Regularization of Wavelets Approximations*

Anestis Antoniadis Jianqing Fan

December 9, 1999

Abstract

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 (1994) are specific members of nonlinear regularized wavelet estimators. They correspond to the lower and upper bound 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 (ROSE), which are the one-step estimators of the penalized least-squares problems using the NRSI as initial estimators. Two other approaches, the graduated nonconvexity algorithm and wavelet networks, are also introduced to handle penalized least-squares problems. The newly introduced approaches are also illustrated by a few numerical examples.

Anestis Antoniadis is Professor, Laboratoire de Modélisation et Calcul, Université Joseph Fourier, 38041 Grenoble CEDEX 9, France. Jianqing Fan is Professor, Department of Statistics, University of California, Los Angeles, CA 90095-1554 and The Chinese University of Hong Kong. Fan's research was partially supported by NSF grant DMS-9804414. The major part of the research was conducted while Fan visited the Université Joseph Fourier. He is grateful to the generous support of the University.

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 *et al.* (1994), Hall and Patil (1995), Neumann and Spokoiny (1995), Antoniadis (1996) and Wang (1996). Further references can be found in the recent survey papers by Donoho *et al.* (1995) and Antoniadis (1997) and recent 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 should 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 *et al.* (1997), the transformation method of Cai and Brown (1997), and the isometric method of Sardy *et al.* (1998). However, it poses some challenges to extend these methods to other statistical contexts such as generalized additive models and generalized ANOVA 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 $\{t_1, \cdots, t_n\}$:

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

where f is the 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 $\{i/2^J, i = 0, \cdots, 2^{J-1}\}$. Let $\mathbf{W}$ be a given wavelet transform and $\boldsymbol{\theta} = \mathbf{W}\mathbf{f}$ be the wavelet transform of $\mathbf{f}$. Since $\mathbf{W}$ 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 over parameterized 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. Since 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 i^{th} 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

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

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 in order to produce sparse solutions. See Theorem 1 for necessary conditions. It poses some challenges to optimize such a high-dimensional nonconvex function.

Our over parameterization approach is complementary to the over-complete wavelet library methods of Chen *et al.* (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 over-complete family of non-orthogonal 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 can be easily found. Theorem 1 gives necessary conditions for the solution to be unique and to be 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. Since the penalized least-squares estimators perform comparably with the oracle estimator, following the similar but easier calculation as that of Donoho *et al.* (1995), we can show that the penalized least-squares estimators 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. One possible method is to apply graduated nonconvexity (GNC) algorithm introduced by Blake and Zisserman (1987) and Blake (1989) and ameliorated by Nikolova (1999a) and Nikolova *et al.* (1999) in the imaging analysis context. The algorithm contains good ideas in optimizing high-dimensional nonconvex functions, but its implementation depends on a number of tuning parameters. It is reasonably fast, but not nearly as fast as 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 coefficient-wise 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. 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 non-equispaced signals is to design adaptively non-orthogonal wavelets to avoid overparameterizing problems. A viable approach is the wavelet networks proposed by Mallat *et al.* (1999). We will briefly discuss this in Section 6.

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 similar manner 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 have been successfully used by Tibshirani (1996), Barron *et al.* (1999) and Fan and Li (1999) for variable selections. Thus, they should also be viable for wavelet applications to other statistical problems. While there is no conceptual difficulty of 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.

The paper 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 relegated 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 down the class of candidate solutions.

We will first briefly discuss on wavelet interpolation by using a regularized wavelet method. This will serve as a crude initial value to our proposed penalized least-squares method. We will 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 $\{t_i, i = 1, \cdots, n\}$ 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

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

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

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

Since 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 j-th resolution level and k-th dyadic location ($k = 0, \cdots, 2^{j-1}$). A Sobolev norm of f with degree of smoothness s can be expressed as

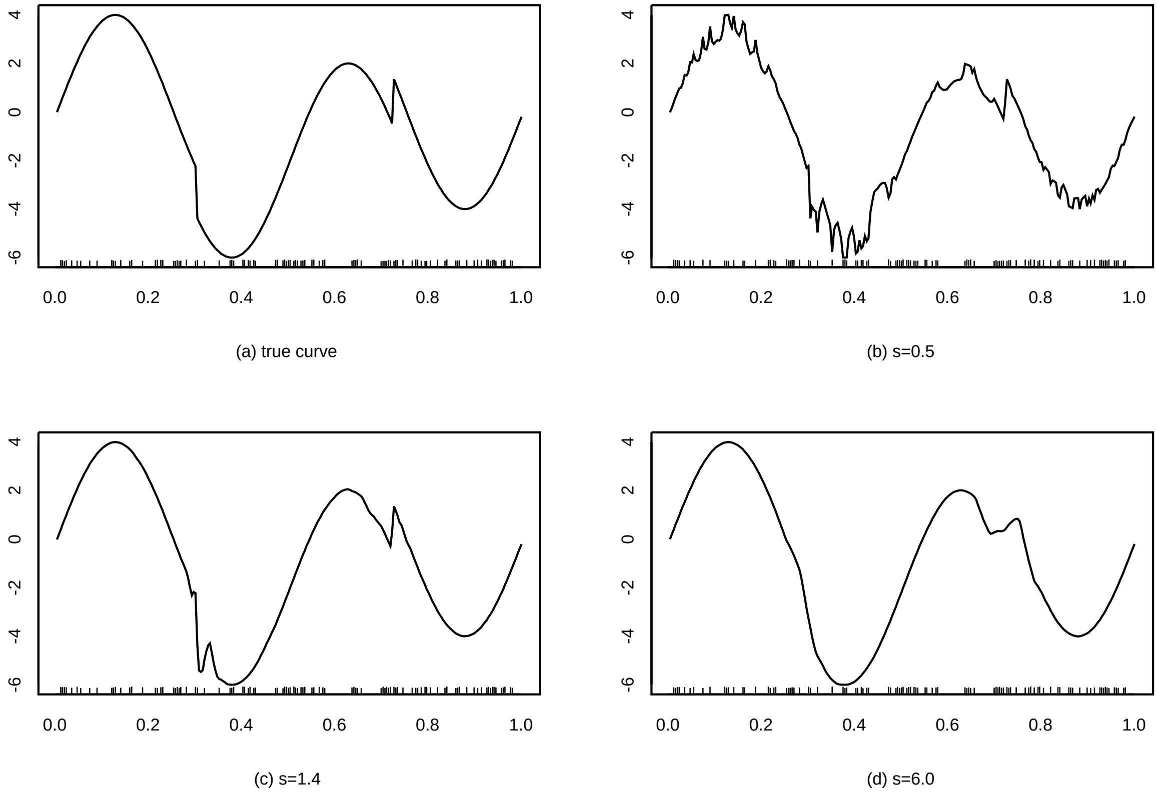
$$\|\boldsymbol{\theta}\|_S^2 = \sum_j 2^{2sj} \|\boldsymbol{\theta}_{j\cdot}\|^2,$$

where $\boldsymbol{\theta}_{j\cdot}$ 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 (see Rao) 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 that θ_i is associated with. An advantage of the method of frame is that it does not involve the choice of regularization parameter.

When $s = 0$, $\boldsymbol{\theta} = \mathbf{A}^T\mathbf{f}_n$ by orthogonality. In this case, the interpolator is particularly easy to compute.



(a) true curve

(b) s=0.5

(c) s=1.4

(d) s=6.0

Figure 1: Illustration of wavelet interpolations by the method of frame. As degrees of smoothness s gets larger, the interpolated functions get smoother. (a) The target function and sampling points (tick marks); (b) – (d) wavelet interpolations with $s = 0.5$, $s = 1.4$ and $s = 6.0$.

As an illustration of how the regularized wavelet interpolations work, we took a hundred data points (located at the tick marks) from the function depicted in Figure 1(a). Figures 1 (b)–(d) show how the method of frame works for different values of s. As s increases, the interpolated functions get smoother and 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

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

This leads 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 have 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 are in general nonlinear. See the results of Nikolova (1999b) 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 $n = 2^J$, 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 i^{th} 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 since $p(|\theta|)$ 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, Johnstone, Hock and Stern, 1992). A clipped-L_1 penalty

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

leads to a mixture of soft and hard thresholding rule (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

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

(see Figure 2(b)) the solution is the hard-thresholding rule (Antoniadis, 1997). This is a smoother penalty function than $p_\lambda(|\theta|) = |\theta| I(|\theta| < \lambda) + \lambda/2 I(|\theta| \geq \lambda)$ 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, while 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 following quadratic spline penalty, called smoothly clipped absolute deviation (SCAD) penalty (see Figure 2(c)):

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

leading to the piecewise linear thresholding

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

In Fan and Li (1999), it is recommended to use $a = 3.7$ based on a Bayesian argument. This thresholding estimator is in the same spirit to that in Gao and Bruce (1997). This penalty function does not over penalize large values of $|\theta|$ and hence does not create excessive biases when the wavelet coefficients are large. Recently, Nikolova (1999b) suggests 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 of them are concave on $[0, \infty)$ and do not intend to over penalize large $|\theta|$. Other possible loss functions include L_p penalty

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

As to be shown in Section 3.1, the choice $p \leq 1$ is a necessary condition for the solution to be a thresholding estimator, while $p \geq 1$ 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 would like to 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 to estimate functions with irregularities

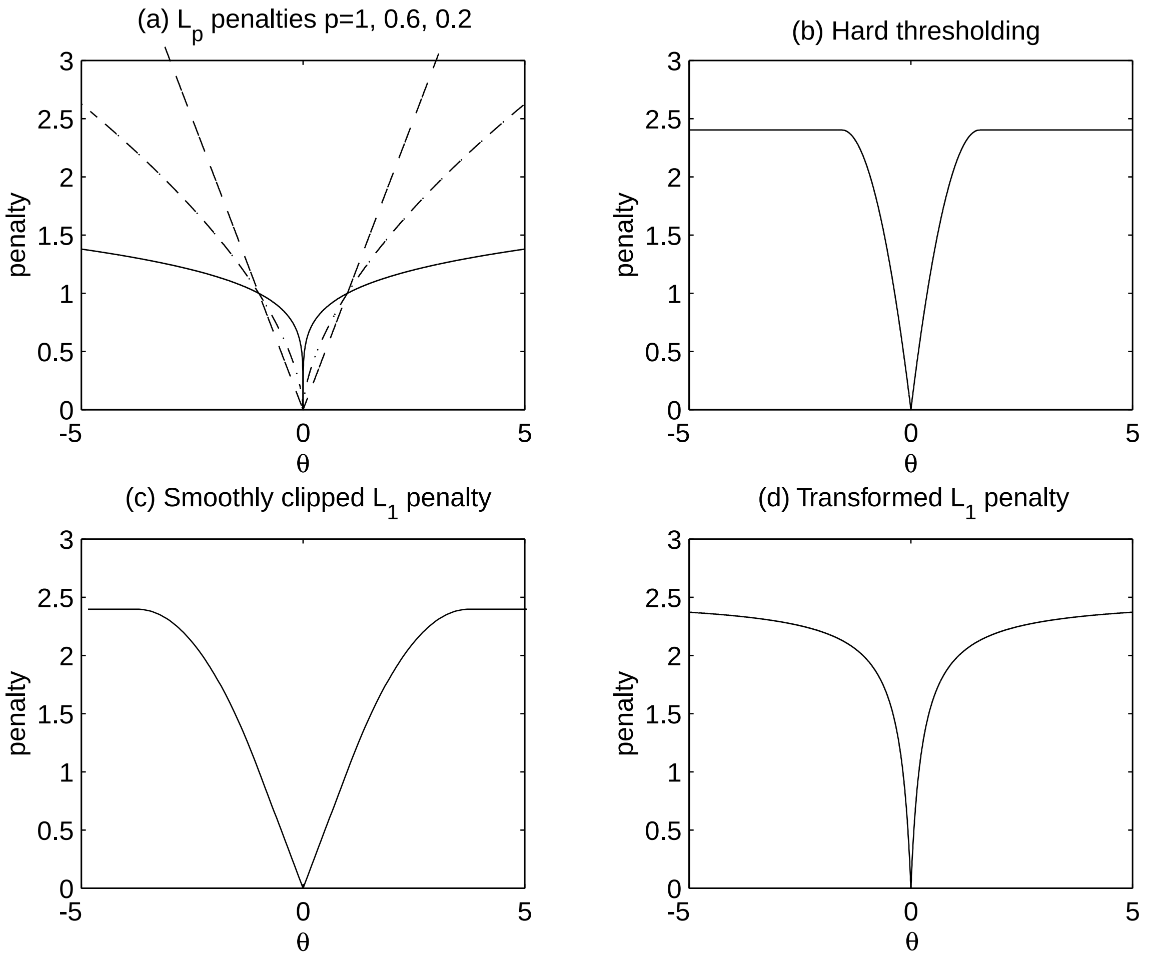
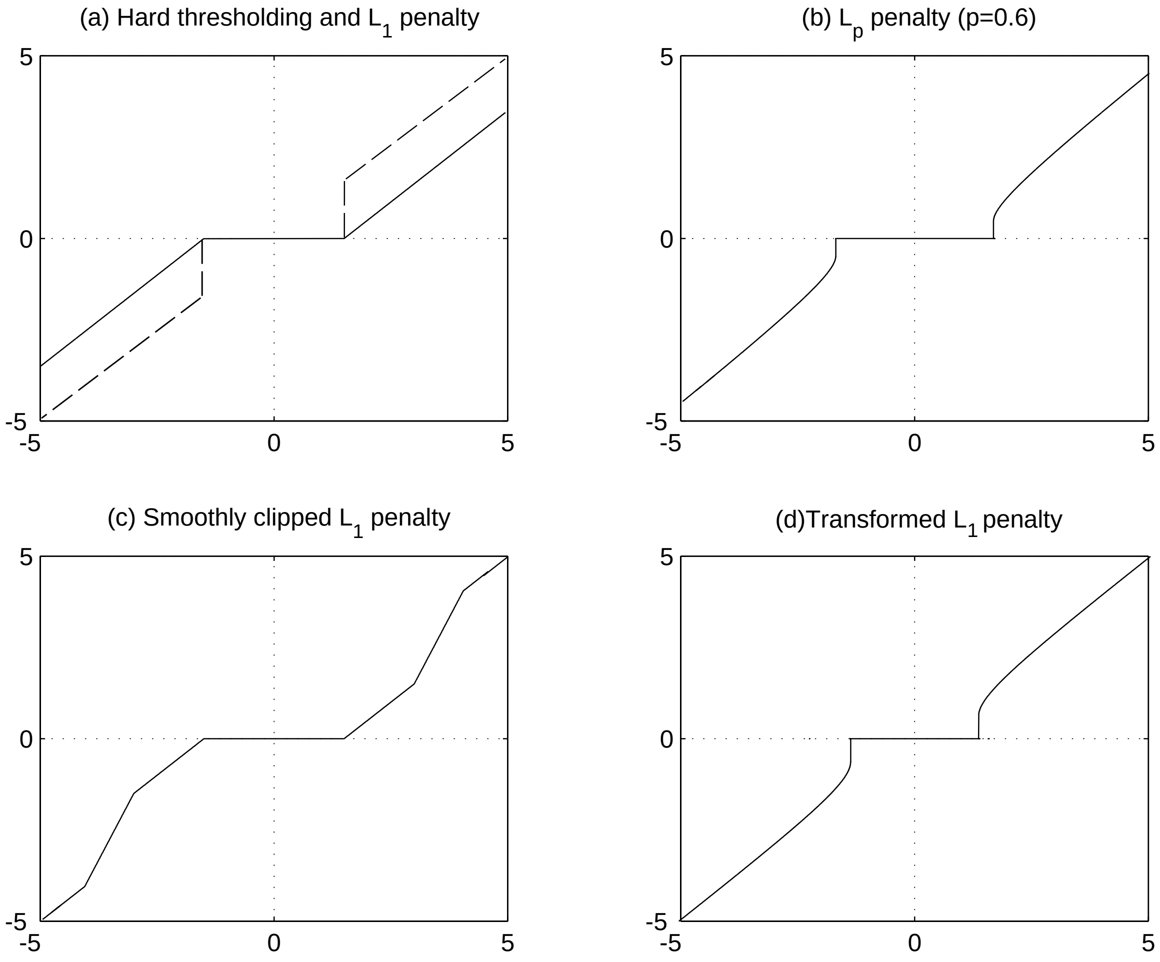


Figure 2: Examples of typical penalty functions that preserve sparsity. (a) L_p penalty with $p = 1$ (long dash), $p = 0.6$ (short dash) and $p = 0.2$ (solid); (b) Hard-thresholding penalty (2.8); (c) SCAD (2.9) with $a = 3.7$; (d) Transformed L_1-penalty (2.11) with $b = 3.7$.

and to denoise images with sharp features. However, there have not been systematically studied yet. We first study the properties of penalized least-squares estimators and then examine the extent to which they can mimic oracle in choosing a subset of wavelet bases.

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:



(a) Hard thresholding and L_1 penalty

(b) L_p penalty (p=0.6)

(c) Smoothly clipped L_1 penalty

(d)Transformed L_1 penalty

Figure 3: Examples of penalized least-squares estimators that possess thresholding properties. (a) The penalized L_p estimators with $p = 1$ (solid) and the hard-thresholding estimator (dashed); (b) the penalized L_p estimator with $p = 0.6$; (c) the penalized SCAD estimator (2.10); (d) the penalized transformed L_1 estimator with $b = 3.7$.

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

(i) The solution to the minimization problem (3.1) exists and is unique. It is anti-symmetric: $\hat{\theta}(-z) = -\hat{\theta}(z)$.

(ii) The solution satisfies

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

where $p_0 = \min_{\theta \ge 0}\{\theta + p'_\lambda(\theta)\}$. *Moreover,* $|\hat{\theta}(z)| \le |z|$.

(iii) If $p'_\lambda(\cdot)$ *is nonincreasing, then for* $|z| > p_0$, *we have*

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

(iv) 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.*

(v) 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 the above results. When $p'_\lambda(0+) > 0$, $p_0 > 0$. Thus, for $|z| \le p_0$, the estimate is thresholded to zero. 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 (iii). 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 would like to add that 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 using different thresholding parameters. Further, the amount of shrinkage gradually tapes 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 quite 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 (iv) 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(iii), 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

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

For wavelet applications, the thresholding parameter p_0 will be in the order of magnitude of the maximum of the Gaussian errors. Thus, we only consider 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 that p satisfies conditions in Theorem 1 and $p'_\lambda(0+) > 0$. Then*

(i) $R_p(\theta, p_0) \le 1 + \theta^2$.

(ii) If $p'_\lambda(\cdot)$ is nonincreasing, then

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

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

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

Note that properties (i) – (iv) 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 or not, 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 will be the sample size.

When $p_0 = \sqrt{2 \log n}$, the universal thresholding proposed by Donoho and Johnstone (1994), by property (iii) of Theorem 2,

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

which is larger than the ideal risk. To bound the risk of 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 since 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. That is,

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

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,

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

It is noted by Donoho and Johnstone (1994) that the universal thresholding is somewhat too large. This is also observed in practice. In this section, we propose a new universal thresholding policy, which 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 (ii), for any penalized least-squares estimator, we have

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

provided that $p_0 \leq \sqrt{2\log n}$. 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 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*

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

for the universal thresholding

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

with $n \geq 4$ and $c \geq 1$ 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

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

Thus, the penalized least-squares perform comparably with the oracle estimator within a logarithmic order. We would like to remark that this conclusion holds for the thresholding parameter $p_0 = \sqrt{\alpha \log n}$ for any $\alpha \geq 2$. 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

$c = 1$

n	p_n	$a^{\dagger}_{n,c}$	$(2\log n)^{1/2}$	$\Lambda^*_{n,c}$	$\Lambda^*_n(\mathrm{DJ})$	Λ_{n,c,a_n}	$b^{\ddagger}_{n,c}$
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

$c = 16$

n	p_n	$a^{\dagger}_{n,c}$	$(2\log n)^{1/2}$	$\Lambda^*_{n,c}$	$\Lambda^*_n(\mathrm{DJ})$	Λ_{n,c,a_n}	$b^{\ddagger}_{n,c}$
64	0.791	1.160	2.884	1.346	3.124	1.879	12.396
128	0.951	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

$^{\dagger}a_n = (2\log n - \log(1 + c^2 \log n))^{1/2}$ — the new thresholding parameter;

$^{\ddagger}b_n = 2\log n + 2(\log n)^{1/2}$ — the upper bound of minimax risk.

Table 1: Coefficient p_n and related quantities for the SCAD penalty for several values of c and n. The coefficient $\Lambda^*_n(\mathrm{DJ})$ is the one computed by Donoho and Johnstone in their Table 2 for the soft-thresholding estimator using the universal thresholding p_0.

1 were computed using a grid search over p_0 with increments 0.001. For a given p_0, the supremum over θ was computed 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 in 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 of these bounds are much sharper than the upper bound $b_{n,c}$. For $c = 1$, $\Lambda^*_{n,c}$ for the SCAD estimator is somewhat smaller than that of the soft-thresholding estimator $\Lambda^*_n(\mathrm{DJ})$.

3.4 Performance of regularized wavelet estimators

The above oracle inequalities can directly be applied to the regularized wavelet estimators defined via (2.3) when the sampling points are equispaced and $n = 2^J$. Suppose that 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 benchmark of our comparison. For simplicity of presentation, we assume that $i_0 = 1$.

By translating the problem in the function space into the wavelet domain, using 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*

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

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

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

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 assume that the noise level is of $1/n$. By simplified calculations of Donoho, Johnstone, Kerkyacharian and Picard (1995), we have

Theorem 4 *Suppose that 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.

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 being done with smoothing splines. However, the estimators derived in this way, while easy to compute, are linear. They tend to oversmooth sharp features such as jumps and short aberrations of regression functions and will not in general recover such important attributes of regression functions. In contrast, nonlinear regularization methods such as the ones studied in the previous sections can recover efficiently 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 not anymore orthonormal.

Finding a solution to the minimization problem (2.3) cannot be done by using classical optimization algorithms, since the penalized loss $\ell(\boldsymbol{\theta})$ to be minimized is nonconvex, nonsmooth and high-dimensional. In this section, we introduce a regularized one-step estimator (ROSE) to solve approximately the minimization problem (2.3). It is related to 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 advantages of the orthonormality of the wavelet matrix $\mathbf{W}$. Let us consider again 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 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 don't know $\boldsymbol{\theta}$ and one possible way is to iteratively optimize (4.1). While 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 (2000) 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 either using fully iterative method or 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

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

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

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

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 will call the components of $\mathbf{W}\hat{\mathbf{Y}}_0^*$ 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 component thresholding should be applied. Details are given in §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 called Nonlinear Regularized Sobolev Interpolator (NRSI).

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 also be approximated by using the properties of wavelets.

As to be shown in §4.3, the NRSI possess 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

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

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

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

by componentwise minimization technique. The resulting procedure is a one-step procedure with a good initial estimator. This procedure will be called a Regularized One-Step Estimator (ROSE). According to Bickel (1975), Robinson (1988) and Fan and Chen (2000), such a procedure is as good as fully-iterated procedure when the initial estimators are good enough. Formal technical derivations of the statement is beyond the scope of this paper.

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 method in Johnstone and Silverman (1997), who have recently 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 noise level at the j^{th} 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

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

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

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

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

$$E(\hat\theta_i - \theta_i^*)^2 \le (2\log n + 2\log^{1/2} n)[c\sigma^2 v_i/n + \min(\theta_i^{*2}, \sigma^2 v_i)], \tag{4.5}$$

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

In the above thresholding, one can also take $p_i = \sqrt{2v_i\log N}\,\sigma$. The result (4.5) continues to hold. The constant 2 in p_i can also replaced by any constant that is no smaller than 2.

In practice, the value of σ^2 is usually unknown and needs to be estimated. In the complete orthogonal case, Donoho *et al.* (1995) have 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 0.6745. 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 zero due to large gap in the design leading to values of synthetic wavelet coefficients that are also close to zero. Taking these into account we suggest and have used the following estimator

$$\hat\sigma = \mathrm{MAD}\{Z_{J-1,k}^*/\sqrt{v_{J-1,k}} : v_{J-1,k} > 0.0001\}/0.6745,$$

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

$$R_p(f) = n^{-1}\mathbb{E}\{\|\mathbf{A}\hat{\boldsymbol{\theta}}_1 - \mathbf{A}\boldsymbol{\theta}\|^2\} = n^{-1}\mathbb{E}\{\|\mathbf{A}\hat{\boldsymbol{\theta}}_1 - \mathbf{A}\boldsymbol{\theta}^*\|^2\} \leq n^{-1}\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2. \tag{4.6}$$

By (4.5), the mean-squared errors are bounded as follow:

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

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

where $\mathrm{tr}(\mathbf{V})$ *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 $B^r_{p,q}$ for the specific case with $s = 0$. To this end, we need some technical conditions. First of all, 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 $\{t_1, \cdots, t_n\}$. Assume that there exists a distribution function $G(t)$ with density $g(t)$, which is bounded away from zero 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 r^{th} 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 technicality, which does not provide us insightful understanding, we assume

$$W_{j,k}(i) = 2^{-(J-j)/2}\psi(2^j i/N - k), \qquad \forall i, j, k.$$

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 two simulated data sets and two real data examples. The NRSI is used as an initial estimate. The ROSE method is employed with the SCAD penalty.

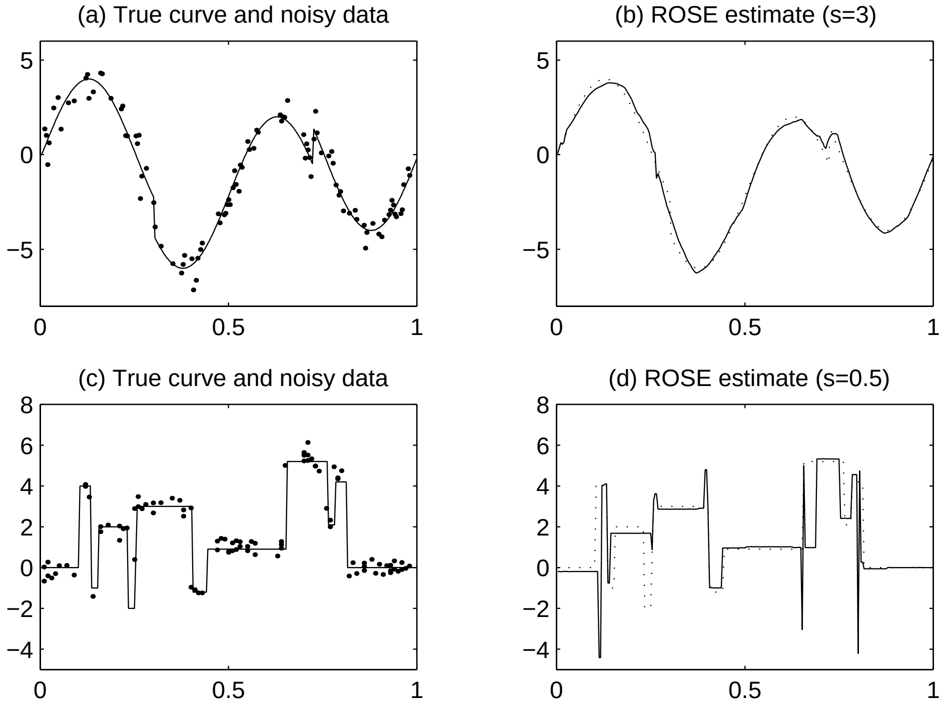


Figure 4: Estimates by using ROSE for two simulated data. (a) and (c) Simulated data and true regressions; (b) and (d) Estimate by using ROSE (solid curve) and true regressions (dashed curves).

For simulated data, we use the functions "heavisine" and "blocks" 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 $\sigma = 0.5$ for the heavisine function and $\sigma = 0.3$ for the "blocks" 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. The simulated data and the testing functions are shown in Figures 4(a) and 4(c). The ROSE estimates were computed using the symmlets of order 6 and $s = 3$ for the heavisine function and the Haar wavelets and $s = 0.5$ for the "blocks" function. 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. 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 $s = 3$.

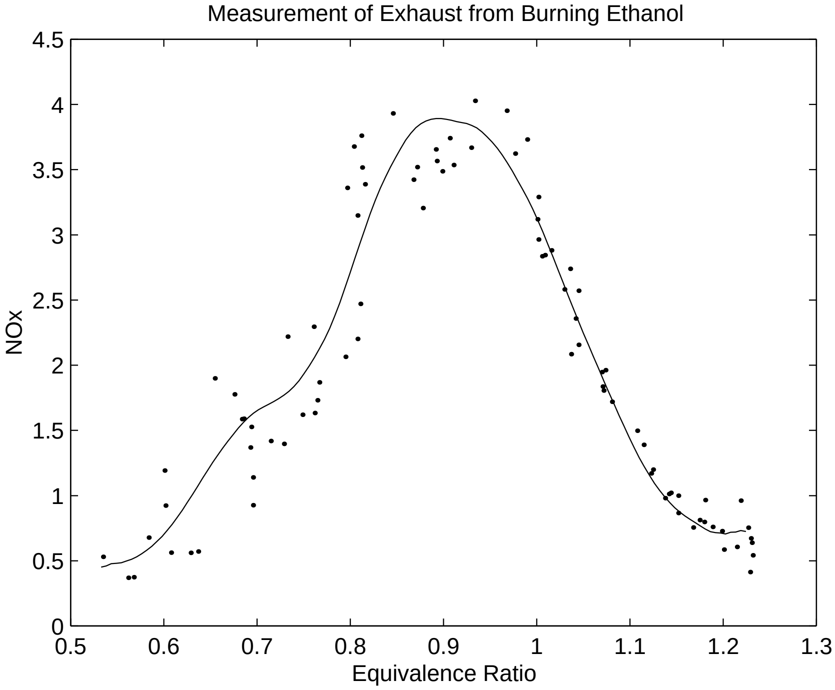


Figure 5: Measurements of exhaust from burning ethanol: observed data (points) and ROSE estimates (solid curve).

We now apply our regularized wavelet estimators to the ethanol data. This data set has also been discussed extensively by Brinkman (1981) and Chambers and Hastie (1992). The data which are displayed in Figure 5 consist of 88 measurements from an experiment in which ethanol was burned in a single cylinder automobile test engine. The concentration of the sum of nitric oxide (NO) and nitrogen dioxide (NO_2) in engine exhaust, normalized by the work done by the engine, is related to equivalent ratio, a measure of the richness of the air/ethanol mix. To analyze this data set we have used a wavelet interpolation with symmlets of order 6, $N = 128$ and a smoothness index $s = 1.5$. The resulting estimate by our procedure is plotted on the figure as a solid line and seems to fit the data well, despite the fact that the variance of the data does not appear to be constant (the data exhibit more variation for smaller equivalence ratios).

The next figure shows another data set that has been analyzed extensively in the field of nonparametric regression. It has been 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 Symmlets of order six were used. Both estimates exhibit large high frequency phenomena. The second row in Figure 6 displays a robust estimate obtained by cleaning first 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 data set with a 256 point Sobolev interpolation with Symmlets of order 6, $s = 3$ and a SCAD penalty. Both estimates are obviously less disturbed by the outliers in the crash data set. There are not anymore high frequency phenomena. This example shows that ROSE by itself is quite robust to outliers.

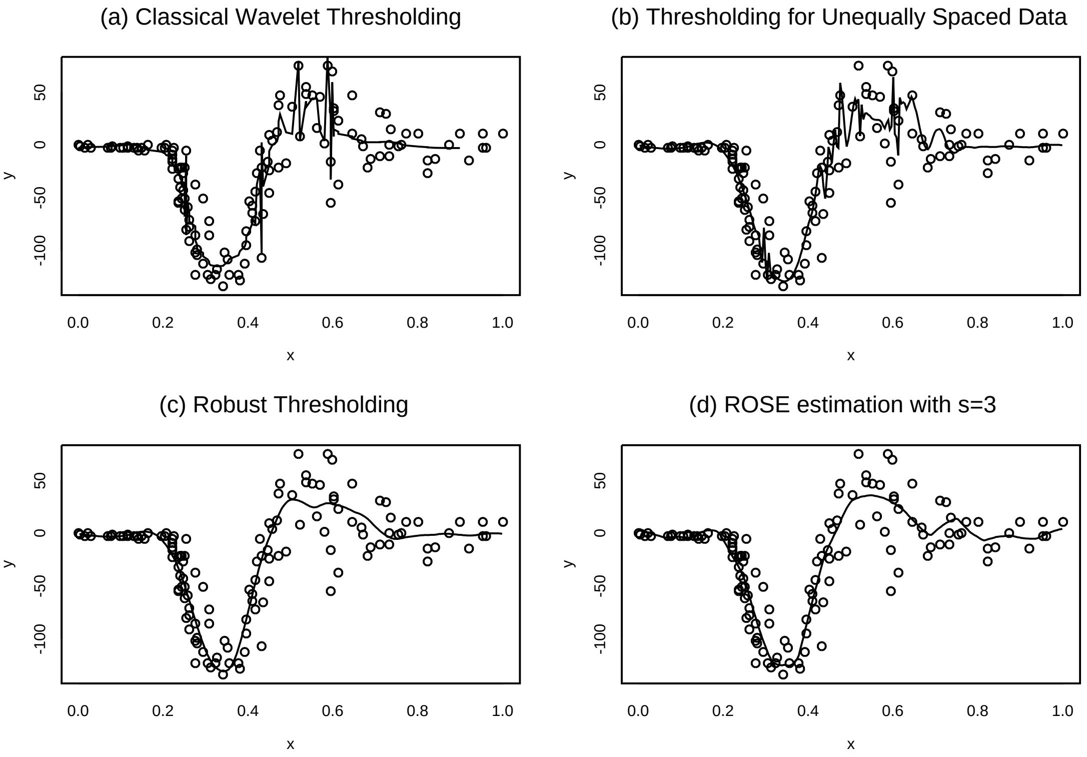


Figure 6: Crash data with several wavelet estimates

6 Other approaches

In this section, we offer two alternative approaches to estimate regression functions from nonequispaced samples. The first approach is to use the graduated nonconvexity (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. The second approach is to design a nonorthogonal wavelet, called wavelet networks, that adapts to nonuniform designs.

6.1 Graduated nonconvexity algorithm

The graduated nonconvexity algorithm was developed in the image processing context (see Blake and Zisserman 1987, Blake 1989). It is capable of minimizing a broad range of nonconvex functions. Basically, the GNC algorithm can be seen as a deterministic relaxation technique (Blake 1989) which 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 which 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 really 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:

(a) the functions $\ell_r(\boldsymbol{\theta})$ are C^1-continuous in $\boldsymbol{\theta}$ and continuous in r;

(b) the concavity of ℓ_r is relaxed monotonously when r decreases;

(c) there exists $r_0 > 0$ such that ℓ_r is strictly convex for any $r \leq r_0$;

(d) $\lim_{r\to 1} \ell_r(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta})$.

Thus, the function ℓ_r has a unique minimum for $r \leq r_0$. When r increases to one, the local minima progressively approaches 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}. Afterwards, 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 §2.3 is closely dependent 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 has been given recently by Nikolova *et al.* (1999). Below 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

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

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 non-negative. They usually reach their maximum concavity at some finite point. More precisely, let

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

and 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.* 1999):

(a) the functions $p_r(|t|)$ are C^1-continuous in t and for any t fixed they are continuous in r;

(b) $p_r(|t|)$ should not stray too much from $p(|t|)$ for each r and $\lim_{r\to 1} p_r(|t|) = p(|t|)$;

(c) the maximum concavity of $p_r(|t|)$, occurring at T_r, is required to increase continuously and strictly monotonously towards 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 towards T as r increases towards 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 in Blake and Zisserman (1987) for the relaxation of a clipped quadratic penalty.

In order to ensure the convexity of initial approximation

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

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

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

where $P_r(\boldsymbol{\theta}) = \sum_{i\geq i_0} p_r(|\theta_i|)$ and $P_r''(\boldsymbol{\theta})$ is its corresponding Hessian matrix. Since 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 it is done in Nikolova *et al.* (1999).

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. In order 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

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

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

An 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.* (1999). 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

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

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 7(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):

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

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 7(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 7(b).

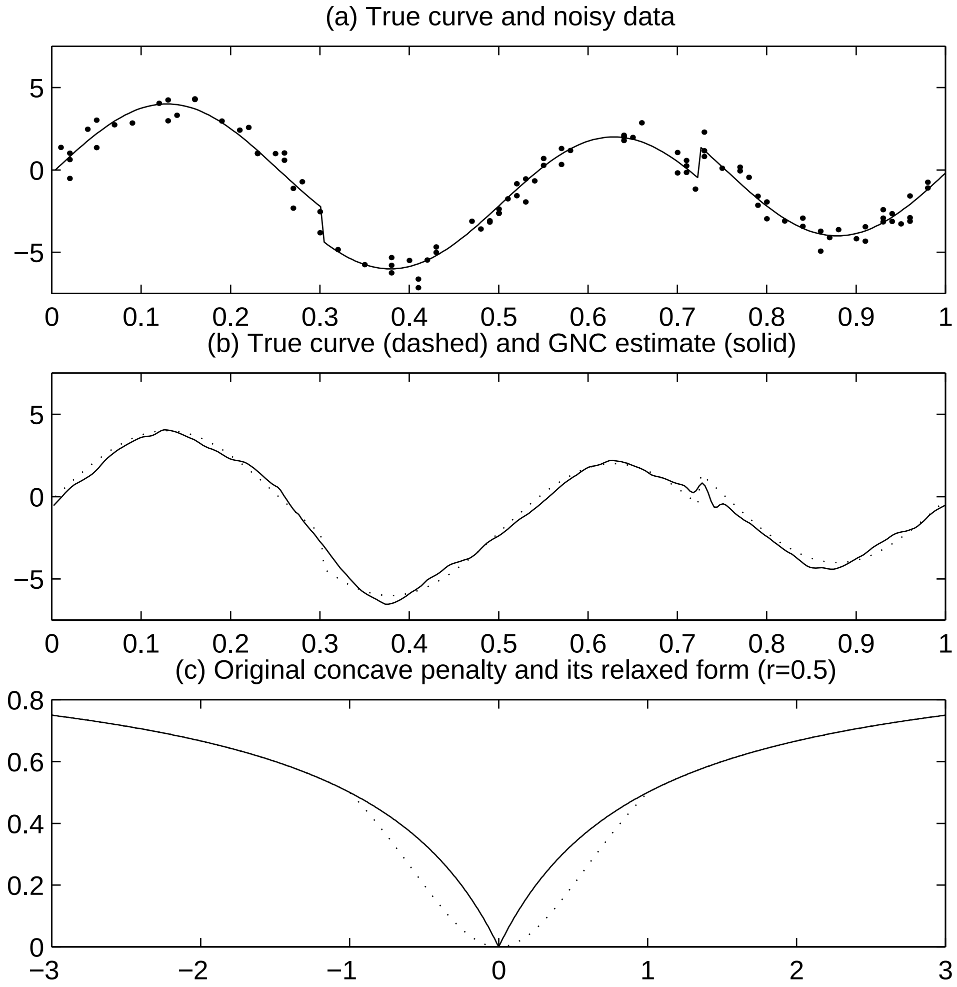


Figure 7: Illustration of the GNC algorithm. (a)The data (points) and the true regression function (solid curve) (b) The unknown function is computed by solving (2.3) using the GNC algorithm; dashed curve – the true function; solid curve – estimated function; (c) Relaxing the concave penalty (2.11) with $b = 3.7$ (solid curve) by using a relaxing function p_r with $r = 0.5$ (dashed curve) defined by equation (6.3).

6.3 Wavelet networks

Another possible approach to handle irregular designs is to interpolate the original data by a wavelet network interpolator as the one described in Bernard *et al.* (1998). The algorithm can be divided into two main parts. The first part of the algorithm is designed to choose a subfamily from a triadic wavelet basis of same cardinality as the number of observations. This step depends only on the design points and allows us to approximate the data by an interpolation of the form $\mathbf{A}\boldsymbol{\theta}$ where this time $\mathbf{A}$ is an $n \times n$ invertible square sparse matrix, while $\boldsymbol{\theta}$ is the vector of wavelet coefficients that are retained through the interpolation. The second part of the algorithm consists in regularizing the estimation of $\boldsymbol{\theta}$ with our penalization method and is closely related to the method derived by Fan and Li (1999) for penalized model selection.

References

Antoniadis, A. (1996), “Smoothing noisy data with tapered coiflets series”, *Scand. J. Statist.*, **23**, 313–330.

Antoniadis, A. (1997), “Wavelets in Statistics: A Review” (with discussion), *Italian Jour. Statist.*, to appear.

Antoniadis, A., Grégoire, G. and McKeague, I. (1994), “Wavelet methods for curve estimation”, *J. Am. Statist. Ass.*, **89**, 1340–1353.

Antoniadis, A., Grégoire, G. and Vial, P. (1997), “Random design wavelet curve smoothing”, *Statistics & Probability Letters*, **35**, pp. 225–232.

Barron, A., Birgé, L. and Massart, P. (1999), “Risk bounds for model selection via penalization”, *Probab. Theory Related Fields*, **113**, 301–413.

Brinkman, N. D. (1981), “Ethanol–A single-cylinder engine study of efficiency and exhaust emissions”, *SAE Transactions*, **90**, 1410–1424.

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”, *J. Amer. Statist. Assoc.*, **70**, 428–433.

Blake A. (1989), “Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **11**, 2–12.

Blake, A. and A. Zisserman (1987), *Visual reconstruction*, MIT Press, Cambridge.

Breiman, L. (1995), "Better subset regression using the nonnegative garotte", *Technometrics*, **37**, 373–384.

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

Chambers, J.M. and Hastie, T.J. (1992), *Statistical models in S*, Wadsworth and Brooks, Pacific Grove.

Chen, S. C, Donoho, D. L and Sanders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**, 1, 33–61.

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", *The 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)", *J. Roy. Statist. Soc. Ser. B*, **54**, 41-81.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). " Wavelet shrinkage: asymptopia? (With discussion.)", *J. Roy. Statist. Soc. Ser. B*, **57**, 301–369.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995), "Density estimation by wavelet thresholding", *Annals of Statistics*, **24**, 508-539.

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 Jour. Statist.*, to appear.

Fan, J. and Chen, J. (2000), "One-step local quasi-likelihood estimation", *J. Roy. Statist. Soc. Ser. B*, to appear.

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", *Ann. Statist.*, **23**, 905–928.

Hall, P. and Turlach, B.A. (1997), "Interpolation methods for nonlinear wavelet regression with irregularly spaced design", *The Annals of Statistics*, **25**, 1912–1925.

Johnstone, I. M. and Silverman, B. W. (1997), "Wavelet threshold estimators for data with correlated noise", *J. Roy. Statist. Soc. Ser. B*, **59**, 319–351.

Kovac, A. and Silverman, B. W. (1998), "Extending the scope of wavelet regression methods by coefficient-dependent thresholding", Technical report, Department of Statistics, University of Bristol.

Neumann, M.H. and Spokoiny, V.G. (1995), "On the efficiency of wavelet estimators under arbitrary error distributions", *Math. Methods Statist.*, **4**, 2, 137–166.

Nikolova, M. (1999a). "Markovian reconstruction using a GNC approach", *IEEE Transactions on Image Processing*, **8**, 9, 1204–1220.

Nikolova (1999b). "Local strong homogeneity of a regularized estimator", To appear in SIAM.

Nikolova, M., Idier, J. and Mohammad-Djafari, A. (1999), "Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF", to appear in *IEEE Image Processing.*

Ogden, T. (1997). *Essential wavelets for statistical applications and data analysis.* Birkhauser Boston, Boston.

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

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. (1998) "Wavelet shrinkage for unequally spaced data", Mathsoft research report 41, to appear in JCGS.

Silverman, B. W. (1985), "Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion)", *J. Roy. Statist. Soc. Ser. B*, **47**, 1–52.

Tibshirani, R. (1995), "Regression shrinkage and selection via the lasso", *J. Roy. Statist. Soc. Ser. B*, **57**, 267–288.

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

Wang, Y. (1996), "Function estimation via wavelet shrinkage for long-memory data", *The Annals of Statistics*, **24**, 466–484.

Appendix: Proofs

A.1 Proof of Theorem 1

The existence of the solution has already been 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

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

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

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

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 above 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 together with (A.1) prove (iii). 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 zero. The continuity at other location 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.

A.2 Proof of Theorem 2

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

$$E(\hat{\theta} - \theta)^2 \leq E(Z - \theta)^2 I(\hat{\theta} \not\in [0, \theta]) + \theta^2 P(\hat{\theta} \in [0, \theta]) \leq 1 + \theta^2. \tag{A.2}$$

To prove (ii), 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 (iii), 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(iii),

$$|\hat{\theta} - Z| \le p_0.$$

Thus,

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

This establishes (ii).

The result in (iii) follows directly from the fact that

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

To show (iv), using the fact that $R_p'(0, p_0) = 0$ due to symmetry, we have by the Taylor expansion that

$$R_p(\theta, p_0) \le R_p(0, p_0) + \frac{1}{2} \sup_{0 \le \eta \le 1} R_p''(\eta, p_0)\theta^2, \qquad \text{for } \theta \in [-1, 1]. \tag{A.3}$$

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

$$R_p'(\theta, p_0) = \int_{-\infty}^{\infty} (\theta + z - 2\hat{\theta})\phi(z - \theta)dz = 2\theta - 2\int_{-\infty}^{\infty} \hat{\theta}\phi(z - \theta)dz.$$

and

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

Using the same arguments as those in the proof of (ii), we have for $\theta > 0$

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

Noticing 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 (A.3), the result (iv) follows for $\theta \in [-1, 1]$. For θ outside this interval, (iv) follows from (A.2). This completes the proof.

A.3 Proof of Lemma 1

For $|\theta| > 1$, by (3.4), we have for $n \geq 4$

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

Thus, we need to show that the inequality holds for $\theta \in [0, 1]$. First of all, by Theorem 2 (iv),

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

Let

$$g(\theta) = (R_p(0, p_0) + 2\theta^2)/(c/n + \theta^2).$$

If $R_p(0, p_0) \leq 2c/n$, then

$$g(\theta) \leq 2 \leq 2\log n.$$

Hence the result holds. When $R_p(0, p_0) > 2c/n$, $g(\theta)$ is monotonically decreasing and hence

$$g(\theta) \leq g(0) = c^{-1} n R_p(0, p_0).$$

By Theorem 2 (iii), we have

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

Using the fact that for $p_0 > 1.14$,

$$\pi^{-1/2}(1 + p_0^{-2}) \leq 1.$$

we conclude that

$$g(\theta) \leq 2c^{-1} d^{1/2} (\log n) + 2c^{-1} (\log n)^{1/2}.$$

A.4 Proof of Theorem 4

Write $\mathbf{Z} = (Z_{j,k})$ and $\boldsymbol{\theta} = (\theta_{j,k})$, $j = 0, \cdots, J-1$, $k = 1, \cdots, 2^j$, where $Z_{j,k}$ and $\theta_{j,k}$ are the wavelet coefficients at the jth resolution level. Then, $Z_{j,k} \sim N(\theta_{j,k}, n^{-1})$. 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

$$E(\hat{\theta}^o_{j,k} - \theta_{j,k})^2 = \min(\theta^2_{j,k}, n^{-1}) = n^{-1}\{\min(\sqrt{n}|\theta_{j,k}|, 1)\}^2. \tag{A.4}$$

Choose an integer J_0 such that $2^{J_0} = n^{1/(2r+1)}$. Then, it follows from (A.4) that

$$\sum_{j=0}^{J_0}\sum_k E(\hat{\theta}^o_{j,k} - \theta_{j,k})^2 \leq 2^{J_0+1}/n = O(n^{-2r/(2r+1)}). \tag{A.5}$$

For $p \leq 2$, by (A.4), we have

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

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

$$C^{p/q}n^{-1+p/2}\sum_{j=J_0+1}^{J-1} 2^{-jap} = O(n^{-1+p/2}2^{-J_0ap}) = O(n^{-2r/(2r+1)}), \tag{A.6}$$

where $a = r + 1/2 - 1/p$. Combination of (A.5) and (A.6) 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 $\|\boldsymbol{\theta}_{j\cdot}\|_2 \leq 2^{(1/2-1/p)j}\|\boldsymbol{\theta}_{j\cdot}\|_p$, since $\boldsymbol{\theta}_{j\cdot}$ has 2^j elements. It follows from this $B^r_{p,q} \subset B^r_{2,q}$. The conclusion follows from the result for the case $p = 2$.

A.4 Proof of Theorem 6

As one naturally expects, rigorous proof of this theorem involves a lot of technicalities, such as approximating discrete summations by their continuous integrations for wavelet coefficients below 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 certain level. For above reasons, we will only outline the key ideas of the proof without taking care of non-intrinsic parts of technicalities. Hence, the key ideas and the intrinsic parts of the proofs are highlighted.

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

$$R_p(f) \leq [2\log n + 2(\log n)^{1/2}][N/n^2 + \sum_{j=1}^{J-1}\sum_{k=1}^{2^j}\min(\theta^{*2}_{j,k}, n^{-1}). \tag{A.7}$$

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,

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

Since 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 (A.7) , we have

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

The result follows from the proof of Theorem 4.