Robust Wavelet Denoising

Sylvain Sardy, Paul Tseng, and Andrew Bruce

***Abstract*—For extracting a signal from noisy data, waveshrink and basis pursuit are powerful tools both from an empirical and asymptotic point of view. They are especially efficient at estimating spatially inhomogeneous signals when the noise is Gaussian. Their performance is altered when the noise has a long tail distribution, for instance, when outliers are present.**

We propose a robust wavelet-based estimator using a robust loss function. This entails solving a nontrivial optimization problem and appropriately choosing the smoothing and robustness parameters. We illustrate the advantage of the robust wavelet denoising procedure on simulated and real data.

***Index Terms*—Basis pursuit, block coordinate relaxation, interior point, robustness, wavelet, waveshrink.**

I. Introduction

SUPPOSE we observe a signal $\underline{s} = (s_1, s_2, \cdots, s_N)$ generated from

$$s_n = f(x_n) + \sigma z_n, \qquad n = 1, \cdots, N \tag{1}$$

where the equally spaced sampling locations x_n are points on the line for one dimensional (1-D) signals or on a grid for images. For now, we assume that the z_ns are identically and independently distributed Gaussian random variables with mean zero and variance one. Our goal is to denoise the signal $\underline{s}$, i.e., to find a good estimate $\hat{\underline{f}}$ of the underlying signal $\underline{f} = (f(x_1), \cdots, f(x_N))$. The hat on top of a letter is the notation used throughout this paper to indicate the estimate of the corresponding parameter. Waveshrink [1] and basis pursuit [2] are two nonparametric expansion based estimators. They assume that f can be well represented by a linear combination of P wavelet basis functions ϕ_p, namely

$$f(x) \approx \sum_{p=1}^{P} \alpha_p \phi_p(x) \tag{2}$$

where $\underline{\alpha} = (\alpha_1, \cdots, \alpha_P)$ are the wavelet coefficients. Waveshrink is defined for orthonormal wavelets only (i.e., $P = N$), whereas basis pursuit can also use an "overcomplete" basis (i.e., $P \gg N$). The advantage of an overcomplete wavelet dictionary is discussed by Chen *et al.* [2]. The goal of Waveshrink and basis pursuit is to estimate the wavelet coefficients for $\hat{\underline{f}} = \Phi\hat{\underline{\alpha}}$ to have a good mean squared error

$$\mathrm{MSE}(\hat{\underline{f}}, \underline{f}) = \frac{1}{N}\mathrm{E}\|\hat{\underline{f}} - \underline{f}\|_2^2$$

where the expectation is taken over $\underline{s}$.

Waveshrink uses orthonormal wavelets, which has two important consequences: First, the least squares estimate is simply $\hat{\underline{\alpha}}^{\mathrm{LS}} = \Phi'\underline{s}$, where Φ is the matrix of discretized ϕ_p, and Φ' denotes the transpose of Φ; second, $\hat{\underline{\alpha}}^{\mathrm{LS}}$ is an unbiased estimate of $\underline{\alpha}$, and its covariance matrix is $\sigma^2 I$ so that the estimated least squares coefficients are independent if the noise is Gaussian. For a smaller mean squared error at the cost of introducing some bias, Donoho and Johnstone [1] apply the hard or the soft function

$$\begin{aligned} \eta_\lambda^{\mathrm{hard}}(x) &= x \cdot 1(|x| > \lambda) \\ \eta_\lambda^{\mathrm{soft}}(x) &= \mathrm{sign}(x) \cdot (|x| - \lambda)_+ \end{aligned} \tag{3}$$

where $1(x \in A)$ is the identity function on A, and where x_+ is x for $x > 0$ and zero otherwise. For Gaussian noise, the shrinkage can be applied to $\hat{\underline{\alpha}}^{\mathrm{LS}}$ component-wise because its components are independent.

The hard and soft estimates are, interestingly, the closed-form solution to two optimization problems that are, in general, difficult to solve unless Φ is orthonormal.

- *Best Subset:* $\eta_\lambda^{\mathrm{hard}}(\hat{\underline{\alpha}}^{\mathrm{LS}})$ is "the best subset of size k" with λ as the $(N-k)$th smallest element of $|\hat{\underline{\alpha}}^{\mathrm{LS}}|$ in the sense that it minimizes the residual sum of squares among all sets with k nonzero wavelet coefficients.
- *l_1-Penalized Least Squares:* $\eta_\lambda^{\mathrm{soft}}(\hat{\underline{\alpha}}^{\mathrm{LS}})$ is the closed-form solution to the following optimization problem:

$$\min_{\underline{\alpha}} \tfrac{1}{2}\|\underline{s} - \Phi\underline{\alpha}\|_2^2 + \lambda\|\underline{\alpha}\|_1. \tag{4}$$

  This property leads to the relaxation algorithm of Section II-A and to the definition of basis pursuit.

When Φ is no longer orthonormal but overcomplete, the least squares estimate no longer has independent components, and the shrinkage idea cannot be applied as such. Basis pursuit generalizes to that situation using the optimization problem formulation (4), whose solution is not trivial when Φ is not orthonormal.

The selection of the smoothing parameter λ is important. Several ways of selecting λ have been proposed for Gaussian noise. They are based on a minimax argument (see Donoho and Johnstone [1] for real-valued noise and Sardy [3] for complex-valued noise) or on minimizing the Stein unbiased risk estimate (SURE) (see Donoho and Johnstone [4]). Nason [5] selects the smoothing parameter by cross validation.

The predictive performance of waveshrink and basis pursuit deteriorates when the noise is not Gaussian. It is because of the l_2 loss function in (4). It arises naturally as the log-likelihood

Manuscript received February 1, 2000; revised February 22, 2001. This work was supported in part by an SBIR Phase I contract with the Naval Air Warfare Center at China Lake, CA. The associate editor coordinating the review of this paper and approving it for publication was Prof. Paulo S. R. Diniz.

S. Sardy is with the Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland.

P. Tseng is with the Department of Mathematics, University of Washington, Seattle, WA 98195 USA.

A. G. Bruce is with the MathSoft, Inc., Seattle, WA 98109 USA.

Publisher Item Identifier S 1053-587X(01)03879-X.

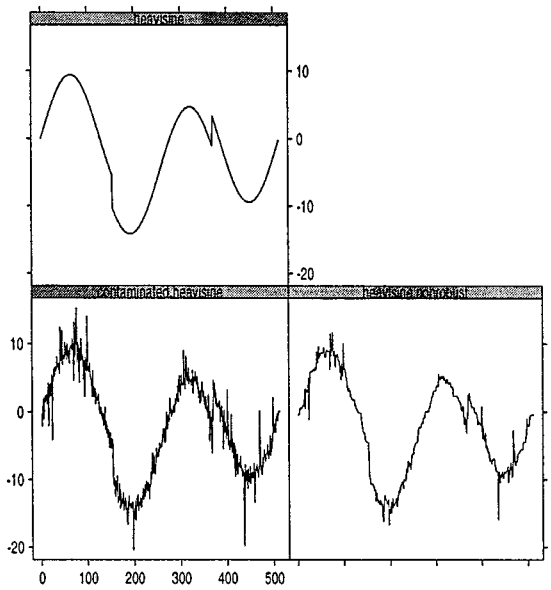

Fig. 1. Nonrobust estimation of *heavisine.* (Top) True signal. (Bottom left) Noisy signal. (Bottom right) Nonrobust estimate.

when the noise is Gaussian, but it is no longer appropriate when the departure from Gaussianity is too strong. In such a case, the quadratic loss function pulls the estimated function toward the outliers. We illustrate this phenomenon on a 1-D signal in Fig. 1; the true and noisy signal (90% standard Gaussian noise and 10% Gaussian noise with a standard deviation of 4) are plotted on the left side; on the right side, basis pursuit gives poor estimation near the outliers. The aim of this paper is to develop a robust wavelet-based estimator that is less affected by a long-tailed noise.

Some work has already been done in this direction. Bruce *et al.* [6] preprocess the estimation of the wavelet coefficients by a "fast and robust smooth/cleaner" at each multiresolution level to downweight the effect of the outliers in the estimation of the wavelet coefficients. Kovac and Silverman [7] preprocess the original signal to remove "bad" observations by means of a rough statistical test involving a running median smoother with a window of, say, 5; their procedure has the drawback of losing information by throwing out "bad" observations. For the situation of a *known* symmetric long tail noise, Averkamp and Houdré [8] derive minimax rules to select the smoothing parameter. Krim and Schick [9] derive a robust estimator of the wavelet coefficients based on minimax description length; their assumption of independent noise in the wavelet domain is not realistic, however.

In this paper, we propose a different approach that has the advantages of having a simple definition, of assuming a realistic independent contamination in the measurements $\underline{s}$, and of being able to deal with overcomplete wavelets as well. Its challenges are in finding an efficient algorithm (see Section II) and in choosing appropriately two tuning parameters (see Section III). Since the nonrobust behavior is due to the l_2 loss function, we simply replace it by a robust loss function ρ and define the coefficient estimate $\hat{\underline{\alpha}}_\lambda$ of the robust wavelet denoising procedure as the solution to

$$\min_{\underline{\alpha}} ||\underline{s} - \Phi\underline{\alpha}||_\rho + \lambda||\underline{\alpha}||_1 \tag{5}$$

where $||\underline{w}||_\rho = \sum_{n=1}^N \rho(w_n)$. We use the Huber loss function [10], which is a hybrid between l_2 for small residuals and l_1 for large residuals, namely

$$\rho(w) = \begin{cases} w^2/2, & |w| \le \tau \\ \tau \cdot |w| - \tau^2/2, & |w| > \tau \end{cases} \tag{6}$$

where $\tau > 0$ is some cutpoint. Using an even loss function, we implicitly assume that the noise is symmetric around zero. Note that when $\tau \longrightarrow 0$ and $\tau \longrightarrow \infty$, it becomes the l_1 and l_2 loss functions, respectively, both of which have the advantage of being convex.

Our proposal of a robust wavelet denoising procedure raises two issues. On the one hand, we must solve the nontrivial optimization problem defined in (5) for a given pair (λ, τ); on the other hand, we must select the smoothing parameter λ and the cutpoint τ. In Section II, we propose two algorithms to solve the robust wavelet denoising optimization problem (5): a block coordinate relaxation algorithm and an interior point algorithm. In Section III, we discuss the problem of selecting the cutpoint τ and the smoothing parameter λ. In Section IV, we give the result of a simulation to compare the efficiency of the two algorithms and to compare the denoising performance of the robust versus the nonrobust estimators. In Section V, we illustrate robust basis pursuit on two real data sets. We conclude the paper in Section VI.

II. Two Optimization Algorithms

A. Block Coordinate Relaxation (BCR) Algorithm

The BCR algorithm relies on the following observation. The Huber loss function (6) may be rewritten as

$$\rho(w) = \min_{w_a + w_b = w} w_a^2/2 + \tau|w_b|. \tag{7}$$

This nontrivial fact can be inferred from results on infimal convolution as discussed in Rockafellar [11, ch. 16]. Interestingly, (5) becomes

$$\min_{\underline{\alpha}, \underline{w}_b} \tfrac{1}{2}||\underline{s} - (\Phi\underline{\alpha} + \underline{w}_b)||_2^2 + \tau||\underline{w}_b||_1 + \lambda||\underline{\alpha}||_1. \tag{8}$$

The reformulated problem (8) has a separable structure to which the BCR algorithm of Sardy *et al.* [12] can be applied. It solves exactly a succession of subproblems using the soft shrinkage function (3).

BCR Algorithm:

1) Choose an initial guess for $\underline{\beta} = (\underline{\alpha}, \underline{w}_b)$, e.g., $\underline{\beta} = \underline{0}$;
2) Partition $B = [\Phi, I]$ into two matrices: $B^\perp$ of N orthonormal columns; $\overline{B}$ of the remaining $P - N$ columns. Define $\underline{\beta}^\perp$ and $\overline{\beta}$ as the corresponding coefficients in $\underline{\beta}$;
3) Define the residual vector $\underline{v} = \underline{s} - B\underline{\beta}$. Find the improved $\underline{\beta}^\perp$ by solving the subproblem

$$\underline{\beta}^\perp = \arg \min_{\underline{b} \in \Re^N} \tfrac{1}{2}||\underline{v} - B^\perp\underline{b}||_2^2 + \lambda||\underline{b}||_1$$

using soft shrinkage;
4) If convergence criterion is not met, go to step 1;

The BCR algorithm assumes that the matrix (here $[\Phi, I]$) is the union of a finite number of orthonormal matrices $B^{\perp}$. This assumption is verified for many wavelet dictionaries including nondecimated wavelets, wavelet packets, local cosine packets, chirplets [13], and brushlets [14]. Sardy *et al.* [12] propose two strategies for choosing $B^{\perp}$ in step 2 of the BCR algorithm and prove convergence for real and complex-valued signals.

B. Interior Point (IP) Algorithm

The interior point algorithm has the advantage of not requiring Φ to be the union of orthonormal blocks. It does not apply to complex-valued signals, however, and is computationally less efficient than the BCR algorithm (see Section IV-A).

1) Transformation to Quadratic Programming: First, we rewrite the optimization problem (5) as

$$\min_{\underline{\alpha},\underline{w}} \sum_n \rho(w_n) + \lambda \|\underline{\alpha}\|_1 \quad \text{subject to} \quad \Phi\underline{\alpha} + \underline{w} = \underline{s}. \tag{9}$$

By attaching Lagrange multipliers to the linear constraints, this, in turn, can be written as

$$\min_{\underline{\alpha},\underline{w}} \max_{\underline{y}} \sum_n \rho(w_n) + \lambda \sum_p |\alpha_p| + (\underline{s} - \Phi\underline{\alpha} - \underline{w})'\underline{y}.$$

The dual to this problem, which is obtained by exchanging the order of "min" and "max," is

$$\max_{\underline{y}} \left\{ \sum_n \min_{w_n} [\rho(w_n) - w_n y_n] + \sum_p \min_{\alpha_p} [\lambda|\alpha_p| - \alpha_p(\underline{\Phi}_p'\underline{y})] \right\} + \underline{s}'\underline{y} \tag{10}$$

where $\underline{\Phi}_p$ is the pth column of Φ.

Since the objective function in (9) is convex, defined everywhere, and the constraint is linear, it is known from convex duality theory (see, e.g., Rockaffelar [11, th. 28.2 and 28.4]) that the duality gap between the primal (9) and the dual (10) problems is zero. Using (6) and some algebra, the dual problem (10) is

$$\min_{\underline{y}} \sum_n \tfrac{1}{2} y_n^2 - \underline{s}'\underline{y} \quad \text{with} \quad \begin{cases} -\tau \le y_n \le \tau \\ -\lambda_p \le \underline{\Phi}_p'\underline{y} \le \lambda_p. \end{cases} \tag{11}$$

This is a quadratic programming problem. Notice that in the case of $\tau = 0$, the dual problem (11) is a linear programming problem, implying that the primal problem can be transformed into a linear programming problem. For brevity, we omit the derivation (see Sardy [15]).

2) Interior Point Algorithm: We solve the quadratic programming problem using a primal-dual log-barrier interior point algorithm inspired by Chen *et al.* [2]. The log-barrier subproblem corresponding to (11) is

$$\min_{\underline{y}} \sum_n \tfrac{1}{2} y_n^2 - \underline{s}'\underline{y} - \mu \sum_n \log(\tau - y_n) - \mu \sum_n \log(\tau + y_n) - \mu \sum_p \log(\lambda - \underline{\Phi}_p'\underline{y}) - \mu \sum_p \log(\lambda + \underline{\Phi}_p'\underline{y})$$

where μ is the log-barrier penalty that is chosen identically for all the penalty terms. Letting $\underline{e}^n$ be the nth canonical basis vector, the first-order optimality condition for the log-barrier subproblem is

$$\underline{y} - \underline{s} + \mu \sum_n \frac{1}{\tau - y_n}\underline{e}^n - \mu \sum_n \frac{1}{\tau + y_n}\underline{e}^n + \mu \sum_p \frac{1}{\lambda - \underline{\Phi}_p'\underline{y}}\underline{\Phi}_p - \mu \sum_p \frac{1}{\lambda + \underline{\Phi}_p'\underline{y}}\underline{\Phi}_p = \underline{0}.$$

Letting

$$\begin{aligned} u_n^+ &= \tau - y_n, \quad r_n^+ = \mu/(\tau - y_n) \\ u_n^- &= \tau + y_n, \quad r_n^- = \mu/(\tau + y_n) \\ v_p^+ &= \lambda - \underline{\Phi}_p'\underline{y}, \quad t_p^+ = \mu/(\lambda - \underline{\Phi}_p'\underline{y}) \\ v_p^- &= \lambda + \underline{\Phi}_p'\underline{y}, \quad t_p^- = \mu/(\lambda + \underline{\Phi}_p'\underline{y}) \end{aligned}$$

and letting

$$\underline{z} = (\underline{u}^+, \underline{u}^-, \underline{v}^+, \underline{v}^-), \quad \underline{x} = (\underline{r}^+, \underline{r}^-, \underline{t}^+, \underline{t}^-)$$
$$A = [I, -I, \Phi, -\Phi] \quad \text{and} \quad \underline{c} = (\tau\underline{1}, \lambda\underline{1})$$

the first-order optimality condition is a set of nonlinear equations

$$\begin{aligned} -A'\underline{y} - \underline{z} + \underline{c} &=: \underline{r}_x = \underline{0} \\ \underline{s} - A\underline{x} - \underline{y} &=: \underline{r}_y = \underline{0} \\ \mu\underline{1} - XZ\underline{1} &=: \underline{r}_z = \underline{0} \end{aligned} \tag{12}$$

where $X = \text{diag}(\underline{x})$ and $Z = \text{diag}(\underline{z})$ with $\underline{x} > 0$ and $\underline{z} > 0$. The variables $\underline{x}$, $\underline{y}$, and $\underline{z}$ are called, respectively, the primal, the dual, and the dual slack variables. This IP problem could alternatively have been derived from (8) and its dual; for instance, $\underline{r}^+ - \underline{r}^-$ corresponds to $\underline{w}_b$, and $\underline{y}$ corresponds to $\underline{w}_a$ in (7).

In an interior point approach, one typically takes a single Newton step to solve the nonlinear system (12) inexactly and then decreases μ. More precisely, given $\mu > 0$, $\underline{x} > 0$, $\underline{y}$, and $\underline{z} > 0$, one computes the Newton direction $(\Delta\underline{x}, \Delta\underline{y}, \Delta\underline{z})$, which is obtained by solving the following system of linear equations:

$$\begin{aligned} A'\Delta\underline{y} + \Delta\underline{z} &= \underline{r}_x \\ A\Delta\underline{x} + \Delta\underline{y} &= \underline{r}_y \\ Z\Delta\underline{x} + X\Delta\underline{z} &= \underline{r}_z \end{aligned} \tag{13}$$

and then, one updates the variables according to

$$\underline{x}^{new} = \underline{x} + \gamma\beta_x\Delta\underline{x} \tag{14}$$
$$\underline{y}^{new} = \underline{y} + \gamma\beta_z\Delta\underline{y} \tag{15}$$
$$\underline{z}^{new} = \underline{z} + \gamma\beta_z\Delta\underline{z} \tag{16}$$

where $\gamma \in (0,1)$, $\beta_x > 0$, and $\beta_z > 0$ are chosen to maintain $\underline{x}^{new} > 0$ and $\underline{z}^{new} > 0$. A popular choice is

$$\beta_x = \min_{p:\Delta x_p < 0} \{-x_p/\Delta x_p\} \tag{17}$$
$$\beta_z = \min_{p:\Delta z_p < 0} \{-z_p/\Delta z_p\} \tag{18}$$

and empirically, the choice of $\gamma = .99$ has worked well. The parameter μ may be updated in many ways. For example, Chen *et al.* [2] suggested $\mu^{new} = (1 - \min(\gamma, \beta_x, \beta_z))\mu$. Typically, only a small number of interior-point iterations is required to obtain a solution of desired accuracy.

3) Conjugate Gradient Solver for the Newton Step: Most of the computational effort is spent in computing the Newton direction at each iteration. From (13), we have that $\Delta\underline{y}$ is the solution of

$$(I + ADA')\Delta\underline{y} = (\underline{r}_y - A(Z^{-1}\underline{r}_z - D\underline{r}_x)) \qquad (19)$$

where $D = Z^{-1}X$ is a diagonal matrix. The dual slack and primal Newton directions are then obtained by $\Delta\underline{z} = \underline{r}_x - A'\Delta\underline{y}$ and $\Delta\underline{x} = Z^{-1}(\underline{r}_z - X\Delta\underline{z})$. We adopt the algorithm of Chen *et al.* [2] and use the conjugate gradient method to solve the dense $N \times N$ system (19). Because multiplication by A and A' are typically fast (on the order of $N \log N$ or $N(\log N)^2$ operations), the conjugate gradient method is attractive. In practice, however, the number of conjugate gradient iterations required to solve (19) accurately can become very large as $(\underline{x}, \underline{y}, \underline{z})$ approaches a solution, thus degrading the performance of the IP algorithm.

4) Finding an Initial Point: The IP algorithm requires an initial point $(\underline{x}^0, \underline{y}^0, \underline{z}^0)$ satisfying $\underline{x}^0 > 0$ and $\underline{z}^0 > 0$, which ideally would not be too far from the solution. Let the ridge regression estimate $\underline{\alpha} = \hat{\underline{\alpha}}_\lambda^{\text{ridge}}$ (obtained by replacing $||\underline{\alpha}||_1$ in (4) by $||\underline{\alpha}||_2^2$) be an initial guess for the coefficients. Let $\underline{\alpha}_+ = \max(\underline{\alpha}, \underline{0})$ and $\underline{\alpha}_- = \max(-\underline{\alpha}, \underline{0})$. With $\hat{\underline{s}} = \Phi\underline{\alpha}$ and $\underline{r} = \underline{s} - \hat{\underline{s}}$, let $\underline{r}_+ = \max(\underline{r}, \underline{0})$ and $\underline{r}_- = \max(-\underline{r}, \underline{0})$. Then, the primal variables $\underline{x}^0 = (\underline{r}_+, \underline{r}_-, \underline{\alpha}_+, \underline{\alpha}_-) + .1\underline{1}$ are positive. In addition, let $\underline{y} = A \cdot \text{sign}((\underline{r}_+, \underline{r}_-, \underline{\alpha}_+, \underline{\alpha}_-))$, and let $\overline{u} = 1.1||(\Phi'\underline{y}, \underline{y})||_\infty$. Then, the dual variables $\underline{y}^0 = \min(\lambda, \tau)\underline{y}/\overline{u}$ satisfy $-\lambda\underline{1} < \Phi'\underline{y}^0 < \lambda\underline{1}$, $-\tau\underline{1} < \underline{y}^0 < \tau\underline{1}$, and the dual slack variables $\underline{z}^0 = \underline{c} - A'\underline{y}^0$ are positive.

5) Convergence: Although there have been many convergence studies of IP algorithms, the algorithms that work well in practice, including the one described above, often have no guarantee of convergence. Specifically, convergence requires the existence of positive constants τ_1, τ_2, τ_3 such that $||\underline{r}_x||_2 + ||\underline{r}_y||_2 \le \tau_3\mu$ and $\tau_1\mu \le x_p z_p \le \tau_2\mu$, for $p = 1, \cdots, 2P$, at all iterations. We can enforce convergence by updating μ in a more conservative manner, but this would slow down its convergence in practice. (See, e.g., Kojima *et al.* [16] for discussions of these issues in linear programming problems.)

A stopping rule for the IP algorithm is when all of the following conditions are satisfied for a small $\epsilon_1 > 0$:

$$\begin{aligned} \text{Primal feasibility:} \quad & ||\underline{r}_y||_2/(1 + ||\underline{x}||_2) < \epsilon_1 \\ \text{Dual feasibility:} \quad & ||\underline{r}_x||_2/(1 + ||\underline{y}||_2) < \epsilon_1 \\ \text{Duality gap:} \quad & \underline{z}'\underline{x}/(1 + ||\underline{x}||_2||\underline{y}||_2) < \epsilon_1. \end{aligned} \qquad (20)$$

III. Selection of λ and τ

The selection of the smoothing parameter λ and the cutpoint τ is a difficult problem. Two different situations can be distinguished.

In one situation, the pair (τ, λ) can be tuned "by eye." For instance, in the first application of Section V, the noise is non-Gaussian, the signal to recover is known to be an aircraft. and the signal-to-noise ratio (SNR) is inherent to the infrared sensor used. In that situation, the smoothing parameter λ and the cutpoint τ can be tuned on a training set of images and then used on future images.

In the other situation, the underlying signal is not known, and neither is the SNR; therefore, an automatic selection of the pair (τ, λ) is needed. Our radar application of Section V is an example of this situation. Several procedures have been developed to select the smoothing parameter λ for non-Gaussian noise. Nason [5] observes, on a simulation using i.i.d., Student t_3 noise, that the l_2-based cross validation gives a better prediction than the minimax or SURE rules derived for Gaussian noise. With smoothing splines, Xiang and Wahba [17] develop a generalized cross validation criterion for a differentiable smoothness penalty and for a known noise distribution in the exponential family. The knowledge of a specific noise distribution is also required by Averkamp and Houdré [8], who develop a minimax rule for specific long tail noise distributions; their selection can only do so much to cope with the problem of using the unnatural l_2 loss function. Crouse *et al.* [18] propose wavelet-Markov models for the dependence between scales, and they employ an EM algorithm to estimate their parameters by maximum likelihood. EM is typically slow and gets trapped into local maxima, however [this cannot happen with our convex cost function (5)].

In this paper, we propose a pragmatic approach that does not require the specific knowledge of the noise distribution. First, an estimate of scale $\hat{\sigma}$ is required. We use the median absolute deviation of the high-frequency wavelet coefficients of Donoho and Johnstone [4]. This estimate of scale is robust to features of the underlying signal and to outliers especially if, as suggested by Kovac and Silverman [7], the Haar wavelet (which support has length two) is used. For the cutpoint τ in (6), we follow Huber [10] and choose $\hat{\tau} = c\hat{\sigma}$. The default in software packages is often $c = 1.345$ based on the following statistical consideration: Suppose you observe i.i.d. data with mean μ. Then, the asymptotic relative efficiency of the Huber estimate of μ is asymptotically 95% efficient with respect to the sample average when the noise is Gaussian (G). We can read this value on the continuous curve (G) of Fig. 2. The relative efficiency is also plotted as a function of c for the distributions used in the simulation. Fig. 2 gives a guideline for the selection of c in our nonparametric regression problem: We see that a value of c between one and three gives overall a good efficiency. Based on the simulation of Section IV, we recommend using a cutpoint of at least $c = 2.0$. Finally, for the smoothing parameter λ, we use the minimax λ_N^* developed by Donoho and Johnstone [1] since the residuals within $\pm\hat{\tau}\hat{\sigma}$ do not depart dramatically from Gaussian residuals.

IV. Simulations

A. Computational Efficiency

Empirically, Sardy *et al.* [12] found the BCR algorithm to be more efficient than the IP algorithm at solving basis pursuit.

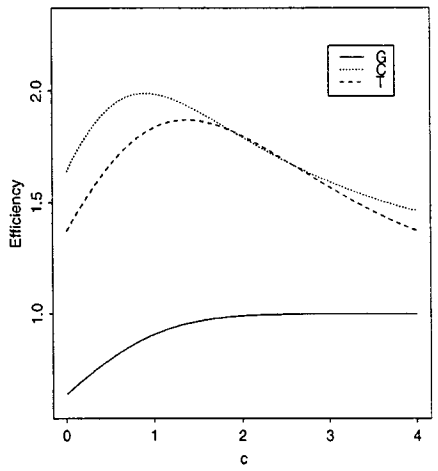

Fig. 2. Asymptotic relative efficiency as a function of c of the Huber estimate with respect to the sample average of an i.i.d. sample generated from the distributions used in the simulation: Gaussian (G), Gaussian mixture (C), and Student t_3 with three degrees of freedom (T).

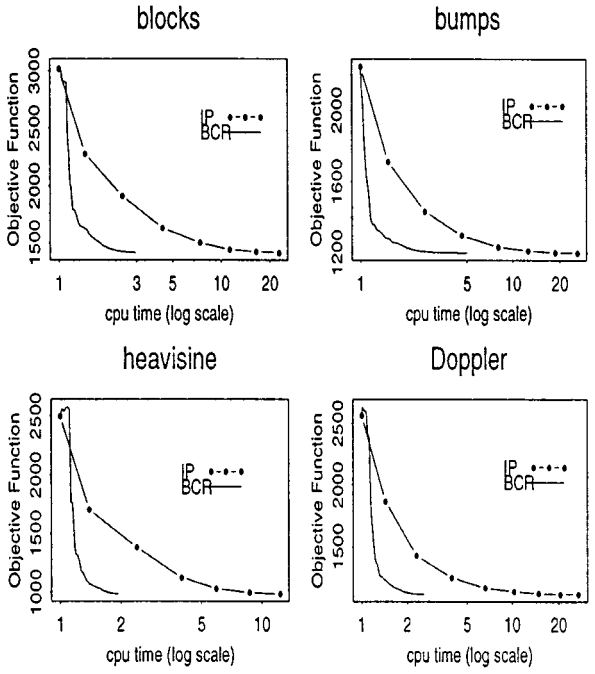

Fig. 3. Decrease in the robust basis pursuit objective function as a function of the CPU time (in seconds) for the BCR algorithm and the IP algorithm.

We observe the same advantage for solving its robust version (5). The reason is that the matrix $[\Phi, I_N]$ now used only has to be augmented by the N columns of the identity matrix. Fig. 3 illustrates, on four contaminated signals, the superiority of the BCR algorithm that is up to 10 times faster than the IP algorithm in achieving the desired precision [$\epsilon_1 = 0.02$ in (20)].

B. Predictive Performance

To illustrate graphically on a 1-D signal the advantage of using a robust procedure, Fig. 4 shows the robust estimation of *heavisine* for the same contaminated data as in Fig. 1. The robust procedure gives a better reconstruction; in particular, it preserves the two discontinuities and downweights the influence of the outliers.

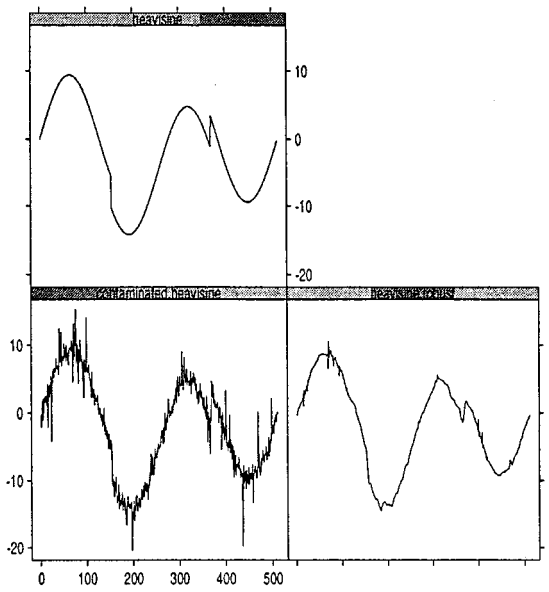

Fig. 4. Robust estimation of *heavisine*. (Top) True signal. (Bottom left) Noisy signal. (Bottom right) Robust estimate to compare with Fig. 1.

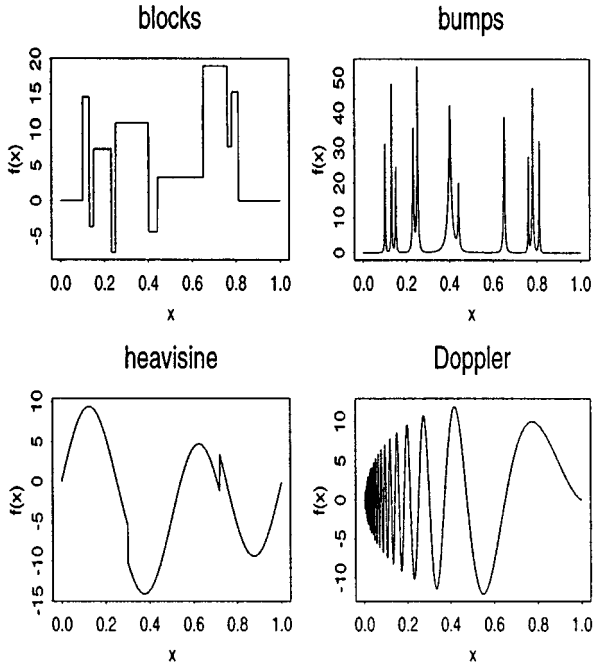

Fig. 5. Four signals used in the Monte Carlo experiment.

We perform a Monte Carlo experiment to evaluate the relative performance of the nonrobust and robust wavelet-based estimators with three noise scenarios:

- (G) standard Gaussian noise;
- (C) a mixture of 90% Gaussian(0, 1) and 10% Gaussian(0, 16) at random locations;
- (T) Student t noise with three degrees of freedom.

We use the four test functions plotted in Fig. 5 and defined in Donoho and Johnstone [1, Tab. 1]: *blocks*, *bumps*, *heavisine*, and

TABLE I
RELATIVE PREDICTIVE PERFORMANCE OF BASIS PURSUIT ($c = \infty$) AND ROBUST BASIS PURSUIT ($c = 2.0, c = 1.345$). ESTIMATED MEAN SQUARED ERROR ($\times 100$) WITH A STANDARD ERROR OF ROUGHLY 3%. (G: GAUSSIAN; C: CONTAMINATED; T: STUDENT t_3

		blocks			bumps			heavisine			Doppler		
N	c	G	C	T	G	C	T	G	C	T	G	C	T
1024	∞	**46**	109	137	**47**	**103**	**130**	**17**	71	91	**33**	85	109
	2.0	77	**103**	**114**	190	220	240	**17**	27	33	34	**49**	**58**
	1.345	141	167	179	860	884	913	20	**26**	**30**	47	59	67
4096	∞	**21**	69	91	**18**	62	85	**7**	50	68	**11**	55	76
	2.0	33	**44**	**49**	24	**35**	**38**	**7**	11	13	**11**	**17**	**20**
	1.345	58	67	72	104	114	109	9	**11**	**12**	15	19	22

Doppler. We normalize them such that their "standard deviation" is equal to 7

$$\int_0^1 (f(x) - \overline{f})^2 \, dx = 49, \quad \text{where} \quad \overline{f} = \int_0^1 f(x) \, dx. \quad (21)$$

We choose two sample sizes of $N = 1024$ and $N = 4096$, and the "s8" wavelet packet dictionary with all but four levels. The minimax smoothing parameters λ_N^* are $\lambda_{1024}^* = 2.232$ and $\lambda_{4096}^* = 2.594$. Following the discussion of Section III, the smoothing parameter is set to $\lambda = \lambda_N^* \hat{\sigma}$ and the cutpoint of the Huber loss function to $\hat{\tau} = c\hat{\sigma}$ with $c = 1.345$ and $c = 2.0$.

For each combination of noise (G, C, T) of underlying function (*blocks, bumps, heavisine, Doppler*), of sample size ($N = 1024$, $N = 4096$), and of procedure (nonrobust, robust), we estimate the MSE by averaging (40)(1024)/N model errors (i.e., 40 for $N = 1024$ and 10 for $N = 4096$ to get the same number of points is generated for the two sample sizes). Table I reports the estimated MSEs of the competing estimators for the 24 scenarios.

In light of Table I, a cutpoint of at least $c = 2$ is advisable for robust basis pursuit; the standard value of $c = 1.345$ derived in the parametric context from asymptotic considerations is not large enough. With a cutpoint of $c = 2$, the gain in efficiency can be dramatic for non-Gaussian noise using robust basis pursuit. Its counterperformance on the *bumps* signal is due to the nature of the signal whose features are difficult to distinguish with noise in the upper tail when the sampling is light ($N = 1024$); with an heavier sampling ($N = 4096$), robust basis pursuit again beats the nonrobust estimator for non-Gaussian noise.

V. APPLICATIONS

The first data is an image taken by a long-wavelength infrared sensor. Just visible above the center of the image is an A-37 trainer aircraft flying above the Sierra Nevada at some distance from the sensor platform. The 128 by 128 original image plotted in the top left of Fig. 6 clearly shows some "outlier pixels." A standard median filter (with a 3×3 window) gets rid of the bad pixels but does not preserve the aircraft well (top right). The two bottom plots of Fig. 6 show the denoised image using (left) basis pursuit and (right) robust basis pursuit. The robust denoising procedure has the definite advantage of cleaning the image of the bad pixels while preserving the outline of the airplane. To clean the image with the two wavelet-based techniques, we used

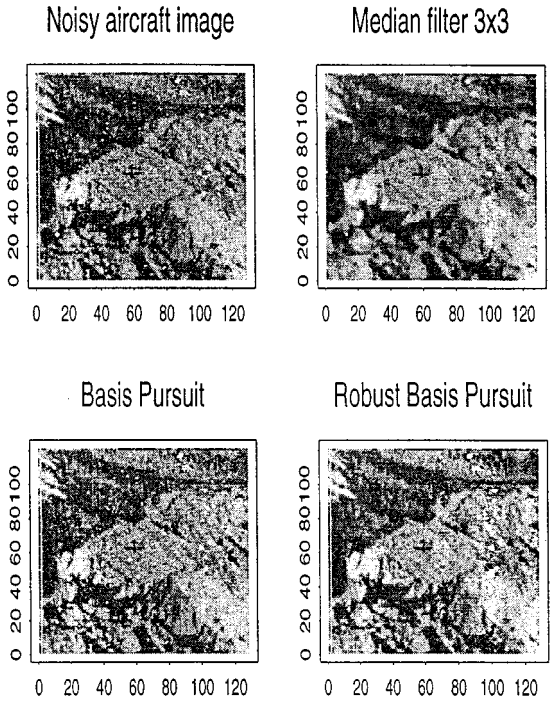

Fig. 6. Top: (Left) Noisy infrared sensor data and (Right) 3×3 robust median filter denoised image. Bottom: (Left) Nonrobust and (Right) robust wavelet based denoised images.

the 2-D nondecimated wavelet transform with the "s8" wavelet. In this application, we know the underlying image we want to recover; therefore, we tried several values of λ and τ (which was feasible in a finite time thanks to the BCR algorithm) and chose $\hat{\lambda} = \hat{\tau} = 0.4$ for the best visually appealing reconstruction of the aircraft. Using this parameter setting $\hat{\lambda} = \hat{\tau} = 0.4$, the robust procedure can be used to clean future images.

The second data are radar glint observations and consist of $N = 512$ angles of a target in degrees. The signal contains a number of glint spikes, causing the apparent signal to behave erratically. From physical considerations, a good model for the true signal is a low-frequency oscillation about 0°. The estimated standard deviation is $\hat{\sigma} = 8.4$. To get a nice "noise-free" visual display, we choose the universal threshold $\lambda = \hat{\sigma}\sqrt{2 \log N} = 29.6$, and for the robust version, we choose $\hat{\tau} = 2.0\hat{\sigma} = 11.3$. Fig. 7 shows the (top) original signal and (left: nonrobust; right: robust) the denoised estimates at the bottom. The robust estimate is a low-frequency oscillation, as expected, whereas the nonrobust estimate remains jagged. Note that the

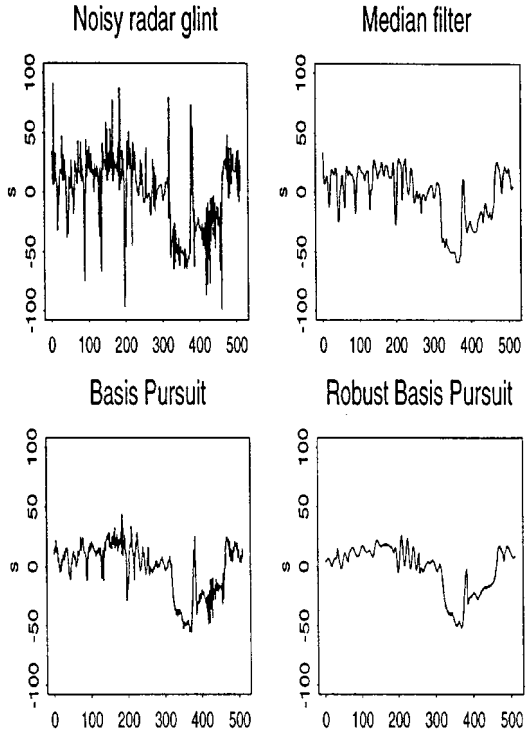

Fig. 7. Top: (Left) Noisy radar glint data and (Right) robust median filter estimate. Bottom: (Left) Nonrobust and (right) robust wavelet-based estimate (with local cosine packet).

median filter's estimate still shows a highly oscillating denoised signal.

VI. Conclusion

We have proposed a robust version of basis pursuit by replacing the nonrobust l_2 loss function by the so-called Huber loss function. We solved the corresponding nontrivial optimization problem for a given smoothing parameter $\lambda > 0$ and a given cutpoint $\tau \geq 0$ with an efficient and converging block coordinate relaxation method that is empirically faster than an interior point competitor. The two techniques are available in the wavelet module of the `Splus` statistical software; the BCR algorithm can otherwise be easily implemented. We proposed a rule to choose the smoothing parameter λ and the cutpoint of the Huber loss function τ. We showed on a particular simulation that robust basis pursuit has a good predictive performance with both Gaussian and long-tailed symmetric additive noise; in particular, we recommend using a cutpoint of at least $c = 2$ for the Huber loss function. As illustrated with two applications, robust basis pursuit has a definite advantage over both a nonrobust wavelet-based estimator and a median filter estimator.

Acknowledgment

The authors wish to thank the associate editor and five anonymous referees for their careful review and A. Davison for providing help with Fig. 2.

References

[1] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[3] S. Sardy, "Minimax threshold for denoising complex signals with waveshrink," *IEEE Trans. Signal Processing*, vol. 48, pp. 1023–1028, Apr. 2000.

[4] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

[5] G. P. Nason, "Wavelet function estimation using cross-validation," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. New York: Springer-Verlag, 1995, pp. 261–280.

[6] A. G. Bruce, I. M. Donoho, and R. D. Martin, "Denoising and robust nonlinear wavelet analysis," *Wavelet Applicat.*, vol. 2242, Apr. 1994.

[7] A. Kovac and B. W. Silverman, "Extending the scope of wavelet regression methods by coefficient-dependent thresholding," *J. Amer. Statist. Assoc.*, vol. 95, pp. 172–183, 2000.

[8] R. Averkamp and C. Houdré, "Wavelet thresholding for non (necessarily) gaussian noise: A preliminary report," Georgia Inst. Technol., Atlanta, Tech. Rep., 1996.

[9] H. Krim and I. C. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Trans. Inform. Theory*, vol. 45, pp. 898–908, May 1999.

[10] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[11] R. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

[12] S. Sardy, A. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *J. Comput. Graph. Statist.*, vol. 9, no. 2, pp. 361–379, 2000.

[13] S. Mann and S. Haykin, "Adaptive "chirplet" transform: And adaptive generalization of the wavelet transform," *Opt. Eng.*, vol. 31, no. 6, pp. 1243–1256, June 1992.

[14] F. G. Meyer and R. R. Coifman, "Biorthogonal brushlet bases for directional image compression," in *Proc. 12th Int. Conf. Anal. Optimization Syst.*, 1996.

[15] S. Sardy, "Regularization Techniques for Linear Regression with a Large Set of Carriers," Ph.D. dissertation, Univ. Washington, Seattle, 1998.

[16] M. Kojima, N. Megiddo, and S. Mizuno, "A primal-dual exterior point algorithm for linear programming," *Math. Progr.*, vol. 61, pp. 261–280, 1993.

[17] D. Xiang and G. Wahba, "A generalized approximate cross validation for smoothing splines with nongaussian data," *Statistica Sinica*, vol. 6, pp. 675–692, 1996.

[18] M. S. Crouse, R. G. Baraniuk, and R. D. Nowak, "Signal estimation using wavelet-markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, 1997, pp. 3429–3432.

Sylvain Sardy received the Maîtrise d'Ingénièrie Mathématique degree from the University of Franche-Comté, Besancon, France, in 1989. He then received the M.S. degree in mathematics in 1991 and the M.S. degree in statistics in 1992, both from Utah State University, Logan. After doing his military service for France at the Nuclear Engineering Department of the Massachusetts Institute of Technology, Cambridge, he received the Ph.D. degree in statistics from the University of Washington, Seattle.

He is now a Postdoctoral Assistant with the Swiss Federal Institute of Technology (EPFL), Lausanne, in the Mathematics Department (DMA).

Paul Tseng received the B.Sc. degree in engineering mathematics from Queen's University, Kingston, ON, Canada, in 1981 and the Ph.D. degree in operations research from the Massachusetts Institute of Technology (MIT), Cambridge, in 1986.

From 1986 to 1987, he was a University Research Fellow with the Department of Management Science, University of British Columbia, Vancouver, BC, Canada. From 1987 to 1990, he was a research associate with the Laboratory for Information and Decision Systems at MIT. He joined the Department of Mathematics, University of Washington, Seattle, in 1990, where he is currently a Professor. His research interests are in optimization and algorithms.

Andrew Bruce, photograph and biography not available at time of publication.