GEOPHYSICS, VOL. 64, NO. 5 (SEPTEMBER-OCTOBER 1999); P. 1648–1656, 6 FIGS.

Tutorial

Signal and noise separation: Art and science

Tadeusz J. Ulrych*, Mauricio D. Sacchi‡, and J. Michael Graul**

ABSTRACT

The separation of signal and noise is a central issue in seismic data processing. The noise is both random and coherent in nature, the coherent part often masquerading as signal. In this tutorial, we present some approaches to signal isolation, in which stacking is a central concept. Our methodology is to transform the data to a domain where noise and signal are separable, a goal that we attain by means of inversion. We illustrate our ideas with some of our favorite transformations: wavelets, eigenvectors, and Radon transforms. We end with the notion of risk, baseball, and the Stein estimator.

INTRODUCTION

The purpose of this article is to present, in tutorial fashion, some approaches to the ubiquitous problem of signal and noise separation that we have found useful. Since it is not detail that we are primarily interested in presenting, we adhere to the reasonable principle of parsimony, both in words and in mathematical expression. We begin with some condensed notation.

Let **D** represent the record of our seismic experiment in whatever form. **D** contains all the information that is at our disposal. Our task is to chisel away at **D** to unearth **S**, the signal therein contained. We are artisans, vainly following Michelangelo. **D** is the rock, **S** is our David. Of course, since we all know the adage "one man's signal is another man's managerial rebuke," we must define **S**. We define signal as that energy that is coherent from trace to trace. Noise, **N**, on the other hand, is that energy that is incoherent from trace to trace. Unfortunately, the most expensive noise is also spatially coherent, and we must modify the above definition. We define signal as that energy that is most coherent and desirable for our interpretation of primary reflected arrivals. For example, Rayleigh waves and multiples can be coherent "undesired signal" or "noise," yet such arrivals would be undesirable for an analysis of primary reflections. This definition immediately introduces one of the central themes of this article: resolution. In order to extract the signal from the background of noise, we must be able to resolve the difference, often very subtle and highly dependent on the acquisition, between the coherent signal and noise. We can now write the model,

$$\mathbf{D} = \mathbf{S} + \mathbf{N}_c + \mathbf{N}_r, \tag{1}$$

where $\mathbf{N}_c$ and $\mathbf{N}_r$ represent the coherent and incoherent noise components, respectively, and $\mathbf{N}_c + \mathbf{N}_r = \mathbf{N}$. From now on, for parsimony of notation, the quantities **D**, **S**, and **N** represent matrices.

In the permanent and ubiquitous task of the identification and suppression of **N**, two processes share the honors for most significant quantum leaps forward. Variable area plotting [according to one of us (Graul)] and stacking. It is the latter that this article is mainly about.

METHODOLOGY

D lives in space and time, the *x*-*t* domain. It is characteristic of this domain that **S** and **N** are intertwined and, consequently, are not only difficult to separate but also to identify. In order to accomplish these tasks, **D** must be mapped into a domain where the characteristics distinguishing signal and noise map **S** and **N** into separate spaces. In operator form,

$$\mathcal{T}\mathbf{m} = \mathbf{d}, \tag{2}$$

where $\mathcal{T}$ is the linear or nonlinear transformation, **m** is the vector of model parameters in the transformed domain, and **d** is a data vector realization from **D**.

Manuscript received by the Editor September 4, 1998; revised manuscript received April 19, 1999.

*University of British Columbia, Department of Earth and Ocean Science, 129-2219 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada. E-mail: ulrych@eos.ubc.ca.
‡University of Alberta, Department of Physics, Avadh Bhatia Physics Laboratory, Edmonton, Alberta T6G 2J1, Canada. E-mail: sacchi@phys.ualberta.ca.
**Texseis Inc., 2618 Hollow Hook, Houston, Texas 7708-3814. E-mail: mgraul@snipertg.com.

The transformation expressed by equation (2) is guided by a principle that is, to our minds, fundamental and lucidly expressed by the late Edwin Jaynes, whom we honor for his many profound contributions. Jaynes wrote (personal communication, 1981), "A problem in inference that is posed as an overdetermined problem is badly posed." That is, the model **m** is nonuniquely related to the data. As such, it may be recovered only when the inverse transformation is regularized, a topic we discuss in more detail later. A more common way, perhaps, of expressing this nonuniqueness is to say that the inverse problem is underdetermined, **m** contains more elements than **d** (fewer equations than unknowns).

Now comes the second part of our methodology. The manner of regularizing the transformation in equation (2), expressed in a form that emphasizes the inverse nature of the problem, follows Bayesian principles. We impose a priori information by means of Bayes' theorem, and in this manner obtain, from an infinity of possibilities, that model which honors the data and satisfies our prior beliefs. A pictorial representation of our two part methodology is illustrated in Figure 1.

PRINCIPLES OF STACKING

In much of what follows, the transformation in equation (2) is linked in some indirect, often obscure manner, to stacking, and we find it illuminating to deal briefly with the essentials of this concept. Let us suppose, for purposes of illustration only, that in equation (1) $\mathbf{N}_c = 0$, **S** is perfectly aligned and, most fortuitously, **N** is such that at every time sample its average value, taken spatially, is zero. Then, even with noise of Himalayan proportions, **S** may be recovered exactly by the simple process of stacking. In a less perfect world, signal stacks constructively and random noise destructively such that, for the special case of Gaussian noise, the noise amplitude is attenuated by the square root of the number of traces. Stacking involves a very special estimation of the first moment of a probability distribution. It is the maximum likelihood estimator, which we write as

$$\delta^0(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} x_i, \tag{3}$$

where the x_i are M samples of the random variable **x**. δ^0 is a very special estimator indeed. Statisticians have shown that, given $\mathbf{x}|\mu \overset{\text{ind}}{\sim} N(\mu, 1)$ (symbolically representing that **x** is independently distributed and comes from a normal or Gaussian probability distribution with mean μ and variance 1), δ^0 has lowest risk of any linear or nonlinear unbiased estimator, a characteristic that should surely appeal to all. We will return to this fascinating topic again later.

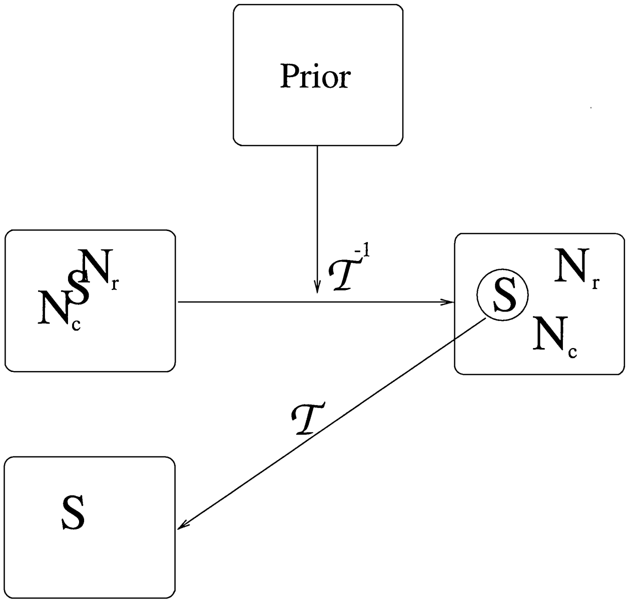

FIG. 1. Pictorial representation of the processing methodology.

Trimmed means

An important issue in stacking is the concept of robustness. It occasionally happens that our data are infected with a few large errors and, consequently, the tails of the underlying distribution are heavier than those of the Gaussian distribution. In such circumstances, δ^0 is much influenced by such errors, and the estimate is said to be not resistant or nonrobust. Overcoming such effects requires the concept of order statistics. The most widely known of such statistics is the median, computed as the middle value of $M + 1$ ordered numbers. Clearly, the median is considerably less sensitive to outliers and for this reason is often used in robust data processing. A point worth noting is that for symmetric distributions the median is equal to the mean.

Consider first the concept of a trimmed mean that is identified by the proportion that is trimmed from each end of the ordered sample. Thus, the 10% trimmed mean of a sample of 20 points is the mean of the remaining 16 points. Note that the median is approximately a 50% trimmed mean. Sometimes, a specified trimming might entail a fraction of an observation. In this case, a weight is assigned to the remaining partially trimmed data, and the resulting estimator is called an α-trimmed mean, which has seen application in seismic data processing.

Let $y_1 \le y_2 \cdots \le y_M$ be the ordered data points. Define k to be the integer that is less than or equal to αM where $0 \le \alpha \le 0.5$, and assign $r = \alpha M - k$. The α-trimmed mean, $T(\alpha)$, is defined by

$$T(\alpha) = \frac{1}{M(1-2\alpha)}\left[(1-r)(y_{k+1} + y_{M-k}) + \sum_{i=k+2}^{M-k-1} y_i\right]. \tag{4}$$

For the sake of completeness, we point out that the α-trimmed mean, just like the median, is an L-estimator, which is a weighted estimator of order statistics. L-estimators, in particular L-moments that are robust estimates of the moments of probability distribution functions, are seeing much press in recent times. An excellent review of some robust data analysis techniques may be found in Kleiner and Graedel (1980). For those seeking L-moments punishment, there is Ulrych et al., (1999).

Weighted stack

At this point we wish to mention the weighted stack and, in particular, the version introduced by Schoenberger (1996). The idea is very appealing. Let us represent a seismic section consisting of M traces with N points per trace by the N row × M column matrix **D**. The application of conventional stacking may be written as $\mathbf{Dw} = \hat{\mathbf{s}}$, where $\mathbf{w}$ is a vector of M weights, each equal to $1/M$, and $\hat{\mathbf{s}}$ is the stack, a N-element vector that is an estimate of the signal. Schoenberger suggested determining $\mathbf{w}$ so that, first, $\hat{\mathbf{s}}$ is an unbiased estimate and, second, $\mathbf{N}_c$ is attenuated because the weights are designed to reject those frequencies at which the coherent energy is present. The former leads to $\sum w_i = 1$; the latter to a matrix equation similar in form to the least squares normal equations.

SOME SPECIAL TRANSFORMATIONS

All of us have our favorite transformations. Here are a few of ours.

Eigenvector

We deal briefly with two eigenvector decompositions that have proved to be of particular value.

Eigenimages.—Consider our data $\mathbf{D}$ to be, as above, a $N \times M$ matrix comprising M traces and N points per trace. $\mathbf{D}$ admits the Lanczos decomposition

$$\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{M} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right eigenvector matrices, respectively, and $\boldsymbol{\Sigma}$ is the matrix of singular values. In the summation representation, the σ_i are the individual, positive singular values and, in general, so ordered that $\sigma_1 > \sigma_2 > \cdots > \sigma_m$. Equation (5) represents a projection of $\mathbf{D}$ onto an orthonormal basis, specifically a basis composed of weighted rank one matrices $\mathbf{u}_i \mathbf{v}_i^T$, called eigenimages. Clearly, the contribution of a particular eigenimage in the reconstruction of $\mathbf{D}$ is proportional to the magnitude of the associated singular value. When the singular values are ordered in decreasing magnitude, it is possible, depending of course on the data, to reconstruct the matrix $\mathbf{D}$ using only the first few eigenimages. For this reason, $\mathbf{D}$ is preprocessed (by means of NMO) prior to transformation so that the signal is as horizontal as possible. As a result, the signal exhibits maximum coherence from trace to trace, and often the first few eigenimages will contain the separated signal. Random noise is dispersed equally among all the eigenimages and, thus, keeping only the first few eigenimages results in the attenuation of both $\mathbf{N}_r$ and $\mathbf{N}_c$ (see Freire and Ulrych, 1988, for details).

This process is very similar to that of stacking in the case where the noise is Gaussian and the signal is well-aligned. We believe, however that when this is not the case, eigenimage decomposition has an advantage since the mapping is based on the covariance matrix of the data (this also applies to a comparison of eigenimage with f-k filtering). The relationship to the covariance structure of the data is important, and a more detailed look is appropriate.

The covariance (or variance-covariance) matrix is a matrix which contains all the individual trace variances along the main diagonal and the covariances between traces in the off-diagonal elements. An inspection of the matrix $\mathbf{C}_D = \mathbf{D}^T\mathbf{D}$ will quickly convince the reader that $\mathbf{C}_D$ is a weighted estimate of the zero-lag covariance matrix of the data $\mathbf{D}$. Indeed, the first element of this matrix is $c_{11} = \sum_{i=1}^{N} d_{1i}^2$, just the weighted estimate of the variance of a zero-mean vector. Since it can be easily shown that the singular values σ_i in equation (5) are just the positive square roots of the eigenvalues of $\mathbf{C}_D$, the fact that the eigenimage decomposition of $\mathbf{D}$ is based on the covariance structure of the data becomes clear.

The difference between eigenimage decomposition and stacking is obvious when we consider AVO processing. In this case, stacking is forbidden, yet the first eigenimage attenuates noise as well as preserving the telltale amplitude variations. In fact, the first eigenimage may be visualized as a stack weighted with the right eigenvector that is destacked (a term introduced to us by Guus Berkhout, personal communication, 1997) with weights determined from the left eigenvector.

Another interesting aspect of the eigenimage decomposition is that, under certain circumstances, considerable data compression may be realized. This is particularly so when geology is fairly gentle, in which case common offset processing may result in better than 95% data compression.

In Figure 2, we illustrate the principle of eigenimage analysis with a very simple example: a synthetic section with Gaussian noise that exhibits an amplitude variation with offset (AVO) signature. Figure 2a is the noisy section with AVO. The first eigenimage is illustrated in Figure 2b, where the attenuation of the random noise is evident. The spectrum of singular values is shown in Figure 2d. The eigenvectors $\mathbf{u_1}$ and $\mathbf{v_1}$ in this example have a clear physical meaning. The eigenvector $\mathbf{v_1}$ represents the trace-to-trace amplitude variation or the AVO effect (Figure 2d), and the eigenvector $\mathbf{u_1}$ is a scaled estimate of the source wavelet (Figure 2e). The latter is true when the waveform is properly flattened. If this is not the case, more than one eigenimage will be required to properly model the waveform.

Karhunen-Loéve.—Although the eigenimage decomposition that we have described is also known in the literature as the Karhunen-Loéve decomposition, we would like to differentiate the two approaches. Eigenimages are based on the eigenvectors of the zero-lag 2-D data covariance matrix. The Karhunen-Loéve transform was introduced in order to decorrelate the expansion coefficients, a_i, in the decomposition

$$\mathbf{x} = \sum_{i=1}^{N} a_i \mathbf{v}_i, \tag{6}$$

where $\mathbf{v}_i$ are the eigenvectors associated with $\mathbf{R}$, the Toeplitz autocovariance matrix of the process $x(t)$. Since the eigenvectors form an orthonormal basis, the coefficients, a_i, are uncorrelated and may be determined from

$$a_i = \mathbf{v}_i^T \mathbf{x}. \tag{7}$$

As we have seen above, the eigenimage decomposition is computed from the zero-lag covariance matrix of a multichannel process. The Karhunen-Loéve decomposition, on the other hand, requires a covariance matrix, $\mathbf{R}$, that contains all required lags of the autocorrelation of the single channel process, $\mathbf{x}$. It is clear, in fact, that the eigenvectors that are required in the two approaches are quite different.

In the example illustrated here, we adapt the Karhunen-Loéve decomposition to a multichannel input by computing $\mathbf{R}$ as the average matrix for all traces. The filtering is accomplished by retaining only a subset of the computed eigenvectors. The results shown in Figure 3 consider the filtering of events with an AVO signature. Figures 3a and 3b show the input **S** and **D**, respectively. Figures 3c and 3d show the filtered output and the residual section. It is clear that, in this case, unlike the case of eigenimage filtering, NMO correction is not required prior to decomposition. The importance of the issue of correctly determining the number of eigenvectors required in the reconstruction (a particularly challenging task) is evident in Figure 3d, where a trace of the signal remains.

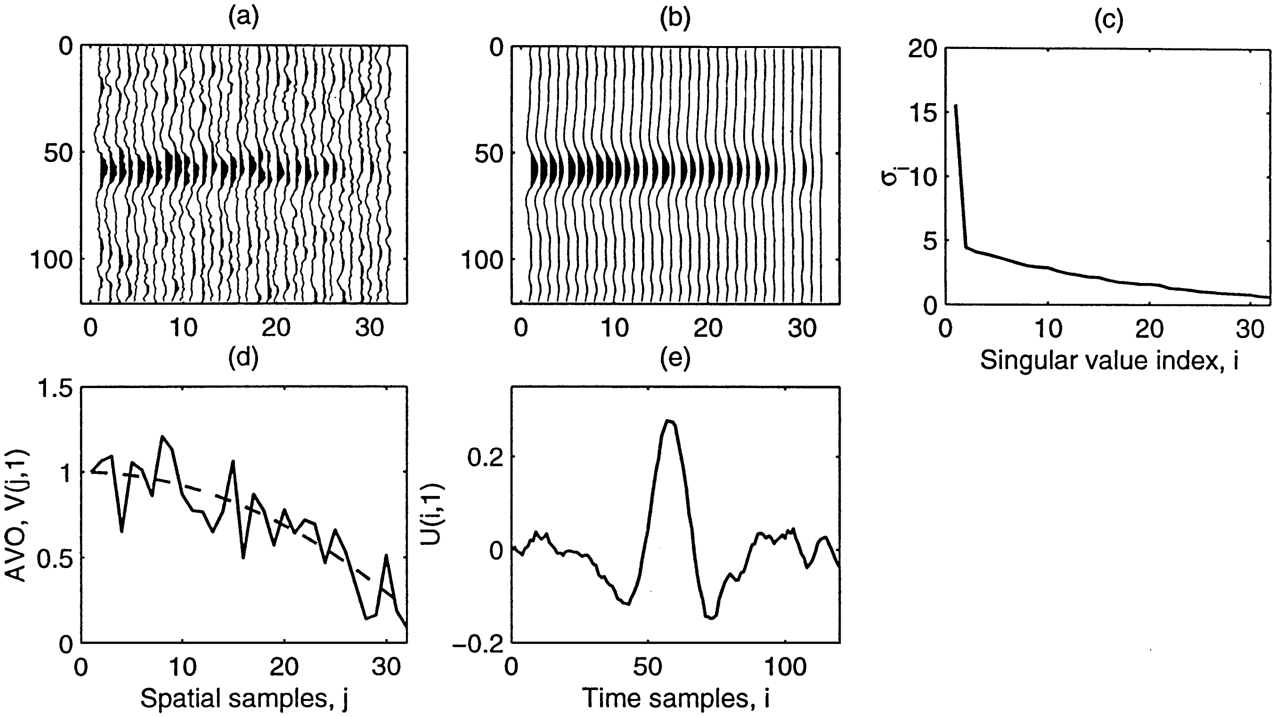

FIG. 2. Eigenimage decomposition: (a) input section with AVO signature, (b) the "clean image" computed by means of the first eigenimage, (c) distribution of singular values, (d) the first right eigenvector, $\mathbf{v}_1$, is an estimate of the scaled AVO effect, (e) the first left eigenvector, $\mathbf{u}_1$, is an estimate of the source wavelet.

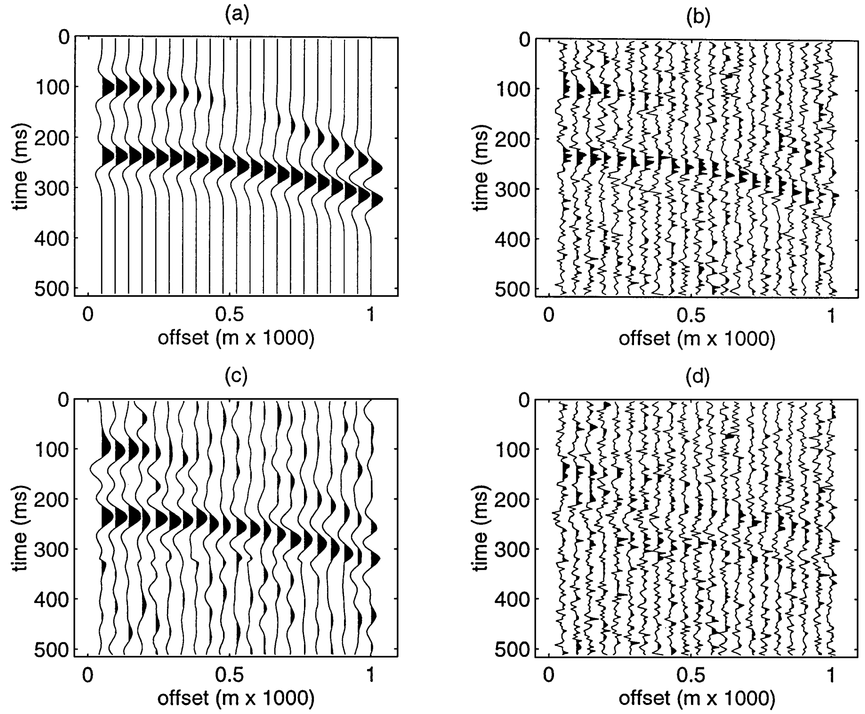

FIG. 3. Karhunen-Loéve filtering: (a) input signal, (b) input data, (c) Karhunen-Loéve filtered section, (d) residual section.

Radon transform

The signal in $\mathbf{D}$ can be approximated, to first order, by hyperbolic events. As such, we seek a mapping from x-t to a domain where a hyperbolic event will map into a point. In principle, the Radon hyperbolic transform will achieve this end. It turns out that the hyperbolic form of the Radon transform, compared to its parabolic cousin, is very much more time-consuming to implement due to the lack of time invariance. Specifically, whereas the parabolic transform (introduced by Hampson, 1986) is linear and allows the use of the fast Fourier transform (FFT), the hyperbolic transform is not, and no decoupling of frequency components would occur.

Therefore, one prefers to transform via the parabolic form after a correction to $\mathbf{D}$ such that hyperbolas become, more or less, parabolas. A little arithmetic is now required. We define the parabolic Radon transform as a linear transformation from the data space into the parabolic Radon domain for a particular frequency. In matrix notation we have

$$\tilde{\mathbf{m}} = \mathbf{L}\mathbf{d}, \tag{8}$$

where $\tilde{\mathbf{m}}$ and $\mathbf{d}$ are vectors that define the model space (the Radon domain) and the data (a common midpoint or shot gather), respectively. We can say that $\mathbf{L}$ is an operator that stacks the data along parabolic paths. We have placed a tilde (˜) over the $\mathbf{m}$ to indicate that the mapping according to equation (8) yields a low-resolution or smooth model. In the stacking process, reflections with different moveout should collapse into points in the Radon space, therefore individual events can be identified and, subsequently, removed. Unfortunately, it turns out that the operator $\mathbf{L}$ does not possess enough resolution to properly distinguish events with similar moveout. This is a simple consequence of the frequency bandwidth and the finite aperture in seismic acquisition.

Instead of using the operator $\mathbf{L}$ to map reflections into the parabolic Radon domain, we prefer to pose the mapping as an inverse problem where now, $\mathbf{d}$, the data, are viewed as the result of the transformation (Sacchi and Ulrych, 1995):

$$\mathbf{d} = \mathbf{L}'\mathbf{m}. \tag{9}$$

The operator $\mathbf{L}'$ is called the adjoint operator ($\mathbf{L}$ and $\mathbf{L}'$ define an adjoint pair). In other words, $\mathbf{L}'$ is an operator that maps a point in the parabolic Radon space into a data parabola. The advantages of using equation (9) over equation (8) are twofold. First, by posing our problem as an inverse problem, we can choose a strategy to enhance the resolution of the transformation (by resolution we mean the ability of the transformation to recognize events with similar moveout curves). Second, by selecting the proper regularization scheme, random noise can be attenuated. A word here about regularization. As we have pointed out, our inverse problem, when correctly posed, is underdetermined. The matrix $\mathbf{L}$ in equation (8) does not have an inverse. An infinity of solutions exist. Of course, one does not abandon the problem in despair. As is well known, a solution may indeed be found, one of the infinite solutions possible, by regularizing the problem. Essentially this consists of building a cost function that, upon optimization, will yield the sought after model parameters. The cost function, J, is in general made up of two parts and can be written (first, in words; we present a Bayesian cost function a little later) as

$$\mathbf{J} = \lambda(\text{model norm}) + (\text{data constraints}); \tag{10}$$

λ is what Bayesians call a hyperparameter. Since we are rather partial to Bayes, we will maintain this nomenclature here. A hyperparameter controls the inversion. Its value determines how well the data constraints are adhered to, at the same time minimizing the model norm.

As fledgling Bayesians, we pose our inverse problem as one of inference and build our cost function by allowing prior information to guide us to a solution that, while honoring the data, exhibits the characteristics that we require (more on this later). In order to incorporate our prior constraints, we use Bayes' theorem which, simply stated, is

$$p(\mathbf{m}\,|\,\mathbf{d}) \propto p(\mathbf{d}\,|\,\mathbf{m})p(\mathbf{m}), \tag{11}$$

where $p(\mathbf{m}\,|\,\mathbf{d})$ is the posterior probability of the model given the data, $p(\mathbf{d}\,|\,\mathbf{m})$ is the likelihood, and $p(\mathbf{m})$ is the prior probability of the model. As is common, we take the likelihood to be Gaussian. It is $p(\mathbf{m})$ that gives the solution that special flavor. We wish to impose sparseness or limited support, the characteristic we mentioned above, that will simulate an extended aperture and allow us to identify and, hopefully attenuate, $\mathbf{N}_c$. Choosing a Cauchy distribution for $p(\mathbf{m})$ (see Sacchi and Ulrych, 1995, for a rationale and details), we maximize $p(\mathbf{m}\,|\,\mathbf{d})$. This implies minimizing the cost function J which, taking the logarithm of $p(\mathbf{m}\,|\,\mathbf{d})$ and substituting the relevant Gaussian and Cauchy densities, is given by

$$J = \lambda \sum \ln\left(1 + \frac{m_i^2}{\sigma_c^2}\right) + (\mathbf{L}'\mathbf{m} - \mathbf{d})^T \mathbf{C}^{-1}(\mathbf{L}'\mathbf{m} - \mathbf{d}), \tag{12}$$

where λ and σ_c^2 are the hyperparameters we talked about and $\mathbf{C}$ is the data covariance matrix. Equation (12) is nonlinear and is solved iteratively.

It is clear that other regularization strategies can be chosen. A popular method to solve an inverse problem is by the introduction of a smoothing operator. In Bayesian terms, this is equivalent to adopting a Gaussian prior distribution for the unknown $\mathbf{m}$. Such a distribution is smooth in the sense that the probability of the model parameters being very different from each other is rather small. Smoothness also follows from the fact that the Gaussian distribution is neutral-tailed (unlike the Cauchy that is heavy-tailed). There is a 99.7% probability that an observation from this distribution will be within three standard deviations of the mean. The Gaussian prior leads to the well known "damped least-squares solution." In Figure 4, we examine the importance of choosing the proper prior information in the computation of the parabolic Radon transform. We consider a model that consists of a primary and multiple event, shown in the τ-q domain in Figure 4a (the primary is the first event), where q is the parameter associated with the parabolic equation (much as p is associated with the linear event in the linear τ-p Radon transform). This model is mapped into the x-t domain in Figure 4c to simulate two reflections with parabolic moveout. To make things more complicated, the data are unevenly sampled and severely aperture-limited, as shown in Figure 4b. Our task is summarized as follows: given the data in the x-t domain (Figure 4b), recover the τ-q model and, from the model, reconstruct the evenly sampled and "infinite" aperture data (Figure 4c). In Figure 4d, we show the τ-q model computed using the least-squares approach; in Figure 4g, we show the τ-q model retrieved using the Cauchy prior. It is clear that the Cauchy prior yields a result that is more consistent with the true model, whereas the least-squares approach fails to distinguish the positions of the two events. This is clearly a consequence of using a prior that smoothes our image, quite contrary to the Cauchy prior that is able to sharpen, or properly deblur, the image. Advantages of the sparse prior go further. We can use the inverted models (Figures 4d and 4g) to reconstruct the offset spaces shown in Figures 4e and 4h. In this case, as is emphasized by the residual sections in Figures 4f and 4i, the Cauchy solution has enabled us to better recover missing near and far offset traces.

The role of the prior model is central. It allows sparseness to be imposed on the solution in a manner similar to the well-known maximum entropy estimator (Burg, 1975; Ulrych and Bishop, 1975). In as much as the truncated input autocorrelation function is extended when the maximum entropy spectrum is Fourier transformed, so the aperture of the input data is extended by the inversion of the model in the τ-q domain.

Wavelet transform processing

The Fourier transform is an indispensable tool that we use in all aspects of data processing and analysis. This transformation uses a harmonic basis that is stationary in time. In effect, this transformation represents a point in time by the superposition of an infinite number of harmonics of infinite duration with an infinite range in scale. To localize events both in time and frequency, Gabor (1946) introduced the short time Fourier transform. Recently (Mallat, 1998), based on a synthesis of ideas from work in the 1920s and 1930s, a particularly flexible transform, the wavelet transform (WT), has seen much press and many applications. The WT is to us what binoculars are to bird watchers. It allows us to localize **D** both in time and in scale, thus enabling us to see details that the Fourier transform is unable to resolve. We apply the WT here in a particularly simple fashion.

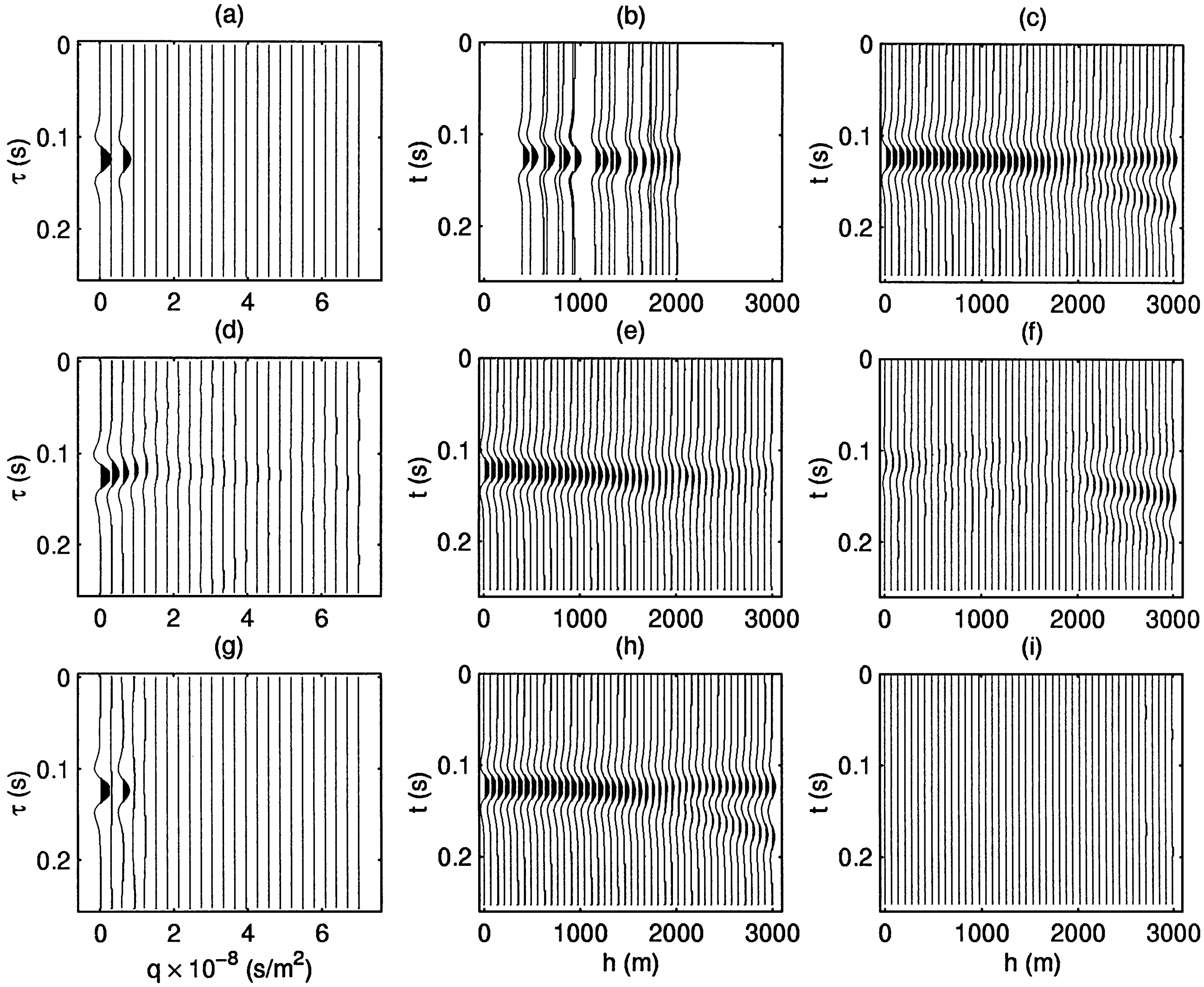

FIG. 4. Radon transform and multiple attenuation: (a) "ideal" τ-q map for a primary-multiple pair, (b) the input unevenly sampled and aperture-limited data, (c) the desired evenly sampled data, (d) the τ-q space computed via inversion using damped least-squares, (e) the reconstructed desired data, (f) the residual panel, (g) the τ-q space computed via inversion using the Cauchy prior, (h) the reconstructed desired data, (i) the residual panel.

A much more complete view of the many possibilities is seismic applications has been presented by Foster et al. (1994).

The first step is to choose an appropriate basis. In this case, we have chosen the D20 wavelet basis of Daubechies (see Mallat, 1998). We write this orthogonal basis as the dilated and translated family,

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t - 2^j n}{2^j}\right) \qquad j, n \in \mathcal{Z}, \tag{13}$$

where $\mathcal{Z}$ is the set of integers; j is called the scale factor, since $\psi_{j,n}(t)$ is of length 2^j; n is called the translation or shift factor and is scale dependent. The wavelet coefficients, $w(j, n)$, are computed in the usual manner by projecting the data, $d(t)$, onto the basis functions:

$$w(j, n) = \langle \psi_{j,n}(t), d(t)\rangle = \sum_t \psi_{j,n}(t)d(t). \tag{14}$$

The wavelet coefficients at the three lowest scales are normalized to $N(0, 1)$ by means of a robust estimate of the standard deviation and a threshold coefficient, ν, is estimated very simply (Donoho, 1994) as $\nu = \sqrt{2\log n_j}$, where n_j is the number of coefficients at scale j. The wavelet coefficients at scale j below the threshold ν are zeroed, and the filtered data are recovered by an inverse WT. The results for the same example that was used in illustrating Karhunen-Loéve filtering are shown in Figure 5. Figures 5a and 5b are equivalent to Figures 3a and 3b and are repeated for convenience. Figure 5c is very close to the result shown in Figure 3c. Differences may be seen in the residual panels, where the small residual signal remaining in the Karhunen-Loéve result is not evident in Figure 5d.

We see that this approach does not use the coherence of **S** as a criterion. It does, however, impose the criterion of smoothness. This naturally leads to a hybrid algorithm that extracts **S** by imposing maximum coherence in x and maximum smoothness in t. We are exploring such an approach at present.

Stein processing

We began this article with some comments regarding the basics of stacking. We end on the same note. This section is aimed, primarily, for cerebral stimulation and for a somewhat more in-depth discussion concerning risk. It illustrates our scientific quest: understand the past, implement the present, and explore the future. This is a bit of the future. Hopefully, applications to seismic processing will also emerge with time.

There is no doubt that, given a realization of M samples from $N(\mu, 1)$, the only admissible estimator of μ is δ^0 defined in equation (3). The question is, is this also true if more than one realization is available? An illuminating example, from the scientific discipline of baseball, that will serve to illustrate the discussion is given by Efron and Morris (1977). It goes like this. After the first 45 at bats in the 1970 season, Roberto Clemente obtained 18 hits; his average at that point in the season was, therefore, .400. Another great, Thurmon Munson, was in an early slump and managed only an average of .178. Faced with the question, what will the respective averages be at the end of the season, the only admissible answer is .400 and .178. Clemente and Munson, however, were batting in the majors with many other batters. James and Stein (1961) (J-S) in the early 1960s proved a very controversial theorem. An average estimated by δ^0 after 40 at bats, given that there are more than two other batters in the league, is not an admissible estimator. In other words, Clemente's and Munson's averages are better estimated taking into account what other batters are doing, providing that there are three or more batters. Put in a different but equivalent manner, for this case δ^0 is not the estimator with lowest risk. Let us pay some attention to this notion of risk.

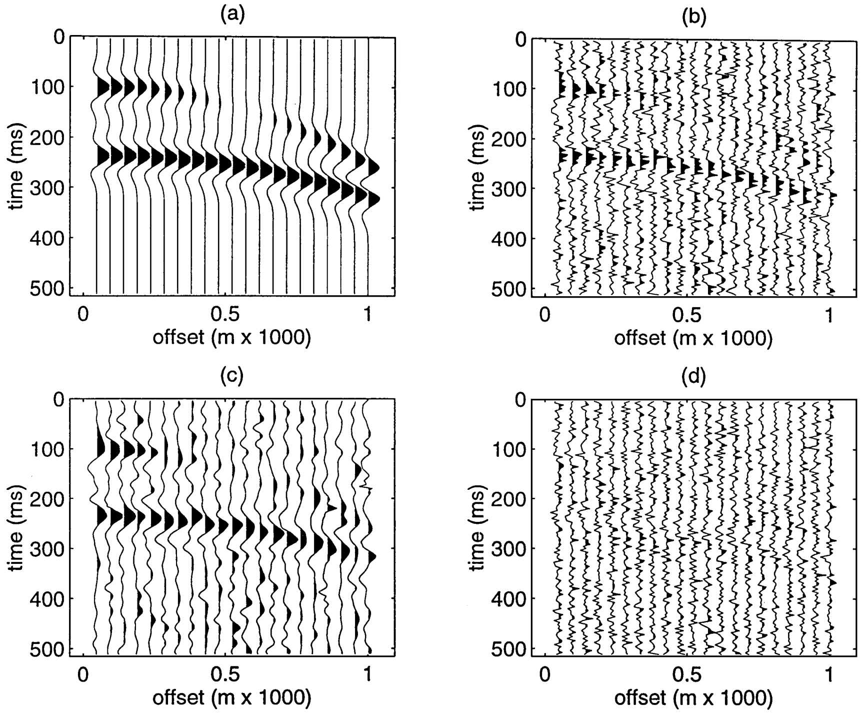

FIG. 5. Wavelet transform filtering: (a) input signal, (b) input data, (c) WT filtered section, (d) residual section.

Life, as some great sage once said, is an inverse problem. In particular it is an ill-posed inverse problem. Our task, it seems, is to try and find solutions which, from the infinity of possibilities, are ones that offer lowest risk when pursued. Statisticians use two terms when dealing with this subject: risk and efficiency. We have considered the maximum likelihood estimator for estimating baseball averages at the end of the season. The properties of $\boldsymbol{\delta}^0$ are well known. It is an unbiased (it is not partial to any preferred value) estimator of the population mean with variance σ^2/M, where σ^2 is the actual variance associated with the probability distribution function. The question we pose is whether there is not some other estimator that can obtain an estimate with lower variance (or with lower risk) than $\boldsymbol{\delta}^0$. Perhaps the median? It turns out that the variance of the sample median for a Gaussian distribution is $\pi\sigma^2/2M$. Since the efficiency of the estimator is defined in terms of the size of sample required to achieve a certain accuracy, the efficiency of the median compared to the mean is $2/\pi = 64\%$. In terms of risk, the risk of the median compared to that of the mean is the ratio of the respective variances, i.e., 1.57.

Let us now look at the J-S estimator. Consider, for a given μ_i,

$$\mathbf{x}_i \mid \mu_i \overset{\text{ind}}{\sim} N(\mu_i, 1) \qquad i = 1, 2, \ldots, \geq 3. \tag{15}$$

The variance is taken to be unity for convenience through an appropriate scale transformation. We wish to determine the unknown vector of means $\boldsymbol{\mu}^T = (\mu_1, \mu_2, \ldots, \mu_k)$ so as to minimize the loss, L, defined in the usual way as

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{k} (\hat{\mu}_i - \mu_i)^2, \tag{16}$$

where $\hat{\boldsymbol{\mu}}$ is the estimate of $\boldsymbol{\mu}$. We now formally define the risk, $R(\cdot, \cdot)$, for the MLE estimator as

$$R(\boldsymbol{\mu}, \boldsymbol{\delta}^0) = E_\mu\left[\sum_{i=1}^{k} \left(\boldsymbol{\delta}^0(x_i) - \mu_i\right)^2\right], \tag{17}$$

where $E_\mu[\cdot]$ is the expectation (the average) over the distribution in equation (15). It turns out that $R(\boldsymbol{\mu}, \boldsymbol{\delta}^0) = k$ for every value of $\boldsymbol{\mu}$. It is constant. This is the lowest risk that was considered possible prior to James and Stein. In 1961, James and Stein gave us a new estimator, $\boldsymbol{\delta}^1$ (for $k \geq 3$)—the J-S estimator, which is now much in use in finite population sampling. It is, for each score x_i,

$$\boldsymbol{\delta}^1(x_i) = \boldsymbol{\delta}^0(\mathbf{x}) + \kappa\left(x_i - \boldsymbol{\delta}^0(\mathbf{x})\right), \tag{18}$$

where κ is called the shrinkage factor and is computed as a function of the deviation of each score from the grand average $\boldsymbol{\delta}^0(\mathbf{x})$. For the normalized distribution of equation (15),

$$\kappa = 1 - \frac{(k-2)}{\sum_k \left(x_i - \boldsymbol{\delta}^0\right)^2}. \tag{19}$$

It turns out that in this case, the risk $R(\boldsymbol{\mu}, \boldsymbol{\delta}^1) < k$ for all values of $\mathbf{x}$.

What is this magical shrinkage factor κ? Efron and Morris (1977) describe it as follows. Designating $\sum_k (x_i - \boldsymbol{\delta}^0)^2 = S$ for convenience, we see that, as S decreases (in other words, as the deviation of each score from the grand average decreases), κ tends to 0 and the individual averages are shrunk towards the grand average. As the data deviations from the grand average increase on the other hand, S increases, κ tends to 1, and not much shrinkage ensues. The shrinkage is, therefore, controlled by the data themselves and by the initial J-S hypothesis that the individual averages are not far from the grand average.

Returning now to the baseball example, in order to compare the risks associated with $\boldsymbol{\delta}^0$ and $\boldsymbol{\delta}^1$, we need to know the actual batting potentials. Of course, these are forever unknown, but we can obtain a pretty good estimate by waiting to the end of the season. By this means, Efron and Morris (1977) show that the error in the maximum likelihood estimate is 3.5 times higher than that of the J-S estimate. Clemente's average is decreased to .294, to be compared with his season's end average of .348. Munson's average, on the other hand, is increased to a J-S estimate of .247, and compares with .320 at the end of the season. For 16 of the 18 players considered by Efron and Morris, the J-S estimated average after 45 at bats is closer to the "true" average than the conventionally estimated figure. Fascinating. Can we apply the Stein concept to seismic processing? Perhaps we can replace the batter with the seismic trace and compute a stack that is of lower risk. In fact, in shrinking x_i, the J-S estimator takes into account the effect of the variance associated with each batter and, in that sense, allows the fold to be incorporated into the calculation.

We are in the process of testing this method of stacking on seismic data. Specifically, our attempt is to improve the signal-to-noise ratio in $\mathbf{D}$ without stacking for the purpose of, say, AVO analysis. We use the J-S estimator as a preprocessing step that allows us to shrink the noise variance in the section. Each batter becomes a time point associated with a trace. We thus consider M batters for each of the N points in the NMO-corrected gather, compute the appropriate shrinkage locally, and then the new sample value.

Figure 6 illustrates a tentative result. Figure 6a is the simulated AVO, Figure 6b is Figure 6a with noise, and Figure 6c shows the result of Stein preprocessing. For comparison purposes, we have processed the same example using eigenimage decomposition. The result obtained by keeping only the first two eigenimages is shown in Figure 6d. Although the eigenimage result has a somewhat higher signal-to-noise ratio, the J-S estimator has considerably shrunk the noise variance in very inexpensive fashion.

Clearly, applying the J-S estimator locally as we have done has similarities to moving average filtering. Specifically, if $\kappa = 1$, $\boldsymbol{\delta}^1(x_i) = \boldsymbol{\delta}^0(\mathbf{x})$ and the J-S estimator is precisely the maximum likelihood average. Clearly, however, there are other ways of applying $\boldsymbol{\delta}^1$. Our aim in introducing the Stein estimator in this article is to suggest a possible new avenue for research. We do not mean to imply that such a procedure will be superior to, say, eigenimage decomposition or WT filtering. We think, however, that the mathematical justification of the J-S estimator and the fact that that it may be easily and inexpensively implemented makes it worthy of attention.

SUMMARY

There is no manner of increasing the information content of the recorded $\mathbf{D}$. All we can hope to do is to expose $\mathbf{S}$ by attenuating $\mathbf{N}$. We attempt to do this by mapping $\mathbf{D}$ into a domain where, because of the different characteristics of $\mathbf{S}$, $\mathbf{N}_c$, and $\mathbf{N}_r$ (such as coherence, bandwidth, fractal dimension, scale etc.), we can separate these quantities into distinct subspaces and thereby attenuate the undesired components. In all cases, the forward and inverse mappings, derived on the basis of scientific

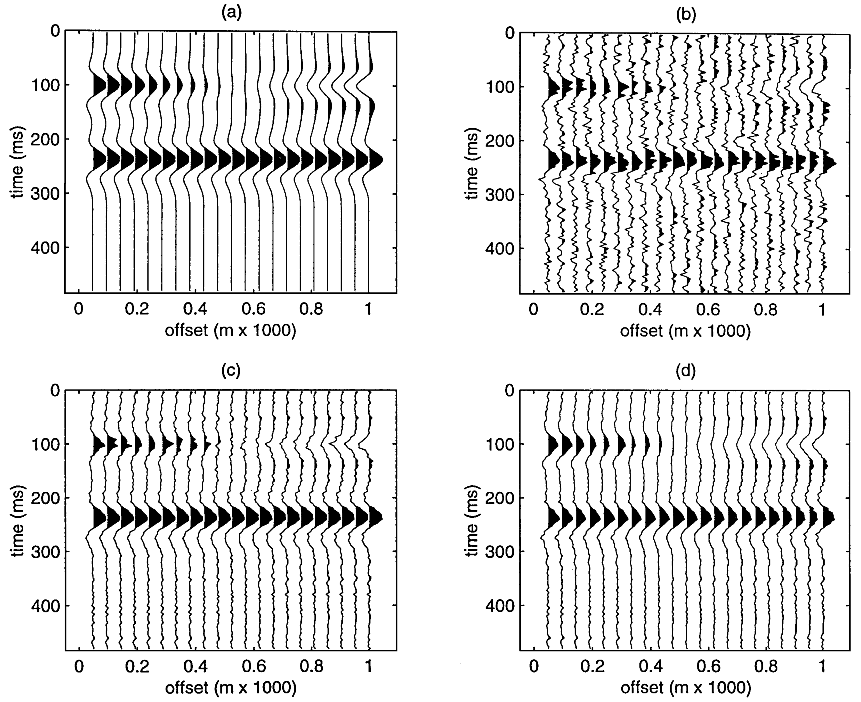

FIG. 6. Stein processing: (a) input signal, (b) input data, (c) stein processed section, (d) eigenimage processed section.

principles, require some parameters to make them work. Setting these parameters is, in general, an art. It is these details that, according to Arthur Weglein (personal communication, 1985), seldom see the light of publication.

We have briefly explored various techniques of signal-to-noise enhancement. Stacking is one such mapping that, however, is not always desirable, since it may destroy important signal attributes. Eigenimage and Karhunen-Loéve decomposition, Radon transformation, wavelet thresholding, and Stein processing can lead to signal and noise separation without, we hope, signal distortion. In such manner we may, perhaps, expose **S**, our David, without, we fervently hope, by so doing damaging **E** that gave birth to **D** in the first place.

REFERENCES

Burg, J. P., 1975, Maximum entropy spectral analysis: Ph.D. thesis, Stanford Univ.

Donoho, D. L., 1994, De-noising by soft-thresholding: manuscript, Department of Statistics, Stanford Univ. Available at www.stat.stanford.edu/~donoho/Reports/index.html.

Efron, B., and Morris, C., 1977, Stein's paradox in statistics: Scientific American, May, 119–128.

Foster, D. J., Mosher, C. C., and Hassanzadeh, S., 1994, Wavelet transform methods for geophysical applications: 64th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1465–1468.

Freire, S. L., and Ulrych, T. J., 1988, Application of singular value decomposition to vertical seismic profiling: Geophysics, **53**, 778–785.

Gabor, D., 1946, Theory of communication: J. IEEE, **93**, 429–457.

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

James, W., and Stein, C., 1961, Estimation with quadratic loss: Proc. 4th Berkeley symp. math. stat. and prob., **1**, 361–379.

Kleiner, B., and Graedel, T. E., 1980, Exploratory data analysis in the geophysical sciences: Rev. Geophys. Space Phys., **18**, 699–717.

Mallat, S., 1998, A wavelet tour of signal processing: Academic Press.

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

Schoenberger, M., 1996, Optimum weighted stack for multiple suppression: Geophysics, **61**, 891–901.

Ulrych, T. J., and Bishop, T. N. 1975, Maximum entropy spectral analysis and autoregressive decomposition: Rev. Geophys. Space Phys., **13**, 183–200.

Ulrych, T. J., Velis, D. R., Woodbury, A.D., and Sacchi, M. D., 1999, L-moments and C-moments: Stoch. Envir. Res. and Risk Assessment, accepted for publication.