GEOPHYSICS, VOL. 68, NO. 1 (JANUARY-FEBRUARY 2003); P. 386–399, 9 FIGS.
10.1190/1.1543224

Latest views of the sparse Radon transform

Daniel Trad*, Tadeusz Ulrych*, and Mauricio Sacchi‡

ABSTRACT

The Radon transform (RT) suffers from the typical problems of loss of resolution and aliasing that arise as a consequence of incomplete information, including limited aperture and discretization. Sparseness in the Radon domain is a valid and useful criterion for supplying this missing information, equivalent somehow to assuming smooth amplitude variation in the transition between known and unknown (missing) data. Applying this constraint while honoring the data can become a serious challenge for routine seismic processing because of the very limited processing time available, in general, per common midpoint. To develop methods that are robust, easy to use and flexible to adapt to different problems we have to pay attention to a variety of algorithms, operator design, and estimation of the hyperparameters that are responsible for the regularization of the solution.

In this paper, we discuss fast implementations for several varieties of RT in the time and frequency domains. An iterative conjugate gradient algorithm with fast Fourier transform multiplication is used in all cases. To preserve the important property of iterative subspace methods of regularizing the solution by the number of iterations, the model weights are incorporated into the operators. This turns out to be of particular importance, and it can be understood in terms of the singular vectors of the weighted transform. The iterative algorithm is stopped according to a general cross validation criterion for subspaces. We apply this idea to several known implementations and compare results in order to better understand differences between, and merits of, these algorithms.

INTRODUCTION

In contrast to other common transformations, like the Fourier and orthogonal wavelet transforms, the Radon transform (RT) operator is not orthogonal. As a consequence, the problem of applying the forward and inverse transform without loss of data is not trivial. Many different methods have been developed for obtaining the RT, but the most commonly used is that of inversion. The misfit between observed and predicted data is minimized subject to a constraint of obtaining the smallest model in a least-squares sense, an approach usually leading to what is known as zero-order regularization. This procedure constitutes the standard (nonsparse) RT, often used with linear or parabolic basis functions (Hampson, 1986; Beylkin, 1987).

The decrease in resolution as a consequence of limited aperture is somewhat attenuated by the use of stochastic inversion in place of zero-order regularization, originally proposed by Thorson and Claerbout (1985), leading to the high-resolution time-domain RT. Sacchi and Ulrych (1995) implemented the sparse RT in the frequency domain by means of Bayes rule using a Cauchy form probability-density function (pdf), which is now generally used in seismic processing. Since then, some other variants of the sparse RT have been developed. Cary (1998) noted the superiority of the time domain algorithms because of their ability to enforce simultaneously time and Radon parameter sparseness. Another important advantage is the ability of time domain methods to deal with time variant problems in the data and model by using time variant model and data weights. Cary also proposed the use of frequency domain operators in the time domain algorithms to obtain a better waveform conservation. K. Marfurt and T. Nemeth (unpublished work), following previous work from Stoffa et al. (1981) and Yilmaz and Taner (1994), noted the capability of time domain algorithms for the sparse RT to deal with aliasing. Herrmann et al. (1999) proposed the so-called dealiased RT, which prevents aliasing by carrying model weights from

Manuscript received by the Editor November 15, 2001; revised manuscript received March 27, 2002.
*University of British Columbia, Department of Earth and Ocean Sciences, 2219 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada. E-mail: dtrad@geop.ubc.ca; ulrych@eos.ubc.ca
‡University of Alberta, Institute for Geophysical Research, Department of Physics, Edmonton, Alberta T6G 2J1, Canada. E-mail: sacchi@phys.ualberta.ca.

low (nonaliased) frequencies to high (potentially aliased) frequencies. This algorithm also improves efficiency by limiting iterations. All these papers showed that not only resolution but also aliasing problems can be attenuated to some degree by the use of sparseness criteria.

Enforcing sparseness brings new problems and new possibilities. The problems are the increase in computation time, possibly the introduction of artifacts and the difficulty to set inversion parameters. The possibilities arise because a sparse RT can be used for other purposes that are not possible to implement with the standard smooth RT, for example, improving continuity and separation of events and removing some artifacts like aliasing that require a high degree of dispersion in the model space.

In spite of the fact that the sparse RT is superior to the standard RT in the sense that it represents a generalization, sometimes the sparse RT is difficult to implement because of its sensitivity to the hyperparameters required to regularize the inversion. The problem is more serious in the frequency domain, where one hyperparameter per frequency has to be automatically chosen. If the hyperparameter for one single frequency is incorrect, the results are affected in the whole Radon domain.

In this paper, we have applied to the RT a series of concepts and ideas often used in optimization. We first give an interpretation of sparseness in terms of tailoring the model space basis functions to the expected model. Then we show how to decrease the sensitivity of the RT to the hyperparameters, while still achieving sparseness in the solution. We do this by incorporating the model weights into the operators and using an automatic stopping criterion in the iterative algorithms as a regularizer. The solution is then built by a limited number of basis functions in the model space, but it is still sparse because these basis functions are tailored for a sparse representation of the desired solution. We present applications and examples of this implementation for different known RT algorithms in frequency, time, and time-frequency domains, as well as some other problems and solutions that can be addressed by operator design and sparseness.

SPARSE INVERSION

Given any transformation $\mathbf{m}$ on some data $\mathbf{d}$ [for example the RT, $\mathbf{m} = RT(\mathbf{d})$], the problem can be cast as the inversion for the model that generates the data under the action of the operator $\mathbf{L}$ (Claerbout, 1992):

$$\mathbf{d} = \mathbf{L}\mathbf{m}. \tag{1}$$

This model does not need to be physical. It simply represents the result of the transform applied on the data. Casting the problem in this manner, the transform can be found by applying the vast arsenal of tools available from inverse theory. To invert operators, we define a cost or objective function, which is a mathematical expression that measures the undesired characteristics of the model. There are many possibilities, but most often we pursue a model that honors the data and has a minimum of information not required by the data. This statement of goals is commonly presented as

$$\begin{aligned} &\text{minimize } \|\mathbf{W}_m\mathbf{m}\|_p^p \\ &\text{subject to } \|\mathbf{W}_d(\mathbf{d} - \mathbf{L}\mathbf{m})\|_q^q = \phi_d, \end{aligned} \tag{2}$$

where ϕ_d is some estimate of the noise level in the data plus a residual due to the failure of the proposed model to explain the data. $\mathbf{W}_d$ is a matrix of data weights, often a diagonal matrix containing the inverse of the standard deviation of the data, and $\mathbf{W}_m$ is a matrix of model weights that we can design in order to enhance our preference regarding the model, for example resolution or smoothness. In equation (2), p and q indicate that different norms can be applied to measure the norm of vectors. A sparse model can be obtained by simply choosing a norm that does not penalize very strongly large elements contained by the model. Conversely, a norm that penalizes large elements leads to smooth models. These model and data weights can be related to model and data covariances matrices by means of a Bayes formulation.

A common approach for obtaining a sparse transform is to choose a ℓ_1 norm for the model and a ℓ_2 norm for the data misfit. This mixed norm problem can be easily transformed to a $\ell_2 - \ell_2$ problem by using model dependent model weight matrices with elements proportional to

$$[\mathbf{W}_m]_{ii} = \frac{1}{\sqrt{m_i}}, \tag{3}$$

since

$$\|\mathbf{m}\|_1^1 = \sum_i |m_i| = \mathbf{m}^T\mathbf{W}_m^T\mathbf{W}_m\mathbf{m} = \|\mathbf{W}_m\mathbf{m}\|_2^2. \tag{4}$$

To avoid division by zero, a minimum threshold has to be chosen for $\mathbf{W}_m$. This threshold constitutes the first hyperparameter and depends on the size of the model. As a robust measure of the size of $\mathbf{m}$, we have used with some success a quantile of the distribution. The p quantile of $\mathbf{m}$ is the value of $\mathbf{m}$ where its cumulative distribution takes the value p. It can be obtained simply by sorting the data and then taking the value located at the index $p \times N$, where N is the number of elements [see for example Lupton (1993)]. The smaller the quantile, the sparser the transform. Hence, this number defines the trade-off between sparseness and smoothness. The same order of quantile can be used for all frequencies. The actual values will be different, but all of them represent the same relative size of the model.

We can now set the misfit and undesired characteristics of the solution in a cost function and, by minimizing it, obtain the model that better approximates the desired solution. Thus, the model can be found by solving the system of equations

$$\left(\lambda\mathbf{W}_m^T\mathbf{W}_m + \mathbf{L}^T\mathbf{W}_d^T\mathbf{W}_d\mathbf{L}\right)\mathbf{m} = \mathbf{L}^T\mathbf{W}_d^T\mathbf{W}_d\mathbf{d}, \tag{5}$$

where λ is the second trade-off hyperparameter that will allow a different weight to be assigned to the misfit and model constraints. Most variants of the sparse RT share the same basic system of nonlinear equations (5). Among the many different methods of solution, the one most often used because of its simplicity and efficiency is the iteratively re-weighted least squares (IRLS) algorithm (Scales et al, 1988). In this method the nonlinear system of equations is solved iteratively by fixing the model weights to some previous estimation and applying a linear minimization at every iteration (referred to from now on as external iterations). The purpose of these external iterations is to update the model weights, by using an approximated solution to the problem in hand. Because the model weights are fixed for every iteration, a simple algorithm for solving linear

problems can be applied. A common choice is some subspace method, in general a conjugate gradient (CG) algorithm. In this case, there are also iterations for the linear solver, referred to as internal iterations.

By applying a right preconditioning, the modeling equation (1) becomes

$$\mathbf{d} = \mathbf{L}\mathbf{W}_m^{-1}\mathbf{W}_m\mathbf{m}, \tag{6}$$

and the optimization problem (2) is now

$$\begin{aligned} &\text{minimize } \|\tilde{\mathbf{m}}\|_p^p \\ &\text{subject to } \left\|\mathbf{W}_d\left(\mathbf{d} - \mathbf{L}\mathbf{W}_m^{-1}\tilde{\mathbf{m}}\right)\right\|_q^q = \phi_d, \end{aligned} \tag{7}$$

where $\tilde{\mathbf{m}} = \mathbf{W}_m\mathbf{m}$. The minimization of the cost (7) produces the following system of equations:

$$\left(\lambda\mathbf{I} + \mathbf{W}_m^{-T}\mathbf{L}^T\mathbf{W}_d^T\mathbf{W}_d\mathbf{L}\mathbf{W}_m^{-1}\right)\tilde{\mathbf{m}} = \mathbf{W}_m^{-T}\mathbf{L}^T\mathbf{W}_d^T\mathbf{W}_d\mathbf{d}. \tag{8}$$

Hence, the effect of the right preconditioning is to set the model weights as part of the modeling rather than a penalizing factor in the cost function.

The system (8) can be solved very efficiently by setting $\lambda = 0$ (no regularization) and letting the number of internal iterations in the CG algorithm play the role of regularizer. The system of equations is only partially solved at every external iteration, because the conjugate gradient algorithm is stopped before the solution is complete.

Note that the hyperparameter λ is no longer used, but a good stopping criterion for the CG algorithm is required instead. The stopping criterion for the external iterations is less important because in real data processing the minimum of the cost function is never achieved and generally only a few (2–4) iterations are applied, with reasonable results. Hence, we will discuss only the stopping criterion for the CG algorithm. A stopping criteria based on generalized cross validation (GCV) for subspaces (Haber, 1997) has proven to be very efficient, giving excellent stability to the CG algorithm. The GCV function measures the dependence of the solution on local information (for example, one data point) rather than global information. Basically, the GCV function is a weighted sum of the data misfits obtained by solving the problem with successive elimination of data points (one point eliminated at a time). The iterative algorithm is stopped when the GCV function reaches a minimum. The GCV function for the solution of $\mathbf{d} = \mathbf{L}\mathbf{W}_m^{-1}\mathbf{W}_m\mathbf{m}$ can be computed as

$$GCV(\lambda) = \frac{\sum_{k=1}^{N}\left[d(\lambda)_k - d_k^{obs}\right]^2}{\sum_{k=1}^{N}(1 - C_{k,k})^2}, \tag{9}$$

where $C_{k,k}$ are the diagonal elements of the resolution matrix,

$$\mathbf{C} = \mathbf{C}(\lambda) = \tilde{\mathbf{L}}(\tilde{\mathbf{L}}^T\tilde{\mathbf{L}} + \lambda\mathbf{I})\tilde{\mathbf{L}}^T, \tag{10}$$

where $\tilde{\mathbf{L}} = \mathbf{L}\mathbf{W}_m^{-1}$ and λ is the regularization parameter that controls the trade-off between the achieved misfit and the minimization of the model norm. When the number of iterations in the CG algorithm plays the role of the hyperparameter λ, however, the GCV function can be simplified. Because of the connection between truncated CG and truncated singular value decomposition (SVD) through Ritz polynomials (Hansen, 1998), the GCV function can be approximated by (Haber, 1997)

$$GCV(iter) = \frac{\sum_{k=1}^{N}\left[d(iter)_k - d_k^{obs}\right]^2}{(N - iter)^2}. \tag{11}$$

This last approximation does not involve any extra operation inside the CG algorithm other than a simple division per iteration and is computationally much more tractable.

SPARSENESS AND SINGULAR VECTORS

The effect of the model weights on the final solution is to transform the singular vectors of the kernel such that they approximate the desired solution. The similarity between the singular vectors of the kernel and the true or desired solution is a very desirable property because it allows the iterative linear solver to map the desired solution in the first few iterations, leaving the noise and nondesired characteristic of the solution to be mapped at a later stage. Therefore, regularization imposed by the number of iterations becomes very efficient.

Even though the kernel of the transformation is defined uniquely, model weights have the highly appealing property of transforming the kernel so that its right singular vectors (the vectors that span the model space) become closer to the model used to calculate the model weights. In fact, at every external iteration, these singular vectors become better approximations to the true model, and a smaller number of internal iterations is required. This statement can be understood from the point of view that the number of internal iterations in CG plays a similar role in the solution as the number of significant singular values used to build the solution in truncated SVD (Hansen, 1998).

We now illustrate this phenomenon by means of a simple example. Two Ricker wavelets (Figure 1b) in the Radon space generate two parabolic events by using the inverse parabolic RT (PRT) (Figure 1a). The wavelets are located very closely in the RT space, so that they produce events with very similar curvature. The RT of these data was computed by using the least-squares RT and the high-resolution RT (Figures 1f and 1g, respectively). We chose a particular frequency to compare the singular vectors of the original kernel $\mathbf{L}$ and the singular vectors for the weighted kernel $\mathbf{L}\mathbf{W}_m^{-1}$. The model and the singular vectors are shown in Figure 1. Figure 1c displays the real part of the true solution (i.e., the original events in the Radon space). Figure 1d presents the real part of the third singular vector of the RT kernel $\mathbf{L}$. This singular vector is similar to a harmonic function and, as such, is not suitable for a parsimonious representation of the sparse model shown in Figure 1c. Many singular vectors are required to construct the desired sparse solution. Figure 1e presents the real part of the third singular vector of the weighted RT kernel $\mathbf{L}\mathbf{W}_m^{-1}$. The singular vector has become a localized function, much better able to represent sparse events with very few components. As a consequence, a small number of iterations is required when using iterative solvers like CG. For ideal clean large-aperture data, both kernels should lead to the same solution, but the presence of noise and other nonideal conditions make the final solutions very different. In these conditions, either the iterative scheme is stopped before completion or some regularization for the system of equations is applied. In both situations, the final solution is built by only some of the singular vectors. For the

nonweighted kernel, this implies smoothing or lack of resolution. For the weighted kernel, the solution is basically the same because the most important information is obtained in the first few iterations.

The example shows that a more parsimonious solution can be obtained by updating our knowledge about the model in every iteration and incorporating this information into the operator of the transform. Parsimony in the solution is important because there is a limitation, due mainly to the noise, on the number of singular vectors that we can use to build the model. Therefore, parsimony leads not only to a more resolved model but also allows us to decrease the influence of noise in the solution. As a practical consequence, the number of internal iterations decreases as the number of external iterations increases.

IMPLEMENTATIONS

Because, in general, very fast direct methods of solving equations (5) and (8) are not available, as there are for the standard

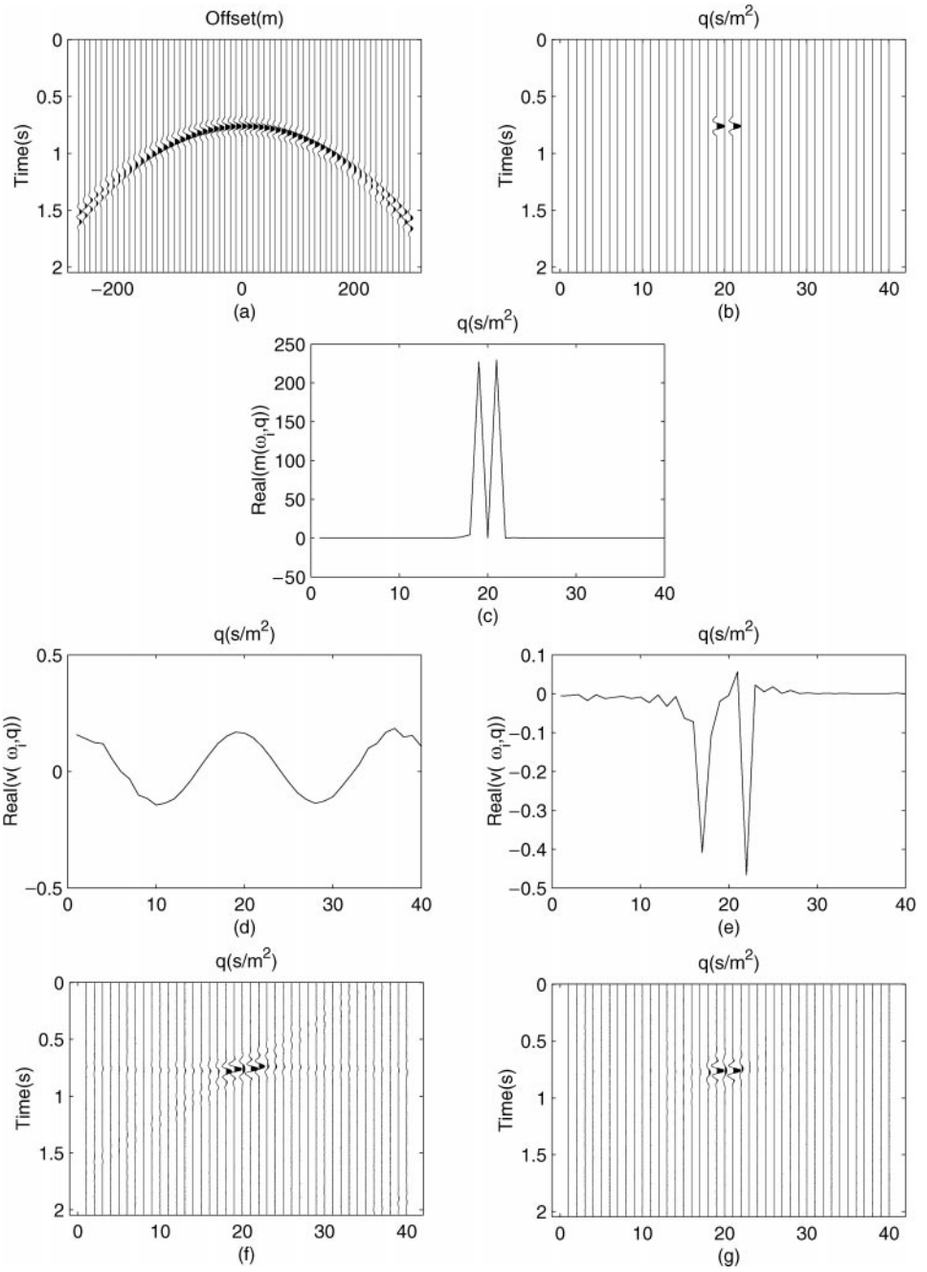

FIG. 1. (a) Two parabolic events with very similar curvature. (b) The ideal PRT of (a). (c) The ideal solution (real part) for a particular frequency. (d) Third right singular vector of the RT forward kernel (real part). (e) Third right singular vector of the model weighted RT kernel (real part). (f) Nonsparse RT obtained with the nonweighted kernel. (g) Sparse RT obtained with the weighted kernel.

(nonsparse) RT, iterative methods are usually preferred to solve these system of equations. CG is a common choice because of very well-known convergence properties, closely related to the nature of the SVD solution (Hansen, 1998). When solving the system of equations (5), the cost of the CG method is dominated by the action of the operator

$$\mathcal{A} = \left(\lambda \mathbf{W}_m^T \mathbf{W}_m + \mathbf{L}^T \mathbf{W}_d^T \mathbf{W}_d \mathbf{L}\right) \tag{12}$$

acting on a vector ($\lambda \neq 0$ in general). In Sacchi and Porsani (1999) and in Schonewille and Duijndam (2001), this operation is performed by circular convolutions in the Fourier domain. The same idea can be applied when solving the system of equations (8), where it is necessary to apply the transformed operator

$$\tilde{\mathcal{A}} = \mathbf{W}_m^{-T} \mathbf{L}^T \mathbf{W}_d^T \mathbf{W}_d \mathbf{L} \mathbf{W}_m^{-1}, \tag{13}$$

where the hyperparameter λ has been set to zero as mentioned before because the number of iterations, given by GCV, can be used instead to regularize. The required steps to compute the action of the operator $\tilde{\mathcal{A}}$ on a vector are:

1) At a given iteration, multiply the residual vector by the diagonal model weight matrix $\mathbf{W}_m^{-1}$.
2) Multiply the resulting vector by the matrix $\mathbf{L^T W}_d^T \mathbf{W}_d \mathbf{L}$ using fast Fourier transforms (FFTs), which requires first the expansion of this matrix to make it circulant (Sacchi and Porsani, 1999).
3) Multiply the resulting vector from the previous step by the diagonal model weight matrix $\mathbf{W}_m^{-1}$.

The sparse RT calculated with this scheme in the frequency domain is very fast. It has a disadvantage, however, because the model weights computed in the frequency domain are coupled for all times when, in fact they should not be. As a consequence, the same weights are imposed for all events (at different times) in a particular trace in the data and Radon spaces. One event at one particular time in the Radon space favors events in that trace for all times. As a result, artifacts are often generated for traces with high-energy events. These artifacts can be seen in the sparse RT in areas where the smooth RT shows no energy. When the hyperparameters are properly set, this problem can often be neglected, because the energy of the artifacts becomes much smaller than the energy of the main events. This is, however, one of the causes for the sensitivity of the results to the hyperparameters.

Cary (1998) recognized this weakness and proposed to compute time-invariant RTs in the time domain but applying frequency domain operators. Frequency domain operators have two important advantages over their time domain counterparts. First, if the computation is done properly inside the bandwidth of the signal, the waveforms are well preserved. Secondly, the action of the forward and adjoint operators can be computed by means of circular convolution in the Fourier domain. Thus, this approach combines the best of both methods, flexibility in the model and data weights, time sparseness, speed, and good waveform preservation. Because it imposes sparseness in the Radon coordinate and in time, the RT is very clean, with few artifacts. The Radon frequency-domain operator is applied in cascade with a forward and inverse FFT:

$$\mathbf{d} = \mathcal{F}^{-1} \mathbf{L} \mathcal{F} \mathbf{m}. \tag{14}$$

In this equation, $\mathbf{d}$ and $\mathbf{m}$ are vectors in the time domain, and $\mathcal{F}$ and $\mathcal{F}^{-1}$ are forward and inverse Fourier transforms. The same CG method can be used, but every time the forward or adjoint operators are required in the algorithm, the three operators are applied in sequence. In particular, the weighted operator $\tilde{\mathcal{A}}$ now becomes

$$\tilde{\mathcal{A}} = \mathbf{W}_m^{-T} \mathcal{F}^T \mathbf{L}^T \mathcal{F}^{-T} \mathcal{F}^{-1} \mathbf{L} \mathcal{F} \mathbf{W}_m^{-1}, \tag{15}$$

where we have not included $\mathbf{W}_d^T \mathbf{W}_d$ for the sake of simplicity. We can simplify equation (15) because, for the Fourier transform with sampling $\Delta f = 1/(N \Delta t)$ (the commonly used FFT),

$$\mathcal{F}^{-1} = \mathcal{F}^T \text{ and } \mathcal{F}^{-T} = \mathcal{F}. \tag{16}$$

Hence,

$$\tilde{\mathcal{A}} = \mathbf{W}_m^{-T} \mathcal{F}^{-1} \mathbf{L}^T \mathbf{L} \mathcal{F} \mathbf{W}_m^{-1}. \tag{17}$$

Thus, the CG matrix vector multiplication can be applied as

1) At a given iteration, multiply the residual vector by the diagonal model weight matrix $\mathbf{W}_m^{-1}$.
2) Take the FFT of the result.
3) Multiply the resulting vector by the matrix $\mathbf{L^T W}_d^T \mathbf{W}_d \mathbf{L}$ using FFTs, which requires that $\mathbf{L^T W}_d^T \mathbf{W}_d \mathbf{L}$ be made circulant.
4) Take the inverse FFT of the result.
5) Multiply the resulting vector from the previous step by the diagonal model weight matrix $\mathbf{W}_m^{-1}$.

Full sparseness in time, as well as in q (i.e., assuming that events map to spikes in time), is an unrealistic constraint. Partial sparseness can be imposed by applying another operator, convolution with a wavelet in the forward operator, and correlating with the same wavelet in the adjoint. This is equivalent to assuming a nondiagonal covariance matrix because the model parameters are correlated in time. Of course, this procedure requires some knowledge of the wavelet. If this is not the case, an approximated zero-phase and zero-mean wavelet can be used with the purpose of band-filtering the RT space. However, this is not a critical point because in practice we always impose only partial sparseness by truncating the external iterations before achieving the minimum of the originally proposed cost function. Therefore, the convolution operator can be usually neglected.

Hence, in every iteration of CG least squares (CGLS), it is necessary to apply four operators for the forward sequence (convolution, FFT, Radon transform, and inverse FFT) and for the adjoint (FFT, Radon transform, inverse FFT, and correlation). Results with this method are usually very clean and artifact free. The price to pay is, of course, more computing time. However, application of all these operators is very fast, and the main burden is that the convergence of CG is slower due to the large number of variables in the time domain. It also requires some extra memory space for storing the three dimensional Radon operator [$L(\omega, h, q)$, where ω is frequency, h is offset, and q is the Radon parameter].

Time-variant RTs (hyperbolic, elliptical, and other variants) require very large (and sparse) operators in the time domain. The design of the operators makes a large difference in time domain methods because of their complexity and flexibility. Time domain methods produce the most sparse results, because of

the similarity of their basis functions to the seismic reflections. Similar to what happens in migration methods, time domain algorithms can compete very efficiently with frequency domain algorithms when only part of the output space is required. For example, multiple attenuation by RT becomes very fast for deep marine data, as the required output space involves only those events following the first multiples.

Another aspect that leads to different implementations is the procedure used to estimate the model weights. A fairly simple and efficient procedure is to take the model information not from a previous iteration but from previous frequencies (Herrmann et al., 1999). This implementation avoids external iterations (partially or completely, depending on the data) and also attenuates alias artifacts because low frequencies are less demanding in terms of offset sampling and are, in general, not aliased. Therefore, if the model weights are transported from low frequencies to high frequencies, it is less likely that they will be affected by aliasing. The assumption behind this method is that the spectrum of the traces in the Radon space changes slowly. This may be true for the wavelet itself, but is certainly not true after convolution with the reflectivity series. Hence, some smoothing is required. Our implementation of this method calculates the first 10–15 frequencies of the RT using IRLS. For the following frequencies, the model weights are computed as a function of an average of the model space inside a sliding window that extends from the last computed frequency toward low frequencies. Therefore, external iterations are required only for the first frequencies. The degree of sparseness can be controlled by the number of external iterations over the first frequencies, by the number of frequencies computed using IRLS, and/or by the length of the window used for the model weights. A combination of all these methods seems to work fine in general.

Avoiding external iterations is very important for time domain methods because of the already large computation time. When the CMPs are computed along a seismic line, it is usually a good approximation to carry the model weights from CMP to CMP. A semblance function also produces a good approximation to the model weights (Stoffa et al., 1981; Yilmaz and Taner 1994).

APPLICATIONS AND EXAMPLES

Although the achievement of high resolution in the RT has many applications, facilitation of multiple removal is probably the most important. In the first example, we test the improvement in resolution for the previously discussed algorithms in a very simple synthetic gather. Figure 2 shows an example of multiple attenuation using sparse RT. The synthetic gather for the resolution tests consists of four events following a given velocity trend, simulating primaries, and three events with the velocity of the first primary, simulating multiples. Similar to what happens in surface reverberations, one of these multiples has its polarity reversed. The destructive interference of this event with a primary located at the same two-way traveltime produces almost zero amplitude at near offsets and a change of shape of the wavelet as the difference in moveout increases with offset. In all the examples we apply a simple (time invariant) automatic muting in the Radon space to eliminate the primaries, and from this filtered space we recover the multiples. These multiples are subtracted from the data. Events with residual moveout smaller than 0.9×10^{-8} s/m^2 (the units represent residual moveout per offset square) have been removed by the mute process. Figure 3 shows the RTs for this synthetic gather, computed using the nonsparse RT (Figure 3a), the frequency domain CG algorithm (Figure 3b), the time domain CG with frequency domain operators (Figure 3c), and the hyperbolic time-domain RT (HRT) (Figure 3d). In the time domain examples (Figures 3c–d) the RT mapping contains spikes rather than wavelets because the convolution with the wavelet has been included into the operators. In all the CG algorithms, the operators were applied using the FFT circular convolution method. The gain in resolution is considerable because the data are almost noise free and very simple in nature. In this situation, it is possible to tune the hyperparameters to obtain very high resolution without generating noticeable artifacts. In all the sparse transforms, we have used a quantile of 10% as a robust measure of the data and five external iterations. The number of internal iterations, automatically determined by GCV, decreases as the model weights become closer to the sparse solution, but for this example ranges from 30 to 10.

With real data, however, other aspects come into play. Different ways to implementing a sparse RT sometimes yield very different results. Figure 4 shows an example of multiple attenuation on a CMP gather from the Gulf of Mexico. In this example, we have used the fast frequency-domain CG. Figure 5 presents the same example but using the HRT. Because of the flexibility of the time domain algorithm, the RT space could have been computed only below 3.5 s (there are no multiples above this time). The HRT suffers from moveout stretching as normal moveout (NMO) correction (and hence PRT) does. Therefore, we apply a stretching filter to the data before the transformation. The separation of primaries and multiples has been performed efficiently by both methods. The PRT preserved the waveform and complicated nature of the events in a better manner, demanding less time as well. Yet, the HRT has separated more clearly the primaries and multiples and achieved more sparseness in the RT space.

We have obtained similar separation of primaries and multiples by using all the discussed methods. However, a closer examination of the RT space shows differences between different implementations. Figure 6 presents a comparison between the RT of the same gather computed by the standard RT using Levinson algorithm (Figure 6a), the fast frequency-domain CG (Figure 6b), the time domain CG (Figure 6c), and the hyperbolic RT with irregular RT space (Figure 6d) as discussed in Trad et al. (2002). The standard RT is very clean (sparse) in the vertical direction (along time), but shows the typical truncation artifacts (Kabir and Marfurt, 1999) and reduced lateral resolution (along q). The result with the fast sparse RT in Figure (6b) shows high lateral resolution, but the frequency dependence of the model weights gives rise to artifacts in the vertical direction. These artifacts become more evident when a clip less than 100% is used in the displays, otherwise they are difficult to note. This problem can be attenuated by the relaxation of the sparseness constraint, although that approach also decreases the lateral resolution. The time-domain sparse PRT approach overcomes this problem (Figure 6c). The RT looks very clean in both directions, in particular when the iterative algorithm is stopped at early stages, giving mainly a reconstruction of the strongest events that have to be attenuated. The computation time increases by about four times, even though FFT matrix

vector multiplication is used. Whether this improvement in vertical resolution pays off the extra time required is data and goal dependent. The HRT (Figure 6d) shows a better separation of the events, mainly because the similarity of the basis functions and the seismic events leads to better resolution. Note that the parameter along the horizontal axis has different meaning for the HRT than for the PRTs.

These differences between multiple removal in different implementations are not easy to note in a final section, after NMO and stack. Differences between PRT and HRT are more evident. Figure 7a contains the NMO + stack section of the Mississippi Canyon data set, without multiple attenuation. Figures 7b and 7c show the same section with multiple attenuation using PRT and HRT, respectively. Some improvements can be seen using the HRT, because of the large moveout difference between primaries and multiples in this data set. The differences in the shallow part are due to the fact that the HRT has been used to predict multiples below 3 s only, because the surface multiples are mainly in depth. In the PRT, we have used a quantile of 50%, three external iterations, and a maximum of 30 external iterations. For the HRT, we changed the quantile to 80%.

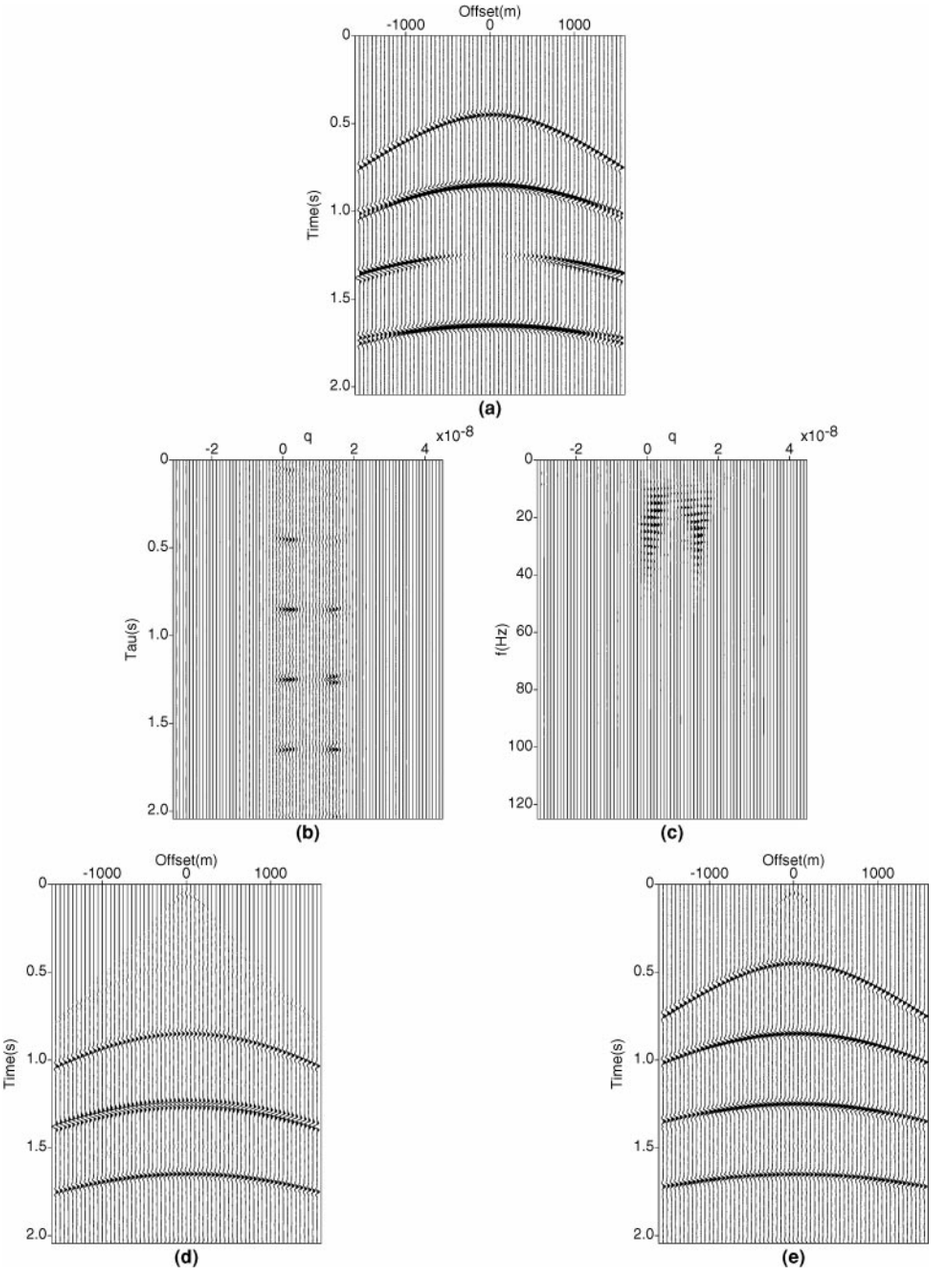

FIG. 2. (a) Synthetic gather. (b) Sparse RT. (c) Spectrum of (b). (d) Recovered multiples after muting in (b). (e) Primaries obtained by subtracting (d) from (a).

PROBLEMS AND SOLUTIONS

Aliasing in the RT due to poor offset sampling is frequency dependent. As a consequence, aliasing yields artifacts that seem uncoherent in the time domain representation of the RT. This characteristic makes sparse time-domain algorithms capable of attenuating aliasing (Yilmaz and Taner, 1994). Also, the distance between aliased versions of a true event is larger for the low frequencies than for the high frequencies. Therefore, as mentioned previously, frequency domain methods where the model weights are carried out from frequency to frequency are also successful in attenuating aliasing (Herrmann et al., 1999). Figure 8a shows a gather with two parabolic events. The alias of the first event at 0.4 s will map into the same range of q values as the second event at 0.8 s. As a result of the frequency dependency of aliasing, the RT contains a series of artifacts when normal frequency domains methods are used (Figure 8b). Time domain algorithms can, when the sparseness constraint is enforced, yield aliasing free RTs (Figure 8c). The same good results can be obtained by a frequency-domain dealiased RT (Figure 8d).

Another serious problem in the RT is the existence of the null space of the transform (i.e., when part of the data has a zero mapping into the transformed domain). This null space arises mainly when the range of spanned basis functions, given by the sampling and aperture in the RT space, is insufficient to synthesize some of the data events, either noise or signal. From this point of view, it is interesting that a combined linear–pseudohyperbolic Radon transform (Trad et al., 2001) decreases the size of the model space without increasing the nullspace of the transform. An advantage of using a dual RT operator is the freedom to choose different parameters for the two parts of the model space (e.g., sampling interval in the Radon domain, maximum frequency to use, regular or irregular Radon sampling, etc.). If the data contain linear and hyperbolic events, the sampling for the hyperbolic space cannot fulfill the aliasing condition for linear and hyperbolic events simultaneously. The dual RT does not have this problem; each part of the model space fulfills the aliasing condition for the events that it is expected to map. To show its capability, we remove ground roll in a real data set: shot gather number 25 from Yilmaz (1987) (Figure 9a, reprinted from Trad et al., 2001).

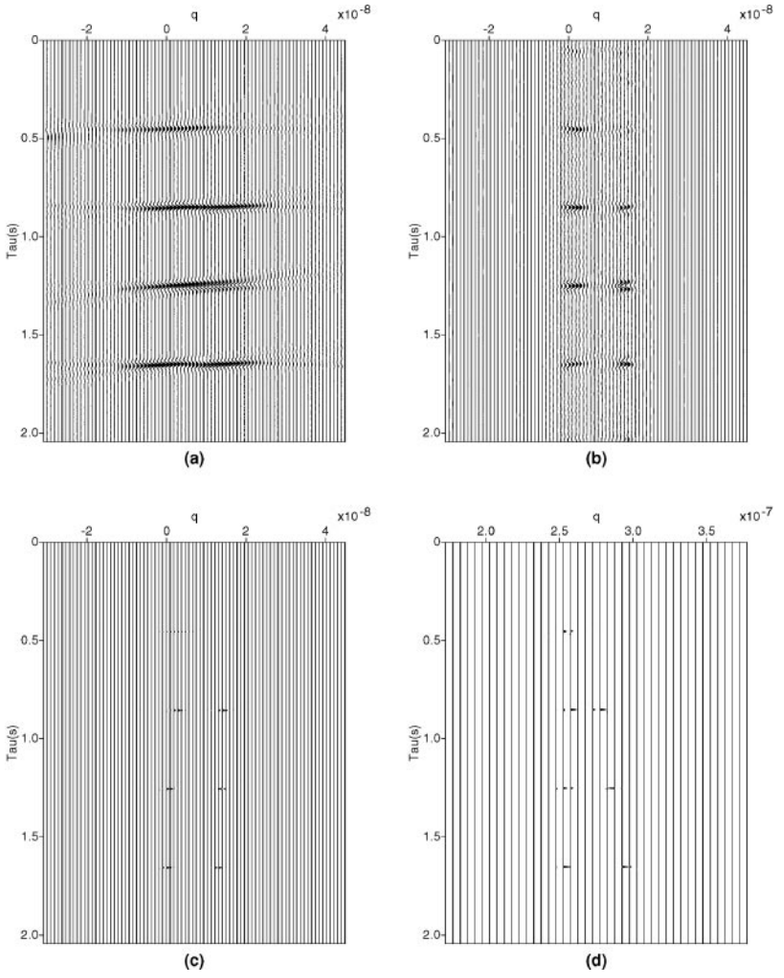

FIG. 3. (a) Nonsparse PRT. (b) Frequency-domain sparse RT. (c) Mixed frequency-time domain PRT (d) HRT.

Because a shot gather is in general asymmetric, the PRT requires a different model for negative and positive offsets (this is another problem where a RT with two different operators can be used). Figure 9b shows 320 traces in the Radon domain. The first 160 traces correspond to the negative offsets, traces 161–320 correspond to the positive offsets. Each half shows the two spaces side by side; the linear RT is on the left, the pseudohyperbolic is on the right. Applying again a mute in the pseudohyperbolic space or, equivalently, using only the linear operator of the hybrid RT, the linear ground roll can be modeled (Figure 9c) and separated by subtraction (Figure 9d). In this example, we used a quantile of 10%, three external iterations, and a maximum of 50 internal iterations. The bandwidth for the parabolic RT was 80 Hz, but for the linear RT it was only 40 Hz because the ground roll has mainly low frequencies. A very small part of the ground roll remains in the small offset traces and can perhaps be removed using additional filtering. This part of the data has not been predicted for two reasons: the slopes become very large (approaching infinity), and the linear events are heavily aliased. In fact, *f-k* filtering has similar difficulty in removing the ground roll in the near zero-offset traces.

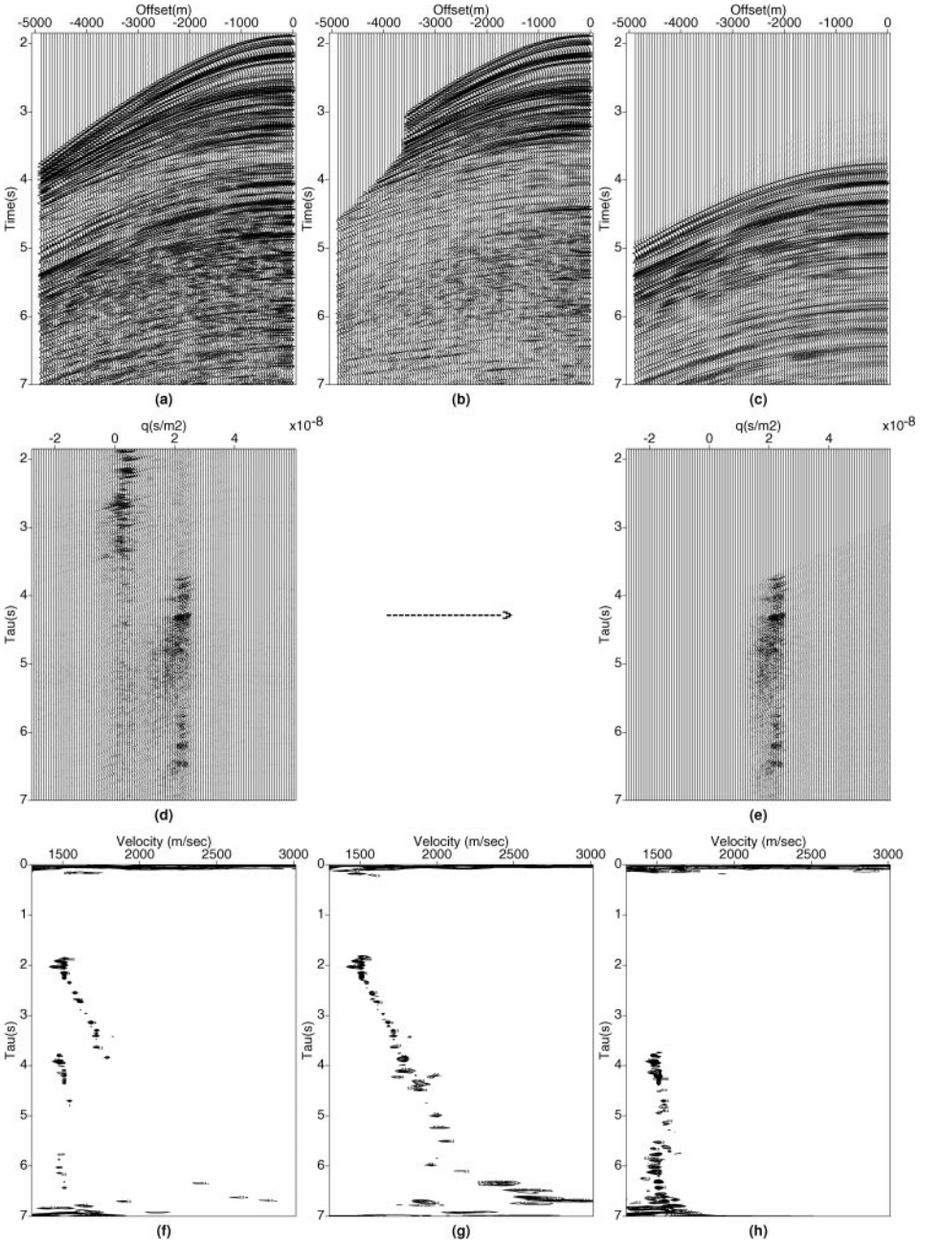

FIG. 4. (a) Marine CMP. (b) Primaries obtained by subtraction. (c) Multiples obtained from inverse RT of (e). (d) Sparse PRT. (e) RT after mute of primaries. (f-g-h) Semblance analysis for (f) data, (g) primaries, and (h) multiples.

CONCLUSIONS

Problems in the RT appear as a result of limited information, including discrete sampling in offset and time, discrete velocity parameterization, finite aperture in offset and velocities, and missing data, in particular due to low fold. As a consequence, the conventional RT suffers from low resolution, artifacts, aliasing, and a nontrivial null space. Even though none of these problems can be completely solved because they arise from the character of discrete signals, many of them can be attenuated by means of the sparse RT. From the examples considered in this paper, we can conclude that the performance of the sparse RT in separating primaries from multiples and interpolating gaps, is superior to the performance of the standard RT. Nonetheless, its application to real data is nontrivial, and well-designed algorithms are required. Not only must the algorithms be fast and robust, they must also be easy to apply.

Many different implementations have been developed for the sparse RT. Differences arise due to the domains where the RT is computed, how the model weights are obtained, and how the operators are designed. One problem common to all of them is the difficulty in choosing the hyperparameters. This

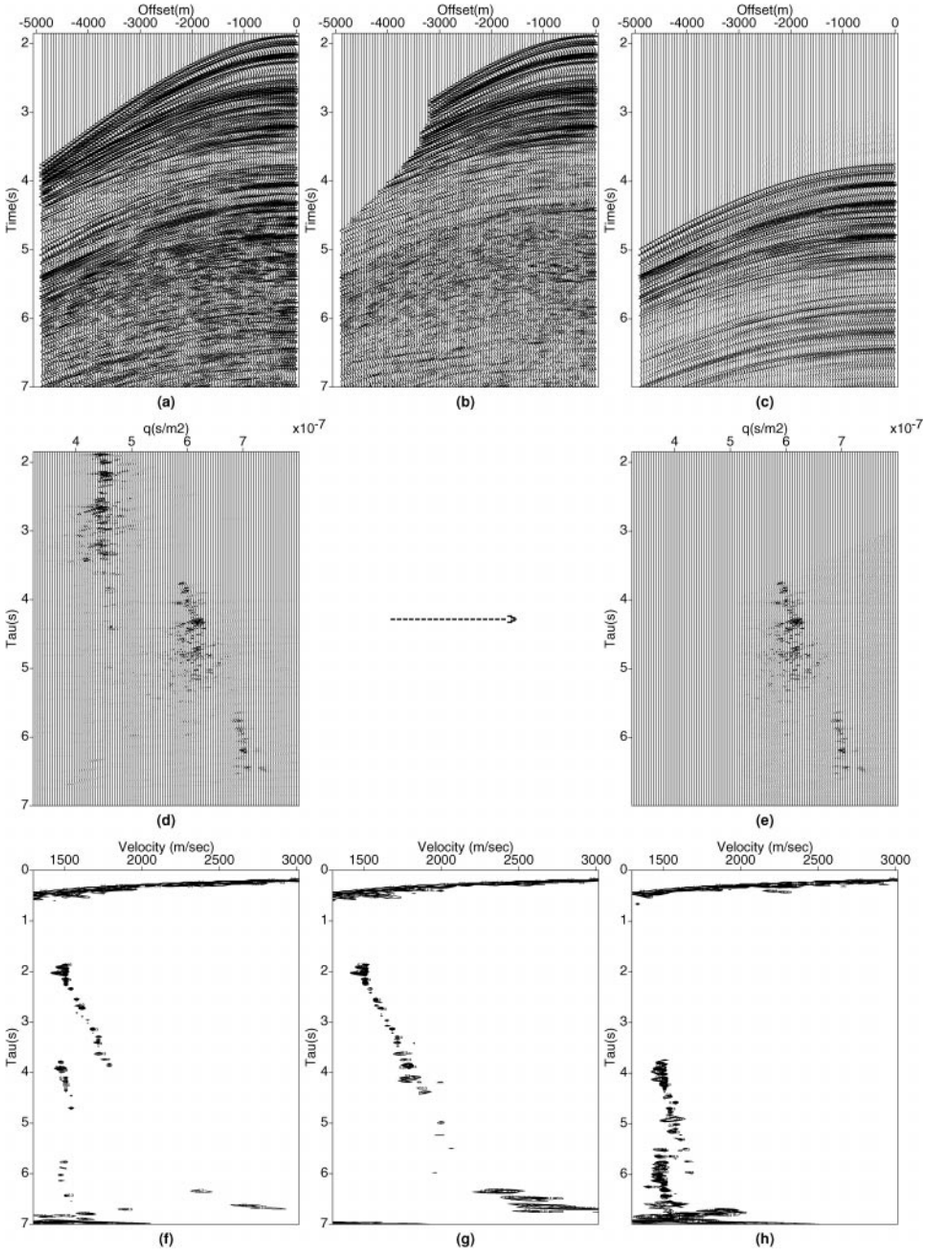

FIG. 5. (a) Marine CMP. (b) Primaries obtained by subtraction. (c) Multiples obtained from inverse RT of (e). (d) Sparse HRT. (e) RT after mute of primaries. (f-g-h) Semblance analysis for (f) data, (g) primaries, and (h) multiples.

is indeed a very difficult problem because the computation time must be kept low, and hence many well-known (and slow) methods of estimating the hyperparameters cannot be used. A simple but partial solution to this problem resides in incorporating the model weights into the operators and using the number of iterations as a regularizer. If a good stopping criterion is applied, the effect of the hyperparameters is less critical. A combined GCV algorithm appears to be very practical. The GCV function stops the inner iterations, acting as a filter for the mapping of poorly resolved events in the data. The number of external iterations acts as a tradeoff between sparseness and smoothness. A good stopping criterion for the external iterations is less of a problem, because the IRLS algorithm tends to converge to a stable solution and the minimum of the cost function is never reached in practical applications.

In this paper, we have used the ideas of fast conjugate gradient algorithms, IRLS method, inversion by transformation to standard form and GCV to implement several known varieties of the RT and have compared results. Frequency-domain sparse RTs provide a robust fast tool for interpolation and separation of events. The main weakness of these implementations is the introduction of artifacts, a problem that is efficiently addressed by time domain methods. Time domain methods allow the enforcement of time sparseness, leading to cleaner transforms, but some price has to be paid in terms of computation time.

Operator design and computation of the model weights are two wide areas of research. We have discussed some practical aspects of both topics. We have also exemplified how the flexibility of the RT to separate events with different shape can be extended by creating hybrid transforms with more than one operator. This implementation leads to a very flexible transform, in the sense that two different sets of basis functions can be used and the two Radon spaces can be designed with different characteristics to avoid aliasing and other artifacts. One application of this hybrid approach that we have illustrated is the separation of coherent noise.

ACKNOWLEDGMENTS

We are particularly grateful to Eldad Haber, who has played an important role in the development of this work and been an

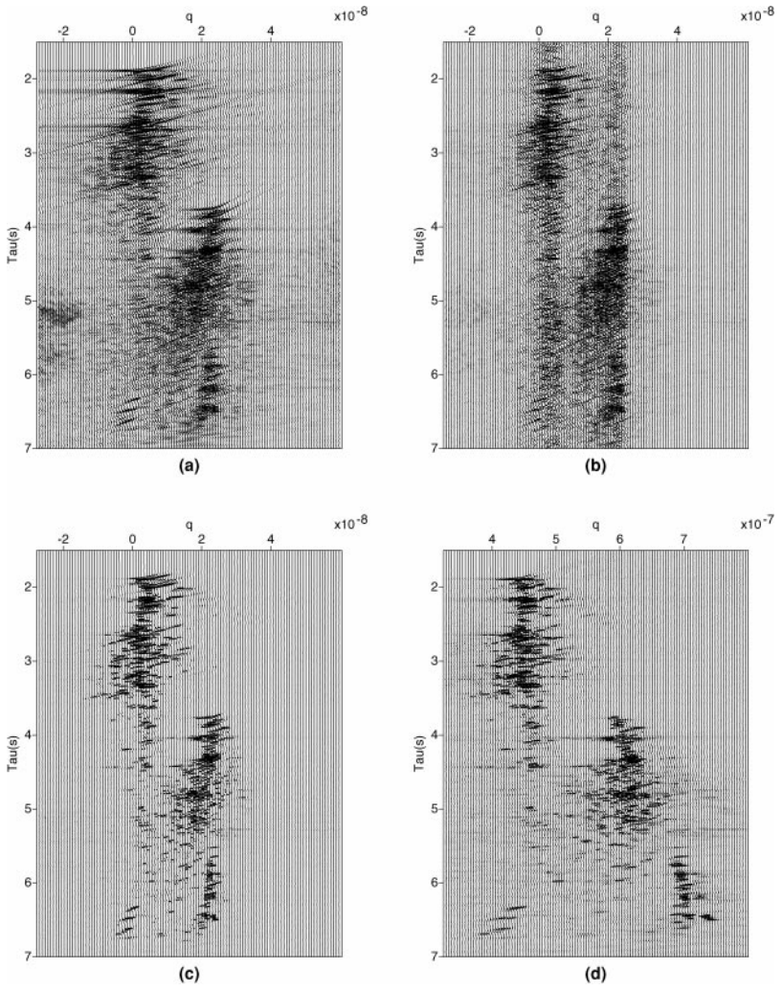

FIG. 6. (a) Nonsparse PRT. (b) Frequency domain sparse RT. (c) Mixed frequency-time domain PRT. (d) HRT.

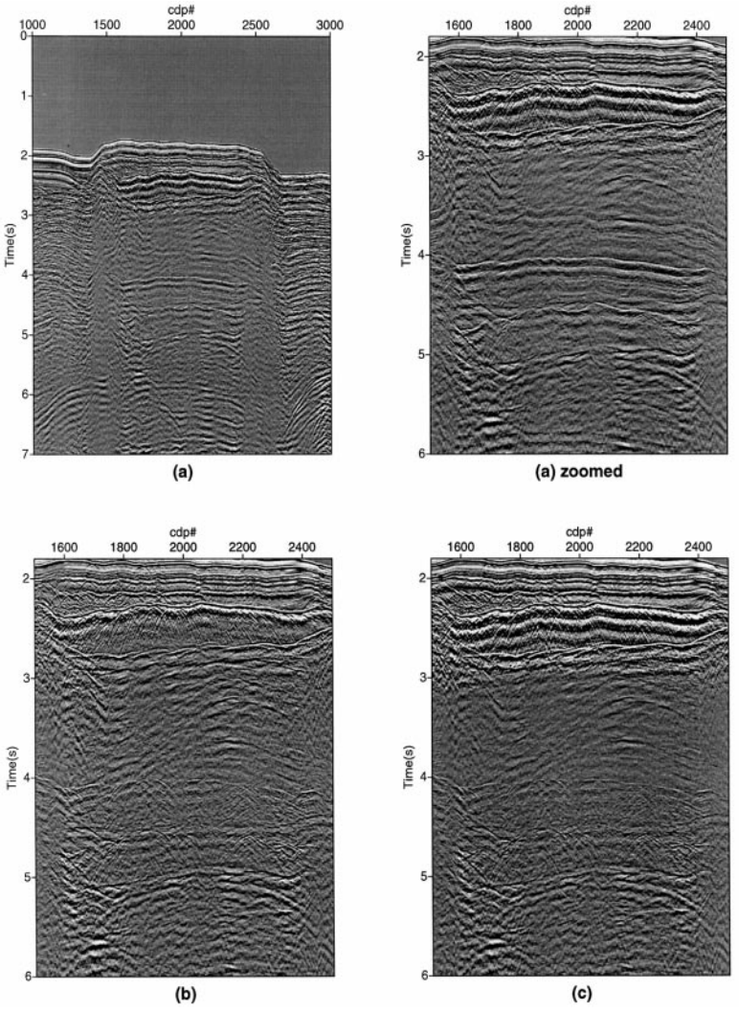

FIG. 7. Mississippi Canyon data set: (a) NMO + stack. (b) NMO + PRT + stack. (c) NMO + HRT + stack.

unending source of algebraic tools and inspiration for us. We are also in debt to the associate editor of *Geophysics* and three reviewers for their valuable suggestions. Finally, we acknowledge Western Digital for the Mississippi Canyon data set.

REFERENCES

Beylkin, G., 1987, Discrete Radon transform: IEEE Trans. Acoust., Speech, and Sig. Proc., **35**, 162–172.

Cary, P., 1998, The simplest discrete Radon transform: 68th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1999–2002.

Claerbout, J., 1992, Earth sounding analysis: Processing versus inversion: Blackwell Scientific Publ., Inc.

Haber, E., 1997, Numerical strategies for the solution of inverse problems: Ph.D. thesis, Univ. of British Columbia.

Hampson, D., 1986, Inverse velocity stacking for multiple elimination: J. Can. Soc. Expl. Geophys., **22**, 44–55.

Hansen, P. 1998, Rank-defficient and discrete ill-posed problems: Numerical aspects of linear inversion: Soc. Ind. Appl. Math.

Herrmann P., Mojesky, T., and Magesan, M., 1999, Amplitude preserving Radon demultiple: Beyond sampling and aperture limitations: Nat. Conv., Can. Soc. Expl. Geophys., Abstracts, 73–74.

Kabir, M. M. N., and Marfurt, K. J., 1999, Toward true amplitude multiple removal: The Leading Edge, **18**, 66–73.

Lupton, R., 1993, Statistics in theory and practice: Princeton Univ. Press.

Sacchi, M., and Porsani, M., 1999, Fast high resolution parabolic RT: 69th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1477–1480.

Sacchi, M., and Ulrych, T., 1995, High-resolution velocity gathers and offset space reconstruction: Geophysics, **60**, 1169–1177.

Scales, J., Gersztenkorn, A., and Treitel, S., 1988, Fast l_p solution of large, sparse, linear systems: Application to seismic travel time tomography: J. Comp. Phys., **75**, 314-333.

Schonewille, M., and Duijndam, A., 2001, Parabolic Radon transform, sampling and efficiency: Geophysics, **66**, 667–678.

Stoffa, P. L., Buhl, P., Diebold, J. B., and Wenzel, F., 1981, Direct mapping of seismic data to the domain of intercept time and ray parameter—A plane-wave decomposition: Geophysics, **46**, 255–267.

Thorson, R., and Claerbout, J., 1985, Velocity-stack and slant-stack stochastic inversion: Geophysics, **50**, 2727–2741.

Trad, D., Sacchi, M., and Ulrych, T., 2001, A hybrid linear-hyperbolic Radon transform: J. Seis. Expl., **9**, 303–318.

Trad, D., Ulrych, T., and Sacchi M., 2002, Accurate interpolation with high-resolution time-variant Radon transforms: Geophysics, **67**, 644–656.

Yilmaz, O., 1987, Seismic data processing.

Yilmaz, O., and Taner, M. T., 1994, Discrete plane-wave decomposition by least-mean-square-error method: Geophysics, **59**, 973–982.

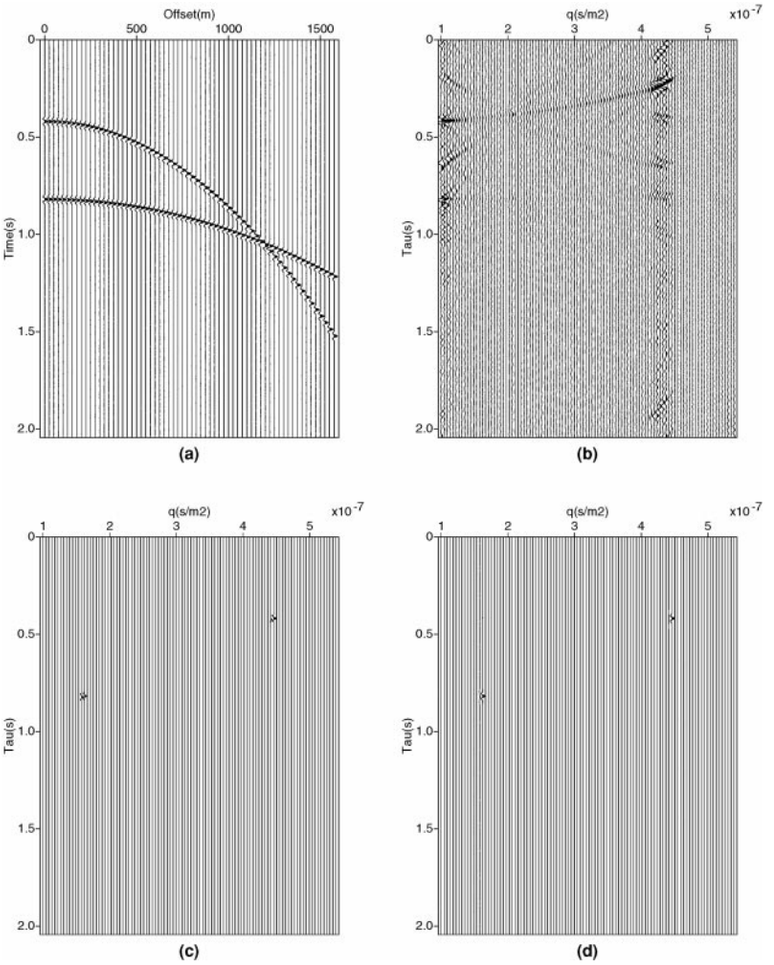

FIG. 8. (a) Nonsparse PRT. (b) Frequency-domain sparse RT. (c) Sparse time-domain PRT (d) Dealiased PRT.

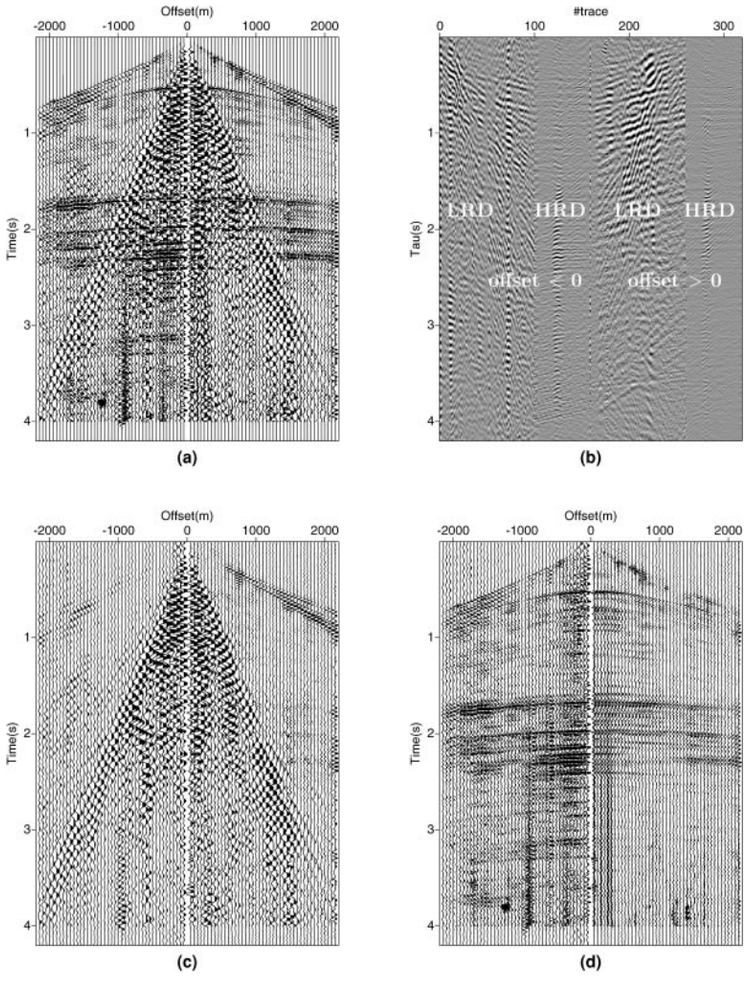

FIG. 9. (a) Shot gather. (b) Combined linear + pseudohyperbolic Radon domain, for negative and positive offsets. (c) Recovered gather from linear space only. (d) Signal obtained by subtracting (c) from (a).