Regularization of Wavelet Approximations

Anestis ANTONIADIS and Jianqing FAN

In this paper, we introduce nonlinear regularized wavelet estimators for estimating nonparametric regression functions when sampling points are not uniformly spaced. The approach can apply readily to many other statistical contexts. Various new penalty functions are proposed. The hard-thresholding and soft-thresholding estimators of Donoho and Johnstone are specific members of nonlinear regularized wavelet estimators. They correspond to the lower and upper envelopes of a class of the penalized least squares estimators. Necessary conditions for penalty functions are given for regularized estimators to possess thresholding properties. Oracle inequalities and universal thresholding parameters are obtained for a large class of penalty functions. The sampling properties of nonlinear regularized wavelet estimators are established and are shown to be adaptively minimax. To efficiently solve penalized least squares problems, nonlinear regularized Sobolev interpolators (NRSI) are proposed as initial estimators, which are shown to have good sampling properties. The NRSI is further ameliorated by regularized one-step estimators, which are the one-step estimators of the penalized least squares problems using the NRSI as initial estimators. The graduated nonconvexity algorithm is also introduced to handle penalized least squares problems. The newly introduced approaches are illustrated by a few numerical examples.

KEY WORDS: Asymptotic minimax; Irregular designs; Nonquadratic penality functions; Oracle inequalities; Penalized least-squares; ROSE; Wavelets.

1. INTRODUCTION

Wavelets are a family of orthogonal bases that can effectively compress signals with possible irregularities. They are good bases for modeling statistical functions. Various applications of wavelets in statistics have been made in the literature. See, for example, Donoho and Johnstone (1994), Antoniadis, Grégoire, and McKeague (1994), Hall and Patil (1995), Neumann and Spokoiny (1995), Antoniadis (1996), and Wang (1996). Further references can be found in the survey papers by Donoho et al. (1995), Antoniadis (1997), and Abramovich, Bailey, and Sapatinas (2000) and books by Ogden (1997) and Vidakovic (1999). Yet, wavelet applications to statistics are hampered by the requirements that the designs are equispaced and the sample size be a power of 2. Various attempts have been made to relax these requirements. See, for example, the interpolation method of Hall and Turlach (1997), the binning method of Antoniadis, Grégoire, and Vial (1997), the transformation method of Cai and Brown (1997), the isometric method of Sardy et al. (1999), and the interpolation method to a fine regular grid of Kovac and Silverman (2000). However, it poses some challenges to extend these methods to other statistical contexts, such as generalized additive models and generalized analysis of variance models.

In an attempt to make genuine wavelet applications to statistics, we approach the denoising problem from a statistical modeling point of view. The idea can be extended to other statistical contexts. Suppose that we have noisy data at irregular design points $\{t_1, \ldots, t_n\}$:

$$Y_i = f(t_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

where f is an unknown regression to be estimated from the noisy sample. Without loss of generality, assume that the function f is defined on $[0, 1]$. Assume further that $t_i = n_i/2^J$ for some n_i and some fine resolution J that is determined by users. Usually, $2^J \geq n$ so that the approximation errors by moving nondyadic points to dyadic points are negligible. Let $\mathbf{f}$ be the underlying regression function collected at all dyadic points $\{i/2^J, i = 0, \ldots, 2^J - 1\}$. Let $\mathbf{W}$ be a given wavelet transform and $\boldsymbol{\theta} = \mathbf{W}\mathbf{f}$ be the wavelet transform of $\mathbf{f}$. Because $\mathbf{W}$ is an orthogonal matrix, $\mathbf{f} = \mathbf{W}^T\boldsymbol{\theta}$.

From a statistical modeling point of view, the unknown signals are modeled by $N = 2^J$ parameters. This is an overparameterized linear model, which aims at reducing modeling biases. One can not find a reasonable estimate of $\boldsymbol{\theta}$ by using the ordinary least squares method. Because wavelets are used to transform the regression function f, its representation in wavelet domain is sparse; namely, many components of $\boldsymbol{\theta}$ are small, for the function f in a Besov space. This prior knowledge enables us to reduce effective dimensionality and to find reasonable estimates of $\boldsymbol{\theta}$.

To find a good estimator of $\boldsymbol{\theta}$, we apply a penalized least squares method. Denote the sampled data vector by $\mathbf{Y}_n$. Let $\mathbf{A}$ be $n \times N$ matrix whose ith row corresponds to the row of the matrix $\mathbf{W}^T$ for which signal $f(t_i)$ is sampled with noise. Then, the observed data can be expressed as a linear model

$$\mathbf{Y}_n = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n) \tag{1.1}$$

where $\boldsymbol{\epsilon}$ is the noise vector. The penalized least squares problem is to find $\boldsymbol{\theta}$ to minimize

$$2^{-1}\|\mathbf{Y}_n - \mathbf{A}\boldsymbol{\theta}\|^2 + \lambda \sum_{i=1}^{N} p(|\theta_i|) \tag{1.2}$$

for a given penalty function p and regularization parameter $\lambda > 0$. The penalty function p is usually nonconvex on $[0, \infty)$ and irregular at point zero to produce sparse solutions. See Theorem 1 for necessary conditions. It poses some challenges to optimize such a high-dimensional nonconvex function.

Anestis Antoniadis is Professor, Laboratoire de Modélisation et Calcul, Université Joseph Fourier, 38041 Grenoble Cedex 9, France. Jianqing Fan is Professor of Statistics and Chairman, Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong. Antoniadis's research was supported by the IDOPT project, INRIA-CNRS-IMAG. Fan's research was partially supported by NSF grants DMS-9803200, DMS-9977096, and DMS-0196041 and RGC grant CUHK-4299/00P of HKSAR. The major part of the research was conducted while Fan visited the Université Joseph Fourier. He is grateful for the generous support of the university. The authors thank Arne Kovac and Bernard Silverman for providing their Splus procedure makegrid required for some of the data analyses. The Splus procedure is available at http://www.stat.mathematik.uni-essen.de/~kovac/makegrid.tar. The authors thank the editor, the associate editor, and the referees for their comments, which substantially improved this article.

Journal of the American Statistical Association
September 2001, Vol. 96, No. 455, Theory and Methods

Our overparameterization approach is complementary to the overcomplete wavelet library methods of Chen, Donoho, and Sanders (1998) and Donoho et al. (1998). Indeed, even when the sampling points are equispaced, one can still choose a large N ($N = O(n \log n)$, say), to have better ability to approximate unknown functions. Our penalized method in this case can be viewed as a subbasis selection from an overcomplete family of nonorthogonal bases, consisting of N columns of the matrix $\mathbf{A}$.

When $n = 2^J$, the matrix $\mathbf{A}$ becomes a square orthogonal matrix $\mathbf{W}^T$. This corresponds to the canonical wavelet denoising problems studied in the seminal paper by Donoho and Johnstone (1994). The penalized least squares estimator (1.2) can be written as

$$2^{-1}\|\mathbf{W}\mathbf{Y}_n - \boldsymbol{\theta}\|^2 + \lambda \sum_{i=1}^{N} p(|\theta_i|).$$

The minimization of this high-dimensional problem reduces to componentwise minimization problems, and the solution can be easily found. Theorem 1 gives necessary conditions for the solution to be unique and continuous in wavelet coefficients. In particular, the soft-thresholding rule and hard-thresholding rule correspond, respectively, to the penalized least squares estimators with the L_1 penalty and the hard-thresholding penalty (2.8) discussed in Section 2. These penalty functions have some unappealing features and can be further ameliorated by the smoothly clipped absolute deviation (SCAD) penalty function and the transformed L_1 penalty function. See Section 2.3 for more discussions.

The hard-thresholding and soft-thresholding estimators play no monopoly role in choosing an ideal wavelet subbasis to efficiently represent an unknown function. Indeed, for a large class of penalty functions, we show in Section 3 that the resulting penalized least squares estimators perform within a logarithmic factor to the oracle estimator in choosing an ideal wavelet subbasis. The universal thresholding parameters are also derived. They can easily be translated in terms of regularization parameters λ for a given penalty function p. The universal thresholding parameter given by Donoho and Johnstone (1994) is usually somewhat too large in practice. We expand the thresholding parameters up to the second order, allowing users to choose smaller regularization parameters to reduce modeling biases. The work on the oracle inequalities and universal thresholding is a generalization of the pioneering work of Donoho and Johnstone (1994). It allows statisticians to use other penalty functions with the same theoretical backup.

The risk of the oracle estimator is relatively easy to compute. Because the penalized least squares estimators perform comparably with the oracle estimator, following the similar but easier calculation to that of Donoho et al. (1995), we can show that the penalized least squares estimators with simple data-independent (universal) thresholds are adaptively minimax for the Besov class of functions, for a large class of penalty functions.

Finding a meaningful local minima to the general problem (1.2) is not easy, because it is a high-dimensional problem with a nonconvex target function. A possible method is to apply the graduated nonconvexity (GNC) algorithm introduced by Blake and Zisserman (1987) and Blake (1989) and ameliorated by Nikolova (1999) and Nikolova, Idier, and Mohammad-Djafari (in press) in the imaging analysis context. The algorithm contains good ideas on optimizing high-dimensional nonconvex functions, but its implementation depends on a several tuning parameters. It is reasonably fast, but it is not nearly as fast as the canonical wavelet denoising. See Section 6 for details. To have a fast estimator, we impute the unobserved data by using regularized Sobolev interpolators. This allows one to apply coefficientwise thresholding to obtain an initial estimator. This yields a viable initial estimator, called nonlinear regularized Sobolev interpolators (NRSI). This estimator is shown to have good sampling properties. By using this NRSI to create synthetic data and apply the one-step penalized least squares procedure, we obtain a regularized one-step estimator (ROSE). See Section 4. Another possible approach to denoise nonequispaced signals is to design adaptively nonorthogonal wavelets to avoid overparameterizing problems. A viable approach is the wavelet networks proposed by Bernard, Mallat, and Slotine (1999).

An advantage of our penalized wavelet approach is that it can readily be applied to other statistical contexts, such as likelihood-based models, in a manner similar to smoothing splines. One can simply replace the normal likelihood in (1.2) by a new likelihood function. Further, it can be applied to high-dimensional statistical models such as generalized additive models. Details of these require a lot of new work and hence are not discussed here. Penalized likelihood methods were successfully used by Tibshirani (1995), Barron, Birgé, and Massart (1999), and Fan and Li (1999) for variable selections. Thus, they should also be viable for wavelet applications to other statistical problems. When the sampling points are equispaced, the use of penalized least squares for regularizing wavelet regression were proposed by Solo (1998), McCoy (1999), Moulin and Liu (1999), and Belge, Kilmer, and Miller (2000). In Solo (1998), the penalized least squares with an L_1 penalty is modified to a weighted least squares to deal with correlated noise, and an iterative algorithm is discussed for its solution. The choice of the regularization parameter is not discussed. By analogy to smoothing splines, McCoy (1999) used a penalty function that simultaneously penalizes the residual sum of squares and the second derivative of the estimator at the design points. For a given regularization parameter, the solution of the resulting optimization problem is found by using simulated annealing, but there is no suggestion in her work of a possible method of choosing the smoothing parameter. Moreover, although the proposal is attractive, the optimization algorithm is computationally demanding. In Moulin and Liu (1999), the soft- and hard-thresholded estimators appeared as Maximum a Posteriori (MAP) estimators in the context of Bayesian estimation under zero-one loss, with generalized Gaussian densities serving as a prior distribution for the wavelet coefficients. A similar approach was used by Belge et al. (2000) in the context of wavelet domain image restoration. The smoothing parameter in Belge et al. (2000) was selected by the L-curve criterion (Hansen and O'Leary 1993). It is known, however, (Vogel 1996) that such a criterion can lead to nonconvergent solutions, especially when the function to be recovered presents some irregularities. Although there is no conceptual difficulty in applying the penalized wavelet method to other statistical problems, the dimensionality involved is usually very high. Its fast implementations require some new ideas, and the GNC algorithm offers a generic numerical method.

This article is organized as follows. In Section 2, we introduce Sobolev interpolators and penalized wavelet estimators. Section 3 studies the properties of penalized wavelet estimators when the data are uniformly sampled. Implementations of penalized wavelet estimators in general setting are discussed in Section 4. Section 5 gives numerical results of our newly proposed estimators. Two other possible approaches are discussed in Section 6. Technical proofs are presented in the Appendix.

2. REGULARIZATION OF WAVELET APPROXIMATIONS

The problem of signal denoising from nonuniformly sampled data arises in many contexts. The signal recovery problem is ill posed, and smoothing can be formulated as an optimization problem with side constraints to narrow the class of candidate solutions.

We first briefly discuss wavelet interpolation by using a regularized wavelet method. This serves as a crude initial value to our proposed penalized least squares method. We then discuss the relation between this and nonlinear wavelet thresholding estimation when the data are uniformly sampled.

2.1 Regularized Wavelet Interpolations

Assume for the moment that the signals are observed with no noise, i.e., $\boldsymbol{\epsilon} = 0$ in (1.1). The problem becomes an interpolation problem, using a wavelet transform. Being given signals only at the nonequispaced points $\{t_i, i = 1, \ldots, n\}$ necessarily means that we have no information at other dyadic points. In terms of the wavelet transform, this means that we have no knowledge about the scaling coefficients at points other than t_i's. Let

$$\mathbf{f}_n = (f(t_1), \ldots, f(t_n))^T$$

be the observed signals. Then, from (1.1) and the assumption $\boldsymbol{\epsilon} = 0$, we have

$$\mathbf{f}_n = \mathbf{A}\boldsymbol{\theta}. \tag{2.1}$$

Because this is an underdetermined system of equations, there exist many different solutions for $\boldsymbol{\theta}$ that match the given sampled data $\mathbf{f}_n$. For the minimum Sobolev solution, we choose the $\mathbf{f}$ that interpolates the data and minimizes the weighted Sobolev norm of f. This would yield a smooth interpolation to the data. The Sobolev norms of f can be simply characterized in terms of the wavelet coefficients $\boldsymbol{\theta}$. For this purpose, we use double array sequence $\theta_{j,k}$ to denote the wavelet coefficient at the jth resolution level and kth dyadic location ($k = 1, \ldots, 2^{j-1}$). A Sobolev norm of f with degree of smoothness s can be expressed as

$$\|\boldsymbol{\theta}\|_S^2 = \sum_j 2^{2sj} \|\boldsymbol{\theta}_{j.}\|^2,$$

where $\boldsymbol{\theta}_{j.}$ is the vector of the wavelet coefficients at the resolution level j. Thus, we can restate this problem as a wavelet-domain optimization problem: Minimize $\|\boldsymbol{\theta}\|_S^2$ subject to constraint (2.1). The solution (Rao 1973) is what is called the normalized method of frame whose solution is given by

$$\boldsymbol{\theta} = \mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{f}_n,$$

where $\mathbf{D} = \mathrm{Diag}(2^{-2sj_i})$ with j_i denoting the resolution level with which θ_i is associated. An advantage of the method of frame is that it does not involve the choice of regularization parameter (unless s is regarded as a smoothing parameter). When $s = 0$, $\boldsymbol{\theta} = \mathbf{A}^T\mathbf{f}_n$ by orthogonality. In this case, the interpolator is particularly easy to compute.

As an illustration of how the regularized wavelet interpolations work, we took 100 data points (located at the tick marks) from the function depicted in Figure 1(a). Figure 1, (b)–(d) show how the method of frame works for different values of s. As s increases, the interpolated functions become smoother. In fact, for a large range of values of s, the wavelet interpolations do not create excessive biases.

2.2 Regularized Wavelet Estimators

Assume now that the observed data follow model (1.1). The traditional regularization problem can be formulated in the wavelet domain as follows. Find the minimum of

$$2^{-1}\|\mathbf{Y}_n - \mathbf{A}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_S^2, \tag{2.2}$$

The resulting estimation procedure parallels standard spline-smoothing techniques. Several variants of such a penalized approach for estimating less regular curves via their wavelet decomposition have been suggested by several authors (Antoniadis 1996, Amato and Vuza 1997, Dechevsky and Penev 1999). The resulting estimators are linear estimators of shrinkage type with a level dependent shrinking of the empirical wavelet coefficients. Several data-driven methods were proposed for the determination of the penalty parameter λ, and we refer the readers to the cited papers for rigorous treatments on the choice of the regularization parameter for such linear estimators. The preceeding leads to a regularized linear estimator.

In general, one can replace the Sobolev norm by other penalty functions, leading to minimizing

$$\ell(\boldsymbol{\theta}) = 2^{-1}\|\mathbf{Y}_n - \mathbf{A}\boldsymbol{\theta}\|^2 + \lambda \sum_{i \geq i_0} p(|\theta_i|) \tag{2.3}$$

for a given penalty function $p(\cdot)$ and given value i_0. This corresponds to penalizing wavelet coefficients above certain resolution level j_0. Here, to facilitate the presentation, we changed the notation $\theta_{j,k}$ from a double array sequence into a single array sequence θ_i. The problem (2.3) produces stable and sparse solutions for functions p satisfying certain properties. The solutions, in general, are nonlinear. See the results of Nikolova (2000) and Section 3 below.

2.3 Penalty Functions and Nonlinear Wavelet Estimators

The regularized wavelet estimators are an extension of the soft- and hard-thresholding rules of Donoho and Johnstone (1994). When the sampling points are equally spaced and

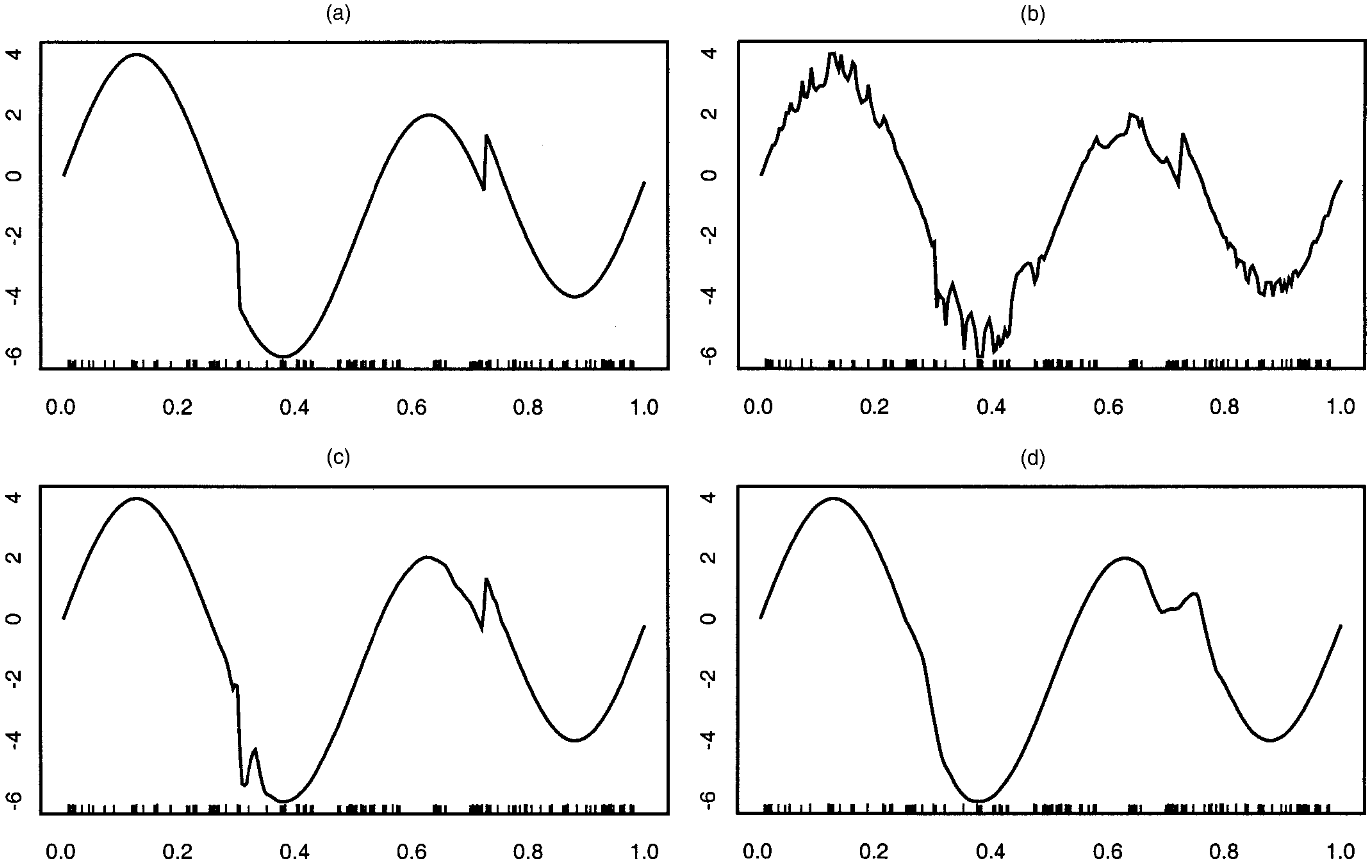

Figure 1. Wavelet Interpolations by Method of Frame. As degrees of smoothness s become larger, the interpolated functions become smoother. (a) The target function and sampling points, true cruve (tick marks); (b)–(d) wavelet interpolations with s = .5,s = 1.4, and s = 6.0.

$n = 2^J$, the design matrix $\mathbf{A}$ in (2.1) becomes the inverse wavelet transform matrix $\mathbf{W}^T$. In this case, (2.3) becomes

$$2^{-1}\sum_{i=1}^{n}(z_i-\theta_i)^2+\lambda\sum_{i\geq i_0}p(|\theta_i|), \tag{2.4}$$

where z_i is the ith component of the wavelet coefficient vector $\mathbf{z} = \mathbf{W}\mathbf{Y}_n$. The solution to this problem is a componentwise minimization problem, whose properties are studied in the next section. To reduce abuse of notation, and because $p(|\theta|)$ is allowed to depend on λ, we use p_λ to denote the penalty function λp in the following discussion.

For the L_1-penalty [Figure 2(a)],

$$p_\lambda(|\theta|)=\lambda|\theta|, \tag{2.5}$$

the solution is the soft-thresholding rule (Donoho et al. 1992). A clipped L_1-penalty

$$p(\theta)=\lambda\min(|\theta|,\lambda) \tag{2.6}$$

leads to a mixture of soft- and hard-thresholding rules (Fan 1997):

$$\hat{\theta}_j=(|z_j|-\lambda)_+I\{|z_j|\leq 1.5\lambda\}+|z_j|I\{|z_j|>1.5\lambda\}. \tag{2.7}$$

When the penalty function is given by

$$p_\lambda(|\theta|)=\lambda^2-(|\theta|-\lambda)^2I(|\theta|<\lambda), \tag{2.8}$$

[see Figure 2(b)], the solution is the hard-thresholding rule (Antoniadis 1997). This is a smoother penalty function than $p_\lambda(|\theta|) = |\theta|I(|\theta| < \lambda) + \lambda/2I(|\theta| \geq \lambda)$ suggested by Fan (1997) and the entropy penalty $p_\lambda(|\theta|) = 2^{-1}\lambda^2 I\{|\theta| \neq 0\}$, which lead to the same solution. The hard-thresholding rule is discontinuous, whereas the soft-thresholding rule shifts the estimator by an amount of λ even when $|z_i|$ stands way out of noise level, which creates unnecessary bias when θ is large. To ameliorate these two drawbacks, Fan (1997) suggests using the quadratic spline penalty, called the smoothly clipped absolute deviation (SCAD) penalty [see Figure 2(c)],

$$p'_\lambda(\theta)=I(\theta\leq\lambda)+\frac{(a\lambda-\theta)_+}{(a-1)\lambda}I(\theta>\lambda)$$
$$\text{for } \theta>0 \text{ and } a>2, \tag{2.9}$$

leading to the piecewise linear thresholding

$$\hat{\theta}_j=\begin{cases}\operatorname{sgn}(z_j)(|z_j|-\lambda) & \text{when } |z_j|\leq 2\lambda,\\ +\frac{(a-1)z_j-a\lambda\operatorname{sgn}(z_j)}{a-2} & \text{when } 2\lambda<|z_j|\leq a\lambda,\\ z_j & \text{when } |z_j|>a\lambda.\end{cases} \tag{2.10}$$

Fan and Li (1999) recommended using $a = 3.7$ based on a Bayesian argument. This thresholding estimator is in the same spirit as that of Gao and Bruce (1997). This penalty function does not overpenalize large values of $|\theta|$ and hence does not create excessive biases when the wavelet coefficients are large.

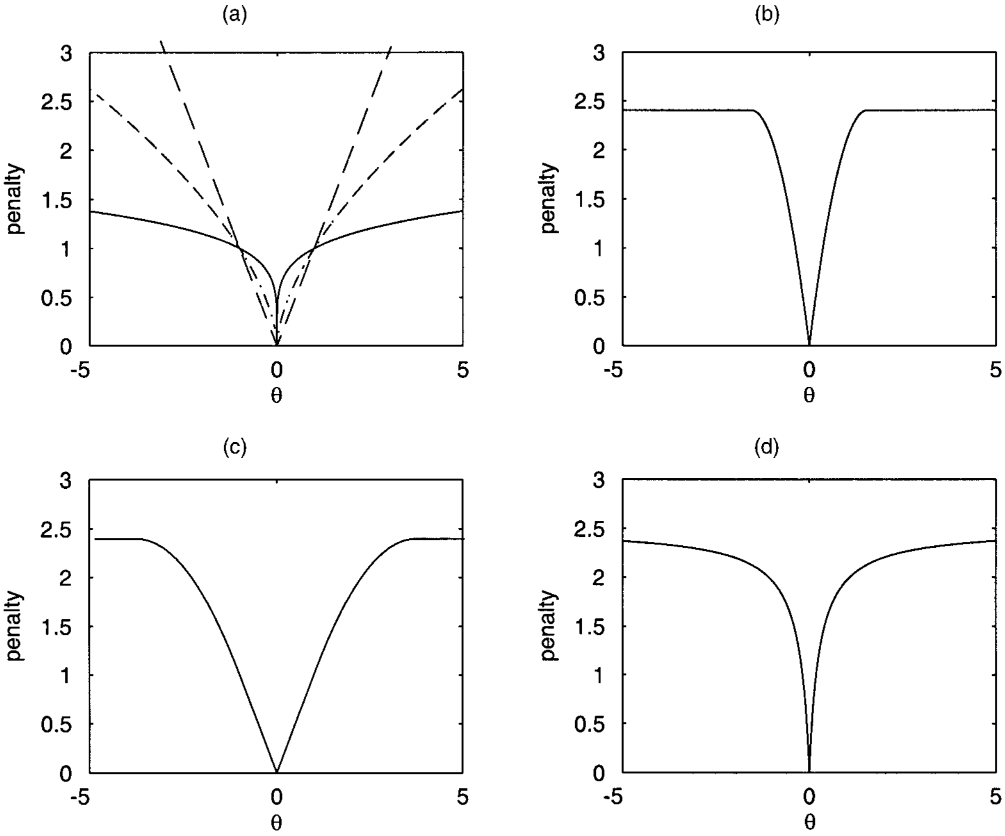

Figure 2. Typical Penalty Functions That Preserve Sparsity. (a) L_p-penalty with $p = 1$ (long dash), $p = .6$ (short dash), and $p = .2$ (solid); (b) hard-thresholding penalty (2.8); (c) SCAD (2.9) with $a = 3.7$; (d) transformed L_1-penalty (2.11) with $b = 3.7$.

Nikolova (1999b) suggested the following transformed L_1-penalty function [see Figure 2(d)]:

$$p_\lambda(|x|) = \lambda b|x|(1+b|x|)^{-1} \qquad \text{for some } b > 0. \tag{2.11}$$

This penalty function behaves quite similarly to the SCAD suggested by Fan (1997). Both are concave on $[0, \infty)$ and do not intend to overpenalize large $|\theta|$. Other possible functions include the L_p-penalty introduced (in image reconstruction) by Bouman and Sauer (1993):

$$p_\lambda(|\theta|) = \lambda|\theta|^p \qquad (p \geq 0). \tag{2.12}$$

As shown in Section 3.1, the choice $p \leq 1$ is a necessary condition for the solution to be a thresholding estimator, whereas $p \geq 1$ is a necessary condition for the solution to be continuous in $\mathbf{z}$. Thus, the L_1-penalty function is the only member in this family that yields a continuous thresholding solution.

Finally, we note that the regularization parameter λ for different penalty functions has a different scale. For example, the value λ in the L_1-penalty function is not the same as that in the L_p-penalty $(0 \leq p < 1)$. Figure 2 depicts some of these penalty functions. Their componentwise solutions to the corresponding penalized least squares problem (2.4) are shown in Figure 3.

3. ORACLE INEQUALITIES AND UNIVERSAL THRESHOLDING

As mentioned in Section 2.3, there are many competing thresholding policies. They provide statisticians and engineers a variety of choices of penalty functions with which to estimate functions with irregularities and to denoise images with sharp features. However, these have not yet been systematically studied. We first study the properties of penalized least squares estimators, and then we examine the extent to which they can mimic oracle in choosing the regularization parameter λ.

3.1 Characterization of Penalized Least Squares Estimators

Let $p(\cdot)$ be a nonnegative, nondecreasing, and differentiable function on $(0, \infty)$. The clipped L_1-penalty function (2.6) does not satisfy this condition and will be excluded in the study. All other penalty functions satisfy this condition. Consider the following penalized least squares problem: Minimize with respect to θ

$$\ell(\theta) = (z - \theta)^2/2 + p_\lambda(|\theta|) \tag{3.1}$$

for a given penalty parameter λ. This is a componentwise minimization problem of (2.4). Note that the function in (3.1) tends to infinity as $|\theta| \to \infty$. Thus, minimizers do exist. Let $\hat{\theta}(z)$ be a solution. The following theorem gives the necessary conditions (indeed, they are sufficient conditions, too) for the solution to be thresholding, to be continuous, and to be approximately unbiased when $|z|$ is large.

Theorem 1. Let $p_\lambda(\cdot)$ be a nonnegative, nondecreasing, and differentiable function in $(0, \infty)$. Further, assume that the

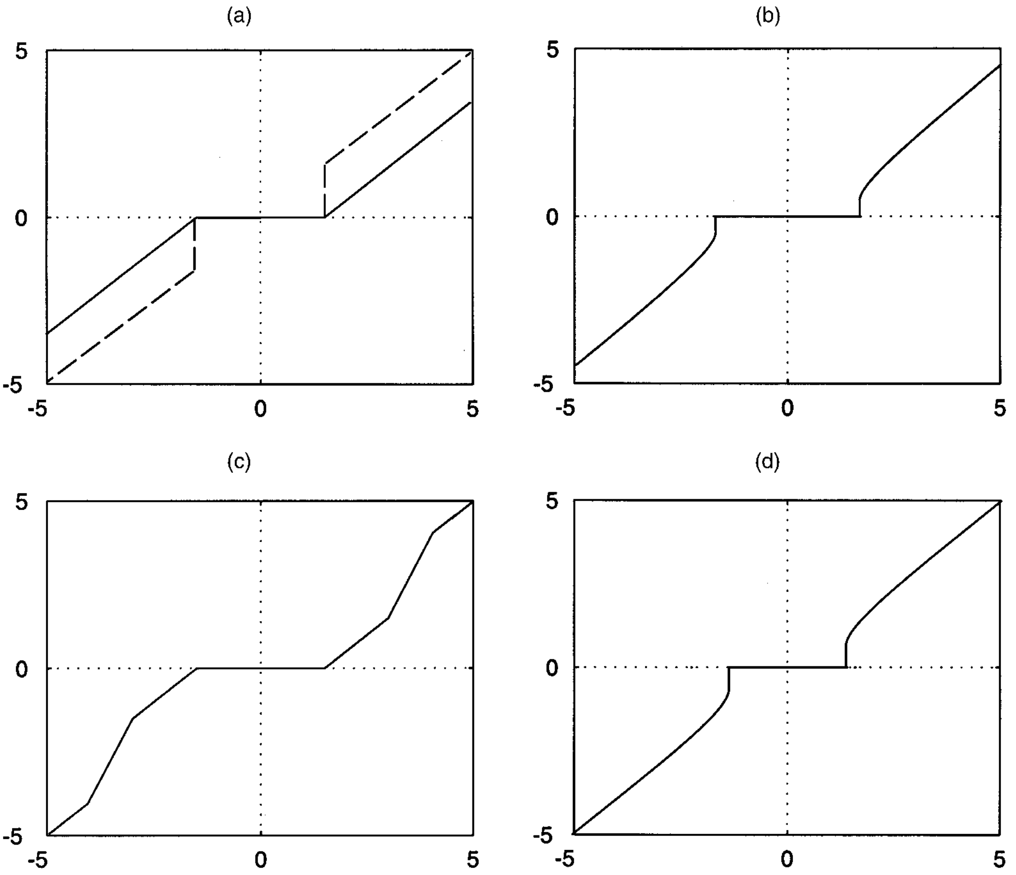

Figure 3. Penalized Least Squares Estimators That Possess Thresholding Properties. (a) The penalized L_1 estimator and the hard-thresholding estimator (dashed); (b) the penalized L_p estimator with $p = .6$; (c) the penalized SCAD estimator (2.10); (d) the penalized transformed L_1 estimator with $b = 3.7$.

function $-\theta - p'_\lambda(\theta)$ is strictly unimodal on $(0, \infty)$. Then we have the following results.

1. The solution to the minimization problem (3.1) exists and is unique. It is antisymmetric: $\hat{\theta}(-z) = -\hat{\theta}(z)$.
2. The solution satisfies

$$\hat{\theta}(z) = \begin{cases} 0 & \text{if } |z| \le p_0, \\ z - \text{ sgn } (z) p'_\lambda(|\hat{\theta}(z)|) & \text{if } |z| > p_0, \end{cases}$$

   where $p_0 = \min_{\theta \ge 0}\{\theta + p'_\lambda(\theta)\}$. Moreover, $|\hat{\theta}(z)| \le |z|$.
3. If $p'_\lambda(\cdot)$ is nonincreasing, then for $|z| > p_0$, we have

$$|z| - p_0 \le |\hat{\theta}(z)| \le |z| - p'_\lambda(|z|).$$

4. When $p'_\lambda(\theta)$ is continuous on $(0, \infty)$, the solution $\hat{\theta}(z)$ is continuous if and only if the minimum of $|\theta| + p'_\lambda(|\theta|)$ is attained at point zero.
5. If $p'_\lambda(|z|) \to 0$, as $|z| \to +\infty$, then

$$\hat{\theta}(z) = z - p'_\lambda(|z|) + o(p'_\lambda(|z|)).$$

We now give the implications of these results. When $p'_\lambda(0+) > 0$, $p_0 > 0$. Thus, for $|z| \le p_0$, the estimate is thresholded to 0. For $|z| > p_0$, the solution has a shrinkage property. The amount of shrinkage is sandwiched between the soft-thresholding and hard-thresholding estimators, as shown in result 3. In other words, the hard- and soft-thresholding estimators of Donoho and Johnstone (1994) correspond to the extreme cases of a large class of penalized least squares estimators. We add that a different estimator $\hat{\theta}$ may require different thresholding parameter p_0 and, hence the estimator $\hat{\theta}$ is not necessarily sandwiched by the hard- and soft-thresholding estimators by using different thresholding parameters. Further, the amount of shrinkage gradually tapers off as $|z|$ gets large when $p'_\lambda(|z|)$ goes to zero. For example, the penalty function $p_\lambda(|\theta|) = \lambda r^{-1}|\theta|^r$ for $r \in (0, 1]$ satisfies this condition. The case $r = 1$ corresponds to the soft-thresholding. When $0 < r < 1$,

$$p_0 = (2 - r)\{(1 - r)^{r-1}\lambda\}^{1/(2-r)},$$

and when $|z| > p_0$, $\hat{\theta}(z)$ satisfies the equation

$$\hat{\theta} + \lambda\hat{\theta}^{r-1} = z.$$

In particular, when $r \to 0$,

$$\hat{\theta} \to \hat{\theta}_0 \equiv (z + \sqrt{z^2 - 4\lambda})/2 = z/(1 + \lambda z^{-2}) + O(z^{-4}).$$

The procedure corresponds basically to the Garotte estimator in Breiman (1995). When the value of $|z|$ is large, one is certain that the observed value $|z|$ is not noise. Hence, one does not wish to shrink the value of z, which would result in underestimating θ. Theorem 1, result 4, shows that this property holds when $p_\lambda(|\theta|) = \lambda r^{-1}|\theta|^r$ for $r \in (0, 1)$. This ameliorates the property of the soft-thresholding rule, which always shifts the estimate z by an amount of δ. However, by Theorem 1, result 4, the solution is not continuous.

3.2 Risks of Penalized Least Squares Estimators

We now study the risk function of the penalized least squares estimator $\hat{\theta}$ that minimizes (3.1). Assume $Z \sim N(\theta, 1)$. Denote by

$$R_p(\theta, p_0) = E\{\hat{\theta}(Z) - \theta\}^2.$$

Note that the thresholding parameter p_0 is equivalent to the regularization parameter λ. We explicitly use the thresholding parameter p_0 because it is more relevant. For wavelet applications, the thresholding parameter p_0 will be in the order of magnitude of the maximum of the Gaussian errors. Thus, we consider only the situation where the thresholding level is large.

In the following theorem, we give risk bounds for penalized least squares estimators for general penalty functions. The bounds are quite sharp because they are comparable with those for the hard-thresholding estimator given by Donoho and Johnstone (1994). A shaper bound will be considered numerically in the following section for a specific penalty function.

Theorem 2. Suppose p satisfies conditions in Theorem 1 and $p'_\lambda(0+) > 0$. Then

1. $R_p(\theta, p_0) \le 1 + \theta^2$.
2. If $p'_\lambda(\cdot)$ is nonincreasing, then

$$R_p(\theta, p_0) \le p_0^2 + \sqrt{2/\pi}\, p_0 + 1.$$

3. $R_p(0, p_0) \le \sqrt{2/\pi}(p_0 + p_0^{-1}) \exp(-p_0^2/2)$.
4. $R_p(\theta, p_0) \le R_p(0, \theta) + 2\theta^2$.

Note that properties 1–4 are comparable with those for the hard-thresholding and soft-thresholding rules given by Donoho and Johnstone (1994). The key improvement here is that the results hold for a larger class of penalty functions.

3.3 Oracle Inequalities and Universal Thresholding

Following Donoho and Johnstone (1994), when the true signal θ is given, one would decide whether to estimate the coefficient, depending on the value of $|\theta|$. This leads to an ideal oracle estimator $\hat{\theta}_o = ZI(|\theta| > 1)$, which attains the ideal L_2-risk $\min(\theta^2, 1)$. In the following discussions, the constant n can be arbitrary. In our nonlinear wavelet applications, the constant n is the sample size.

As discussed, the selection of λ is equivalent to the choice of p_0. Hence, we focus on the choice of the thresholding parameter p_0. When $p_0 = \sqrt{2 \log n}$, the universal thresholding proposed by Donoho and Johnstone (1994), by property (3) of Theorem 2,

$$R_p(0, p_0) \le \sqrt{2/\pi}\{(2 \log n)^{1/2} + 1\}/n \quad \text{when } p_0 \ge 1,$$

which is larger than the ideal risk. To bound the risk of the nonlinear estimator $\hat{\theta}(Z)$ by that of the oracle estimator $\hat{\theta}_o$, we need to add an amount cn^{-1} for some constant c to the risk of the oracle estimator, because it has no risk at point $\theta = 0$. More precisely, we define

$$\Lambda_{n,c,p_0}(p) = \sup_\theta \frac{R_p(\theta, p_0)}{cn^{-1} + \min(\theta^2, 1)}$$

and denote $\Lambda_{n,c,p_0}(p)$ by $\Lambda_{n,c}(p)$ for the universal thresholding $p_0 = \sqrt{2 \log n}$. Then, $\Lambda_{n,c,p_0}(p)$ is a sharp risk upper bound for using the universal thresholding parameter p_0. That is,

$$R_p(\theta, p_0) \le \Lambda_{n,c,p_0}(p)\{cn^{-1} + \min(\theta^2, 1)\}. \tag{3.2}$$

Thus, the penalized least squares estimator $\hat{\theta}(Z)$ performs comparably with the oracle estimator within a factor of $\Lambda_{n,c,p_0}(p)$. Likewise, let

$$\Lambda^*_{n,c}(p) = \inf_{p_0} \sup_\theta \frac{R_p(\theta, p_0)}{cn^{-1} + \min(\theta^2, 1)}$$

and

$$p_n = \text{the largest constant attaining } \Lambda^*_{n,c}(p).$$

Then, the constant $\Lambda^*_{n,c}(p)$ is the sharp risk upper bound using the minimax optimal thresholding p_n. Necessarily,

$$R_p(\theta, p_n) \le \Lambda^*_{n,c}(p_n)\{cn^{-1} + \min(\theta^2, 1)\}. \tag{3.3}$$

Donoho and Johnstone (1994) noted that the universal thresholding is somewhat too large. This is observed in practice. In this section, we propose a new universal thresholding policy, that takes the second order into account. This gives a lower bound under which penalized least squares estimators perform comparably with the oracle estimator. We then establish the oracle inequalities for a large variety of penalty functions. Implications of these on the regularized wavelet estimators are given in the next section.

By Theorem 2, property 2, for any penalized least squares estimator, we have

$$R_p(\theta, p_0) \le 2 \log n + \sqrt{4/\pi}(\log n)^{1/2} + 1 \tag{3.4}$$

if $p_0 \le \sqrt{2 \log n}$. This is a factor of $\log n$ order larger than the oracle estimator. The extra $\log n$ term is necessary because thresholding estimators create biases of order p_0 at $|\theta| \approx p_0$. The risk in $[0, 1]$ can be better bounded by using the following lemma.

Lemma 1. If the penalty function satisfies conditions of Theorem 1 and $p'_\lambda(\cdot)$ is nonincreasing and $p'_\lambda(0+) > 0$, then

$$R_p(\theta, p_0) \le (2 \log n + 2 \log^{1/2} n)\{c/n + \min(\theta^2, 1)\}$$

for the universal thresholding

$$p_0 = \sqrt{2 \log n - \log(1 + d \log n)}, \qquad 0 \le d \le c^2,$$

with $n \ge 4$ and $c \ge 1$ and $p_0 > 1.14$.

The results in Donoho and Johnstone (1994) correspond to the case $c = 1$. In this case, one can take the new universal thresholding as small as

$$p_0 = \sqrt{2\log n - \log(1+\log n)}. \tag{3.5}$$

Letting $c = 16$, we can take

$$p_0 = \sqrt{2\log n - \log(1+256\log n)}. \tag{3.6}$$

This new universal thresholding rule works better in practice.

A consequence of Lemma 1 is that

$$\Lambda_{n,c}(p)^* \le \Lambda_{n,c}(p) \le 2\log n + 2\log^{1/2} n. \tag{3.7}$$

Thus, the penalized least squares perform comparably with the oracle estimator within a logarithmic order. We remark that this conclusion holds for the thresholding parameter $p_0 = \sqrt{\alpha \log n}$ for any $\alpha \ge 2$. The constant factor in (3.7) depends on the choice of α, but the order of magnitude does not change.

The SCAD penalty leads to an explicit shrinkage estimator. The risk of the SCAD estimator of θ can be found analytically. To better gauge its performance, Table 1 presents the minimax risks for the SCAD shrink estimator, using the optimal thresholding and the new universal thresholding (3.5) and (3.6) for $c = 1$ and $c = 16$ and for several sample sizes n. The numerical values in Table 1 were computed by using a grid search over p_0 with increments .001. For a given p_0, the supremum over θ was computed by using a Matlab nonlinear minimization function.

Table 1 reveals that the new universal thresholding $a_{n,c}$ is much closer to the minimax thresholding p_n than that of the universal thresholding. This is particularly the case for $c = 16$. Further, the sharp minimax risk bound Λ^*_{n,c,a_n} with $c = 16$ is much smaller than the one with $c = 1$, used by Donoho and Johnstone (1994). The minimax upper bound Λ_{n,c,a_n} produced by new universal thresholding with $c = 16$ is closer to $\Lambda^*_{n,c}$. All these bounds are much sharper than the upper bound $b_{n,c}$.

Table 1. Coefficient p_n and Related Quantities for SCAD Penalty for Several Values of c and n

n	p_n	$a_{n,c}$[a]	$(2\log n)^{1/2}$	$\Lambda^*_{n,c}$	$\Lambda^*_n(DJ)$	Λ_{n,c,a_n}	$b_{n,c}$[b]
			$c = 1$				
64	1.501	2.584	2.884	3.086	3.124	7.351	12.396
128	1.691	2.817	3.115	3.657	3.755	8.679	14.110
256	1.881	3.035	3.330	4.313	4.442	10.004	15.800
512	2.061	3.234	3.532	5.013	5.182	11.329	17.472
1024	2.241	3.434	3.723	5.788	5.976	12.654	19.129
2048	2.411	3.619	3.905	6.595	6.824	13.978	20.772
			$c = 16$				
64	.791	1.160	2.884	1.346	3.124	1.879	12.396
128	.951	1.606	3.115	1.738	3.755	3.046	14.110
256	1.121	1.957	3.330	2.153	4.442	4.434	14.800
512	1.311	2.258	3.532	2.587	5.182	5.694	17.472
1024	1.501	2.526	3.723	3.086	5.976	7.055	19.129
2048	1.691	2.770	3.905	3.657	6.824	8.411	20.772

NOTE: The coefficient Λ^*_n(DJ) is computed by Donoho and Johnstone (1994) in their table 2 for the soft-thresholding estimator using the universal thresholding p_0.

[a] $a_{n,c} = (2\log n - \log(1+c^2\log n))^{1/2}$: the new thresholding parameter.

[b] $b_n = 2\log n + 2(\log n)^{1/2}$: the upper bound of minimax risk.

For $c = 1$, $\Lambda^*_{n,c}$ for the SCAD estimator is somewhat smaller than that of the soft-thresholding estimator Λ^*_n(DJ).

3.4 Performance of Regularized Wavelet Estimators

The preceding oracle inequalities can be directly applied to the regularized wavelet estimators defined via (2.3) when the sampling points are equispaced and $n = 2^J$. Suppose the data are collected from model (1.1). For simplicity of presentation, assume that $\sigma = 1$. Then, the wavelet coefficients $\mathbf{Z} = \mathbf{W}\mathbf{Y}_n \sim N(\boldsymbol{\theta}, I_n)$. Let

$$R_p(\hat{f}_p, f) = n^{-1}\sum_{i=1}^{n}\{\hat{f}_p(t_i) - f(t_i)\}^2$$

be the risk function of the regularized wavelet estimator $\hat{f}_p$. Let $R(\hat{f}_o, f)$ be the risk of the oracle wavelet thresholding estimator, which selects a term to estimate, depending on the value of unknown wavelet coefficients. Namely, $\hat{f}_o$ is the inverse wavelet transform of the ideally selected wavelet coefficients $\{Z_i I(|\theta_i| > 1)\}$. This is an ideal estimator and serves as a benchmark for our comparison. For simplicity of presentation, we assume that $i_0 = 1$.

By translating the problem in the function space into the wavelet domain, invoking the oracle inequalities (3.3) and (3.7), we have the following results.

Theorem 3. With the universal thresholding $p_0 = \sqrt{2\log n}$, we have

$$R_p(\hat{f}_p, f) \le \Lambda_{n,c}(p)\{cn^{-1} + R(\hat{f}_o, f)\}.$$

With the minimax thresholding p_n, we have the sharper bound:

$$R_p(\hat{f}_p, f) \le \Lambda^*_{n,c}(p)\{cn^{-1} + R(\hat{f}_o, f)\}.$$

Further, $\Lambda_{n,c}(p)$ and $\Lambda^*_{n,c}(p)$ are bounded by (3.7).

The risk of the oracle estimator is relatively easy to compute. Assume that the signal f is in a Besov ball. Because of simple characterization of this space via the wavelet coefficients of its members, the Besov space ball $B^r_{p,q}(C)$ can be defined as

$$B^r_{p,q} = \left\{f \in L_p : \sum_j \left(2^{j(r+1/2-1/p)}\|\boldsymbol{\theta}_{j\cdot}\|_p\right)^q < C\right\}, \tag{3.8}$$

where $\boldsymbol{\theta}_{j\cdot}$ is the vector of wavelet coefficients at the resolution level j. Here, r indicates the degree of smoothness of the underlying signal f. Note that the wavelet coefficients $\boldsymbol{\theta}$ in the definition of the Besov space are continuous wavelet coefficients. They are approximately a factor of $n^{1/2}$ larger than the discrete wavelet coefficients $\mathbf{W}\mathbf{f}$. This is equivalent to assuming that the noise level is of order $1/n$. By simplified calculations of Donoho et al. (1995), we have the following theorem.

Theorem 4. Suppose the penalty function satisfies the conditions of Lemma 1 and $r + 1/2 - 1/p > 0$. Then, the maximum risk of the penalized least squares estimator $\hat{f}_p$ over the Besov ball $B^r_{p,q}(C)$ is of rate $O(n^{-2r/(2r+1)}\log n)$ when the universal thresholding $\sqrt{2n^{-1}\log n}$ is used. It also achieves the rate of convergence $O(n^{-2r/(2r+1)}\log n)$ when the minimax thresholding $p_n/\sqrt{n}$ is used.

Thus, as long as the penalty function satisfies conditions of Lemma 1, regularized wavelet estimators are adaptively minimax within a factor of logarithmic order.

4. PENALIZED LEAST SQUARES FOR NONUNIFORM DESIGNS

The Sobolev wavelet interpolators introduced in Section 2 could be further regularized by a quadratic penalty in analogy with what is being done with smoothing splines. However, the estimators derived in this way, although easy to compute, are linear. They tend to oversmooth sharp features such as jumps and short aberrations of regression functions and, in general, will not recover such important attributes of regression functions. In contrast, nonlinear regularization methods, such as the ones studied in the previous sections, can efficiently recover such attributes. Our purpose in this section is to naturally extend the results of the previous sections to the general situation, in which the design matrix is no longer orthonormal.

Finding a solution to the minimization problem (2.3) cannot be done by using classical optimization algorithms, because the penalized loss $\ell(\boldsymbol{\theta})$ to be minimized is nonconvex, nonsmooth (because of the singularity of p at the origin), and high-dimensional. In this section, we introduce a ROSE to solve approximately the minimization problem (2.3). It is related to the one-step likelihood estimator and hence is supported by statistical theory (Bickel 1975; Robinson 1988).

4.1 Regularized One-Step Estimator

The following technique is used to avoid minimizing high-dimensional nonconvex functions and to take advantage of the orthonormality of the wavelet matrix $\mathbf{W}$. Let us again consider equation (1.1), and let us collect the remaining rows of the matrix $\mathbf{W}^T$ that were not collected into the matrix $\mathbf{A}$ into the matrix $\mathbf{B}$ of size $(N-n)\times N$. Then, the penalized least squares in expression (2.3) can be written as

$$\ell(\boldsymbol{\theta}) = 2^{-1}\|\mathbf{Y}^* - \mathbf{W}^T\boldsymbol{\theta}\|^2 + \sum_{i\geq i_0} p_\lambda(|\theta_i|),$$

where $\mathbf{Y}^* = (\mathbf{Y}_n^T, (\mathbf{B}\boldsymbol{\theta})^T)^T$. By the orthonormality of the wavelet transform,

$$\ell(\boldsymbol{\theta}) = 2^{-1}\|\mathbf{W}\mathbf{Y}^* - \boldsymbol{\theta}\|^2 + \sum_{i\geq i_0} p_\lambda(|\theta_i|). \tag{4.1}$$

If $\mathbf{Y}^*$ were given, this minimization problem can be easily solved by componentwise minimizations. However, we do not know $\boldsymbol{\theta}$, and one possible way is to iteratively optimize (4.1). Although this is a viable idea, we are not sure if the algorithm will converge. A one-step estimation scheme avoids this problem, and its theoretical properties can be understood. Indeed, in a completely different context, Fan and Chen (1999) show that the one-step method is as efficient as the fully iterative method, both empirically and theoretically, as long as the initial estimators are reasonably good. In any case, some good estimates of $\boldsymbol{\theta}$ are needed, by using either a fully iterative method or a one-step method.

We now use our Sobolev wavelet interpolators to produce an initial estimate for $\boldsymbol{\theta}$ and hence for $\mathbf{Y}^*$. Recall that $\hat{\boldsymbol{\theta}} = \mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{Y}_n$ was obtained via wavelet interpolation. Let

$$\widehat{\mathbf{Y}}_0^* = (\mathbf{Y}_n^T, (\mathbf{B}\hat{\boldsymbol{\theta}})^T)^T$$

be the initial synthetic data. By the orthonormality of $\mathbf{W}$, it is easy to see that

$$\hat{\boldsymbol{\theta}}^* = \mathbf{W}\widehat{\mathbf{Y}}_0^* \sim N(\boldsymbol{\theta}^*, \sigma^2\mathbf{V}), \tag{4.2}$$

where

$$\mathbf{V} = \mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-2}\mathbf{A}\mathbf{D} \quad \text{and} \quad \boldsymbol{\theta}^* = \mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{D}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\theta}$$

is the vector of wavelet coefficients. We call the components of $\mathbf{W}\widehat{\mathbf{Y}}_0^*$ the empirical synthetic wavelet coefficients. Note that $\boldsymbol{\theta}^*$ is the wavelet interpolation of the signal $\mathbf{f}_n$. It does not create any bias for the function f at observed data points, and the biases at other points are small (see Figure 1).

The empirical synthetic wavelet coefficients are nonstationary with a known covariance structure $\mathbf{V}$. Componentwise thresholding should be applied. Details are given in Section 4.2. Let $\hat{\boldsymbol{\theta}}_1^*$ be the resulting componentwise thresholding estimator. The resulting estimate $\hat{f}_1 = \mathbf{W}^T\hat{\boldsymbol{\theta}}_1^*$ is an (NRSI).

We do not have an automatic choice for the smoothing Sobolev interpolation parameter s. Although the interpolated function becomes smoother as s increases, it does not remove the noise in the observed signal, because the interpolated function must necessarily pass through all observed points. The regularization employed by NRSI yields reasonable interpolators allowing some errors in matching the given sample points.

As noted in Section 2, when $s = 0$, $\hat{\boldsymbol{\theta}} = \mathbf{A}^T\mathbf{Y}_n$ is easy to compute. In this case, the covariance matrix $\mathbf{V} = \mathbf{A}^T\mathbf{A}$ is also easy to compute. Its diagonal elements can be approximated by using the properties of wavelets.

As shown in Section 4.3, the NRSI possesses good sampling properties. One can also regard this estimator $\hat{\boldsymbol{\theta}}_1$ as an initial estimator and use it to create the synthetic data

$$\widehat{\mathbf{Y}}_1^* = (\mathbf{Y}_n^T, (\mathbf{B}\hat{\boldsymbol{\theta}}_1)^T)^T.$$

With the synthetic data, one can now minimize the penalized least squares

$$\ell(\boldsymbol{\theta}) = 2^{-1}\|\mathbf{W}\widehat{\mathbf{Y}}_1^* - \boldsymbol{\theta}\| + \sum_{i\geq i_0} p_\lambda(|\theta_i|) \tag{4.3}$$

by componentwise minimization technique. The resulting procedure is a one-step procedure with a good initial estimator. This procedure is the ROSE. By this one-step procedure, the interaction of λ with the parameter s is largely reduced, and, with the proper choice of the threshold, a large range of values of s yields estimates with minimal distorsion. According to Bickel (1975), Robinson (1988), and Fan and Chen (1999), such a procedure is as good as a fully iterated procedure when the initial estimators are good enough. Formal technical derivations of the statement are beyond the scope of this article.

4.2 Thresholding for Nonstationary Noise

As shown in (4.2), the noise in the empirical synthetic wavelet coefficients is not stationary, but their covariance matrix is known up to a constant. Thus, we can employ coefficient-dependent thresholding penalties to the empirical synthetic wavelet coefficients. This is an extension of the method of Johnstone and Silverman (1997), who extended wavelet thresholding estimators for data with stationary correlated Gaussian noise. In their situation, the variances of the wavelet coefficients at the same level are identical, so that they threshold the coefficients level by level with thresholds of the order $\sqrt{2\log N}\sigma_j$, where σ_j is a robust estimate of the noise level at the jth resolution of the wavelet coefficients.

Let v_i be the ith diagonal element of the matrix $\mathbf{V}$. Then, by (4.2), the ith synthetic wavelet coefficient, denoted by Z_i^*, is distributed as

$$Z_i^* \sim N(\theta_i^*, v_i\sigma^2). \tag{4.4}$$

The coefficient-dependent thresholding wavelet estimator is to apply

$$p_i = \sqrt{2v_i \log n}\,\sigma \tag{4.5}$$

to the synthetic wavelet coefficient Z_i^*. This coefficient-dependent thresholding estimator corresponds to the solution of (2.3) with the penalty function $\sum_{i\geq i_0}^N p_{\lambda_i}(|\theta_i|)$, where the regularization parameter λ_i is chosen such that p_i is the thresholding parameter for the ith coefficient:

$$\min_{\theta\geq 0}\{\theta + p'_{\lambda_i}(\theta)\} = p_i.$$

Invoking the oracle inequality with $c = 1$, the risk of this penalized least squares estimator is bounded by

$$E(\hat{\theta}_i - \theta_i^*)^2 \leq (2\log n + 2\log^{1/2} n) \\ \times [c\sigma^2 v_i/n + \min(\theta_i^{*2}, \sigma^2 v_i)]. \tag{4.6}$$

Averaging these over i, we obtain an oracle inequality similar to that of Donoho and Johnstone (1998) in the uniform design setting.

In the preceding thresholding, one can take $p_i = \sqrt{2v_i \log N}\,\sigma$. The result (4.6) continues to hold. The constant 2 in p_i can be replaced by any constant that is no smaller than 2.

In practice, the value of σ^2 is usually unknown and must be estimated. In the complete orthogonal case, Donoho et al. (1995) suggested the estimation of the noise level by taking the median absolute deviation of the coefficients at the finest scale of resolution and dividing it by .6745. However, in our setting, it is necessary to divide each synthetic wavelet coefficient by the square root of its variance v_i. Moreover, it can happen that some of these variances are close to 0 due to a large gap in the design, leading to values of synthetic wavelet coefficients that are also close to 0. Taking these into account, we suggest and have used the estimator

$$\hat{\sigma} = \mathrm{MAD}\{Z_{J-1,k}^*/\sqrt{v_{J-1,k}} : v_{J-1,k} > .0001\}/.6745,$$

where $Z_{J-1,k}^*$ is the synthetic wavelet coefficients at the highest resolution level $J-1$ and $v_{J-1,k}$ is its associated variance.

4.3 Sampling Properties

The performance of regularized wavelet estimators is assessed by the mean squared risk:

$$R_p(f) = n^{-1}\sum_{i=1}^{n} \mathbb{E}\{\hat{f}_p(t_i) - f(t_i)\}^2.$$

In terms of the wavelet transform for the NRSI, it can be expressed as

$$\begin{aligned} R_p(f) &= n^{-1}\mathbb{E}\{\|\mathbf{A}\hat{\boldsymbol{\theta}}_1 - \mathbf{A}\boldsymbol{\theta}\|^2\} \\ &= n^{-1}\mathbb{E}\{\|\mathbf{A}\hat{\boldsymbol{\theta}}_1 - \mathbf{A}\boldsymbol{\theta}^*\|^2\} \leq n^{-1}\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2. \end{aligned} \tag{4.7}$$

By (4.6), the mean squared errors are bounded as follows.

Theorem 5. Assume that the penalty function p satisfies the condition in Lemma 1. Then, the NRSI with coefficient-dependent thresholding satisfies

$$R_p(f) \leq n^{-1}(2\log n + 2\log^{1/2} n) \\ \times \left[c\sigma^2 \mathrm{tr}(\mathbf{V})/n + \sum \min(\theta_i^{*2}, \sigma^2 v_i)\right],$$

where tr($\mathbf{V}$) is the trace of matrix $\mathbf{V}$.

Note that when $s = 0$, the matrix $\mathbf{V} = \mathbf{A}^T\mathbf{A} \leq I_N$. Hence, $\mathrm{tr}(\mathbf{V}) \leq N$ and $v_i \leq 1$.

The NRSI was used only as an initial estimator to the penalized least squares estimator (1.2). We consider its performance over the Besov space $B_{p,q}^r$ for the specific case with $s = 0$. To this end, we need some technical conditions. First, we assume that $N/n = O(\log^a n)$ for some $a > 0$. Let G_n be the empirical distribution function of the design points $\{t_1, \ldots, t_n\}$. Assume that there exists a distribution function $G(t)$ with density $g(t)$, which is bounded away from 0 and infinity such that

$$G_n(t) \to G(t) \qquad \text{for all } t \in (0,1) \text{ as } n \to \infty.$$

Assume further that $g(t)$ has the rth bounded derivative. When r is not an integer, we assume that the $[r]$ derivative of g satisfies the Lipschitz condition with the exponent $r - [r]$, where $[r]$ is the integer part of r.

To ease the presentation, we now use double indices to indicate columns of the wavelet matrix $\mathbf{W}$. Let $W_{j,k}(i)$ be the element in the ith row and the (j,k)th column of wavelet matrix $\mathbf{W}^T$, where j is the resolution level and k is the dyadic location. Let ψ be the mother wavelet associated with the wavelet transform $\mathbf{W}$. Assume that ψ is bounded with a compact support and has first $r-1$ vanishing moments. Then,

$$W_{j,k}(i) \approx 2^{-(J-j)/2}\psi(2^j i/N - k)$$

for i, j, k not too close to their boundaries. To avoid unnecessary technicalities, which do not provide us insightful understanding, we assume

$$W_{j,k}(i) = 2^{-(J-j)/2}\psi(2^j i/N - k) \quad \text{for all } i, j, k.$$

As in Theorem 4, we assume that $\sigma^2 = n^{-1}$.

Theorem 6. Suppose that the penalty function satisfies the conditions of Lemma 1 and $r + 1/2 - 1/p > 0$. Then, the maximum risk of the nonlinear regularized Sobolev interpolator over a Besov ball $B^r_{p,q}$ is of rate $O(n^{-2r/(2r+1)} \log n)$ when the universal thresholding rule is used. It achieves the rate of convergence $O(n^{-2r/(2r+1)} \log n)$ when the minimax thresholding $p_n/\sqrt{n}$ is used.

5. NUMERICAL EXAMPLES

In this section, we illustrate our penalized least squares method by using three simulated datasets and two real data examples. The NRSI is used as an initial estimate. The ROSE method is employed with the SCAD penalty and with coefficient-dependent universal thresholding penalties given by (4.5).

For simulated data, we use the functions heavisine, blocks, and doppler in Donoho and Johnstone (1994) as testing functions. The noise level is increased so that the signal-to-noise ratio is around 4. This corresponds to taking $\sigma = .5$ for the heavisine function, $\sigma = .3$ for the blocks function, and $\sigma = .4$ for the doppler function. A random sample of size 100 is simulated from model (1.1). The design points are uniformly distributed on $[0, 1]$, but they are not equispaced. For the doppler function, we set the x_i's by choosing 100 design points from a standard normal distribution and then rescaling and relocating the order statistics of the absolute values of the 100 samples (this yields a denser sampling of the first portion of the Doppler signal, which is the region over which it is varying rapidly). The simulated data and the testing functions are shown in Figure 4, (a), (c), (e). The ROSE estimates were computed by using the symmlets of order 6 and $s = 3$ for the heavisine and the doppler functions and the Haar wavelets and $s = .5$ for the blocks function. As an indication of the relative computation speed of the NSRI and ROSE methods, we report the CPU time in seconds needed by the MATLAB process to run NSRI, and then ROSE, for producing one estimate of the heavisine function under such a setting: CPU-NSRI = 1.6144e-01 and CPU-ROSE = 1.1840e-02. The larger value for the NSRI CPU time is due to the initialization of the Sobolev interpolation over the appropriate grid.

As one can see from the figures, the blocks data are very sensitive to small gaps in the design because they have a lot of discontinuities. The resulting fit for the doppler function oversmooths the first portion of the Doppler signal, because the random design does not catch the rapid oscillations of the signal very well. On the other hand, the fit for the heavisine case is much smoother and better. Note, however, the bias in the discontinuous parts of the heavisine function due to the wavelet NRSI initial estimate with $s = 3$.

We now report on Monte Carlo experiments conducted to compare the performance of ROSE with Hall and Turlach's interpolation method (HALL/TURL) and the one of Kovac and Silverman (KOVAC/SILV). We used as test functions the heavisine and blocks functions, normalized such that their standard deviations were equal to 5. We set up the x_i's by choosing $n = 100$ design points from a standard normal distribution and then rescaling and relocating their order statistics such that the first and last values were 0 and 1, with a discretization (used for Sobolev interpolation) of the $[0, 1]$ interval into a regular grid of length $N = 256$. For each set of x_i's so chosen, we simulated a noisy function with $\sigma^2 = 1$ (i.e., the standard deviation of the noise is 5 times smaller than the standard deviation of the function). Each noisy function was then used to estimate the true f over the selected x_i's by using each of the three wavelet estimation techniques. The quality of the estimate was measured by computing the observed risk, namely,

$$\widehat{R}(\hat{f}, f) = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{f}(x_i) - f(x_i)\big)^2.$$

To eliminate effects that might be attributable to a particular choice of the x_i's, we reselected the x_i's for each simulation. We repeated the above 200 times to obtain 200 observed risks for each function and each wavelet estimation method. We used SCAD to select the threshold for ROSE (with $s = 3$ for heavisine and $s = .5$ for blocks), SURE to select the threshold for the Silverman and Kovac algorithm, and a universal threshold adjusted to the maximal spacing of the design points for the Hall and Turlach procedure, as advocated by them. We assumed knowledge of the noise variance σ^2 in computing the thresholds.

Figure 5 displays the boxplots of the observed risks and for the heavisine and blocks functions. For the heavisine, ROSE appears to be the best technique, whereas KOVAC/SILV looks slightly better for the blocks. The universal thresholding rule advocated by Hall and Turlach leads to a larger risk than do the other two methods. We can conclude that the ROSE wavelet scheme works well.

As an example of our regularized wavelet estimation procedure to an actual unequally sampled time series, we consider the problem of estimating the light curve for the variable star RU Andromeda. This time series was obtained from the American Association of Variable Star Observers international database, which is maintained by J. A. Mattei and is accessible at www.aavso.org. It was used by Sardy et al. (1999) to illustrate the performance of their Haar wavelet–based procedures for denoising unequally sampled noisy signals. The observed magnitude values for this star are indicated in Figure 6 by small dots, which range in time from Julian Day 2,449,004 to 2,450,352 (January 1993 to mid 1996). The magnitudes of this star are measured at irregularly spaced times because of blockage of the star by sunlight, weather conditions, and availability of telescope time. There were 295 observations in all, three of which were reported as upper limits on the star's magnitude and hence were eliminated because their error properties are quite different from the remaining observations. Of the 292 remaining observations, we selected 100 observations at random from the first 256 values to conform to the assumption made throughout this article.

The ROSE method is employed with the Symmlets of order 6, $s = 3$, and the SCAD penalty with coefficient-dependent

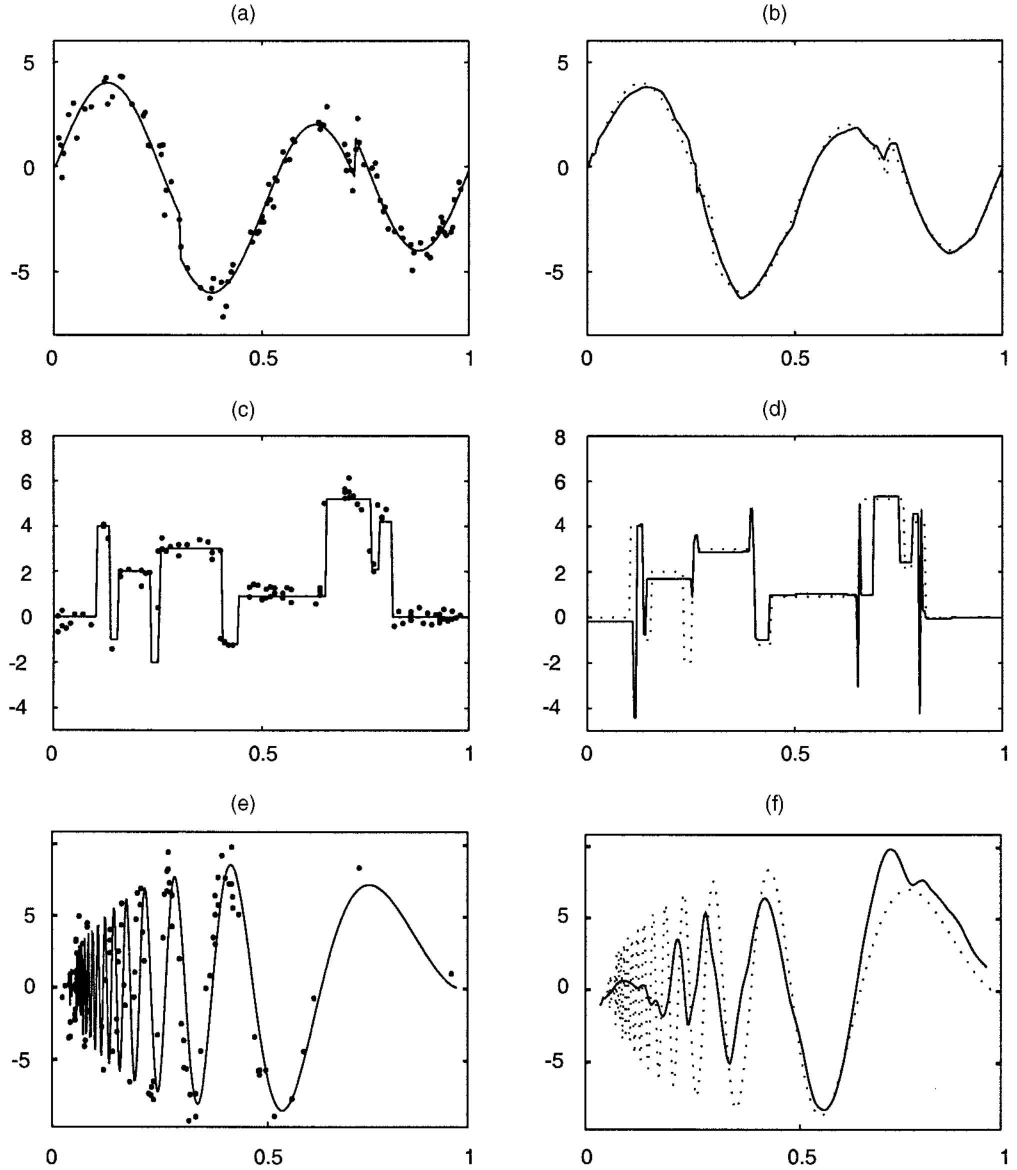

Figure 4. Estimates by Using ROSE for Two Simulated Data. (a), (c), (e) Simulated data and true regressions; (b), (d), (f) estimate by using ROSE (solid curve) and true regressions (dashed curves) with $s = 3$, $s = .5$, $s = 3$.

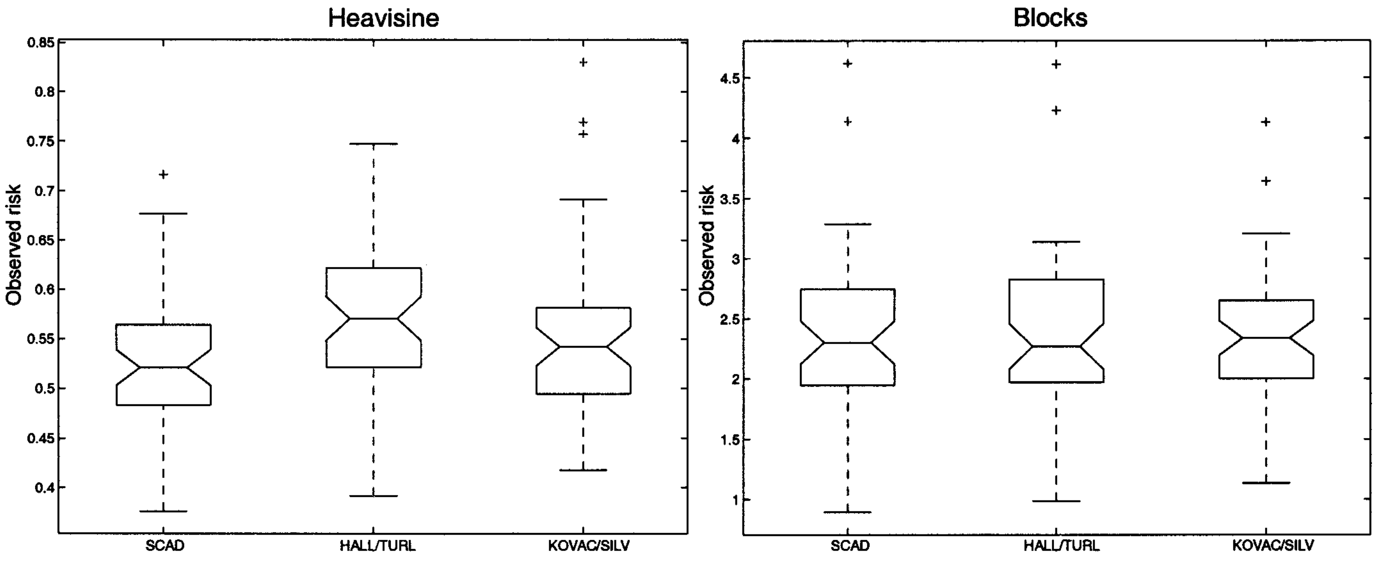

Figure 5. Boxplots of Observed Risks and for Heavisine and Blocks Functions for Each Wavelet Method.

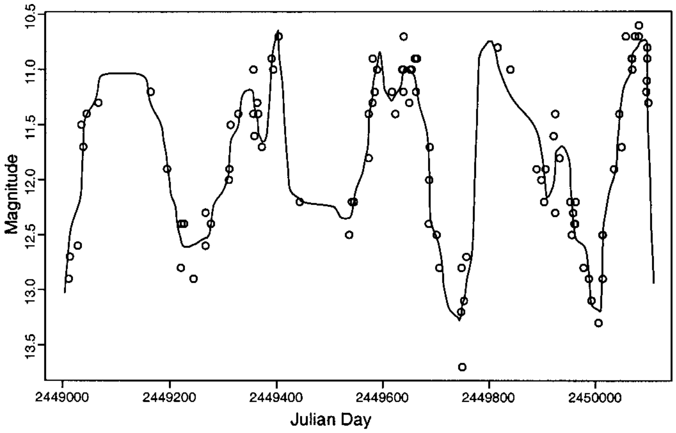

Figure 6. Light Curve for Variable Star RU Andromeda. Observed data (points) and ROSE estimates (solid curve).

universal thresholds given by (4.5). The estimated light curve is indicated in Figure 6 by a solid line. Compared with the findings in Sardy et al. (1999), the estimated light curve tracks the overall light variations quite nicely.

Figure 7 shows another dataset that was analyzed extensively in the field of nonparametric regression. It was discussed by Silverman (1985) and consists of 133 observations from a crash test and shows the acceleration of a motorcyclist's head during a crash. Classical wavelet thresholding or the interpolation method of Hall and Turlach (1997) for unequally spaced data produce wiggly estimates, like those in the first row of Figure 5. In both cases, VisuShrink was applied, and the Symmlets of order 6 were used. Both estimates exhibit large high-frequency phenomena. The second row in Figure 7 displays a robust estimate obtained by cleaning the data from outliers and extreme observations by median filtering and then using wavelet thresholding on linearly interpolated data on a regular grid, as suggested by Kovac and Silverman (1999), and the ROSE estimate on the same dataset with a 256-point Sobolev interpolation using the Symmlets of order 6, $s = 3$, and the SCAD penalty with coefficient-dependent universal thresholds given by (4.5). Both estimates are obviously less disturbed by the outliers in the crash dataset; there are no longer any high-frequency phenomena. This example shows that ROSE by itself is quite robust to outliers.

6. OTHER APPROACHES

In this section, we present an alternative approach to estimate regression functions from nonequispaced samples by using the GNC algorithm to find a local minimum of the penalized least squares problem (2.3). This method is more computationally intensive than the NRSI and ROSE, and its implementations depend on a number of tuning parameters. Nevertheless, it offers nice ideas for optimizing high-dimensional nonconvex functions.

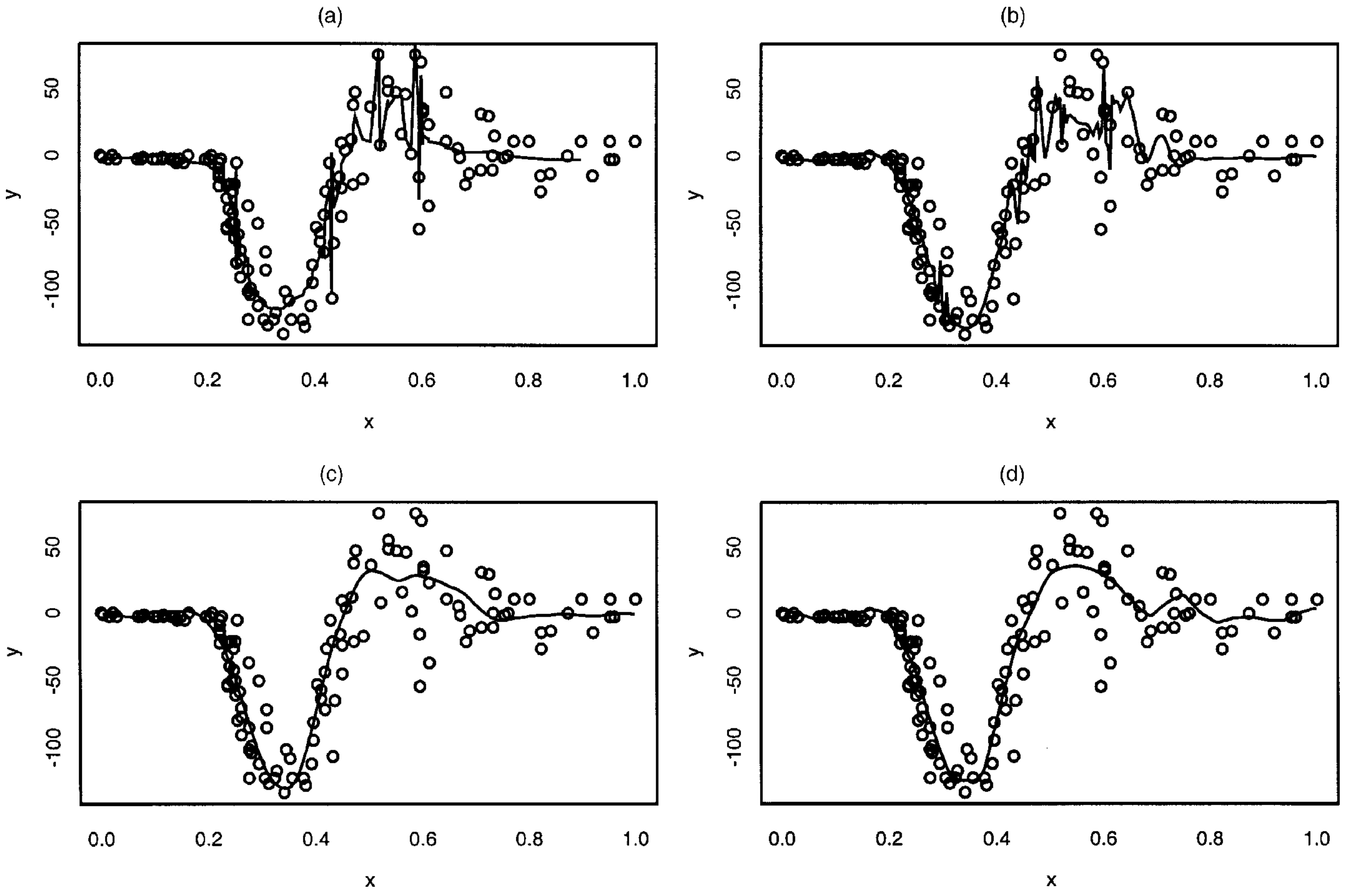

Figure 7. Crash Data With Several Wavelet Estimates. (a) Classical Wavelet Thresholding, (b) Thresholding for Unequally Spaced Data, (c) Robust Thresholding, (d) ROSE estimation with s = 3.

6.1 Graduated Nonconvexity Algorithm

The graduated nonconvexity algorithm was developed in the image processing context (Blake and Zisserman 1987; Blake 1989). It can minimize a broad range of nonconvex functions. Basically, the GNC algorithm can be seen as a deterministic relaxation technique (Blake 1989) that substitutes a sequence of local minimizations along a sequence of approximate (relaxed) functions ℓ_{r_k} for the minimization of ℓ. Here, $\{r_k\}_{k=0}^K$ is an increasing sequence of positive relaxation parameters that are similar to the cooling temperatures in the simulated annealing. The first relaxed objective function ℓ_{r_0} is strictly convex, and hence its minimization can be found by using standard techniques. A local minimizer of $\ell_{r_k}(\boldsymbol{\theta})$ serves as the initial value for minimization of $\ell_{r_{k+1}}(\boldsymbol{\theta})$. The last one fits the function ℓ, which is the object that we want to minimize.

The GNC algorithm requires the family of relaxed functions ℓ_r, depending on a parameter $r \in (0, 1)$, to satisfy the following conditions:

1. The functions $\ell_r(\boldsymbol{\theta})$ are C^1-continuous in $\boldsymbol{\theta}$ and continuous in r.
2. The concavity of ℓ_r is relaxed monotonously when r decreases.
3. There exists $r_0 > 0$ such that ℓ_r is strictly convex for any $r \le r_0$.
4. $\lim_{r\to 1} \ell_r(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta})$.

Thus, the function ℓ_r has a unique minimum for $r \le r_0$. When r increases to one, the local minima progressively approach a local minima of the object function ℓ.

The implementation of the algorithm depends on the choice of relaxation sequence $\{r_k\}_{k=0}^K$. The GNC minimization starts from calculating the unique minimum $\hat{\boldsymbol{\theta}}_{r_0}$ of ℓ_{r_0}. Afterward, for each r_k, an intermediate minimum $\hat{\boldsymbol{\theta}}_{r_k}$ of ℓ_{r_k} is calculated by a local descent method in a vicinity of previously obtained intermediate minimum; namely, $\hat{\boldsymbol{\theta}}_{r_k}$ is obtained by iterating a local decent algorithm with the initial value $\hat{\boldsymbol{\theta}}_{r_{k-1}}$. The final estimate is $\hat{\boldsymbol{\theta}}_{r_K}$.

The closeness of the ultimate estimate $\hat{\boldsymbol{\theta}}_{r_K}$ to the global minimum of ℓ depends critically on the sequence of relaxed functions. It is therefore reasonable to require that the relaxed functions ℓ_r closely approximate the original functional ℓ.

6.2 Applications to Penalized Least Squares

The success of a GNC optimization to compute estimates corresponding to nonsmooth penalties in Section 2.3 closely depends on the pertinence of the approximation involved in the relaxed penalized functions. An extension of the GNC algorithm to ill posed linear inverse problems and a systematic way to calculate initializations for which a local minimization of ℓ provides meaningful estimates was given by Nikolova et al. (1999). Here, we briefly summarize key ideas in Nikolova et al. (1999) and extend the GNC algorithm to our case. To facilitate notation, we drop the dependence of notation λ and rewrite (2.3) as

$$\ell(\boldsymbol{\theta}) = 2^{-1}\|\mathbf{Y}_n - A\boldsymbol{\theta}\|^2 + \sum_{i\ge i_0} p(|\theta_i|). \tag{6.1}$$

In our applications, the nonconvexity comes from nonconvexity penalties. Hence, we need only to relax the penalized term in (6.1). Penalty functions satisfying the conditions of Theorem 1 have strictly concave parts, but their concavity vanishes at infinity, namely, the second derivative at infinity is nonnegative. They usually reach their maximum concavity at some finite point. More precisely, let

$$p''(t) = \lim_{\varepsilon\to 0} \varepsilon^{-2}\{p(t+\varepsilon) + p(t-\varepsilon) - 2p(t)\} \qquad \text{for } t > 0,$$

and let T be the largest minimizer of $p''(\cdot)$ over $t > 0$. That is, T is the location where the maximum concavity $\inf_{t\in\mathbb{R}^+} p''(t)$ of the function p occurs. Given such a penalty function, a relaxed penalty p_r should satisfy the following conditions (Nikolova et al., to appear):

1. The functions $p_r(|t|)$ are C^1-continuous in t, and for any t fixed they are continuous in r.
2. $p_r(|t|)$ should not stray too much from $p(|t|)$ for each r and $\lim_{r\to 1} p_r(|t|) = p(|t|)$.
3. The maximum concavity of $p_r(|t|)$, occurring at T_r, is required to increase continuously and strictly monotonously toward 0 as $r \to r_0$ so that p_{r_0} is a convex function.

An appropriate choice of a relaxed penalty is usually based on the closeness of T_r to the original T and the way T_r decreases toward T as r increases toward 1. One way to construct such relaxed penalties p_r is to fit splines in the vicinity of the points where p is not differentiable and nonconvex. This technique was proposed by Blake and Zisserman (1987) for the relaxation of a clipped quadratic penalty.

To ensure the convexity of initial approximation

$$\ell_r(\boldsymbol{\theta}) = 2^{-1}\|\mathbf{Y}_n - A\boldsymbol{\theta}\|^2 + \sum_{i\ge i_0} p_r(|\theta_i|),$$

it is necessary to find an r such that the Hessian matrix of ℓ_r is nonnegative definite for any $\boldsymbol{\theta}$:

$$A^T A + P_r''(\boldsymbol{\theta}) > 0 \quad \text{for all } \boldsymbol{\theta},$$

where $P_r(\boldsymbol{\theta}) = \sum_{i\ge i_0} p_r(|\theta_i|)$ and $P_r''(\boldsymbol{\theta})$ is its corresponding Hessian matrix. Because the matrix $A^T A$ is singular and p_r has its concave parts, such a condition is difficult to fulfill. Thus, some modifications on family of relaxation p_r for r near r_0 are needed. A possible way to do this is to render convexity of the initial relaxed penalty p_r, as done by Nikolova et al. (in press).

Take a number $\rho \in (r_0, 1)$. With slight abuse of notation, modify the definition of P_r for $r \in [r_0, \rho]$ as

$$P_r(\boldsymbol{\theta}) = P_\rho(\boldsymbol{\theta}) + \frac{\rho - r}{\rho - r_0} Q(\boldsymbol{\theta}),$$

where $Q(\boldsymbol{\theta}) = \sum_i q(|\theta_i|)$ for a convex function q. To ensure the convexity of P_{r_0}, Q has to compensate for the nonconvex parts of P_ρ, and at the same time Q should not deform P_ρ too much. The auxiliary penalty q should be C^1-continuous and symmetric with $q(0) = 0$. A possible choice of the function q is given by

$$q(|t|) = \{p_\rho(u_\rho) - p_\rho(|t|) \\ + (|t| - u_\rho)\dot{p}_\rho(u_\rho)\} I(|t| \geq u_\rho), \qquad (6.2)$$

where $u_\rho > 0$ is such that p_ρ is strictly convex over the interval $|t| < u_\rho$.

An illustration, let us consider the transformed L_1 penalty function (2.11), which has been used in the context of image processing for restoring blurred images by Nikolova et al. (in press). For this type of penalty, the maximum concavity occurs at $T = 0$ with the minimum of the second derivative $-2b^2$. Consider the family of relaxed functions

$$p_r(|t|) = \begin{cases} \frac{b_r t^2}{1+c_r t^2} & \text{if} \quad |t| < \frac{1-r}{r}, \\ \frac{b|t|}{1+b|t|} & \text{if} \quad |t| \geq \frac{1-r}{r} \end{cases} \qquad (6.3)$$

with $b_r = \frac{2rb}{1-r}$ and $c_r = \frac{r(r+2b-2br)}{(1-r)^2}$. The penalty and its relaxed form are depicted in Figure 8(c). The constants b_r and c_r are determined by the C^1-continuity of the function p_r. The maximum concavity occurs at $T_r = \frac{1}{c_r} < \frac{1-r}{r}$ with the minimum of the second derivative $-rb/(1-r)$. This initial choice of relaxed functions are not convex for all $r > 0$. Thus, we appendix a convex term according to (6.3):

$$q(|t|) = \begin{cases} 0 & \text{if } |t| < u_\rho, \\ \frac{b_\rho}{4c_\rho} - p_\rho(|t|) + (|t| - u_\rho)g & \text{if } |t| \geq u_\rho, \end{cases}$$

where $u_\rho = \frac{1}{\sqrt{3c_\rho}}$ and $g = \frac{9b_\rho}{8\sqrt{3c_\rho}}$. As an illustration of the GNC algorithm, we simulated 100 data points from the heavisine function with a signal-to-noise ratio about 3. The data and the true regression function are shown in Figure 8(a). We apply the GNC algorithm with the number of relaxing steps $K = 40$ to solve the penalized least squares problem (2.3) with $\lambda = 6$ and penalty function (2.11) with $b = 3.7$. The GNC algorithm found a reasonably good estimate, which is superimposed as a solid line to the true function (dashed curve) in Figure 8(b).

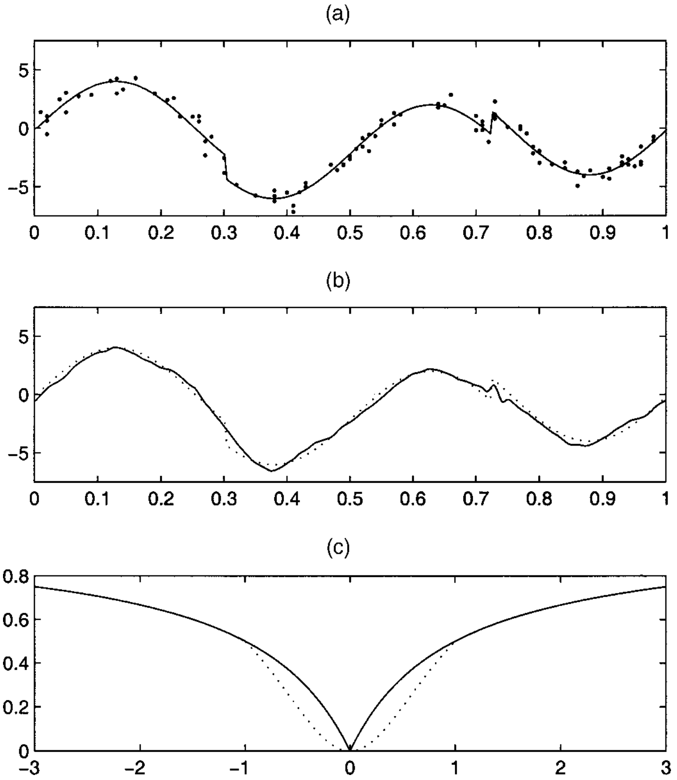

Figure 8. GNC Algorithm. (a) The data (points) and the true regression function (solid curve). (b) The unknown function is computed by solving (2.3) using the GNC algorithm; dashed curve: the true function; solid curve: estimated function. (c) Relaxing the concave penalty (2.11) with b = 3.7 (solid curve) by using a relaxing function p_r with r = .5 (dashed curve) defined by Equation (6.3).

APPENDIX A: PROOF OF THEOREM 1

The existence of the solution was noted. When $z = 0$, it is clear that $\hat{\theta}(z) = 0$ is the unique minimizer. Without loss of generality, assume that $z > 0$. Then, for all $\theta > 0$, $\ell(-\theta) > \ell(\theta)$. Hence, $\hat{\theta}(z) \geq 0$. Note that for $\theta > 0$,

$$\ell'(\theta) = \theta - z + p'_\lambda(\theta).$$

When $z < p_0$, the function ℓ is strictly increasing on $(0, \infty)$ because the derivative function is positive. Hence, $\hat{\theta}(z) = 0$. When the function $\ell'(\theta)$ is strictly increasing, there is at most one zero-crossing, and hence the solution is unique. Thus, we only need to consider the case that $\ell'(\theta)$ has a valley on $(0, \infty)$ and $z > p_0$. In this case, there are two possible zero-crossings for the function ℓ' on $(0, \infty)$. The larger one is the minimizer because the derivative function at that point is increasing. Hence, the solution is unique and satisfies

$$\hat{\theta}(z) = z - p'_\lambda(\hat{\theta}(z)) \leq z. \qquad \text{(A.1)}$$

Thus, $\hat{\theta}(z) \leq z - p'_\lambda(z)$ when $p'_\lambda(\cdot)$ is nonincreasing. Let θ_0 be the minimizer of $\theta + p'_\lambda(\theta)$ over $[0, \infty)$. Then, from the preceding argument, $\hat{\theta}(z) > \theta_0$ for $z > p_0$. If $p'_\lambda(\cdot)$ is nonincreasing, then

$$p'_\lambda(\hat{\theta}(z)) \leq p'_\lambda(\theta_0) \leq \theta_0 + p'_\lambda(\theta_0) = p_0.$$

This and (A.1) prove result 3. It is clear that continuity of the solution $\hat{\theta}(z)$ at the point $z = p_0$ if and only if the minimum of the function $|\theta| + p'_\lambda(|\theta|)$ is attained at 0. The continuity at other locations follows directly from the monotonicity and continuity of the function $\theta + p'_\lambda(\theta)$ in the interval $(0, \infty)$. The last conclusion follows directly from (A.1). This completes the proof.

APPENDIX B: PROOF OF THEOREM 2

First, $R_p(\theta, p_0)$ is symmetric about 0 by Theorem 1, result 1. Thus, we can assume without loss of generality that $\theta \geq 0$. By Theorem 1, results 1 and 2,

$$E(\hat{\theta} - \theta)^2 \leq E(Z - \theta)^2 I(\hat{\theta} \notin [0, \theta]) + \theta^2 P(\hat{\theta} \in [0, \theta]) \\ \leq 1 + \theta^2. \qquad \text{(B.1)}$$

To prove result 2, we note that $E(\hat{\theta} - \theta)^2 = 1 + 2E(Z - \theta)\ (\hat{\theta} - Z) + E(\hat{\theta} - Z)^2$. For $Z > \theta$, we have $\hat{\theta} \leq Z$ by Theorem 1, result 3, which implies that $(Z - \theta)(\hat{\theta} - Z) \leq 0$. Similarly, for $Z < 0$,

$(Z-\theta)(\hat{\theta}-Z) \le 0$. Thus, $E(\hat{\theta}-\theta)^2 \le 1+2E(\theta-Z)(Z-\hat{\theta})I(0 \le Z \le \theta)+E(\hat{\theta}-Z)^2$. By Theorem 1, result 3, $|\hat{\theta}-Z| \le p_0$. Thus,

$$E(\hat{\theta}-\theta)^2 \le 1+2p_0E(\theta-Z)I(Z\le\theta)+p_0^2 \le 1+p_0\sqrt{2/\pi}+p_0^2.$$

This establishes result 2.

Result 3 follows directly from the fact that

$$R_p(0,p_0) \le EZ^2I\{|Z| \ge p_0\}.$$

To show result 4, using the fact that $R'_p(0,p_0)=0$ due to symmetry, we have by the Taylor expansion that

$$R_p(\theta,p_0) \le R_p(0,p_0) + \frac{1}{2}\sup_{0\le\eta\le1} R''_p(\eta,p_0)\theta^2 \qquad \text{for } \theta \in [-1,1]. \qquad \text{(B.2)}$$

We now compute the second derivative. Let $\phi(\cdot)$ be the standard normal density. Then, by simple calculation, we have

$$\begin{aligned} R'_p(\theta,p_0) &= \int_{-\infty}^{\infty}(\theta+z-2\hat{\theta})\phi(z-\theta)dz \\ &= 2\theta - 2\int_{-\infty}^{\infty}\hat{\theta}\phi(z-\theta)dz \end{aligned}$$

and $R''_p(\theta,p_0)=2+2E\hat{\theta}(\theta-Z)$.

By using the same arguments as those in the proof of result 2, we have for $\theta>0$

$$R''_p(\theta,p_0) \le 2+2E\hat{\theta}(\theta-Z)I(0\le Z\le\theta).$$

Noting that $\hat{\theta}=0$ for $|Z|\le p_0$, we have for $p_0 \ge 1$ $R''_p(\theta,p_0)\le 2$. For the general case, using the fact that $|\hat{\theta}|\le|Z|$, we have for $\theta\in[0,1]$

$$\begin{aligned} R''_p(\theta,p_0) &\le 2+2\theta E(\theta-Z)I(0\le Z\le\theta) \\ &= 2+\sqrt{2/\pi}\theta(1-\exp(-\theta^2/2)) \le 4. \end{aligned}$$

By (B.2), result 4 follows for $\theta\in[-1,1]$. For θ outside this interval, 4 follows from (B.1). This completes the proof.

APPENDIX C: PROOF OF LEMMA 1

For $|\theta|>1$, by (3.4), we have for $n\ge4$

$$R_p(\theta,p_0) \le 2\log n + 2(\log n)^{1/2}.$$

Thus, we need to show that the inequality holds for $\theta\in[0,1]$. First, by Theorem 2, result 4,

$$R_p(\theta,p_0) \le R_p(0,\theta)+2\theta^2.$$

Let $g(\theta)=(R_p(0,p_0)+2\theta^2)/(c/n+\theta^2)$. If $R_p(0,p_0)\le 2c/n$, then $g(\theta)\le 2\le 2\log n$. Hence, the result holds. When $R_p(0,p_0) > 2c/n$, $g(\theta)$ is monotonically decreasing and hence $g(\theta)\le g(0)=c^{-1}nR_p(0,p_0)$. By Theorem 2, result 3, we have

$$\begin{aligned} g(\theta) &\le nc^{-1}p_0(1+p_0^{-2})\sqrt{2/\pi}\exp(-p_0^2/2) \\ &\le 2\pi^{-1/2}c^{-1}(1+p_0^{-2})(\log n)^{1/2}(1+d^{1/2}(\log n)^{1/2}). \end{aligned}$$

By using the fact that for $p_0>1.14$, $\pi^{-1/2}(1+p_0^{-2})\le1$, we conclude that $g(\theta)\le 2c^{-1}d^{1/2}(\log n)+2c^{-1}(\log n)^{1/2}$.

APPENDIX D: PROOF OF THEOREM 4

Write $\mathbf{Z}=(Z_{j,k})$ and $\boldsymbol{\theta}=(\theta_{j,k})$, $j=0,\ldots,J-1$, $k=1,\ldots,2^j$, where $Z_{j,k}$ and $\theta_{j,k}$ are the wavelet coefficients at the jth resolution level. Then, by the model assumption, $Z_{j,k}\sim N(\theta_{j,k},n^{-1})$. By Theorem 3, we need only to compute the maximum risk of the oracle estimator $\hat{\theta}^o_{j,k}=Z_{j,k}I(|Z_{j,k}|>n^{-1})$. Note that under the $n^{1/2}$-scale transform between the discrete and continuous wavelet coefficients, the risk function for the oracle estimator becomes

$$R(\hat{f}_o,f)=\sum_{j=1}^{J-1}\sum_{k=1}^{2^j}E(\hat{\theta}^o_{j,k}-\theta_{j,k})^2.$$

Now, the risk for the componentwise oracle estimator is known to be

$$E(\hat{\theta}^o_{j,k}-\theta_{j,k})^2=\min(\theta^2_{j,k},n^{-1})=n^{-1}\{\min(\sqrt{n}|\theta_{j,k}|,1)\}^2. \qquad \text{(D.1)}$$

Choose an integer J_0 such that $2^{J_0}=n^{1/(2r+1)}$. Then, it follows from (D.1) that

$$\sum_{j=0}^{J_0}\sum_k E(\hat{\theta}^o_{j,k}-\theta_{j,k})^2 \le 2^{J_0+1}/n = O(n^{-2r/(2r+1)}). \qquad \text{(D.2)}$$

For $p\le2$, by (D.1), we have

$$\sum_{j=J_0+1}^{J-1}\sum_k E(\hat{\theta}^o_{j,k}-\theta_{j,k})^2 \le n^{-1}\sum_{j=J_0+1}^{J-1}\sum_k(\sqrt{n}|\theta_{j,k}|)^p.$$

By the definition of the Besov ball, the last expression is bounded by

$$\begin{aligned} C^{p/q}n^{-1+p/2}\sum_{j=J_0+1}^{J-1}2^{-jap} &= O(n^{-1+p/2}2^{-J_0ap}) \\ &= O(n^{-2r/(2r+1)}), \qquad \text{(D.3)} \end{aligned}$$

where $a=r+1/2-1/p$. A combination of (D.2) and (D.3) yields

$$R_p(\hat{f}_p,f)=\sum_{j=0}^{J-1}\sum_k E(\hat{\theta}^o_{j,k}-\theta_{j,k})^2=O(n^{-2r/(2r+1)})$$

uniformly for all $\boldsymbol{\theta}\in B^r_{p,q}(C)$. We now need only to deal with the case $p>2$. Note that $\|\boldsymbol{\theta}_{j\cdot}\|_2 \le 2^{(1/2-1/p)j}\|\boldsymbol{\theta}_{j\cdot}\|_p$, because $\boldsymbol{\theta}_{j\cdot}$ has 2^j elements. It follows from this that $B^r_{p,q}\subset B^r_{2,q}$. The conclusion follows from the result for the case $p=2$.

APPENDIX E: PROOF OF THEOREM 6

As one expects, rigorous proof of this theorem involves a lot of technicalities, such as approximating discrete summations by their continuous integrations for wavelet coefficients below a certain resolution level. In fact, some of these approximations at high-resolution levels are not valid, and one can modify the estimator slightly without estimating wavelet coefficients above a certain level. For these reasons, we will only outline the key ideas of the proof without taking care of nonintrinsic parts of technicalities. Hence, the key ideas and the intrinsic parts of the proofs are highlighted.

As noted, $v_{j,k}\le1$ because $\mathbf{V}\le I_N$. By Theorem 5 and noting the factor $n^{-1/2}$ difference between the discrete and continuous wavelet coefficients, we have

$$\begin{aligned} R_p(f) &\le [2\log n+2(\log n)^{1/2}] \\ &\quad\times\left[N/n^2+\sum_{j=1}^{J-1}\sum_{k=1}^{2^j}\min(\theta^{*2}_{j,k},n^{-1})\right]. \qquad \text{(E.1)} \end{aligned}$$

Thus, we need only to show that $\boldsymbol{\theta}^* \in B^r_{p,q}$. Note that $\boldsymbol{\theta}^* = \mathbf{A}^T \mathbf{f}_n$. Thus,

$$\begin{aligned}\theta^*_{j,k} &= 2^{-J)2} \int_0^1 \psi_{j,k}(t) f(t) \, dG_n(t) \\ &= 2^{-J)2} \int_0^1 \psi_{j,k}(t) f(t) g(t) \, dt (1 + o(1)).\end{aligned}$$

Because f is in the Besov ball $B^r_{p,q}(C)$ and g is continuously differentiable with a derivative bounded away from 0, it follows that fg also belongs to a Besov ball $B^r_{p,q}(C')$ with $C' \geq C$. The factor $2^{-J)2}$ is the difference between the discrete and continuous wavelet coefficients. Therefore, $\boldsymbol{\theta}^* \in B^r_{p,q}(C')$. By (E.1), we have

$$R_p(f) = O(\log n) \left[N) n^2 + \sum_{j,k} \min(\theta^{*2}_{j,k}, n^{-1}) \right].$$

The result follows from the proof of Theorem 4.

[Received December 1999. Revised November 2000.]

REFERENCES

Abramovich, F., Bailey, T. C., and Sapatinas, T. (2000), "Wavelet Analysis and Its Statistical Applications," *The Statistician*, 49, 1–29.

Amato, U., and Vuza, D. T. (1997). "Wavelet Approximation of a Function From Samples Affected By Noise," *Revue Roumaine de Mathématiques Pures et Appliquées*, 42, 481–493.

Antoniadis, A. (1996), "Smoothing Noisy Data With Tapered Coiflets Series," *Scandinavian Journal of Statistics*, 23, 313–330.

Antoniadis, A. (1997), "Wavelets in Statistics: A Review" (with discussion), *Italian Journal of Statistics*, 6, 97–144.

Antoniadis, A., Grégoire, G., and McKeague, I. (1994), "Wavelet Methods for Curve Estimation," *Journal of the American Statistical Association*, 89, 1340–1353.

Antoniadis, A., Grégoire, G., and Vial, P. (1997), "Random Design Wavelet Curve Smoothing," *Statistics and Probability Letters*, 35, pp. 225–232.

Barron, A., Birgé, L., and Massart, P. (1999), "Risk Bounds for Model Selection Via Penalization," *Probability Theory and Related Fields*, 113, 301–413.

Belge, M., Kilmer, M. E., and Miller, E. L. (2000), "Wavelet Domain Image Restoration With Adaptive Edge-Preserving Regularization," *IEEE Transactions on Image Processing*, 9, 597–608.

Bernard, C., Mallat, S., and Slotine, J. J. (1999), "Wavelet Interpolation Networks," Preprint, Centre de Mathématiques Appliquées, Ecole Polytechnique, France.

Bickel, P. J. (1975), "One-Step Huber Estimates in Linear Models," *Journal of the American Statistical Association*, 70, 428–433.

Blake, A. (1989), "Comparison of the Efficiency of Deterministic and Stochastic Algorithms for Visual Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 2–12.

Blake, A., and Zisserman, A. (1987), *Visual Reconstruction*, Cambridge, MA: MIT Press.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garotte," *Technometrics*, 37, 373–384.

Bouman, C., and Sauer, K. (1993), "A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation," *IEEE Transactions on Image Processing*, 2, 3, 296–310.

Cai, T. T., and Brown, L. D. (1998), "Wavelet Shrinkage for Nonequispaced Samples," *Annals of Statistics*, 26, 1783–1799.

Chen, S. C., Donoho, D. L., and Sanders, M. A. (1998), "Atomic Decomposition by Basis Pursuit," *SIAM Journal of Scientific Computing*, 20, 1, 33–61.

Dechevsky, L. T., and Penev, S. I. (1999), "Weak Penalized Least Squares Wavelet Regression Estimation," Technical Report S99-1, Department of Statistics, School of Mathematics, University of New South Wales, Australia.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Donoho, D. L., and Johnstone, I. M. (1998), "Minimax Estimation Via Wavelet Shrinkage," *Annals of Statistics*, 26, 879–921.

Donoho, D. L., Johnstone, I. M., Hock, J. C., and Stern, A. S. (1992), "Maximum Entropy and the Nearly Black Object" (with discussions), *Journal of the Royal Statistical Society,* Ser. B, 54, 41–81.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia?" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 57, 301–369.

Donoho, D. L., Vetterli, M., DeVore, R. A., and Daubechies, I. (1998), "Data Compression and Harmonic Analysis," Technical report, Department of Statistics, Stanford University.

Fan, J. (1997), "Comment on 'Wavelets in Statistics: A Review' by A. Antoniadis," *Italian Journal of Statistics*, 6, 97–144.

Fan, J., and Chen, J. (1999), "One-Step Local Quasi-Likelihood Estimation," *Journal of the Royal Statistical Society,* Ser. B, 61, 927–943.

Fan, J., and Li, R. (1999), "Variable Selection via Penalized Likelihood," Technical report, Department of Statistics, UCLA.

Gao, H. Y., and Bruce, A. G. (1997), "WaveShrink with Firm Shrinkage," *Statistica Sinica*, 7, 855–874.

Hall, P., and Patil, P. (1995). "Formulae for Mean Integrated Squared Error of Nonlinear Wavelet-Based Density Estimators," *Annals of Statistics*, 23, 905–928.

Hall, P., and Turlach, B. A. (1997), "Interpolation Methods for Nonlinear Wavelet Regression With Irregularly Spaced Design," *Annals of Statistics*, 25, 1912–1925.

Hansen, P. C., and O'Leary, D. P. (1993), "The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems," *SIAM Journal of Scientific Computing*, 14, 1487–1503.

Johnstone, I. M., and Silverman, B. W. (1997), "Wavelet Threshold Estimators for Data With Correlated Noise," *Journal of the Royal Statistical Society,* Ser. B, 59, 319–351.

Kovac, A., and Silverman, B. W. (2000), "Extending the Scope of Wavelet Regression Methods by Coefficient-Dependent Thresholding," *Journal of the American Statistical Association*, 95, 172–183.

McCoy, E. J. (1999). "Wavelet Regression: A Penalty Function Approach," in *Proceeding of the 52nd session of the International Statistical Institute*.

Moulin, P., and Liu, J. (1999), "Analysis of Multiresolution Denoising Schemes Using Generalized Gaussian and Complexity Priors," *IEEE Transactions on Information Theory*, 45, 9, 909–919.

Neumann, M. H., and Spokoiny, V. G. (1995), "On the Efficiency of Wavelet Estimators Under Arbitrary Error Distributions," *Mathematical Methods of Statistics*, 4, 2, 137–166.

Nikolova, M. (1999), "Markovian Reconstruction Using a GNC Approach," *IEEE Transactions on Image Processing*, 8, 1204–1220.

Nikolova, M. (1999), "Local Strong Homogeneity of a Regularized Estimator," *SIAM Journal on Applied Mathematics*, 61, 633–658.

Nikolova, M., Idier, J., and Mohammad-Djafari, A. (in press), "Inversion of Large-Support Ill-Posed Linear Operators Using a Piecewise Gaussian MRF," submitted to *IEEE Transactions on Image Processing*.

Ogden, T. (1997), *Essential Wavelets for Statistical Applications and Data Analysis*, Boston: Birkhauser.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: Wiley.

Robinson, P. M. (1988), "The Stochastic Difference Between Econometric and Statistics," *Econometrica*, 56, 531–547.

Sardy, S., Percival, D. B., Bruce A. G., Gao, H.-Y., and Stuelzle, W. (1999), "Wavelet Shrinkage for Unequally Spaced Data," *Statistics and Computing*, 9, 65–75.

Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 47, 1–52.

Solo, V. (1998), "Wavelet Signal Estimation in Coloured Noise With Extension to Transfer Function Estimation," in *Proceedings of the 37th IEEE Conference on Decision and Control*, 4, 3940–3941.

Tibshirani, R. (1995), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Ser. B, 57, 267–288.

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*. New York: Wiley.

Vogel, C. R. (1996), "Non-Convergence of the L-Curve Regularization Parameter Selection Method," *Inverse Problems*, 12, 535–547.

Wang, Y. (1996), "Function Estimation via Wavelet Shrinkage for Long-Memory Data," *Annals of Statistics*, 24, 466–484.

Discussion

Brani VIDAKOVIC

Anestis Antoniadis and Jianqing Fan deserve congratulations for a wonderful and illuminating paper.

Links among wavelet-based penalized function estimation, model selection, and now actively explored wavelet-shrinkage estimation are intriguing and attracted the attention of many researchers. Antoniadis and Fan provide numerous references. The nonlinear estimators resulting as optimal in the process of regularization, for some specific penalty functions, turn out to be the familiar hard- or soft-thresholding rules, or some of their sensible modifications. Simply speaking, the penalty function determines the estimation rule, and in many cases, a practicable and ad hoc shrinkage rule can be linked to a regularization process under a reasonable penalty function. The authors explore the nature of penalty functions resulting in thresholding-type rules. They also show that for a large class of penalty functions, corresponding shrinkage estimators are adaptively minimax and have other good sampling properties.

My discussion is directed toward the link of the regularization problem and Bayesian modeling and inference in the wavelet domain, which is only hinted at by Antoniadis and Fan.

1. BAYES WAVELET MODELING

Any decision made about the model, including an estimate, a test, or a prediction, should take into account available prior information and possible costs of inaccurate actions. Bayesian decision theory is concerned with devising actions that minimize the average cost to the decision maker using a coherently obtained posterior that incorporates both observations and the a priori information. Some of the benefits of Bayesian modeling in the wavelet domain are now well understood, and a variety of methods, based on the Bayes estimation of the signal part in an observed wavelet coefficient, can incorporate particular information about unknown signals (smoothness, periodicity, and self-similarity, for instance).

It is now a standard practice in wavelet shrinkage to specify a location model on wavelet coefficients, elicit a prior on their locations (the signal part in wavelet coefficients), exhibit the Bayes estimator for the locations, and, if resulting Bayes estimators are shrinkage, apply the inverse wavelet transformation to such estimators.

In considering this model-induced shrinkage, the main concern is, of course, performance of induced shrinkage rules, measured by the realized mean square error, whereas the match between models and data in the wavelet domain is paid no special attention. It is certainly desirable for selected models to describe our empirical observations well for the majority of signals and images. At the same time, the calculation of shrinkage rules should remain inexpensive. Our experience is that the realistic but complicated models, for which the rules are obtained by expensive simulations, are seldom accepted bypractitioners, despite their reportedly good performance. The two desirable goals of simplicity and reality can be achieved simultaneously, and Bayesian interpretation of regularization provides a way, which is the point of my discussion.

The authors consider a paradigmatic normal location model with known variance, in which a typical wavelet coefficient z is modeled as $\phi(z-\theta)$, where θ is the signal part. The choice of prior θ often is based on inspecting the empirical realizations of coefficients of the pure signals (noiseless data). Lack of intuition on links between function features and nature of wavelet coefficients and a great variety of possible signals call for use of automatic priors.

Berger and Müller indicated in late 1993 (personal communication) that priors from the ϵ-contamination family are suitable for the signal part in the wavelet domain because the resulting Bayes rules are close in shape to standard thresholding rules. The point mass at zero, $\delta(0)$, in

$$\pi(\theta) = \epsilon\delta(0) + (1-\epsilon)\xi(\theta) \tag{1}$$

induces nonlinear shrinkage and models sparsity, whereas $\xi(\theta)$ is a spread distribution that models wavelet coefficients with large energies (squared magnitudes). This spread distribution can be improper. Besides, adequate changes in ϵ provide a possibility of levelwise adaptive rules.

Various priors on the signal part were proposed by many authors. Papers by Abramovich, Sapatinas, and Silverman (1998), Clyde, Parmigiani, and Vidakovic (1998), Chipman, Kolaczyk, and McCulloch (1997), Vidakovic (1998a), and many others propose priors with different degrees of intricacy but that are in spirit similar to the Berger–Müller proposal (1). An overview can be found in Vidakovic (1998b).

Interesting automatic (objective) priors were proposed as well. Berger and Pericchi (1996) demonstrated that in the context of Bayesian model selection, in testing that the signal part is 0, Jeffreys's prior is

$$\pi(\theta,\sigma) = \frac{1}{\sigma}\left[\epsilon\delta(0) + (1-\epsilon)\frac{1}{\pi\sigma(1+\theta^2/\sigma^2)}\right],$$

and the intrinsic prior is

$$\pi(\theta,\sigma) = \frac{1}{\sigma}\left[\epsilon\delta(0) + (1-\epsilon)\frac{1-\exp(-\theta^2/\sigma^2)}{2\sqrt{\pi}[\theta^2/\sigma]}\right].$$

The shrinkage rules, involving Bayes factors, in both cases can have simple approximations.

Brani Vidakovic is Associate Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205. This research was supported in part by NSF grant DMS-0072585 at Georgia Tech.

Journal of the American Statistical Association
September 2001, Vol. 96, No. 455, Theory and Methods

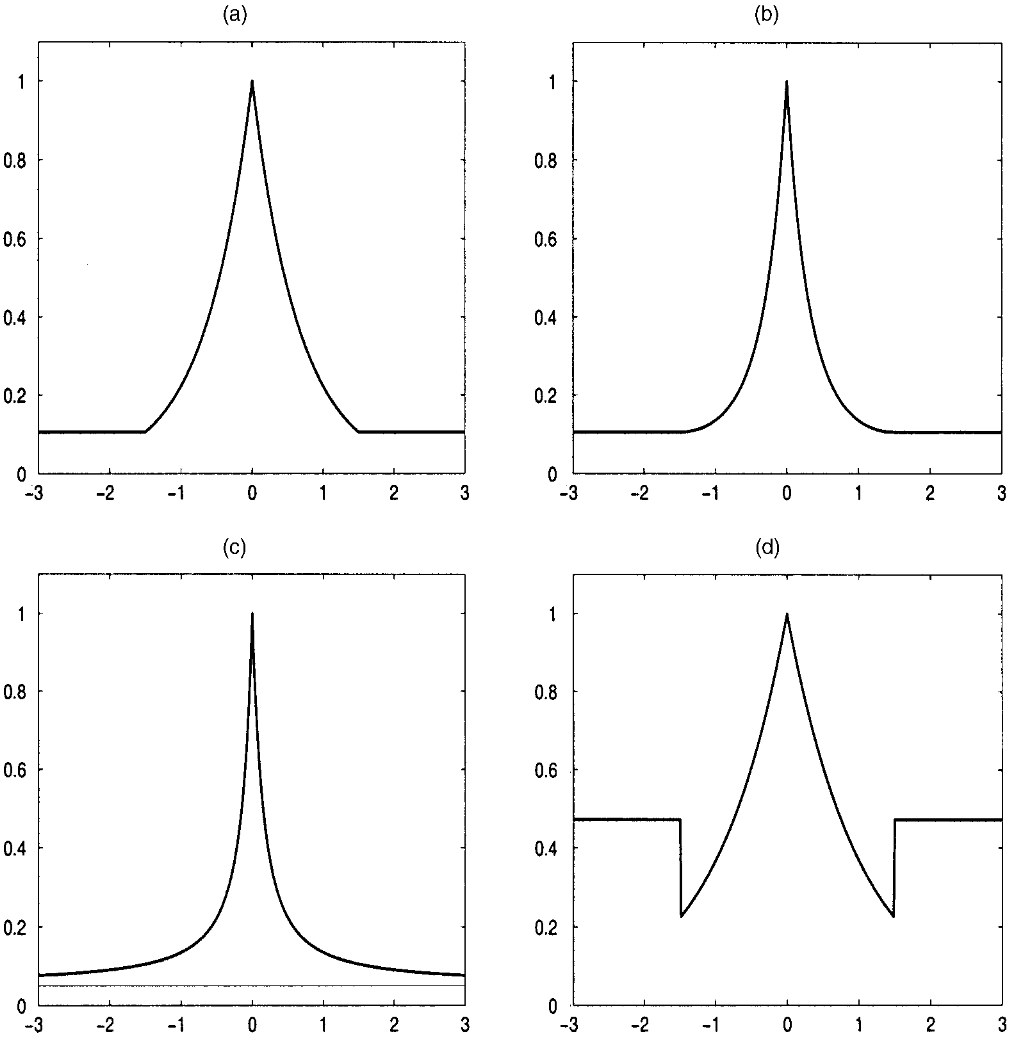

Figure 1. MAP Priors to Penalties (AF 2.6), (AF 2.8), (AF 2.11), and Penalty From Fan (1997). In all cases, $\lambda = 1.5$, and for the prior in panel (d), $b = 2$.

2. MAP PRINCIPLE

All information in Bayesian inference is contained in the posterior, and posterior location measures (mean, median, mode) are standard Bayes rules for the location parameters. Typically, it is more difficult to exhibit the mean or median of a posterior than the value at which the posterior is maximized, a posterior mode. This is because for the mean or median, an exact expression for the posterior is needed. MAP rules that maximize the posterior also maximize, at the same time, the product of the likelihood and prior, and they are typically shrinkage rules.

Given an observation z, the posterior distribution of θ is proportional to

$$\pi(\theta|z) \propto \phi(z-\theta)\cdot\pi(\theta). \tag{2}$$

Let $s(\theta) = -\log\pi(\theta)$ be the score of the prior. Notice that the posterior is maximized at the same argument at which

$$s(\theta) - \log\phi(z-\theta) = \frac{1}{2\sigma^2}(z-\theta)^2 + s(\theta) \tag{3}$$

is minimized. If $s(\theta)$ is strictly convex and differentiable, the minimizer of (3) is a solution $\hat{\theta}$ of

$$s'(\theta) + \frac{1}{\sigma^2}(\theta - z) = 0.$$

One finds

$$\hat{\theta} = h^{-1}(z), \qquad h(u) = u + \sigma^2 s'(u). \tag{4}$$

Generally, the inversion in (4) may not be analytically feasible, but a solution may be achieved via an approximate sequence of invertible functions. The authors provide examples of prior distributions on θ for which an analytical maximization is possible. Some additional solvable cases can be found in Fan (1997), Hyvärinen (1998), and Wang (1999).

For example, if $\pi(\theta) = \frac{1}{\sqrt{2}}e^{-\sqrt{2}|\theta|}$, then $s'(\theta) = \sqrt{2}\ \text{sign}(\theta)$, and $\hat{\theta}(d) = \ \text{sign}(d)\max(0, |z| - \sqrt{2}\sigma^2)$.

For

$$\pi(\theta) \propto e^{-a\theta^2/2 - b|\theta|}, \qquad a, b > 0,$$

i.e., if $s'(\theta) = a\ \theta + b\ \ \mathrm{sign}(\theta)$, the MAP rule is

$$\hat{\theta}(d) = \frac{1}{1+\sigma^2 a}\ \mathrm{sign}(d)\ \max(0, |d| - b\,\sigma^2).$$

If π is a super Gaussian probability density,

$$\pi(\theta) \propto \left[\sqrt{\alpha(\alpha+1)} + \left|\frac{\theta}{b}\right|\right]^{\alpha+3},$$

the corresponding MAP rule is

$$\hat{\theta}(d) = \mathrm{sign}(d) \times \max\left(0, \frac{|d|-ab}{2} + \frac{1}{2}\sqrt{(|d|+ab)^2 - 4\sigma^2(\alpha+3)}\right), \tag{5}$$

where $a = \sqrt{\alpha(\alpha+1)/2}$, and $\hat{\theta}(d)$ is set to 0 if the square root in (5) is imaginary.

Leporini and Pesquet (1998) explore cases for which the prior is an exponential power distribution $[\mathcal{EPD}(\alpha, \beta)]$. If the noise also has an $\mathcal{EPD}(a, b)$ distribution with $0 < \beta < b \leq 1$, this MAP solution is a hard-thresholding rule. If $0 < \beta \leq 1 < b$, then the resulting MAP rule is

$$\hat{\theta}(d) = d - \left(\frac{\beta a^b}{b\alpha^\beta}\right)^{1/(b-1)} |d|^{(\beta-1)/(b-1)} + o(|d|^{(\beta-1)/(b-1)}).$$

The same authors also consider the Cauchy noise and explore properties of the resulting rules. When the priors are hierarchical (mixtures), Leporini, Pesquet, and Krim (1999) demonstrated that the MAP solution can be degenerated, and they suggested the maximum generalized marginal likelihood method. Some related derivations can be found in Chambolle et al. (1998) and Pesquet et al. (1996).

3. PENALTIES IN THE MAP CONTEXT

What are the common properties of priors linked to some penalty functions considered by Antoniadis and Fan? It is interesting that the priors look like histograms of typical wavelet coefficients, corresponding to noiseless signals and images. Such empirical densities exhibit sharp, double exponential–like peaks around 0 and fairly flat tails.

On the other hand, shapes of the priors are in the spirit of the standard modeling family (1), where the point mass at 0 is softened by a peak at 0. The tail parts are in some of the examples improper (flat).

As an illustration, we consider the priors corresponding to (AF 2.6), (AF 2.8), and (AF 2.11) and the penalty suggested in Fan (1997), $p_\lambda(\theta) = |\theta|\mathbf{1}(|\theta| < \lambda) - \lambda/2\mathbf{1}(|\theta| \geq \lambda)$. They are

$$\begin{aligned}
\pi(\theta) &\propto e^{-\lambda\cdot\min(|\theta|,\lambda)},\\
\pi(\theta) &\propto e^{-\lambda^2+(|\theta|-\lambda)^2\mathbf{1}(|\theta|<\lambda)},\\
\pi(\theta) &\propto e^{-\lambda b|\theta|(1+b|\theta|)^{-1}},\\
\pi(\theta) &\propto e^{-|\theta|\mathbf{1}(|\theta|<\lambda)-\lambda/2\mathbf{1}(|\theta|\geq\lambda)},
\end{aligned}$$

and they are depicted in Figure 1.

In conclusion, I point to some benefits of the MAP point of view on regularized wavelet estimation:

1. honest statistical models whose marginals well match the observations
2. possible incorporation of prior information
3. use of Bayesian machinery to exhibit solutions in cases when simple, closed-form solutions are impossible.

Finally, I thank the editor for the kind invitation to discuss this important paper.

REFERENCES

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998), "Wavelet Thresholding via a Bayesian Approach," *Journal of the Royal Statistical Society*, Ser. B, 60, 725–749.

Berger, J, and Pericchi, L. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Chambolle, A., DeVore, R. A., Lee N.-Y., and Lucier, B. J. (1998), "Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal Through Wavelet Shrinkage," *IEEE Transactions on Image Processing*, 7, 319–355.

Chipman, H., Kolaczyk, E., and McCulloch, R. (1997), "Adaptive Bayesian Wavelet Shrinkage," *Journal of the American Statistical Association*, 92, 1413–1421.

Clyde, M. A., Parmigiani, G., and Vidakovic, B. (1998), "Multiple Shrinkage and Subset Selection in Wavelets," *Biometrika,* 85, 391–402.

Fan, J. (1997), Comment on "Wavelets in Statistics: A Review," *Italian Journal of Statistics*, 6, 97–144.

Hyvärinen, A. (1998), "Sparse Code Shrinkage: Denoising of Nongaussian Data by Maximum Likelihood Estimation," Technical Report A51, Helsinki University of Technology, Finland.

Leporini, D., and Pesquet, J.-C. (1998), "Wavelet Thresholding for Some Classes of Non-Gaussian Noise," Technical report, CNRS/Université Paris-Sud, France.

Leporini, D., Pesquet, J.-C., and Krim, H. (1999), "Best Basis Representations with Prior Statistical Models," in *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, *Lecture Notes in Statistics*, 141, New York: Springer Verlag, pp. 109–113.

Pesquet, J.-C., Krim, H., Leporini, D., and Hamman, E. (1996), "Bayesian Approach to Best Basis Selection," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5, pp. 2634–2637.

Vidakovic, B. (1998a), "Nonlinear Wavelet Shrinkage With Bayes Rules and Bayes Factors," *Journal of the American Statistical Association*, 93, 173–179.

Vidakovic, B. (1998b), "Wavelet-Based Nonparametric Bayes Methods," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, *Lecture Notes in Statistics*, 133, New York: Springer-Verlag, pp. 133–155,

Wang, Y. (1999), "An Overview of Wavelet Regularization," in *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, *Lecture Notes in Statistics*, 141, New York: Springer Verlag, pp. 109–113.

Discussion

Pierre MOULIN

I congratulate the authors on a well-written and insightful article on nonparametric wavelet regression. Of the many interesting aspects of the article, the characterization and risk analysis of penalized least squares estimators particularly caught my attention. In this discussion, I comment on closely related work in the statistical signal and image processing literature.

1. CHARACTERIZATION OF PENALIZED LEAST SQUARES ESTIMATORS

Several of the problems considered take the form (2.4)

$$\min_{\theta}\left\{\frac{1}{2}\sum_{i=1}^{n}(z_i-\theta_i)^2+\lambda\sum_{i\geq i_0}p(|\theta_i|)\right\}, \qquad (1)$$

where $\{\theta_i\}$ are the unknown wavelet coefficients and $\{z_i\}$ are the transformed data. This yields a set of independent one-dimensional optimization problems whose solution is of the form $\hat{\theta}_i = F_\lambda(z_i)$ for $i \geq i_0$. Problems of the form (2.4) have also been encountered in the statistical signal processing literature. In particular, the papers by Nikolova (in press) and Moulin and Liu (1999) contain results that are analogous to Theorem 1 and eatablish the existence of thresholding effects in the shrinkage function F_λ, when the penalty $p(\cdot)$ is nondifferentiable at the origin. A very interesting by product of Theorem 1 is that the soft- and hard-thresholding estimators are extreme cases of a broad class of penalized least squares estimators.

2. RELATED WORK IN STATISTICAL SIGNAL AND IMAGE PROCESSING LITERATURE

Nonsmooth (including nonconvex) penalties such as $p(|\theta|) = |\theta|^r, r \in (0, 1]$, have been popular in the image processing literature, because of the edge-preserving properties of the resulting estimators. See German and Reynolds (1992),Bouman and Sauer (1993), Nikolova (1996), Simoncelli (1996), and Charbonnier et al. (1997). An analysis of the resulting shrinkage functions was presented by Moulin and Liu (1999). The complexity penalty $p(|\theta|) = \mathcal{I}_{|\theta|\neq 0}$ (which the authors call entropy penalty, for reasons unknown to me) has close connections to the model selection problem. Recently, Nowak and de Figueiredo (in press) studied the penalty $p(|\theta|) = \ln(c + |\theta|)$, which is derived from Jeffreys's prior. From a practical point of view, numerous image denoising experiments showed that the choice of the shape of the shrinkage function has a relatively mild effect on the quality of the estimates, if the asymptotic bias $|F_\lambda(Z_i) - Z_i|$ is 0 or at least small and if the threshold discontinuity is relatively benign; for instance, the squared L^2 risks obtained by using the shrinkage functions derived from the Generalized Gaussian (GG) family $p(|\theta|) = |\theta|^r$, with $r \in (\frac{1}{2}, 1]$, are typically within 15% of each other. Another useful property of nonsmooth penalties is that they yield sparse estimates for signals and images, because of the thresholding effects discussed earlier.

Although the use of fixed universal shrinkage functions is elegant and is asymptotically nearly opltimal for a variety of function classes, substantial performance improvements were obtained by applying level-dependent shrinkage functions to the empirical wavelet coefficients. For instance, Donoho and Johnstone's SureShrink algorithm performs much better than VisuShrink. A related idea appears in a paper by Chambolle et al. (1998), which explores the use of Besov penalties on wavelet coefficients. The criterion they minimize is

$$\min_{\theta}\left\{\frac{1}{2}\sum_{jk}(Z_{jk}-\theta_{jk})^2+\lambda\sum_{jk}2^{jrs}|\theta_{jk}|^r\right\}, \qquad (2)$$

where $j \in \mathbb{Z}$ is the scale and $k \in \mathbb{Z}$ is the location parameter. This criterion penalizes lack of smoothness of the function, measured by using L^r norm $(r \geq 1)$ and smoothness index s. This results in scale-dependent shrinkage functions. Nevertheless, Chambolle et al. observed that the resulting estimator. Increased flexibility in the estimator can be obtained by minimizing the criterion

$$\min_{\theta}\left\{\frac{1}{2}\sum_{jk}(Z_{jk}-\theta_{jk})^2+\sum_{jk}\lambda_j|\theta_{jk}|^r\right\}, \qquad (3)$$

where the parameters $\{\lambda_j\}$ are selected in a data-driven way, see Moulin and Liu (1999) and Belge, Kilmer, and Miller (2000).

Lastly, Simoncelli (1996) studied the minimum mean squared error estimator for a problem in which $\{Z_{jk}\}$ are the sum of $\{\theta_{jk}\}$ and white Gaussian noise and $\{\theta_{jk}\}$ follow a GG distribution. This yields continuous shrinkage functions whose aspect is nevertheless fairly similar to the shrinkage functions derived from a MAP criterion.

2.1 Wavelet-Based Interpolation

A variety of problems in computer vision, such as recovery of motion fields and disparity fields, surface interpolation, and edge detection (Blake and Zisserman 1997, Bertero, Poggio, and Torre 1988), involve interpolation, from data acquired on an irregular grid. Other problems, such as resolution enhancement, involve interpolation from data acquired on an irregular

Pierre Moulin is Associate Professor, Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 (E-mail: *moulin@ifp.uiuc.edu*).

grid. Other problems, such as resolution. In all cases, accurate recovery of jumps, edges, and other singularities presents major problems. Wavelet methods have been used to construct interpolation algorithms based on a characterization of signals and images in the wavelet domain; see, e.g., Mallat and Hwang (1992), Ford and Etter (1998), Carey, Chuang, and Hemami (1999), and Vazquez, Konrad, and Dubois (2000). So far, the emphasis has been placed on the ability to recover singularities and on the visual aspect of the interpolated functions. The analytical framework developed in Section 4 of the article provides a new perpective on these applications. The recent paper by Choi and Baraniuk (1999), which formulates the interpolation problem as a penalized least squares problem by using a Besov norm as a roughness penalty, is closely related to the authors' approach. Choi and Baraniuk's optimization criterion is convex, and their results seem excellent.

2.2 Nonadditive Penalties

Although the authors developed a powerful framework for generalizing Donoho and Johnstone's thresholding methods and analyzing penalized least squares estimators, it would be useful to further generalize that analysis to nonadditive penalties on the wavelet coefficients. The advantages of such penalties were demonstrated in several papers in the image processing literature. For instance some of the best denoising results to date were obtained by exploiting the dependencies between wavelet coefficients. This can be done conveniently by using statistical modeling of these coefficients. One constructs a tractable prior (usually containing a few hyperparameters that are empirically estimated from the data) and then evaluates or approximates the MAP estimator. Excellent results have been obtained by Crouse, Nowak, and Baraniuk (1998), using hidden Markov models, and by Simoncelli (1997, 1999) and Mıhçak et al. (1999, 1999b), using models that capture the spatial clustering of significant wavelet coefficients.

REFERENCES

Bertero, M., Poggio, T. A., and Torre, V. (1988), "Ill-Posed Problems in Early Vision," *Proceedings of the IEEE*, 76, pp. 869–889.

Carey, G. K., Chuang D. B., and Hemami, S. S. (1997), "Regularity-Preserving Image Interpolation," in *Proceedings of the IEEE International Conference on Image Processing*,

Chambolle, A., DeVore, R. A., Lee, N., and Lucier, B. (1998), "Nonlinear Wavelet Image Provessing: Variational Problems, Compression, and Noice Removal Through Wavelet Shrinkage," *IEEE Transactions in Image Processing*, 7, pp. 319–335.

Charbonnier, P., Blanc-Féraud, L., Aubert, G., and Barlaud, M., (1997), "Deterministic Edge-Preserving Regularization in Computed Imaging," *IEEE Transactions in Image Processing*, 6, pp. 298–311.

Choi, H., and Baraniuk, R. G. (1999), "Interpolation and Denoising of Nonuniformly Sampled Data Using Wavelet-Domain Processing," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1645–1648,

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998), "Wavelet-Based Statistical Signal Processing Using Hidden Markov Models," *IEEE Transactions in Signal Processing*, 46, pp. 886–902.

Ford, C., and Etter, D. M. (1998), "Wavelet Basis Reconstruction of Nonuniformly Sampled Data," *IEEE Transactions on Circuits and Systems II. Analog and Digital Signal Processing*, 45, pp. 1165–1168.

Geman, S., and Reynolds, G. (1992), "Constrained Restoration and Recovery of Discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, pp. 367–383.

Li, S. Z., (1995), "On Discontinuity-Adaptive Smoothness Priors in Computer Vision," *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 17, pp. 576–586.

Mallat, S. G., and Hwang, W.-L. (1992), "Singularity Detection and Processing with Wavelets," *IEEE Transactions in Information Theory*, 38, pp. 617–643.

Mıhçak, M. K., Kozintsev, I. and Ramchandran, K. (1999), "Spatially Adaptive Statical Modeling of Wavelet Image Coefficients, and Its Application to Denoising," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3253–3256.

Mıhçak, M. K., Kozintsev, I. Ramchandran, K., and Moulin, P. (1999b), "Low-Complexity Image Denoising Based on Statistical Modeling of Wavelet Coefficients," *IEEE Signal Processing Letters*, 6, pp. 300–303.

Nikolova, M. (1996), "Regularisation Functions and Estimators," in *Proceedings of the IEEE International Conference on Image Processing*, (vol. 2), pp. 457–460.

Nowak, R. D., and de Figueiredo, M. A. T. (in press) "Wavelet Based Image Estimation: An Empirical Bayes Approach Using Jeffreys' Noninformative Prior," *IEEE Transactions in Image Processing*.

Simoncelli, E. P. (1996), "Noise Removal via Bayesian Wavelet Coring," in *Proceedings of the IEEE International Conference on Image Processing* (vol. 1) pp. 379–382.

Simoncelli, E. P. (1997), "Statistical Models for Images: Compression, Restortion and Synthesis," in *Proceedings of the 31st ASILOMAR Conference*, pp. 673–678.

Simoncelli, E. P. (1999), "Modeling the Joint Statistics of Images in the Wavelet Domain," in *SPIE Conference 3813 on Wavelet Applications in Signal and Image Processing VII*.

Vazquez, C., Konrad, J., and Dubois, E. (2000), "Wavelet-Based Reconstruction of Irregularly Sampled Images: Application to Stereo Imaging," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 319–322.

Discussion

T. Tony Cai

Professors Antoniadis and Fan are to be congratulated for their valuable work on the penalized least squares method for wavelet regression. In this very interesting article the authors present a general characterization of the penalized least squares estimators for a class of additive smooth penalty functions. A main contribution of this article is the unified and systematic derivation of the oracle inequalities and minimax properties for a whole class of wavelet estimators. An important advantage of the approach taken by the authors is its generality. As demonstrated nicely by the authors, the penalized

T. Tony Cai is Assistant Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104 supported in part by NSF grant DMS-0072578.

© 2001 American Statistical Association
Journal of the American Statistical Association
September 2001, Vol. 96, No. 455, Theory and Methods

least squares method can be used successfully to treat both equispaced and nonequispaced samples. It can also be readily applied to other nonparametric function estimation problems using wavelets.

My remarks are primarily confined to the discussion of extensions of the penalized least squares method presented.

As in the article, I use in the following discussion both a single index i and the more conventional double indices (j, k) for wavelet coefficients.

1. PENALIZED LEAST SQUARES

The general penalized least squares can be written as

$$\ell(\boldsymbol{\theta}) = \|\mathbf{Y}_n - \mathbf{A}\boldsymbol{\theta}\|^2 + \lambda p(\boldsymbol{\theta}). \qquad (1)$$

Assuming for the moment that the sample is equispaced and the sample size is a power of 2, then the matrix A is the inverse discrete wavelet transform W^{-1}, and (1) can be equivalently written as

$$\ell(\boldsymbol{\theta}) = \|\mathbf{Z}_n - \boldsymbol{\theta}\|^2 + \lambda p(\boldsymbol{\theta}), \qquad (2)$$

where $\mathbf{Z}_n = W\mathbf{Y}_n$ is the empirical wavelet coefficients. The estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ minimizes $\ell(\boldsymbol{\theta})$ in (2). The performance of the penalized least squares estimator depends on the penalty $p(\boldsymbol{\theta})$ and the regularization parameter λ. Let us first consider the effect of the penalty function $p(\boldsymbol{\theta})$.

Among possible choices of $p(\boldsymbol{\theta})$, the additive penalty $p(\boldsymbol{\theta}) = \sum_i p(|\theta_i|)$ is the most intuitive choice. When an additive penalty is used, the minimization problem becomes separable. Minimizing (2) is equivalent to minimizing $\ell(\theta_i) = \|z_i - \theta_i\|^2 + \lambda p(|\theta_i|)$ for each coordinate i. The resulting penalized least squares estimator in this case is separable. That is, the estimate of any coordinate θ_i depends solely on the empirical wavelet coefficient z_i, not on any other coefficients z_j. This is intuitively appealing. However, separable estimators have their drawbacks. The difficulty arises through the need to guard against false positives about the presence of true significant coordinates (corresponding to irregularities of the regression function f). Consequently it is unvoidable to introduce a logarithmic factor into both the thresholding constant λ and the convergence rate. As a result, the estimator is often oversmoothed. The problem is unavoidable for separable estimators, because decisions about individual terms are based on a relatively low level of information. See Cai (1999a, 2000b) and Hall, Kerkyacharian, and Picard (1999).

Therefore, there are true benefits to considering a more general penalty function $p(\boldsymbol{\theta})$ in (2). One possibility is to use a blockwise additive penalty. First, we divide the coefficients into nonoverlapping blocks, and then we define a blockwise additive penalty

$$p(\boldsymbol{\theta}) = \sum_b p(\theta_{(b)}), \qquad (3)$$

where $\theta_{(b)}$ denotes the coefficient vector in the bth block. For example, one can divide the coefficients into equal-length blocks of size L. Denote the indices in the bth block by $(b) = \{(b-1)L+1, \ldots, bL\}$ and denote by $\theta_{(b)} = (\theta_{(b-1)L+1}, \ldots, \theta_{bL})$ and $z_{(b)} = (z_{(b-1)L+1}, \ldots, z_{bL})$ the true and empirical wavelet coefficient vectors in the bth block, respectively. Let the penalty be

$$p(\theta_{(b)}) = \lambda I\{\theta_{(b)} \neq 0\}. \qquad (4)$$

Then the solution to the penalized least squares problem (2) is a block thresholding estimator:

$$\hat{\theta}_i = z_i I\{\|z_{(b)}\|_{\ell^2} > \lambda\} \quad \text{for } i \in (b).$$

Block thresholding estimators increase the estimation accuracy by pooling information about neighboring coefficients to make a simultaneous thresholding decision. The estimators were shown to enjoy some desirable properties for appropriately chosen block size L and thresholding constant λ. For example, the estimators can attain the exact minimax convergence rate without a logarithmic penalty and enjoy good numerical performance (Hall et al. 1999; Cai 1999b).

The penalty in (4) is only a particular choice and can be replaced by other penalty functions. The block size can be allowed to vary from level to level. It will be interesting to study the properties of estimators derived from this class of blockwise additive penalty functions using the same unified approach as in this article.

1.1 From Equispaced to Nonequispaced

Wavelet methods for nonequispaced data in the literature are mostly based on interpolation and approximation, either in the original function domain or in the wavelet domain. The regularized Sobolev interpolators used in this article can be regarded as approximation in the wavelet coefficient space. As indicated by Figure 1 in the article, the regularization parameter s appears to be an important parameter. It will be interesting to find an empirical rule for choosing the value of s.

The approach of the regularized Sobolev interpolators has the advantage of naturally extending the results for equispaced data to nonequispaced data. For example, one can easily extend the penalized least squares equation (4.3) to other penalty functions, such as a blockwise additive penalty. This will result in a block thresholding estimator based on the synthetic data.

1.2 Choice of the Regularization Parameter λ

Besides the penalty function $p(\boldsymbol{\theta})$, the regularization parameter λ plays a crucial role in the performance of the resulting estimators. In fact, some of the results in this article appear to indicate that the performance of the penalized least squares estimators do not differ significantly for all the smooth penalty functions under consideration. This means that it might be even more important to choose λ in an optimal way than to choose the penalty function. Penalized least squares were considered by Chambolle et al. (1998) in the context of image processing. The regularization parameter λ in Chambolle et al. (1998) was chosen empirically by minimizing an upper bound of the risk. A similar approach can be used here for the general class of smooth penalty functions. The advantage of this approach is that it is computationally simple and fast. Another possibility is to use cross validation, which is computationally more intense.

2. OTHER FUNCTION ESTIMATION PROBLEMS

The penalized least squares approach can be applied to other statistical contexts. In particular, it can be used in statistical inverse problems where one observes a function of the interest indirectly. For example, let us consider the estimation of the derivative of a regression function. Suppose we observe

$$y_i = f(i)n) + \epsilon_i, \qquad i = 1, \ldots, n(= 2^J), \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

and we wish to estimate the derivative f' under the mean squared error $E\|\hat{f}' - f'\|_2^2$. By using the Vaguelet wavelet decomposition approach (Abramovich and Silverman 1998; Cai to appear), one can first construct an appropriate estimator $\hat{f}$ of f and then use its derivative $\hat{f}'$ as an estimator of f'. Suppose f has the standard orthonormal wavelet expansion

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} \theta_{j,k} \psi_{j,k}(x),$$

where ϕ and ψ are the father and mother wavelets, respectively. Let

$$\hat{f}(x) = \sum_{k=1}^{2^{j_0}} \hat{\xi}_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(x).$$

Then, under regularity conditions, the risk $E\|\hat{f}' - f'\|_2^2$ of $\hat{f}'$ as an estimator of f' equals

$$\begin{aligned} E\Bigg(& \sum_{k=1}^{2^{j_0}} 2^{j_0} (\hat{\xi}_{j_0,k} - \xi_{j_0,k}) \phi'_{j_0,k}(x) \\ & + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} 2^j (\hat{\theta}_{j,k} - \theta_{j,k}) \psi'_{j,k}(x) + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} 2^j \theta_{j,k} \psi'_{j,k}(x) \Bigg)^2 \\ \asymp\ & 2^{2j_0} E\|\hat{\xi}_{j_0,\cdot} - \xi_{j_0,\cdot}\|_{\ell^2}^2 + \sum_{j=j_0}^{J-1} 2^{2j} E\|\hat{\theta}_{j,\cdot} - \theta_{j,\cdot}\|_{\ell^2}^2 \\ & + \sum_{j=J}^{\infty} 2^{2j} \|\theta_{j,\cdot}\|_{\ell^2}^2, \end{aligned}$$

where $\psi'_{j,k}(x) = 2^{j)2} \psi'(2^j x - k)$, $\theta_{j,\cdot} = (\theta_{j,1}, \ldots, \theta_{j,2^j})$, and $\phi'_{j_0,k}$, $\xi_{j_0,\cdot}$ and $\hat{\theta}_{j,\cdot}$ are defined analogously.

Ignoring higher order approximation errors, this problem is in principle equivalent to the following normal mean problem. Given

$$z_{j,k} = \theta_{j,k} + \epsilon_{j,k}, \quad \epsilon_{j,k} \overset{\text{iid}}{\sim} N(0,\ n^{-1}\sigma^2)$$

with $k = 1, \ldots, 2^j$, $j = j_0, \ldots, J-1$, and $2^J = n$, we wish to estimate $\boldsymbol{\theta}$ under the risk

$$R(\hat{\boldsymbol{\theta}},\ \boldsymbol{\theta}) = \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} 2^{2j} E(\hat{\theta}_{j,k} - \theta_{j,k})^2.$$

The minimax convergence rate over a Besov ball $B^r_{p,q}(C)$ in this case is $n^{2(r-1))(1+2r)}$. The penalized least squares approach can be used to construct near-optimal estimators of $\boldsymbol{\theta}$. For example, by using an additive penalty $p_\lambda(\boldsymbol{\theta})$ with $p_0 = (6 \log n)^{1)2}$ as defined in Theorem 1 [instead of the standard choice $p_0 = (2 \log n)^{1)2}$], the resulting estimator is adaptively within a logarithmic factor of the minimax risk over a wide range of Besov balls $B^r_{p,q}(C)$, assuming that the additive penalty p in (2) satisfies the conditions of Lemma 1. The blockwise additive penalty (4) can also be used here to obtain an exact rate-optimal block thresholding estimator of $\boldsymbol{\theta}$.

This problem can also be reformulated as a more general penalized weighted least squares problem. By renormalizing $z_{j,k}$, $\theta_{j,k}$, and $\epsilon_{j,k}$, one can equivalently consider the following heteroscedastic normal mean problem. Given

$$z^*_{j,k} = \theta^*_{j,k} + \epsilon^*_{j,k}, \quad \epsilon^*_{j,k} \overset{\text{iid}}{\sim} N(0,\ 2^{2j} n^{-1} \sigma^2)$$

with $k = 1, \ldots, 2^j$, $j = j_0, \ldots, J-1$, and $2^J = n$, we wish to estimate $\boldsymbol{\theta}^*$ under the conventional mean squared error $E\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_{\ell^2}^2$. In this case, estimators of $\boldsymbol{\theta}^*$ can be constructed by solving a penalized weighted least squares problem. The estimator $\hat{\boldsymbol{\theta}}^*$ minimizes

$$\ell(\boldsymbol{\theta}^*) = \sum_{j=j_0}^{J-1} \sum_k 2^{-2j} (z^*_{j,k} - \theta^*_{j,k})^2 + \lambda p(\boldsymbol{\theta}^*). \tag{5}$$

The weights here are inverse proportional to the variances of $z^*_{j,k}$. For general inverse problems, the weights 2^{-2j} in (5) are replaced by a^{-2j} for some $a > 0$.

3. CONCLUSION

Penalized least-squares provides a unified framework for many seemingly different wavelet thresholding rules in different nonparametric function estimation contexts. This general framework enables us to systematically study a large class of wavelet estimators simultaneously. Clearly, there is much work ahead of us. This paper provides us a means for generating new ideas that can guide our future research in this area. I thank the authors for their clear and imaginative work.

REFERENCES

Abramovich, F., and Silverman, B. W. (1998), "Wavelet Decomposition Approaches to Statistical Inverse Problems," *Biometrika*, 85, 115–129.

Cai, T. (1999a), "Adaptive Wavelet Estimation: A Block Thresholding and Oracle Inequality Approach," *Annals of Statistics*, 27, 898–924.

Cai, T. (1999b), "On Block Thresholding in Wavelet Regression: Adaptivity and the Choice of Block Size and Threshold Level," Technical report, University of Pennsylvania, Dept. of Statististics.

Cai, T. (in press), "On Adaptive Estimation of a Derivative and Other Related Linear Inverse Problems," submitted to *Journal of Statistical Planning and Inference*.

Cai, T. (2000b), "On Adaptability and Information-Pooling in Nonparametric Function Estimation," Technical report, University of Pennsylvania, Dept. of Statistics.

Chambolle, A., DeVore, R., Lee, N., and Lucier, B. (1998), "Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal Through Wavelet Shrinkage," *IEEE Transactions in Image Processing*, 70, 319–335.

Hall, P., Kerkyacharian, G., and Picard, D. (1999), "On the Minimax Optimality of Block Thresholded Wavelet Estimators," *Statistica Sinica*, 9, 33–50.

Discussion

V. SOLO

The authors are to be congratulated on a stimulating piece of work on wavelet methods for function estimation. There are three main contributions: extension of wavelet asymptotics for nonquadratically penalized least squares estimates to handle new kinds of penalty; a new two-stage estimation algorithm (ROSE) for wavelet fitting with irregularly spaced data; and the bringing of attention to other types of nonquadratic penalties less well known in the statistics literature.

To apply an estimator beyond the situation in which it first appears, it is useful to exhibit it as an instance of an estimation principle. In this vein, the point of departure in this article, taking wavelet function fitting beyond its usual domain of regularly gridded data, is the recognition that soft thresholding is the solution to an l_1 penalized least squares problem (Donoho et al. 1992; DeVore and Lucier 1992). This leads the authors to replace the l_1 penalty with something more flexible, and so to ROSE.

Yet, other kinds of nonquadratic penalties are possible, including robust-statistics-type (Huber) penalties (Stevenson, Schmitz, and Delp 1994; Li 1995). Other nonwavelet formulations are possible, too, such as total variation denoising (Rudin, Osher, and Fatemi 1992), which is a nonwavelet method in which an L_1 penalty is applied to the function derivative. There are also important methods based on smooth approximations to the L_1 penalty (Vogel and Oman 1996).

I have also used the l_1 penalized formulation to extend a wavelet-based estimation to deal with correlated noise (Solo 1998) [in a manner different from Johnstone and Silverman (1997)] and, in joint work with a Ph.D. student, to deal with optical flow estimation (i.e., motion estimation from image sequences) (Ng and Solo 1999).

Turning to Section 4.1 and ROSE, we now point out that some of the discussion and computation can be simplified. The idea of starting computations with an interpolator is a nice one, which I have independently used in the context of inverse problems. However, the Sobolev weighting is a clever and novel variation I have not seen before. With the interpolator $\hat{\theta} = DA^T(ADA^T)^{-1}Y$, we can use the interpolation property $A\hat{\theta}_D = Y$ plus the orthogonality

$$I_{N\times N} = WW^T = A^TA + B^TB$$

to find that in (4.2)

$$\begin{aligned} \hat{\theta}^{*T} &= Y_D^{*T}W^T \\ &= \left[Y^T, \hat{\theta}_D^T B^T\right]\begin{pmatrix} A \\ B \end{pmatrix} \\ &= Y^TA + \hat{\theta}_D^T B^T B \\ &= \hat{\theta}_D^T(A^TA + B^TB) \\ &= \hat{\theta}_D^T. \end{aligned}$$

So $\hat{\theta}_D$ already consists of the empirical synthetic wavelet coefficients. Thus, once the Sobolev interpolation is carried out, no extended data $\widehat{Y}_0^*$ need to be generated.

Similarly, we can avoid computing the extended data $\widehat{Y}_1^*$ as follows. We develop an expression for the second-step empirical wavelet coefficients in (4.3),

$$\begin{aligned} (\hat{\theta}^E)^T &= \widehat{Y}_1^{*T}W^T \\ &= \left[Y^T, \hat{\theta}_1^{*T}B^T\right]\begin{pmatrix} A \\ B \end{pmatrix} \\ &= Y^TA + \hat{\theta}_1^{*T}B^TB \\ &= Y^TA + \hat{\theta}_1^{*T}(I - A^TA) \\ &= \hat{\theta}_1^{*T} + (Y - A\hat{\theta}_1^*)^TA \\ \Rightarrow \hat{\theta}^E &= \hat{\theta}_1^* + A^Te_1^*, \end{aligned}$$

where e_1^* are residuals. Let us finally note that $\hat{\theta}^E$ is an interpolator because ($AA^T = I_{n\times n}$ and so)

$$A\hat{\theta}^E = A\hat{\theta}_1^* + e_1^* = Y.$$

The steps of the simplified algorithm are then as follows.

1. Get $\hat{\theta}^* = \hat{\theta}_D$.
2. Do weighted theresholding of $\hat{\theta}^*$ to get $\hat{\theta}_1^*$ (which could be called $\hat{\theta}_{\mathrm{NRSI}}^*$).
3. Calculate the residuals $e_1^* = Y - A\hat{\theta}_1^*$. (This avoids calculation of $B\hat{\theta}_1^*$.)
4. Calculate the second-step empirical wavelet coefficients $\hat{\theta}^E = \hat{\theta}_1^* + A^Te_1^*$. (More computations are saved here.)
5. Calculate $\hat{\theta}_1^E$, i.e., $\hat{\theta}_{\mathrm{ROSE}}^E$, by weighted thresholding of $\hat{\theta}^E$.

To plot the final estimator, we need the function estimate $\hat{f}_{\mathrm{ROSE}} = W^T\hat{\theta}_{\mathrm{ROSE}}^E$.

If we denote the componentwise threshold operator by H, we can see that if the iteration converges (say, to $\hat{\theta}_R$), it must satisfy

$$\hat{\theta}_R = H(\hat{\theta}_R + A^T(Y - A\hat{\theta}_R)).$$

It would be interesting to be able to say something about this equation; for example, how does its solution relate to the optimizer of (1.2)?

With all these estimation procedures, there is the problem of estimating tuning parameters. The authors present an

V. Solo is Professor, School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia.

asymptotic rule for the threshold $p_0 = \sigma v_j \sqrt{2 \log n}$. However, for traditional wavelet thresholding, SURE (Stein's unbiased risk estimator) was also suggested (Donoho and Johnstone 1995). I subsequently showed (Solo 1996) how SURE provides a powerful method for generating tuning parameter estimators for arbitrary linear or nonlinear ill-conditioned inverse problems with linear or nonlinear regularization. This approach was illustrated for total variation denoising in Solo (1999, 2000) and for optical flow in Ng and Solo (1997).

For regularly gridded data with $e = y - \theta$, SURE is given by

$$\widehat{R} = \frac{\|e\|^2}{n} - \frac{2\gamma\sigma^2}{n} + \sigma^2,$$

$$\gamma = \Sigma_1^n \frac{\partial e_i}{dy_i},$$

$$= n - \Sigma_1^n \frac{\partial \hat{\theta}_i}{\partial y_i}.$$

From Theroem 1, result 2 (replace z with y), we find

$$\frac{\partial \hat{\theta}}{\partial y} = \frac{I(|y| \geq p_0)}{1 + p''_\lambda(|\hat{\theta}|)}.$$

And so

$$\gamma = n - \Sigma_{\hat{\theta}_i \neq 0} \frac{1}{1 + p''_\lambda(|\hat{\theta}_i|)}.$$

With a weighting as used by the authors, we get

$$\gamma = n - \Sigma_{\hat{\theta}_i \neq 0} \frac{1}{1 + v_i p''_\lambda(|\hat{\theta}_i|)}.$$

The empirical behavior of the consequent approximate SURE remains to be seen.

REFERENCES

DeVore, R., and Lucier, B. (1992), "Fast Wavelet Technique for Near Optimal Image Processing," in *IEEE MilCom Conference Record*, pp. 1129–1135.

Donoho, D., and Johnstone, I. (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224.

Donoho, D. L., Johnstone, I. M., Hock, J. C., and Stern, A. S. (1992), "Maximum Entropy and the Nearly Black Object," *Journal of the Royal Statistical Society*, Ser. B, 54, 41–81.

Johnstone, I., and Silverman, B. (1997), "Wavelet Threshold Estimators for Data With Correlated Noise," *Journal of the Royal Statistical Society*, Ser. B, 59, 319–352.

Li, S. (1995), "On Discontinuity-Adaptive Smoothness Priors in Computervision," *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 17, 576–586.

Ng, L., and Solo, V. (1997), "A Data-Driven Method for Choosing Smoothing Parameters in Optical Flow Problems," in *Proceedings of IEEE International Conference on Image Processing 1997*, pp. III:360–363.

Ng, L., and Solo, V. (1999), "Optical Flow Estimation Using Adaptive Wavelet Zeroing," in *Proceedings of IEEE International Conference on Image Processing 1999*, Kobe, Japan.

Rudin, L., Osher, S., and Fatemi, E. (1992), "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D*, 60, 259–268.

Solo, V. (1996), "A SURE-Fired Way to Choose Smoothing Parameters in Ill-Conditioned Inverse Problems," in *Proceedings of IEEE International Conference on Image Processing 1996*. (vol. III), pp. 89–92.

Solo, V. (1998), "Wavelet Signal Estimation in Coloured Noise With Extension to Transfer Function Estimation," in *Proceedings of the 37th IEEE Conference on Decision and Control*, pp. 3040–3041.

Solo, V. (1999), "Selection of Regularization Parameters for Total Variation Denoising," in *Proceedings of International Conference on Acoustics Speech and Signal Processing 1999*, IEEE, vol. III, no. 2239, pp. 3.

Solo, V. (2000), "Total Variation Denoising in Coloured Noise," in *Proceedings of International Conference on Acoustics Speech and Signal Processing 2000*, IEEE.

Stevenson, R., Schmitz, B., and Delp, E. (1994), "Discontinuity Preserving Regularization of Inverse Visual Problems," *IEEE Transactions in Systems Man and Cybernetics*, 24, 455–469.

Vogel, C., and Oman, M. E. (1996), "Iterative Methods for Total Variation Denoising," *SIAM Journal on Scientific and Statistical Computing,* vol. 17, no. 1, pp. 227–238.

Rejoinder

A. Antoniadis and J. Fan

We are very grateful to the editor, Martin Tanner, for organizing this discussion, and we would like to thank all discussants for their insightful and stimulating comments, touching both theoretical and practical aspects, and for offering some original insights and providing several relevant references from the Bayesian, signal, and image processing literature. Their contributions are very helpful for our understanding of this work from various perspectives.

1. NONQUADRATIC PENALIZED LEAST SQUARES

We are most grateful to Professor Solo for providing an efficient simplification of our two-step algorithm for computing the ROSE estimator. Such a simplification fully exploits the structure of the discrete wavelet transform matrix W. It helps a great deal for our understanding of the two-step estimation procedure and leads to a much more efficient method for computing the ROSE estimator. By using his notation, Professor Solo raises an insightful question on the existence of the solution to the equation

$$\hat{\theta}_R = H\left(\hat{\theta}_R + A^T\left(Y - A\hat{\theta}_R\right)\right) \qquad (1)$$

and its relation to the original minimization problem (1.2). The system of Equation (1) must be satisfied if the iterative solution to the original problem (1.2) converges. In an attempt to answer this question, let us consider a specific case with the L_1-penalty function, namely, minimizing

$$J(\boldsymbol{\theta}) = \|\mathbf{Y}_n - A\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{N} |\theta_i|.$$

Journal of the American Statistical Association
September 2001, Vol. 96, No. 455, Theory and Methods

The criterion $J(\boldsymbol{\theta})$ is exactly the same as the one studied by Alliney and Ruzinsky (1994) and also by Solo (1998). Because in our case the matrix A has full rank, $J(\boldsymbol{\theta})$ is a strictly convex functional of $\boldsymbol{\theta}$ and therefore has a unique minimum. Following the development in Alliney and Ruzinsky (1994), the necessary conditions for $\tilde{\boldsymbol{\theta}}$ to be the minimizer are

$$r_k + \lambda \ \mathrm{sgn}(\tilde{\theta}_k) = 0, \quad \tilde{\theta}_k \neq 0, \quad |r_k| \leq \lambda, \quad \tilde{\theta}_k = 0,$$

where $\mathbf{r} = A^T \mathbf{Y}_n - A^T A\tilde{\boldsymbol{\theta}}$. If one uses the Gauss–Seidel kind of iterative algorithm, as in Alliney and Ruzinsky (1994), to determine the minimum point $\tilde{\boldsymbol{\theta}}$ of $J(\boldsymbol{\theta})$, one can easily see, using their expression (4.1) and the remark following their theorem 4, that the solution indeed satisfies Equation (1). The same conclusion is true if the minimization is performed with the constraint that $\boldsymbol{\theta}$ must vary within our interpolation Sobolev space. However, for more general penalty functions, the situation is more complicated because the functional $J(\boldsymbol{\theta})$ is no longer convex and therefore does not necessarily have a unique minimum. The question then is whether the solution to (1) is a local minimizer to the problem (2.1). More work is needed. However, Solo's question on the existence of the solutions to equation (1) can be answered positively. For continuous thresholding rules, by the fixed-point theorem, there exists a solution to equation (1).

2. CHOICE OF THE REGULARIZATION PARAMETER

We wholeheartedly agree with the emphasis by professors Cai, Moulin, and Solo on the importance of the choice of the regularization parameter in practice. The minimax optimal thresholding parameter and the universal thresholding rule, introduced by Donoho and Johnstone (1994) and extended further in this article, provide a simple solution from the minimax point of view. They are not necessarily optimal for a given denoising problem, but nevertheless they serve as a starting point for choosing an optimal regularization parameter. In the context of choosing the regularization parameter λ for the hard- and soft-thresholding wavelet estimators, a variety of methods, such as SURE and cross validation, have been introduced. Many of these can be extended to the penalized least squares problem with a nonquadratic penalty.

Professor Solo suggests a data-based estimator of the regularization parameter λ in the same spirit as the SURE criterion proposed by Donoho and Johnstone (1995). He provides a simple SURE formula for general penalty functions. His disarmingly simple method deserves more thorough studies, including careful and detailed simulation studies, such as those in Bruce and Gao (1996), for the classical WaveShrink procedure. We, however, speculate that this SURE method shares the same drawback as Donoho and Johnstone's SureThresh procedure in the situations of extreme sparsity of the wavelet coefficients. This drawback could possibly be addressed by using a hybrid scheme. Similar to one of Donoho and Johnstone's procedures, the following heuristic can be used: If a set of empirical wavelet coefficients are judged to be sparsely represented, then the hybrid scheme defaults to the universal thresholding rule; otherwise, the SURE criterion is used to select a thresholding value.

Another possible way to address the optimal choice of the regularization parameter is the cross-validation criterion as suggested by Professor Cai. The cross-validation criterion has been widely employed as a data-driven procedure for choosing regularization parameters. As correctly pointed out by Professor Cai, the cross-validation method is usually intensive in computation, unless there are some updating formulae that allow us to calculate the leave-one-out estimators based on the estimators using the full dataset. Such formulas are easy to derive for the projection-type of estimators, as, for example, in the wavelet-based linear methods, suggested by Antoniadis (1996) and independently by Amato and Vuza (1997). By using a weighted quadratic penalty, the parameter λ is chosen by cross validation in Amato and Vuza (1997). The approach in Antoniadis (1996) is based on risk minimization, and it depends on a preliminary consistent estimator of the noise level. Of course, the preceding linear wavelet methods are not designed to handle spatially inhomogeneous functions with low degree of regularities. For such functions, an appropriate penalty is the one of Chambolle et al. (1998) leading to nonlinear thresholding or nonlinear shrinkage methods. Recall that the leave-one-out principle underlying cross validation is achieved when an appropriate compatibility condition holds. In our setting, if $\hat{\boldsymbol{\theta}}_i^{-(i)}$ denotes the ith component of the estimator of $\boldsymbol{\theta}$ based on the sample without the ith observation, compatibility means that $\hat{\boldsymbol{\theta}}_i^{-(i)} = \tilde{\boldsymbol{\theta}}_i$ holds, where $\tilde{\boldsymbol{\theta}}_i$ is the estimator based on the sample with the ith observation replaced by its fitted value $(A\hat{\boldsymbol{\theta}}^{-(i)})_i$. Under such a condition, the cross-validation functional can be expressed in terms of the ordinary residuals. Unfortunately, for nonlinear shrinkage or thresholding, the compatibility condition is violated. One way to proceed is to pretend that it holds approximately, that is, $\hat{\boldsymbol{\theta}}_i^{-(i)} \simeq \tilde{\boldsymbol{\theta}}_i$, as done in Jansen, Malfait, and Bultheel (1997). In our case, the GCV criterion advocated by these authors takes the form

$$\mathrm{GCV}(\lambda) = \frac{\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_R\|^2/N}{(N_0/N)^2},$$

where $\hat{\boldsymbol{\theta}}$ is the wavelet coefficients from the Sobolev interpretor, $\hat{\boldsymbol{\theta}}_R$ is the vector of the ROSE coefficients with the regularization parameter λ, and N_0/N is the proportion of wavelets coefficients replaced by 0. It is clear that this is an area where further theoretical and practical work is needed.

3. BLOCKWISE SEPARABLE PENALTY FUNCTIONS

We greatly appreciate the points made by Professors Cai and Moulin on nonadditive or separable penalties, touching an area that is currently theoretically explored by many researchers. Our wavelet regularization procedure with additive penalty achieves adaptivity through term-by-term thresholding of the empirical wavelet coefficients. However, as pointed out by Professor Cai, this term-by-term rule is not optimal: it removes too many terms from the empirical wavelet expansion. As a result, the resulting wavelet estimators contain too much bias and have suboptimal L^p-risk ($1 \leq p < \infty$) in terms of rate of convergence. One way to increase estimation precision is to pull information from neighboring empirical wavelet coefficients. In other words, empirical wavelet coefficients could be thresholded or shrunk in blocks (or groups) rather than individually. As a result, the amount of information available for estimating the length of the empirical wavelet coefficients within a block would be an order of magnitude larger than in the case of a term-by-term thresholding rule. This would allow thresholding decisions to be made more accurately and permit convergence rates to be improved. This is the spirit of blockwise thresholding rules outlined in the discussion by Professor Cai. The blockwise thresholding estimators or their extensions can be regarded as the penalized least squares with blockwise additive penalty functions. This is elucidated below.

Professor Cai has correctly pointed out that the blockwise penalty function

$$p(\theta_{(b)}) = \lambda I\{\theta_{(b)} \neq 0\}$$

results in a blockwise thresholding estimator

$$\hat{\theta}_i = z_i I\{\|z_{(b)}\| \geq \lambda\} \quad \text{for } i \in (b).$$

This complexity penalty function is discontinuous, making computation more involved. It is a limit of the family of entropies $\sum_{(b)} \|\theta_{(b)}\|^{\alpha}$ as $\alpha \to 0$. (This partially answers the query raised by Professor Moulin regarding why this complexity penalty is called an entropy penalty in our text.) An important contribution of this work is to replace the indicator function by a smoother function, called the hard-thresholding penalty function. This enhances the degree of regularity of the penalized least squares criterion and hence facilitates the computation. In the blockwise inference context, a natural extension of our work is to minimize, in the wavelet coefficients domain, the following penalized least squares:

$$\sum_{(b)} \|z_{(b)} - \theta_{(b)}\|^2 + \sum_{(b)} p_\lambda(\|\theta_{(b)}\|), \tag{2}$$

where $p_\lambda(\cdot)$ is a penalty function given in Theorem 1. Similar to equation (3) of Professor Moulin's contribution, the flexibility can be further enhanced by introducing a weight $\lambda_{(b)}$ in the penalty part of (2) or more generally by using a block-dependent penalty function $p_\lambda^{(b)}(\|\theta_{(b)}\|)$. The solution to (2) and its generation is blockwise separable, leading to minimizing

$$\|z_{(b)} - \theta_{(b)}\|^2 + p_\lambda(\|\theta_{(b)}\|). \tag{3}$$

The solution to problem (3) is surprisingly simple. Let us drop the subscript (b). Problem (3) can be written as

$$\min_r \left\{ \min_{\|\theta\|=r} \|z - \theta\|^2 + p_\lambda(r) \right\}. \tag{4}$$

The minimum of the terms inside the brackets is achieved at $\hat{\theta}_r = rz/\|z\|$. Substituting this into (4), we need to minimize

$$\min_r \{(\|z\| - r)^2 + p_\lambda(r)\}. \tag{5}$$

The solution problem to (5) is characterized in Theorem 1. Let $\hat{r}(\|z\|)$ be the solution to (5). Then, the minimizer to the problem (3) is given by

$$\hat{\theta}(z) = \hat{r}(\|z_{(b)}\|) z_{(b)} / \|z_{(b)}\|.$$

In particular, when $\|z_{(b)}\| \leq \min_{r \geq 0}\{r + p'_\lambda(r)\}$, $\hat{\theta}(z) = 0$. Specifically, if the hard-thresholding penalty function (2.8) is used,

$$\hat{\theta}(z_{(b)}) = z_{(b)} I(\|z_{(b)}\| \geq \lambda),$$

and when the L_1-penalty function (2.5) is employed,

$$\hat{\theta}(z_{(b)}) = \left(1 - \lambda/\|z_{(b)}\|\right)_+ z_{(b)}.$$

The former is the same as the blockwise thresholding rule given in Professor Cai's discussion, with the advantage that the penalty function is smoother.

The blockwise penalized least squares (2) admit nice Bayesian interpretation. They model the correlated wavelet coefficients in the same block by using some correlated prior distributions [spherically symmetric distributions in (2)]. Abramovich, Besbeas, and Sapatinas (2000) considered Bayesian wavelet block shrinkage and blockwise thresholding estimators and indicated that they outperform existing classical blockwise thresholding estimators in terms of mean squared error via simulation studies. It would be interesting to see, with the remarks made by Professor Vidakovic, if these estimators can have a similar MAP interpretation.

We thank Professor Cai and Professor Moulin for suggesting some ways to select the smoothing parameters, block length, and threshold level. As Professor Cai points out convincingly, there is much work ahead of us for analyzing such procedures.

4. BAYESIAN WAVELET MODELING

We are very grateful to Professors Moulin and Vidakovic for drawing more attention to the links of the regularization problem with Bayesian modeling and inferences in the wavelet domain and for providing relevant references in the literature. Professor Vidakovic's identification of the priors gives not only nice Bayesian interpretation of our proposed procedures but also the possibility of deriving adaptive choice of thresholding rules.

When prior distributions are a scale mixture of two distributions, an important distinction should be made between the two types of mixtures: a scale mixture of two normal distributions [considered by Chipman, Kolaczyk, and McCulloch (1997)] and a scale mixture of a normal distribution with a point mass at zero [considered by Clyde, Parmigiani, and Vidakovic (1998) and Abramovich, Sapatinas, and Silverman (1998)]. The former prior distributions lead to nonvanishing Bayesian estimates of wavelet coefficients, yet the latter prior distributions can result in the Bayesian estimates of wavelet coefficients that possess bona fide thresholding properties. In both cases, as long as the prior is proper, the resulting Bayesian estimates of wavelet coefficients will have bias in the frequentist analysis, even when the true wavelet coefficients are way above the noise level. This kind of unnecessary bias can be attenuated by using the prior distributions with heavier tails; this would result in penalty functions with flatter tails.

Consider the hierarchal model

$$\theta \mid \gamma \sim N(0, \gamma\tau^2),$$
$$\gamma \sim \text{ Bernoulli}(\delta).$$

The binary random variable γ determines whether the wavelet coefficient is nonzero ($\gamma = 1$), arising from a $N(0, \tau^2)$ distribution, or zero ($\gamma = 0$), arising from a point mass at zero. It is assumed that γ has a Bernoulli distribution with $P(\gamma = 1) = 1 - P(\gamma = 0) = \delta$ for some fixed hyperparameter $0 \leq \delta \leq 1$. The prior parameters δ and τ^2 can depend on the resolution level, resulting in level-dependent wavelet threshold and shrinkage estimators. Once z is observed, the posterior distribution on the wavelet coefficient θ is given by

$$\theta \mid \gamma, z, \sigma^2 \sim N\left(\gamma \frac{\tau^2}{\sigma^2 + \tau^2} z, \gamma \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right).$$

To incorporate model uncertainty about which of the wavelet coefficients θ are zero, we now average over all possible γ. The marginal posterior distribution of θ conditionally on σ^2 is then given by

$$\theta \mid z, \sigma^2 \sim p(\gamma = 1 \mid z, \sigma^2)\, N\left(\frac{\tau^2}{\sigma^2 + \tau^2} z, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) + (1 - p(\gamma = 1 \mid z, \sigma^2))\, \delta(0),$$

where $\delta(0)$ is a point mass at zero. It is not difficult to see that

$$p(\gamma = 1 \mid z, \sigma^2) = \frac{1}{1 + O(z, \sigma^2)},$$

where the posterior odds ratio $O(z, \sigma^2)$ is given by

$$O(z, \sigma^2) = \frac{1 - \delta}{\delta} \frac{(\sigma^2 + \tau^2)^{1)2}}{\sigma} \exp\left(-\frac{\tau^2 z^2}{2\sigma^2(\sigma^2 + \tau^2)}\right).$$

This leads to thresholding or shrinkage, according to the Bayes rules that are used.

ADDITIONAL REFERENCES

Abramovich, F., Besbeas, P., and Sapatinas, T. (2000), "Empirical Bayes Approach to Block Wavelet Function Estimation," Technical report, Dept. of Mathematics and Statistics, University of Cyprus.

Alliney, S., and Ruzinsky, S. A. (1994), "An Algorithm for the Minimization of Mixed ℓ_1 and ℓ_2 Norms With Application to Bayesian Estimation," *IEEE Transactions in Signal Processing*, 42, 618–627.

Amato, U., and Vuza, D. T. (1997), "Wavelet Approximation of a Function From Samples Affected by Noise," *Revue Roumanie de Mathematiques pures et Appliquées*, 42, 481–493.

Bruce, A. G., and Gao, H.-Y. (1996), "Understanding WaveShrink: Variance and Bias Estimation," *Biometrika*, 83, 727–745.

Chipman, H. A., Kolaczyk, E. D., and McCullogh, R. E. (1997), "Adaptive Bayesian Wavelet Shrinkage," *Journal of the American Statistical Association*, 92, 440, 1413–1421.

Jansen, M., Malfait, M., and Bultheel, A. (1997), "Generalised Cross-Validation for Wavelet Thresholding," *Signal Processing*, 56, 33–44.