Accelerated large-scale inversion with message passing

Felix J. Herrmann, the University of British Columbia, Canada

SUMMARY

To meet current-day challenges, exploration seismology increasingly relies on more and more sophisticated algorithms that require multiple paths through all data. This requirement leads to problems because the size of seismic data volumes is increasing exponentially, exposing bottlenecks in IO and computational capability. To overcome these bottlenecks, we follow recent trends in machine learning and compressive sensing by proposing a sparsity-promoting inversion technique that works on small randomized subsets of data only. We boost the performance of this algorithm significantly by modifying a state-of-the-art ℓ_1-norm solver to benefit from message passing, which breaks the build up of correlations between model iterates and the randomized linear forward model. We demonstrate the performance of this algorithm on a toy sparse-recovery problem and on a realistic reverse-time-migration example with random source encoding. The improvements in speed, memory use, and output quality are truly remarkable.

INTRODUCTION

Modern-day exploration seismology relies on the solution of large-scale optimization problems typically requiring multiple passes through all data. Because seismic inversions also depend on long-offset and full-azimuth sampling, this leads to exponentially growing costs in the collection, storage, and processing of these data volumes. During this talk, we discuss how recent insights from compressive sensing, message passing in machine learning, and convex optimization can be used to reduce these costs by working on small randomized subsets of data (e.g. small numbers of super shots) while minimizing the number of passes through all data. To be more specific, we adapt approximate message passing (AMP, see e.g. Donoho et al., 2009) to the seismic situation where realities of acquisition and the large problem size make it difficult to meet the assumptions of AMP in practice. To arrive at our formulation, we first briefly outline ideas behind randomized dimensionality reduction and compressive sensing (CS Candès et al., 2006; Donoho, 2006). Next, we describe how to solve large-scale sparsity-promoting problems, followed by a discussion on AMP that aims to improve the convergence of the solvers. To meet challenges specific to exploration seismology, we adapt AMP by proposing a nontrivial modification of the spectral-gradient method SPGℓ_1 (van den Berg and Friedlander, 2008). As demonstrated by the examples, this modification allow us to reap the benefits of message passing without sacrificing flexibility of the ℓ_1-norm solver.

RANDOMIZED DIMENSIONALITY REDUCTION

Following trends in many research areas, there has been a recent push to randomize certain aspects of seismic acquisition. Examples of this trend are simultaneous and continuous marine acquisition (see e.g. Hampson et al., 2008; Herrmann et al., 2011; Mansour et al., 2012) and coil sampling (Moldoveanu, 2010). The randomization in sources and coil centers matters because it shapes interferences into relatively harmless noise in the appropriate domain. This "interference noise" is then removed by some nonlinear (e.g. median) filtering or by improving the sampling (e.g., by increasing the fold).

COMPRESSIVE SENSING

Unfortunately, relying on simple filtering alone in situations where high degrees of sub-sampling are desirable is inadequate because the interference may lead to unacceptable degradation. To overcome this problem, we follow a different strategy by recognizing randomized dimensionality reduction as an instance of compressive sensing (see e.g. Candès et al., 2006; Donoho, 2006). In this approach, filtering is replaced by sparsity-promoting inversion, which involves recovery of fully sampled data by an iterative algorithm (see Herrmann, 2010; Herrmann et al., 2011, for a detailed discussion of this topic in exploration seismology). This approach works well as long as the data permits a sparse representation in some transformed domain and can lead to significant cost reductions in acquisition and to efficient computations (Herrmann and Li, 2012). For noise-free observation, the recovery in both cases involves solving the convex sparsity-promoting program (Basis Pursuit)

$$\text{BP}: \qquad \underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b}. \qquad (1)$$

According to CS, this program recovers $k-$sparse or compressible vectors (that permit an accurate reconstruction from the k-largest entries) $\mathbf{x}_0 \in \mathbb{C}^N$ from incomplete measurements $\mathbf{b} \in \mathbb{C}^n$ with $n \ll N$. This recovery, which requires the inversion of the underdetermined $n \times N$ sensing matrix $\mathbf{A}$, depends on the sparsity and undersampling parameters $\varepsilon = k/N$ and $\delta = n/N$ (Donoho et al., 2011). So far, our challenges have been (i) the design of practical sampling schemes that favour recovery, e.g., the design of randomized acquisition or source encoding; (ii) implementation of fast ($\mathscr{O}(N \log N)$) sparsifying transforms that exploit structure, e.g. curvelets in seismic exploration; (iii) the development of algorithms that are frugal in the number of matrix-vector multiplies they require to solve BP.

Solutions by cooling

Even though CS has led to major breakthroughs in many research areas including dynamic MRI (Gamper et al., 2008), the application of its principles to exploration seismology has been extremely challenging because of the large scale of our problems, which easily involve vectors with billions of unknowns, and the less-than-ideal sensing matrices—i.e., matrices that lack proper normalization and that may not be close to Gaussian matrices. To overcome this problem, 'optimizers' (see e.g. van den Berg and Friedlander, 2008; Hennenfent et al., 2008) proposed a continuation method that employs properties of the Pareto tradeoff curve, which traces the ℓ_2-norm of the residual as a function of the ℓ_1-norm of the solution. This approach results in a solution during which a series of ℓ_1-norm constrained LASSO (Tibshirani, 1997) subproblems—i.e., $\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2$ s.t. $\|\mathbf{x}\|_1 \leq \tau$ with τ the prescribed ℓ_1-norm—is solved. Each subproblem involves gradient updates,

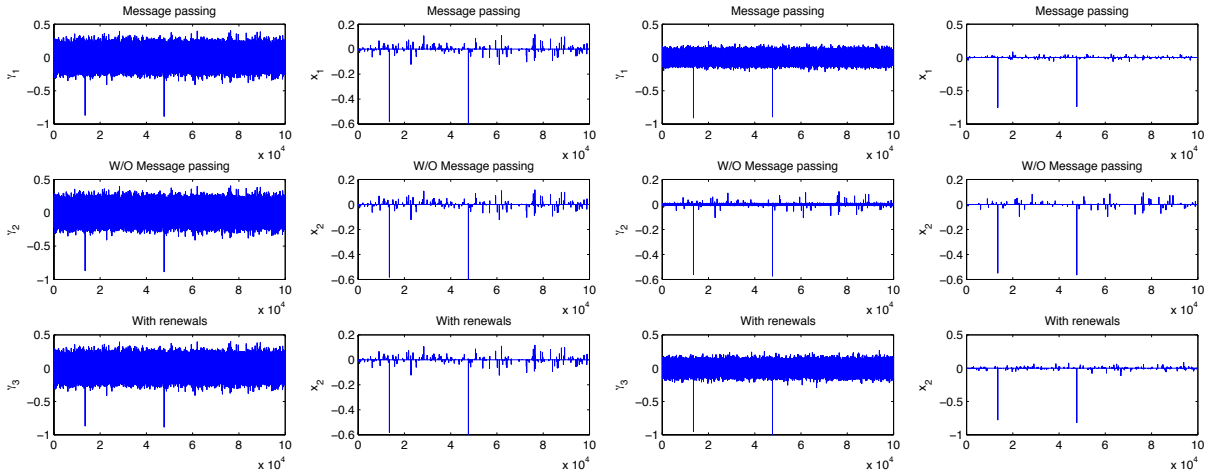

Figure 1: Model iterates after one (left) and and two (right) iterations before and after soft thresholding. Notice the spurious spiky artifacts that develop after the first iteration in the second row. These correlations are removed by the message term (top row) or by drawing a new sampling matrix and data for each iteration (bottom row).

$\mathbf{x}^{t+1} = \mathbf{x}^t + \mathbf{A}^H(\mathbf{b} - \mathbf{A}\mathbf{x}^t)$ with t the iteration number, and a nonlinear projection promoting sparsity. Not withstanding the success of these continuation methods in improving the convergence of solvers for BP, rapid convergence for truly large problems remains a major challenge.

Solutions by approximate message passing

To better understand typical behavior of solvers for BP, let us consider the first two iterations of iterative soft thresholding (Hennenfent et al., 2008) plotted in the second row of Fig. 1 before and after soft thresholding for $\mathbf{A}$ a Gaussian 10×10^5 matrix and $\mathbf{x}$ a vector with two ones at random locations. For this particular choice, we find Gaussian interference after the first iteration—i.e., $\mathbf{A}^H\mathbf{A}\mathbf{x}_0 = \mathbf{x}_0 + \mathbf{w}$ with $\mathbf{w}$ zero-mean Gaussian noise with variance $n^{-1}\|\mathbf{x}_0\|_2^2$ (Montanari, 2012). Clearly, this property favors recovery by soft thresholding because this elementwise operation ($\eta(x,\theta) = \mathrm{sgn}(x)\max(|x| - \theta, 0)$, with $\theta > 0$ the threshold value.) solves $\min_y \frac{1}{2}(y-x)^2 + \theta|x|$. Unfortunately, this property does not persist after the first iteration because dependencies build up between the model iterate $\mathbf{x}^t$ and the matrix $\mathbf{A}$. These correlations cause spurious artifacts that can not easily be separated by thresholding and this results in slow convergence because the algorithm spends many iterations to remove these wrongly identified entries. Using arguments from machine learning and coding theory, Donoho et al. (2009) address this issue by including an approximate message-passing (AMP) term in iterative soft thresholding schemes that solve BP. After initializing the model iterate and residue to $\mathbf{x}^0 = 0$, $\mathbf{r}^0 = \mathbf{b}$, the algorithm proceeds as follows:

$$\begin{aligned} \mathbf{x}^{t+1} &= \eta\left(\mathbf{A}^H\mathbf{r}^t + \mathbf{x}^t; \theta_t\right) \\ \mathbf{r}^t &= \mathbf{b} - \mathbf{A}\mathbf{x}^t + \tfrac{I_t}{n}\mathbf{r}^{t-1}. \end{aligned} \tag{2}$$

The function $\eta(\cdot;\theta_t)$ is a tuned iteration-dependent soft thresholding, and I_t is the number of entries that survived the threshold at the previous iteration. According to Montanari (2012) inclusion of this extra term is statistically equivalent to

$$\begin{aligned} \mathbf{x}^{t+1} &= \eta\left(\mathbf{A}^H(t)\mathbf{r}^t + \mathbf{x}^t; \theta_t\right) \\ \mathbf{r}^t &= \mathbf{b}(t) - \mathbf{A}(t)\mathbf{x}^t. \end{aligned} \tag{3}$$

In his argument, drawing new Gaussian sensing matrices $\mathbf{A}(t)$ and data $\mathbf{b}(t)$ for each iteration has the same effect as including the extra term in Equation 2. In both cases (juxtapose rows one and three with row two in Figure 1), correlations between the sensing matrix and model iterate are broken. This results in a shaping of the interferences into Gaussian noise for each iteration, which leads to improved convergence of the algorithm (see Herrmann, 2012, for a discussion on AMP).

Large dimensionality: a blessing in disguise

The theory behind the message term in Eq. 2 is involved and relies on asymptotic cancellation of the correlations, which provably holds for unit-norm column-normalized Gaussian matrices in the large-scale limit ($N \to \infty$). In that limit, the linear system decouples (Montanari, 2012) and recovering each entry of the model vector boils down to estimating x_i^t using the property that $\left(\mathbf{x}^t + \mathbf{A}^H\mathbf{r}^t\right)_i = (\mathbf{x} + \tilde{\mathbf{w}})_i$ for $i = 1 \cdots N$. As we mentioned before, this can be done by soft thresholding as long as $\{\tilde{w}_i\}_{i=1\cdots N}$ are asymptotically Gaussian in the large-scale limit and do not have too large a variance depending on (ε, δ).

Even though we have control over the subsampling ratio (δ), controlling the sparsity (ε), column normalization, and Gaussian-like behavior are much more difficult to manage. At this point, we take a leap of faith and we assume that we can come up with randomized sampling schemes that lead to matrices that behave approximately Gaussian in the large-scale limit.

COOLING AND MESSAGE PASSING COMBINED

To avoid reliance on the delicate interplay between the message-passing term and the tuned thresholding scheme that hinges on sensing matrices given by normalized Gaussian matrices we propose, motivated by the statistical equivalence of Equations 2

Result: Estimate for the model $\mathbf{x}^{t+1}$

1 $\mathbf{x}^0, \widetilde{\mathbf{x}} \longleftarrow \mathbf{0}$ and $t, \tau^0 \longleftarrow 0$; // Initialize
2 **while** $t \leq T$ **do**
3 $\quad \mathbf{A} \longleftarrow \mathbf{A} \sim P(\mathbf{A})$; // Draw new sensing matrix
4 $\quad \mathbf{b} \longleftarrow \mathbf{A}\mathbf{x}$; // Collect new data
5 $\quad \mathbf{x}^{t+1} \longleftarrow \text{spgl1}(\mathbf{A}, \mathbf{b}, \tau^t, \sigma = 0, \mathbf{x}^t)$; // Reach Pareto
6 $\quad \tau^t \longleftarrow \|\mathbf{x}^{t+1}\|_1$; // New initial τ value
7 $\quad t \longleftarrow t + \Delta T$; ; // Add # of iterations of spgl1
8 **end**

Algorithm 1: Modified SPGℓ_1 with message passing.

and 3, to draw new sensing matrices $\mathbf{A}$ and corresponding data vectors $\mathbf{b}$ after each LASSO-subproblem is solved. This approach has the advantage that it does not require normalization of the sensing matrix. The proposed method is summarized in Algorithm 1 and involves a relative minor but non-trivial modification of SPGℓ_1 (van den Berg and Friedlander, 2008). Following earlier work by Herrmann and Li (2012), we exit SPGℓ_1 as soon as the Pareto curve is reached, which corresponds to the solution of a LASSO sub-problem. At this point, we select a new statistically independent experiment (lines 3–4) by drawing a new copy of the sensing matrix ($\mathbf{A} \sim P(\mathbf{A})$ with $P(\cdot)$ denoting the probability distribution for the random matrix $\mathbf{A}$), followed by a warm start of SPGℓ_1 with the previous model iterate and ℓ_1-norm of the previous iteration as starting value for the τ parameter. Because SPGℓ_1 itself constitutes the solution of a series of LASSO subproblems, followed by warm starts, removing lines (3–4) does not alter the SPGℓ_1.

MESSAGE PASSING IN EXPLORATION SEISMOLOGY

At this point, the reader may be inclined to dismiss the proposed algorithm because it requires multiple independent randomized experiments. On first thought this requirement seems to defeat the purpose of CS, which aims to reduce and not increase the number of measurements. While this argument certainly holds when acquiring data physically in the field, the situation is quite different when we are dealing with imaging and full-waveform inversion (FWI) problems that extract information from seismic-data volumes that are already collected. To reduce the size of these "abundant" data volumes and the pertinent excessive demands on computational resources, (random) source encodings (Romero et al., 2000) have been employed (see also Krebs et al., 2009; Haber et al., 2012; van Leeuwen et al., 2011) with some success. In these approaches, the induced error is controlled either by averaging over past model iterates or by growing the number of encoded sources (van Leeuwen and Herrmann, 2011; Aravkin et al., 2011). Our approach aims to do the same for randomized linear inversion problems. However, the way in which we accomplish the error control differs fundamentally because we for each iteration we rely on transform-domain denoising via ℓ_1-norm minimization instead. The performance of the proposed algorithm is demonstrated with two examples:

Accelerated sparse recovery

To set the stage, let us first examine a recovery example for $N = 248759$ with a Gaussian sensing matrix and a realistic compressible complex-valued curvelet vector with coefficients

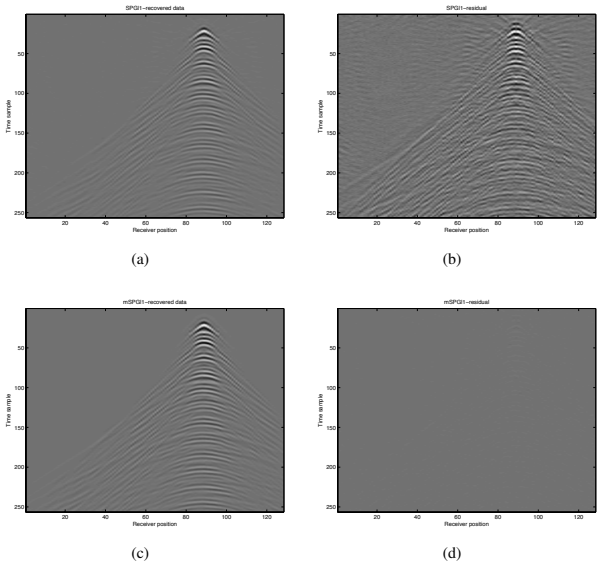

Figure 2: Recovery of a compressible curvelet vector using SPGℓ_1 (a) and modified SPGℓ_1 (c). Notice the striking difference in quality after transforming the estimated curvelet vector back into the shot domain. The improvements are especially visible for the resisdues in (b) and (d). The grey scale for the residues is devided by 10.

obtained by solving a sparsity promoting program on a shot record. We measure the resulting not so sparse vector $\mathbf{x}$ with $\varepsilon = 0.31$ using a Gaussian matrix with $\delta = 0.13$ and i.i.d. entries sampled from $N(0, 1/n)$. Results after $T = 500$ iterations with and without messaging are included in Figure 2 and show a remarkable improvement (SNRs of 15 dB versus 44 dB).

Accelerated imaging

While the previous example certainly illustrates the potential uplift of modifying SPGℓ_1, it does not really reflect a situation relevant to exploration seismology. To study improvements by message passing, we carefully revisit an imaging experiment by Herrmann and Li (2012). To avoid issues with the linearization, we generate data using the linearized Born scattering operator for an accurate velocity model and for 350 sequential sources and 701 receivers sampled at 20 m and 10 m, respectively. To simulate data from a 409×1401 gridded velocity perturbation sampled at 5 m, ten random frequencies are selected from $20 - 50$ Hz. To reduce this data volume, we reduce the 350 sequential sources to only three (opposed to 17 in Herrmann and Li, 2012) simultaneous sources by hitting the source direction with a 3×350 Gaussian matrix. This reduction in the number of sources decreases the cost of linearized modeling and migration hundred fold. By inverting the dimensionality-reduced linearized Born scattering operator compounded with the curvelet synthesis matrix, we remove the source crosstalk running our algorithm for $T \approx 500$ iterations with and without messaging. Again, we observe in Figure 3 significant improvements from the messaging for a less-than-ideal sensing matrix with $N = 4213324$, a small subsampling parameter $\delta \approx 0.006$, and a large sparsity parameter $\varepsilon = 0.32$. This is remarkable because the number of PDE solves is the same in both cases.

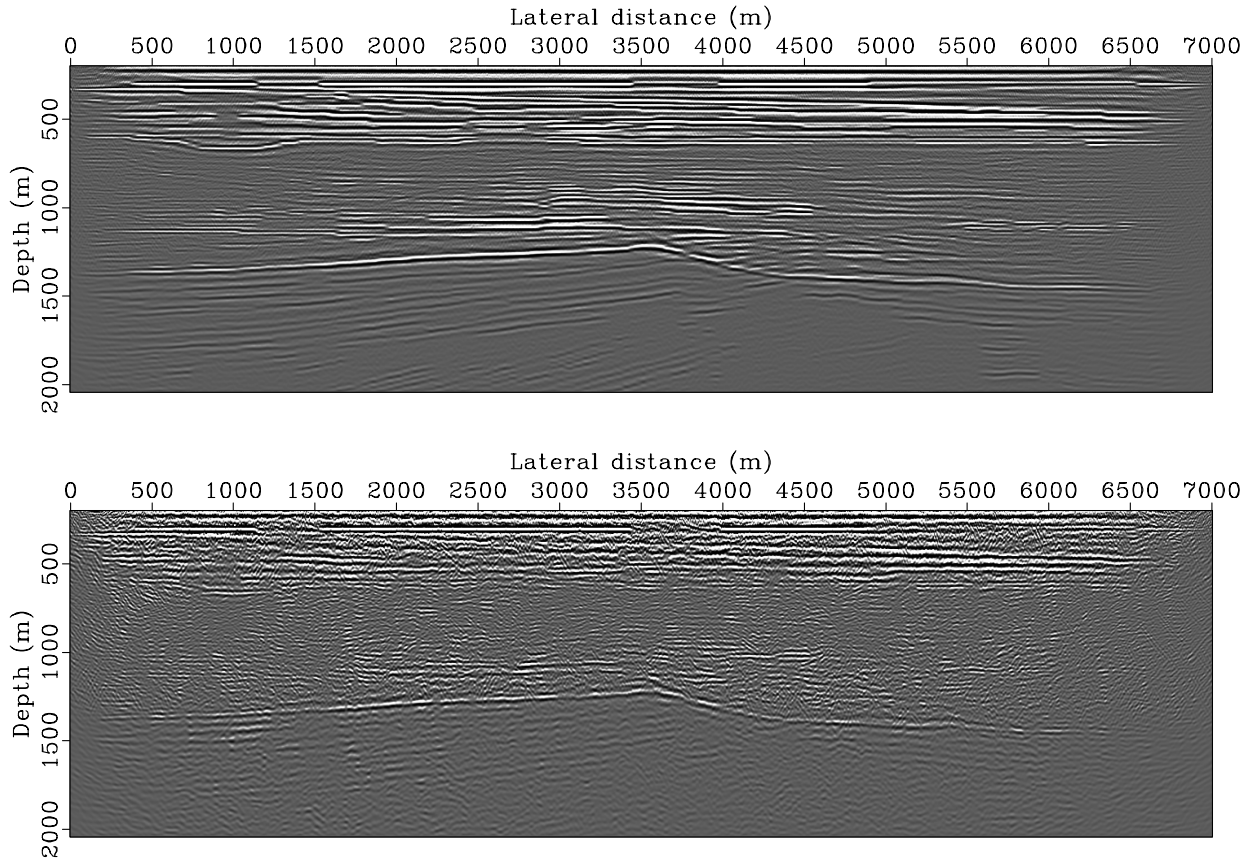

Figure 3: Imaging with (top) and without (bottom) 'messaging'.

DISCUSSION AND CONCLUSIONS

At first glance, the proposed method may seem to be some sort of incarnation of stochastic-gradients where the cost of gradient evaluations of separable optimization problems are reduced by randomized sampling. Whilst this approach is shared by the method presented here, the ideas from message passing allow us to truly exploit the underlying structure of the seismic imaging problem by turning each sub-problem of the solver into a "denoising" problem using linearity w.r.t. the sources and abundance of source experiments to draw from. The effects of drawing new randomized simultaneous shots are dramatic and similar to improvements attained by true approximate message passing reported elsewhere. In either case, high quality results are attained while working on small sub-problems with a total computational cost of only $3-5$ iterations of least-squares involving all data for each iteration. However, this low number of iterations is typically not enough to get satisfactory results warranting our alternative approach.

Even though each dimensionality-reduced sub-problem (e.g. simultaneous-source experiment) contains the same amount of information, the proposed algorithm allows us to accelerate and even improve inversion results without increasing the problem size. Not only does this feature remove IO bottlenecks but it also reduces costs of matrix-vector multiplications (e.g., the number of PDE solves that have to be computed in parallel). In addition, drawing new experiments helps the solver to recover challenging small curvelet coefficients close to the nullspace. This is helpful in difficult cases where the sub-sampling parameter is small and where the model is not particularly sparse. In our imaging example, this explains how we were able to benefit from message passing by exploiting near invariance of curvelets under the wave-equation Hessian. This invariance leads to sub-problems that are very much like denoising problems.

Whilst rerandomizations are not strictly necessary, AMP itself hinges on very strict conditions on the sensing matrix that are difficult to meet in seismic reality. The condition of the large-scale limit ($N \rightarrow \infty$), however, is not difficult to meet as long as the entries of the sensing matrix are zero mean and have finite moments. This opens interesting perspectives on the application of message passing to exploration seismology. For more applications of message passing, we refer to other contributions by the authors to these proceedings.

ACKNOWLEDGMENTS

Thanks to Charles Jones (BG) for the Compass velocity model and Xiang Li for help with the imaging example. This work was supported by NSERC (CRD 375142) and by the sponsors of the Seismic Laboratory for Imaging and Modelling (SLIM).

REFERENCES

Aravkin, A. Y., M. P. Friedlander, F. J. Herrmann, and T. van Leeuwen, 2011, Robust inversion, dimensionality reduction, and randomized sampling. (To appear in Mathematical Programming).

Candès, E., J. Romberg, and T. Tao, 2006, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information: IEEE Trans. Inform. Theory, **52**, 489–509.

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

Donoho, D., I. Johnstone, and A. Montanari, 2011, Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising: Arxiv preprint arXiv:1111.1041.

Donoho, D., A. Maleki, and A. Montanari, 2009, Message passing algorithms for compressed sensing: Proceedings of the National Academy of Sciences, **106**, 18914–18919.

Gamper, U., P. Boesiger, and S. Kozerke, 2008, Compressed sensing in dynamic MRI: Magnetic Resonance in Medicine, **59**, 365–373.

Haber, E., M. Chung, and F. J. Herrmann, 2012, An effective method for parameter estimation with PDE constraints with multiple right hand sides. (To appear in SIAM Journal of Optimization).

Hampson, G., J. Stefani, and F. Herkenhoff, 2008, Acquisition using simultaneous sources: The Leading Edge, **27**, 918–923.

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**, A23–26.

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

———, 2012, Pass on the message: recent insights in large-scale sparse recovery: Presented at the EAGE technical program.

Herrmann, F. J., M. P. Friedlander, and O. Yilmaz, 2011, Fighting the curse of dimensionality: compressive sensing in exploration seismology. to appear in IEEE Signal Processing Magazine.

Herrmann, F. J., and X. Li, 2012, Efficient least-squares imaging with sparsity promotion and compressive sensing. (To appear in Geophysical Prospecting).

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.

Mansour, H., H. Wason, T. T. Lin, and F. J. Herrmann, 2012, Simultaneous-source marine acquisition with compressive sampling matrices. (To appear in Geophysical Prospecting).

Moldoveanu, N., 2010, Random sampling: A new strategy for marine acquisition: SEG Technical Program Expanded Abstracts, **29**, 51–55.

Montanari, A., 2012, Graphical models concepts in compressed sensing: Cambridge University press. (Compressed Sensing: Theory and Applications).

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.

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

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

van Leeuwen, T., A. Y. Aravkin, and F. J. Herrmann, 2011, Seismic waveform inversion by stochastic optimization: International Journal of Geophysics, article ID: 689041.

van Leeuwen, T., and F. J. Herrmann, 2011, Fast waveform inversion without source encoding. Submitted for publication.