One-norm regularized inversion: learning from the Pareto curve

Gilles Hennenfent and Felix J. Herrmann, Earth & Ocean Sciences Dept., the University of British Columbia

ABSTRACT

Geophysical inverse problems typically involve a trade off between data misfit and some prior. Pareto curves trace the optimal trade off between these two competing aims. These curves are commonly used in problems with two-norm priors where they are plotted on a log-log scale and are known as L-curves. For other priors, such as the sparsity-promoting one norm, Pareto curves remain relatively unexplored. First, we show how these curves provide an objective criterion to gauge how robust one-norm solvers are when they are limited by a maximum number of matrix-vector products that they can perform. Second, we use Pareto curves and their properties to define and compute one-norm compressibilities. We argue this notion is key to understand one-norm regularized inversion. Third, we illustrate the correlation between the one-norm compressibility and the performance of Fourier and curvelet reconstructions with sparsity promoting inversion.

INTRODUCTION

Many inverse problems in geophysics are ill posed (Parker, 1994)—their solutions are not unique or are acutely sensitive to changes in the data. To solve this kind of problem stably, additional information must be introduced. This technique is called *regularization* (see, e.g., Phillips, 1962; Tikhonov, 1963).

Specifically, when the solution of an ill-posed problem is known to be (almost) sparse, Oldenburg et al. (1983) and others have observed that a good approximation to the solution can be obtained by using one-norm regularization to promote sparsity. More recently, results in information theory have breathed new life into the idea of promoting sparsity to regularize ill-posed inverse problems. These results establish that, under certain conditions, the sparsest solution of a (severely) underdetermined linear system can be *exactly* recovered by seeking the minimum one-norm solution (Candès et al., 2006; Donoho, 2006; Rauhut, 2007). This has led to tremendous activity in the newly established field of *compressed sensing* and a resurgence of one-norm regularized inversion.

In this publication, we demonstrate how the Pareto curve can be used to make quantitative assessments regarding the performance of one-norm regularized, transform-

In original form 2008 April 9.

based recovery from incomplete data. First, we outline the general problem formulation and associated convex optimization problems. Second, we give a brief overview of the Pareto curve and its properties. Third, we discuss two possible applications of this curve to gain further insights in one-norm regularized inverse problems and make quantitative assessments regarding their success.

PROBLEM STATEMENT

Consider the following underdetermined system of linear equations

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}, \tag{1}$$

where the n-vectors $\mathbf{y}$ and $\mathbf{n}$ represent observations and additive noise, respectively. The n-by-N matrix $\mathbf{A}$ is the modeling operator that links the model $\mathbf{x}_0$ to the noise-free data given by $\mathbf{y} - \mathbf{n}$. We assume that $N \gg n$ and that $\mathbf{x}_0$ has few nonzero or significant entries. We use the terms "model" and "observations" in a broad sense, so that many linear geophysical problems can be cast in the form shown in Equation 1.

Because $\mathbf{x}_0$ is assumed to be (almost) sparse, one can promote sparsity as a prior via one-norm regularization to overcome the singular nature of $\mathbf{A}$ when estimating $\mathbf{x}_0$ from $\mathbf{y}$. The challenge is to balance at best the data misfit, defined as $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$, with the regularization term, $\|\mathbf{x}\|_1$. One of the following three convex optimization approaches

$$\begin{aligned}
\mathrm{QP}_\lambda: &\qquad \min_{\mathbf{x}} \ \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1,\\
\mathrm{BP}_\sigma: &\qquad \min_{\mathbf{x}} \ \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \le \sigma,\\
\mathrm{LS}_\tau: &\qquad \min_{\mathbf{x}} \ \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \le \tau,
\end{aligned}$$

can be used to this effect.

QP_λ closely relates to quadratic programming (QP) and is probably the most commonly used in geophysics. However, it is generally not clear how to choose the Lagrange multiplier $\lambda \ge 0$ such that the solution of QP_λ is, in some sense, optimal. The basis pursuit (BP) denoise problem (Chen et al., 1998) is often preferred when an estimate of the noise $\sigma \ge 0$ in the data is available. The LASSO problem (Tibshirani, 1996) is of lesser interest because an estimate of the one norm of the solution $\tau \ge 0$ is typically not available for geophysical problems.

To gain insights into one-norm regularized inversion, we propose to look at the Pareto curve. In the context of the two-norm—i.e., Tikhonov—regularization, the Pareto curve is commonly used where it is plotted on a log-log scale and is known as L-curve (Lawson and Hanson, 1974).

PARETO CURVE

The Pareto curve traces, for a specific pair of **A** and **y**, the optimal tradeoff in the space spanned by the data misfit and the one-norm regularization term. Figure 1 gives a schematic illustration. Point ① clarifies the connection between the three parameters of QP_λ, BP_σ, and LS_τ (see van den Berg and Friedlander, 2008; Hennenfent et al., 2008, for more details). The coordinates of a point on the Pareto curve are (τ, σ) and the slope of the tangent at this point is $-\lambda$. The end points of the curve—points ② and ③—are two special cases. When $\tau = 0$, the solution of LS_τ is $\mathbf{x} = 0$ (point ②). It coincides with the solution of BP_σ with $\sigma = \|\mathbf{y}\|_2$. When $\sigma = 0$, the solution of BP_σ (point ③) coincides with the solutions of LS_τ, where $\tau = \tau_{BP_0}$, and QP_λ, where $\lambda = 0^+$—i.e., λ infinitely close to zero from above. These relations are formalized as follows in van den Berg and Friedlander (2008):

Result 1. *The Pareto curve i) is convex and decreasing, ii) is continuously differentiable, and iii) has a negative slope* λ*.*

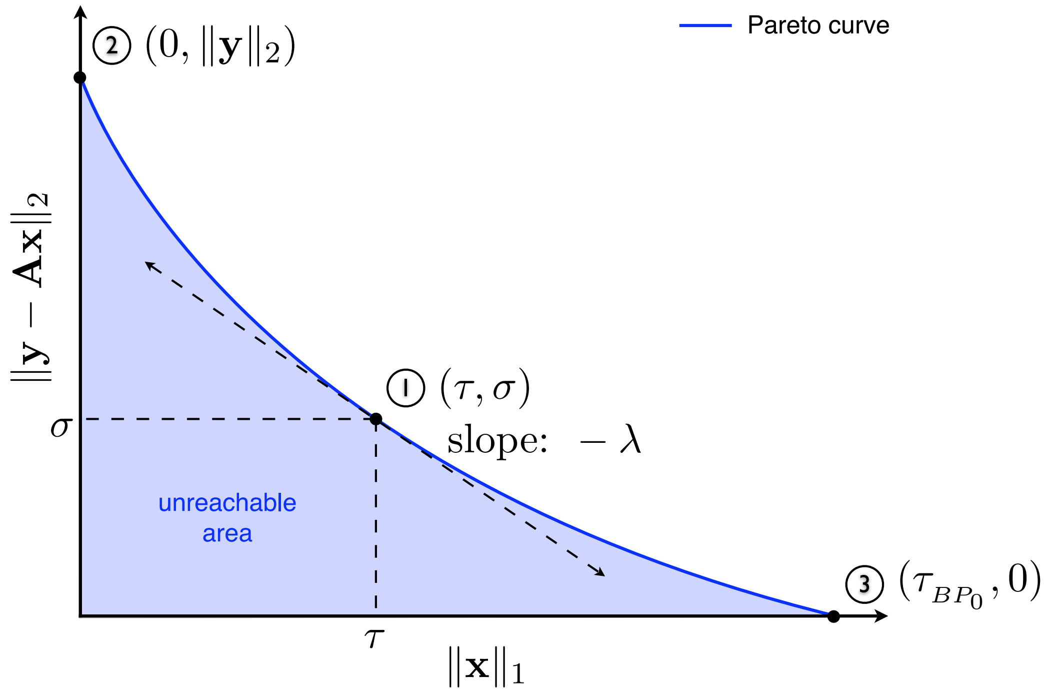


Figure 1: Schematic illustration of a Pareto curve. Point ① exposes the connection between the three parameters of QP_λ, BP_σ, and LS_τ. Point ② corresponds to the trivial solution, i.e., $\mathbf{x} = 0$, and point ③ to a solution of BP_σ with $\sigma = 0$.

For large-scale geophysical applications, it is not practical (or even feasible) to sample the entire Pareto curve. However, its regularity, as implied by Result 1, means that it is possible to obtain a good approximation to the curve with very few interpolating points, as illustrated by Hennenfent et al. (2008) on a wavefield reconstruction problem using curvelets (Herrmann and Hennenfent, 2008).

APPLICATIONS

Comparison of one-norm solvers

To understand the robustness of different one-norm solvers that are limited by a maximum number of matrix-vector products that they can perform, Hennenfent et al. (2008) propose to track on a graph the data misfit versus the one norm of successive iterates for each solver. The Pareto curve serves as a reference and gives a rigorous yardstick for measuring the quality of the solution path generated by each algorithm.

Figure 2 shows the Pareto curve and the BP_0 solution paths of four one-norm solvers: iterative soft thresholding (IST) introduced by Daubechies et al. (2004), IST extension to include cooling (ISTc - Figueiredo and Nowak, 2003; Herrmann and Hennenfent, 2008), the spectral projected-gradient ($\mathrm{SPG}\ell_1$) algorithm introduced by van den Berg and Friedlander (2008), and iterative re-weighted least-squares (IRLS - Gersztenkorn et al., 1986), which uses a quadratic approximation to the one-norm regularization function. The maximum number of iterations is small compared to the size of the problem and fixed. This roughly equates to using the same number of matrix-vector products for each solver. The starting vector provided to each solver is the zero vector, and hence the paths start at $(0, \|\mathbf{y}\|_2)$—point ② in Figure 1. The problem considered is a benchmark problem that is typically used in the compressed sensing literature (Donoho et al., 2006). The matrix $\mathbf{A}$ is taken to have Gaussian independent and identically-distributed entries; a sparse solution $\mathbf{x}_0$ is randomly generated, and the "observations" $\mathbf{y}$ are computed according to Equation 1.

Whereas $\mathrm{SPG}\ell_1$ provides a fairly accurate approximation to the BP_0 solution, those computed by IST, ISTc, and IRLS suffer from larger errors. IST solves QP_{0^+}. Because there is hardly any regularization, IST first works towards minimizing the data misfit. When the data misfit is sufficiently small, the effect of the one-norm penalization starts, yielding a change of direction towards the BP_0 solution. The limited number of matrix-vector products terminates the procedure early. The data misfit at the candidate solution is small but the one norm is completely incorrect. ISTc solves QP_λ for a decreasing sequence $\lambda_i \to 0$. The starting vector for QP_{λ_i} is the solution of $\mathrm{QP}_{\lambda_{i-1}}$, which is by definition on the Pareto curve when each subproblem is accurately solved. This explains why ISTc so closely follows the curve, at least at the beginning of the path. Towards the end, ISTc accumulates small errors because there are not enough iterations to solve each subproblem to sufficient accuracy. IRLS suffers from similar difficulties.

Compression of seismic energy

Another insightful application of the Pareto curve is when the matrix $\mathbf{A}$ is defined as the synthesis operator of a (redundant) transform, e.g., Fourier or curvelet (Candès et al., 2005, and references therein) transform. In this case, the Pareto curve measures

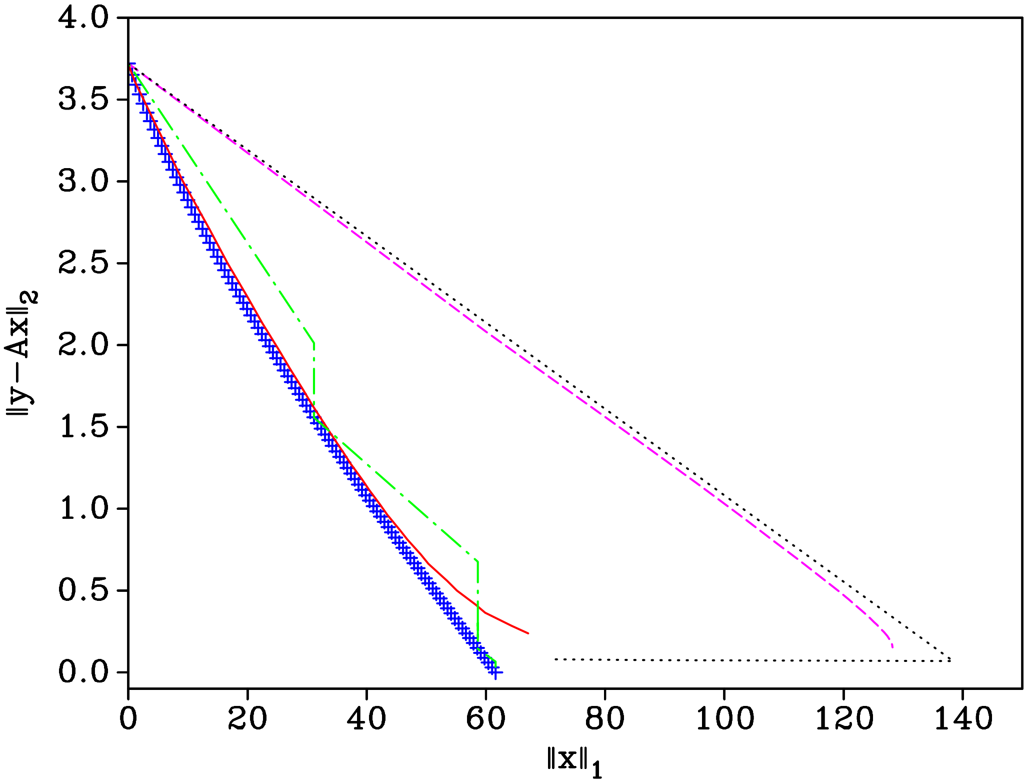


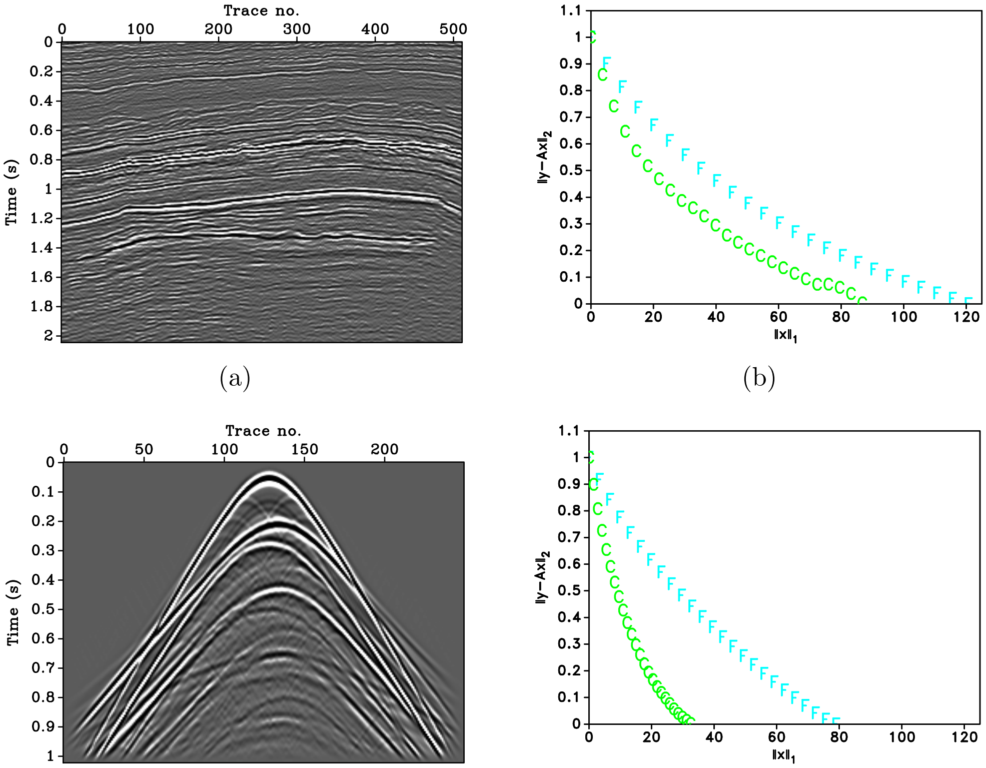
Figure 2: Pareto curve and optimization paths (same, limited number of iterations) of four solvers for a BP_0 problem. The symbols + represent a sampling of the Pareto curve. The solid (—) line is the solution path of ISTc, the chain ($- \cdot -$) line the path of SPGLℓ_1, the dashed ($- -$) line the path of IST, and the dotted ($\cdots$) line the path of IRLS.

the compressibility, in the one-norm sense, of the data $\mathbf{y}$ in a given transform domain. This can be used to compare different transforms.

In Figure 3, we compare the Fourier transform with the curvelet transform. The synthesis operators have unit-norm columns. In the first experiment, $\mathbf{y}$ is defined as the real migrated image depicted in Figure 3(a). Figure 3(b) shows the corresponding Pareto curves approximated with seven interpolating points. In the second experiment, $\mathbf{y}$ is defined as the synthetic shot gather depicted in Figure 3(c) and Figure 3(d) displays the corresponding Pareto curves also approximated with seven interpolating points.

In both experiments, the curvelet transform compresses better the seismic energy. The improvement on the Fourier transform is particularly sizable in the case of the shot gather. Although these observations do not, by themselves, explain the success of sparsity-promoting formulations using the curvelet transform, they provide a better justification than the arguments proposed by Candès et al. (2005) and Hennenfent and Herrmann (2006). These arguments, rooted in the empirical decay rate for the magnitude-sorted transform coefficients, link a higher rate to a better performance of one-norm recovery. Whereas this reasoning is viable for non-redundant transforms, it looses some of its strength for redundant signal expansions, where more coefficients are needed to approximate the signal. This confronts us with a dilemma because results using the overcomplete curvelet transform generally show an improvement over Fourier-based techniques, an observation that would be consistent with increased sparsity. In this case, an argumentation can be made either by considering decay rates as a function of the percentage of the total coefficients (Hennenfent and Herrmann, 2006) or by working with a reduced transform-domain coefficient vector to compensate for the redundancy of the transform.(Candès et al., 2005)

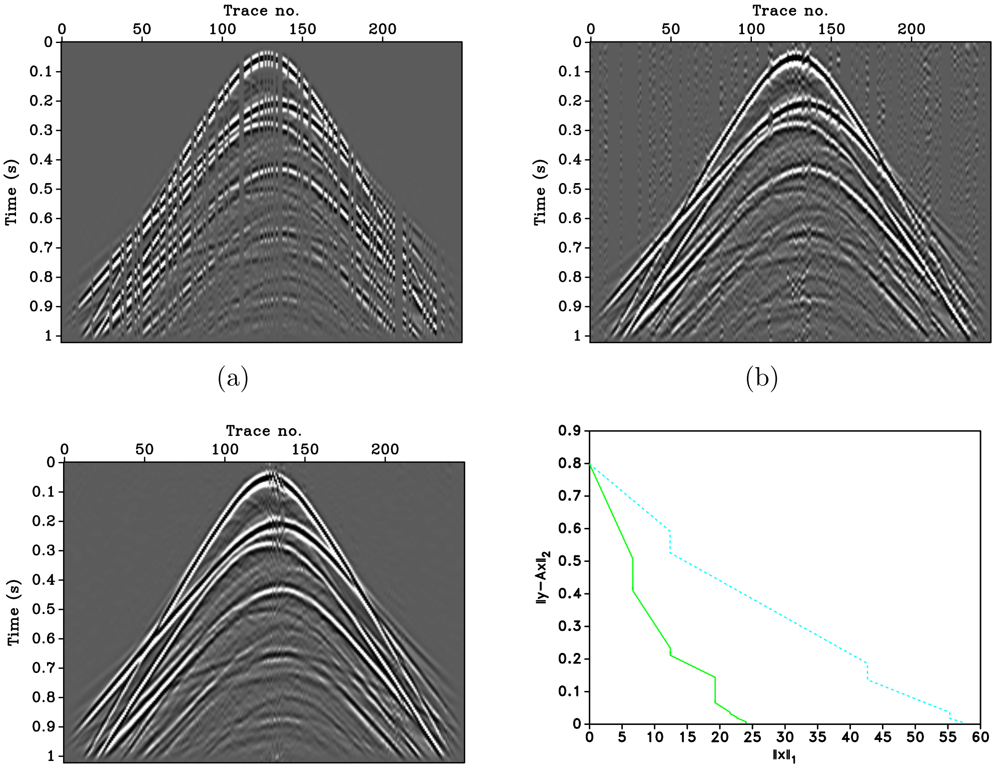
Unfortunately, these explanation artifices are rather unsatisfactory. Therefore we empirically study whether there exists a link between the one-norm compressibility and the performance of sparsity-promoting algorithms. The test problem is the reconstruction of an incomplete, noise-free wavefield. The methods considered are the Fourier reconstruction with sparsity-promoting inversion (FRSI - Zwartjes, 2005) and its curvelet counterpart (CRSI - Herrmann and Hennenfent, 2008). Both methods are implemented using SPGℓ_1. The simulated input data (Figure 3(c)) is the shot gather in Figure 3(c) with randomly-missing traces. Figures 4(b) and 4(c) show the FRSI and CRSI results, respectively. The corresponding signal-to-noise ratios (SNR) are 9.25 dB and 13.55 dB. Figure 4(d) displays the SPGℓ_1 paths to the BP$_0$ solutions. Note how these paths behave similarly to the Pareto curves in Figure 3(d). Furthermore, the offset between the BP$_0$ solutions remains remarkably constant. The restriction operator used in FRSI and CRSI did not fundamentally change the structure of the problems.



(a) (b)

(c) (d)

Figure 3: One-norm compressibility in the Fourier and curvelet domains. (a) Real migrated image and (b) corresponding Pareto curves for both transforms. (c) Synthetic shot gather and (d) corresponding Pareto curves for both transforms. The symbols "F" trace Fourier curves, and "C" curvelet curves.



(a)

(b)

(c)

(d)

Figure 4: Noise-free wavefield reconstruction using the Fourier or curvelet transform. (a) Simulated acquired data, (b) FRSI result (SNR = 9.25 dB), (c) CRSI result (SNR = 13.55 dB), and (d) SPGℓ_1 paths to the BP$_0$ solutions of FRSI (dash line) and CRSI (solid line).

CONCLUSIONS

The current resurgence of successful one-norm regularized inverse problems motivates the need of a better understanding of the inner workings of these problems. The sheer size of geophysical applications makes this task difficult though. We show that the Pareto curve, thanks to its properties, is a practical and insightful tool. This curve serves, for example, as a reference in order to make an informed decision on how to best truncate the solution process of one-norm solvers and reduce the amount of computation. The Pareto curve and its properties are also useful in defining and computing the one-norm compressibility of a signal in a transform domain, as we illustrate using Fourier and curvelets. Furthermore, we observe that our extended notion of compressibility can be used to make quantitative assessments regarding the performance of one-norm regularized, transform-based recovery from incomplete data. This prospect is particularly exciting and can possibly be leveraged to other one-norm regularized inversions.

ACKNOWLEDGMENTS

The authors thank E. van den Berg, Ö. Yilmaz, and M. P. Friedlander for fruitful discussions. This publication was prepared using Madagascar (`rsf.sf.net`), a package for reproducible computational experiments, SPGℓ_1 (`cs.ubc.ca/labs/scl/spgl1`), a one-norm solver, and Sparco (`cs.ubc.ca/labs/scl/sparco`), a suite of linear operators and problems for testing algorithms for sparse signal reconstruction. This research was in part financially supported by NSERC Discovery Grant 22R81254 of F.J.H. and by CRD Grant DNOISE 334810-05 of F.J.H., and was carried out as part of the SINBAD project with support, secured through ITF, from the following organizations: BG Group, BP, Chevron, ExxonMobil, and Shell.

REFERENCES

van den Berg, E. and M. P. Friedlander, 2008, Probing the Pareto frontier for basis pursuit solutions: Technical Report TR-2008-01, UBC Computer Science Department. (`http://www.optimization-online.org/DB_HTML/2008/01/1889.html`).

Candès, E. J., L. Demanet, D. L. Donoho, and L. Ying, 2005, Fast discrete curvelet transforms: Multiscale Modeling and Simulation, **5**, 861–899.

Candès, E. J., J. Romberg, and T. Tao, 2006, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information: IEEE Transactions on Information Theory, **52**, no. 2, 489–509.

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

Daubechies, I., M. Defrise, and C. De Mol, 2004, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint: Communications on Pure and Applied Mathematics, **LVII**, 1413–1457.

Donoho, D. L., 2006, Compressed sensing: IEEE Transactions on Information Theory, **52**, no. 4, 1289–1306.

Donoho, D. L., Y. Tsaig, I. Drori, and J.-L. Starck, 2006, Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit: Technical Report TR-2006-2, Stanford Statistics Department. (`http://stat.stanford.edu/~idrori/StOMP.pdf`).

Figueiredo, M. and R. Nowak, 2003, An EM algorithm for wavelet-based image restoration: IEEE Transactions on Image Processing, **12**, no. 8, 906–916.

Gersztenkorn, A., J. B. Bednar, and L. Lines, 1986, Robust iterative inversion for the one-dimensional acoustic wave equation: Geophysics, **51**, no. 2, 357–369.

Hennenfent, G. and F. J. Herrmann, 2006, Seismic denoising with non-uniformly sampled curvelets: Computing in Science and Engineering, **8**, no. 3, 16–25.

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.

Herrmann, F. J. and G. Hennenfent, 2008, Non-parametric seismic data recovery with curvelet frames: Geophysical Journal International, **173**, 233–248.

Lawson, C. L. and R. J. Hanson, 1974, Solving least squares problems: Prentice Hall.

Oldenburg, D., T. Scheuer, and S. Levy, 1983, Recovery of the acoustic impedance from reflection seismograms: Geophysics, **48**, no. 10, 1318–1337.

Parker, R. L., 1994, Geophysical inverse theory: Princeton University Press.

Phillips, D. L., 1962, A technique for the numerical solution of certain integral equations of the first kind: Journal of the Association for Computing Machinery, **9**, no. 1, 84–97.

Rauhut, H., 2007, Random sampling of sparse trigonometric polynomials: Applied and Computational Harmonic Analysis, **22**, no. 1, 16–42.

Tikhonov, A. N., 1963, Solution of incorrectly formulated problems and regularization method: Soviet mathematics - Doklady, **4**, 1035–1038.

Tibshirani, R., 1996, Regression shrinkage and selection via the LASSO: Journal Royal Statististics, **58**, no. 1, 267–288.

Zwartjes, P. M., 2005, Fourier reconstruction with sparse inversion: PhD thesis, Delft University of Technology.

In original form 2008 April 9.