Geophys. J. Int. (2007) **170,** 781–799 doi: 10.1111/j.1365-246X.2007.03360.x

Non-linear primary-multiple separation with directional curvelet frames

Felix J. Herrmann,[1] Urs Böniger[1,*] and Dirk Jacob (Eric) Verschuur[2]

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

[2]*Laboratory of Acoustical Imaging and Sound Control, Faculty of Applied Sciences, Delft University of Technology,* PO Box 5046, 2600 *GA, Delft, The Netherlands. E-mail: d.j.verschuur@tudelft.nl*

Accepted 2007 January 10. Received 2007 January 10; in original form 2006 March 14

SUMMARY

Predictive multiple suppression methods consist of two main steps: a prediction step, during which multiples are predicted from seismic data, and a primary-multiple separation step, during which the predicted multiples are 'matched' with the true multiples in the data and subsequently removed. This second separation step, which we will call the estimation step, is crucial in practice: an incorrect separation will cause residual multiple energy in the result or may lead to a distortion of the primaries, or both. To reduce these adverse effects, a new transformed-domain method is proposed where primaries and multiples are separated rather than matched. This separation is carried out on the basis of differences in the multiscale and multidirectional characteristics of these two signal components. Our method uses the curvelet transform, which maps multidimensional data volumes into almost orthogonal localized multidimensional prototype waveforms that vary in directional and spatio-temporal content. Primaries-only and multiples-only signal components are recovered from the total data volume by a non-linear optimization scheme that is stable under noisy input data. During the optimization, the two signal components are separated by enhancing sparseness (through weighted ℓ_1-norms) in the transformed domain subject to fitting the observed data as the sum of the separated components to within a user-defined tolerance level. Whenever, during the optimization, the estimates for the primaries in the transformed domain correlate with the predictions for the multiples, the recovery of the coefficients for the estimated primaries will be suppressed while for regions where the correlation is small the method seeks the sparsest set of coefficients that represent the estimation for the primaries. Our algorithm does not seek a matched filter and as such it differs fundamentally from traditional adaptive subtraction methods. The method derives its stability from the sparseness obtained by a non-parametric (i.e. not depending on a parametrized physical model) multiscale and multidirectional overcomplete signal representation. This sparsity serves as *prior* information and allows for a Bayesian interpretation of our method during which the log-likelihood function is minimized while the two signal components are assumed to be given by a superposition of prototype waveforms, drawn independently from a probability function that is weighted by the predicted primaries and multiples. In this paper, the predictions are based on the data-driven surface-related multiple elimination method. Synthetic and field data examples show a clean separation leading to a considerable improvement in multiple suppression compared to the conventional method of adaptive matched filtering. This improved separation translates into an improved stack.

Key words: curvelet transform, denoising, ℓ_1-norm, multiples, non-linear optimization, sparseness.

INTRODUCTION

In complex areas, multiple suppression techniques based on moveout filtering fail because the assumptions on the hyperbolic moveout in the CMP-offset domain are not met. Furthermore, the occurrence of shallow, high-velocity layers can lead to small moveout differences between primaries and multiples which are difficult to interpret. These complications may result in unsatisfactory separation of primaries and multiples.

In cases where moveout filtering based methods fail, 'wave-equation'-based predictive methods (Verschuur *et al.* 1992;

*Presently at: Universität Potsdam, Institut für Geowissenschaften, Karl-Liebknecht-Strasse 24–25, 14476 Golm, Germany.

Fokkema & van den Berg 1993; Weglein *et al.* 1997) have shown considerable improvements. Wave-equation methods consist of two main steps: the multiple-prediction and the primary-multiple separation step. The separation step is often referred to as adaptive subtraction, during which imperfections in the predictions, such as the water bottom reflectivity (see e.g. Berryhill & Kim 1986; Wiggins 1988; Lokshtanov 1999) or source and receiver characteristics (Verschuur *et al.* 1992; Berkhout & Verschuur 1997; Ikelle *et al.* 1997), are absorbed by a matched-filtering procedure. This procedure is important because predictions for surface-related as well as internal multiples based on 2-D input data (see e.g. Verschuur *et al.* 1992; Berkhout & Verschuur 1997; Coates & Weglein 1996) are often inaccurate in situations where the subsurface displays 3-D complexity. Other complications determining the success of multiple attenuation include source–receiver directivity, ghosts and the obliquity factor (which gives rise to an effective directivity); unbalanced amplitudes of multiple predictions that consist of mixtures of different-order multiples (Verschuur & Berkhout 1997; Chen *et al.* 2004) and incomplete data, for example, due to missing near offsets or unequal source and receiver spacing both of which may give rise to artefacts in the predicted multiples (see Verschuur 2006).

Several attempts have been made to improve multiple elimination by either increasing the accuracy of the multiple predictions or by devising a more robust subtraction/separation methodology. Examples of the first approach are methods based on model-driven time delays, as proposed by Ross (1997) and Ross *et al.* (1997), or methods based on data-driven time delays by Ikelle & Yoo (2000). Decomposition of the predicted multiples into coherent and incoherent components is an example of the second approach (Kabir 2003), where the incoherent signal component is assumed to mostly contain diffracted multiples. In that approach, both components are simultaneously subtracted from the input data. Another example is the approach taken by Wang (2003) who improves the adaptive subtraction by introducing additional local time and phase shifts.

By allowing the matched filter to be non-stationary, yielding an estimated wavelet that varies, significant improvements have been achieved in multiple suppression. There are however limits on the performance of this non-stationary matched filtering technique amongst which (i) the allowable degree of non-stationarity of the error in the multiple prediction, that is, the degree with which the matched filter is allowed to vary; (ii) the inability of the matched filter to handle different errors in the phase, location, dip and frequency characteristics of the different multiple predictions and its difficulty to handle situations where different predicted multiples overlap; (iii) the presence of noise and possible edge effects on the estimates for the local matched filter and (iv) the ability to stably apply this filter.

The primary goal of this paper is to present an alternative primary-multiple separation scheme that recovers both signal components from imperfect predictions for the multiples and from noisy data. These imperfections may include shifts, phase rotations and unknown non-stationarity of the source and receiver characteristics. With the proposed method, we aim to (i) remove the sensitivity of matched filtering to the accuracy of the predicted multiples; (ii) avoid the creation of spurious artefacts and (iii) limit possible distortions of the estimated primaries.

This paper builds upon the extensive body of literature where sparsity of certain signal representations is exploited (see for instance, the seminal work by Claerbout & Muir 1973), a concept widely employed in the geophysical sciences with applications ranging from deconvolution (Oldenburg *et al.* 1981; Ulrych & Walker 1982; Levy *et al.* 1988; Sacchi *et al.* 1994) to filtering based on high-resolution Fourier (Sacchi & Ulrych 1996; Zwartjes & Gisolf 2006) and Radon transforms (Trad *et al.* 2003) and adaptive subtraction for multiple attenuation (Guitton & Verschuur 2004). In the Bayesian context, these methods correspond to invoking long-tailed (Cauchy) distributions which also promote sparsity and lead to similar formulations. Recent developments in the theory of stable signal recovery (Donoho & Tsaig 2006; Candès *et al.* 2006b; Donoho 2006; Donoho *et al.* 2006), and signal separation by morphological component analysis (MCA) (Starck *et al.* 2004; Elad *et al.* 2005) derive from the same principles and provide additional insights on the conditions for recovery and signal separation and on new multiscale and multidirectional transforms that are sparse. As with the MCA, our approach of signal-separation derives from seeking a representation that is sparse for the two signal components. A signal is considered sparse in a representation when the magnitude-sorted coefficients in the transformed domain decay rapidly. It is shown that this sparsity not only reduces the 'dimensionality' of the problem but it also leads to a separation scheme that is relatively insensitive to errors in the predicted signal components. The idea of primary-multiple separation by non-linear optimization dates back to earlier work by the authors (see e.g. Herrmann & Verschuur 2004, 2005) and can be seen as an extension of the energy-norm based work of Nemeth & Bube (2001), Trad (2001) to the non-linear case. Our early results were based on thresholding in the curvelet domain and this paper extends these results towards a formulation based on non-linear optimization.

Outline

First, we discuss the canonical denoising problem for orthornormal and overcomplete sparsity representations. We show that this denoising problem can be cast into an optimization problem that can be solved by iterative thresholding. During the optimization, sparsity in the transformed domain is exploited by minimizing the ℓ_1-norm on the coefficients of the transformed-domain vector, referred to as the sparsity vector. It is shown that the coherent signal component can be stably recovered from the noise by virtue of sparsity. We show that this recovery can be generalized to the problem of separating two coherent signal components in the presence of noise given a prediction for these components. Again, sparsity is exploited leading to a separation scheme that is stable and relatively insensitive to errors in the predicted components. After formulating the separation problem in terms of a non-linear optimization problem, the appropriate domain for primary-multiple separation is selected by comparing the performance in the physical, Fourier, wavelet and curvelet domains. It is demonstrated that the curvelet transform not only obtains the best sparsity on the two signal components but it is also shown that this transform leads to a separation based on the local multiscale and multidirectional characteristics of the two signal components. We conclude by illustrating our algorithm on synthetic as well as real data, while making comparisons with the conventional adaptive subtraction method.

STABLE SIGNAL RECOVERY FROM OVERCOMPLETE REPRESENTATIONS

Mathematically, the problem of primary-multiple separation corresponds to a joint estimation of the primary and multiple signal components from noisy data

$$\mathbf{y} = \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{n}, \tag{1}$$

given a prediction $\check{\mathbf{s}}_2$ for the multiples. We used the symbol $\check{}$ to denote predicted quantities (as opposed to estimated quantities that

are the output of the presented separation procedure) that serve as input to our method. The recorded total data set includes the unknown primaries, $\mathbf{s}_1$, and multiples, $\mathbf{s}_2$, and is represented by $\mathbf{y} = \mathbf{s} + \mathbf{n}$ with $\mathbf{s} := \mathbf{s}_1 + \mathbf{s}_2$. The additional noise term, $\mathbf{n}$, is included to allow for possible deviations with respect to this signal model. These deviations can be caused by measurement errors or by unmodelled signal components. The noise term is given by a zero-centred discrete white Gaussian noise process, that is, with for each sample $n_i \in N(0, \sigma)$ with σ the standard deviation.

Our primary-multiple separation method derives from a generalization of the classical denoising problem, where the deterministic coherent signal component is recovered from noisy data by exploiting sparsity in a transformed domain (see e.g. Starck *et al.* 2004; Elad *et al.* 2005; Candès *et al.* 2006b; Donoho *et al.* 2006). Before deriving our extension towards primary-multiple separation, we first describe non-linear signal recovery for sparse orthonormal and later for sparse overcomplete signal representations.

Denoising by non-linear optimization

Denoising aims to recover the unknown noise-free data M-vector,[1] that is, the vector $\in \mathbb{R}^M$, from noisy data

$$\mathbf{y} = \mathbf{s} + \mathbf{n}, \tag{2}$$

with the noise term defined as before. Following recent developments in theoretical signal processing, the noise-free data $\mathbf{s}$ can be recovered from noisy and possibly incomplete measurements $\mathbf{y}$ when the data volume permits a sparse representation with respect to some possibly overcomplete signal representation, that is,

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n} \tag{3}$$

for a sparse vector $\mathbf{x}_0$ with its magnitude-sorted entries decaying rapidly. In this expression, the sparsity vector is related to the data through the sparsity synthesis or composition matrix $\mathbf{A}$. The recovery of $\mathbf{x}_0$ involves the following non-linear optimization problem

$$\mathbf{P}_1 : \quad \begin{cases} \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon \\ \widehat{\mathbf{s}} = \mathbf{A}\widehat{\mathbf{x}}, \end{cases} \tag{4}$$

which is shown to be remarkable robust under noise and missing data (Elad *et al.* 2005; Candès *et al.* 2006b; Donoho *et al.* 2006). The 'hat' symbol $\widehat{\ }$ is reserved for estimates found through optimization. As long as the sparsity synthesis matrix $\mathbf{A}$ adheres to certain conditions, the solution of this optimization problem lies within the noise level (see e.g. Elad *et al.* 2005; Candès *et al.* 2006b). This optimization problem $\mathbf{P}_1$ is the constrained variation of the basis-pursuit denoising algorithm (Chen *et al.* 2001).

As part of the optimization, the sparsity vector is fitted within the tolerance ϵ. This tolerance depends on the noise level given by the standard deviation of the noise vector $\mathbf{n}$. Since $n_{1\ldots M} \in N(0, \sigma^2)$, the probability of $\|\mathbf{n}\|_2^2$ exceeding its mean by plus or minus two standard deviations is small. The $\|\mathbf{n}\|_2^2$ is distributed according the χ^2-distribution with mean $M \cdot \sigma^2$ and standard deviation $\sqrt{2M} \cdot \sigma^2$. By choosing $\epsilon^2 = \sigma^2(M + \nu\sqrt{2M})$ with $\nu = 2$, we remain within the mean plus or minus two standard deviations.

[1] Seismic data is represented in terms of vectors that contain the seismic data volumes lexicographically sorted. The length of the M-vector corresponds to the total number of samples in the seismic data volume.

Denoising with orthonormal sparsity representations

The non-linear optimization problem $\mathbf{P}_1$ permits an explicit solution when the sparsity matrix is orthonormal in which case the transpose of the sparsity matrix corresponds to its inverse. Following Donoho (1995), Mallat (1997), Chen *et al.* (2001), $\mathbf{P}_1$ is solved by an element-wise soft thresholding procedure

$$\widehat{\mathbf{s}} = \mathbf{S}^T T_\lambda (\mathbf{S}\mathbf{y}), \tag{5}$$

with

$$T_\lambda(x) := \operatorname{sgn}(x) \cdot \max(0, |x| - |\lambda|). \tag{6}$$

This thresholding solves

$$\widehat{\mathbf{s}} = \arg\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{s}\|_2^2 + \lambda\|\mathbf{S}\mathbf{s}\|_1, \tag{7}$$

which by virtue of the orthonormality is equivalent to

$$\begin{cases} \min_{\mathbf{v}} \|\mathbf{u} - \mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_1 \\ \widehat{\mathbf{s}} = \mathbf{S}^T\widehat{\mathbf{v}}, \end{cases} \tag{8}$$

with $\mathbf{u} := \mathbf{S}\mathbf{y}$ and $\mathbf{v} := \mathbf{S}\mathbf{s}$. $\mathbf{P}_1$ is solved by setting the Lagrange multiplier λ to $\lambda = \sqrt{\epsilon^2/(M + \nu\sqrt{2M})} \cdot \sqrt{2\log M} = \sigma \cdot \sqrt{2\log M}$ (Donoho 1995; Lee & Lucier 2001; Daubechies *et al.* 2005).

Denoising with overcomplete sparsity representations

Denoising based on orthonormal transforms often does not give the most pleasing results. Compared to decimated orthonormal wavelets, non-decimated wavelets are known to give superior denoising results for functions with point-singularities (see e.g. Coifman & Donoho 1995; Starck *et al.* 2004). For non-decimated wavelets which are translation invariant, the synthesis matrix contains more columns than rows, that is, $\mathbf{A} := \mathbf{S}^\dagger = \mathbf{W}^T \in \mathbb{R}^{M \times N}$ with $N = M \cdot \log M \gg M$ is overcomplete. The symbol † is used to denote the pseudoinverse, that is, $\mathbf{S}^\dagger := (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$ with $\mathbf{S}$ the redundant transform and $\mathbf{S}^T$ its transpose. This property holds because the non-decimated wavelet transform is a tight frame[2] that preserves energy and for which we have $\mathbf{S}^T\mathbf{S} = \mathbf{I}$. Conversely, $\mathbf{S}\mathbf{S}^T$ is a projection—frames are redundant signal representations—making it difficult to recover the unknown sparsity N-vector $\mathbf{x}_0$ from $\mathbf{s} = \mathbf{A}\mathbf{x}_0$. The recovery problem becomes underdetermined and eqs (7) and (8) are no longer equivalent, an observation also made by Elad (2006).

Since the constrained optimization problem $\mathbf{P}_1$ extends to overcomplete representations, this formulation is used to recover the sparsity vector $\mathbf{x}_0$. Following Elad *et al.* (2005), the optimization problem $\mathbf{P}_1$, is replaced by a series of simpler optimization problems

$$\mathbf{P}_\lambda : \quad \begin{cases} \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ \widehat{\mathbf{s}} = \mathbf{A}\widehat{\mathbf{x}}. \end{cases} \tag{9}$$

These optimization problems depend on the Lagrange multiplier λ, which is not known. A cooling method is used where $\mathbf{P}_\lambda$ is solved for a Lagrange multiplier λ that is slowly decreased from a large starting value. The optimal $\widehat{\mathbf{x}}$ is found for the largest λ for which $\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2 \leq \epsilon$. During the optimization, the underdetermined frame matrix $\mathbf{A}$ is inverted by imposing the sparsity promoting ℓ_1-norm. This norm regularizes the inverse problem of finding the unknown coefficient vector (see also Daubechies *et al.* 2005). We refer to Donoho *et al.* (2006) and Tropp (2006) for the recovery conditions for eqs (4) and (9).

[2] A tight frame is a redundant signal representation that preserves energy.

Table 1. The cooling method with $|\mathbf{A}^T\mathbf{y}|_\infty > \lambda_1 > \lambda_2 > \cdots$ the series of decreasing Lagrange multipliers. The inner loop is repeated L times.

Initialize:
$m = 0$; $\mathbf{x}^0 = \mathbf{A}^T\mathbf{y}$;
Choose: L, $|\mathbf{A}^T\mathbf{y}|_\infty > \lambda_1 > \lambda_2 > \cdots$
while $\|\mathbf{y} - \mathbf{A}\mathbf{x}^m\|_2 > \epsilon$ **do**
  $m = m + 1$;
  $\mathbf{x}^m = \mathbf{x}^{m-1}$;
  for $l = 1$ to L **do**
    $\mathbf{x}^m = T_{\lambda_m}(\mathbf{x}^m + \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^m))$ {Iterative thresholding}
  end for
end while
$\widehat{\mathbf{s}} = \mathbf{A}\mathbf{x}^m$.

Solution by the cooling method based on iterative thresholding

Following Daubechies *et al.* (2005) and Elad *et al.* (2005) and ideas dating back to Figueiredo & Nowak (2003), eq. (9) is solved by an iterative thresholding technique that derives from the Landweber descent method. After m iterations of the outer cooling loop, during which the Lagrange multiplier is lowered, estimations for the coefficient vector are computed for fixed λ by the following inner loop

$$\mathbf{x}^{m+1} = T_\lambda[\mathbf{x}^m + \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^m)], \tag{10}$$

with $\lambda = \lambda_m$. As shown by Daubechies *et al.* (2005), this iteration for fixed λ converges to the solution of the subproblem in eq. (9) for m large enough and $\|\mathbf{A}\| < 1$. The cost of each iteration is a synthesis and subsequent analysis. The details of the cooling algorithm are presented in Table 1.

PRIMARY-MULTIPLE SEPARATION BY NON-LINEAR OPTIMIZATION

A primer: primary estimation by thresholding

In earlier work by the authors (see e.g. Herrmann & Verschuur 2004, 2005), it was shown that thresholding in the curvelet domain with a threshold defined in terms of the magnitude of the predicted multiples leads to an effective primary-multiple separation scheme. As shown below, curvelets transform seismic data volumes into a representation that allows for a separation based on scale (dominant frequency), location and angle. For those events in the data where the primaries and multiples overlap, the algorithm separates on the basis of the magnitudes of the predicted multiples in the transformed domain. These properties explain the early success of our primary-multiple separator in which the primaries are estimated with

$$\widehat{\mathbf{s}}_1 = \mathbf{S}^T T_{\mathbf{w}}(\mathbf{S}\mathbf{y}), \tag{11}$$

where $\mathbf{S}$ is the sparsity transform and $T_{\mathbf{w}}$ the soft thresholding operator with a threshold $\mathbf{w}$ that varies element by element, that is, $T^s_w(x) := \mathrm{sgn}(x) \cdot \max(0, |x| - |\mathrm{w}|)$.

Even though the above estimator cannot be expected to perform well for overcomplete signal representations for which eqs (7) and (8) are no longer equivalent, soft thresholding with a varying threshold corresponds to the first iteration of the iterative method defined earlier. This method solves $\mathbf{P}_1$ for a weighted ℓ_1-norm, $\|\mathbf{x}\|_{1,\mathbf{w}} := \sum_{\mu\in\mathcal{M}} |w_\mu x_\mu|$. For each element in the index set $\mu \in \mathcal{M}$ of the coefficient vector, the weighted ℓ_1-norm penalty behaves as

$$|w_\mu x_\mu| \propto \begin{cases} |x_\mu|^2 & \text{when} \quad x_\mu \sim w_\mu \\ |x_\mu| & \text{when} \quad x_\mu \gg w_\mu. \end{cases} \tag{12}$$

This behaviour corresponds to that of a ℓ_2-norm penalty whenever the coefficients for the predicted multiples are close to those of the total data while the penalty term acts as a ℓ_1-norm otherwise. The ℓ_2-norm penalizes the outliers while the ℓ_1-norm promotes the outliers bringing out the primaries.

Sparsity-domain primary-multiple separation

Motivated by recent results on the stable signal recovery from overcomplete representations (see e.g. Starck *et al.* 2004; Elad *et al.* 2005), the primary-multiple separation problem is formulated in terms of a non-linear optimization problem. The solution of this problem provides simultaneous estimates for the multiples and primaries given predictions for the multiples. As in stable signal recovery, the method exploits sparsity in a transformed domain for both signal components. In that respect our method differs fundamentally from matched filtering (see e.g. Verschuur & Berkhout 1997), since it exploits a representation that is sparse, that is, a transform that leads to a rapid decay for the magnitude-sorted coefficients in the sparsity vectors for the two signal components.

Sparse signal model

Following the ideas of MCA (see e.g. Starck *et al.* 2004), an augmented sparsity synthesis matrix is defined consisting of an inverse transform for the synthesis of each of the two signal components in eq. (1). Again the data is described as a sparse superposition of now two sparsity matrices one for each signal component,

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}, \tag{13}$$

with

$$\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2] \quad \text{and} \quad \mathbf{x}_0 = [\mathbf{x}_{01} \quad \mathbf{x}_{02}]^T \tag{14}$$

the augmented sparsity synthesis matrix and sparsity vector, respectively. In this formulation, the subscripts 1 and 2 are reserved for primaries and multiples. The above signal model with the coefficients of $\mathbf{x}_0$ sparse, forms the basis of MCA. Even though MCA was initially designed to separate signal components that are sparse in different sparsity representations, we show that this method can be extended to signal components with similar characteristics. Because primaries and multiples are both solutions of the wave equation, we cannot expect to find a generic overcomplete signal representation that separates these two components without providing *prior* information on the wave arrivals. We argue that these signal components can still be separated as long as there exist reasonable predictions for the signal components. These predictions are used as weights that allow us to recover the two signal components using the same signal representation for each component.

The weighted ℓ_1-norm optimization problem

If the two signal components permit a sparse representation then the predicted multiples can be used as weights in the sparsity promoting ℓ_1 norm. These weights drive the two signal components

apart during the optimization and $\mathbf{x}_0$ can be recovered to reasonable accuracy.[3] The $\mathbf{w}$-weighted optimization problem becomes

$$\mathbf{P}_{\mathbf{w}}: \quad \begin{cases} \min_{\mathbf{x}} \|\mathbf{x}\|_{\mathbf{w},1} \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon \\ \widehat{\mathbf{s}}_1 = \mathbf{A}_1\widehat{\mathbf{x}}_1 \quad \text{and} \quad \widehat{\mathbf{s}}_2 = \mathbf{A}_2\widehat{\mathbf{x}}_2 \\ \text{given:} \quad \check{\mathbf{s}}_2 \quad \text{and} \quad \mathbf{w}(\mathbf{y}, \check{\mathbf{s}}_2) \end{cases} \tag{15}$$

with $\mathbf{w} = [\mathbf{w}_1, \quad \mathbf{w}_2]^T$ the weighting vectors with strictly positive weights defined in terms of the predicted multiples. The estimates for the primaries and multiples are computed from the sparsity vector that minimizes $\mathbf{P}_{\mathbf{w}}$. During the optimization, the sparsity vector is recovered by minimizing the weighted ℓ_1 norm subject to a recovery that is within the tolerance.

The weights

The weighting vectors are based on an *a priori* prediction for the multiples, obtained by surface-related multiple elimination (SRME) prediction (see e.g. Berkhout & Verschuur 1997) or by other means. The corresponding prediction for the primaries is obtained through simple subtraction. When the predicted multiples consist of surface-related multiples only, the predicted primaries are computed according to

$$\check{\mathbf{s}}_1 = \mathbf{y} - \check{\mathbf{s}}_2. \tag{16}$$

When different predictions, such as surface-related and internal multiples or multiples of different order, are available, the predicted primaries are computed according

$$\check{\mathbf{s}}_1 = \mathbf{y} - \sum_{p=1}^{P} \check{\mathbf{s}}_{2,p}, \tag{17}$$

where the sum runs over the $P > 1$ different multiple contributions. The entries in the weighting vectors are defined in terms of the predicted primaries, multiples and the noise level (Donoho 1995; Chen *et al.* 2001),

$$\begin{cases} \mathbf{w}_1 := \max\left(\sigma \cdot \sqrt{2\log N},\ C_1|\check{\mathbf{u}}_1|\right) \\ \mathbf{w}_2 := \max\left(\sigma \cdot \sqrt{2\log N},\ C_2|\check{\mathbf{u}}_2|\right). \end{cases} \tag{18}$$

The σ is set as before, while $\check{\mathbf{u}}_1 := \max\left(\{|\widehat{\mathbf{x}}_{2,p}|\}_{p=1,\cdots,P}\right)$ and $\check{\mathbf{u}}_2 := |\check{\mathbf{x}}_1|$ with the transform-domain vectors given by $\check{\mathbf{x}}_{2,p} := \mathbf{A}_2^T\check{\mathbf{s}}_{2,p}$ and $\check{\mathbf{x}}_1 := \mathbf{A}_1^T\check{\mathbf{s}}_1$. The constants C_1 and C_2 normalize the ℓ_2-norms for the primaries and the multiples. The above definition for the weights is designed to separate the primaries and multiples whenever their predictions exceed the noise level. If the predicted coefficients are smaller then the noise level, the weights are set to remove the incoherent noise as during ordinary denoising discussed before.

Solution by the block-relaxation method

Following Elad *et al.* (2005), the constrained optimization problem $\mathbf{P}_{\mathbf{w}}$ is solved through a series of simpler optimization problems. Because the synthesis matrix consists of two parts, use is made of the block-relaxation method introduced by Bruce *et al.* (1998). Each signal component is recovered separately, while keeping the other component fixed. As with the solution of $\mathbf{P}_1$, a cooling method is used where the Lagrange multiplier is gradually lowered.

[3] For an orthonormal sparsity representation, this recovery can be expected to be within the noise level when the two sparsity vectors $\mathbf{x}_{01}$ and $\mathbf{x}_{02}$ are disjoint, that is, $x_{1,\mu} = 0$ when $x_{2,\mu} \neq 0$ or $x_{2,\mu} = 0$ when $x_{1,\mu} \neq 0$ for $\mu \in \mathcal{M}$.

Table 2. The primary-multiple separation by optimization algorithm.

Initialize: $\mathbf{x}_1^0 = \mathbf{A}_1^T\check{\mathbf{s}}_1$, $\mathbf{x}_2^0 = \mathbf{A}_2^T\check{\mathbf{s}}_2$, m = 0, $R^0 = 1$;
Choose: L, $\lambda_1 \geq \lambda_2 \geq \cdots$
while $\|\mathbf{y} - \mathbf{A}\mathbf{x}^m\|_2 > \epsilon$ and $R^m < R^{m-1}$ **do**
  $m = m + 1$, $l = 0$;
  $\mathbf{x}^m = \mathbf{x}^{m-1}$;
  $\mathbf{w}_1 := \max(\lambda \cdot \sigma, C_{1,s}|\check{\mathbf{u}}_1|)$,$\mathbf{w}_2 := \max(\lambda \cdot \sigma, C_{2,s}|\check{\mathbf{u}}_2|)$ {Set the weights}
  $R^m(\mathbf{z}_1, \mathbf{z}_2) = \frac{\mathbf{z}_1^T\mathbf{z}_2}{\|\mathbf{z}_1\|\|\mathbf{z}_2\|}$; {Compute decorrelation}
  while $l \leq L$ and ϵ and $R^m < R^{m-1}$ **do**
    $l = l + 1$;
    $\mathbf{s}_2 = \mathbf{A}_2\mathbf{x}_2^m$; {Synthesize}
    $\mathbf{r}_1 = \mathbf{y} - \mathbf{s}_2$; {Calculate residual}
    $\mathbf{x}_1^m = \mathbf{x}_1^m + \mathbf{A}_1^T(\mathbf{r}_1 - \mathbf{A}_1\mathbf{x}_1^m)$; {Descent update}
    $\mathbf{x}_1^m = \text{sign}(\mathbf{x}_1^m) \cdot \max(0, |\mathbf{x}_1^m| - |\lambda_m \cdot \mathbf{w}_1|)$; {Soft threshold}
    $\mathbf{s}_1 = \mathbf{A}_1\mathbf{x}_1^m$; {Synthesize}
    $\mathbf{r}_2 = \mathbf{y} - \mathbf{s}_1$; {Calculate residual}
    $\mathbf{x}_2^m = \mathbf{x}_2^m + \mathbf{A}_2^T(\mathbf{r}_2 - \mathbf{A}_2\mathbf{x}_2^m)$; {Descent update}
    $\mathbf{x}_2^m = \text{sign}(\mathbf{x}_2^m) \cdot \max(0, |\mathbf{x}_2^m| - |\lambda_m \cdot \mathbf{w}_2|)$; {Soft threshold}
  end while
end while

The outer loop

At each iteration of the outer loop the Lagrange multiplier is set at $\lambda = \lambda_m$ and the following optimization problems are solved

$$\widehat{\mathbf{x}}_j = \arg\min_{\mathbf{x}_j} \frac{1}{2}\|\mathbf{y} - \mathbf{A}_j\mathbf{x}_j - \sum_{i\neq j}\mathbf{A}_i\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_{1,\lambda\cdot\mathbf{w}_j} \quad j = 1, \cdots, J, \tag{19}$$

assuming the other components $\mathbf{x}_i$ for $i \neq j$ to be known. The λ_m is the cooling parameter after m iterations with $\lambda_1 > \lambda_2 > \cdots$. For the primary-multiple separation, $J = 2$, and each component is solved with the iterative Landweber descent method (Vogel 2002), supplemented with soft thresholding (Starck *et al.* 2004; Daubechies *et al.* 2005).

The inner loop

At the mth iteration for the outer loop, the subproblem for the jth component of the sparsity vector is solved (see Table 2) by iterations on

$$\mathbf{x}_j^m = T_{\lambda\cdot\mathbf{w}_j}\left[\mathbf{x}_j^m + \mathbf{A}_j^T\left(\mathbf{s} - \mathbf{A}_j\mathbf{x}_j^m - \sum_{i\neq j}\mathbf{A}_i\mathbf{x}_i\right)\right], \tag{20}$$

with the threshold set by the jth-component of the weighting vector, that is, $\mathbf{w}_j$. This inner loop is repeated L times unless the stopping criterion is met.

Stopping criterion

Because signal separation is the primary goal, lowering the Lagrange multiplier until the data is approximated to within the tolerance is not sufficient. We, therefore, introduce an additional stopping criterion that measures the improvements in the degree of decorrelation. The algorithm proceeds as long as this cross-correlation,

$$R^m(\mathbf{z}_1, \mathbf{z}_2) := \frac{\mathbf{z}_1^T\mathbf{z}_2}{\|\mathbf{z}_1\|\|\mathbf{z}_2\|} \tag{21}$$

does not increase during the iterations. This expression measures the average degree of correlation between the residues after m iterations, $\mathbf{z}_1 := \mathbf{A}_1^T(\mathbf{y} - \mathbf{x}_2^m)$, and $\mathbf{z}_2 := \mathbf{A}_2^T(\mathbf{y} - \mathbf{x}_1^m)$ in the transformed domain.

$R(\mathbf{z}_1, \mathbf{z}_2)$ equals unity when $\mathbf{z}_1 = \mathbf{z}_2$ and is zero when the two signal components are disjunct, that is, when $z_{1,\mu} = 0$ when $z_{2,\mu} \neq 0$ or $z_{2,\mu} = 0$ when $z_{1,\mu} \neq 0$ for $\mu \in \mathcal{M}$.

The algorithm

First, the entries in the sparsity vector are initialized according the predictions for the primaries and multiples, $\mathbf{x}_1^0 = \mathbf{A}_1^T \check{\mathbf{s}}_1$ and $\mathbf{x}_2^0 = \mathbf{A}_2^T \check{\mathbf{s}}_2$ as given in equation (16) or (17). After each iteration in the outer loop, the cooling parameter is decreased. The algorithm is completed when the degree of decorrelation between the residues of the two signal components no longer increases, that is, when $R^m \geq R^{m-1}$. The details of the algorithm are summarized in Table 2.

When and why should this signal separation algorithm work?

Probabilistic interpretation

The above signal-separation by optimization approach derives from a signal model (*cf.* eq. 13) given by a sparse superposition of prototype waveforms corrupted by Gaussian noise. During the optimization, the mismatch between the observations and reconstructed data is minimized in the ℓ_2 sense. This quadratic term is known as the log-likelihood function. This function is jointly minimized with the weighted ℓ_1-norm that serves as the *prior*. This weighted ℓ_1-norm corresponds to a Laplace probability distribution for the coefficients in the sparsity vector $\mathbf{x}$. Each coefficient is drawn independently from a probability density function proportional to $\exp(-\text{Const} \cdot |w_m x_m|)$. The Laplace distribution is known to generate sparse sequences (Starck *et al.* 2004; Elad 2006). This probability density function is weighted by the predictions for the primaries and multiples and has the tendency to reduce the probability of drawing an element for one signal component if the prediction for the other component is large.

Estimation by an oracle attenuation

Consider the following stylized signal separation problem

$$\mathbf{y} = \mathbf{s} + \mathbf{e} \tag{22}$$

with $\mathbf{e}$ a random coloured Gaussian noise term. Following Mallat (1997), we can write

$$\sigma_\mu^2 = \mathbf{E}\left\{\|e_\mu\|^2\right\} = \langle \mathbf{K}\mathbf{b}_\mu,\ \mathbf{b}_\mu \rangle \tag{23}$$

for the stochastic expectation (denoted by the $\mathbf{E}\{\}$) for the variance of the coloured noise in an orthonormal basis $\mathbf{B} = \{\mathbf{b}_\mu\}_{\mu \in \mathcal{M}}$. This variance depends on the basis vectors and on the covariance $\mathbf{K}$. By applying the following shrinkage,

$$\widehat{\mathbf{s}} = \sum_{\mu \in \mathcal{M}} a_\mu x_\mu \mathbf{b}_\mu \tag{24}$$

with

$$a_\mu = \frac{|x_\mu|^2}{|x_\mu|^2 + \sigma_\mu^2} \tag{25}$$

the expectation for the ℓ_2 error (the risk), that is, the $\mathbf{E}\{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2\}$, is minimized. Even though the above estimator in eq. (24) with so-called oracle attenuation attains a lower bound for the error, its practical value is limited since it requires information on the true coefficients x_μ of the unknown signal $\mathbf{s}$. A risk close to the minimal risk can be achieved by applying a diagonal soft-thresholding procedure (see e.g. Mallat 1997)

$$\widehat{\mathbf{s}} = \sum_{\mu \in \mathcal{M}} T_{w_\mu}(\langle y,\ \mathbf{b}_\mu \rangle)\mathbf{b}_\mu = \mathbf{S}^T T_{\mathbf{w}}\,(\mathbf{S}\mathbf{y})\,, \tag{26}$$

with $w_\mu = \sigma_\mu \sqrt{2 \log N}$ and $T_{\mathbf{w}}$ the soft-thresholding operation as defined earlier. This estimator does not require knowledge of the coefficients x_μ.

This diagonal formulation only works efficiently for signals $\mathbf{s}$ that are sparse in the basis $\mathbf{B}$ and for coloured noise terms that are nearly independent in the same orthonormal basis. For a Gaussian $\mathbf{e}$, a near diagonalization of the covariance is enough to guarantee sufficient uncorrelated coefficients and the diagonal estimator based on soft thresholding is close to optimal (see e.g. Mallat 1997).

The above derivation translates to our seismic signal separation problem by

(i) assuming that multiples (primaries) present in the input data act as a coherent Gaussian noise term (*cf.* eq. 22) for the unknown model, that is, the primaries (multiples). This noise term is assumed to be diagonal in the transformed domain.

(ii) replacing the soft-thresholding, valid for orthonormal bases, to the weighted minimization problems of the type $\mathbf{P}_{\mathbf{w}}$ (*cf.* eq. 15). The close relationship between soft thresholding, $\mathbf{P}_\lambda$ and the extension to overcomplete signal representations was explained earlier;

As long as there exists a sparse representation for the two signal components (primaries and multiples), our $\mathbf{w}$-weighted formulation of the primary-multiple separation problem (*cf.* eq. 15) can be expected to perform well as long as there is not too much overlap between the two signal components in the transformed domain.

The cooling

The above formulation for the primary-multiple separation is based on a recovery algorithm that exploits sparsity of a certain representation for the primaries and the multiples. This sparsity is promoted by the weighted ℓ_1-norm penalty term on the sparsity vector. The algorithm starts with a large Lagrange multiplier that emphasizes the *prior*. As a result, the recovered sparsity vector is nearly empty and contains primarily separated signal components. As the Lagrange multiplier is lowered, the penalty is relaxed and the sparsity vectors are allowed to pick up more events to fit the data. Because of the weighting, the sparsity vectors for each component are discouraged to pick up events from the other component. Hence, the signal is separated. As the cooling parameter is lowered, there is less emphasis on the *prior* and the two sparsity vectors will pick up more events in the data which will lead to an increased cross-correlation between the two sparsity vectors. To accommodate this aspect, the stopping criterion is based on the residues. Only when the correlation between the residues in the transformed domain increases, the algorithm will be terminated.

The sparsity

The algorithm's performance depends on the degree of sparseness achieved by the signal representation. This sparseness guarantees recovery and separation. Not only the relative number of entries that need to be separated is reduced but the probability of having two large entries at the same location in the two sparsity vectors is also

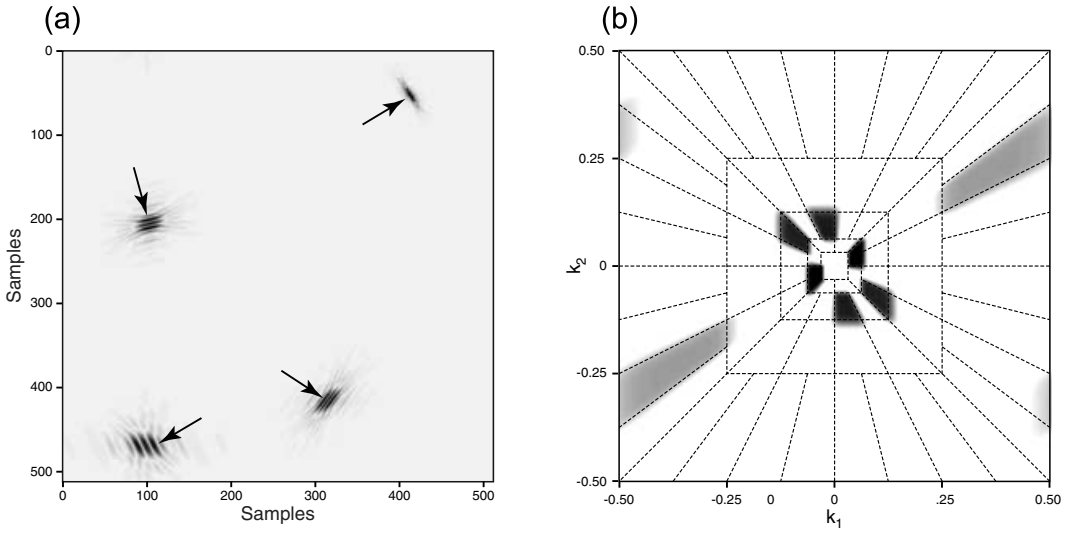

Figure 1. Spatial and frequency representation of curvelets. (a) Four different curvelets in the spatial domain at three different scales. (b) Dyadic partitioning in the frequency domain, where each wedge corresponds to the frequency support of a curvelet in the spatial domain. This figure illustrates the microlocal correspondence between curvelets in the physical and Fourier domain. Curvelets are characterized by rapid decay in the physical space and of compact support in the Fourier space. Note the correspondence between the orientation of curvelets in the two domains. The 90° rotation is a property of the Fourier transform.

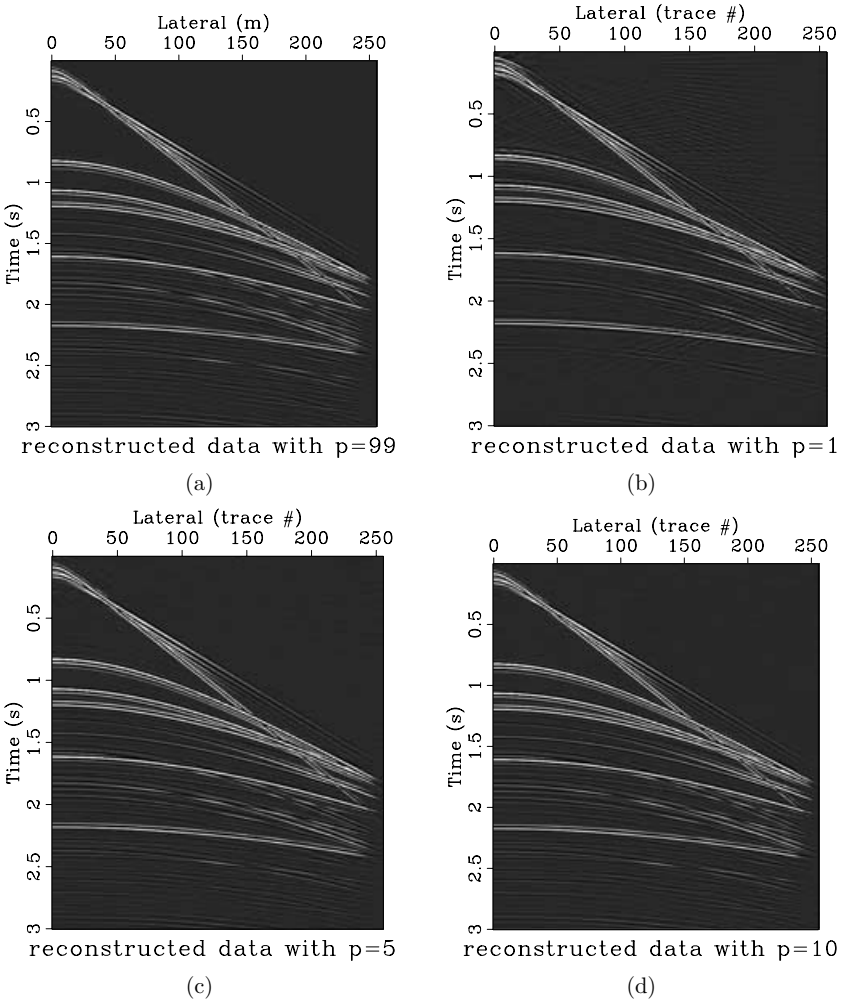

Figure 2. Illustration of the non-linear approximation by the curvelet representation for idealized synthetic data. (a) original idealized synthetic data; (b)–(d) approximations of the data in (a) with $p = 1, 5$ and 10 per cent of the total number of curvelet coefficients. The reconstruction with >1 per cent barely adds more detail.

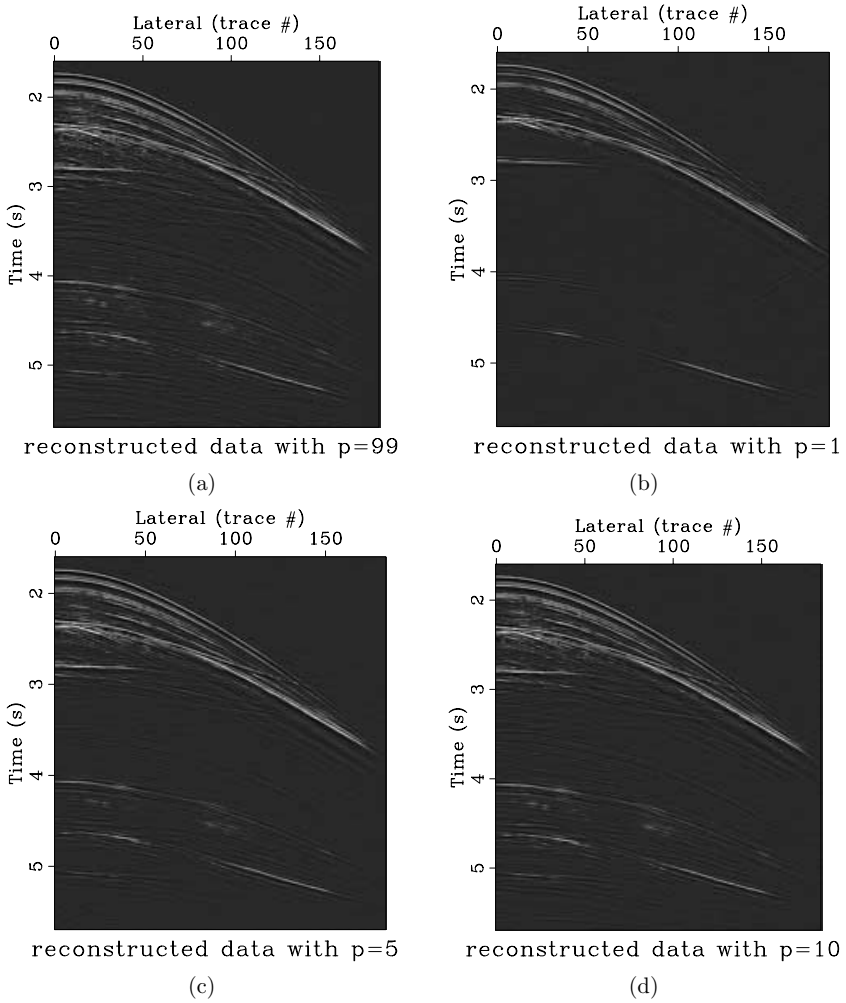

Figure 3. Illustration of the non-linear approximation by the curvelet representation for complex real data. (a) original real data; (b–d) approximations of the data in (a) with $p = 1, 5$ and 10 per cent of the total number of curvelet coefficients. Observe that the partial curvelet reconstruction for the real data performs almost as well as the synthetic data example.

diminished. The attainable sparsity and hence the performance of the algorithm depends on how well the transform is able to locally capture wavefronts, which on its turn depends on how much the prototype waveforms locally look like 'little waves'. The capability to achieve high degrees of sparsity is intrinsically linked to a near diagonalization of the signal's covariance[4] (Donoho 1993; Mallat 1997). For instance, the discrete wavelet transform is known to be an unconditional basis for certain function classes. Unconditional bases near diagonalize the covariance and are also sparse. Since both signal components are sparse in the same basis, the separation based on the thresholding (*cf.* eq. 11) corresponds to a shrinkage with an 'oracle' given by the predicted signal component (Donoho 1995; Mallat 1997). Depending on the accuracy of the prediction, shrinkage is near optimal in an unconditional basis. Even though the concept of an unconditional basis does not apply to overcomplete signal representations, we assume that some of its properties carry over to sparse overcomplete signal representations. In that

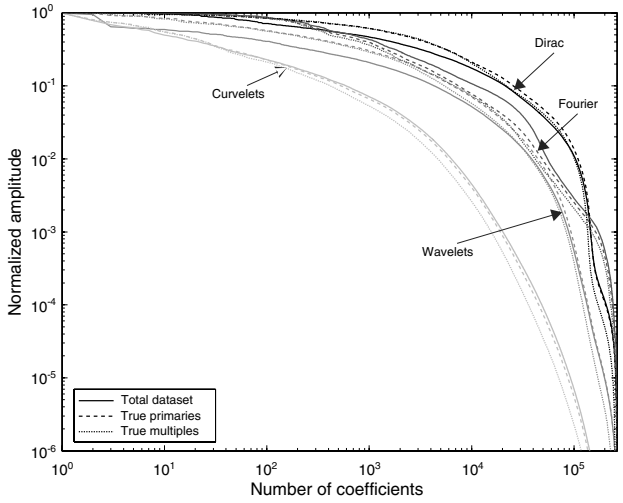

Figure 4. This plot shows the empirical decay rates for the total data **s**, the true primaries $\mathbf{s}_1$ and the true multiples $\mathbf{s}_2$ in the transformed domains defined by the Dirac **Id**, discrete wavelet **W**, Fourier **F** bases and the curvelet frame **C**. The overcomplete curvelet transform decays the fastest.

[4] In this case, we assume that the predictions for the primaries and multiples are drawn from a random process.

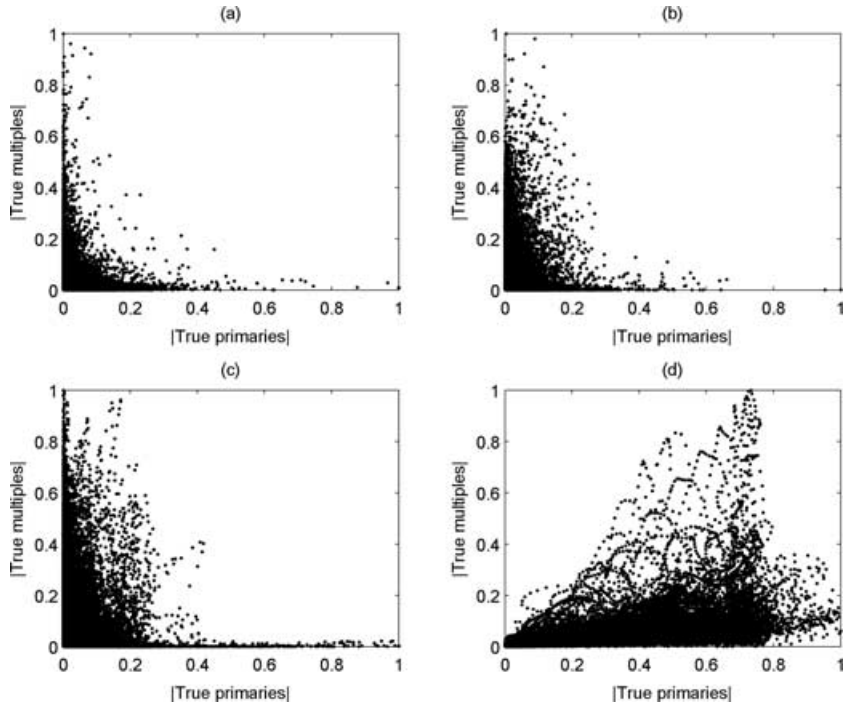

Figure 5. Cross-plots of the absolute values of the coefficients for the true primaries and multiples in the different transformed domains. (a) True coefficients in the curvelet domain. (b) True coefficients in the wavelet domain. (c) True coefficients in the Dirac domain. (d) True coefficients in the Fourier domain. Note the improved decorrelation for the curvelet coefficients which are predominantly aligned with the axes as opposed to the Dirac, Fourier and to a lesser extend the wavelet coefficients.

case, the single thresholding needs to be replaced by an iterative method.

OVERCOMPLETE SPARSITY REPRESENTATIONS FOR SEISMIC DATA

So far, our emphasis has been directed towards the formulation of a stable signal-separation scheme based on the assumption that there exists a generic non-data-adaptive representation for primaries and multiples that is sparse and leads to a near diagonalization of the covariance for these two signal components. Does such a representation exist for data containing wave fronts? With other words, do the results obtained with discrete wavelet transforms for piece-wise smooth 1-D functions carry over to higher dimensions? We argue that the recently developed curvelet transform (Candès & Donoho 2000b, 2004) is definitely a candidate. We will show that curvelets obtain high theoretical and empirical non-linear approximation rates for seismic data and we will also provide arguments why curvelets are the appropriate choice for a seismic signal representation that does not contain any information on the location and dips of the wave arrivals. Before comparing the performance of our separation algorithm with different signal representations, including the Fourier and wavelet transform, a brief introduction to the curvelet transform is given. For details the reader is referred to the Appendix A and to Candès *et al.* (2006a) and Ying *et al.* (2005) for the numerical implementation of the curvelet transform.

The curvelet transform: the appropriate domain for seismic data?

The key point of this paper is to separate primaries from multiples in a sparse non-parametric transformed domain. With a non-parametric transform, we refer to a transform that does *not* assume *a priori* information, for example, velocities or dips. Furthermore, no assumptions will be made regarding the shape, direction and frequency content of the arriving waveforms.

Until recently, defining a non-parametric transform that is sparse on seismic data has been difficult. The wavefronts present in seismic data, that may include caustics, lead to a slow decay for the Fourier coefficients and can also not be efficiently represented by the discrete wavelet transform because wavelets are not directional by construction. The prototype waveforms that make up these transforms are simply not rich enough to sparsely represent seismic data, they either lack a multiscale structure or directionality.

The recently developed curvelet transform (see e.g. Candès & Donoho 2004) compose signals in terms of waveforms that are multiscale and multidirectional. Because the rows of the transform contain prototype waveforms that behave locally like 'little waves', the curvelet transform obtains near optimal sparsity on bandwith-limited[5] seismic data (Candès *et al.* 2006a; Hennenfent & Herrmann 2006). The curvelet transform is overcomplete because the number of rows with waveforms exceeds the number of samples in the image. By using the fast discrete curvelet transform (FDCT by wrapping, see e.g. Candès *et al.* 2006a; Ying *et al.* 2005), data is perfectly reconstructed after decomposition by applying the adjoint of the curvelet transform, that is, we have $\mathbf{r} = \mathbf{C}^T\mathbf{C}\mathbf{r}$ for arbitrary $\mathbf{r}$. The computational cost of the FDCT is of the same order as the FFT. The curvelet transform matrix and its adjoint are given by $\mathbf{C}$ and $\mathbf{C}^T$ and these transforms define the sparsity synthesis and analysis matrices according $\mathbf{A} := \mathbf{C}^T \in \mathbb{R}^{M\times N}$ and $\mathbf{A}^T := \mathbf{C}$ as described

[5] Because of band limitation of the source, seismic data volumes are always limited in bandwidth containing 'wavefronts' that are relatively smooth in the direction along the wave arrivals and oscillatory across.

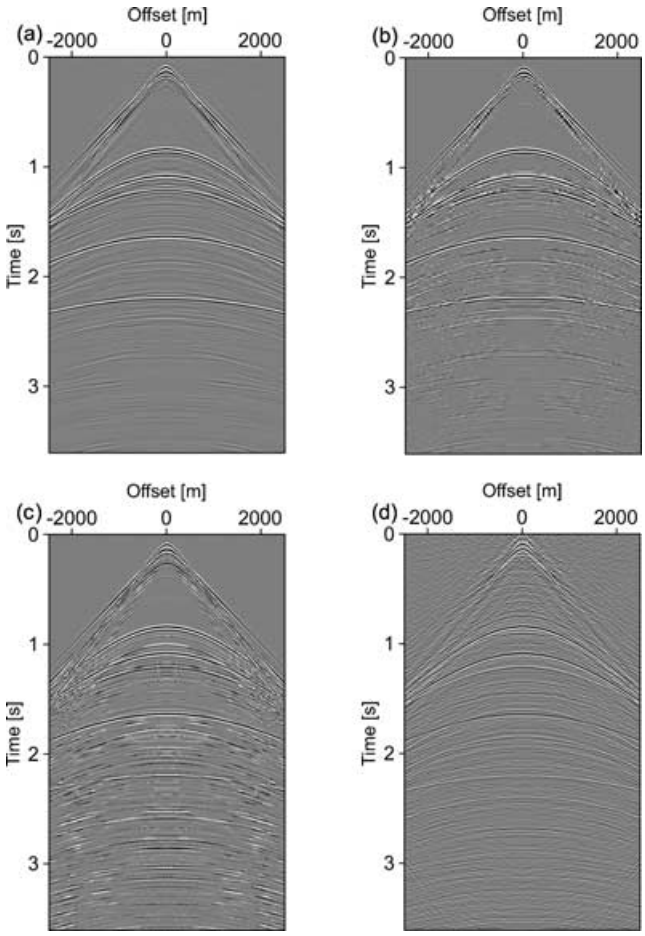

Figure 6. Comparison of the primary estimation performance for the synthetic data set of Fig. 8, based on the curvelet transform (**C**) and the orthonormal Dirac basis (**Id**), the discrete wavelet (**W**) and Fourier (**F**) transforms. The input data for the algorithm are the total data (Fig. 8a) and the predicted multiples (Fig. 8b). The comparisons are made for the iterative block relaxation method with fixed settings for the parameters: (a) estimate for the primaries using curvelets, (b) wavelets, (c) Dirac and (d) Fourier. Visual inspection of these estimates for the primaries confirm the findings listed in Table 3.

Table 3. Performance of the signal-separation for the Dirac, Fourier and wavelet bases and for the overcomplets curvelet frame. These numbers are computed for the 2-D synthetic example depicted in Fig. 6. Curvelet frames clearly score better on all fronts, including the estimation error in terms of the normalized ℓ_2-difference between the predicted and true signal components, $\varepsilon_{1,2}(\mathbf{s}) := \frac{\|\mathbf{s}-\widehat{\mathbf{s}}\|}{\|\mathbf{s}\|}$.

Separation basis	$R(\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2)$	$\varepsilon(\mathbf{s}_1)$	$\varepsilon(\mathbf{s}_2)$
Dirac (**Id**)	0.97163	3.2846	10.9635
Fourier (**F**)	0.02586	0.7937	0.6208
Wavelet (**W**)	0.01025	0.3518	0.4622
Curvelet (**C**)	0.00415	0.2172	0.3129

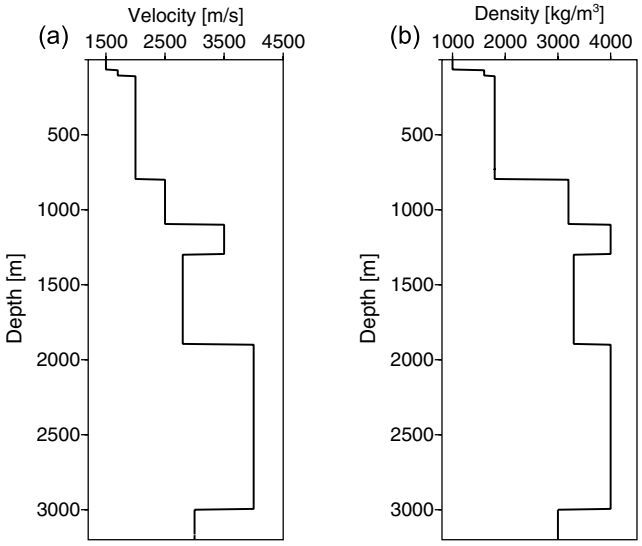

Figure 7. Vertical cross-sections of a horizontally layered acoustic model used to generate seismic reflection data. (a) Velocity profile. (b) Density profile.

in the preceding section. For this choice of curvelet transform, the pseudoinverse equals the adjoint, that is, $\mathbf{C}^T = \mathbf{C}^\dagger$, which means that the FDCT by wrapping is a numerical isometry, that is, the collection of curvelets in the overcomplete signal representation $\mathbf{A}$ forms a tight frame with moderate redundancy (a factor of roughly 8 in two dimensions). Even though the energy is preserved, that is, $\|\mathbf{r}\| = \|\mathbf{C}\mathbf{r}\|$, the curvelet representation is overcomplete ($N \gg M$) and hence $\mathbf{C}\mathbf{C}^T$ is a projection, which makes it difficult to recover the sparsity vector $\mathbf{x}_0$ from $\mathbf{r} = \mathbf{C}^T\mathbf{x}_0$. The ℓ_1-norm linear programs presented earlier overcome this underdetermination problem leading to a stable recovery of the unknown $\mathbf{x}_0$, provided this vector is sufficiently sparse. Besides these properties, what makes curvelets the appropriate domain for seismic signal separation?

Curvelet properties

Curvelets are directional frames that represent a tiling of the 2-/3-D frequency plane into multiscale and multi-angular wedges (see Fig. 1). Because the directional sampling increases every-other scale doubling, curvelets become more and more anisotropic for finer and finer scales. They become 'needle-like' as illustrated in Fig. 1. Curvelets are localized in both domains and are oscillatory in one and smooth in the other direction. Even though curvelets are not of compact support (non-zero over a finite interval) in the physical domain, they are of rapid decay with an effective support given by ellipsoids parametrized by a width $\propto 2^{j/2}$, length $\propto 2^j$ and angle $\theta = 2\pi l 2^{\lfloor j/2 \rfloor}$ with j the scale and l the angular index with the number of angles doubling every other scale doubling (see Fig. 1). Curvelets are indexed by the multi-index $\mu := (j, l, \mathbf{k}) \in \mathcal{M}$ with $\mathcal{M}$ the multi-index set running over all scales, j, angles, l, and positions $\mathbf{k}$ (see for details Candès *et al.* 2006a; Ying *et al.* 2005).

By virtue of their anisotropic shape, curvelets are well adapted to detect wavefronts because locally aligned curvelets strongly correlate with wavefronts. This alignment leads to large curvelet coefficients and a concentration of the wavefield's energy in the transformed domain. As the examples included in Figs 2 and 3 suggest, this alignment property holds even in real-data situations, where the smoothness assumption along the wavefronts may be violated. In the marine setting, the lack of smoothness seems less of an issue because of the 'wavefront healing' that occurs as the wavefield travels through the water column.

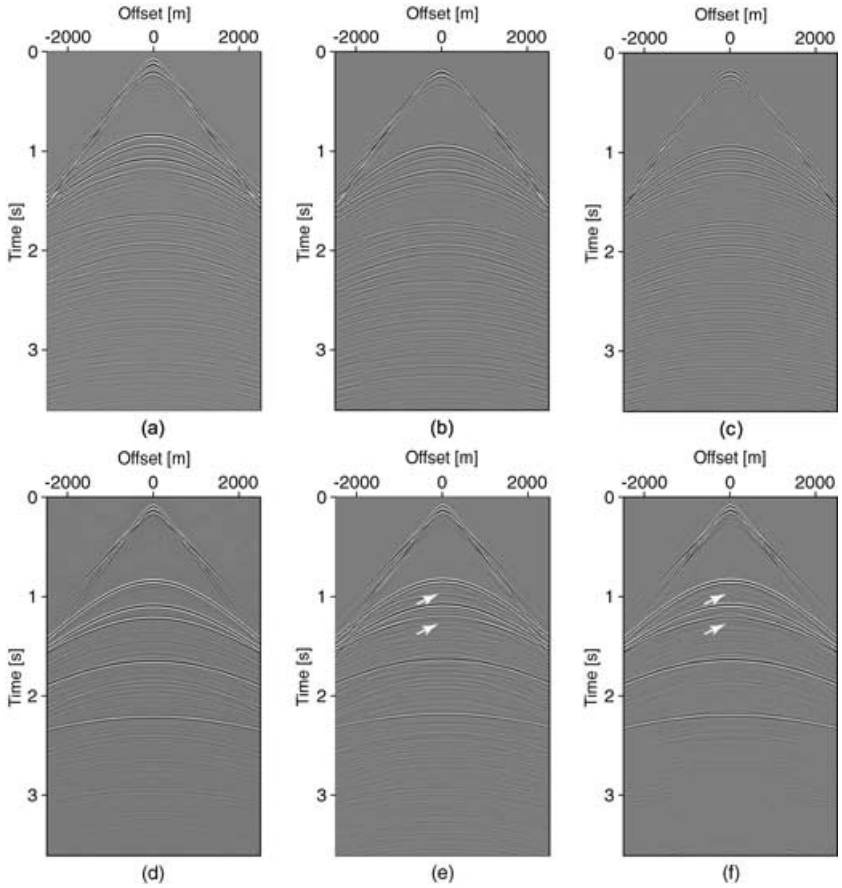

Figure 8. 2-D noise-free synthetic example, based on the medium plotted in Fig. 6. (a) Input data with all multiples. (b) True surface-related multiples. (c) SRME predicted multiples. (d) True data without surface-related multiples. (e) The predicted primaries with matched filtering. (f) The predicted primaries with curvelet-based separation. Note the significant improvement obtained with the curvelet-based method.

Non-linear approximation rates

The above construction accomplishes near optimal non-linear approximation rates for functions with wavefronts (see e.g. Candès & Donoho 2004; Hennenfent & Herrmann 2006). These compression rates measure the asymptotic decay of the ℓ_2-norm difference between the original data and a partial reconstruction from the largest coefficients. Theoretically, in two dimensions, Fourier only attains, besides log -like terms, a decay rate for the error with the largest K coefficients of $\mathcal{O}(K^{-1/2})$ for functions that are twice-differentiable except for wavefronts[6] situated along piecewise twice differentiable curves. For comparison curvelets obtain the near optimal rate.[7]

Plots for the empirical decay rates for the magnitude-sorted coefficients for the primaries, $\mathbf{s}_1$, multiples, $\mathbf{s}_2$ and total data set, $\mathbf{s}$ are included in Fig. 4. These plots clearly show that the curvelet coefficients decay the fastest amongst the different transforms. There is also not much difference between the decay rates for the different signal components for curvelets whereas there are observable differences amongst the signal components for the other transforms.

[6] Wavefronts correspond to singularities and are not differentiable.

[7] For singularities on 2-D surfaces in three dimensions these numbers are up to log factors $\mathcal{O}(K^{-1/3})$ for Fourier and $\mathcal{O}(K^{-1})$ for curvelets (Demanet 2005, personal communication) $\mathcal{O}(K^{-2})$.

Curvelets and waves

Because of their second dyadic partitioning (*cf.* Fig. 1)—the partitioning of the dyadic coronae into angular wedges (Stein 1993; Smith 1998; Candès & Demanet 2005; Candès *et al.* 2006a) and their parabolic scaling relation—curvelets are known to remain invariant under high-frequency asymptotic wave propagation. This invariance means that primaries as well as multiples are sparse. This property gives rise to a near dispersion-free propagation of curvelets and supports the claim that the covariance matrices for the two signal components are nearly diagonal in the curvelet domain. As shown earlier, this property underlies the weighting that produces the signal separation.

Comparison sparsity domains for primary-multiple separation

The following transforms in two dimensions are compared: the orthonormal Dirac, that is, $\mathbf{A}_1^T = \mathbf{A}_2^T = \mathbf{S} := \mathbf{Id}$; the orthonormal discrete Fourier transform, $\mathbf{S} := \mathbf{F}$; the orthonormal discrete wavelet transform, $\mathbf{S} := \mathbf{W}$, and the overcomplete FDCT with wrapping $\mathbf{S} := \mathbf{C}$.

Fig. 5 includes cross-plots for the normalized absolute values of the transform-domain coefficients of the true primaries and multiples pertaining to a 2-D synthetic data set. Not only does the curvelet parametrization by scale, location and angle(s) help to separate the signal components but the transform also

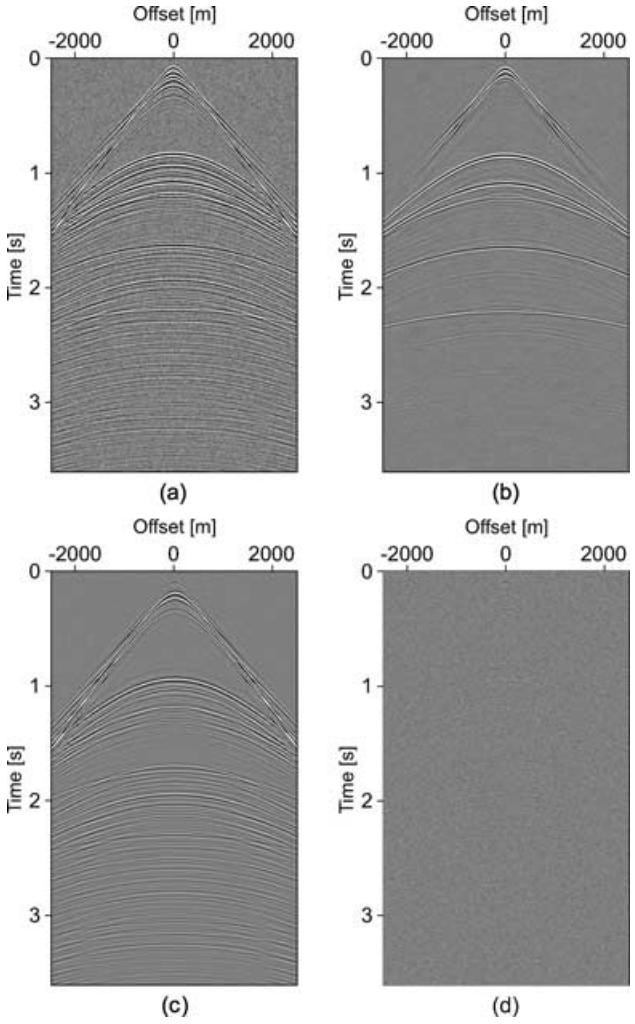

Figure 9. 2-D noisy synthetic example, based on the medium plotted in Fig. 6 and the predicted multiples plotted in Fig. 8. (a) Noisy input data with all multiples and signal-to-noise ratio of 23 dB. (b) Estimated primaries. (c) Estimated multiples. (d) The estimated noise according to the residue $\widehat{\mathbf{r}} = \mathbf{y} - \widehat{\mathbf{s}}$ with $\widehat{\mathbf{s}} = \widehat{\mathbf{s}}_1 + \widehat{\mathbf{s}}_2$. The example clearly demonstrates that primaries and multiples can be separated in the presence of Gaussian noise, provided there is a reasonable accurate prediction for the multiples.

decorrelates, that is, the cross-plots for the coefficients of the primaries and multiples are more concentrated along the axes for curvelets than for the other transforms as shown in Fig. 5. This property makes it easier to separate the coefficients through iterative thresholding.

Comparisions between the performance of the proposed separation method are summarized in Fig. 6 for a synthetic 2-D data example. These results are obtained by running the outer loop of the block-relaxation method for five decreasing Lagrange multipliers and with $L = 1$ for the inner loop (see Table 2). The decreasing Langrange multipliers are kept the same for the estimations based on the different transforms. The improved performance for the curvelet transform is already apparent in this example even though the algorithm was not run to convergence. Besides visible improvement, there is also an improved performance in the degree of decorrelation achieved by the algorithm, that is, the $R(\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2)$, and in the relative ℓ_2-norm difference between the estimated and true signal components. The results are listed in Table 3 and confirm the visible improvements of the results presented in Fig. 6.

COMPARISON CURVELET-BASED SIGNAL SEPARATION AND MATCHED FILTERING

In this section, we study the performance of our curvelet-based primary-multiple separation by means of a series of stylized examples. First, we study the case of a 1-D medium yielding strong correlations between the primaries and multiples where the prediction is accurate but where the separation is a challenge. The second example describes the situation where the prediction is inaccurate which corresponds to the situation where 2-D predicted multiples are used for data with 3-D structure. The next synthetic example is included to demonstrate our algorithm with the 3-D curvelet transform. We conclude by applying the method to a real data set with the 3-D curvelet transform.

Synthetic 2-D examples

First, we return to the data set presented earlier in Fig. 6, which concerns a 2-D shot record with surface-related multiples generated by a horizontally layered medium with seven layers as plotted in Fig. 7. Estimations for the primaries by matched filtering and curvelet-based separation are presented in Fig. 8 and compared with the true primaries. The primary signal component is estimated from the total data set plotted in Fig. 8(a) using the SRME prediction for the multiples as plotted in Fig. 8(c). For comparison, plots for the true multiples and primaries are also included in Figs 8(b) and (d). The results for the matched-filter and curvelet-based estimations are included in Figs 8(e) and (f).

This example can be considered as a worst-case scenario for the traditional wave-equation based prediction and adaptive subtraction methods, because the section is generated by a laterally invariant velocity model where at the near-offset primaries and multiples correlate in the physical domain. Indeed, these difficulties are reflected in the estimates obtained by matched filtering plotted in Fig. 8(e). In this matched-filter result, with filters estimated within overlapping time-offset windows, clear distortions in the estimated primaries can be observed as well as remnants from multiples, when compared to the true primaries plotted in Fig. 8(d). The arrows in Fig. 8(e) point to remaining multiple energy. Correlations between the primaries and multiples are apparently the main cause of these distortions. As can be observed from Fig. 8(f), the nonlinear curvelet-based separation result in Fig. 8(f) is much better and resembles the true primaries very well, in particular near the arrows.

We included Fig. 9, to illustrate the robustness of our separation method for noisy input data. The multiple prediction is the same as in Fig. 8 and the weighting vector is defined according to eq. (18). Estimates for the primaries and multiples from the noisy input data, Fig. 9(a) with signal-to-noise ratio of 23 dB, are plotted in Figs 9(b) and (c). The estimate for the incoherent noise, given by the residue $\widehat{\mathbf{r}} = \mathbf{y} - \widehat{\mathbf{s}}$ is included in Fig. 9(d). These results demonstrate the robustness of the separation method under Gaussian noise. Each signal component, the primaries, multiples and the incoherent noise, are recovered. The estimated noise contains very little coherent energy. Compared to the noise-free case, the degree of correlation $R(\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_1) = 0.069$ and the relative errors in the estimated primaries and multiples, $\varepsilon(\mathbf{s}_1) = 0.1836$ and $\varepsilon(\mathbf{s}_2) = 0.1788$. These errors are close to the values obtained for the noise-free example (see Table 3). The results for the relative errors are even better than for the noise-free case because the algorithm is run to convergence in this case.

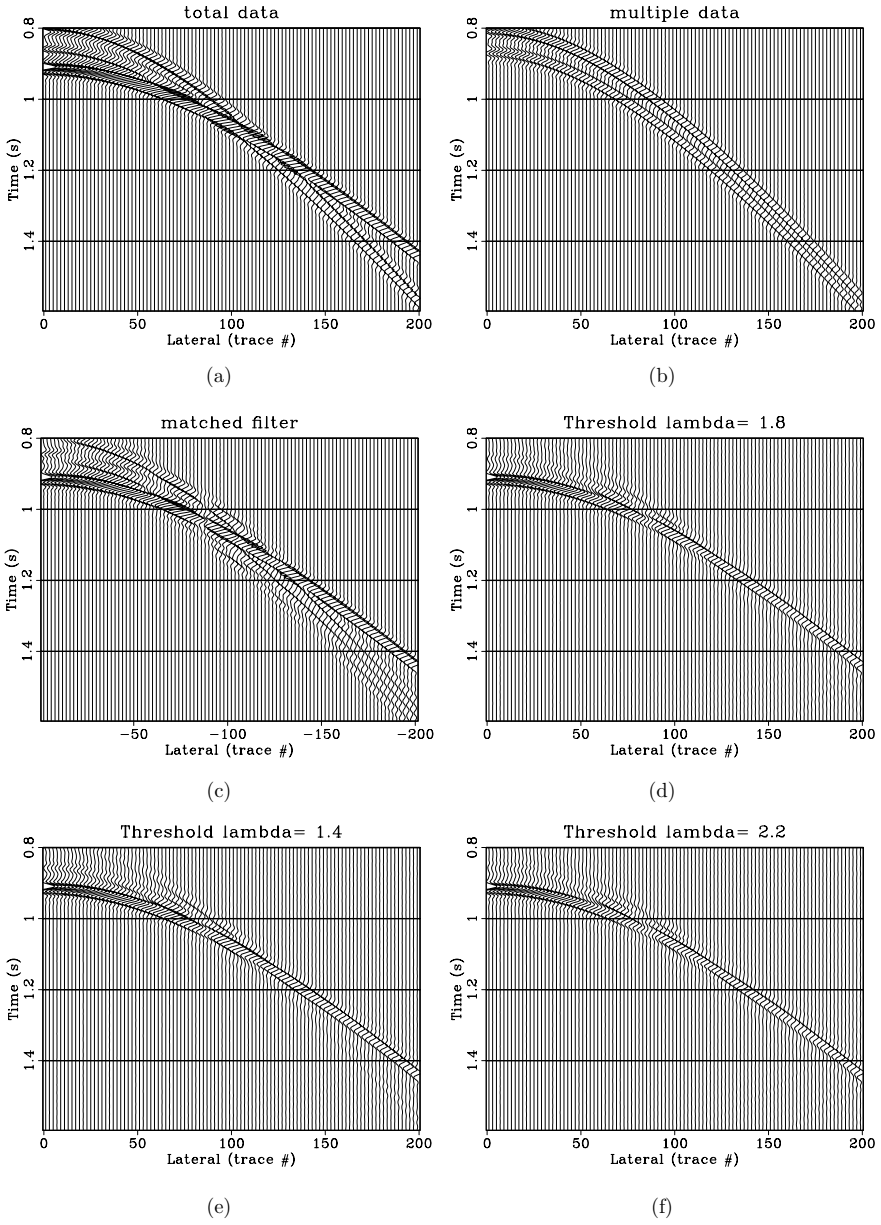

Figure 10. Example of primary-multiple separation through soft thresholding in the curvelet domain for predicted multiples with moveout errors. (a) The total data with primaries and multiples. (b) The predicted multiples containing moveout errors. (c) The result obtained with least-squares adaptive subtraction with localized windows. (d)–(f) The result obtained with a single curvelet-domain soft thresholding with $\lambda = 1.8, 1.4$ *and* 2.2.

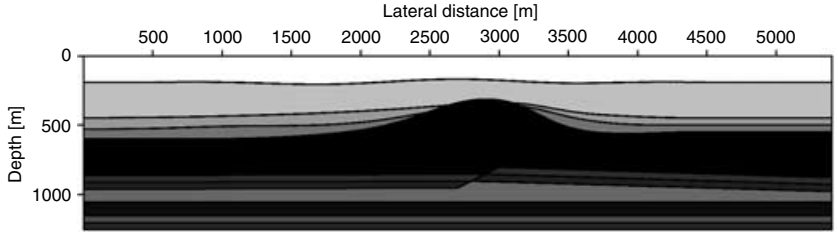

Figure 11. 2-D subsurface model used to generate synthetic shot records with an acoustic finite difference modelling algorithm.

To further illustrate the performance of curvelet-domain primary-multiple separation, we include an example where two multiples overlap with one primary (see Fig. 10a) and where the prediction for one of the multiple events has an erroneous moveout (see Fig. 10b). In that situation, the primary-multiple separation will be difficult and the result for the matched filtering with local windows leads to substantial residual multiple energy and dimming of the primary amplitudes (Fig. 10c). Soft thresholding (*cf.* eq. 11 with $\lambda = 1.4$),

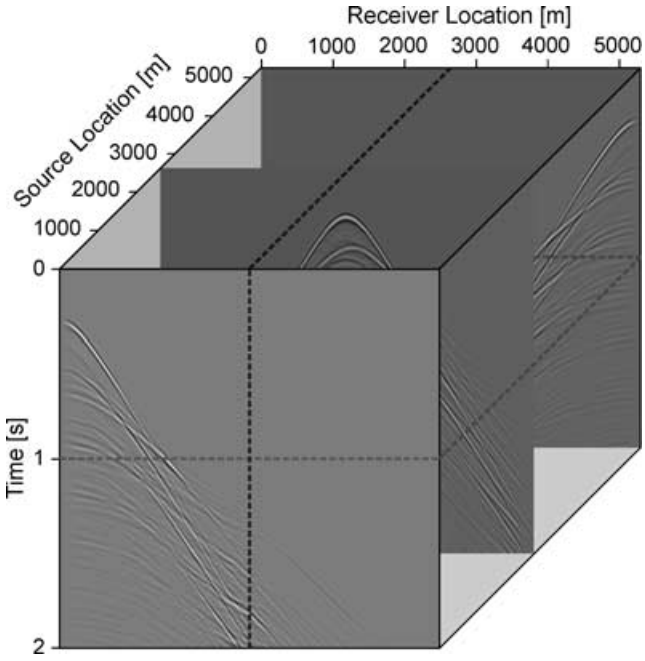

Figure 12. Shot records modelled in the subsurface model of Fig. 11. The data is generated with a fixed receiver spread of 361 receivers, resulting in a 361 × 361 cube of traces.

on the other hand, leads to a clear separation with moderate loss of primary energy (see Fig. 10d).

The 2-D synthetic examples discussed so far demonstrate the robustness of transformed-domain non-linear thresholding techniques to errors in the predicted multiples and noise. Not only is our method robust under moderate shifts and phase rotations, it also handles errors in the moveout of the predicted multiples. This behaviour is a direct consequence of the multiscale and multidirectional behaviour of the prototype waveforms that define the transform. In particular, the locality of the waveforms allows for an improved adaptation difficult to accomplish with matched filtering.

The examples also showed that we have control over the separation by varying the Lagrange multiplier (the threshold, see Figs 10e and f) and the expected noise level σ. Finally, there is also the possibility to add the estimated noise (the residue defined by the difference of the total data and the sum of the estimated primaries and multiples) after completion of the algorithm. This noise term is non-linear in the parameters and may contain spurious primary energy.

Synthetic 3-D example

The performance of our algorithm on 3-D data volumes is demonstrated for data generated by a subsurface velocity model with 2-D inhomogeneities as plotted in Fig. 11. This velocity model consists of a high-velocity layer, which represents salt, surrounded by sedimentary layers and a water bottom that is not completely flat. Using an acoustic finite-difference modelling algorithm, 361 shots with 361 receivers are simulated on a fixed receiver spread with receivers located from 0 to 5400 m with steps of 15 m. The complete pre-stack data set can be represented as a 3-D volume along the shot, receiver and time coordinates (see Fig. 12). From this data volume, surface-related multiples are predicted and subsequently removed with the iterative matched-filter method, as described by Verschuur & Berkhout (1997).

Comparison of the final result, obtained after three iterations, shows good suppression of the multiples in the shot records before and after matched filtering as shown Figs 13(a) and (b), respectively. These results, however, suffer from small residuals due to the property that least-squares subtraction always has the tendency to reduce the multiple energy towards the noise level. As a consequence small residuals remain in situations where the multiple prediction is not perfect. Because of limitations in the acquisition (e.g. limited aperture), predictions for the multiples will unfortunately always contain errors that lead to residual multiple energy.

Because the data volumes in shot, receiver and time coordinates contain wavefronts with continuity along all three coordinate directions, a separation algorithm is used which employs the 3-D curvelet transform. This 3-D transform has the distinct advantage that it (i) exploits the continuity of the wavefronts that make up the primary and multiple arrivals in all three coordinate directions; (ii) separates the signal components on the basis of scale, two angles and location in three dimensions and (iii) exploits the theoretical near optimality of the non-linear approximation rate of curvelets for functions that contain wavefronts along piecewise smooth sheets.

For data with imprints of lateral velocity variations, our separation approach will take advantage of differences in the 3-D

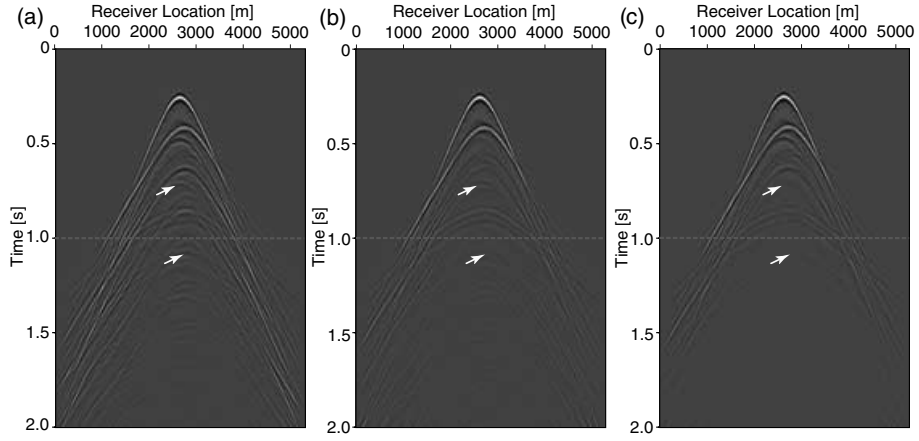

Figure 13. Comparison of curvelet-based separation and matched filtering for the middle shot record of the data set in Fig. 12. (a) Input shot with all multiples. (b) Result with adaptive subtraction. (c) Result with curvelet-based separation. The arrows point at internal multiples that should not be removed. The dashed line indicates the location of the time slice (Fig. 12).

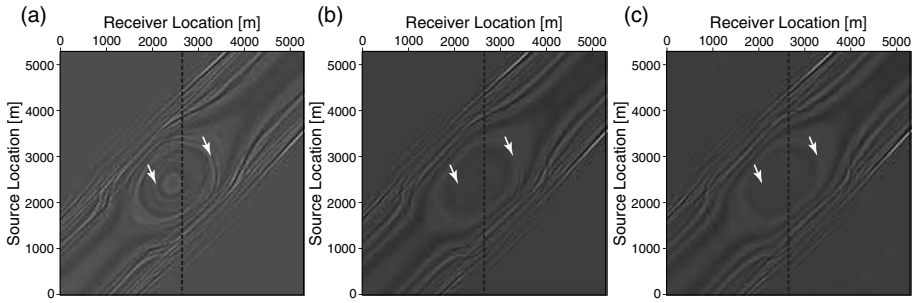

Figure 14. Comparison of curvelet-based separation and matched filtering for a time slice of the data set in Fig. 12. (a) Input shot with all multiples. (b) Result with adaptive subtraction. (c) Result with curvelet-based separation. The arrows point at (residual) multiples.

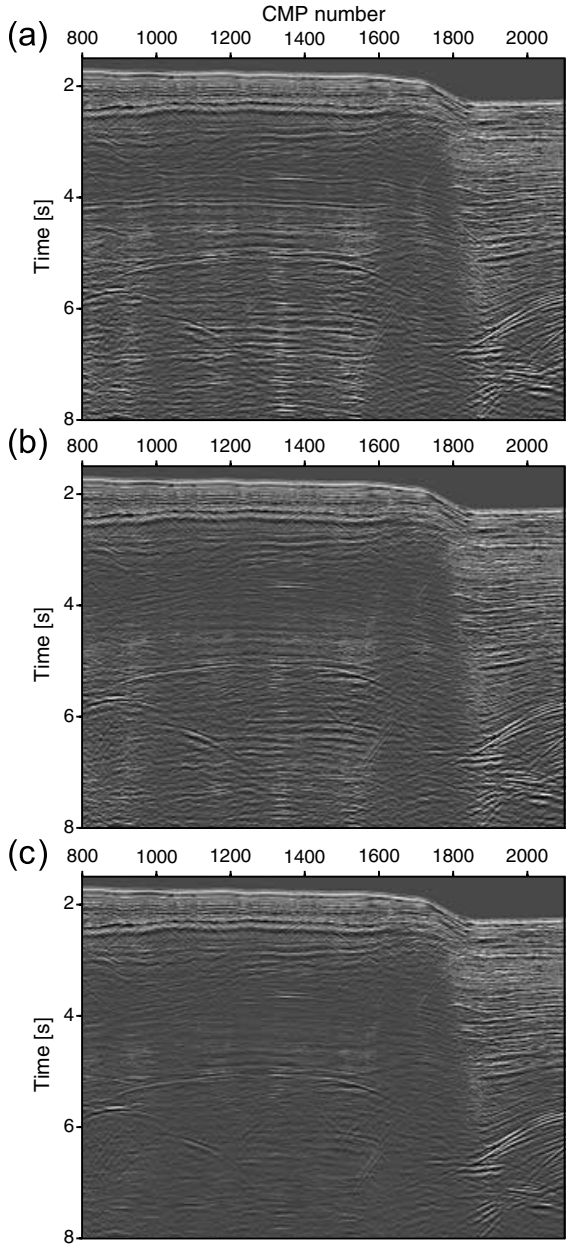

Figure 15. Field data from the gulf of Mexico. (a) Stack with multiples. (b) Stack after surface-related multiple suppression in the pre-stack domain. (c) Stack after curvelet-based primary-multiple separation. Note that the curvelet-based results show considerably less multiple energy, while retaining the primaries (including all diffractions around 6 s).

structure in the wavefields for the primaries and multiples. As the shot records plotted in Fig. 13(c) suggest, the curvelet-based result shows a better multiple suppression compared to the result obtained by matched filtering in Fig. 13(b). The estimated primaries are much cleaner, without a notable distortion of the primaries. Note that the internal multiples are preserved during the separation, as indicated by the arrows in Fig. 13. This preservation of the internal multiples is consistent because this component was not predicted and should not be removed at this stage. To appreciate the 3-D aspects of curvelet-based separation, time slices are included in Fig. 14 along the horizontal dashed line depicted in Fig. 12 and 13.

Real 3-D Gulf data example

To demonstrate the potential of the separation method for multiple suppression, a real-data example is included involving pre-stack field data from a 2-D line acquired at the Mississippi Canyon in the Gulf of Mexico. Results after stacking the data volumes with multiples, the stack after matched filtering and after curvelet-based separation are presented in Fig. 15. Surface-related multiples were estimated from this line which is represented as a 3-D data volume, that is, a shot-offset-time volume. These predicted multiples were used by the least-squares matched filtering technique in the shot domain as well as by our non-linear separation algorithm. Application of matched filtering to the pre-stack data results in an acceptable multiple suppression as can be seen in Fig. 15(b). However, these results are not optimal because of 3-D effects in the subsurface, which lead to erroneous predictions for the multiples utilizing a 2-D prediction algorithm. These imperfections result in residual multiples in the output section and consequently in the stack. Curvelet-based separation is less susceptible to these prediction errors and leads to a better multiple suppression result for the stack in Fig. 15(c). As opposed to matched filtering in the time domain, the performance of our algorithm derives from a direct exploitation of the full 3-D spatio-temporal character of the primary and multiple wavefields. This property lies at the heart of the notable improvement for the multiple suppression.

The improvements for the separation are also visible in the pre-stack domain as can be seen from Fig. 16, which includes the data volumes for the input data, predicted multiples and the output generated by matched filtering and curvelet-based separation. Again note in Fig. 16(d) the improved suppression of multiple energy and better recovery of primary reflections, especially in the area around shot 800–1000 below 4.5 s.

For a final comparison between the two methods, time slices at 4.7 s, as indicated by the dashed line in Fig. 16, are plotted in Fig. 17 for these four data sets before stack. Again, clutter can be observed in the subtraction results for matched filtering included in Fig. 17(c).

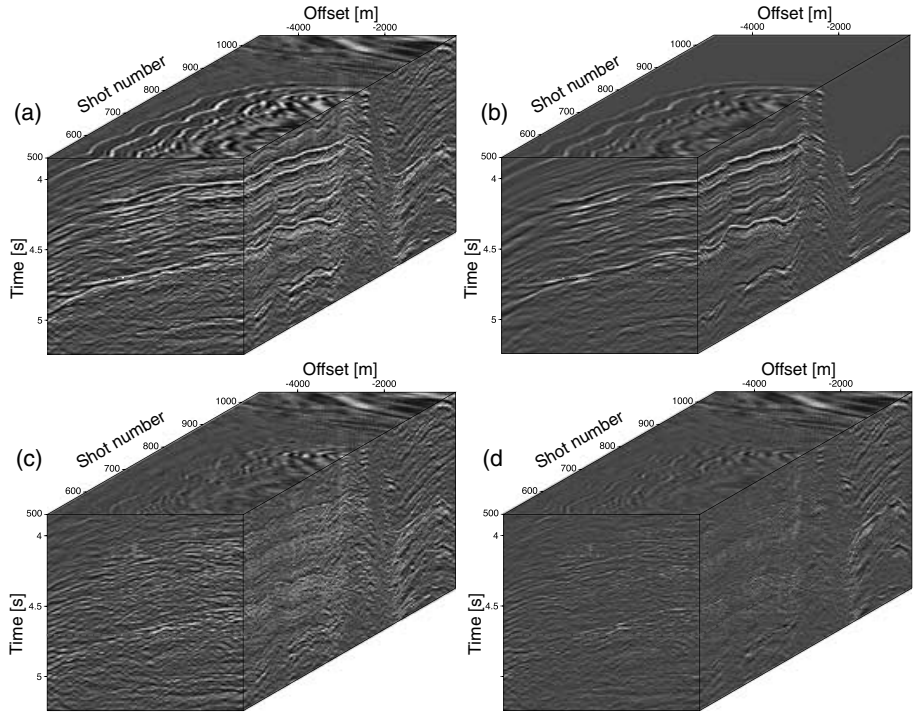

Figure 16. Comparison of adaptive subtraction and curvelet-based separation in the pre-stack domain. As the 2-D pre-stack data can be represented as a 3-D shot-offset-time volume, the curvelet-based separation can take advantage of all axes to discriminate between the primaries and the multiples. (a) Input data with multiples. (b) Predicted multiples. (c) Adaptive subtraction result. (d) Curvelet-based separation result. Note that the curvelet-based result has less residual noise, while retaining the primaries.

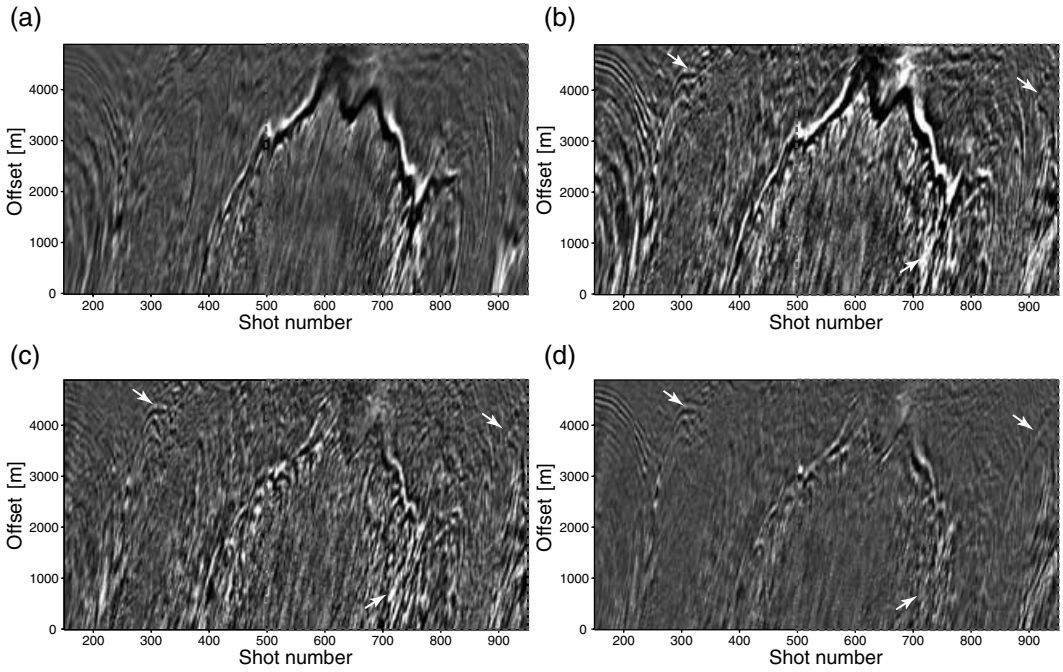

Figure 17. Time slices through the pre-stack volumes as shown in Fig. 16. The dashed line indicates the source-offset area displayed in Fig. 16.

Curvelet-based separation, on the other hand, is much cleaner as can be seen from Fig. 17(d). The arrows point at primaries that are recovered from multiple interference.

Even though the curvelet-based results appear cleaner, some dimming of primary energy may be observed in areas where the primaries and multiples exactly overlap in the curvelet domain. So far, our results were based on only one iteration of the block-relaxation method for 3-D curvelets and this dimming is expected. When the iterative procedure is carried out to convergence the results are expected to improve.

CONCLUSIONS

The success of separating coherent signal components with a generic non-parametric transform largely depends on the sparsity of the to-be-separated components in the transformed domain. We argued that curvelets, which decompose multidimensional data into multiscale and multidirectional prototype waveforms, are the appropriate domain for primary-multiple separation, given predictions for the multiples that may contain errors. We showed that primaries and multiples can be separated by solving a non-linear optimization problem during which the weighted ℓ_1-norms for the two signal components are jointly minimized. This weighted ℓ_1-norm optimization problem corresponds for orthonormal transforms to a simple element-wise thresholding with an 'oracle' defined in terms of the predicted signal component. Thresholding corresponds in this case, to Wiener shrinkage which is arguably optimal when the transform is optimally sparse. We showed that the curvelet transform obtains the best non-linear approximation rate for primaries as well as for multiples.

A simpler more intuitive explanation why the proposed algorithm works is that curvelets look like 'little' localized waves indexed by scale, orientation and location allowing for a localized separation based on differences in the spatial as well as frequency content of the two signal components. Because our method is non-adaptive, it differs fundamentally from matched filtering algorithms, which aims at separating the signal components by adapting the predicted signal components locally with respect to the true signal components. In our method there is no such adaptation and our algorithm derives its stability from the multiscale and multidimensional structure of curvelets.

Application of our algorithm to synthetic and real data examples demonstrates a notable improvement. Not only are the primaries and multiples better separated but our algorithm also proved to be insensitive to Gaussian noise and to errors in the prediction. This stability is a well-documented feature of transforms that are sparse and allows for a stable signal recovery from noisy and incomplete data.

The performance of the algorithm depends on the degree of correlation between the two to-be-separated signal components in the transformed domain. In cases where the two signal components overlap in the transformed domain a slight dimming of the reconstructed desired signal (i.e. the primaries) may occur. This dimming becomes less prominent when the algorithm is iterated sufficiently many times. In particular, the algorithm led to good results when using the 3-D curvelet transform. The separation can take advantage of the full 3-D structure of the primary and multiple wavefields. Since our method is generic, it can be used to separate arbitrary signal components as long as there is a reasonable prediction for one of the signal components and a transform that is sparse.

Furthermore, the method can handle non-stationary errors in the predicted multiples, for example, by neglecting out-of-plane events in multiple prediction. Whereas matched-filtering techniques have problems separating the signal components, curvelet-based separation is less sensitive to these errors and this leads to a notable improvement in both the pre-stack domain as well as after stacking.

ACKNOWLEDGMENTS

The authors would like to thank the authors of the Fast Discrete Curvelet Transform (FDCT) for making this code available at www.curvelet.org. This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and Collaborative Research and Development Grant DNOISE (334810-05) of Felix J. Herrmann and was carried out as part of the SINBAD project with support, secured through the Industry Technology Facilitator (ITF), from the following organizations: BG Group, BP, Chevron, ExxonMobil and Shell. The authors would also like to thank the Institute of Pure and Applied Mathematics at UCLA supported by the NSF under grant DMS-9810282. The authors thank WesternGeco for providing the field data. Finally, we would like to thank Laurent Demanet and an anonymous reviewer for their suggestions that helped to improve the paper.

REFERENCES

Berkhout, A.J. & Verschuur, D.J., 1997, Estimation of multiple scattering by iterative inversion, part I: theoretical considerations, *Geophysics,* **62,** 1586–1595.

Berryhill, J.R. & Kim, Y.C., 1986, Deep-water peg legs and multiples: emulation and suppression, *Geophysics,* **51,** 2177–2184.

Bruce, A.G., Sardy, S. & Tseng, P., 1998, Block coordinate relaxation methods for nonparametric signal de-noising, *Int. Soc. Opt. Eng.,* **3391,** 75–86.

Candès, E., Demanet, L., Donoho, D. & Ying, L., 2006a, Fast discrete curvelet transforms, *SIAM Multiscale Model. Simul.,* **5,** 861–899.

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

Candès, E.J. & Demanet, L., 2005, The curvelet representation of wave propagators is optimally sparse, *Comm. Pure Appl. Math,* **58,** 1472–1528.

Candès, E.J. & Donoho, D., 2004, New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities, *Comm. Pure Appl. Math.,* **57,** 219–266.

Candès, E.J. & Donoho, D.D., 2000a, Curvelets—a surprisingly effective nonadaptive representation for objects with edges. *Curves and Surfaces*, Vanderbilt University Press, Nashville.

Candès, E.J. & Donoho, D.D., 2000b, Recovering Edges in III-posed problems: optimality of curvelet frames, *Ann. Statist.,* **30,** 784–842.

Chen, J., Baysal, E. & Yilmaz, O., 2004, Weighted subtraction for diffracted multiple attenuation: 74th Ann. Internat. Mtg., SEG, Expanded Abstracts, 1329–1332, Soc. Expl. Geophys.

Chen, S.S., Donoho, D.D. & Saunders, M.S., 2001, Atomic decomposition by basis pursuit, *SIAM J. Scientif. Comput.,* **43,** 129–159.

Claerbout, J. & Muir, F., 1973, Robust modeling with erratic data, *Geophysics,* **38,** 826–844.

Coates, R.T. & Weglein, A.B., 1996, Internal multiple attenuation using inverse scattering: results from prestack 1 and 2-D acoustic and elastic synthetics: 66th Ann. Internat. Mtg., SEG, Expanded Abstracts, 1522–1525, Soc. Expl. Geophys.

Coifman, R.R. & Donoho, D.L., 1995, Translation-invariant de-noising, in *Wavelets and Statistics*, Springer-Verlag, New York, pp. 125–150.

Daubechies, I., Defrise, M. & de Mol, C., 2005, An iterative thresholding algorithm for linear inverse problems with a sparsity constrains, *Comm. Pure Appl. Math.,* **57,** 1413–1457.

Do, M. & Vetterli, M., 2002, Contourlets, in *Beyond Wavelets*, Elsevier Inc., Burlington.

Donoho, D., 1993, Unconditional bases are optimal bases for data compression and statistical estimation, *App l. Comput. Harmon. Anal.,* 100–115.

Donoho, D., Elad, M. & Temlyakov, V., 2006, Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Trans. Inform. Theory,* **52,** 6–18.

Donoho, D.L., 1995, De-noising by soft thresholding, *IEEE Trans. Inform. Theory,* **41,** 613–627.

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

Donoho, D.L. & Tsaig, Y., 2006, Extensions of compressed sensing. *Signal Processing,* **86**(3), 549–571.

Elad, M., 2006, Why simple shrinkage is still relevant for redundant representations, *IEEE Trans. Inform. Theory,* **52**(2), 5559–5569.
Elad, M., Starck, J., Querre, P. & Donoho, D., 2005, Simulataneous cartoon and texture image inpainting using morphological component analysis (MCA), *Appl. Comput. Harmon. Anal.,* **19,** 340–358.
Figueiredo, M. & Nowak, R., 2003, An EM algorithm for wavelet-based image restoration, *IEEE Trans. Image Proc.,* **12,** 906–916.
Fokkema, J.T. & van den Berg, P.M., 1993, *Seismic Applications of Acoustic Reciprocity*, Elsevier, Amsterdam.
Guitton, A. & Verschuur, D.V., 2004, Adaptive subtraction of multiples using the l^1-norm, *Geophys Prospect,* **52,** 27–27.
Hennenfent, G. & Herrmann, F.H., 2006, Seismic denoising with nonuniformly sampled curvelets, *Comp. Sci. &Eng.,* **8,** 16–25.
Herrmann, F.J. & Verschuur, D.V., 2005, Robust curvelet-domain primary-multiple separation with sparseness constraints. *Presented at the, EAGE., 67th Conference & Exhibition Proceedings,* pp. 226.
Herrmann, F.J. & Verschuur, E., 2004, Curvelet-domain multiple elimination with sparseness constraints. *Presented at the 74rd Annual Internat. Mtg.*, SEG, Soc. Expl. Geophys., Expanded abstracts.
Ikelle, L., Roberts, G. & Weglein, A., 1997, Source signature estimation based on the removal of first-order multiples, *Geophysics,* **62,** 1904–1920.
Ikelle, L.T. & Yoo, S., 2000, An analysis of 2D and 3D inverse scattering multiple attenuation: 70th Ann. Internat. Mtg., SEG, Expanded Abstracts, 1973–1976, Soc. Expl. Geophys.
Kabir, M.M. N., 2003, Weighted subtraction for diffracted multiple attenuation: 73rd Ann. Internat. Mtg., SEG, Expanded Abstracts, 1941–1944, Soc. Expl. Geophys.
Lee, N.-Y. & Lucier, B.L., 2001, Wavelet methods for inverting the Radon transform with noisy data, *IEEE Trans. Image Proc.,* **10,** 79–94.
Levy, S., Oldenburg, D. & Wang, J., 1988, Subsurface imaging using magnetotelluric data, *Geophysics,* **53,** 104–117.
Lokshtanov, D., 1999, Multiple suppression by data-consisten deconvolution, *The Leading Edge,* **18,** 115–119.
Mallat, S.G., 1997, *A Wavelet Tour of Signal Processing*, Elsevier Inc., Burlington.
Nemeth, T. & Bube, K., 2001, An operator decomposition approach for the separation of signal and coherent noise in seismic wavefields, *Inverse problems,* **17,** 533–551.
Oldenburg, D.W., Levy, S. & Whittall, K.W., 1981, Wavelet estimation and deconvolution, *Geophysics,* **46,** 1528–1542.
Ross, W.S., 1997, Multiple suppression: beyond 2-D. part I: theory: 67th Ann. Internat. Mtg., Expanded Abstracts, 1387–1390, Soc. Expl. Geophys.
Ross, W.S., Yu, Y. & Gasparotto, F.G., 1997, Multiple suppression: beyond 2-D. part II: application to subsalt multiples: 67th Ann. Internat. Mtg., Expanded Abstracts, 1391–1394, Soc. Expl. Geophys.
Sacchi, M. & Ulrych, T., 1996, Estimation of the discrete Fourier transform, a linear inversion approach, *Geophysics,* **61,** 1128–1136.
Sacchi, M.D., Velis, D.V. & Cominguez, A.C., 1994, Minimum entropy deconvolution with frequency-domain constraints, *Geophysics,* **59,** 938–945.
Smith, H., 1998, A Hardy space for Fourier integral operators, *J. Geom. Anal.,* **8**(4), 629–653.
Starck, J.L., Elad, M. & Donoho, D., 2004, Redundant multiscale transforms and their application to morphological component separation, *Adv. Imaging Electron Phys.,* **132,** 288–348.
Stein, E., 1993, Harmonic analysis: real-variable methods, in *Orthogonalityand Oscillatory Integrals*, Princeton University Press, New Jersey.
Trad, D., 2001, Implementations and applications of the sparse Radon transform, *PhD thesis*, University of British Columbia.
Trad, D., Ulrych, T. & Sacchi, M., 2003, Latest views of the sparse radon transform, *Geophysics,* **68,** 386–399.
Tropp, T., 2006, Just relax: convex programming methods for subset selection and sparse approximation, *IEEE Trans. Inform. Theory.,* **52**(3), 1030–1051.
Ulrych, T.J. & Walker, C., 1982, Analytic minimum entropy deconvolution, *Geophysics,* **47,** 1295–1302.
Verschuur, D.J., 2006, Seismic Multiple Removal Techniques: Past, Present and Future, EAGE publications b.v., Houten, the Netherlands.
Verschuur, D.J. & Berkhout, A.J., 1997, Estimation of multiple scattering by iterative inversion, part II: practical aspects and examples, *Geophysics,* **62,** 1596–1611.
Verschuur, D.J., Berkhout, A.B. & Wapenaar, C.P.A., 1992, Adaptive surface-related multiple elimination, *Geophysics,* **57,** 1166–1177.
Vogel, C., 2002, *Computational Methods for Inverse Problems*, SIAM, Philadelphia.
Wang, Y., 2003, Multiple subtraction using an expanded multichannel matching filter, *Geophysics,* **68,** 346–354.
Weglein, A.B., Carvalho, F.C. & Stolt, P.S., 1997, An iverse scattering series method for attenuating multiples in seismic reflection data, *Geophysics,* **62,** 1975–1989.
Wiggins, J.W., 1988, Attenuation of complex water-bottom multiples by wave-equation-based prediction and subtraction, *Geophysics,* **53,** 1527–1539.
Ying, L., Demanet, L. & Candés, E., 2005, 3D discrete curvelet transform, 591 413, SPIE.
Zwartjes, P. & Gisolf, A., 2006, Fourier reconstruction of marine-streamer data in four spatial coordinates, *Geophysics,* **71,** 171–186.

APPENDIX A:

Curvelet properties

Curvelets correspond to a partitioning of the 2- or 3-D Fourier plane by highly anisotropic elements that obey a parabolic scaling principle (Smith 1997; Candès & Donoho 2000b; Do & Vetterli 2002; Ying *et al.* 2005; Candès *et al.* 2006a) width $\propto$ length^2. In this appendix, we limit ourselves to discussing the 2-D curvelet transform only. Compared to ordinary separable [8] discrete wavelets, which have location, scale and gender indices, curvelets have indices that discretize the scale, a, $0 < a < 1$; orientation θ, $\theta \in [-\pi/2, \pi/2]$ and location $\mathrm{b} \in \mathbb{R}^2$. Consequently, discrete curvelets represent a family of directional prototype waveforms that are made out of a combination of translations, rotations and parabolic scalings. These three operations take a directional wavelet, $\varphi(\mathrm{x})$, which contains a bump in one direction and wavelet-like [9] oscillations in the other, to a three-index family

$$\varphi_\mu(\mathrm{x}) \approx \varphi\left(\mathrm{D}_a \mathrm{R}_\theta\left[\mathrm{x} - \mathrm{b}\right]\right). \tag{A1}$$

The index-set $\mu = \{j, l, \mathbf{k}\}$ rules the discretization of the scale, $a = 2^{-j}$, scale-dependent orientation, $\theta_{j,l} = 2\pi l 2^{\lfloor j/2 \rfloor}$ and location, $\mathbf{R}_{\theta_{j,l}} \mathrm{b}_\mathrm{k}^{j,l}$ with $\mathbf{R}_\theta$ denoting a rotation over θ radians and $\mathbf{D}_a$ a parabolic scaling yielding,

$$\varphi_a(\mathrm{x}) \approx \varphi\left(\mathbf{D}_a \mathrm{x}\right) \quad \text{with} \quad \mathbf{D}_a = \begin{pmatrix} 1/a & 0 \\ 0 & 1/\sqrt{a} \end{pmatrix}. \tag{A2}$$

With this sampling of the 2-D frequency plane, also known as a second dyadic partitioning (Stein 1993; Smith 1997), a redundant tight frame is created that is

(i) *Multiscale* with elements living in different dyadic corona in the 2-D Fourier plane, that is, $\mathrm{k} \in [2^j, 2^{j+1}]$ with j the dyadic scale and k the wave number (dual of the space variable x).

(ii) *Multidirectional* with elements living on wedges oriented according $\theta = \pi l 2^{-j/2}$ with $l = 0, 1, 2, \ldots, 2^{\lfloor j/2 \rfloor} - 1$ the index for

[8] Separable transforms are multidimensional transforms that are made off 1-D transforms along each coordinate direction independently. See for example, Mallat (1997).
[9] Curvelets are like Meyer wavelets with infinite vanishing moments in the direction of the oscillations.

the angles. Within each corona the number of orientations doubles *every other scale*, yielding orientations $\propto \frac{1}{\sqrt{\text{scale}}}$. See Fig. 1.

(iii) *Highly anisotropic*, obeying the following scaling law width $\propto \text{length}^2$ with width $\propto 2^{-j/2}$ and length $\propto 2^{-j}$.

(iv) *Strictly localized* in the Fourier domain with each curvelet located in the symmetric wedge

$$W_{j,l} = \left\{\pm \mathrm{k},\ 2^j \leq |\mathrm{k}| \leq 2^{(j+1)},\ |\theta - \theta_j| \leq \pi l 2^{-\lfloor j/2 \rfloor}\right\}. \tag{A3}$$

(v) *Localized (rapid decay) and smooth in space.*

and for which

(i) *Fast* $\mathcal{O}(N \log N)$ *algorithms* (see Candès *et al.* 2006a; Ying *et al.* 2005, for detail on the numerical implementation of the 2- and 3-D Fast Curvelet Transform, the FDCT) exist that decompose 2- or 3-D sampled images which are lexicographically stored in a M-sample vector $\mathbf{s}$, that is,

$$\mathbf{x} = \{\langle \mathbf{s},\ \varphi_\mu \rangle\}_{\mu \in \mathcal{M}} := \mathbf{C}\mathbf{s}, \tag{A4}$$

with the brackets $\langle \cdot, \cdot \rangle$ denoting the discrete inner product. In this expression, $\mathbf{C} \in \mathbb{R}^{N \times M}$ is a rectangular matrix with $N \gg M$. This matrix represents the implementation of the FDCT with curvelets on its rows.

(ii) An explicit construction exits for the adjoint that equals the pseudoinverse, that is, $\mathbf{C}^T = \mathbf{C}^\dagger$, yielding the following composition

$$\mathbf{s} = \sum_{\mu \in \mathcal{M}} \langle \mathbf{s},\ \varphi_\mu \rangle \varphi_\mu := \mathbf{C}^T \mathbf{C}\mathbf{s}. \tag{A5}$$

Because the frame is numerically tight there is also conservation of the energy, that is, $\|\mathbf{s}\|_2^2 = \|\mathbf{x}\|_2^2$.

Theoretical non-linear approximation rates for curvelet frames

Optimality of a signal representation refers to accomplishing (close to) optimal non-linear approximation rates for a certain class of functions. The non-linear approximation error is given by the energy difference between the original function and its reconstruction using the largest K entries in the sorted sparsity vector. The faster this error decays as a function of the number of largest entries, the higher the non-linear approximation rate.

Non-linear approximation rates measure the asymptotic decay of the L_2-norm difference between the original function and the partial reconstruction from the K largest coefficients. For example, wavelets obtain optimal non-linear approximation rates for models that are represented by 1-D Bounded Variation functions, that is, $f \in \mathbf{BV}[0, 1]$ if $\|f\|_{BV} := \int |f'|(t)|dt < \infty$. For these functions, the K-term approximation

$$f_K^W = \mathbf{W}^T \mathbf{x}_{I_K} \tag{A6}$$

decays as

$$\left\| f - f_K^W \right\|_2^2 \propto K^{-2} \tag{A7}$$

as opposed to a decay of K^{-1} for linear approximations based on the Fourier transform (see e.g. Mallat 1997). In this equation, $\mathbf{x} = \mathbf{W}\, f$, represents a vector with the wavelet coefficients of the function f, I_K is the index set of the first K largest coefficients ($\{I_K : x_{I(1)} \geq x_{I(2)} \geq \cdots \geq x_{I(K)}\}$) and $\mathbf{W}^T$ is the transpose $=$ inverse of the orthonormal discrete wavelet transform. For the above function space, this decay rate proves to be optimal.

For 2-D functions with singularities on piece-wise $\mathbb{C}^2$ curves, the non-linear approximation rates are sub optimal for both Fourier and separable wavelets (Candès & Donoho 2000a,b),

$$\left\| f - f_K^F \right\|_2^2 \propto K^{-\frac{1}{2}} \tag{A8}$$

and

$$\left\| f - f_K^W \right\|_2^2 \propto K^{-1}, \tag{A9}$$

compared to the rate obtained by adaptive triangulation

$$\left\| f - f_K^O \right\|_2^2 \propto K^{-2}. \tag{A10}$$

Recently introduced curvelets (Candès & Donoho 2000a,b) obtain

$$\left\| f - f_K^C \right\|_2^2 \propto C K^{-2} (\log K)^3, \tag{A11}$$

with C a constant. This rate is close to the above optimal rate. This rate for the infinite dimensional case, f is a continuous-valued function, is expected to carry over to the finite-dimensional case for sampled functions. Indeed, the empirical non-linear approximation rates for seismic data suggest that curvelets are optimal for data with multidimensional wave fronts.