Curvelet-based non-linear adaptive subtraction with sparseness constraints

Felix Herrmann and Peyman Moghaddam, EOS, University of British Columbia

Abstract

In this paper an overview is given on the application of directional basis functions, known under the name Curvelets/Contourlets, to various aspects of seismic processing and imaging, which involve adaptive subtraction. Key concepts in the approach are the use of (i) directional basis functions that localize in both domains (e.g. space and angle); (ii) non-linear estimation, which corresponds to localized muting on the coefficients, possibly supplemented by constrained optimization. We will discuss applications that include multiple, ground-roll removal and migration denoising.

Introduction

Everybody would agree that a seismologists dream would be the existence of a basis-function decomposition which localizes in space and angle; is sparse for seismic data and images; is well behaved under certain (imaging) operators and is independent on the local phase characteristics. To make a long story short, this dream has come true, at least in part for 2-D data, with the recent introduction of directional non-separable wavelets, known under the names Curvelets or Contourlets [4, 2]. As their names suggest, these basis functions are designed to optimally represent images with edges on piece-wise smooth curves ($f \in C^2$ with a finite number of jumps to be precise). For our application, this optimality corresponds to a non-adaptive (the transform does not depend on the function to be transformed, like the FFT) decomposition that is near optimal (read as sparse as you can get, i.e. the non-linear approximation rate is near optimal) for seismic reflection data as well as for seismic images, both of which tend to consist of events that are relatively smooth in the tangential direction and oscillatory/singular across.

For some time, orthogonal separable discrete wavelet transforms looked like they would deliver onto above's dream, since they represented a transform that is optimal for non-stationary piece-wise smooth 1-D functions. Indeed, for this class of functions, wavelets form an unconditional basis and consequently lead to an almost diagonalization of the data's covariance. Even though, wavelets have been very successful in many different fields, their application to functions and operators that involve directivity, e.g. seismic data with wavefronts and non-stationary directional filters, such as the normal operators for seismic imaging, has been less successful. For instance, discrete wavelets transforms have not really made an impact in seismic processing and processing because their lack of directional selectivity renders them less effective. Curvelets (from now on I will use the name Curvelets to refer to both Curvelets and Contourlets), on the other hand, remain more or less invariant under imaging/modeling operators, while they maintain near optimal sparse representations for functions with singularities on curves. These properties make Curvelets the ideal representations in which to formulate seismic processing and imaging problems [8, 11, see also other contributions by the authors to the Proceedings of this Conference]/

The success of identifying and effectively removing coherent (noise) components from data depends on the interplay of two factors: (i) the ability to accurately model the noise & signal components and (ii) the ability to separate/filter these components in a possibly noisy environment. Adaptive-subtraction methods aim to robustly remove the noise component, given imperfections in the noise prediction and the existence of an additional incoherent noise term. Without being all inclusive, applications of adaptive subtraction range from multiple, ground-roll elimination [14] to the robust computation of 4D-difference cubes and the removal of noise after migration [8, and other contributions by the authors to the Proceedings of this Conference].

We cast the adaptive subtraction method into the general framework of inversion theory. We argue that one is much better shape to solve an inversion problem if one is able to sparsely represent data, noise and denoised data (the model). Given the appropriate choice of basis functions, allows us to remove the bulk of the noise using a diagonal minimax estimation operator, based on thresholding [9, 5]. Sparseness/minimal structure is imposed by solving an additional global optimization problem on the estimated coefficients that minimizes an additional penalty functions [2, 13].

First, we will formulate the adaptive subtraction problem as a least-squares inverse problem. Then, we will recast this problem onto optimal basis functions and present the non-linear estimation by thresholding rule that is based on the *magnitudes* of the coefficients only. Basic properties of Curvelets are presented next, followed by a description of the global optimization techniques we employ. We conclude by presenting a number of examples.

The inverse problem underlying adaptive subtraction

Virtually any linear problems in seismic processing and imaging can be seen as a special case of the following generic problem: how to obtain $\mathbf{m}$ from data $\mathbf{d}$ in the presence of (coherent) noise $\mathbf{n}$:

$$\mathbf{d} = \mathbf{K}\mathbf{m} + \mathbf{n}, \tag{1}$$

in which, for the adaptive subtraction problem, $\mathbf{K}$ and $\mathbf{K}^*$, represent the respective time convolution and correlation with an effective wavelet $\mathbf{\Phi}$ that minimizes the following functional [see e.g. 6]

$$\min_{\mathbf{\Phi}} = \|\mathbf{d} - \mathbf{\Phi} \overset{t}{*} \mathbf{m}\|_p, \tag{2}$$

where $\mathbf{d}$ represents data with coherent noise; $\mathbf{m}$ the predicted noise and $\hat{\mathbf{n}} = \min_{\mathbf{\Phi}} \|\mathbf{d} - \mathbf{\Phi} * \mathbf{m}\|_p$ the noise-free data, obtained by minimizing the L^p-norm. For $p = 2$, Eq. 2 corresponds to a standard least-squares problem where the energy of the 'noise' is minimized.

Seismic data and images are notoriously non-stationary and non-Gaussian, which may give rise to a loss of coherent events in the denoised component when employing the L^2-norm. The non-stationarity and non-Gaussianity have with some success been addressed by solving Eq. 2 within sliding windows and for $p = 1$ [6]. In this paper, we follow a different strategy replacing the above variational problem by

$$\hat{\mathbf{m}}: \quad \min_{\mathbf{m}} \frac{1}{2} \|\mathbf{C}_{\mathbf{n}}^{-1/2} (\mathbf{d} - \mathbf{m})\|_2^2 + \mu J(\mathbf{m}). \tag{3}$$

In this formulation, the necessity of estimating a wavelet, $\mathbf{\Phi}$, has been omitted. Noisy data is again represented by $\mathbf{d}$ but now $\mathbf{m}$ is the noise-free model on which an additional penalty function is imposed, such as a L^1-norm, i.e. $J(\mathbf{m}) = \|\mathbf{m}\|_1$. The noise term is now given by the predicted noise and this explains the emergence of the covariance operator, whose kernel is given by

$$\mathbf{C}_{\mathbf{n}} = E\{\mathbf{n}\mathbf{n}^T\}. \tag{4}$$

Question now is can we find a basis-function representation which is (i) sparse and *local* on the model (noise-free data), $\mathbf{m}$, data, $\mathbf{d}$ and predicted noise $\mathbf{n}$ and (ii) almost diagonalizes the covariance of the model and the predicted noise. Answer is affirmative, even though the condition of locality is non-trivial. For instance, decompositions in terms of principle components/Karhuhnen-Loéve-basis (KL) or independent components may diagonalize the covariance operators but these decomposition are generally **not** *local* and not necessary the same for both noise and model.

By selecting a basis-function decomposition that is local, sparse and almost diagonalizing (i.e. $\mathbf{C}_{\tilde{\mathbf{n}}} \approx \mathrm{diag}\{\mathbf{C}_{\tilde{\mathbf{n}}}\} = \mathrm{diag}\{\mathbf{\Gamma}^2_{\tilde{\mathbf{n}}}\}$), we find a solution to the above denoising problem that minimizes (mini) the maximal (max) mean-square-error given the worst possible Bayesian *prior* for the model [9, 5]. This minimax estimator minimizes

$$\hat{\mathbf{m}}_0 : \quad \min_{\mathbf{m}} \frac{1}{2} \|\mathbf{\Gamma}^{-1}_{\tilde{\mathbf{n}}} \left(\tilde{\mathbf{d}} - \tilde{\mathbf{m}}\right) \|_2^2 \tag{5}$$

by a simple diagonal non-linear thresholding operator

$$\hat{\mathbf{m}} = \mathbf{B}^\dagger \mathbf{\Gamma} \Theta_{\boldsymbol{\lambda}} \left(\mathbf{\Gamma}^{-1} \mathbf{B} \mathbf{d}\right) = \mathbf{B}^\dagger \Theta_{\boldsymbol{\lambda} \mathbf{\Gamma}} \left(\tilde{\mathbf{d}}\right). \tag{6}$$

The threshold is set to $\lambda \mathrm{diag}\{\mathbf{\Gamma}\}$ (we dropped the subscript $\tilde{\mathbf{n}}$) with λ an additional control parameter, which sets the confidence interval (e.g. the confidence interval is 95 % for $\lambda = 3$) (de)-emphasizing the thresholding. This thresholding operation removes the bulk of the noise by shrinking the coefficients that are smaller then the noise to zero and brings us in the convex region of the global optimization problem of Eq. 3. The symbol † indicates (pseudo)-inverse, allowing us to use Frames rather then orthonormal bases. Before going onto detail on how to impose the additional penalty term, let us first focus on the selection of the appropriate basis-function decomposition that accomplishes the above task of replacing the adaptive subtraction problem, given in Eq. 2, by its diagonalized counterpart (cf. Eq. 6).

Basis-function for seismic imaging and processing

Non-linear estimators of the type given in Eq. 6 are known to asymptotically obtain minimax [5] for particular combinations of basis functions and classes of functions. For example, thresholding of the wavelet coefficients for piece-wise smooth functions (to be more specific functions that lie in particular Besov spaces) is minimax ands hence gives results with near optimal SNR. Can these results be extended to functions that contain singularities/edges on curves and to cases where an (imaging operator) is involved? In other words, can we find basis functions that form an unconditional basis for seismic data, which generally can be considered to be the result of the (repeated) action of the scattering operator its adjoint (both asymptotically Fourier Integral Operators (FIO)) and the composition of these two operators, the normal operator (which is an elliptic Pseudo Differential Operator (ΨDO))?

To answer this question let us first explain what is meant by an unconditional basis. Without being overly technical, (almost) unconditional bases obey the following relationship

$$\|\tilde{\mathbf{f}}'_{\boldsymbol{\mu}}\|_{discretefancy} \leq \|\tilde{\mathbf{f}}_{\boldsymbol{\mu}}\|_{discretefancy}, \quad \forall \mu \tag{7}$$

for their discrete norms. These discrete norms on the μ-indexed coefficients can be quite intricate (e.g. Besov norms, i.e. non-L^2) hence the name fancy. In an unconditional basis, these discrete norms are equivalent to norms on the original continuous function, i.e. $\|f\|_{continuousfancy} \asymp \|\tilde{\mathbf{f}}_\mu\|_{discretefancy}$. Moreover, these discrete norms decay whenever we shrink the coefficients, i.e. $\|\tilde{\mathbf{f}}'_\mu\| = \|s_\mu \tilde{\mathbf{f}}_\mu\|$ with $|s_\mu| \leq 1$. Implications, of this property are profound because it means that the (fancy) norm *always* decreases, irrespective of the sign or phase of s_μ. This decrease only depends on the *magnitude* of s_μ. Moreover, the function can be shown to remain in the same "class", which means as much that the edges are preserved. Translated to our language, this property means that we no longer need to know the *'phase'* of our predicted noise. It suffices to only know the local standard deviation (proportional to *magnitude* of the Curvelet coefficients of the predicted noise) of the predicted noise, which is used to define the threshold in the non-linear estimator (cf. Eq. 6). We used the quotes for the 'phase' because we are only referring to local phase rotations over the ('compact') support of the localized basis functions. This local phase is different from the phase in the Fourier coefficients, which completely governs the location of certain events. Needless to say Fourier bases are **not** unconditional.

Besides the favorable unconditional-basis property (near) unconditional bases also obtain almost optimal non-linear approximation rates. Curvelets as proposed by [4] and [2] respectively, constitute a family of relatively new nearly unconditional non-separable wavelet bases that sparsely represent functions with singularities (read wavefronts) that lie on piece-wise smooth curves. For these type of functions, Curvelets obtain near optimal non-linear approximation rates (NAR's). For comparison, the first m-term reconstruction of the sorted 2-D Fourier coefficients obtains for a 2-D function $\mathbf{x}$ with curved singularities [3, 2, 4] the following NAR (measured in the L^2-norm),

$$\|\mathbf{x} - \tilde{\mathbf{x}}_m\|_2^2 \propto m^{-1/2}$$

while separable wavelets (ordinary $2-D$ discrete wavelet transforms, again ordered in decreasing size over their index set) obtain

$$\|\mathbf{x} - \tilde{\mathbf{x}}_m\|_2^2 \propto m^{-1}.$$

This improvement for wavelets has mainly been responsible for heir success in signal/image processing [9]. Clearly, the further improvement in NRA by Curvelets, which obtain (ordered over μ-index in decreasing order)

$$\|\mathbf{x} - \tilde{\mathbf{x}}_m\|_2^2 \propto Cm^{-2}(\log m)^3,$$

will likely open up a whole suit of successful applications. Besides the logarithmic term, the above NRA is equivalent to the best possible approximation rate (e.g. by an adaptive triangular meshing, given the location of the edges/reflection events) attainable for this type of functions [2, 4].

In addition to the optimal NRA, Curvelets almost diagonalize FIO's and ΨDO's [1] the asymptotic "building blocks" of seismic modeling and imaging. Consequently, we can expect the covariance matrices of seismic data and possible sources of coherent noise such as multiples and ground-roll to be almost diagonalized by Curvelets, i.e. Curvelets are close to the KL-bases which, by definition, are the basis that diagonalizes the Covariance operator.

So how do Curvelets obtain such a high non-linear approximation rate? Without being all inclusive [see for details [2, 4]], the answer to this question lies in the fact that Curvelets are

- *multi-scale*, i.e. they live in different dyadic corona (see Figure 1) in the FK-domain.
- *multi-directional*, i.e. they live on wedges within these corona (see Figure 1).
- *anisotropic*, i.e. they obey the following scaling law width $\propto$ length^2.
- *directional selective* with # orientations $\propto \frac{1}{\sqrt{\text{scale}}}$.
- *local* both in (x, t) and (k, f).
- almost *orthogonal*, they are *tight* frames with a moderate redundancy. Contourlets implement the pseudo-inverse in closed-form while Curvelets provide the transform and its adjoint, yielding a pseudo-inverse computed by iterative Conjugate Gradients.

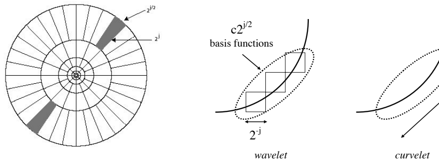


Fig. 1: **Left:** Curvelet partitioning of the frequency plane [modified from [3]]. **Right:** Comparison of non-linear approximation rates Curvelets and Wavelets [modified from [4]].

In Figure 1, a number of Curvelet properties are detailed [adapted from [3, 2, 4]]. Figure 1 describes the Curvelet partitioning of the frequency-wavenumber plane. One Curvelet lives in a wedge and becomes more *directional selective* and *anisotropic* for the higher frequencies (while moving away from the origin). Curvelets are localized in both the space (or (x, t)) and spatial (KF)-domains and have, as consequence of their partitioning, the tendency to align themselves with curves/wavefronts (see Figure 1). As such they are more flexible then a representation yielded by e.g. high-resolution Radon [as described by e.g. [13]], because they are local and able to follow any piece-wise smooth curve. Only Curvelets that align with reflectors yield large inner products and this explains their success in solving the denoising problem for models that contain events on curves. The noise is canceled because the basis functions adapt themselves locally to the dip and hence optimizing their overlap with the data integrating out the incoherent noise.

To demonstrate the insensitivity of Curvelet denoising to variations in the local phase, we include an example where we compare the thresholding results (cf. Eq. 6) for ground-roll removal given two noise predictions that are 90-degrees out of phase. We used the noise predicted by high-res parabolic Radon and we subsequently took the Hilbert transform to apply the phase shift. The results are summarized in Fig. 2.

Sparseness constraints by global optimization

Having some basic understanding why Curvelets work, we now turn our attention towards the issue how to impose additional sparseness constraints ($J(\mathbf{m})$) that reflect our *prior* knowledge. By applying the thresholding operator (cf. Eq. 6) with the threshold set according to (i) the standard deviation of the Curvelet transform for the predicted noise: $\lambda|\mathbf{Bn}|$ with $\mathbf{B}$ and $\mathbf{n}$ the predicted noise and the Curvelet transform, respectively; (ii) a Monte-Carlo sampling for the diagonal of the Covariance operator of the noise. The first method uses actual predictions for the noise, which are either based on some physical model [6] (e.g. physical models for multiples or ground roll [10]) or on noise predicted by conventional denoising techniques such as high-res Radon [13]. The second method uses information on an operator that colors the noise. Migration denoising is an example of the second method where the migration operator colors the noise. Refer for this latter application and multiple elimination to [7, 8] or to other contributions of the first author to the Proceedings of this Conference.

It can be shown that thresholding (cf. Eq.6) brings us in the convex of the denoising problem formulated in Eq. 3 for $J(\mathbf{m} = \|\mathbf{m}\|_1)$. Remains to impose the additional *prior* information residing in $J(\mathbf{m})$. By setting this constraint to the L^1-norm we hope to (i) impose minimum structure; (ii) remove possible estimation and side-band artifacts, enhancing the continuity along the imaged reflectors. To accomplish these goals, we formulate the following constrained optimization problem [2]

$$\hat{\mathbf{m}}: \quad \min_{m} J(\mathbf{m}) \quad \text{s.t.} \quad |\tilde{\mathbf{m}} - \hat{\tilde{\mathbf{m}}}_0|_\mu \le \mathbf{e}_\mu, \quad \forall\mu, \tag{8}$$

where our initial guess estimated by thresholding, $\hat{\tilde{\mathbf{m}}}_0$, is updated, subject to the constraints

$$\mathbf{e}_\mu = \begin{cases} \mathbf{\Gamma}_\mu & \text{if} \quad |\hat{\tilde{\mathbf{m}}}_0|_\mu \quad \ge \quad |\lambda\mathbf{\Gamma}|_\mu \\ \boldsymbol{\lambda\Gamma}_\mu & \text{if} \quad |\hat{\tilde{\mathbf{m}}}_0|_\mu \quad < \quad |\lambda\mathbf{\Gamma}|_\mu. \end{cases} \tag{9}$$

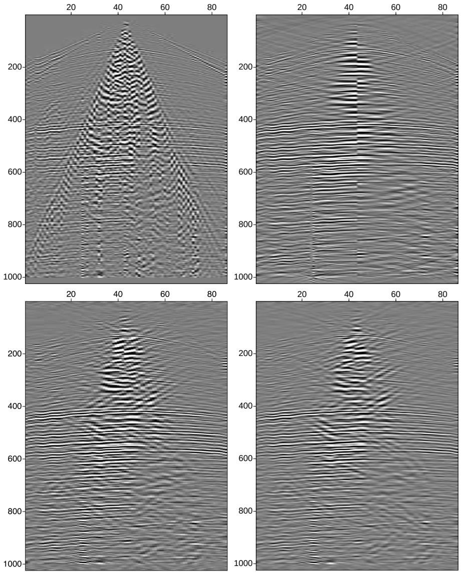


Fig. 2: Illustration of the robustness of adaptive subtraction by thresholding applied to ground-roll removal for the Yilmaz' ozdata20 dataset. **Left panel:** noisy data with ground roll; **Second panel:** high-res Parabolic Radon denoised example, which was used to predict the noise [13]. Notice the discontinuity at zero offset. **Third panel:** denoised data by thresholding with noise predicted by Radon; **Fourth panel:** the same but now for 90 degrees phase rotated predicted noise. The insensitivity of thresholding w.r.t. local phase is clear from this example.

This global optimization procedure is solved using an Augmented-Lagrangian technique [12, 7] with the Lagrangian multipliers initialized by the thresholded result, $\hat{\tilde{\mathbf{m}}}_0$. The optimization runs over all the index space μ and the tolerance differs by λ for those coefficients that have initially been thresholded to obtain the first estimate and those that survived the thresholding. The coefficients that were perceived to be noise are allowed to vary more as part of the optimization.

Applications

Application of the presented non-linear adaptive subtraction method are numerous. It can be applied to problems involving coherent or incoherent noise removal. As shown in an another paper submitted to these Proceedings, our method can readily be extended to include an (imaging) operator [7]. The applications can roughly be divided into *adaptive subtraction* for

- **multiple elimination**, where predicted multiples and multiple-free data are represented by the coherent noise $\mathbf{n}$ and model $\mathbf{m}$, respectively [in these proceedings and in 8]. See Fig. 3 for an example.
- **ground-roll elimination**, where ground roll-free data is the model $\mathbf{m}$ and ground roll the noise $\mathbf{n}$ [14]. See Fig. 2 for an example.
- **computation 4-D difference cubes**, where one vintage dataset is the noise of the other and *vice versa* [see 11].

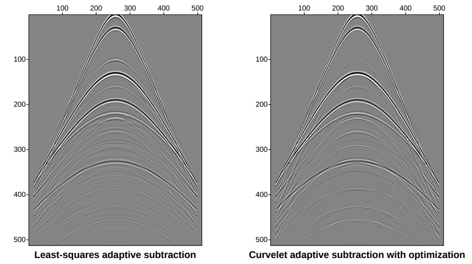
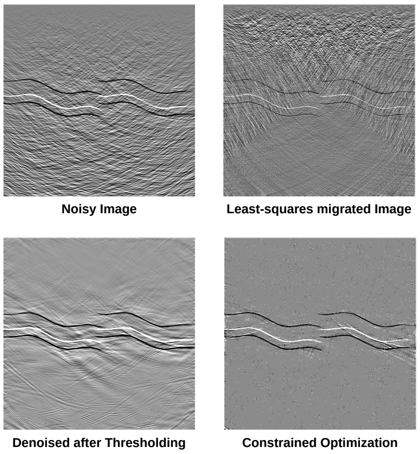


Fig. 3: Synthetic example of multiple elimination. **Left panel:** Windowed least-squares method [see e.g. 6]. **Right panel:** Our sparseness-constrained denoising based on initial thresholding followed by the constrained optimization (cf. Eq. 8).


Noisy Image

Least-squares migrated Image

Denoised after Thresholding

Constrained Optimization

Fig. 4: Synthetic common-offset Kirchoff-migration example. **Top row:** Noisy migrated and least-square migrated image with 10 interations of CG. Notice the coloring of the noise and the failure of CG to correct for the normal operator. **Bottom row:** Thresholded image (left) and constrained optimized least-squares image (right). Notice, the significant improvement in the signal-to-noise ratio by the thresholding. The constraint optimization (cf. Eq. 8) removes the artifacts and restores the least-squares amplitudes, leading to a drastically improved image.

and *sparseness-constrained least-squares migration* and AVO-inversion, where an operator involved (e.g. least-squares migration, where the migration operator colors the noise [7, see also these Proceedings]). As we can see from the example shown in Fig. 4, thresholding followed by optimization yields a substantially improved image, where most of the coherent energy has been removed. Conversely, least-squares migration by Conjugate gradients concentrates the energy of the noise towards regions that are not well insonified and fails to restore the amplitudes. Our method, on the other hand, removes most of the artifacts and restores the amplitudes for the larger angles.

Conclusions

Why do Curvelets seem to be the appropriate choice to formulate problems in seismic processing and imaging. The answer to this question is relatively simple. Curvelets provide an almost unconditional basis not only for functions with singularities on curves but also for functions that are generated by operators that approximately solve the wave equation. The *superior* non-linear approximation rate, locality both in position and dip, allow for the construction of non-linear estimators, based thresholding and possibly supplemented by global optimization. Unlike subtraction, thresholding simply puts a certain "feature" to zero if it belongs to the noise. The decision to mute only depends on the magnitude of the noise and our method is therefore robust under local phase rotations. In addition, the unconditional basis property corresponds to an almost diagonalization of the covariance operators (and also imaging operators), which explains their success of bringing us close to the solution of the sparseness-constrained adaptive subtraction problem. Because Curvelets are a non-adaptive transformation, they form an attractive alternative to data-adaptive methods, while allowing for maximal flexibility in dealing with non-stationary data.

Acknowledgments

The authors thank Emmanuel Candés and David Donoho for making an early version of their Curvelet code (Digital Curvelet Transforms via Unequispaced Fourier Transforms, presented at the ONR Meeting, University of Minnesota, May, 2003) available. We also would like to thank Mauricio Sacchi for his migration code and Yarham Carson and Daniel Trad for helping to make the ground role example. This work was in part financially supported by a NSERC Discovery Grant.

References

[1] E. J. Candès and L. Demanet. Curvelets and fourier integral operators. 2002. URL `http://www.acm.caltech.edu/~emmanuel/publications.html`. to appear Comptes-Rendus de l'Academie des Sciences, Paris, Serie I.

[2] E. J. Candès and F. Guo. New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction. *Signal Processing*, pages 1519–1543, 2002. URL `http://www.acm.caltech.edu/~emmanuel/publications.html`.

[3] Emmanual J. Candès and David L. Donoho. Recovering Edges in Ill-posed Problems: Optimality of Curvelet Frames. *Ann. Statist.*, 30:784–842, 2000.

[4] M. Do and M. Vetterli. *Beyond wavelets*, chapter Contourlets. Academic Press, 2002.

[5] David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998. URL `citeseer.nj.nec.com/donoho92minimax.html`.

[6] A. Guitton. Adaptive subtraction of multiples using the l^1-norm. *Geophys Prospect*, 52(1):27–27, 2004. URL `http://www.blackwell-synergy.com/links/doi/10.1046/j.1365-2478.2004.00401.x/abs`.

[7] Felix J. Herrmann and Peyman Moghaddam. Curvelet-domain least-squares migration with sparseness constraints. In *Expanded Abstracts*. EAGE, 2004.

[8] Felix J. Herrmann and Eric Verschuur. Separation of primaries and multiples by non-linear estimation in the curvelet domain. In *Expanded Abstracts*. EAGE, 2004.

[9] S. G. Mallat. *A wavelet tour of signal processing*. Academic Press, 1997.

[10] George A. McMechan and Mathew J. Yedlin. Analysis of dispersive waves by wave field transformation. *Geophysics*, 46, 1981. URL `http://link.aip.org/link/?GPY/46/869/1`.

[11] Felix J. Herrmann Moritz Beyreuther, Jamin Cristall. Curvelet denoising of 4d seismic. In *Expanded Abstracts*. EAGE, 2004.

[12] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer, 1999.

[13] D. Trad, T. Ulrych, and M. Sacchi. Latest views of the sparse radon transform. *Geophysics*, 68(1):386–399, 2003.

[14] Daniel Trad Yarham Carson and Felix J. Herrmann. Curvelet processing and imaging: adaptive ground roll removal. In *Expanded Abstracts*. CSEG, 2004.