|_2$ and of LS_τ with $\tau = 0$; point ③ coincides with the solution of BP_σ with $\sigma = 0$ and of LS_τ with $\tau = 0$. Left- and right-hand limits can be used to define the value of λ at points ② and ③. The relevance of this curve in the seismic context is discussed by [20].

The Pareto curve can be interpreted as the graph of the value function

$$\phi(\tau) = \inf_{\|x\|_1 \le \tau} \{\|Ax-b\|_2\}.$$

Let x_τ be the optimal solution of LS_τ, and let $r_\tau = b - Ax_\tau$ be the corresponding residual. Let $\bar{\tau}$ be the smallest value of τ at which the graph first touches the horizontal axis. (This is guaranteed if A has full rank.) The function ϕ and the Pareto curve is characterized by the following theorem, due to van den Berg and Friedlander [16].

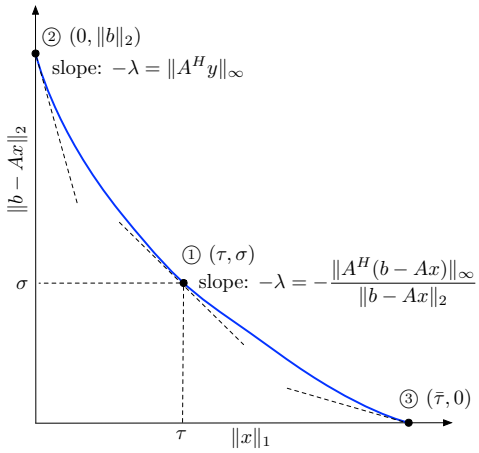
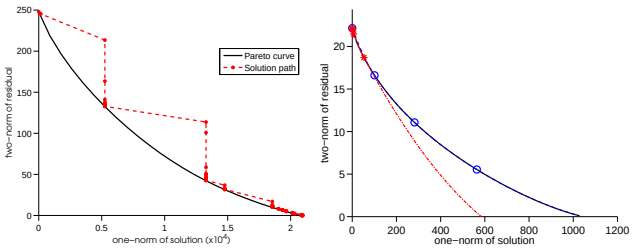


Fig. 1: (Adapted from [20].) Schematic illustration of a Pareto curve. Point ① exposes the connection between the three parameters of QP_λ, BP_σ, and LS_τ. Point ③ corresponds to a solution of BP_σ with $\sigma = 0$.



(a) (b)

Fig. 2: (Adapted from [16].) (a) A typical Pareto curve, and the path taken by the SPGL1 algorithm; (b) approximating the Pareto curve from a few samples.

Theorem 3. *Suppose that A is full rank. Then*

1) *The function ϕ is convex and nonincreasing.*
2) *For all $\tau \in (0, \bar{\tau})$, ϕ is continuously differentiable, $\phi'(\tau) = -\lambda_\tau$, where $\lambda_\tau = \|A^H y_\tau\|_\infty$ and $y_\tau = r_\tau / \|r_\tau\|_2$.*
3) *For $\tau \in [0, \bar{\tau}]$, $\|x_\tau\|_1 = \tau$, and ϕ is strictly decreasing.*

The solid curve in Fig. 2(a) graphs the Pareto curve for a seismic interpolation problem similar to that shown in Fig. 6.

Although QP_λ has proven to be the most used approach, it is generally not clear how to choose the parameter λ such its solution gives a desired misfit. This difficulty is illustrated by Fig. 2(b). The solid (black) curve is the true Pareto curve; the three solid (red) dots are solution/residual pairs of QP_λ that correspond to equally spaced values of λ between $\|A^H b\|_\infty$ and 0. This is typical behavior: even though the values of λ are equally spaced, the resulting samples are not equally spaced, and a quadratic interpolation/extrapolation (dotted red line) based on these samples severly underestimates the curve. On the other hand, the solution/residual pairs of BP_σ (blue circles) for equally spaced samples of σ between $\|b\|_2$ and 0

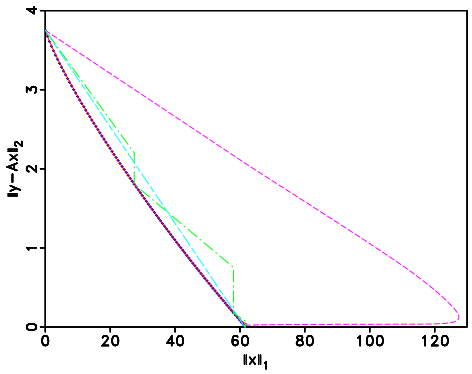


Fig. 3: (Adapted from [20].) Pareto curve and solution paths of four solvers for a BP_σ, with $\sigma = 0$. The symbols + represent a sampling of the Pareto curve. The solid (—) line, obscured by the Pareto curve, is the solution path of IST with cooling, the chain (– · –) line the path of SPGL1, the dashed (– –) line the path of IST, and the dotted $(\cdots)$ line the path of IRLS.

yield good coverage, and an estimate of the curve based on these samples (blue solid line) closely approximates the true curve.

C. Pareto root-finding

The SPGL1 approach for BP_σ is based on approximately solving a sequence of subproblems LS_τ, using a spectral projected-gradient method;at each iteration k it refines the estimate τ_k such that $\tau_k \to \tau_\sigma$, which causes LS_τ and BP_σ to share a solution. The sequence of estimates τ_k is derived by simply applying Newton's method to find a root of the nonlinear equation

$$\phi(\tau) = \sigma.$$

Theorem 3 is central to this approach because it describes how the gradient of ϕ, needed for Newton's method, is related to the solutions of the LS_τ subproblems. Practical refinements are needed that allow for LS_τ to be solved only approximately [16, §3], [21, §3]. Fig. 3 shows how SPGL1 and IST (with cooling) closely follow the Pareto curve; however, SPGL1 requires significantly fewer matrix multiplies.

IV. Compressive seismic-data acquisition

Perhaps it is too early to claim that CS will constitute a paradigm shift in seismic acquisition. The first breakthrough was the identification of seismic data regularization and simultaneous/continuous acquisition as instances of CS [22]. Further encouraging progress has been made in the selection of the sparsifying transform and the design of randomized sampling schemes that are realizable in the field.

We discuss progress in each of these areas by means of carefully designed examples that include real field data.

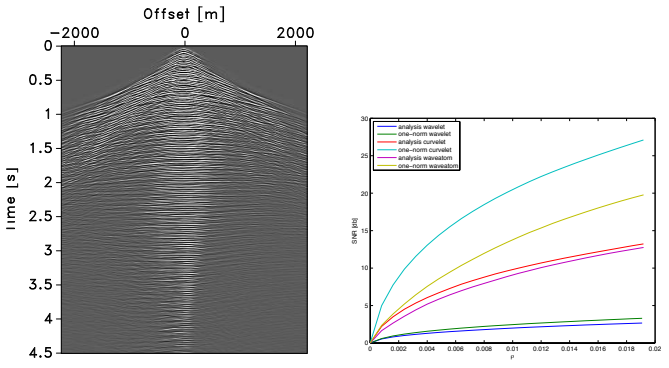
A. Selection of the sparsifying transform

CS leverages structure within signals to reduce the required sampling rates. Typically, this structure translates into compressible representations, using an appropriate transform, that concentrate the signal's energy into a small percentage of large coefficients. The size of seismic data volumes, along with the complexity of its high-dimensional and highly directional wavefront-like features, makes it difficult to find a transform that accomplishes this task.

We thus only consider transforms that are fast ($\mathcal{O}(N \log N)$), multiscale (split the Fourier spectrum into dyadic frequency bands), and multidirectional (split the Fourier spectrum into second dyadic angular wedges). For completeness, we also include separable 2-D wavelets in our study. Unlike wavelets, which compose curved wavefronts into a superposition of multiscale "fat dots" with limited directionality, curvelets [13] and wave atoms [14] compose wavefields as a superposition of highly anisotropic, localized, and multiscale waveforms, which obey the so-called parabolic-scaling principle. For curvelets, this principle translates into a support where length is proportional to the square of the width. At fine scales, this leads to needle-like curvelets. Curvelets, with their near invariance under wave propagation, are thus highly suitable for compressing seismic data. Wave atoms share with curvelets this invariance, and they are also anisotropic because their wavelength depends quadratically on their width. While curvelets are optimal for data with delta-like wavefronts, wave atoms are more appropriate for compressing data with oscillatory wavefronts. Seismic data sits somewhere between these two extremes, and we include both transforms in our study. Since no other tilings of phase space share the hyperbolic scaling principle, we will aside from wavelets not consider other candidates for non-adaptive sparsifying transforms.

1) Approximation error: For an appropriately chosen representation magnitude-sorted transform-domain coefficients often decay rapidly. For orthonormal bases, the decay rate is directly linked to the decay of the nonlinear approximation error. This error can be expressed by $\sigma_k(f)_{\ell_2} := \min_{x \in \Sigma_k^N} \|f - S^H x\|_2$, where f_k is optimal argument, which gives the best k-term approximation in the ℓ_2 sense.When S is orthonormal, f_k is uniquely determined by taking the largest-in-magnitude k-entries of Sx. Unfortunately, such a direct way of finding f_k is not available when S is redundant, because redundant expansions are not unique: there are many coefficient sequences that explain the discrete data f, and these different sequences may have varying decay rates.

To address this issue, we use an alternative definition for the nonlinear approximation error, which is based on the solution of a sparsity-promoting program. With this definition, the k-term sparse approximation error is computed by taking the k-largest coefficients from the vector that solves $\text{minimize}_x \ \|x\|_1 \ \text{ subject to } \ S^H x = f$, where the $P \times N$ matrix S is the redundant ($P > N$) curvelet analysis operator. This solution is typically sparser than the vector obtained by applying the analysis operator S directly. To be able to compare various redundant transforms with different degrees of redundancy, we study the signal-to-noise ratio $\text{SNR}(\rho) = -20 \log \frac{\|f - f_{\rho P}\|}{\|f\|}$, where $\rho = k/P$ is a compression ratio. A smaller ratio implies a larger fraction of ignored coefficients and sparser transform-coefficient vector, which leads to a smaller SNR. In our study, we include $f_{\rho P}$ that are derived



(a) (b)

Fig. 4: (Adapted from [7]) Signal-to-noise ratios (SNRs) for the nonlinear approximation errors of a common-receiver gather (a) from a Gulf of Suez data set. The SNRs (b) are plotted as a function of the sparsity ratio $\rho \in (0,\, 0.02]$. The plots include curves for the errors obtained from the analysis and one-norm minimized synthesis coefficients. Notice the significant improvement in SNRs for the synthesis coefficients obtained by sparsity promotion.

from either the analysis coefficients, i.e., the largest ρP coefficients of Sf, or from the synthesis coefficients that are solutions of the above sparsity-promoting program.

2) Empirical approximation errors: Parametrizing the SNR by ρ allows us to compare the recovery quality of seismic data using various transforms, such as wavelets, curvelets, and wave atoms. Figure 4 compares the performance of these transforms on a common-receiver gather extracted from a Gulf of Suez dataset. Our results in Figure 4 clearly show that curvelets and wave atoms benefit significantly from sparsity promotion, though wave atoms lag behind curvelets. This effect is most pronounced for synthesis coefficients. Because wavelets are orthogonal, they can not benefit, as expected. Note that the observed behavior is consistent with the degree of redundancy of each transform: the curvelet transform has the largest redundancy (a factor of about eight in 2-D), wave atoms have only a redundancy of two, and wavelets are not redundant. This suggests that seismic data processing [12] including sparse recovery from subsampling would potentially benefit most from curvelets. However, this may not be the only factor that determines the performance of CS.

B. Acquisition schemes

Before discussing the application of CS to realistic data examples, we briefly discuss differences between recovery from missing shots, which is an instance of seismic data regularization, and recovery from simultaneous data. The seismic data regularization problem can be considered as the seismic-version of inpainting.

Mathematically, sequential and simultaneous acquisition only differ in the definition of the measurement basis. For sequential-source acquisition, this sampling matrix is given by the Kronecker product of two identity bases—i.e., $I \stackrel{\text{def}}{=} I_{N_s} \otimes I_{N_t}$, which is the $N \times N$ identity matrix where $N = N_s N_t$, the product of the number of shots N_s and the number of time samples N_t. For simultaneous acquisition, where all sources fire simultaneously, this matrix is given by $M \stackrel{\text{def}}{=} G_{N_s} \otimes I_{N_t}$ with G_{N_s} an $N_s \times N_s$ Gaussian matrix with *i.i.d.* entries. In both cases, we use a restriction operator $R \stackrel{\text{def}}{=} R_{n_s} \otimes I_{N_t}$ to model the collection of incomplete data by reducing the number of shots to $n_s \ll N_s$. This restriction acts on the source coordinate only. For both recovery experiments, we use 2-D curvelets as the sparsifying transform S (cf. II-C). Note that these two randomized samplings do not necessarily reflect practical sampling scenarios but are merely intended to demonstrate the impact of different randomized sampling strategies on the recovery performance.

CS predicts superior recovery for compressive-sampling matrices with smaller coherence. This coherence depends on the interplay between the restriction, measurement, and synthesis matrices. To make a fair comparison, we keep the randomized restriction matrix the same and compare the recoveries for measurement matrices given by the identity or by a random Gaussian matrix. Physically, the first CS experiment corresponds to surveys with sequential shots missing. The second CS experiment corresponds to simultaneous-source experiments with half of the experiments missing. Examples of both measurements for the real common-receiver gather of Figure 4 are plotted in Figure 5. Both CS experiments are using the same amount of data and illustrate the importance of the choice of the compressive-sensing matrix, which determines acquisition.

Comparing the recovery quality for data for both experiments confirms the insight from CS that states that incoherent measurement matrices favor sparsity-promoting recovery. This opens the possibility of designing efficient acquisition strategies in the field, or of dimensionality reduction in the computer.

Example: coil sampling. The quality of 3-D seismic imaging and full-waveform inversion depends largely on azimuthal coverage. While full-azimuth acquisition is getting within reach on land, full-azimuth sampling in marine remains challenging because of physical constraints on the length, number, and maneuverability of towed streamer arrays with hydrophones.

Moldoveanu addresses these physical limitations of marine acquisition by adopting Hennenfent and Herrmann's jittered-sampling approach [23] to devise a full-coverage randomized coil-sampling strategy where several vessels shoot in tandem while navigating along coils whose centers are randomized [24]. This new sampling paradigm not only avoids time consuming turns at the boundaries of the acquisition area but it also removes severe sampling issues in the cross-line direction related to the number of streamers each vessel can tow.

To illustrate the performance of sparsity-promoting recovery for this type of sampling, we consider a binned common-azimuth/offset volume simulated from a synthetic model with 66% missing data. The recovery results are included in Figure 6 for a cross section. To limit the memory imprint and to exploit parellel computer architectures, we use curvelets in the lateral directions and wavelets along the time axis to sparsify. The curvelets exploit continuity of the wavefield

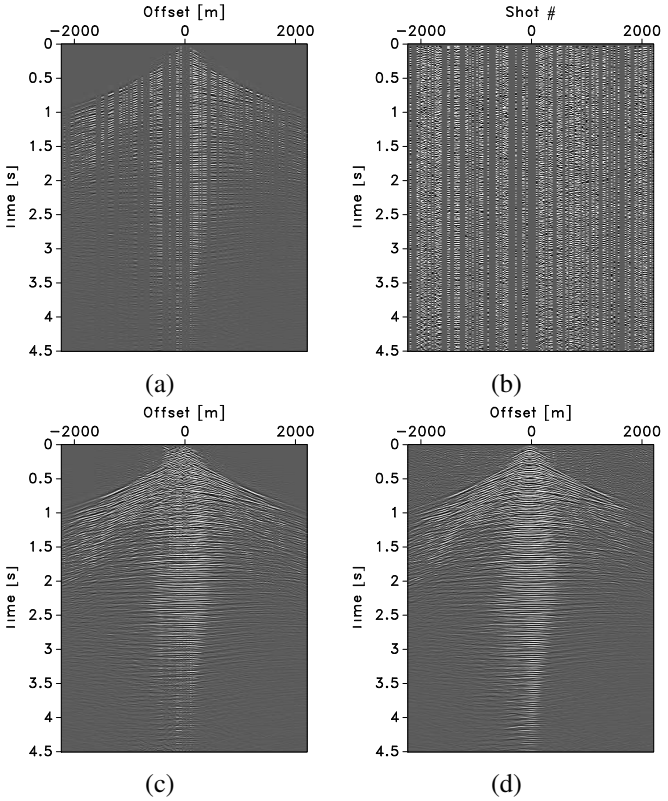


Fig. 5: (Adapted from [7]) Recovery from a compressively-sampled common-receiver gather with 50% of the sources missing. **(a)** Receiver gather with sequential shots selected uniformly at random. **(b)** The same but for random simultaneous shots. **(c)** Recovery from incomplete data in **(a)**. **(d)** The same but now for the data in **(b)**. Notice the remarkable improvement in the recovery from simultaneous data.

along the source-receiver directions while wavelets capture the transient nature of the wavefronts. While this approach does not exploit continuity of the wavefield in all directions, it leads to a feasible algorithm that recovers vectors with several billions of transform coefficients with a relatively low number of iterations of the one-norm solver SPGL1 [16]. The results show excellent recovery from this sampling even in regions of large complexity. This example is a good illustration of the validity of this technology on industry-type data volumes. However, this approach, which is still expensive, is not the only possibility offered by compressive sensing. For instance, compressive sensing also gives unique insights (simultaneous) random time-dithered acquisition, where acquisition is made more efficient by compressing the average interleave time between shots possibly in combination with randomly dispersed transducers on the sea bottom. Remember, the performance of these instances of randomized acquisition rely on choices for the sparsifying transforms and solutions of large-scale sparsity-promoting programs to recover fully sampled data.

V. Compressive seismic computation

We have so far concentrated on applying CS to seismic acquisition. While the invocation of CS in acquisition potentially reaps major increases in efficiency, CS can also be applied to increase the efficiency of wavefield simulations, imaging, and inversion.

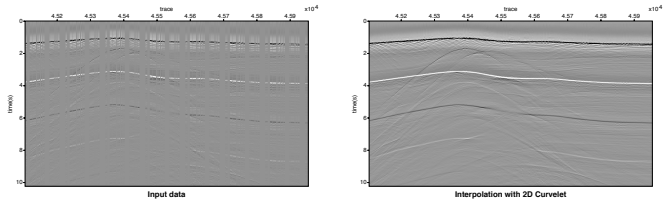


Fig. 6: Synthetic common-azimuth/offset example of coil sampling. The data is simulated with finite-differences on the SEAM model. Cross-section with 66% missing traces and recovered section. Notice the excellent recovery even in regions with strong complexity.

A. Compressive simulation

To simulate seismic surveys, one needs for each source experiment to solve a large linear system that discretizes the underlying wave equation. Because seismic surveys consist of many source experiments, we must reduce the number of PDE solves by exploiting linearity of the wave equation in the sources. This linearity allows us to combine sequential sources into a smaller number of "supershots", each consisting of a random superposition of all sequential shots. Neelamani et al. [25] and Herrmann et al. [22] identify this principle, also known as "phase encoding" [2], as an instance of CS, and demonstrate that it can be used to make wave simulations more efficient by reducing the number of sources.

This technique allowed us to significantly speedup simulations with the time-harmonic Helmholtz equation. Used in combination with randomized importance sampling in the temporal frequency band, we achieved speedups proportional to the subsampling ratio. As shown in Figure 7, sequential simulations can be recovered from a 4-times reduced data volume by solving a sparsity-promoting program with a cost of $\mathcal{O}(n^3 \log n)$, where n the number of sources, receivers, depth levels, and frequencies. This additional cost is very small compared to the cost of solving the Helmholtz system, which is $\mathcal{O}(n^4)$.

B. Compressive imaging

Even though useful for wavefield simulations, compressive simulation is not particularly suitable for making wave-equation based seismic imaging more efficient because we would have to solve a sparsity-promoting program for each PDE solve. To overcome this problem, [26] proposes to image directly with simultaneous data, making the least-squares migration problem

$$\underset{\Delta x}{\text{minimize}} \ \frac{1}{2K}\sum_{i=1}^{K}\|\Delta b_i - A_i \Delta x\|_2^2 = \frac{1}{2K}\|\Delta b - A\Delta x\|_2^2 \tag{V.1}$$

more efficient. Here, Δb_i are the vectorized monochromatic shot records with linearized data (residue), A_i is the linearized

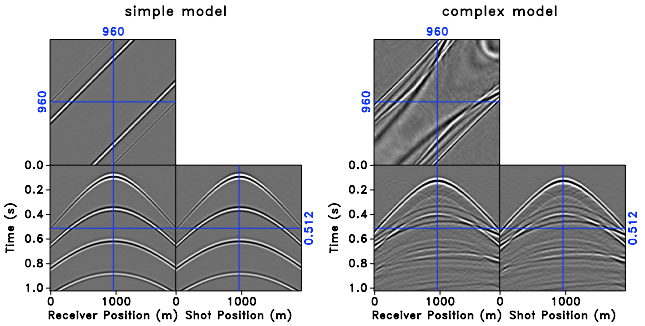


Fig. 7: (Adapted from [22]) Recovery from compressive simulations for simple and complex velocity models. **(a)** Seismic line for the simple model (with a SNR of 28.1 dB) from 25 % of the samples with the ℓ_1-solver running to convergence. **(b)** The same for the complex model but now with a SNR of 18.2 dB.

Born scattering matrix, and Δx the unknown seismic image with the medium perturbations. (The quantities Δb and A aggregate the data across experiments.) In certain cases, the residual $\Delta b_i - A_i \Delta x$ can be interpreted as a linearization of the forward map defined by the wave equation; see (V.2).

This is a difficult problem because each iteration requires a large number of PDE solves, in particular, $4K$ solves, where $K = N_f \cdot N_s$, and N_f and N_s are the number of frequencies and sources. In order to do the inversion, we must be careful to limit the cost of each matrix-vector multiply, which we accomplish by dimensionality reduction. In particular, we use the same supershots as defined in section V-A.

The optimization problem defined by (V.1), however, differs fundamentally from the standard CS problem because now the system is overdetermined—there are more equations then unknowns. As before, we turn this expensive to evaluate overdetermined problem into an underdetermined problem by replacing sequential sources by a reduced number of simultaneous sources. Again, curvelets are used to mitigate the source crosstalk related to this dimensionality reduction.

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. To increase the convergence of SPGL1 for these expensive iterations, we use recent insights from approximate-message passing [27], which are aimed at removing correlations that develop between the model iterate and the randomized Born-scattering operator. Inspired by [28], we remove these correlations by selecting different subsets of random source-encoded supershots [29] after each time SPGL1 reaches the Pareto curve, see Figure 2(a).

To demonstrate the uplift from drawing renewals for the randomized subsets, we include imaging results for the BG Compass model in Figure 8. The experiments are carried out with only 17 simultaneous shots (opposed to 350 sequential shots) with 10 frequencies selected from the interval $20 - 50$ Hz. The results after solving 10 subproblems of SPGL1 clearly indicate that renewals lead to superior image quality and improved convergence as reported by [26].

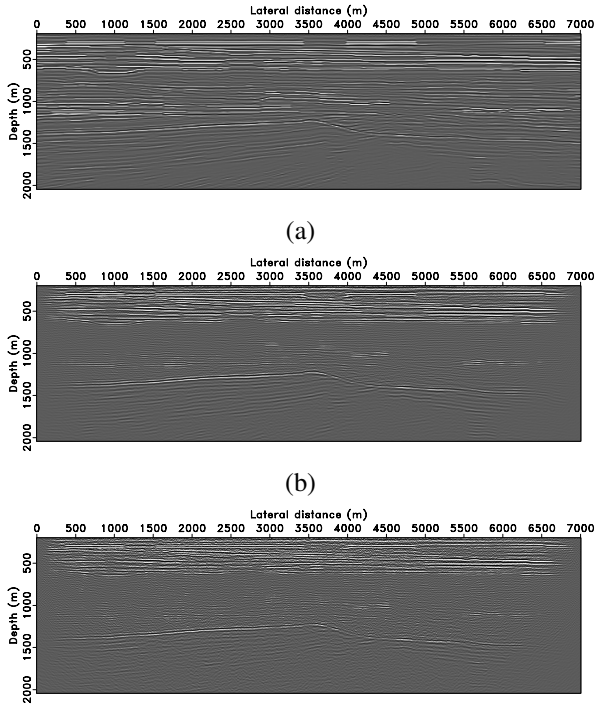


Fig. 8: (Adapted from [26]) Dimensionality-reduced sparsity-promoting imaging from random subsets of 17 simultaneous shots and 10 frequencies. We used the background velocity-model plotted in Figure 9(c) **(a)** True perturbation given by the difference between the true velocity model and the FWI result plotted in Figure 9(c). **(b)** Imaging result with redraws for the supershots. **(c)** The same but without renewals.

We obtained this 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.

C. Compressive inversion

As we reported in earlier work (e.g., see [30]), the cost of computing gradient and Newton updates in full-waveform inversion (FWI) is one of the major impediments that prevents successful adaptation of this industry-size problems. FWI involves the solution of an multi-experiment unconstrained optimization problem (cf. (V.1) for the linearized case):

$$\underset{m}{\text{minimize}} \ \frac{1}{2K} \sum_{i=1}^{K} \|b_i - \mathcal{F}_i[m, q_i]\|_2^2, \qquad \text{(V.2)}$$

with b_i monochromatic shot records with the Earth response to monochromatic sources q_i, and $\mathcal{F}_i[m, q_i]$ represents monochromatic nonlinear forward operators. This operator is parameterized by the velocity m.

To overcome the computational burden of solving (V.2), we follow a similar procedure as outlined in Section V-B but with the difference that we do a new linearization after each of the SPGL1 subproblems.

We test our algorithm on the synthetic model plotted in

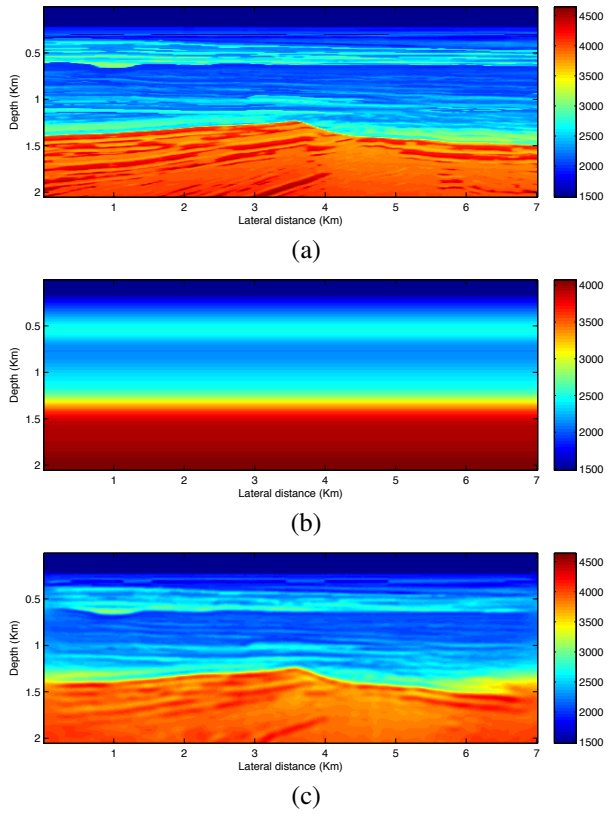


Fig. 9: Full-waveform inversion result. (a) Initial model. (b) True model. (c) Inverted result starting from 2.9Hz with 7 simultaneous shots and 10 frequencies.

Fig. 9(a), which we use to generate data with a source signature given by a $12\,\mathrm{Hz}$ Ricker wavelet. To mimic practice, we use a smooth starting model without lateral information (Fig. 9(b)) and we start the inversion at $2.9\,\mathrm{Hz}$. This means that the seismic data carries relatively little low-frequency information. All simulations are carried out with 350 shot positions sampled at a $20\mathrm{m}$ interval and 701 receiver positions sampled at a $10\mathrm{m}$ interval, yielding an maximum offset of $7\mathrm{km}$. To improve convergence, the inversions are carried out sequentially in 10 overlapping frequency bands on the interval $2.9-22.5\mathrm{Hz}$, each using 7 different simultaneous shots and 10 selected frequencies. For each subproblem, we use roughly 20 iterations of SPGL1 at a cost roughly equivalent to one tenth of the cost of a gradient calculation using all sources. The result for each frequency band after 10 SPGL1 subproblems is depicted in Fig. 9(c). We can see from this result that our inversion captures more or less all discontinuities with a resolution commensurate with the frequency range over which we carry out the inversion. This remarkable result combines the randomized-dimensionality and CS approaches with recent insights from stochastic optimization. As before, drawing independent supershots after solving each SPGL1 subproblem benefited our results [29].

As before, we reduce the computation costs of minimizing or of solving problem (V.2) by randomizing source superposition. Choosing a different collection of supershots for each subproblem gives superior results.

VI. Discussion and conclusions

We discussed possible adaptations of CS to solve outstanding problems in exploration seismology including measures to make acquisition and computations more efficient. The presented results illustrate that we are at the cusp of exciting new developments where acquisition and processing workflows are not hampered by the fear of creating coherent artifacts related to periodic subsampling. Instead, we arrive at a workflow with control over these artifacts. This is accomplished by the following three new design principles, and the slogan "randomize, sparsify, and convexify":

- randomize—break coherent aliases by introducing randomness, e.g., by designing randomly perturbed acquisition grids, or by designing randomized simultaneous sources;
- sparsify—use sparsifying transforms in conjunction with sparsity-promoting programs that separate signal from subsampling artifacts, and that restore amplitudes;
- convexify—relax difficult combinatorial problems into tractable convex optimization problems.

The potential benefits of CS are real and significant. But to realize them, several obstacles need to be surmounted, including the need to overcome the inertia of entrenched engineering practices, and adapting the theoretical framework to practical acquisition schemes and workflows for imaging and inversion.

The seismic application of CS and its extensions rely on solving extremely large system of equations that arise from the physical setting of exploration seismology. This puts pressures on developing large-scale solvers that can handle massive data volumes.

VII. Acknowledgments

We are grateful to Nick Moldoveanu (WesternGeco) for making the coil dataset available and Charles Jones (BG) for providing us with the BG Compass velocity model. We also would like to thank Haneet Wason, Hassan Mansour, Tim Lin, and Xiang Li for preparing the figures. This publication was prepared using CurveLab—a toolbox implementing the Fast Discrete Curvelet Transform, WaveAtom—a toolbox implementing the Wave Atom transform, SPGL1—a solver for large-scale sparse reconstruction, and SPOT—a linear-operator toolbox. The authors were financially supported by CRD Grant DNOISE 334810-05 and by the industrial sponsors of the Seismic Laboratory for Imaging and Modelling: BG Group, BGP, BP, Chevron, ConocoPhilips, Petrobras, PGS, Total SA, and WesternGeco.

References

[1] M. D. Sacchi, T. J. Ulrych, and C. J. Walker, "Interpolation and extrapolation using a high-resolution discrete Fourier transform," *IEEE Transactions on Signal Processing*, vol. 46, no. 1, pp. 31 – 38, January 1998.

[2] L. A. Romero, D. C. Ghiglia, C. C. Ober, and S. A. Morton, "Phase encoding of shot records in prestack migration," *Geophysics*, vol. 65, no. 2, pp. 426–436, 2000. [Online]. Available: http://link.aip.org/link/?GPY/65/426/1

[3] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[4] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, April 1995. [Online]. Available: http://portal.acm.org/citation.cfm?id=207985.207987

[6] R. Saab and Ö. Yılmaz, "Sparse recovery by non-convex optimization – instance optimality," *Applied and Computational Harmonic Analysis*, vol. 29, no. 1, pp. 30–48, 2010.

[7] F. J. Herrmann, "Randomized sampling and sparsity: Getting more information from fewer samples," *Geophysics*, vol. 75, no. 6, pp. WB173–WB187, 2010. [Online]. Available: http://link.aip.org/link/?GPYSA7/75/WB173/1

[8] J. Claerbout and F. Muir, "Robust modeling with erratic data," *Geophysics*, vol. 38, no. 05, pp. 826–844, 1973.

[9] Y. Sun, G. T. Schuster, and K. Sikorski, "A quasi-Monte Carlo approach to 3-D migration: Theory," *Geophysics*, vol. 62, no. 3, pp. 918–928, 1997.

[10] B. Jafarpourand, V. K. Goyal, D. B. McLaughlin, and W. T. Freeman, "Compressed history matching: Exploiting transform-domain sparsity for regularization of nonlinear dynamic data integration problems," *Mathematical Geosciences*, vol. 42, no. 1, pp. 1–27, 2009. [Online]. Available: http://www.springerlink.com/index/10.1007/s11004-009-9247-z

[11] J. Ma, "Compressed sensing for surface characterization and metrology," *Instrumentation and Measurement, IEEE Transactions on*, vol. 59, no. 6, pp. 1600 –1615, june 2010.

[12] G. Hennenfent and F. J. Herrmann, "Seismic denoising with non-uniformly sampled curvelets," *IEEE Comp. in Sci. and Eng.*, vol. 8, no. 3, pp. 16–25, 2006.

[13] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying, "Fast discrete curvelet transforms," *SIAM Multiscale Model. Simul.*, vol. 5, no. 3, pp. 861–899, 2006. [Online]. Available: www.curvelet.org

[14] L. Demanet and L. Ying, "Wave atoms and sparsity of oscillatory patterns," *Applied and Computational Harmonic Analysis*, vol. 23, no. 3, pp. 368–387, 2007.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[16] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[17] ——, "SPGL1: A solver for large-scale sparse reconstruction," Available at http://www.cs.ubc.ca/labs/scl/index.php/Main/Spgl1, June 2007.

[18] F. J. Herrmann and G. Hennenfent, "Non-parametric seismic data recovery with curvelet frames," *Geophysical Journal International*, vol. 173, pp. 233–248, April 2008.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1997.

[20] G. Hennenfent, E. van den Berg, M. P. Friedlander, and F. J. Herrmann, "New insights into one-norm solvers from the pareto curve," *Geophysics*, vol. 73, no. 4, pp. A23–26, 2008. [Online]. Available: http://geophysics.geoscienceworld.org/cgi/content/abstract/73/4/A23

[21] E. van den Berg and M. P. Friedlander, "Sparse optimization with least-squares constraints," *SIAM Journal on Optimization*, vol. 21, no. 4, pp. 1201–1229, 2011. [Online]. Available: http://link.aip.org/link/?SJE/21/1201/1

[22] F. J. Herrmann, Y. A. Erlangga, and T. Lin, "Compressive simultaneous full-waveform simulation," *Geophysics*, vol. 74, p. A35, 2009.

[23] G. Hennenfent and F. J. Herrmann, "Simply denoise: wavefield reconstruction via jittered undersampling," *Geophysics*, vol. 73, no. 3, May-June 2008.

[24] N. Moldoveanu, "Random sampling: A new strategy for marine acquisition," *SEG Technical Program Expanded Abstracts*, vol. 29, no. 1, pp. 51–55, 2010. [Online]. Available: http://link.aip.org/link/?SGA/29/51/1

[25] N. Neelamani, C. Krohn, J. Krebs, M. Deffenbaugh, and J. Romberg, "Efficient seismic forward modeling using simultaneous random sources and sparsity," in *SEG International Exposition and 78th Annual Meeting*, 2008, pp. 2107–2110. [Online]. Available: http://users.ece.gatech.edu/~justin/Publications_files/simulseg2008.pdf

[26] F. J. Herrmann and X. Li, "Efficient least-squares imaging with sparsity promotion and compressive sensing," 08/2011 2011, to appear in Geophysical Prospecting. [Online]. Available: http://slim.eos.ubc.ca/Publications/private/Journals/leastsquaresimag.p%df

[27] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[28] A. Montanari, "Graphical models concepts in compressed sensing," *CoRR*, vol. abs/1011.4328, 2010.

[29] F. J. Herrmann, X. Li, A. Aravkin, and T. van Leeuwen, "A modified, sparsity promoting, Gauss-Newton algorithm for seismic waveform inversion." *Proc. SPIE*, vol. 2011, no. 81380V, 2011. [Online]. Available: http://slim.eos.ubc.ca/Publications/public/Journals/SPIEreport.pdf

[30] X. Li and F. J. Herrmann, "Full-waveform inversion from compressively recovered model updates," vol. 29, no. 1. SEG, 2010, pp. 1029–1033. [Online]. Available: http://link.aip.org/link/?SGA/29/1029/1