

Block Thresholding Wavelet Regression Using SCAD Penalty

Cheolwoo Park
Department of Statistics
University of Georgia
GA 30602, USA

Abstract

This paper concerns wavelet regression using a block thresholding procedure. Block thresholding methods utilize neighboring wavelet coefficients information to increase estimation accuracy. We propose to construct a data-driven block thresholding procedure using the Smoothly Clipped Absolute Deviation (SCAD) penalty. A simulation study demonstrates competitive finite sample performance of the proposed estimator compared to existing methods. We also show that the proposed estimator achieves optimal convergence rates in Besov spaces.

Keywords and Phrases: Besov space; Block thresholding; Convergence rates; Smoothly clipped absolute deviation penalty; Wavelet regression.

1 Introduction

Wavelet methods have shown advantages in nonparametric function estimation in terms of local adaptivity, computational efficiency, and asymptotic optimality. Compared to the traditional linear procedures such as kernel smoothing, wavelet methods achieve (near) optimal convergence rates in Besov spaces. For example, as shown by Donoho and Johnstone (1998), wavelet methods can outperform optimal linear methods, even at the level of convergence rate, in certain Besov spaces.

Traditional wavelet estimators achieve adaptivity through term-by-term thresholding (Donoho and Johnstone, 1994a) of the empirical wavelet coefficients. In term-by-term thresholding procedures each individual empirical wavelet coefficient is compared with a predetermined threshold. Hall et al (1999); Cai (1999); Cai and Silverman (2001) consider

block thresholding which thresholds empirical wavelet coefficients in groups rather than individually. The goal is to improve estimation precision using adjacent wavelet coefficients. Cai and Zhou (2009) propose a data-driven block James-Stein type thresholding procedure, which empirically chooses the block size and threshold level from the data. The estimator enjoys the improved convergence rates over a wide collection of Besov bodies.

In this paper we propose to construct a data-driven block thresholding procedure using the Smoothly Clipped Absolute Deviation (SCAD) penalty. The SCAD penalty is proposed by Antoniadis and Fan (2001), and it has shown superior performances in many statistical contexts as mentioned in Section 2. We show that the proposed estimator possesses the similar theoretical properties as shown in Cai and Zhou (2009) and demonstrate its better finite sample performance in a simulation study.

This paper is organized as follows. Section 2 reviews Besov space, block thresholding wavelet regression, and the SCAD penalty. In Section 3, we introduce the proposed estimator and its asymptotic properties. A simulation study comparing three block thresholding estimators is conducted in Section 4. Proofs are deferred to Section 5.

2 Background

Given the data (t_i, Y_i) for $i = 1, 2, \dots, n$, consider the nonparametric regression model:

$$Y_i = f(t_i) + \sigma\epsilon_i, \quad (1)$$

where σ is the noise level and ϵ_i 's are independent standard normal errors. The objective of this paper is to construct a wavelet-based estimator for the unknown regression function f , an element of Besov spaces $B_{p,q}^s(M)$, which will be defined soon. Since several methods have been proposed to handle irregularly spaced design in wavelet regression (Antoniadis and Pham, 1998; Pensky and Vidakovic, 2001; Chicken, 2003; Nunes et al, 2006; Park and Kim, 2006), we further assume that $t_i = i/n$ and $n = 2^J$ for some positive integer J for simplicity.

Let $\{\phi, \psi\}$ be a pair of compactly supported scaling and wavelet functions, and $\phi_{j,k}(x) = 2^{j/2}\phi(2^jx - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^jx - k)$ be their dilated and translated versions. The collection of dilation and translation of ϕ and ψ , $\{\phi_{j_0,k}, k = 1, \dots, 2^{j_0}; \psi_{j,k}, j \geq j_0 \geq 0, k = 1, \dots, 2^j\}$, generates an orthonormal basis of $L_2[0, 1]$ space (see Daubechies (1992) for more details). Assume that $f \in L_2[0, 1]$ can be expressed as a wavelet series:

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

where $\alpha_{j_0,k} = \langle f, \phi_{j_0,k} \rangle$ are the coefficients of the scaling functions at the coarsest level which capture the global structure of f , and $\theta_{j,k} = \langle f, \psi_{j,k} \rangle$ are the wavelet coefficients which capture the detailed structure of f .

Besov spaces are a rich class of function spaces. They include many traditional smoothness spaces such as Hölder and Sobolev spaces, as well as function classes of significant spatial inhomogeneity such as the Bump Algebra and the Bounded Variation Classes. The r th difference $\Delta_h^{(r)} f$ is defined as

$$\Delta_h^{(r)} f(t) = \sum_{k=0}^r \binom{r}{k} (-1)^k f(t + kh),$$

and the r th modulus of smoothness of f in $L_p[0, 1]$ as

$$W_{r,p}(f; h) = \|\Delta_h^{(r)} f\|_{L_p[0,1-rh]}.$$

The Besov seminorm of index (s, p, q) is defined for $r > s$ by

$$|f|_{B_{p,q}^s} = \left(\int_0^1 \left(\frac{W_{r,p}(f; h)}{h^s} \right)^q \frac{dh}{h} \right)^{1/q},$$

if $q < \infty$ and by

$$|f|_{B_{p,\infty}^s} = \sup_{0 < h < 1} \frac{W_{r,p}(f; h)}{h^s},$$

if $q = \infty$. The Besov Space $B_{p,q}^s$ is the set of all functions $f : [0, 1] \rightarrow R$ with $f \in L_p([0, 1])$ and $|f|_{B_{p,q}^s} < \infty$. (See DeVore and Popov (1988) for more details.)

For a given r -regular wavelet ψ , i.e. it has compact support and r continuous derivatives, with $r > s$, define the sequence seminorm of the wavelet coefficients θ of a function f by

$$\|\theta\|_{b_{p,q}^s} = \left(\sum_{j=j_0}^{\infty} \left(2^{j(s+1/2-1/p)} \left(\sum_k |\theta_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q},$$

where $\theta_{j,k}$ are wavelet coefficients. When p (or q) = ∞ , l_p (resp. l_q) norms are replaced by l_∞ ; for example,

$$\|\theta\|_{b_{\infty,\infty}^s} = \sup_j 2^{j(s+1/2-1/p)} \sup_k |\theta_{j,k}|.$$

The wavelet basis characterizes smoothness of Besov spaces. It is an important fact that the Besov function norm is equivalent to the sequence norm of the wavelet coefficients of f . Donoho and Johnstone (1998) show an equivalence result on the white noise model and the nonparametric regression over the Besov classes $B_{p,q}^s(M)$, which is defined to be the set

of all functions whose Besov norm is less than M . Define the minimum risk of estimating $\boldsymbol{\theta}$ over the Besov body $B_{p,q}^s(M)$ as

$$R^*(B_{p,q}^s(M)) = \inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in B_{p,q}^s(M)} E \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2. \quad (2)$$

Donoho and Johnstone (1998) show that $R^*(B_{p,q}^s(M))$ converges to 0 at the rate of $n^{-2s/(1+2s)}$ as $n \rightarrow \infty$.

Donoho and Johnstone (1994a) introduce nonlinear wavelet thresholding estimators in nonparametric regression, which typically considers term-by-term assessment of the empirical wavelet coefficients. If an estimate of a coefficient is sufficiently large in absolute value, then the corresponding term in the empirical wavelet expansion is retained or shrunk toward zero by an amount equal to the threshold; otherwise it is omitted. It has been shown that the convergence rate of term-by-term wavelet thresholding estimators are asymptotically near optimal over a class of Besov spaces.

Block thresholding methods group adjacent empirical coefficients and further increase estimation precision. The idea is to use more information from the data than a term-by-term threshold rule for estimating the average empirical wavelet coefficient within a block, and making a decision about retaining or discarding it. This would allow threshold decisions to be made more accurately and improve rates of convergence. Hall et al (1997), Hall et al (1999), Cai (1999), Cai and Silverman (2001), Chicken (2003), Park and Kim (2004, 2006), and Cai and Zhou (2009) propose wavelet block thresholding estimators. Cai (1999) shows that the Blockwise James-Stein type estimator attains the exact optimal rate of convergence over the Besov classes with $p > 2$ and advantage over the traditional nonlinear methods with $1 \leq p < 2$. In addition, a block size of order $\log n$ is used since it leads to an estimator which is both globally and locally adaptive. Recently Cai and Zhou (2009) improve the procedure by proposing a data-driven block thresholding approach. The estimator empirically chooses the block size and threshold level at each resolution level by minimizing Stein's unbiased risk estimate. The estimator is sharp-adaptive over a collection of Besov bodies and achieves simultaneously within a small constant factor of the minimax risk over a wide collection of Besov bodies including both $p \geq 2$ (dense) and $p < 2$ (sparse) cases. In Section 3, we will show that the proposed estimator possesses the same properties.

Antoniadis and Fan (2001) introduce nonlinear regularized wavelet estimators with various penalties. The hard and soft thresholding estimators of Donoho and Johnstone (1994a) are specific members of nonlinear regularized wavelet estimators. They apply the SCAD penalty to the wavelet regression and show the oracle inequalities. The SCAD penalty is

given as

$$p_\lambda(|w|) = \lambda \begin{cases} |w| & \text{if } |w| \leq \lambda, \\ -\frac{(w^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)\lambda} & \text{if } \lambda < |w| \leq a\lambda, \\ \frac{1}{2}(a+1)\lambda & \text{if } |w| > a\lambda. \end{cases} \quad (3)$$

The SCAD penalty possesses the following three important properties; (i) unbiasedness for a large true coefficient to avoid excessive estimation bias; (ii) sparsity (estimating a small coefficient as zero) to reduce model complexity; and (iii) continuity to avoid unnecessary variation in model prediction. The SCAD penalty has proved to be successful in many other statistical contexts including penalized regression (Fan and Li, 2001), classification (Zhang et al, 2006), Cox model (Fan and Li, 2002), and varying coefficient models (Wang et al, 2007).

3 Proposed estimator and its asymptotic properties

In Section 3.1 we propose a data-driven block thresholding wavelet estimator using the SCAD penalty, so called *SURE-Block-SCAD estimator*. Section 3.2 presents the asymptotic properties of the proposed estimator.

3.1 SURE-Block-SCAD estimator

The nonparametric regression problem in (1) can be converted to a problem of estimating the wavelet coefficients at each resolution level, which is equivalent to a problem of estimating the mean of a multivariate normal variable. Suppose that, at each resolution level j , we observe

$$x_i = \theta_i + \tau z_i, \quad i = 1, 2, \dots, d \quad (4)$$

where τ is the noise level, z_i 's are i.i.d standard normal random variables, and $d = 2^j$. Without loss of generality we assume $\tau = 1$. The goal is to estimate the mean vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^T$ based on the observations $\mathbf{x} = (x_1, \dots, x_d)^T$. Let L be the length of each block and $m = d/L$ be the number of blocks. For simplicity we assume that d is divisible by L . We estimate the mean $\boldsymbol{\theta}$ by solving the following penalized regression problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^d (x_i - \theta_i)^2 + \sum_{b=1}^m p_\lambda(\|\boldsymbol{\theta}_b\|_2)$$

where $\boldsymbol{\theta}_b = (\theta_{(b-1)L+1}, \dots, \theta_{bL})$, $\|\boldsymbol{\theta}_b\|_2 = \left(\sum_{l=1}^L \theta_{(b-1)L+l}^2\right)^{1/2}$, and p_λ is given in (3). Then, the solution is given by

$$\boldsymbol{\theta}_b(\lambda, L) = \begin{cases} \left(1 - \frac{\lambda}{\|\mathbf{x}_b\|_2}\right)_+ \mathbf{x}_b & \text{if } \|\mathbf{x}_b\|_2 \leq 2\lambda, \\ \frac{a-1-a\lambda/\|\mathbf{x}_b\|_2}{a-2} \mathbf{x}_b & \text{if } 2\lambda < \|\mathbf{x}_b\|_2 \leq a\lambda, \\ \mathbf{x}_b & \text{if } \|\mathbf{x}_b\|_2 > a\lambda, \end{cases} \quad (5)$$

where $\mathbf{x}_b = (x_{(b-1)L+1}, \dots, x_{bL})$. Fan and Li (2001) suggest $a = 3.7$ based on a Bayesian argument and we use this value in this paper. Note that the solution omits small coefficients, shrinks middle-sized coefficients, and retains large coefficients, which is a key of success of the SCAD penalty in many other statistical contexts. The solution (5) is defined through the norm $\|\mathbf{x}_b\|_2$, but it is more natural to use the square of the norm in asymptotic calculations because it is associated with the bias and variance expansion which might be related to thresholding coefficients. Therefore, we modify the solution (5) by replacing $\|\mathbf{x}_b\|_2$ with $S_b^2 = \|\mathbf{x}_b\|_2^2$, which leads to

$$\boldsymbol{\theta}_b(\lambda, L) = \begin{cases} \left(1 - \frac{\lambda}{S_b^2}\right)_+ \mathbf{x}_b & \text{if } S_b^2 \leq 2\lambda, \\ \frac{a-1-a\lambda/S_b^2}{a-2} \mathbf{x}_b & \text{if } 2\lambda < S_b^2 \leq a\lambda, \\ \mathbf{x}_b & \text{if } S_b^2 > a\lambda. \end{cases} \quad (6)$$

Our limited simulation study shows that the modified solution (6) produces smaller mean squared errors (MSE) compared to (5).

We select the block size L and threshold level λ by empirically minimizing Stein's Unbiased Risk Estimate (SURE). Denote $\hat{\boldsymbol{\theta}}_b(\lambda, L) = \mathbf{x}_b + g(\mathbf{x}_b)$ where g is a weakly differentiable function. Stein (1981) shows that

$$E_{\boldsymbol{\theta}_b} \|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2 = E_{\boldsymbol{\theta}_b} [L + \|g\|_2^2 + 2 \nabla \cdot g].$$

It can be shown that $E_{\boldsymbol{\theta}_b} \|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2 = E_{\boldsymbol{\theta}_b} (SURESC(\mathbf{x}_b, \lambda, L))$ where

$$\begin{aligned} SURESC(\mathbf{x}_b, \lambda, L) &= L + (S_b^2 - 2L)I(S_b^2 \leq \lambda) + \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I(\lambda < S_b^2 \leq 2\lambda) \\ &+ \left(\left(\frac{S_b^2 - a\lambda}{a-2} \right)^2 \frac{1}{S_b^2} + \frac{2L}{a-2} - \frac{2a\lambda(L-2)}{S_b^2(a-2)} \right) I(2\lambda < S_b^2 \leq a\lambda). \end{aligned}$$

This implies that the total risk $E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}(\lambda, L) - \boldsymbol{\theta}\|_2^2 = E_{\boldsymbol{\theta}} (SURESC(\mathbf{x}, \lambda, L))$ where

$$SURESC(\mathbf{x}, \lambda, L) = \sum_{b=1}^m SURESC(\mathbf{x}_b, \lambda, L)$$

is an unbiased risk estimate.

In cases of extreme sparsity of the wavelet coefficients the proposed estimator might have a serious drawback as discussed in Donoho and Johnstone (1995). We follow the hybrid scheme suggested by Cai and Zhou (2009). Therefore, the final wavelet coefficients estimator is constructed as follows. Let $T_d = d^{-1} \sum_{i=1}^d (x_i^2 - 1)$, $\gamma_d = d^{-1/2} (\log_2 d)^{3/2}$ and $\lambda_F = 2L \log d$. Denote the minimizers of *SURESC* with a restricted search range by (λ^*, L^*) , that is,

$$(\lambda^*, L^*) = \arg \min_{\max(L-2, 0) \leq \lambda \leq \lambda_F, 1 \leq L \leq \sqrt{d}} \text{SURESC}(\mathbf{x}, \lambda, L).$$

Then, we define the SURE-Block-SCAD estimator $\hat{\boldsymbol{\theta}}^*(\mathbf{x})$ of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}}_b^* = \begin{cases} \hat{\boldsymbol{\theta}}_b(\lambda^*, L^*) & \text{if } T_d > \gamma_d, \\ \begin{cases} \left(1 - \frac{2 \log d}{x_i^2}\right)_+ x_i & \text{if } x_i^2 \leq 2 \log d, \\ \frac{a-1-2a \log d/x_i^2}{a-2} x_i & \text{if } 2 \log d < x_i^2 \leq a \log d, \\ x_i & \text{if } x_i^2 > a \log d, \end{cases} & \text{if } T_d \leq \gamma_d. \end{cases} \quad (7)$$

if $T_d \leq \gamma_d$. When $T_d \leq \gamma_d$, the estimator corresponds to $L = 1$ and it is called the nonnegative garrote estimator (Brieman, 1995).

Cai and Zhou (2009) propose the Blockwise James-Stein estimator (we refer to it as *SURE-Block-JS*) defined as

$$\hat{\boldsymbol{\theta}}_b^{JS}(\lambda, L) = \left(1 - \frac{\lambda}{S_b^2}\right)_+ \mathbf{x}_b, \quad b = 1, 2, \dots, m.$$

Then, its SURE risk is given by $\text{SUREJS}(\mathbf{x}, \lambda, L) = \sum_{b=1}^m \text{SUREJS}(\mathbf{x}_b, \lambda, L)$ where

$$\text{SUREJS}(\mathbf{x}_b, \lambda, L) = L + (S_b^2 - 2L)I(S_b^2 \leq \lambda) + \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I(S_b^2 > \lambda),$$

and choose λ and L by

$$(\lambda^{JS}, L^{JS}) = \arg \min_{\max(L-2, 0) \leq \lambda \leq \lambda_F, 1 \leq L \leq \sqrt{d}} \text{SUREJS}(\mathbf{x}, \lambda, L).$$

The SURE-Block-JS estimator $\hat{\boldsymbol{\theta}}^{JS}(\mathbf{x})$ of $\boldsymbol{\theta}$ is given by $\hat{\boldsymbol{\theta}}_b^{JS} = \hat{\boldsymbol{\theta}}_b(\lambda^{JS}, L^{JS})$ if $T_d > \gamma_d$, and

$$\hat{\theta}_i^{JS} = \left(1 - \frac{2 \log d}{x_i^2}\right)_+ x_i$$

if $T_d \leq \gamma_d$. Note that while this estimator is always biased, the SURE-Block-SCAD estimator in (7) is unbiased if $S_b^2 > a\lambda$.

Figure 1 compares *SUREJS* and *SURESC* under the model (4) with $d = 256$. The simulated x_i 's are displayed in Figure 1(a). We consider ten L values between 1 and $\sqrt{256}$

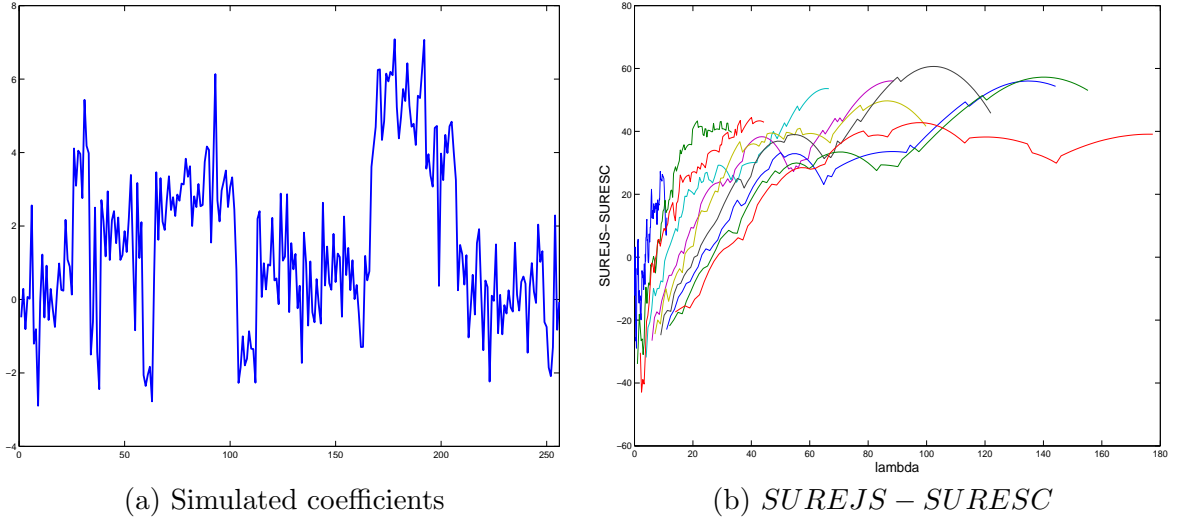


Figure 1: Comparison of SURE risks.

and 100 λ values between $\max(L-2, 0)$ and $2L \log(256)$. In Figure 1(b), the x axis represents λ and the y axis $SUREJS(\mathbf{x}, \lambda, L) - SURESC(\mathbf{x}, \lambda, L)$ for each λ and L . Ten curves correspond to ten different L values. It can be seen that the risk of the proposed estimator is lower than that of the SURE-Block-JS estimator except for small λ values. This can be explained analytically. Let $g(\mathbf{x}_b, \lambda, L) = SUREJS(\mathbf{x}_b, \lambda, L) - SURESC(\mathbf{x}_b, \lambda, L)$. Then,

$$g(\mathbf{x}_b, \lambda, L) = \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I(S_b^2 > a\lambda) + \left\{ \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} - \left(\frac{S_b - a\lambda/S_b}{a-2} \right)^2 - \frac{2L}{a-2} + \frac{2a\lambda(L-2)}{S_b^2(a-2)} \right\} I(2\lambda < S_b^2 \leq a\lambda).$$

It can be shown that $g(\mathbf{x}_b, \lambda, L)$ is negative for $\max(0, L-2) \leq \lambda \leq 2(L-2)$ and positive for $2(L-2) < \lambda \leq S_b^2/a$. When $S_b^2/a < \lambda \leq S_b^2/2$, $g(\mathbf{x}_b, \lambda, L)$ can take either negative or positive values, and $g(\mathbf{x}_b, \lambda, L) = 0$ for $\lambda > S_b^2/2$. Since λ is bounded by $\lambda_F = 2L \log d$, most of λ values fall in the interval $\max(0, L-2) \leq \lambda \leq S_b^2/a$, which explains $g < 0$ for small λ and $g > 0$ for large λ in Figure 1(b). Section 4 compares the finite sample performances of the two estimators.

In what follows we introduce the final form of the wavelet function estimator and its algorithm. In order to obtain the estimator, we first transform the data into the wavelet domain via discrete wavelet transform (DWT): $\tilde{\mathbf{Y}} = Wn^{-1/2}\mathbf{Y}$ where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$,

W is an orthonormal DWT matrix, and

$$\tilde{\mathbf{Y}} = (\tilde{\alpha}_{j_0,1}, \dots, \tilde{\alpha}_{j_0,2^{j_0}}, \dots, \tilde{y}_{j_0,1}, \dots, \tilde{y}_{j_0,2^{j_0}}, \dots, \tilde{y}_{J-1,1}, \dots, \tilde{y}_{J-1,2^{J-1}})^T,$$

are the empirical wavelet coefficients. Since the DWT is an orthogonal transform, $\tilde{y}_{j,k}$ can be expressed as

$$\tilde{y}_{j,k} = \tilde{\theta}_{j,k} + n^{-1/2} \sigma z_{j,k}, \quad j \geq j_0, \quad k = 1, 2, \dots, 2^j \quad (8)$$

where $\tilde{\theta}_{j,k} = E(\tilde{y}_{j,k})$ and $z_{j,k}$'s are independent standard normal random variables. Note that $\tilde{\theta}_{j,k}$'s are the DWT of the sampled function $\{n^{-1/2} f(i/n)\}$ and it approximately equals the true wavelet coefficient $\theta_{j,k}$ of f .

Denote $\tilde{\mathbf{Y}}_j = (\tilde{y}_{j,1}, \dots, \tilde{y}_{j,2^j})^T$ by the empirical wavelet coefficients at resolution level j . Then, the empirical wavelet coefficients using SURE-Block-SCAD at each resolution level j is given by

$$\hat{\boldsymbol{\theta}}_j = n^{-1/2} \sigma \cdot \hat{\boldsymbol{\theta}}^* (n^{1/2} \sigma^{-1} \tilde{\mathbf{Y}}_j)$$

where $\hat{\boldsymbol{\theta}}^*$ is the SURE-Block-SCAD estimator given in (7) and $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_{j,1}, \dots, \hat{\theta}_{j,2^j})^T$. Here, σ can be estimated by the empirical wavelet coefficients at the highest resolution level (see Donoho and Johnstone (1994a)). Then, the estimate of the function f is given by

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

3.2 Asymptotic properties

In this section we present the asymptotic properties of the SURE-Block-SCAD estimator.

Suppose the model (4) with $\tau = 1$. Let $r(\lambda, L) = d^{-1} E \|\hat{\boldsymbol{\theta}}(\lambda, L) - \boldsymbol{\theta}\|_2^2$,

$$R(\boldsymbol{\theta}) = \inf_{\lambda \geq 0, 1 \leq L \leq \sqrt{d}} r(\lambda, L) = \inf_{\lambda \geq \max(0, L-2), 1 \leq L \leq \sqrt{d}} r(\lambda, L), \quad (9)$$

and $R_F(\boldsymbol{\theta}) = r(2 \log d, 1)$. The following theorem gives upper bounds of the risk of the SURE-Block-SCAD estimator $\hat{\boldsymbol{\theta}}^*$ in (7) for individual resolution level, and is used for the proof of Theorem 2.

Theorem 1. *For any constant $\eta > 0$ and some constant $c_\eta > 0$,*

(i)

$$\frac{1}{d} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq R(\boldsymbol{\theta}) + R_F(\boldsymbol{\theta}) I(\mu_d \leq 3\gamma_d) + c_\eta d^{\eta-1/4},$$

uniformly in $\boldsymbol{\theta} \in \mathcal{R}^d$;

(ii)

$$\frac{1}{d} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq R_F + c_\eta d^{-1-\eta},$$

uniformly in $\mu_d \leq \gamma_d/3$.

Suppose that we observe the sequence model in (8). If we set $\hat{\boldsymbol{\theta}}_{j,k} = 0$ for $j \geq J$, the minimax risk among all block SCAD thresholding estimators with all possible block sizes $1 \leq L_j \leq 2^{j/2}$ and threshold levels $\lambda_j \geq 0$ is

$$R_T^*(B_{p,q}^s(M)) = \inf_{\lambda_j \geq 0, 1 \leq L_j \leq 2^{j/2}} \sup_{\boldsymbol{\theta} \in B_{p,q}^s(M)} E \sum_{j=j_0}^{\infty} \|\hat{\boldsymbol{\theta}}_j(\lambda_j, L_j) - \boldsymbol{\theta}_j\|_2^2,$$

where T represents ‘‘block SCAD Thresholding estimators’’. The following theorem shows that the SURE-Block-SCAD estimator adaptively achieves the exact minimax block thresholding risk $R_T^*(B_{p,q}^s(M))$ asymptotically over a wide range of Besov bodies.

Theorem 2. *Suppose that ψ is r -regular. Then,*

$$\sup_{f \in B_{p,q}^s(M)} E_f \|\hat{f}^* - f\|_2^2 \leq R_T^*(B_{p,q}^s(M))(1 + o(1)),$$

or equivalently,

$$\sup_{\boldsymbol{\theta} \in B_{p,q}^s(M)} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq R_T^*(B_{p,q}^s(M))(1 + o(1)),$$

for $1 \leq p, q \leq \infty$, $0 < M < \infty$, and $r \geq s > 4 \left(\frac{1}{p} - \frac{1}{2}\right)_+ + \frac{1}{2}$ with $\frac{2s^2-1/6}{1+2s} > 1/p$.

Theorem 3 shows that the SURE-Block-SCAD estimator achieves nearly optimally adaptive convergence rates over a collection of Besov bodies $B_{p,q}^s(M)$ including both $p \geq 2$ (dense) and $p < 2$ (sparse) cases. The estimator adaptively obtains both the optimal rate and optimal constant with $p = q = 2$. When $p \geq 2$ and $q \geq 2$, the estimator adaptively achieves within a factor of 1.25 of the minimax risk $R^*(B_{p,q}^s(M))$ in (2). When $p < 2$, the maximum risk of the estimator is within a constant factor.

Theorem 3. *Suppose ψ is r -regular.*

(i) *The SURE-Block-SCAD estimator is adaptively sharp minimax over Besov bodies $B_{2,2}^s(M)$ for all $M > 0$ and $r \geq s > 0.88$. That is,*

$$\sup_{\boldsymbol{\theta} \in B_{2,2}^s(M)} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq R^*(B_{2,2}^s(M))(1 + o(1)).$$

(ii) The SURE-Block-SCAD estimator is adaptively, asymptotically within a factor of 1.25 of the minimax risk over Besov bodies $B_{p,q}^s(M)$,

$$\sup_{\boldsymbol{\theta} \in B_{p,q}^s(M)} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq 1.25 R^*(B_{p,q}^s(M))(1 + o(1))$$

for all $p \geq 2$, $q \geq 2$, $M > 0$, and $\frac{2s^2-1/6}{1+2s} > 1/p$ with $r \geq s > 1/2$.

(iii) The SURE-Block-SCAD estimator is asymptotically minimax up to a constant factor $G(p \wedge q)$ over Besov bodies $B_{p,q}^s(M)$ with $1 \leq p, q \leq \infty$, $0 < M < \infty$, and $s > 4\left(\frac{1}{p} - \frac{1}{2}\right)_+ + \frac{1}{2}$ with $\frac{2s^2-1/6}{1+2s} > 1/p$. That is,

$$\sup_{\boldsymbol{\theta} \in B_{p,q}^s(M)} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq G(p \wedge q) R^*(B_{p,q}^s(M))(1 + o(1))$$

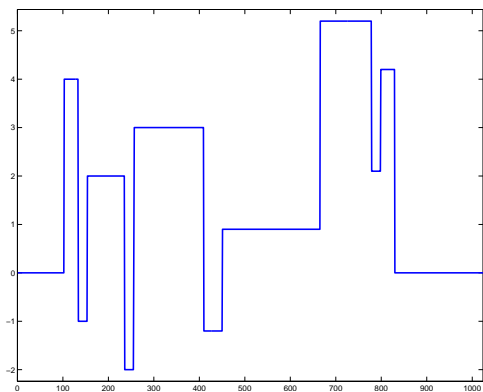
where $G(p \wedge q)$ is a constant depending only on $p \wedge q$.

4 Simulation

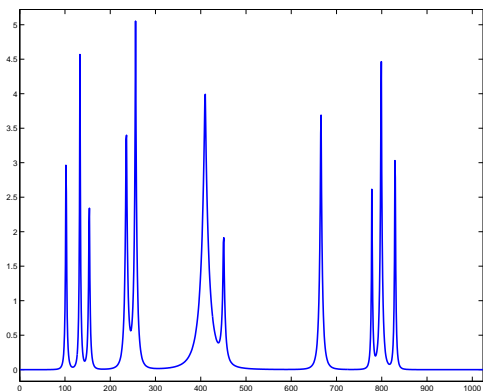
In this section we compare the numerical performance of the proposed method with Sure-Block-JS (Cai and Zhou, 2009) and HardBlock (Cai, 2002). HardBlock is a block thresholding procedure with a hard thresholding rule and shows superior finite sample performances in many cases. We label Sure-Block-SCAD as SCAD, Sure-Block-JS as JS, and HardBlock as HT in this section.

Six test functions, plotted in Figure 2, are used for the comparison of the three wavelet procedures. Sample sizes vary from $n = 256$ to $n = 16384$ and signal-to-noise ratios (SNR) 3 and 7 are considered. We use Daubechies' compactly supported wavelet Symmlet 8. The software package WaveLab for simulations in WaveLab 805 are used (see <http://www-stat.stanford.edu/wavelab/>).

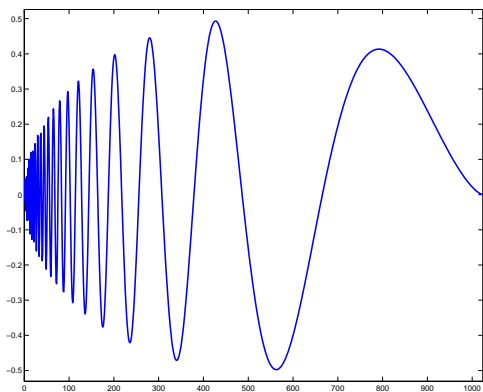
Tables 1 and 3 below report the means and standard errors (in parentheses) of MSE's over 1000 replications for the three block thresholding estimators for SNR 3 and 7, respectively. Tables 2 and 4 report the medians of MSE's and p -values (in parentheses) from Wilcoxon tests for SNR 3 and 7, respectively. We run a Wilcoxon's two-sample test (Wilcoxon, 1945) that compares two medians in order to see the significance of differences between the estimators. The p -values in the SCAD column compare the medians of SCAD and JS, those in JS compare JS and HT, and those in HT compare SCAD and HT, respectively. For example, the first p -value in Table 2, 0.2208, implies that the medians of SCAD and JS for the Blocks example with $n = 256$ are not statistically different.



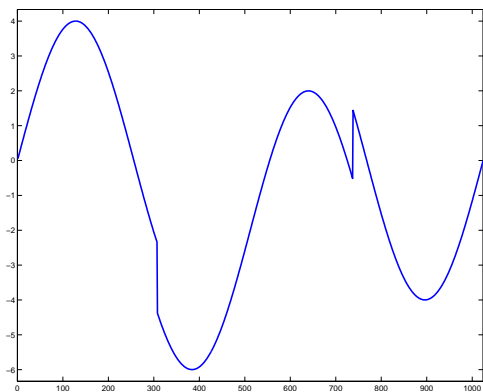
(a) Blocks



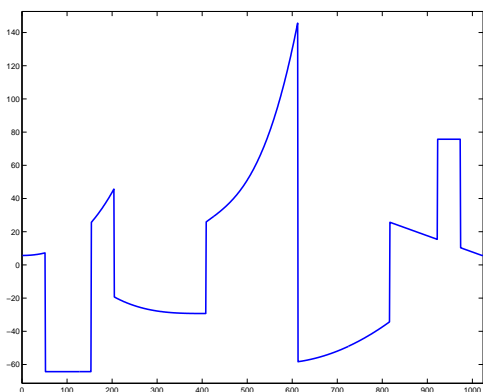
(b) Bumps



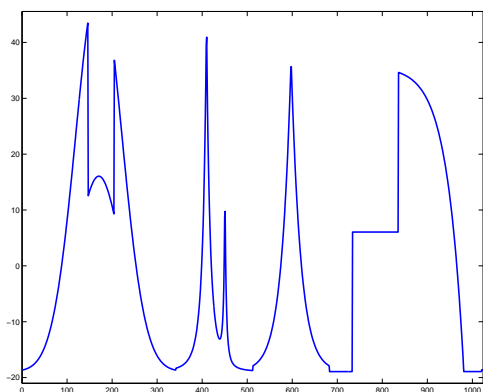
(c) Doppler



(d) HeaviSine



(e) Piece-Polynomial



(f) Piece-Regular

Figure 2: Simulated examples.

Table 1: Summary of MSE (SNR=3)

n	SCAD	JS	HT	SCAD	JS	HT
Blocks	Mean (S.E.)			Bumps		
256	0.4016 (0.0023)	0.4058 (0.0023)	0.4111 (0.0013)	0.0993 (0.0005)	0.0961 (0.0005)	0.0431 (0.0001)
512	0.3137 (0.0018)	0.3200 (0.0020)	0.3902 (0.0008)	0.0810 (0.0005)	0.0805 (0.0005)	0.0441 (0.0001)
1024	0.3029 (0.0020)	0.3132 (0.0021)	0.3749 (0.0006)	0.0659 (0.0005)	0.0684 (0.0005)	0.0441 (0.0001)
2048	0.2368 (0.0025)	0.2657 (0.0027)	0.3533 (0.0004)	0.0502 (0.0006)	0.0521 (0.0006)	0.0424 (0.0001)
4096	0.2610 (0.0021)	0.2732 (0.0021)	0.3531 (0.0003)	0.0538 (0.0007)	0.0579 (0.0007)	0.0426 (0.0000)
8192	0.1825 (0.0023)	0.2048 (0.0026)	0.3546 (0.0002)	0.0433 (0.0006)	0.0478 (0.0006)	0.0427 (0.0000)
16384	0.1444 (0.0024)	0.1628 (0.0027)	0.3599 (0.0001)	0.0338 (0.0005)	0.0367 (0.0006)	0.0432 (0.0000)
n	SCAD	JS	HT	SCAD	JS	HT
Doppler	HeaviSine			HeaviSine		
256	0.0056 (0.0000)	0.0058 (0.0000)	0.0077 (0.0000)	0.1152 (0.0006)	0.1121 (0.0006)	0.7483 (0.0023)
512	0.0044 (0.0000)	0.0045 (0.0000)	0.0075 (0.0000)	0.0773 (0.0004)	0.0759 (0.0004)	0.7706 (0.0017)
1024	0.0045 (0.0001)	0.0053 (0.0001)	0.0077 (0.0000)	0.0519 (0.0003)	0.0509 (0.0003)	0.8024 (0.0012)
2048	0.0039 (0.0001)	0.0039 (0.0001)	0.0077 (0.0000)	0.0339 (0.0002)	0.0339 (0.0002)	0.8111 (0.0009)
4096	0.0033 (0.0001)	0.0035 (0.0001)	0.0078 (0.0000)	0.0251 (0.0001)	0.0244 (0.0001)	0.8279 (0.0006)
8192	0.0034 (0.0001)	0.0035 (0.0001)	0.0080 (0.0000)	0.0185 (0.0001)	0.0175 (0.0001)	0.8433 (0.0005)
16384	0.0031 (0.0001)	0.0034 (0.0001)	0.0081 (0.0000)	0.0122 (0.0001)	0.0122 (0.0001)	0.8566 (0.0003)
n	SCAD	JS	HT	SCAD	JS	HT
Piece-Polynomial	Piece-Regular			Piece-Regular		
256	204.5900 (1.1612)	204.0000 (1.1271)	252.7700 (0.8907)	23.8280 (0.1317)	23.1860 (0.1193)	30.1680 (0.0942)
512	155.7800 (1.1742)	164.2000 (1.8941)	254.4100 (0.5176)	16.1180 (0.0752)	16.3090 (0.1257)	31.9570 (0.0733)
1024	106.7300 (0.4803)	104.9200 (0.5486)	246.1600 (0.3599)	10.1830 (0.0521)	9.9580 (0.0550)	30.6260 (0.0454)
2048	72.7600 (0.2637)	73.6810 (0.5624)	240.9200 (0.2594)	8.7485 (0.0368)	9.5084 (0.1153)	30.2760 (0.0341)
4096	55.3240 (0.2531)	58.9470 (0.7179)	243.7200 (0.1872)	5.5617 (0.0294)	8.6334 (0.1491)	30.4760 (0.0235)
8192	33.3230 (0.1779)	42.0070 (0.7081)	246.2700 (0.1340)	3.8305 (0.0289)	11.0450 (0.1154)	30.8160 (0.0170)
16384	29.0860 (0.1702)	40.7440 (0.7562)	248.3900 (0.0946)	3.0630 (0.0272)	11.5740 (0.0404)	31.2230 (0.0119)

Table 2: Summary of MSE (SNR=3)

n	SCAD	JS	HT	SCAD	JS	HT
Blocks	Median (p -value: SCAD vs JS, JS vs HT, SCAD vs HT)			Bumps		
256	0.3945 (0.2208)	0.4011 (0.0002)	0.4103 (0.0000)	0.0988 (0.0001)	0.0960 (0.0000)	0.0431 (0.0000)
512	0.2999 (0.1430)	0.2986 (0.0000)	0.3903 (0.0000)	0.0792 (0.3606)	0.0784 (0.0000)	0.0440 (0.0000)
1024	0.3020 (0.0032)	0.3068 (0.0000)	0.3742 (0.0000)	0.0641 (0.0001)	0.0668 (0.0000)	0.0440 (0.0000)
2048	0.2306 (0.0000)	0.2621 (0.0000)	0.3537 (0.0000)	0.0463 (0.0520)	0.0478 (0.0000)	0.0424 (0.0000)
4096	0.2452 (0.0075)	0.2488 (0.0000)	0.3529 (0.0000)	0.0507 (0.0000)	0.0542 (0.0000)	0.0427 (0.0000)
8192	0.1863 (0.0000)	0.1952 (0.0000)	0.3547 (0.0000)	0.0420 (0.0000)	0.0471 (0.0000)	0.0428 (0.1365)
16384	0.1525 (0.0000)	0.1592 (0.0000)	0.3599 (0.0000)	0.0302 (0.0001)	0.0330 (0.0000)	0.0432 (0.0000)
n	SCAD	JS	HT	SCAD	JS	HT
Doppler	HeaviSine			HeaviSine		
256	0.0055 (0.0036)	0.0055 (0.0000)	0.0077 (0.0000)	0.1130 (0.0015)	0.1104 (0.0000)	0.7466 (0.0000)
512	0.0041 (0.0035)	0.0042 (0.0000)	0.0075 (0.0000)	0.0766 (0.0117)	0.0752 (0.0000)	0.7686 (0.0000)
1024	0.0034 (0.0000)	0.0040 (0.0000)	0.0076 (0.0000)	0.0513 (0.0196)	0.0503 (0.0000)	0.8022 (0.0000)
2048	0.0036 (0.5173)	0.0036 (0.0000)	0.0077 (0.0000)	0.0333 (0.9067)	0.0334 (0.0000)	0.8122 (0.0000)
4096	0.0029 (0.4573)	0.0031 (0.0000)	0.0078 (0.0000)	0.0245 (0.0008)	0.0239 (0.0000)	0.8274 (0.0000)
8192	0.0031 (0.4188)	0.0031 (0.0000)	0.0080 (0.0000)	0.0183 (0.0000)	0.0172 (0.0000)	0.8435 (0.0000)
16384	0.0027 (0.0051)	0.0031 (0.0000)	0.0081 (0.0000)	0.0118 (0.9667)	0.01117 (0.0000)	0.8565 (0.0000)
n	SCAD	JS	HT	SCAD	JS	HT
Piece-Polynomial	Piece-Regular			Piece-Regular		
256	200.8700 (0.7603)	200.0700 (0.0000)	251.3800 (0.0000)	23.4150 (0.0014)	22.8440 (0.0000)	30.1410 (0.0000)
512	148.6200 (0.1225)	147.0100 (0.0000)	254.2100 (0.0000)	15.9940 (0.0000)	15.5170 (0.0000)	32.0040 (0.0000)
1024	103.9000 (0.0002)	102.3600 (0.0000)	254.8000 (0.0000)	9.9723 (0.0001)	9.7599 (0.0000)	30.5550 (0.0000)
2048	72.1400 (0.0000)	70.2490 (0.0000)	241.1500 (0.0000)	8.6831 (0.0000)	8.3849 (0.0000)	30.2850 (0.0000)
4096	54.4210 (0.0000)	52.1790 (0.0000)	243.6300 (0.0000)	5.4750 (0.0000)	5.8752 (0.0000)	30.4570 (0.0000)
8192	32.2550 (0.0000)	33.2610 (0.0000)	246.2800 (0.0000)	3.7334 (0.0000)	12.4410 (0.0000)	30.8120 (0.0000)
16384	28.1300 (0.0000)	29.8950 (0.0000)	248.3600 (0.0000)	2.9303 (0.0000)	11.5060 (0.0000)	31.2220 (0.0000)

Tables 1 and 2 with SNR 3 show that SCAD stays close to JS for smaller sample sizes, and performs better for larger sample sizes in most cases. Specifically, SCAD shows superior performance for Piece-Polynomial and Piece-Regular. HT produces the smallest MSE's for the Bumps example, but shows inferior performance in many cases, for example Heavisine, Piece-Polynomial, and Piece-Regular, although performing stably with small standard errors. Tables 3 and 4 with SNR 7 also indicate a similar lesson except the improved performance of HT for Blocks, Bumps, and Doppler. But, for other examples, the proposed estimator SCAD shows the best performance. For the comparison of SCAD and JS with SNR 7, SCAD produces lower MSE's in most cases. Therefore, we conclude that the performance of the proposed estimator is satisfactory since it produces small MSE's for large samples, and comparable MSE's for small samples with both low and high SNR's.

Table 3: Summary of MSE (SNR=7)

n	SCAD	JS	HT	SCAD	JS	HT
Blocks		Bumps				
	Mean (S.E.)					
256	0.2559 (0.0025)	0.2460 (0.0025)	0.0779 (0.0003)	0.0378 (0.0003)	0.0391 (0.0004)	0.0094 (0.0000)
512	0.1841 (0.0018)	0.1915 (0.0020)	0.0689 (0.0002)	0.0425 (0.0004)	0.0431 (0.0004)	0.0081 (0.0000)
1024	0.2029 (0.0025)	0.2326 (0.0026)	0.0685 (0.0001)	0.0450 (0.0005)	0.0468 (0.0005)	0.0081 (0.0000)
2048	0.1809 (0.0023)	0.2009 (0.0023)	0.0657 (0.0001)	0.0544 (0.0007)	0.0530 (0.0008)	0.0078 (0.0000)
4096	0.1847 (0.0021)	0.2176 (0.0021)	0.0646 (0.0001)	0.0469 (0.0005)	0.0500 (0.0005)	0.0078 (0.0000)
8192	0.1864 (0.0027)	0.2237 (0.0028)	0.0653 (0.0000)	0.0379 (0.0006)	0.0428 (0.0006)	0.0079 (0.0000)
16384	0.1683 (0.0030)	0.2050 (0.0031)	0.0660 (0.0000)	0.0298 (0.0005)	0.0350 (0.0006)	0.0080 (0.0000)
n	SCAD	JS	HT	SCAD	JS	HT
Doppler		HeaviSine				
256	0.0037 (0.0000)	0.0038 (0.0000)	0.0015 (0.0000)	0.0505 (0.0003)	0.0484 (0.0003)	0.1505 (0.0005)
512	0.0026 (0.0000)	0.0027 (0.0000)	0.0014 (0.0000)	0.0351 (0.0002)	0.0345 (0.0002)	0.1486 (0.0003)
1024	0.0035 (0.0001)	0.0037 (0.0001)	0.0014 (0.0000)	0.0223 (0.0002)	0.0222 (0.0002)	0.1490 (0.0002)
2048	0.0026 (0.0000)	0.0028 (0.0001)	0.0014 (0.0000)	0.0141 (0.0001)	0.0145 (0.0001)	0.1504 (0.0002)
4096	0.0025 (0.0000)	0.0029 (0.0001)	0.0014 (0.0000)	0.0103 (0.0001)	0.0105 (0.0001)	0.1523 (0.0001)
8192	0.0025 (0.0000)	0.0030 (0.0001)	0.0015 (0.0000)	0.0087 (0.0001)	0.0089 (0.0001)	0.1554 (0.0001)
16384	0.0024 (0.0000)	0.0029 (0.0001)	0.0015 (0.0000)	0.0073 (0.0001)	0.0078 (0.0001)	0.1575 (0.0001)
n	SCAD	JS	HT	SCAD	JS	HT
Piece-Polynomial		Piece-Regular				
256	93.5920 (0.7350)	102.1300 (0.9496)	50.0550 (0.1849)	8.9437 (0.0454)	10.2710 (0.0711)	5.9428 (0.0169)
512	68.6690 (0.5836)	81.0160 (1.6833)	47.7330 (0.1163)	5.6982 (0.0351)	9.3241 (0.2215)	5.7726 (0.0142)
1024	45.3440 (0.2684)	48.5270 (0.7482)	46.5030 (0.0734)	3.9517 (0.0240)	3.9363 (0.0434)	5.7123 (0.0089)
2048	29.3480 (0.1712)	38.1590 (0.8833)	46.1560 (0.0530)	3.3522 (0.0185)	4.9898 (0.1351)	5.5367 (0.0060)
4096	22.5710 (0.1683)	33.1900 (0.8427)	45.0380 (0.0351)	2.5698 (0.0230)	11.3090 (0.1114)	5.6481 (0.0044)
8192	21.4680 (0.1726)	42.5120 (0.9474)	45.0940 (0.0247)	2.1977 (0.0301)	11.3880 (0.0347)	5.6836 (0.0031)
16384	18.6880 (0.2056)	57.6930 (0.8288)	45.6340 (0.0173)	1.7638 (0.0298)	10.6070 (0.0319)	5.7442 (0.0022)

5 Proofs

In this section we denote C to represent a generic constant that may vary from place to place.

To prove Theorem 1 we need the following four lemmas and Proposition 1 under the

Table 4: Summary of MSE (SNR=7)

n	SCAD	JS	HT	SCAD	JS	HT
Blocks	Median (p -value: SCAD vs JS, JS vs HT, SCAD vs HT)			Bumps		
256	0.2341 (0.0008)	0.2157 (0.0000)	0.0772 (0.0000)	0.0359 (0.0804)	0.0365 (0.0000)	0.0093 (0.0000)
512	0.1631 (0.1255)	0.1679 (0.0000)	0.0686 (0.0000)	0.0417 (0.7194)	0.0415 (0.0000)	0.0081 (0.0000)
1024	0.1990 (0.0000)	0.2304 (0.0000)	0.0683 (0.0000)	0.0454 (0.0100)	0.0468 (0.0000)	0.0081 (0.0000)
2048	0.1892 (0.0000)	0.1976 (0.0000)	0.0657 (0.0000)	0.0469 (0.1183)	0.0459 (0.0000)	0.0078 (0.0000)
4096	0.1717 (0.0000)	0.1884 (0.0000)	0.0645 (0.0000)	0.0457 (0.0001)	0.0487 (0.0000)	0.0078 (0.0000)
8192	0.1614 (0.0000)	0.1849 (0.0000)	0.0653 (0.0000)	0.0365 (0.0000)	0.0405 (0.0000)	0.0079 (0.0000)
16384	0.1452 (0.0000)	0.2077 (0.0000)	0.0660 (0.0000)	0.0273 (0.0000)	0.0320 (0.0000)	0.0080 (0.0000)
n	SCAD	JS	HT	SCAD	JS	HT
Doppler	HeaviSine			HeaviSine		
256	0.0036 (0.0365)	0.0037 (0.0000)	0.0015 (0.0000)	0.0505 (0.0000)	0.0479 (0.0000)	0.1504 (0.0000)
512	0.0025 (0.0020)	0.0026 (0.0000)	0.0014 (0.0000)	0.0343 (0.0315)	0.0340 (0.0000)	0.1483 (0.0000)
1024	0.0029 (0.2043)	0.0030 (0.0000)	0.0014 (0.0000)	0.0216 (0.9031)	0.0214 (0.0000)	0.1489 (0.0000)
2048	0.0021 (0.0972)	0.0022 (0.0000)	0.0014 (0.0000)	0.0132 (0.0094)	0.0134 (0.0000)	0.1504 (0.0000)
4096	0.0021 (0.0000)	0.0024 (0.0000)	0.0014 (0.0000)	0.0093 (0.0110)	0.0095 (0.0000)	0.1523 (0.0000)
8192	0.0021 (0.0000)	0.0026 (0.0000)	0.0015 (0.0000)	0.0076 (0.0284)	0.0078 (0.0000)	0.1555 (0.0000)
16384	0.0019 (0.0000)	0.0024 (0.0000)	0.0015 (0.0000)	0.0066 (0.0006)	0.0073 (0.0000)	0.1574 (0.0000)
n	SCAD	JS	HT	SCAD	JS	HT
Piece-Polynomial	Piece-Regular			Piece-Regular		
256	87.6650 (0.0000)	93.4090 (0.0000)	49.8390 (0.0000)	8.8843 (0.0000)	10.0810 (0.0000)	5.9413 (0.0000)
512	62.0680 (0.4869)	60.7830 (0.0000)	47.5560 (0.0000)	5.5820 (0.0000)	5.9760 (0.0000)	5.7482 (0.0000)
1024	44.2700 (0.0148)	43.1630 (0.0000)	46.3760 (0.0000)	3.8325 (0.0001)	3.6958 (0.0000)	5.7004 (0.0000)
2048	28.1520 (0.1766)	28.1060 (0.0000)	46.1330 (0.0000)	3.3173 (0.2623)	3.3007 (0.0000)	5.5450 (0.0000)
4096	21.4280 (0.0000)	22.0220 (0.0000)	45.0230 (0.0000)	2.4674 (0.0000)	12.3550 (0.0000)	5.6352 (0.0000)
8192	20.1350 (0.0000)	25.2170 (0.0000)	45.0990 (0.0000)	1.9153 (0.0000)	11.1100 (0.0000)	5.6831 (0.0000)
16384	19.2650 (0.0000)	67.7940 (0.0000)	45.6210 (0.0000)	1.5028 (0.0000)	10.3250 (0.0000)	5.7426 (0.0000)

sequence model (4) with $\tau = 1$. For a given L and λ , set $r_b(\lambda, L) = E_{\boldsymbol{\theta}_b} \|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2$ and define

$$r(\lambda, L) = \frac{1}{d} \sum_{b=1}^m r_b(\lambda, L) = \frac{1}{d} \sum_{b=1}^m E_{\boldsymbol{\theta}_b} \|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2.$$

Let

$$\tilde{R}(\boldsymbol{\theta}) = \inf_{0 \leq \lambda \leq \lambda_F, 1 \leq L \leq \sqrt{d}} r(\lambda, L) = \inf_{\max(L-2, 0) \leq \lambda \leq \lambda_F, 1 \leq L \leq \sqrt{d}} r(\lambda, L). \quad (10)$$

The difference between $\tilde{R}(\boldsymbol{\theta})$ in (10) and $R(\boldsymbol{\theta})$ in (9) is that the range of λ in $\tilde{R}(\boldsymbol{\theta})$ is restricted to be less than equal to λ_F . The result of Lemma 1 implies that the effect of this restriction is negligible for any block size L .

Lemma 1. *For any fixed $\eta > 0$, there exists a constant $C_\eta > 0$ such that for all $\boldsymbol{\theta} \in \mathcal{R}^d$,*

$$\tilde{R}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) \leq C_\eta d^{\eta-1/2}.$$

Proof. Suppose that $R(\boldsymbol{\theta}) = r(\lambda_1, L_1)$. If $\lambda_1 \leq \lambda_F$, $\tilde{R}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) = 0$.

For the proof of the case when $\lambda_1 > \lambda_F$, we introduce the risk of the Sure-Block-JS estimator (Cai and Zhou, 2009),

$$R^{JS}(\boldsymbol{\theta}) = \inf_{\lambda \geq \max(L-2, 0), 1 \leq L \leq \sqrt{d}} r^{JS}(\lambda, L) = r^{JS}(\lambda_2, L_2),$$

where $r^{JS}(\lambda, L) = d^{-1} \sum_{b=1}^m r_b^{JS}(\lambda, L)$, and $r_b^{JS}(\lambda, L) = d^{-1} \sum_{b=1}^m E_{\boldsymbol{\theta}_b} \|\hat{\boldsymbol{\theta}}_b^{JS}(\lambda, L) - \boldsymbol{\theta}_b\|_2^2$. Then,

$$\begin{aligned} \tilde{R}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) &\leq r(\lambda_F, L_1) - r(\lambda_1, L_1) \\ &= \frac{1}{d} \sum_{b=1}^m (r_b(\lambda_F, L_1) - r_b(\lambda_1, L_1)) \\ &= \frac{1}{d} \sum_{b=1}^m (r_b^{JS}(\lambda_F, L_1) - r_b^{JS}(\lambda_2, L_1)) + \frac{1}{d} \sum_{b=1}^m (r_b(\lambda_F, L_1) - r_b^{JS}(\lambda_F, L_1)) \\ &\quad + \frac{1}{d} \sum_{b=1}^m (r_b^{JS}(\lambda_2, L_1) - r_b(\lambda_1, L_1)) \\ &= (I) + (II) + (III) \end{aligned}$$

By Lemma 1 of Cai and Zhou (2009), $(I) \leq C_\eta d^{\eta-1/2}$.

For (II), note that

$$\begin{aligned} r_b^{JS}(\lambda_F, L) - r_b(\lambda_F, L) &= E[\text{SUREJS}(\mathbf{x}_b, \lambda_F, L) - \text{SURESCAD}(\mathbf{x}_b, \lambda_F, L)] \\ &\equiv Eg(\mathbf{x}_b, \lambda_F, L), \end{aligned}$$

where

$$\begin{aligned} g(\mathbf{x}_b, \lambda_F, L) &= \frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} I(S_b^2 > 2\lambda_F) \\ &\quad - \left\{ \left(\frac{S_b - a\lambda_F/S_b}{a-2} \right)^2 + \frac{2L}{a-2} - \frac{2a\lambda_F(L-2)}{S_b^2(a-2)} \right\} I(2\lambda_F < S_b^2 \leq a\lambda_F) \\ &= \frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} I(S_b^2 > a\lambda_F) \\ &\quad + \left\{ \frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} - \left(\frac{S_b - a\lambda_F/S_b}{a-2} \right)^2 - \frac{2L}{a-2} + \frac{2a\lambda_F(L-2)}{S_b^2(a-2)} \right\} I(2\lambda_F < S_b^2 \leq a\lambda_F). \end{aligned}$$

If $S_b^2 > a\lambda_F$,

$$|g(\mathbf{x}_b, \lambda, L)| \leq \left| \frac{\lambda_F^2 - 2\lambda_F(L-2)}{a\lambda_F} \right| = \frac{1}{a} |\lambda_F - 2(L-2)|, \quad (11)$$

and

$$P(S_b^2 > a\lambda_F) = \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} P(\chi_{L+2k}^2 > a\lambda_F),$$

where $\theta_* = \|\boldsymbol{\theta}_b\|_2^2/2$ and χ_m^2 represents the χ^2 random variable with the degrees of freedom m . For $k < (a\lambda_F - L)/2$,

$$\begin{aligned} P(\chi_{L+2k}^2 > a\lambda_F) &\leq \frac{1}{2} \left[\left(\frac{a\lambda_F}{L+2k} \right)^{-1} \exp \left(-\frac{1}{2}(a\lambda_F - L - 2k) \right) \right] \\ &= o(1/d) \end{aligned} \quad (12)$$

by Lemma 2 in Cai (1999). For $k \geq (a\lambda_F - L)/2$,

$$\sum_k \frac{\theta_*^k e^{-\theta_*}}{k!} = o(1/d) \quad (13)$$

by Stirling formula. If $2\lambda_F < S_b^2 \leq a\lambda_F$,

$$\begin{aligned} g(\mathbf{x}_b, \lambda_F, L) \Big|_{S_b^2=2\lambda_F} &= -\frac{4}{a-2} \\ g(\mathbf{x}_b, \lambda_F, L) \Big|_{S_b^2=a\lambda_F} &= \frac{\lambda_F - 2L}{a} - \frac{8}{a(a-2)}, \end{aligned} \quad (14)$$

and

$$P(2\lambda_F < S_b^2 \leq a\lambda_F) \leq P(S_b^2 \geq 2\lambda_F) = o(1/d) \quad (15)$$

using similar arguments in (12) and (13). Combining (11)–(15), we obtain

$$(II) = o\left(\frac{\log d}{d}\right).$$

For (III), note that

$$\begin{aligned} &r_b^{JS}(\lambda_2, L) - r_b(\lambda_1, L) \leq r_b^{JS}(\lambda_F, L) - r_b(\lambda_1, L) \\ = &E \left(\frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} - S_b^2 + 2L \right) I(\lambda_F < S_b^2 \leq \lambda_1) \\ &+ E \left(\frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} - \frac{\lambda_1^2 - 2\lambda_1(L-2)}{S_b^2} \right) I(\lambda_1 < S_b^2 \leq 2\lambda_1) \\ &+ E \left(\frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} - \left(\frac{S_b - a\lambda_1/S_b}{a-2} \right)^2 - \frac{2L}{a-2} + \frac{2a\lambda_1(L-2)}{S_b^2(a-2)} \right) I(2\lambda_1 < S_b^2 \leq a\lambda_1) \\ &+ E \left(\frac{\lambda_F^2 - 2\lambda_F(L-2)}{S_b^2} \right) I(S_b^2 > a\lambda_1) \\ = &(i) + (ii) + (iii) + (iv). \end{aligned}$$

Then, for $\lambda_1 > \lambda_F$

$$\begin{aligned} (i) &\leq 4P(S_b^2 > \lambda_F) = o(1/d), \\ (ii) &\leq \frac{(\lambda_F - \lambda_1)(\lambda_F + \lambda_1 - 2L + 4)}{\lambda_1} \leq 0, \\ (iv) &\leq \lambda_F P(S_b^2 > a\lambda_F) = o(\lambda_F/d). \end{aligned}$$

In (iii), the maximum is achieved when $\lambda_1 = (S_b^2 + (L - 2)(a - 2))/a$, and

$$(iii) \leq \left(\frac{(\lambda_F - L + 2)^2}{2\lambda_1} - \frac{4}{a - 2} \right) P(S_b^2 \geq 2\lambda_F) = o(\lambda_F/d).$$

Therefore,

$$(III) = o\left(\frac{\log d}{d}\right),$$

which completes the proof of Lemma 1. \square

The following lemma is adapted from Donoho and Johnstone (1995).

Lemma 2. *If $\gamma_d^2 d / \log d \rightarrow \infty$, then*

$$\sup_{\mu_d \geq 3\gamma_d} (1 + \mu_d) P(T_d \leq \gamma_d) = o(d^{-1/2}).$$

The following lemma provides bounds for the loss of the SURE-Block-SCAD estimator, and the derivative of the risk function $r_b(\lambda, L)$. The first bound is used in the proof of Theorem 2 and the second in Theorem 1.

Lemma 3. *Let $\{x_i : i = 1, \dots, d\}$ be given in (4) with $\tau = 1$. Then,*

- (i) $\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 \leq 2\lambda_F d + 2\|\mathbf{z}\|_2^2$,
- (ii) $\left| \frac{\partial}{\partial \lambda} r_b(\lambda, L) \right| < 23$ for $\lambda > 0$ and $L \geq 1$.

Proof. (i). Define the event $A_d = \{T_d \leq \gamma_d\}$. Then,

$$\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 = \sum_{b=1}^m \|\hat{\boldsymbol{\theta}}_b(\lambda^*, L^*) - \boldsymbol{\theta}_b\|_2^2 I(A_d) + \|\hat{\boldsymbol{\theta}}^F - \boldsymbol{\theta}\|_2^2 I(A_d^c)$$

where $\hat{\boldsymbol{\theta}}^F$ denotes the non-negative garrote estimator given in (7). It can be shown that

$$\|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2 \leq \begin{cases} 2S_b^2 + 2\|\mathbf{z}_b\|_2^2, & \text{if } S_b^2 \leq \lambda, \\ 2\|\mathbf{z}_b\|_2^2 + 2\frac{\lambda^2}{S_b^2}, & \text{if } \lambda < S_b^2 \leq 2\lambda, \\ 2\|\mathbf{z}_b\|_2^2 + \lambda, & \text{if } 2\lambda < S_b^2 \leq a\lambda, \\ \|\mathbf{z}_b\|_2^2, & \text{if } S_b^2 > a\lambda, \end{cases}$$

where $\mathbf{z}_b = \mathbf{x}_b - \boldsymbol{\theta}_b$ and hence

$$\|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2 \leq 2\|\mathbf{z}_b\|_2^2 + 2\lambda.$$

Since $\lambda^* < \lambda_F$, we have

$$\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 = 2 \sum_b (\lambda^* + \|\mathbf{z}_b\|_2^2) I(A_d) + 2(\lambda_F d + \|\mathbf{z}\|_2^2) I(A_d^c) \leq 2\lambda_F d + 2\|\mathbf{z}\|_2^2.$$

(ii). Denote the density of χ_m^2 by f_m . One can show that $f_m(y) < 1$ for all $y > 0$ by Stirling formula. Then,

$$\begin{aligned} r_b(\lambda, L) &= \|\boldsymbol{\theta}_b\|_2^2 + E_{\boldsymbol{\theta}_b} \left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} \right) I(\lambda < S_b^2 < 2\lambda) \\ &+ E_{\boldsymbol{\theta}_b} \left(\left(\frac{S_b - a\lambda/S_b}{a-2} \right)^2 - \frac{2a\lambda(L-2)}{(a-2)S_b^2} + \frac{2L}{a-2} \right) I(2\lambda < S_b^2 \leq a\lambda) \\ &+ E_{\boldsymbol{\theta}_b} (2L - S_b^2) I(S_b^2 > \lambda), \end{aligned}$$

and

$$\begin{aligned} \left| \frac{\partial}{\partial \lambda} r_b(\lambda, L) \right| &= \left| \int_{\lambda}^{2\lambda} \frac{2\lambda - 2L + 4}{y} f_{L+2k}(y) dy + \frac{\lambda^2 - 2\lambda(L-2)}{\lambda} (f_{L+2k}(2\lambda) - f_{L+2k}(\lambda)) \right. \\ &+ \int_{2\lambda}^{a\lambda} \frac{-2a(y - a\lambda + (L-2)(a-2))}{(a-2)^2 y} f_{L+2k}(y) dy + \frac{4a}{a-2} f_{L+2k}(a\lambda) \\ &\left. - \left(\lambda - 2L + \frac{4a}{a-2} \right) f_{L+2k}(2\lambda) - (2L - \lambda) f_{L+2k}(\lambda) \right| \\ &\leq 2 + \left| (\lambda - 2L + 4)(f_{L+2k}(2\lambda) - f_{L+2k}(\lambda)) \right| + \frac{a}{a-2} + \frac{4a}{a-2} f_{L+2k}(a\lambda) \\ &- \left(\lambda - 2L + \frac{4a}{a-2} \right) f_{L+2k}(2\lambda) - (2L - \lambda) f_{L+2k}(\lambda) \Big| \\ &\leq 2 + \left| 4 - \frac{4a}{a-2} \right| f_{L+2k}(2\lambda) + 4f_{L+2k}(\lambda) + 3 + 9 < 23. \end{aligned}$$

□

The following lemma is used for Proposition 1 and its proof is omitted because it is simple.

Lemma 4. For $\lambda > \max(L-2, 0)$ and $L \geq 1$, we have

- (i) $|SURESC(\mathbf{x}_b, \lambda, L)| \leq \lambda + 5$,
- (ii) $r_b(\lambda, L) \leq \lambda + 5$,
- (iii) $\|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2 \leq 2\lambda + 2\|\mathbf{z}_b\|_2^2$,
- (iv) $\frac{\partial}{\partial \lambda} \|\hat{\boldsymbol{\theta}}_b(\lambda, L) - \boldsymbol{\theta}_b\|_2^2 \leq 9 + \frac{4}{\lambda} \|\mathbf{z}_b\|_2^2$.

In order to prove Theorem 1 we develop Proposition 1. Define

$$\begin{aligned} U(\lambda, L) &= \frac{1}{d} \text{SURESC}(\mathbf{x}, \lambda, L) \\ &= 1 + \frac{1}{d} \sum_{b=1}^m \left[(S_b^2 - 2L) I(S_b^2 \leq \lambda) + \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I(\lambda < S_b^2 < 2\lambda) \right. \\ &\quad \left. + \left\{ \left(\frac{S_b - a\lambda/S_b}{a-2} \right)^2 + \frac{2L}{a-2} - \frac{2a\lambda(L-2)}{(a-2)S_b^2} \right\} I(2\lambda < S_b^2 \leq a\lambda) \right], \end{aligned}$$

and $D(\lambda, L) = \frac{1}{d} \sum_{b=1}^m \|\hat{\theta}_b(\lambda, L) - \theta_b\|_2^2$. Note that both $D(\lambda, L)$ and $U(\lambda, L)$ have the same expectation $r(\lambda, L)$. We want to show that the minimizer (λ^*, L^*) of the risk $U(\lambda, L)$ is close to the ideal threshold level and block size in an asymptotic sense, that is,

$$\Delta_d = \left| ED(\lambda^*, L^*) - \inf_{\lambda, L} r(\lambda, L) \right|$$

is negligible for $\max(L-2, 0) \leq \lambda \leq \lambda_F$ and $1 \leq L \leq \sqrt{d}$. It can be shown that

$$\Delta_d \leq E \sup_{\lambda, L} |D(\lambda, L) - r(\lambda, L)| + 2E \sup_{\lambda, L} |r(\lambda, L) - U(\lambda, L)|, \quad (16)$$

and the following proposition provides the upper bounds for (16).

Proposition 1. *Uniformly in $\theta \in \mathcal{R}^d$, we have*

$$\begin{aligned} E_{\theta} \sup_{\max(L-2, 0) \leq \lambda \leq \lambda_F, 1 \leq L \leq \sqrt{d}} |U(\lambda, L) - r(\lambda, L)| &= o(d^{\eta-1/4}) \\ E_{\theta} \sup_{\max(L-2, 1/\log d) \leq \lambda \leq \lambda_F, 1 \leq L \leq \sqrt{d}} |D(\lambda, L) - r(\lambda, L)| &= o(d^{\eta-1/4}) \end{aligned}$$

for $0 < \eta < 1/2$.

Proof. The framework of the proof follows that of Proposition 2 in Cai and Zhou (2005). Define

$$Z_d(\lambda, L) \equiv U(\lambda, L) - r(\lambda, L) = \frac{1}{d} \sum_{b=1}^m (\text{SURESC}(\mathbf{x}_b, \lambda, L) - r_b(\lambda, L)).$$

Lemma 4 (i) and (ii) imply $|\text{SURESC}(\mathbf{x}_b, \lambda, L) - r_b(\lambda, L)| \leq 2(\lambda + 5)$, and for $r_d = L^{1/2}(\log d)^{1+\rho}$ with $\rho > 1/2$, it can be shown that

$$P(|Z_d(\lambda, L)| > r_d d^{-1/2}) \leq 2 \exp\left(-\frac{r_d^2 L}{2(\lambda + 5)^2}\right).$$

For distinct $0 < \lambda < \lambda' < a\lambda/2$ with $\lambda' - \lambda \leq \delta_d$ and $\delta_d/L = o(r_d d^{-1/2})$, let $N_d(\lambda, \lambda') = \max(\#\{i : \lambda < S_b^2 < \lambda'\}, \#\{i : a\lambda < S_b^2 < a\lambda'\})$. If we define

$$U^{JS}(\lambda, L) \equiv \text{SUREJS}(\mathbf{x}, \lambda, L) = 1 + \frac{1}{d} \sum_{b=1}^m \left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I(S_b^2 > \lambda) + (S_b^2 - 2L) I(S_b^2 \leq \lambda) \right),$$

then

$$\begin{aligned} |U(\lambda, L) - U(\lambda', L)| &\leq |U^{JS}(\lambda, L) - U^{JS}(\lambda', L)| + |U(\lambda, L) - U^{JS}(\lambda, L) - U(\lambda', L) + U^{JS}(\lambda', L)| \\ &\leq \text{(i)} + \text{(ii)}. \end{aligned}$$

By Proposition 2 in Cai and Zhou (2005), (i) $\leq \frac{4}{d}(1 + \delta_d/L)N_d(\lambda, \lambda') + 2\delta_d/L$. Note that

$$\begin{aligned} \text{(ii)} &\leq \frac{1}{d} \sum_{b=1}^m \left| \frac{(\lambda - \lambda')(\lambda + \lambda' - 2(L - 2))}{S_b^2} I(S_b^2 > a\lambda') \right| \\ &\quad + \frac{1}{d} \sum_{b=1}^m \left| \frac{(\lambda - \lambda')(\lambda + \lambda' - 2(L - 2))}{S_b^2} I(a\lambda < S_b^2 \leq a\lambda') \right| \\ &\quad + \frac{1}{d} \sum_{b=1}^m \left| \frac{(a\lambda' - S_b^2)(a\lambda' - S_b^2 + 2(L - 2)(a - 2))}{(a - 2)^2 S_b^2} I(a\lambda < S_b^2 \leq a\lambda') \right| \\ &\leq \frac{2\delta_d}{aL} + \frac{2}{ad} N_d(\lambda, \lambda') + \left(\frac{3}{d} + \frac{12\delta_d}{dL} \right) N_d(\lambda, \lambda') \leq \left(\frac{4}{d} + \frac{12\delta_d}{dL} \right) N_d(\lambda, \lambda') + \frac{\delta_d}{L}, \end{aligned}$$

and thus

$$|U(\lambda, L) - U(\lambda', L)| \leq \left(\frac{8}{d} + \frac{16\delta_d}{dL} \right) N_d(\lambda, \lambda') + \frac{3\delta_d}{L}.$$

Since $|r(\lambda) - r(\lambda')| \leq \delta_d/L$ by Lemma 3 (ii),

$$|Z_d(\lambda, L) - Z_d(\lambda', L)| \leq \frac{8}{d}(1 + 2\delta_d/L)N_d(\lambda, \lambda') + \frac{26\delta_d}{L}.$$

The rest of the proof follows from an analogous argument to the proof of Proposition 2 in Cai and Zhou (2005). \square

Proof of Theorem 1. The proof is similar to that of Theorem 4 in Cai and Zhou (2005).

In proving Theorems 2 and 3 the following lemma provides the bound for the approximation errors between the mean of the empirical wavelet coefficients and true coefficients of $f \in B_{p,q}^s(M)$. This is adapted from Lemma 4 in Cai and Zhou (2009).

Lemma 5. *Let $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_{j,k})$ be the DWT of the sampled function $\{n^{-1/2}f(i/n)\}$. Then, $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \leq Cn^{-2(s-1/p)}$.*

Proof of Theorem 2. We consider the normalized model of (8), that is

$$y'_{j,k} = \theta'_{j,k} + z_{j,k}, \quad j \geq j_0, \quad k = 1, 2, \dots, 2^j$$

where $y'_{j,k} = n^{1/2}\tilde{y}_{j,k}$ and $\theta'_{j,k} = n^{1/2}\tilde{\theta}_{j,k}$. Note that $E_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}\|_2^2 = n^{-1}E_{\boldsymbol{\theta}'}\|\hat{\boldsymbol{\theta}}'^* - \boldsymbol{\theta}'\|_2^2$. Let $\tilde{f} = \sum_{k=1}^{2^j} \tilde{\theta}_{j,k}\phi_{j,k}$. Then, $\sup_{f \in B_{p,q}^s(M)} \|\tilde{f} - f\|_2^2 = o(n^{-2s/(1+2s)})$ by Lemma 5.

Set $0 < \epsilon_0 < 1/(1 + 2s)$ and let J_0 be the largest integer satisfying $2^{J_0} \leq n^{\epsilon_0}$. Then,

$$\begin{aligned} n^{-1}E_{\boldsymbol{\theta}'}\|\hat{\boldsymbol{\theta}}'^* - \boldsymbol{\theta}'\|_2^2 &= \left(\sum_{j \leq J_0} \sum_k + \sum_{J_0 \leq j < J_1} \sum_k + \sum_{j \geq J_1} \sum_k \right) n^{-1}E(\hat{\theta}'_{j,k} - \theta'_{j,k})^2 \\ &= S_1 + S_2 + S_3 \end{aligned}$$

where $J_1 > J_0$ satisfies $2^{J_1} = n^\gamma$ with

$$\frac{1}{1/2 + 2(s - (1/p - 1/2)_+)} < \gamma < \frac{1}{(1 + 2s)(\eta + 3/4)}$$

for sufficiently small $\eta > 0$. According to Theorem 3 of Cai and Zhou (2009),

$$S_1 + S_3 = o\left(n^{-\frac{2s}{1+2s}}\right).$$

We turn to the term S_2 .

$$\begin{aligned} S_2 &\leq \sum_{J_0 \leq j < J_1} n^{-1}2^j R(\boldsymbol{\theta}'_j) + \sum_{J_0 \leq j < J_1} n^{-1}2^j R_F(\boldsymbol{\theta}'_j) I(\mu'_{2^j} \leq 3\gamma_{2^j}) + \sum_{J_0 \leq j < J_1} n^{-1}c_\eta 2^{j(\eta+3/4)} \\ &= S_{21} + S_{22} + S_{23} \end{aligned}$$

where $\mu'_{2^j} = 2^{-j}n\|\boldsymbol{\theta}_j\|_2^2$ and $\gamma_{2^j} = 2^{-j/2}j^{3/2}$. Note that

$$S_{21} \leq R_T^*(B_{p,q}^s(M))$$

and $R_T^*(B_{p,q}^s(M)) \geq R^*(B_{p,q}^s(M)) \geq Cn^{-2s/(1+2s)}$ for some constant $C > 0$. Also,

$$S_{23} \leq Cn^{-1}2^{J_1(\eta+3/4)} = o\left(n^{-\frac{2s}{1+2s}}\right).$$

When $L = 1$ and $\lambda = \lambda_F$, it can be shown that

$$2^j R_F(\boldsymbol{\theta}'_j) \leq \sum_{k=1}^{2^j} (n\theta_{j,k}^2 \wedge \lambda_F) + 8(2 \log 2^j)2^{-j} + o\left(\frac{j}{2^j}\right) \leq n\|\boldsymbol{\theta}_j\|_2^2 + 8(2 \log 2^j)2^{-j} + o\left(\frac{j}{2^j}\right).$$

Then,

$$S_{22} \leq \sum_{J_0 \leq j < J_1} \left(3n^{-1}2^{j/2}j^{3/2} + 8n^{-1}(2 \log 2^j)2^{-j} \right) = o\left(n^{-\frac{2s}{1+2s}}\right),$$

which completes the proof.

Proof of Theorem 3. (i) The proof is similar to that of Theorem 4 in Cai and Zhou

(2009) except one part. It can be shown that

$$\begin{aligned}
n^{-1}2^j R(\boldsymbol{\theta}'_j) &\leq n^{-1} \sum_{b=1}^m E_{\boldsymbol{\theta}'_b} \|\hat{\boldsymbol{\theta}}'_b(L_j - 2, L_j) - \boldsymbol{\theta}'_b\|_2^2 \\
&\leq n^{-1} \sum_{b=1}^m E_{\boldsymbol{\theta}'_b} \|\hat{\boldsymbol{\theta}}'_b(L_j - 2, L_j) - \hat{\boldsymbol{\theta}}'^{JS}_b(L_j - 2, L_j)\|_2^2 \\
&\quad + n^{-1} \sum_{b=1}^m E_{\boldsymbol{\theta}'_b} \|\hat{\boldsymbol{\theta}}'^{JS}_b(L_j - 2, L_j) - \boldsymbol{\theta}'_b\|_2^2 \\
&= n^{-1} \sum_{b=1}^m E_{\boldsymbol{\theta}'_b} \|\hat{\boldsymbol{\theta}}'^{JS}_b(L_j - 2, L_j) - \boldsymbol{\theta}'_b\|_2^2 + o\left(\frac{j}{n}\right).
\end{aligned}$$

Therefore, we can use the result from $E_{\boldsymbol{\theta}'_b} \|\hat{\boldsymbol{\theta}}'^{JS}_b(L_j - 2, L_j) - \boldsymbol{\theta}'_b\|_2^2$.

(ii) We need the following proposition for the proof of Theorem 3 (ii).

Proposition 2. *Let $X \sim N(\mu, 1)$ and let $\mathcal{F}_p(\eta)$ denote the probability measures $F(d\mu)$ satisfying the moment condition $\int |\mu|^p F(d\mu) \leq \eta^p$. Let*

$$\bar{r}(\delta_\lambda^g, \eta) = \sup_{\mathcal{F}_p(\eta)} \left\{ E_F r_g(\mu) : \int |\mu|^p F(d\mu) \leq \eta^p \right\}$$

where $r_g(\mu) = E_\mu(\delta_\lambda^g(x) - \mu)^2$ and

$$\delta_\lambda^g(x) = \begin{cases} \left(1 - \frac{\lambda^2}{x^2}\right)_+ x, & \text{if } x^2 \leq 2\lambda^2, \\ \frac{a-1-2a\lambda^2/x^2}{a-2} x, & \text{if } 2\lambda^2 < x^2 \leq a\lambda^2, \\ x, & \text{if } x^2 > a\lambda^2. \end{cases}$$

If $0 < p < 2$ and $\lambda = \sqrt{2 \log \eta^{-p}}$, then

$$\bar{r}(\delta_\lambda^g, \eta) \leq 2\eta_p \lambda^{2-p} (1 + o(1))$$

as $\eta \rightarrow 0$.

Proof. It follows from a similar argument to the proof of Proposition 16 in Donoho and Johnstone (1994b) that

$$\bar{r}(\delta_\lambda^g, \eta) = \sup_{\mu \geq \eta} \left(\frac{\eta}{\mu}\right)^p r_g(\mu) (1 + o(1)),$$

as $\eta \rightarrow 0$. Note that

$$\begin{aligned}
r_g(\mu) &= E_\mu \left[1 + (x^2 - 2)I(x^2 \leq \lambda^2) + \frac{\lambda^4 + 2\lambda^2}{x^2}I(\lambda^2 < x^2 \leq 2\lambda^2) \right. \\
&\quad \left. + \left\{ \left(\frac{x - a\lambda^2/x}{(a-2)^2} \right)^2 + \frac{2}{a-2} + \frac{2a\lambda^2}{(a-2)x^2} \right\} I(2\lambda^2 < x^2 \leq a\lambda^2) \right] \\
&\leq 1 + (\lambda^2 - 2)P(x^2 \leq \lambda^2) + (2 + \lambda^2)P(\lambda^2 < x^2 \leq 2\lambda^2) + (\lambda^2 + 4)P(2\lambda^2 < x^2 \leq a\lambda^2) \\
&\leq \lambda^2(1 + o(1)),
\end{aligned}$$

which implies that for $\mu \geq \lambda$

$$\left(\frac{\eta}{\mu} \right)^p r_g(\mu) \leq \eta^p \lambda^{2-p} (1 + o(1)),$$

as $\eta \rightarrow 0$. Similarly,

$$\begin{aligned}
r_g(\mu) &= E_\mu \left[x^2 - 1 + \left(\frac{\lambda^4 + 2\lambda^2}{x^2} - x^2 + 2 \right) I(\lambda^2 < x^2 \leq 2\lambda^2) \right. \\
&\quad \left. + \left\{ \left(\frac{x - a\lambda^2/x}{a-2} \right)^2 + \frac{2}{a-2} + \frac{2a\lambda^2}{(a-2)x^2} - x^2 + 2 \right\} I(2\lambda^2 < x^2 \leq a\lambda^2) \right] \\
&\leq \mu^2 + 6P(x^2 \geq \lambda^2).
\end{aligned}$$

Then, it can be show that for $\eta \leq \mu < \lambda$

$$\left(\frac{\eta}{\mu} \right)^p r_g(\mu) \leq \eta^p \lambda^{2-p} (1 + o(1)),$$

as $\eta \rightarrow 0$, which completes the proof. \square

Proof of Theorem 3 (ii). Set $\rho_g(\eta) \equiv \inf_\lambda \sup_{\mathcal{F}_p(\eta)} E_F r_g(\mu)$. Proposition 2 implies that

$$\rho_g(\eta) \leq \bar{r}(\delta_\lambda^g, \eta) \leq 2\eta^p (2 \log \eta^{-p})^{(2-p)/2} (1 + o(1))$$

as $\eta \rightarrow 0$. For $0 < p < 2$, Theorem 15 of Donoho and Johnstone (1994b) shows that

$$\rho(\eta) \equiv \inf_\delta \sup_{\mathcal{F}_p(\eta)} E_F E_\mu (\delta(x) - \mu)^2 = \eta^p (2 \log \eta^{-p})^{(2-p)/2} (1 + o(1))$$

as $\eta \rightarrow 0$. Note that $\rho_g(\eta)/\rho(\eta)$ is bounded as $\eta \rightarrow 0$ and $\rho_g(\eta)/\rho(\eta) \rightarrow 1$ as $\eta \rightarrow \infty$. Both $\rho_g(\eta)$ and $\rho(\eta)$ are continuous on $(0, \infty)$, and thus

$$G(p) = \sup_\eta \frac{\rho_g(\eta)}{\rho(\eta)} < \infty$$

for $0 < p < 2$. Theorems 4 and 5 in Donoho and Johnstone (1998) obtain the asymptotic minimaxity over Besov bodies from the univariate Bayes-minimax estimators. We use a similar argument as in Section 5.3 of Donoho and Johnstone (1998), and then

$$\begin{aligned} R_T^*(B_{p,q}^s(M)) &\leq \inf_{\lambda_j} \sup_{\boldsymbol{\theta} \in B_{p,q}^s(M)} E \sum_{j=j_0}^{\infty} \|\hat{\boldsymbol{\theta}}_j(\lambda_j, 1) - \boldsymbol{\theta}_j\|_2^2 \\ &\leq G(p \wedge q) R^*(B_{p,q}^*(M))(1 + o(1)). \end{aligned}$$

Acknowledgements

The author is grateful to Professor Tony Cai for his useful comments. The work was supported in part by National Security Agency Grant No. H982300810056.

References

- ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*. **96**, 939–967.
- ANTONIADIS, A. and PHAM, D.T. (1998). Wavelet regression for random or irregular design. *Computational Statistics and Data Analysis*. **28**, 353–369.
- BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*. **37**, 373–384.
- CAI, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *The Annals of Statistics*. **27**, 898–924.
- CAI, T. T. (2002). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statistica Sinica*. **12**, 1241–1273.
- CAI, T. T. and SILVERMAN, B. W. (2001) Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya*. **63** 127–148.
- CAI, T. and ZHOU, H. (2005). A data-driven block thresholding approach to wavelet estimation. Technical report, Department of Statistics, University of Pennsylvania.
- CAI, T. and ZHOU, H. (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*. **37** 569–595.

- CHICKEN, E. (2003). Block thresholding and wavelet estimation for nonequispaced samples. *Journal of Statistical Planning and Inference*. **116** 113–129.
- DAUBECHIES, I. (1992) *Ten Lectures on Wavelets* SIAM. Philadelphia, PA.
- DEVORE, R. and POPOV, V. (1988). Interpolation of Besov spaces. *Transactions of the American Mathematical Society*. **305** 397–414.
- DONOHO, D. and JOHNSTONE, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*. **81** 425–455.
- DONOHO, D. and JOHNSTONE, I. M. (1994b). Minimax Risk over l_p -Balls for l_q -error. *Probability Theory and Related Fields*. **99** 277-303.
- DONOHO, D. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*. **90** 1200–1224.
- DONOHO, D. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*. **26** 879-921.
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. **96** 1348–1360.
- FAN, J. and LI, R. (2002). Variable selection for Cox’s proportional hazards model and frailty model. *The Annals of Statistics*. **30** 74–99.
- HALL, P., KERKYACHARIAN, G., and PICARD, D. (1998). Block thresholding rules for curve estimation using kernel and wavelet methods. *The Annals of Statistics*. **26** 922–942.
- HALL, P., KERKYACHARIAN, G., and PICARD, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica*. **9** 33–49.
- HALL, P., PENEV, S., KERKYACHARIAN, G., and PICARD, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statistics and Computing*. **7** 115–124.
- NUNES, M., KNIGHT, M. and NASON, G.P. (2006). Adaptive lifting for nonparametric regression. *Statistics and Computing*. **16** 143–159.
- PARK, C. and KIM, W.-C. (2004). Estimation of a regression function with a sharp change point using boundary wavelets. *Statistics & Probability Letters*. **66** 435–448.

- PARK, C. and KIM, W.-C. (2006). Wavelet estimation of a regression function with a sharp change point in a random design. *Journal of Statistical Planning and Inference*. **136** 2381–2394.
- PENSKY, M. and VIDAKOVIC, B. (2001). On non-equally spaced wavelet regression. *Annals of the Institute of Statistical Mathematics*. **53** 681–690.
- WANG, L. CHEN, G., and LI, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*. **23** 1486–1494.
- WILCOXON, F. (2007). Individual comparisons by ranking methods. *Biometrics Bulletin*. **1** 80–83.
- ZHANG, H., AHN, J., LIN, X., and PARK, C. (2006). Gene selection using support vector machines with nonconvex penalty. *Bioinformatics*. **22** 88–95.