A Data-Driven Block Thresholding Approach To Wavelet Estimation

T. Tony Cai[1] and Harrison H. Zhou[2]
University of Pennsylvania and Yale University

Abstract

A data-driven block thresholding procedure for wavelet regression is proposed and its theoretical and numerical properties are investigated. The procedure 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 class of Besov bodies and achieves simultaneously within a small constant factor of the minimax risk over a wide collection of Besov Bodies including both the dense and sparse cases. The procedure is easy to implement. Numerical results show that it has superior finite sample performance in comparison to the other leading wavelet thresholding estimators.

Keywords: Adaptivity; Besov body; Block thresholding; James-Stein estimator; Nonparametric regression; Stein's unbiased risk estimate; Wavelets.
AMS 1991 Subject Classification: Primary 62G07, Secondary 62G20.

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
The research of Tony Cai was supported in part by NSF Grants DMS-0072578 and DMS-0306576.
[2]Department of Statistics, Yale University, New Haven, CT 06511.

1 Introduction

Consider the nonparametric regression model:

$$y_i = f(t_i) + \sigma\, z_i, \quad i = 1, 2, ..., n, \tag{1}$$

where $t_i = i/n$, σ is the noise level and z_i's are independent standard normal random variables. The goal is to estimate the unknown regression function $f(\cdot)$ based on the sample $\{y_i\}$.

Wavelet methods have demonstrated considerable success in nonparametric regression. They achieve a high degree of adaptivity through thresholding of the empirical wavelet coefficients. Standard wavelet approaches threshold the empirical coefficients term by term based on their individual magnitudes. See, for example, Donoho and Johnstone (1994a), Gao (1998), and Antoniadis and Fan (2001). More recent work has demonstrated that block thresholding, which simultaneously keeps or kills all the coefficients in groups rather than individually, enjoys a number of advantages over the conventional term-by-term thresholding. Block thresholding increases estimation precision by utilizing information about neighboring wavelet coefficients and allows the balance between variance and bias to be varied along the curve which results in adaptive smoothing. The degree of adaptivity, however, depends on the choice of block size and threshold level.

The idea of block thresholding can be traced back to Efromovich (1985) in orthogonal series estimators. In the context of wavelet estimation, global level-by-level thresholding was discussed in Donoho and Johnstone (1995) for regression and in Kerkyacharian, Picard and Tribouley (1996) for density estimation. But these block thresholding methods are not local, so they do not enjoy a high degree of spatial adaptivity. Hall, Kerkyacharian and Picard (1999) introduced a local blockwise hard thresholding procedure with a block size of the order $(\log n)^2$ where n is the sample size. Cai (1999) considered blockwise James-Stein rules and investigated the effect of block size and threshold level on adaptivity using an oracle inequality approach. In particular it was shown that a block size of order $\log n$ is optimal in the sense that it leads to an estimator which is both globally and locally adaptive. Cai and Silverman (2001) considered overlapping block thresholding estimators and Chicken and Cai (2004) applied block thresholding to density estimation.

The block size and threshold level play important roles in the performance of a block thresholding estimator. The local block thresholding methods mentioned above all have fixed block size and threshold and same thresholding rule is applied to all resolution levels regardless of the distribution of the wavelet coefficients. In the present paper, we propose a data-driven approach to empirically select both the block size and threshold at individual

resolution levels. At each resolution level, the procedure, *SureBlock*, chooses the block size and threshold empirically by minimizing Stein's Unbiased Risk Estimate (SURE). By empirically selecting both the block size and threshold and allowing them to vary from resolution level to resolution level, the SureBlock estimator has significant advantages over the more conventional wavelet thresholding estimators with fixed block sizes.

We consider in the present paper both the numerical performance and asymptotic properties of the SureBlock estimator. It is shown that the minimizer of SURE yields a near-optimal estimate of the wavelet coefficient vector at each resolution level which in turn generates an adaptive nonparametric estimator of the regression function f. The theoretical properties of the SureBlock estimator are considered in the Besov sequence space formulation that is by now classical for the analysis of wavelet regression methods. Besov spaces, denoted by $B^{\alpha}_{p,q}$ and defined in Section 4, are a very rich class of function spaces which contain functions of homogeneous and inhomogeneous smoothness. The theoretical results show that the SureBlock estimator automatically adapts to the sparsity of the underlying wavelet coefficient sequence and enjoys excellent adaptivity over a wide range of Besov bodies. In particular, in the "dense case" $p \geq 2$ the SureBlock estimator is sharp adaptive over all Besov bodies $B^{\alpha}_{p,q}(M)$ with $p = q = 2$ and adaptively achieves within a factor of 1.25 of the minimax risk over Besov bodies $B^{\alpha}_{p,q}(M)$ for all $p \geq 2$, $q \geq 2$. At the same time the SureBlock estimator achieves simultaneously within a constant factor of the minimax risk over a wide collection of Besov bodies $B^{\alpha}_{p,q}(M)$ in the "sparse case" $p < 2$. These properties are not shared simultaneously by many commonly used fixed block size procedures such as VisuShrink (Donoho and Johnstone (1994a)), SureShrink (Donoho and Johnstone (1995)) or BlockJS (Cai (1999)).

The SureBlock estimator is completely data-driven and easy to implement. A simulation study is carried out and the numerical findings reconfirm the theoretical results. The numerical results show that SureBlock has superior finite sample performance in comparison to the other leading wavelet estimators. More specifically, SureBlock uniformly outperforms both VisuShrink and SureShrink (Donoho and Johnstone, 1994a and 1995) in all 42 simulation cases in terms of the average squared error. SureBlock procedure is better than BlockJS (Cai (1999)) in 37 out of 42 cases.

The paper is organized as follows. After Section 2 in which basic notation and block thresholding methods are briefly reviewed, we introduce in Section 3 the SureBlock procedure. Asymptotic properties of the SureBlock estimator are presented in Section 4. Numerical implementation and finite-sample performance of the SureBlock estimator are then discussed in Section 5. The numerical performance of SureBlock is compared with

those of VisuShrink (Donoho and Johnstone (1994a)), SureShrink (Donoho and Johnstone (1995)) and BlockJS (Cai (1999)). The proofs are given in Section 6.

2 Wavelets And Block Thresholding

Let $\{\phi,\ \psi\}$ be a pair of father and mother wavelets. The functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. Dilation and translation of ϕ and ψ generate an orthonormal wavelet basis with an associated orthogonal Discrete Wavelet Transform (DWT) which transforms sampled data into the wavelet coefficient domain. A wavelet ψ is called *r-regular* if ψ has r vanishing moments and r continuous derivatives. See Daubechies (1992) and Strang (1992) for details on the discrete wavelet transform and compactly supported wavelets.

Let $\phi_{j,k}(x) = 2^{\frac{j}{2}}\phi(2^j x - k)$, and $\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$. For a given square-integrable function f on $[0,1]$, define the wavelet coefficients of the wavelet expansion of f by

$$\xi_{j,k} = \langle f, \phi_{j,k}\rangle, \qquad \theta_{j,k} = \langle f, \psi_{j,k}\rangle.$$

Suppose we observe the noisy data $Y = (y_1, y_2, ..., y_n)'$ as in (1) and suppose the sample size $n = 2^J$ for some integer $J > 0$. We use the standard device of the discrete wavelet transform to turn the function estimation problem into a problem of estimating wavelet coefficients. Let $\widetilde{Y} = W \cdot Y$ be the discrete wavelet transform of Y. Then $\widetilde{Y}$can be written as

$$\widetilde{Y} = (\widetilde{\xi}_{j_0,1}, \cdots, \widetilde{\xi}_{j_0 2^{j_0}}, \widetilde{y}_{j_0,1}, \cdots, \widetilde{y}_{j_0,2^{j_0}}, \cdots, \widetilde{y}_{J-1,1}, \cdots, \widetilde{y}_{J-1,2^{J-1}})'. \tag{2}$$

where j_0 is some fixed primary resolution level. Here $\widetilde{\xi}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $\widetilde{y}_{j,k}$ are the empirical wavelet coefficients at level j which represent fine structure at scale 2^j. Since the DWT is an orthogonal transform, the $\widetilde{y}_{j,k}$ are independent normal random variables with standard deviation σ. Note that the mean of $n^{-\frac{1}{2}}\widetilde{y}_{j,k}$ equals, up to some "small" approximation errors, the wavelet coefficient $\theta_{j,k}$ of f. See Daubechies (1992). Through the discrete wavelet transform the nonparametric regression problem is then turned into a problem of estimating a high dimensional normal mean vector.

A block thresholding procedure thresholds wavelet coefficients in groups and makes simultaneous decisions on all the coefficients within a block. Fix a block size L and a threshold level λ and divide the empirical wavelet coefficients $\widetilde{y}_{j,k}$ at any given resolution level j into nonoverlapping blocks of size L. Denote (jb) the indices of the coefficients in the b-th block at level j, i.e. $(jb) = \{(j,k) : (b-1)L + 1 \le k \le bL\}$. Let $S^2_{jb} = \sum_{k\in(jb)} \widetilde{y}^2_{j,k}$

denote the sum of squared empirical wavelet coefficients in the block. A block (jb) is deemed important if S_{jb}^2 is larger than the threshold λ and then all the coefficients in the block are retained; otherwise the block is considered negligible and all the coefficients in the block are estimated by zero. The most commonly used rules for a given block are the James-Stein rule

$$\widehat{\theta}_{j,k} = \widetilde{y}_{j,k} \cdot (1 - \frac{\lambda}{S_{jb}^2})_+ \qquad \text{for } (j,k) \in (jb) \tag{3}$$

and the hard thresholding rule

$$\widehat{\theta}_{j,k} = \widetilde{y}_{j,k} \cdot I(S_{jb}^2 > \lambda), \qquad \text{for } (j,k) \in (jb). \tag{4}$$

The estimator of $\{f(t_i), i = 1, 2, \ldots, n\}$ is then given by the inverse discrete wavelet transform of the thresholded wavelet coefficients.

Block thresholding estimators depend on the choice of the block size L and threshold level λ which largely determines the performance of the resulting estimator. It is thus important to choose L and λ in an optimal way.

3 The SureBlock Procedure

As discussed in the previous section the nonparametric regression problem of estimating f can be turned into a problem of estimating a high dimensional normal mean vector by applying the orthogonal discrete wavelet transform to the data. We thus begin in this section by considering estimation of the mean of a multivariate normal random variable.

Suppose that we observe

$$x_i = \theta_i + \sigma z_i, \quad i = 1, 2, ..., d \tag{5}$$

where z_i are independent standard normal random variables and σ is known. Here $x = (x_1, \cdots, x_d)$ represents the observations in one resolution level in wavelet regression. Without loss of generality we assume in this section $\sigma = 1$. We wish to estimate $\theta = (\theta_1, \cdots, \theta_d)$ based on the observations $x = (x_1, \cdots, x_d)$ under the average mean squared error:

$$R(\widehat{\theta},\ \theta) = \frac{1}{d} \sum_{i=1}^{d} E(\widehat{\theta}_i - \theta_i)^2. \tag{6}$$

We shall estimate θ by a blockwise James-Stein estimator with block size L and threshold level λ chosen empirically by minimizing Stein's Unbiased Risk Estimate (SURE). Let $L \geq 1$ be the possible length of each block, and $m = d/L$ be the number of blocks.

(For simplicity we shall assume that d is divisible by L in the following discussion.) Let $\underline{x}_b = (x_{(b-1)L+1}, ..., x_{bL})$ represent observations in the b-th block, and similarly $\underline{\theta}_b = \left(\theta_{(b-1)L+1}, ..., \theta_{bL}\right)$ and $\underline{z}_b = \left(z_{(b-1)L+1}, ..., z_{bL}\right)$. Let $S_b^2 = \|\underline{x}_b\|_2^2$ for $b = 1, 2, ..., m$. The Blockwise James-Stein estimator is given by

$$\widehat{\underline{\theta}}_b(\lambda, L) = \left(1 - \frac{\lambda}{S_b^2}\right)_+ \underline{x}_b, \quad b = 1, 2, ..., m \tag{7}$$

where $\lambda \geq 0$ is the threshold level. Write $\widehat{\underline{\theta}}_b(\lambda, L) = \underline{x}_b + g(\underline{x}_b)$, where g is a function from $I\!R^L$ to $I\!R^L$. Stein (1981) showed that when g is weakly differentiable, then

$$E_{\underline{\theta}_b}\|\widehat{\underline{\theta}}_b(\lambda, L) - \underline{\theta}_b\|_2^2 = E_{\underline{\theta}_b}\left\{L + \|g\|_2^2 + 2\nabla \cdot g\right\}.$$

In our case,

$$g\left(\underline{x}_b\right) = \left(1 - \frac{\lambda}{S_b^2}\right)_+ \underline{x}_b - \underline{x}_b$$

is weakly differentiable. Simple calculations show $E_{\underline{\theta}_b}\|\widehat{\underline{\theta}}_b(\lambda, L) - \underline{\theta}_b\|_2^2 = E_{\underline{\theta}_b}\left(SURE\left(\underline{x}_b, \lambda, L\right)\right)$ where

$$SURE\left(\underline{x}_b, \lambda, L\right) = L + \frac{\lambda^2 - 2\lambda\left(L-2\right)}{S_b^2} I\left(S_b^2 > \lambda\right) + \left(S_b^2 - 2L\right) I\left(S_b^2 \leq \lambda\right). \tag{8}$$

This implies that the total risk $E_\theta\|\widehat{\theta}(\lambda, L) - \theta\|_2^2 = E_\theta SURE\left(x, \lambda, L\right)$ where

$$SURE\left(x, \lambda, L\right) = \sum_{b=1}^{m} SURE\left(\underline{x}_b, \lambda, L\right) \tag{9}$$

is an unbiased estimate of risk.

We choose the block size L^S and threshold level λ^S to be the minimizer of $SURE\left(x, \lambda, L\right)$, i.e.

$$\left(\lambda^S, L^S\right) = \underset{\lambda,\, L}{\arg\min} SURE\left(x, \lambda, L\right).$$

The estimator of θ is then given by (7) with $L = L^S$ and $\lambda = \lambda^S$.

The SURE Block James-Stein estimator has some drawbacks in situations of extreme sparsity of wavelet coefficients. We propose to use a hybrid scheme similar to Donoho and Johnstone (1995) and Johnstone (1999). The hybrid method works as follows. Set $T_d = d^{-1}\sum\left(x_i^2 - 1\right)$, $\gamma_d = d^{-\frac{1}{2}}\log_2^{\frac{3}{2}} d$ and $\lambda^F = 2L\log d$. For a fixed $0 \leq v < 1$ let (λ^*, L^*) denote the minimizers of SURE with an additional restriction on the search range :

$$(\lambda^*, L^*) = \underset{\max\{L-2,0\} \leq \lambda \leq \lambda^F, 1 \leq L \leq d^v}{\arg\min} SURE\left(x, \lambda, L\right). \tag{10}$$

Define the estimator $\widehat{\theta}^*(x)$ of θ by

$$\underline{\widehat{\theta}^*}_b = \underline{\widehat{\theta}}_b(\lambda^*, L^*), \quad \text{if } T_d > \gamma_d, \tag{11}$$

and

$$\widehat{\theta}_i^* = \left(1 - \frac{2\log d}{x_i^2}\right)_+ x_i, \quad \text{if } T_d \le \gamma_d. \tag{12}$$

When $T_d \le \gamma_d$ the estimator is a degenerate Block James-Stein estimator with block size $L = 1$. In this case the estimator is also called the non-negative garrote estimator. See Breiman (1995) and Gao (1998). The SURE approach has also been used for the selection of the threshold level for fixed block size procedures, term-by-term thresholding ($L = 1$) in Donoho and Johnstone (1995) and block thresholding ($L = \log n$) in Chicken (2004).

We now return to the nonparametric regression model (1). As in Section 2 denote by $\underline{\widetilde{Y}}_j = \{\widetilde{y}_{j,k} : \ k = 1, ..., 2^j\}$ and $\underline{\theta}_j = \{\theta_{j,k} : \ k = 1, ..., 2^j\}$ the empirical and true wavelet coefficients of the regression function f at resolution level j. We apply the hybrid SURE Block James-Stein procedure to the empirical wavelet coefficients $\underline{\widetilde{Y}}_j$ level by level and then use the inverse discrete wavelet transform to obtain the estimate of the regression function. More specifically the procedure for wavelet regression, called *SureBlock*, has the following steps.

1. Transform the data into the wavelet domain via the discrete wavelet transform: $\widetilde{Y} = W \cdot Y$.

2. At each resolution level j, Select the block size L_j^* and threshold level λ_j^* by minimizing Stein's unbiased risk estimate and then estimate the wavelet coefficients by the hybrid Block James-Stein rule. More precisely, for $j_0 \le j < J$

$$(\lambda_j^*, L_j^*) = \underset{\max\{L-2,0\} \le \lambda \le \lambda^F, 1 \le L \le 2^{vj}}{\arg\min} SURE\left(\sigma^{-1}\underline{\widetilde{Y}}_j, \lambda, L\right) \tag{13}$$

   where $\underline{\widetilde{Y}}_j$ is the empirical wavelet coefficient vector at resolution level j, and

$$\underline{\widehat{\theta}}_j = \sigma \cdot \widehat{\theta}^*(\sigma^{-1}\underline{\widetilde{Y}}_j) \tag{14}$$

   where $\widehat{\theta}^*$ is the hybrid Block James-Stein estimator given in (11) and (12).

3. Obtain the estimate of the function via the inverse discrete wavelet transform of the denoised wavelet coefficients: $\widehat{f} = W^{-1} \cdot \widehat{\Theta}$.

Theoretical results given in Section 4 show that the pair (L_j^*, λ_j^*) is asymptotically optimal in the sense that the resulting block thresholding estimator adaptively attains the exact minimax block thresholding risk asymptotically.

Note that, in contrast to some other block thresholding methods, both the block size and threshold level of the SureBlock estimator are chosen empirically and vary from resolution level to resolution level. Both the theoretical and numerical results given in the next two sections show that the SureBlock estimator outperforms wavelet thresholding estimators with a fixed block size.

4 Theoretical Properties of SureBlock

In this section we turn to the theoretical properties of the SureBlock estimator in the Besov sequence space formulation that is by now classical for the analysis of wavelet regression methods. The asymptotic results show that the SureBlock procedure is strongly adaptive.

Besov spaces are a very rich class of function spaces and contain as special cases many traditional smoothness spaces such as Hölder and Sobolev Spaces. Roughly speaking, the Besov space $B_{p,q}^{\alpha}$ contains functions having α bounded derivatives in L^p norm, the third parameter q gives a finer gradation of smoothness. Full details of Besov spaces are given, for example, in Triebel (1983) and DeVore and Popov (1988). For a given r-regular mother wavelet ψ with $r > \alpha$ and a fixed primary resolution level j_0, the Besov sequence norm $\|\cdot\|_{b_{p,q}^{\alpha}}$ of the wavelet coefficients of a function f is then defined by

$$\|f\|_{b_{p,q}^{\alpha}} = \|\underline{\xi}_{j_0}\|_p + \left(\sum_{j=j_0}^{\infty}(2^{js}\|\underline{\theta}_j\|_p)^q\right)^{\frac{1}{q}} \tag{15}$$

where $\underline{\xi}_{j_0}$ is the vector of the father wavelet coefficients at the primary resolution level j_0, $\underline{\theta}_j$ is the vector of the wavelet coefficients at level j, and $s = \alpha + \frac{1}{2} - \frac{1}{p} > 0$. Note that the Besov function norm of index (α, p, q) of a function f is equivalent to the sequence norm (15) of the wavelet coefficients of the function. See Meyer (1992).

We shall focus our theoretical discussion to a sequence space version of the SureBlock estimator. Suppose that $n = 2^J$ for some integer J and that we observe sequence data

$$y_{j,k} = \theta_{j,k} + n^{-\frac{1}{2}}\sigma z_{j,k}, \quad j \geq j_0,\ k = 1, 2, \cdots, 2^j \tag{16}$$

where $z_{j,k}$ are independent standard normal random variables. We wish to estimate the mean array θ under the expected squared error

$$R(\hat{\theta}, \theta) = E\sum_{j,k}(\widehat{\theta}_{j,k} - \theta_{j,k})^2 = E\|\widehat{\theta} - \theta\|_2^2.$$

Define the Besov body $B^{\alpha}_{p,q}(M)$ by

$$B^{\alpha}_{p,q}(M) = \{\theta : (\sum_{j=j_0}^{\infty}(2^{js}\|\underline{\theta}_j\|_p)^q)^{1/q} \leq M\} \tag{17}$$

where, as usual, $s = \alpha + \frac{1}{2} - \frac{1}{p} > 0$. The minimax risk of estimating θ over the Besov body $B^{\alpha}_{p,q}(M)$ is

$$R^*(B^{\alpha}_{p,q}(M)) = \inf_{\widehat{\theta}} \sup_{\theta \in B^{\alpha}_{p,q}(M)} E\|\widehat{\theta} - \theta\|_2^2. \tag{18}$$

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

We apply the SureBlock procedure of Section 3 to the array of empirical coefficients $y_{j,k}$ for $j_0 \leq j < J$, to obtain estimated wavelet coefficients $\widehat{\theta}_{j,k}$. Set $\widehat{\theta}_{j,k} = 0$ for $j \geq J$. The minimax risk among all Block James-Stein estimators with all possible block sizes $1 \leq L_j \leq 2^{jv}$ with $0 \leq v < 1$ and threshold levels $\lambda_i \geq 0$ is

$$R^*_T(B^{\alpha}_{p,q}(M)) = \inf_{\lambda_j \geq 0, 1 \leq L_j \leq 2^{jv}} \sup_{\theta \in B^{\alpha}_{p,q}(M)} E \sum_{j=j_0}^{\infty} \|\widehat{\underline{\theta}}_j(\lambda_j, L_j) - \underline{\theta}_j\|_2^2. \tag{19}$$

We shall call $R^*_T(B^{\alpha}_{p,q}(M))$ the minimax block thresholding risk. It is clear that $R^*_T(B^{\alpha}_{p,q}(M)) \geq R^*(B^{\alpha}_{p,q}(M))$. On the other hand, Theorems 2 and 3 below show $R^*_T(B^{\alpha}_{p,q}(M))$ is within a small constant factor of the minimax risk $R^*(B^{\alpha}_{p,q}(M))$.

The following theorem shows that the SureBlock estimator adaptively attains the exact minimax block thresholding risk $R^*_T(B^{\alpha}_{p,q}(M))$ asymptotically over a wide range of Besov bodies.

Theorem 1 *Suppose we observe the sequence model (16). Let $\hat{\theta}^*$ be the SureBlock estimator of θ defined in (11) and (12). Then*

$$\sup_{\theta \in B^{\alpha}_{p,q}(M)} E_\theta\|\widehat{\theta}^* - \theta\|_2^2 \leq R^*_T(B^{\alpha}_{p,q}(M))\,(1 + o\,(1)) \tag{20}$$

for $1 \leq p,\ q \leq \infty$, $0 < M < \infty$, and $\alpha > \frac{1}{1-v}((\frac{1}{p} - \frac{1}{2})_+ + \frac{1}{2}) - \frac{1}{2}$.

Theorems 2 and 3 below make it clear that the SureBlock procedure is indeed nearly optimally adaptive over a wide collection of Besov bodies $B^{\alpha}_{p,q}(M)$ including both the dense ($p \geq 2$)and sparse ($p < 2$) cases. The estimator is asymptotically sharp adaptive over Besov bodies with $p = q = 2$ in the sense that it adaptively attains both the optimal rate and optimal constant. Over Besov bodies with $p \geq 2$ and $q \geq 2$ SureBlock adaptively achieves

within a factor of 1.25 of the minimax risk. At the same time the maximum risk of the estimator is simultaneously within a constant factor of the minimax risk over a collection of Besov bodies $B^{\alpha}_{p,q}(M)$ in the sparse case of $p < 2$.

Theorem 2 *(i). The SureBlock estimator is adaptively sharp minimax over Besov bodies* $B^{\alpha}_{2,2}(M)$ *for all* $M > 0$ *and* $\alpha > \frac{v}{2(1-v)}$. *That is,*

$$\sup_{\theta \in B^{\alpha}_{2,2}(M)} E_{\theta} \|\widehat{\theta}^{*} - \theta\|_2^2 \leq R^{*}(B^{\alpha}_{2,2}(M)) \, (1 + o\,(1)) \tag{21}$$

(ii). The SureBlock estimator is adaptively, asymptotically within a factor of 1.25 of the minimax risk over Besov bodies $B^{\alpha}_{p,q}(M)$,

$$\sup_{\theta \in B^{\alpha}_{p,q}(M)} E_{\theta} \|\widehat{\theta}^{*} - \theta\|_2^2 \leq 1.25 R^{*}(B^{\alpha}_{p,q}(M)) \, (1 + o\,(1)) \tag{22}$$

for all $p \geq 2$, $q \geq 2$, $M > 0$ *and* $\alpha > v/2(1-v)$.

For the sparse case $p < 2$, the SureBlock estimator is also simultaneously within a small constant factor of the minimax risk.

Theorem 3 *The SureBlock estimator is asymptotically minimax up to a constant factor* $G\,(p \wedge q)$ *over a large range of Besov bodies with* $1 \leq p, q \leq \infty$, $0 < M < \infty$, *and* $\alpha > \frac{1}{1-v}((\frac{1}{p} - \frac{1}{2})_{+} + \frac{1}{2}) - \frac{1}{2}$. *That is,*

$$\sup_{\theta \in B^{\alpha}_{p,q}(M)} E_{\theta} \|\widehat{\theta}^{*} - \theta\|_2^2 \leq G\,(p \wedge q) \cdot R^{*}(B^{\alpha}_{p,q}(M)) \, (1 + o\,(1)) \tag{23}$$

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

4.1 Analysis of a Single Resolution Level

The proof of the main results given above relies on a detailed risk analysis of the SureBlock procedure on a single resolution level. In particular we derive the following upper bound for the risk of SureBlock at a given resolution level. The result may also be of independent interest for the problem of estimation of a multivariate normal mean.

Consider the sequence model (5) where $x_i = \theta_i + z_i$, $i = 1, ..., d$ and z_i are independent normal $N(0,1)$ variables. Set $r(\lambda, L) = d^{-1} E \|\hat{\theta}(\lambda, L) - \theta\|_2^2$ and let

$$R\,(\theta) = \inf_{0 \leq \lambda, 1 \leq L \leq d^{v}} r\,(\lambda, L) = \inf_{\max\{L-2,0\} \leq \lambda, 1 \leq L \leq d^{v}} r\,(\lambda, L) \tag{24}$$

and

$$R_F\,(\theta) = r\,(2 \log d, 1)\,. \tag{25}$$

The following theorem gives upper bounds of the risk of the SureBlock estimator.

Theorem 4 *Let the data* $\{x_i, i = 1, 2, ..., d\}$ *be given as in (5) and let* $\hat{\theta}^*$ *be the SureBlock estimator defined in (11) and (12). Set* $\mu_d = \|\theta\|_2^2/d$ *and* $\gamma_d = d^{-\frac{1}{2}} \log_2^{\frac{3}{2}} d$. *Then*

(a). for any constant $\eta > 0$ *and uniformly in* $\theta \in \mathbb{R}^d$,

$$d^{-1} E_\theta \|\hat{\theta}^* - \theta\|_2^2 \leq R(\theta) + R_F(\theta) I(\mu_d \leq 3\gamma_d) + c_\eta d^{\eta - \frac{1}{2} + \frac{v}{2}}; \tag{26}$$

(b). for any constant $\eta > 0$ *and uniformly in* $\mu_d \leq \frac{1}{3}\gamma_d$,

$$d^{-1} E_\theta \|\hat{\theta}^* - \theta\|_2^2 \leq R_F(\theta) + c_\eta d^{-1-\eta} \tag{27}$$

where c_η *is a constant depending only on* η.

In the proof of Theorem 1, we will see that $R(\theta)$, the ideal risk, is the dominating term and all other terms in Equations (26) and (27) are negligible. This result shows that the SureBlock procedure mimics the performance of the oracle procedure where the ideal threshold level and the ideal block size are given.

Theorem 4 can be regarded as a generalization of Theorem 4 of Donoho and Johnstone(1995) from a fixed block size of one to variable block size. This generalization is important because it enables the resulting SureBlock estimator to be not only adaptively rate optimal over a wide collection of Besov bodies across both the dense ($p \geq 2$) and sparse ($p < 2$) cases, but also sharp adaptive over spaces where linear estimators can be asymptotically minimax. This property is not shared by fixed block size procedures such as VisuShrink, SureShrink or BlockJS. The theoretical advantages of the SureBlock estimator over these fixed block size procedures are also confirmed by the numerical results given in the next section.

5 Implementation And Numerical Results

We consider in this section the finite sample performance of the SureBlock estimator. The SureBlock procedure is easy to implement. The following result is useful for numerical implementation as well as the derivation of the theoretical results of the SureBlock estimator.

Proposition 1 *Let* x_i, $i = 1, ..., d$ *and* $SURE(x, \lambda, L)$ *be given as in (5) and (9), respectively. Let the block size* L *be given. Then the minimizer* λ *of* $SURE(x, \lambda, L)$ *is an element of the set* A *where*

$$A \triangleq \left\{x_i^2;\ 1 \leq i \leq d\right\} \cup \{0\}, \quad \text{if } L = 1$$

and

$$A \triangleq \left\{S_i^2\ ;\ S_i^2 \geq L - 2, 1 \leq i \leq m\right\} \cup \{L - 2\}, \text{ if } L \geq 2.$$

Proposition 1 shows that for a given block size L it is sufficient to search over the finite set A for the threshold λ which minimizes $SURE\,(x, \lambda, L)$.

The SureBlock procedure first empirically chooses the optimal threshold and block size (λ^*, L^*) as in (10). The values of (λ^*, L^*) depend on the choice of the parameter v which determines the search range for the block size L. In our simulation studies we choose $v = \frac{3}{4}$. That is,

$$(\lambda^*, L^*) = \underset{\max\{L-2,0\}\le\lambda\le\lambda^F, 1\le L\le d^{\frac{3}{4}}}{\arg\min} SURE\,(x, \lambda, L)\,. \tag{28}$$

The theoretical results given in the next section show that this procedure is adaptively within a constant factor of the minimax risk over a wide collection of Besov bodies.

The noise level σ is assumed to be known in Section 3. In practice σ needs to be estimated from the data. We use the following robust estimator of σ given in Donoho and Johnstone (1994a). The estimator $\hat{\sigma}$ is based on the empirical wavelet coefficients at the highest resolution level,

$$\hat{\sigma} = \frac{1}{.6745}\text{median}(|\tilde{y}_{J-1,k}| : 1 \le k \le 2^{J-1}).$$

In this section we compare the numerical performance of SureBlock with those of VisuShrink (Donoho and Johnstone (1994a)), SureShrink (Donoho and Johnstone (1995)) and BlockJS (Cai (1999)). VisuShrink thresholds empirical wavelet coefficients individually with a fixed threshold level. SureShrink is a term by term thresholding procedure which selects the threshold at each resolution level by minimizing Stein's unbiased risk estimate. In the simulation, we use the hybrid method proposed in Donoho and Johnstone (1995). BlockJS is a block thresholding procedure with a fixed block size $\log n$ and a fixed threshold level. Each of these wavelet estimators has been shown to perform well numerically as well as theoretically. For further details see the original papers.

Six test functions, representing different level of spatial variability, and various sample sizes, wavelets and signal to noise ratios are used for a systematic comparison of the four wavelet procedures. The test functions are plotted in Figure 3 in Appendix. Sample sizes ranging from $n = 256$ to $n = 16384$ and signal-to-noise ratios (SNR) from 3 to 7 were considered. The SNR is the ratio of the standard deviation of the function values to the standard deviation of the noise. Different combinations of wavelets and signal-to-noise ratios yield basically the same results. For reasons of space, we only report in detail the results for one particular case, using Daubechies' compactly supported wavelet *Symmlet* 8 and SNR equal to 7. We use the software package WaveLab for simulations and the procedures MultiVisu (for VisuShrink) and MultiHybrid (for SureShrink) in WaveLab 802 are used (See http://www-stat.stanford.edu/~wavelab/).

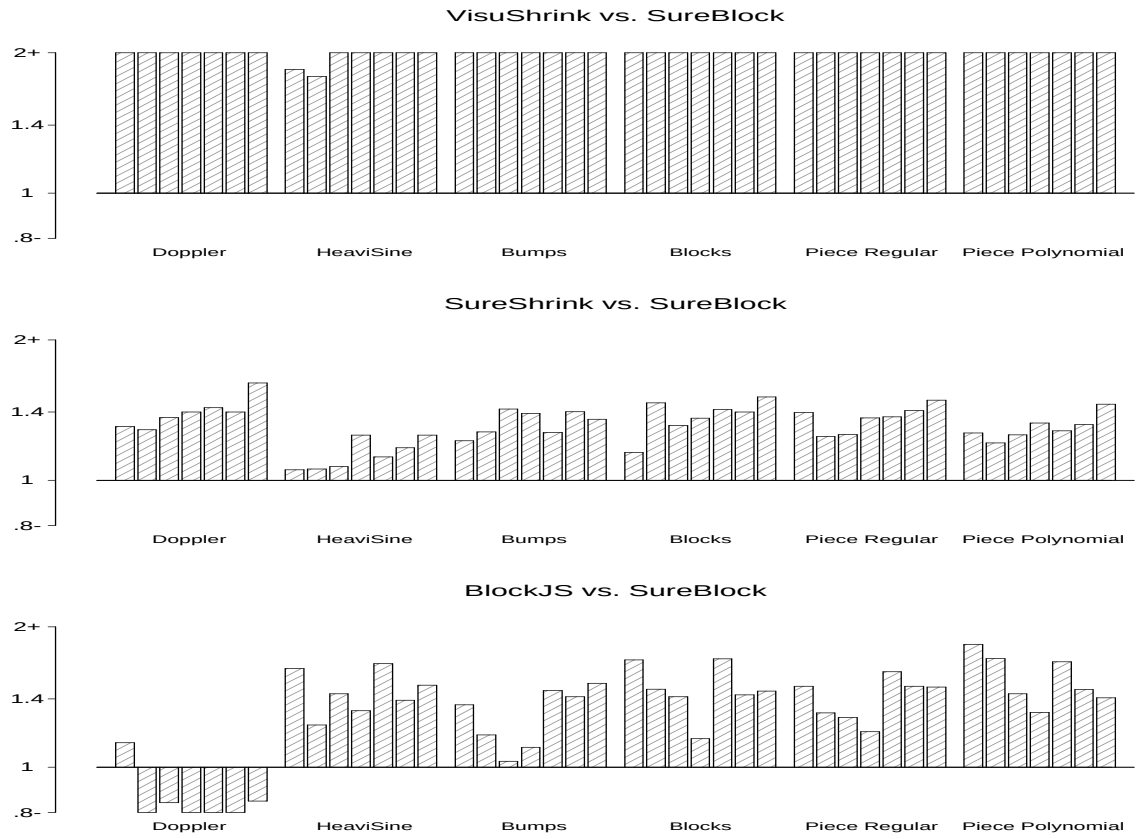

Figure 1: The vertical bars represent the ratios of the average squared errors of various estimators to the corresponding average squared error of SureBlock. The higher the bar the better the relative performance of SureBlock. The bars are plotted on a log scale and are truncated at the value 2 of the original ratio. For each signal the bars are ordered from left to right by the sample sizes ($n = 256$ to 16382).

Table 1 below reports the average squared errors (rounded to three significant digits) over 50 replications for the four thresholding estimators. A graphical presentation is given in Figure 1. SureBlock consistently outperforms both VisuShrink and SureShrink in all 42 simulation cases in terms of the average squared error. SureBlock procedure is better than BlockJS about 88% of times (37 out of 42 cases). SureBlock fails to dominate BlockJS only for the test function "Doppler". For $n = 16384$ the risk ratio of SureShrink to BlockJS is $0.021/0.011 \approx 2$. However, SureBlock is almost as good as BlockJS with a risk ratio $0.012/0.011 \approx 1$. Although SureBlock does not dominate BlockJS for the test function "Doppler" , the improvement of SureBlock over BlockJS is significant for other test functions.

The simulation results show that, by empirically choosing the block size and threshold and allowing them to vary from resolution level to resolution level, the SureBlock estimator has significant numerical advantages over thresholding estimators with fixed block size $L = 1$ (VisuShrink or SureShrink) or $L = \log n$ (BlockJS). These numerical findings reconfirm the theoretical results given in Section 4.

n	VS	SS	BJS	SB	n	VS	SS	BJS	SB
Blocks					*Bumps*				
256	2.635	0.669	0.990	0.583	256	3.360	0.773	0.865	0.636
512	2.143	0.705	0.707	0.481	512	2.758	0.618	0.571	0.487
1024	1.576	0.438	0.473	0.334	1024	1.856	0.537	0.389	0.378
2048	1.038	0.341	0.289	0.251	2048	1.205	0.299	0.237	0.215
4096	0.676	0.204	0.246	0.144	4096	0.745	0.176	0.203	0.139
8192	0.434	0.140	0.143	0.100	8192	0.460	0.108	0.109	0.077
16382	0.279	0.086	0.083	0.057	16382	0.270	0.050	0.056	0.037
Doppler					*HeaviSine*				
256	1.512	0.468	0.405	0.359	256	0.420	0.240	0.371	0.228
512	0.953	0.390	0.208	0.304	512	0.285	0.169	0.197	0.160
1024	0.633	0.237	0.146	0.174	1024	0.207	0.093	0.125	0.087
2048	0.403	0.154	0.074	0.110	2048	0.149	0.070	0.074	0.056
4096	0.222	0.073	0.036	0.051	4096	0.081	0.037	0.055	0.033
8192	0.127	0.035	0.020	0.025	8192	0.058	0.027	0.032	0.023
16382	0.079	0.021	0.011	0.013	16382	0.038	0.015	0.018	0.012
Piece-Polynomial					*Piece-Regular*				
256	2.400	0.735	1.067	0.582	256	2.265	0.839	0.896	0.601
512	1.687	0.518	0.737	0.431	512	1.503	0.530	0.558	0.427
1024	1.195	0.369	0.424	0.295	1024	0.978	0.328	0.335	0.262
2048	0.805	0.264	0.261	0.199	2048	0.705	0.227	0.199	0.167
4096	0.495	0.157	0.207	0.123	4096	0.419	0.134	0.157	0.098
8192	0.332	0.104	0.116	0.079	8192	0.268	0.086	0.091	0.061
16382	0.207	0.064	0.062	0.044	16382	0.169	0.049	0.049	0.033

Table 1: Average squared errors over 50 replications with $SNR = 7$. In the table VS stands for VisuShrink, SS for SureShrink, BJS for BlockJS and SB for SureBlock.

In addition to the comparison of the overall global performance of the wavelet estimators, it is also instructive to compare the performance at individual resolution levels and to examine the value in the empirical selection of block sizes. Table 2 gives the average

squared errors at individual resolution levels over 50 replications for Bumps with $n = 1024$ and $SNR = 7$. Table 2 compares the errors for SureBlock, SureShrink and an estimator, called SureGarrote, which empirically chooses the threshold λ at each level but fixes the block size $L = 1$. SureBlock consistently outperforms SureGarrote and SureShrink at each resolution level. Both SureGarrote and SureShrink have block size $L = 1$, the results indicate that the Non-negative garrote shrinkage is slightly better than soft thresholding in this example.

Resolution Level j	3	4	5	6	7	8	9
SureBlock	0.007	0.018	0.029	0.044	0.062	0.077	0.138
SureGarrote	0.026	0.022	0.029	0.050	0.069	0.081	0.139
SureShrink	0.036	0.043	0.030	0.053	0.080	0.101	0.195

Table 2: Average squared errors at individual resolution levels

Figure 2 shows a typical result of the SureBlock procedure applied to the noisy Bumps signal. The left panel is the noisy signal; the middle panel is a display of the empirical wavelet coefficients arranged according resolution levels; and the right panel is the SureBlock reconstruction (solid line) and the true signal (dotted line). In this example the block sizes chosen by SureBlock are 2, 3, 1, 5, 3, 5 and 1 from the resolution level $j = 3$ to level $j = 9$.

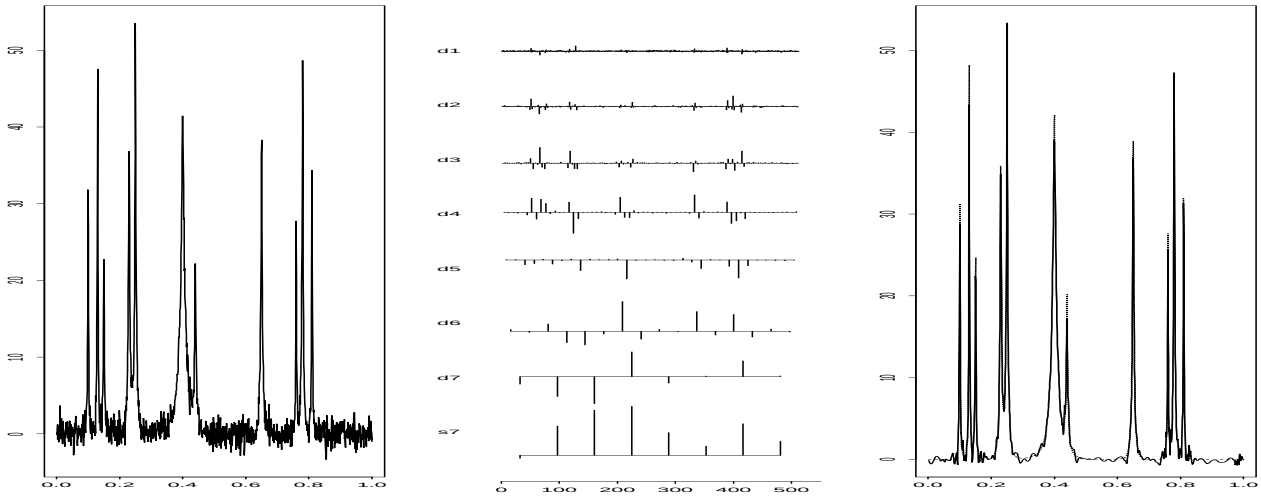

Figure 2: SureBlock procedure applied to a noisy Bumps signal.

6 Proofs

Throughout this section, without loss of generality, we shall assume the noise level $\sigma = 1$. We first prove Theorem 4 and then use it as the main tool to prove Theorem 1. The proofs of Theorems 2 and 3 and Proposition 1 are given later.

6.1 Notation And Preparatory Results

Before proving the main theorems, we need to introduce some notation and collect a few technical results. The proofs of some of these preparatory results are long and tedious. For reasons of space these proofs are omitted here. We refer interested readers to Cai and Zhou (2005) for the complete proofs.

Consider the sequence model (5) with $\sigma = 1$. For a given block size L and threshold level λ, set $r_b(\lambda, L) = E_{\underline{\theta}_b} \|\widehat{\underline{\theta}}_b(\lambda, L) - \underline{\theta}_b\|^2$ and define

$$r(\lambda, L) = \frac{1}{d} \sum_{b=1}^{m} r_b(\lambda, L) = ED(\lambda, L)$$

where $D(\lambda, L) = \frac{1}{d} \sum_{b=1}^{m} \|\widehat{\underline{\theta}}_b(\lambda, L) - \underline{\theta}_b\|_2^2$. Set

$$\widetilde{R}(\theta) = \inf_{\lambda \le \lambda^F, 1 \le L \le d^v} r(\lambda, L) = \inf_{\max\{L-2,0\} \le \lambda \le \lambda^F, 1 \le L \le d^v} r(\lambda, L). \tag{29}$$

The difference between $\tilde{R}(\theta)$ and $R(\theta)$ defined in (24) is that the search range for the threshold λ in $\tilde{R}(\theta)$ is restricted to be at most λ^F. The result given below shows 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 $\theta \in R^d$,*

$$\widetilde{R}(\theta) - R(\theta) \le C_a d^{a - \frac{1}{2}}.$$

The following simple lemma is adapted from Donoho and Johnstone (1995) and is used the proof of Theorem 4.

Lemma 2 *Let $T_d = d^{-1} \sum (x_i^2 - 1)$ and $\mu_d = d^{-1}\|\theta\|_2^2$. If $\gamma_d^2 d / \log d \to \infty$, then*

$$\sup_{\mu_d \ge 3\gamma_d} (1 + \mu_d) P(T_d \le \gamma_d) = o\left(d^{-\frac{1}{2}}\right).$$

We also need the following bounds for the loss of the SureBlock estimator and the derivative of the risk function $r_b(\lambda, L)$. The first bound is used in the proof of Theorem 1 and the second in the proof of Theorem 4.

Lemma 3 *Let* $\{x_i : i = 1, ..., d\}$ *be given as in (5) with* $\sigma = 1$. *Then*

$$\|\hat{\theta}^* - \theta\|_2^2 \leq 2\lambda^F d + 2\|z\|_2^2 \tag{30}$$

and for $\lambda > 0$ *and* $L \geq 1$,

$$\left|\frac{\partial}{\partial\lambda} r_b(\lambda, L)\right| < 6. \tag{31}$$

Finally we develop a key technical result for the proof of Theorem 4. Set

$$U(\lambda, L) \triangleq \frac{1}{d} SURE(x, \lambda, L) = 1 + \frac{1}{d}\sum_{b=1}^{m}\left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I\left(S_b^2 > \lambda\right) + \left(S_b^2 - 2L\right) I\left(S_b^2 \leq \lambda\right)\right).$$

Note that both $D(\lambda, L)$ and $U(\lambda, L)$ have expectation $r(\lambda, L)$.

The goal is to show that the minimizer $\left(\lambda^S, L^S\right)$ of Stein unbiased estimator of risk $U(\lambda, L)$ is asymptotically the ideal threshold level and block size. The key step is to show that the following difference

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

is negligible for $\max\{L-2, 0\} \leq \lambda \leq \lambda^F$ and $1 \leq L \leq d^v$. It is easy to verify that for any two functions g and h defined on the same domain, $|\inf_x g(x) - \inf_x h(x)| \leq \sup_x |g(x) - h(x)|$. Hence,

$$|U(\lambda^S, L^S) - \inf_{\lambda, L} r(\lambda, L)| = |\inf_{\lambda, L} U(\lambda, L) - \inf_{\lambda, L} r(\lambda, L)| \leq \sup_{\lambda, L} |U(\lambda, L) - r(\lambda, L)|$$

and consequently

$$\begin{aligned}
\Delta_d &\leq E\left|D\left(\lambda^S, L^S\right) - r\left(\lambda^S, L^S\right) + r\left(\lambda^S, L^S\right) - U\left(\lambda^S, L^S\right) + U\left(\lambda^S, L^S\right) - \inf_{\lambda, L} r(\lambda, L)\right| \\
&\leq E\sup_{\lambda, L}|D(\lambda, L) - r(\lambda, L)| + 2E\sup_{\lambda, L}|r(\lambda, L) - U(\lambda, L)|. \qquad (32)
\end{aligned}$$

The following proposition provides the upper bounds for the two terms on the RHS of (32) and is a key technical result for the proof of Theorem 4.

Proposition 2 *Uniformly in* $\theta \in \mathbb{R}^d$ *, for* $0 \leq v < 1$ *we have*

$$E_\theta \sup_{\max\{L-2,0\} \leq \lambda \leq \lambda^F, 1 \leq L \leq d^v} |U(\lambda, L) - r(\lambda, L)| = o\left(d^{\eta - \frac{1}{2} + \frac{v}{2}}\right) \tag{33}$$

$$E_\theta \sup_{\max\{L-2,1/\log d\} \leq \lambda \leq \lambda^F, 1 \leq L \leq d^v} |D(\lambda, L) - r(\lambda, L)| = o\left(d^{\eta - \frac{1}{2} + \frac{v}{2}}\right) \tag{34}$$

for any $1 - v > \eta > 0$, *where* $\lambda^F = 2L\log d$.

In the proofs we will denote by C a generic constant that may vary from place to place.

6.2 Proof of Theorem 4

The proof of Theorem 4 is similar to but more involved than those of Theorem 4 of Donoho and Johnstone(1995) and Theorem 2 of Johnstone (1999) because of variable block size.

Without loss of generality, we may assume $\lambda \geq 1/\log d$ since equation (31) implies

$$R(\theta) \leq \inf_{1/\log d \leq \lambda, 1 \leq L \leq d^v} r(\lambda, L) \leq (1 + C/\log d) R(\theta),$$

which means that the ideal risk over $\lambda \geq 1/\log d$ is asymptotically the same as the ideal risk over all $\lambda \geq 0$. Correspondingly, we shall also restrict the minimizer of $U(\lambda, L)$ searched over $[1/\log d, \infty)$ instead of over $[0, \infty)$ in the SureBlock procedure.

Set $T_d = d^{-1} \sum (x_i^2 - 1)$ and $\gamma_d = d^{-\frac{1}{2}} \log_2^{\frac{3}{2}} d$. Define the event $A_d = \{T_d \leq \gamma_d\}$ and decompose the risk of SureBlock procedure into two parts,

$$R_d(\theta) = d^{-1} E_\theta \|\widehat{\theta}^* - \theta\|_2^2 = R_{1,d}(\theta) + R_{2,d}(\theta)$$

where $R_{1,d}(\theta) = d^{-1} E_\theta \left\{ \|\widehat{\theta}^* - \theta\|_2^2 I(A_d) \right\}$ and $R_{2,d}(\theta) = d^{-1} E_\theta \left\{ \|\widehat{\theta}^* - \theta\|_2^2 I(A_d^c) \right\}$.

We first consider $R_{1,d}(\theta)$. On the event A_d, the signal is sparse and $\widehat{\theta}^*$ is the non-negative garrote estimator $\hat{\theta}_i^* = (1 - 2\log d / x_i^2)_+ x_i$ by the definition of $\widehat{\theta}^*$ in equation (12). Decomposing $R_{1,d}(\theta)$ further into two parts with either $\mu_d = d^{-1}\|\theta\|_2^2 \leq 3\gamma_d$ or $\mu_d > 3\gamma_d$ yields that

$$R_{1,d}(\theta) \leq R_F(\theta) I(\mu_d \leq 3\gamma_d) + r_{1,d}(\theta)$$

where $R_F(\theta)$ is the risk of the non-negative garrote estimator and

$$r_{1,d}(\theta) = d^{-1} E_\theta \left\{ \|\widehat{\theta}^* - \theta\|_2^2 I(A_d) \right\} I(\mu_d > 3\gamma_d).$$

Note that on the event A_d, $\|\hat{\theta}^*\|_2^2 \leq \|x\|_2^2 \leq d + d\gamma_d$ and so

$$r_{1,d}(\theta) \leq 2d^{-1}(E\|\widehat{\theta}^*\|_2^2 + \|\theta\|_2^2) P(A_d) I(\mu_d > 3\gamma_d) \leq 2(1 + 2\mu_d) P(A_d) = o(d^{-\frac{1}{2}}) \quad (35)$$

where the last step follows from Lemma 2.

Note that for any $\eta > 0$, Lemma 1 yields that for some constant $C_\eta > 0$

$$\widetilde{R}(\theta) - R(\theta) \leq C_\eta d^{\eta - \frac{1}{2}} \quad (36)$$

for all $\theta \in R^d$, and Equations (32) - (34) yield

$$R_{2,d}(\theta) - \widetilde{R}(\theta) \leq \Delta_d \leq C d^{\eta - \frac{1}{2} + \frac{v}{2}}. \quad (37)$$

The proof of part (a) of the theorem is completed by putting together Equations (35)-(37).

We now turn to part (b). Note first that $R_{1,d}(\theta) \le R_F(\theta)$. On the other hand,

$$R_{2,d}(\theta) = d^{-1}E\left\{\|\widehat{\theta}^* - \theta\|_2^2 I\left(A_d^c\right)\right\} \le Cd^{-1}\left(E\|\widehat{\theta}^* - \theta\|_2^4\right)^{\frac{1}{2}} P^{\frac{1}{2}}\left(A_d^c\right).$$

To complete the proof of (27), it then suffices to show that under the assumption $\mu_d \le \frac{1}{3}\gamma_d$, $E\|\widehat{\theta}^* - \theta\|_2^4$ is bounded by a polynomial of d and $P\left(A_d^c\right)$ decays faster than any polynomial of d^{-1}. Note that in this case $\|\theta\|_2^2 = d\mu_d \le d^{\frac{1}{2}}\log_2^{\frac{3}{2}} d$. Since $\|\hat{\theta}^*\|_2^2 \le \|x\|_2^2$ and $x_i = \theta_i + z_i$,

$$\begin{aligned} E\|\widehat{\theta}^* - \theta\|_2^4 &\le E\left(2\|\widehat{\theta}^*\|_2^2 + 2\|\theta\|_2^2\right)^2 \le E\left(2\|x\|_2^2 + 2\|\theta\|_2^2\right)^2 \\ &\le E\left(4\|\theta\|_2^2 + 2\|z\|_2^2\right)^2 \le 32\|\theta\|_2^4 + 8E\|z\|_2^4 \\ &\le 32d\log_2^3 d + 16d + 8d^2. \end{aligned}$$

On the other hand, it follows from Hoeffding's inequality and Mill's inequality that

$$\begin{aligned} P\left(A_d^c\right) &= P\left(d^{-1}\sum\left(z_i^2 + 2z_i\theta_i + \theta_i^2 - 1\right) > d^{-\frac{1}{2}}\log_2^{\frac{3}{2}} d\right) \\ &\le P\left(d^{-1}\sum(z_i^2 - 1) > \frac{1}{3}d^{-\frac{1}{2}}\log_2^{\frac{3}{2}} d\right) + P\left(d^{-1}\sum 2z_i\theta_i > \frac{1}{3}d^{-\frac{1}{2}}\log_2^{\frac{3}{2}} d\right) \\ &\le 2\exp\left(-C\log_2^3 d\right) + \frac{1}{2}\exp\left(-C\log_2^3 d/\mu_d\right) \end{aligned}$$

which decays faster than any polynomial of d^{-1}. ∎

6.3 Proof of Theorem 1

We now use Theorem 4 as one of the main technical tools to prove Theorem 1. Again we set $\sigma = 1$. To more conveniently use the results given in Theorem 4, we shall renormalize the model (16) so that the noise level is 1. Multiplying by a factor of $n^{\frac{1}{2}}$ on both sides, the sequence model (16) becomes

$$y'_{j,k} = \theta'_{j,k} + z_{j,k}, \quad j \ge j_0,\ k = 1, 2, \cdots, 2^j \tag{38}$$

where $y'_{j,k} = n^{\frac{1}{2}} y_{j,k}$ and $\theta'_{j,k} = n^{\frac{1}{2}}\theta_{j,k}$. Note that $E_\theta\|\hat{\theta}^* - \theta\|_2^2 = n^{-1}E_{\theta'}\|\hat{\theta}'^* - \theta'\|_2^2$ where the expectation on the left side is under model (16) and the right side under the renormalized model (38). We shall use the sequence model (38) in the proof below.

Fix $0 < \varepsilon_0 < 1/(1+2\alpha)$ and let J_0 be the largest integer satisfying $2^{J_0} \le n^{\varepsilon_0}$. Write

$$\begin{aligned} n^{-1}E_{\theta'}\|\hat{\theta}'^* - \theta'\|_2^2 &= \left(\sum_{j\le J_0}\sum_k + \sum_{J_0\le j<J_1}\sum_k + \sum_{j\ge J_1}\sum_k\right) n^{-1}E_{\theta'}\|\hat{\theta}'^* - \theta'\|_2^2 \\ &= S_1 + S_2 + S_3 \end{aligned}$$

where $J_1 > J_0$ is to be chosen later. Since $0 < \varepsilon_0 < 1/(1+2\alpha)$, $n^{\varepsilon_0} n^{-1} \log n = o\left(n^{-2\alpha/(1+2\alpha)}\right)$. It then follows from equation (30) in Lemma 3 that

$$S_1 \leq C\lambda^F n^{\varepsilon_0} n^{-1} = o\left(n^{-\frac{2\alpha}{1+2\alpha}}\right)$$

which is negligible relative to the minimax risk. On the other hand, it follows from (26) in Theorem 4 that

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

where $\mu'_{2^j} = 2^{-j}\|\underline{\theta}'_j\|_2^2 = 2^{-j} n \|\underline{\theta}_j\|_2^2$ and $\gamma_{2^j} = 2^{-\frac{j}{2}} j^{\frac{3}{2}}$. It is obvious that

$$S_{21} = \sum_{J_0 \leq j < J_1} n^{-1} 2^j R(\underline{\theta}'_j) \leq R_T^*(B_{p,q}^\alpha(M)) \tag{39}$$

and $R_T^*(B_{p,q}^\alpha(M)) \geq R^*(B_{p,q}^\alpha(M)) \geq C n^{-2\alpha/(1+2\alpha)}$ for some constant $C > 0$. We shall see that both S_{22} and S_{23} are negligible relative to the minimax risk. Note that

$$S_{23} = \sum_{J_0 \leq j < J_1} n^{-1} c_\eta 2^{j(\eta+1/2+v/2)} \leq C n^{-1} 2^{J_1(\eta+1/2+v/2)}. \tag{40}$$

On the other hand, it follows from the Oracle Inequality (3.10) in Cai (1999) with $L = 1$ and $\lambda = \lambda^F = 2\log(2^j)$ that

$$2^j R_F(\underline{\theta}'_j) \leq \sum_{j=1}^{2^j} (n\theta_{j,k}^2 \wedge \lambda^F) + 8(2\log 2^j)^{-\frac{1}{2}} 2^{-j} \leq n\|\underline{\theta}_j\|_2^2 + 8(2\log 2^j)^{-\frac{1}{2}} 2^{-j}. \tag{41}$$

Recall that $\mu'_{2^j} = 2^{-j} n \|\underline{\theta}_j\|_2^2$ and $\gamma_{2^j} = 2^{-\frac{j}{2}} j^{\frac{3}{2}}$. It then follows from (41) that

$$\begin{aligned} S_{22} &= \sum_{J_0 \leq j < J_1} n^{-1} 2^j R_F(\underline{\theta}'_j) I(\mu'_{2^j} \leq 3\gamma_{2^j}) \leq \sum_{J_0 \leq j < J_1} \left(3n^{-1} 2^{\frac{j}{2}} j^{\frac{3}{2}} + 8n^{-1}(2\log 2^j)^{-\frac{1}{2}} 2^{-j}\right) \\ &\leq C n^{-1} 2^{J_1\left(\eta+\frac{1}{2}\right)} J_1^{\frac{3}{2}}. \end{aligned} \tag{42}$$

Hence if J_1 satisfies $2^{J_1} = n^\gamma$ with some $\gamma < \frac{1}{(1+2\alpha)(\eta+1/2+v/2)}$, then (40) and (42) yield

$$S_{22} + S_{23} = o\left(n^{-\frac{2\alpha}{1+2\alpha}}\right). \tag{43}$$

We now turn to the term S_3. It is easy to check that for $\theta \in B_{p,q}^\alpha(M)$, $\|\underline{\theta}_j\|_2^2 \leq M^2 2^{-2\alpha' j}$ where $\alpha' = \alpha - (\frac{1}{p} - \frac{1}{2})_+ > 0$. Note that if J_1 satisfies $2^{J_1} = n^\gamma$ for some $\gamma > \frac{1}{1/2+2\alpha'}$, then

for all sufficiently large n and all $j \geq J_1$, $\mu'_{2^j} \leq \frac{1}{3}\gamma_{2^j}$ where $\gamma_{2^j} = 2^{-\frac{j}{2}} j^{\frac{3}{2}}$. It thus follows from (27) and (41) that whenever $\alpha > 2(\frac{1}{p} - \frac{1}{2})_+$,

$$\begin{aligned} S_3 &= \sum_{j \geq J_1} \sum_k n^{-1} E_{\theta'} (\hat{\theta}'^*_{j,k} - \theta'_{j,k})^2 \leq \sum_{j \geq J_1} \left(n^{-1} 2^j R_F(\underline{\theta}'_j) + c_\eta n^{-1} 2^{-j\eta} \right) \\ &\leq \sum_{j \geq J_1} \|\underline{\theta}_j\|_2^2 + Cn^{-1} \leq C 2^{-2\alpha' J_1} + Cn^{-1} = o(n^{-\frac{2\alpha}{1+2\alpha}}). \end{aligned} \tag{44}$$

For $\alpha > \frac{1}{1-v}\left(2(\frac{1}{p} - \frac{1}{2})_+ + \frac{1}{2}\right) - \frac{1}{2}$, both Equations (43) and (44) hold, and hence the proof is completed by choosing γ satisfying

$$\frac{1}{1/2 + 2\left(\alpha - (1/p - 1/2)_+\right)} < \gamma < \frac{1}{(1+2\alpha)\left(\eta + 1/2 + v/2\right)}.$$

This is always possible by choosing $\eta > 0$ sufficiently small. ∎

6.4 Proof of Theorem 2

Define the minimax linear risk by $R^*_L(B^\alpha_{p,q}(M)) = \inf_{\hat{\theta}\,\text{linear}} \sup_{\theta \in B^\alpha_{p,q}(M)} E\|\hat{\theta} - \theta\|_2^2$. It follows from Donoho, Liu and MacGibbon (1990) that

$$\begin{aligned} R^*_L(B^\alpha_{p,q}(M)) &= \sup_{\theta \in B^\alpha_{p,q}(M)} \sum_{j=j_0}^{\infty} \sum_k \left(\frac{\theta^2_{j,k}/n}{\theta^2_{j,k} + 1/n} \right), \\ R^*_L(B^\alpha_{p,q}(M)) &= R^*(B^\alpha_{p,q}(M))(1 + o(1)) \quad \text{for } p = q = 2, \end{aligned}$$

and $R^*_L(B^\alpha_{p,q}(M)) \leq 1.25 R^*(B^\alpha_{p,q}(M))(1 + o(1))$, for all $p > 2$ and $q > 2$.

It therefore suffices to establish that the SureBlock procedure asymptotically attains the minimax linear risk, i.e.,

$$\sup_{\theta \in B^\alpha_{p,q}(M)} E_\theta \|\hat{\theta}^* - \theta\|_2^2 \leq R^*_L(B^\alpha_{p,q}(M))(1 + o(1)), \quad \text{for } \alpha > 0,\ p \geq 2 \text{ and } q \geq 2.$$

Recall that in the proof of Theorem 1 it is shown that

$$E_\theta \|\hat{\theta}^* - \theta\|_2^2 \leq S_1 + S_{21} + S_{22} + S_{23} + S_3,$$

where $S_1 + S_{22} + S_{23} + S_3 = o(n^{-2\alpha/(2\alpha+1)})$ and $S_{21} = \sum_{J_0 \leq j < J_1} n^{-1} 2^j R(\underline{\theta}'_j)$ with J_0 and J_1 chosen as in the proof of Theorem 1. Since the minimax risk $R^*(B^\alpha_{p,q}(M)) \asymp n^{-2\alpha/(2\alpha+1)}$, this implies that S_{21} is the dominating term in the maximum risk of the SureBlock procedure. It follows from the definition of $R(\underline{\theta}'_j)$ given in (24) that

$$n^{-1} 2^j R(\underline{\theta}'_j) \leq n^{-1} \sum_b E_{\underline{\theta}'_b} \|\hat{\underline{\theta}}'_b(L_j - 2, L_j) - \underline{\theta}'_b\|_2^2$$

where the RHS is the risk of the blockwise James-Stein estimator with any fixed block size $1 \le L_j \le 2^{jv}$ where $0 \le v < 1$ and a fixed threshold level $L_j - 2$.

Stein's unbiased risk estimate (see for example Lehmann (1983, page 300)) yields that

$$n^{-1}\sum_b E_{\underline{\theta}'_b}\|\underline{\hat{\theta}}'_b(L_j-2,L_j)-\underline{\theta}'_b\|_2^2 \le \sum_b\left(\frac{\|\underline{\theta}_b\|^2 L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}+\frac{2}{n}\right).$$

Hence the maximum risk of SureBlock procedure satisfies

$$\begin{aligned}\sup_{\theta\in B^\alpha_{p,q}(M)} E_\theta\|\hat{\theta}^*-\theta\|_2^2 &\le \sup_{\theta\in B^\alpha_{p,q}(M)}\sum_{J_0\le j<J_1}\sum_b\left(\frac{\|\underline{\theta}_b\|^2 L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}+\frac{2}{n}\right)\cdot(1+o(1))\\ &= \sup_{\theta\in B^\alpha_{p,q}(M)}\left(\sum_{J_0\le j<J_1}\sum_b\frac{\|\underline{\theta}_b\|_2^2 L_j/n}{\|\underline{\theta}_b\|_2^2+L_j/n}+2\sum_{J_0\le j<J_1}\frac{2^j}{nL_j}\right)\cdot(1+o(1)).\end{aligned}$$

Note that in the proof of Theorem 1, J_1 satisfies$2^{J_1}=n^\gamma$ with $\gamma<\frac{1}{(1+2\alpha)(\eta+1/2+v/2)}$. Hence if L_j satisfies $2^{j\rho}\le L_j\le 2^{jv}$ for some $\rho>\frac{1}{2}-\frac{v}{2}-\eta$, then

$$\sum_{j\le J_1}\frac{2^j}{nL_j}\le\frac{1}{n}2\cdot 2^{J_1(\frac{1}{2}+\frac{v}{2}+\eta)}=o\left(n^{-\frac{2\alpha}{1+2\alpha}}\right)$$

and hence

$$\sup_{\theta\in B^\alpha_{p,q}(M)} E_\theta\|\hat{\theta}^*-\theta\|_2^2\le\sup_{\theta\in B^\alpha_{p,q}(M)}\sum_{J_0\le j<J_1}\sum_b\left(\frac{\|\underline{\theta}_b\|^2 L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}\right)\cdot(1+o(1)).$$

To complete the proof it remains to show that

$$\sup_{\theta\in B^\alpha_{p,q}(M)}\sum_{J_0\le j<J_1}\sum_b\left(\frac{\|\underline{\theta}_b\|^2 L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}\right)\le R^*_L(B^\alpha_{p,q}(M))(1+o(1)). \tag{45}$$

Note that

$$\sum_{b=1}^{2^j/L_j}\frac{\|\underline{\theta}_b\|^2 L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}=\frac{L_j}{n}\left(\frac{2^j}{L_j}-\sum_{b=1}^{2^j/L_j}\frac{L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}\right)=\frac{2^j}{n}-\left(\frac{L_j}{n}\right)^2\sum_{b=1}^{2^j/L_j}\frac{1}{\|\underline{\theta}_b\|^2+L_j/n}.$$

Then the simple inequality $(\sum_{i=1}^m a_i)(\sum_{i=1}^m a_i^{-1})\ge m^2$, for $a_i>0,\ 1\le i\le m$ yields that

$$\begin{aligned}\sum_{J_0\le j<J_1}\sum_b\left(\frac{\|\underline{\theta}_b\|^2 L_j/n}{\|\underline{\theta}_b\|^2+L_j/n}\right) &\le \frac{2^j}{n}-\left(\frac{L_j}{n}\right)^2\left(\frac{2^j}{L_j}\right)^2\sum_{b=1}^{2^j/L_j}\left(\|\underline{\theta}_b\|^2+\frac{L_j}{n}\right)\\ &= \frac{2^j/n\sum_{b=1}^{2^j/L_j}\|\underline{\theta}_b\|^2}{\sum_{b=1}^{2^j/L_j}\|\underline{\theta}_b\|^2+2^j/n}=\frac{2^j/n\sum_k|\theta_{j,k}|^2}{\sum_k|\theta_{j,k}|^2+2^j/n}.\end{aligned}$$

Theorem 2 in Cai, Low and Zhao (2000) shows that

$$\sup_{\theta\in B^{\alpha}_{p,q}(M)}\sum_{J_0\le j<J_1}\frac{2^j/n\sum_k|\theta_{j,k}|^2}{\sum_k|\theta_{j,k}|^2+2^j/n}=R^*_L(B^{\alpha}_{p,q}(M))(1+o(1)),$$

and this proves inequality (45). ∎

6.5 Proof of Theorem 3

To establish the theorem, it suffices to show the following result which plays the role similar to that of Proposition 16 in Donoho and Johnstone (1994b).

Proposition 3 *Let $X\sim N(\mu,1)$ and let $\mathcal{F}_p(\eta)$ denote the probability measures $F(d\mu)$ satisfying the moment condition $\int|\mu|^p F(d\mu)\le\eta^p$. Let*

$$\overline{r}(\delta^g_\lambda,\eta)=\sup_{\mathcal{F}_p(\eta)}\left\{E_F r_g(\mu):\int|\mu|^p F(d\mu)\le\eta^p\right\},$$

where $r_g(\mu)=E_\mu(\delta^g_\lambda(x)-\mu)^2$ and $\delta^g_\lambda(x)=\left(1-\frac{\lambda^2}{x^2}\right)_+x$. Let $p\in(0,2)$ and $\lambda=\sqrt{2\log\eta^{-p}}$, then

$$\overline{r}(\delta^g_\lambda,\eta)\le 2\eta^p\lambda^{2-p}(1+o(1))\ \text{ as }\eta\to 0.$$

Proof of Proposition 3 : It follows from an analogous argument to the proof of Proposition 16 in Donoho and Johnstone (1994b) that

$$\overline{r}_g(\delta_\lambda,\eta)=\sup_{\mu\ge\eta}\left(\frac{\eta}{\mu}\right)^p r_g(\mu)(1+o(1)),\quad\text{as }\eta\to 0.$$

Stein's unbiased risk estimate yields that

$$\begin{aligned}r_g(\mu)&=E_\mu\left[1+\left(x^2-2\right)I(|x|\le\lambda)+\left(\frac{2\lambda^2}{x^2}+\frac{\lambda^4}{x^2}\right)I(|x|\ge\lambda)\right]\\&\le 1+\left(\lambda^2-2\right)P(|x|\le\lambda)+\left(2+\lambda^2\right)P(|x|\ge\lambda)\\&\le\lambda^2(1+o(1))\end{aligned}$$

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

$$\left(\frac{\eta}{\mu}\right)^p r_g(\mu)\le\eta^p\lambda^{2-p}(1+o(1)),\quad\text{as }\eta\to 0.\tag{46}$$

Similarly,

$$r_g(\mu)=E_\mu\left[x^2-1+\left(\frac{2\lambda^2}{x^2}+\frac{\lambda^4}{x^2}-x^2+2\right)I(|x|\ge\lambda)\right]\le\mu^2+4P(|x|\ge\lambda).$$

Now the simple inequality $P(|x| \geq \lambda) \leq \frac{2\phi(\lambda)}{\lambda} + \frac{\mu^2}{4} = \frac{\mu^2}{4} + o(\eta^p)$ yields that for $\eta \leq \mu < \lambda$

$$\left(\frac{\eta}{\mu}\right)^p r_g(\mu) \leq 2\eta^p \lambda^{2-p}(1 + o(1)), \quad \text{as } \eta \to 0. \tag{47}$$

The proposition now follows from (46) and (47).

We now return to the proof of Theorem 3. Set $\rho_g(\eta) \triangleq \inf_\lambda \sup_{\mathcal{F}_p(\eta)} E_F r_g(\mu)$. Then Proposition 3 implies that

$$\rho_g(\eta) \leq \overline{r}(\delta^g_\lambda, \eta) \leq 2\eta^p \left(2\log \eta^{-p}\right)^{(2-p)/2}(1 + o(1)), \text{ as } \eta \to 0.$$

For $p \in (0, 2)$, Theorem 18 of Donoho and Johnstone (1994b) shows the univariate Bayes-minimax risk satisfies

$$\rho(\eta) \triangleq \inf_\delta \sup_{\mathcal{F}_p(\eta)} E_F E_\mu (\delta(x) - \mu)^2 = \eta^p \left(2\log \eta^{-p}\right)^{(2-p)/2}(1 + o(1)), \quad \text{as } \eta \to 0.$$

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

$$G(p) = \sup_\eta \frac{\rho_g(\eta)}{\rho(\eta)} < \infty, \text{ for } p \in (0, 2).$$

Theorems 4 and 5 in Section 4 of Donoho and Johnstone (1998) derived the asymptotic minimaxity over Besov bodies from the univariate Bayes-minimax estimators. It then follows from an analogous argument of Section 5.3 in Donoho and Johnstone (1998) that

$$R^*_T(B^\alpha_{p,q}(M)) \leq \inf_{\lambda_j} \sup_{\theta \in B^\alpha_{p,q}(M)} E \sum_{j=j_0}^{\infty} \|\widehat{\underline{\theta}}_j(\lambda_j, 1) - \underline{\theta}_j\|^2 \leq G(p \wedge q) \cdot R^*(B^\alpha_{p,q}(M))(1 + o(1)). \quad \blacksquare$$

6.6 Proof of Proposition 1

Proposition 1 in Section 5 shows that the minimizer of the Stein's unbiased estimator of risk for the block James-Stein estimator is in a finite set A with cardinality $|A| \leq d + 1$. This makes the implementation of the SureBlock procedure easy. In addition, Proposition 1 is needed for the proof of Proposition 2, a key result to prove Theorem 4.

Proof of Proposition 1: We shall consider two separate cases: $L = 1$ and $L \geq 2$. For $L = 1$, we define $A = \{x_i^2;\ 1 \leq i \leq d\} \cup \{0\}$. From equations (8) and (9), Stein's unbiased estimator of risk is

$$SURE(x, \lambda, 1) = \sum_{i=1}^{d} \left(1 + \frac{\lambda^2 + 2\lambda}{x_i^2} I\left(x_i^2 > \lambda\right) + \left(x_i^2 - 2\right) I\left(x_i^2 \leq \lambda\right)\right).$$

Suppose, without loss of generality, that the x_i have been reordered in the order of increasing $|x_i|$. On intervals $x_k^2 \leq \lambda < x_{k+1}^2$,

$$SURE\left(x, \lambda, 1\right) = d + \sum_{i=1}^{k} \left(x_i^2 - 2\right) + \sum_{i \geq k+1} \frac{\lambda^2 + 2\lambda}{x_i^2}$$

is strictly increasing. So is true on the interval $0 \leq \lambda < x_1^2$. Therefore the minimizer λ^S of $SURE\left(x, \lambda, 1\right)$ is either 0 or one of the x_i^2.

We now turn to the case $L \geq 2$. We first show that $\lambda^S \geq L - 2$, or equivalently that

$$SURE\left(x, \lambda, L\right) \geq SURE\left(x, L-2, L\right) \tag{48}$$

uniformly for $0 \leq \lambda \leq L - 2$. It suffices to show that for every block b

$$SURE\left(\underline{x}_b, \lambda, L\right) \geq SURE\left(\underline{x}_b, L-2, L\right), \text{ for } 0 \leq \lambda \leq L-2.$$

Equation (8) gives

$$\begin{aligned} & SURE\left(\underline{x}_b, L-2, L\right) - SURE\left(\underline{x}_b, \lambda, L\right) \\ = & -\frac{\left(\lambda - (L-2)\right)^2}{S_b^2} I\left(S_b^2 > L-2\right) + \left(\frac{S_b^4 - 2LS_b^2 - \lambda^2 + 2\lambda\left(L-2\right)}{S_b^2}\right) I\left(\lambda < S_b^2 \leq L-2\right) \end{aligned}$$

The first term on the right side is nonpositive and the second term can be also shown to be nonpositive. Note that the convex function $h\left(x\right) = x^2 - 2Lx - \left(\lambda^2 - 2\lambda\left(L-2\right)\right)$ is nonpositive at two endpoints $x = \lambda$ and $x = L - 2$ which implies

$$S_b^4 - 2LS_b^2 - \lambda^2 + 2\lambda\left(L-2\right) \leq 0, \text{ for } \lambda < S_b^2 \leq L-2.$$

Thus inequality (48) is established.

We now consider the case $\lambda \geq L-2$ with $L \geq 2$. Let $S_{(1)}^2 \leq S_{(2)}^2 \leq \cdots \leq S_{(m)}^2$ denote the order statistics of $S_i^2, 1 \leq i \leq m$. On the interval $S_{(k)}^2 \leq \lambda < S_{(k+1)}^2$ for some $1 \leq k \leq m$, we consider two separate cases: (i) $L - 2 \leq S_{(k)}^2 \leq \lambda < S_{(k+1)}^2$; (ii) $S_{(k)}^2 \leq L - 2 \leq \lambda < S_{(k+1)}^2$. If $L - 2 \leq S_{(k)}^2 \leq \lambda < S_{(k+1)}^2$, it follows from equation (8) that

$$SURE\left(x, \lambda, L\right) = d + \sum_{b=1}^{k} \left(S_{(b)}^2 - 2L\right) + \sum_{b \geq k+1} \frac{\lambda^2 - 2\lambda\left(L-2\right)}{S_{(b)}^2}$$

which, as a function of λ, achieves minimum at $S_{(k)}^2$. For the second case, $SURE\left(x, \lambda, L\right)$ achieves minimum at $\lambda = L-2$ from the equation above, and this completes the proof. ∎

References

[1] Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations (with discussion). *J. Amer. Statist. Assoc.* **96**, 939-967.

[2] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

[3] Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.

[4] Cai, T., Low, M.G., and Zhao, L. (2000). Blockwise thresholding methods for sharp adaptive estimation. Technical Report, Department of Statistics, University of Pennsylvania.

[5] Cai, T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya Ser. B* **63**, 127-148.

[6] Cai, T. and Zhou, H. (2005). A data-driven block thresholding approach to wavelet estimation. Technical Report, Department of Statistics, University of Pennsylvania. Available at
`http://stat.wharton.upenn.edu/~tcai/paper/SureBlock-TechReport.pdf`.

[7] Chicken, E. (2004). Block-dependent thresholding in wavelet regression. *J. Nonparametric Statist.*, to appear.

[8] Chicken, E. and Cai, T. (2004). Block thresholding for density estimation: Local and global adaptivity. *J. Multivariate Anal.*, in press.

[9] Daubechies, I. (1992). *Ten Lectures on Wavelets* SIAM: Philadelphia.

[10] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305**, 397-414.

[11] Donoho, D.L. and Johnstone, I.M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

[12] Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over l_p-balls for l_q-error. *Prob. Th. Rel. Fields* **99**, 277-303.

[13] Donoho, D. L. and Johnstone, I. M. (1995). Adapt to unknown smoothness via wavelet shrinkage. *J. Amer. Stat. Assoc.* **90** , 1200-1224.

[14] Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

[15] Donoho, D.L., Liu, R.C. and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18**, 1416-37.

[16] Efromovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Theor. Probab. Appl.* **30**, 557-661.

[17] Gao, H.-Y.(1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.* **7**, 469-488.

[18] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9**, 33-50.

[19] Johnstone, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9**, 51-84.

[20] Kerkyacharian, G., Picard, D. and Tribouley, K. (1996). L_p Adaptive density estimation. *Bernoulli* **2**, 229-247.

[21] Lehmann, E.L. (1983). *Theory of Point Estimation.* John Wiley and Sons, New York.

[22] Meyer, Y. (1992). *Wavelets and Operators*. Cambridge University Press, Cambridge.

[23] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-51.

[24] Strang, G. (1992). Wavelet and dilation equations: a brief introduction. *SIAM Review* **31**, 614 - 627.

[25] Triebel, H. (1983). *Theory of Function Spaces*. Birkhäuser Verlag, Basel.

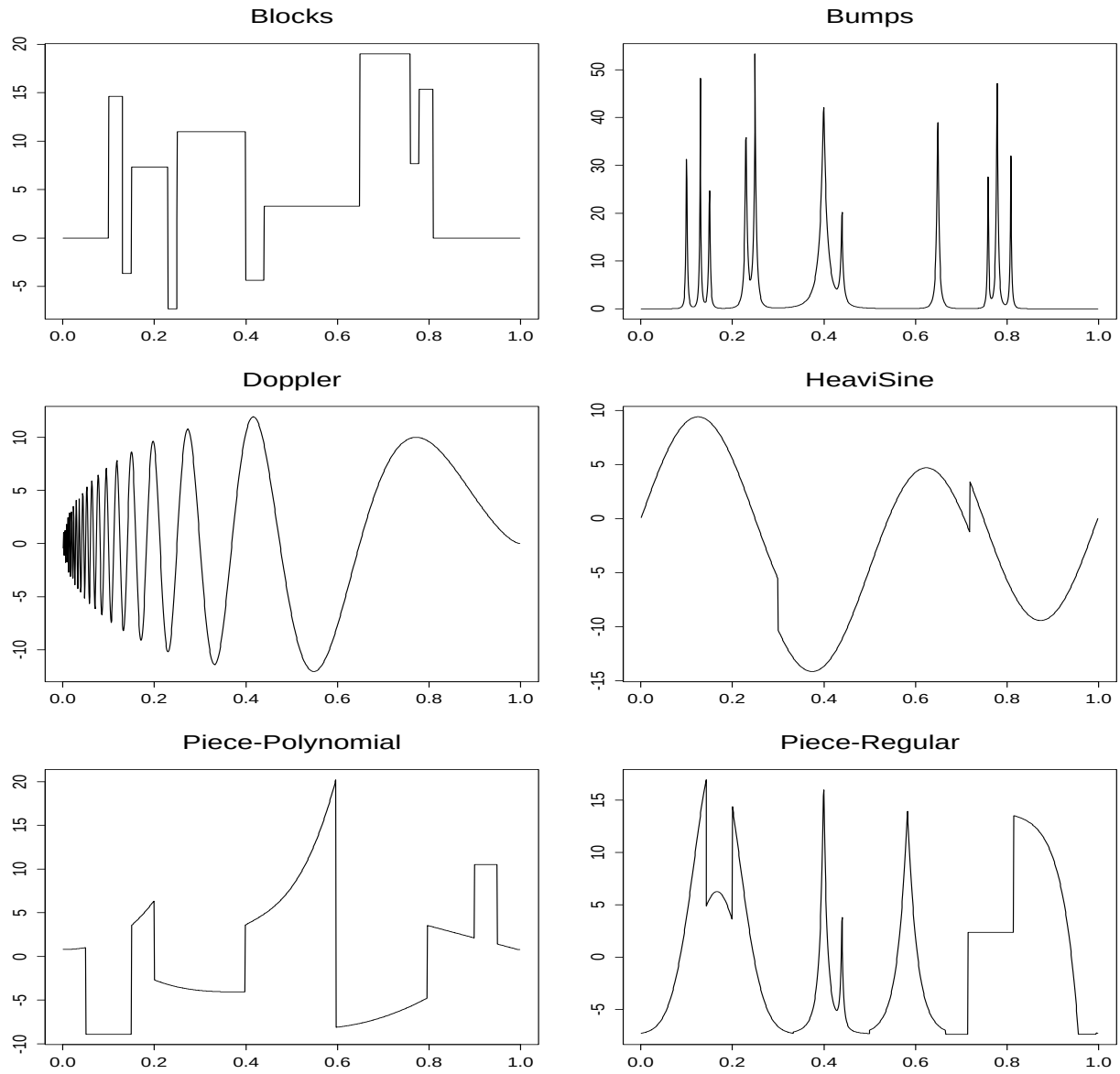

Figure 3: Test functions

7 Appendix: Proofs of Technical Results

We collect in this Appendix the proofs of the technical results stated in Section 6.1.

7.1 Proof of Lemma 1

Recall that $R(\theta)$ and $\widetilde{R}(\theta)$ are defined as following

$$\begin{aligned} R(\theta) &= \inf_{\lambda, 1\leq L\leq n^v} r(\lambda, L) = \inf_{L-2\leq\lambda, 1\leq L\leq n^v} r(\lambda, L) \\ \widetilde{R}(\theta) &= \inf_{\lambda\leq\lambda^F, 1\leq L\leq n^v} r(\lambda, L) = \inf_{L-2\leq\lambda\leq\lambda^F, 1\leq L\leq n^v} r(\lambda, L), \end{aligned}$$

If $R(\theta) = r(\lambda_1, L_1)$ with $\lambda_1 \leq \lambda^F$, $\widetilde{R}(\theta) - R(\theta) = 0$. In the case $R(\theta) = r(\lambda_1, L_1)$ with $\lambda_1 > \lambda^F$, we have $\widetilde{R}(\theta) \leq r\left(\lambda^F, L_1\right)$, which implies

$$\begin{aligned} \widetilde{R}(\theta) - R(\theta) &\leq r\left(\lambda^F, L_1\right) - r(\lambda_1, L_1) \\ &= \frac{1}{d}\sum_{b=1}^{m}\left(r_b\left(\lambda^F, L_1\right) - r_b(\lambda_1, L_1)\right) \end{aligned}$$

then it is enough to prove

$$B = r_b\left(\lambda^F, L\right) - r_b(\lambda_1, L) \leq Cd^{\delta-1}$$

uniformly for all λ_1, b and L. Recall that

$$r_b(\lambda, L) = \|\underline{\theta}_b\|^2 + E_{\underline{\theta}_b} g(\lambda) I\left(S_b^2 > \lambda\right)$$

where

$$g(\lambda) = \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} - S_b^2 + 2L$$

which is an increasing function of λ for $\lambda \geq L-2$, then simple algebra gives

$$B \leq E_{\underline{\theta}_b}\left(\frac{\left(\lambda^F\right)^2 - 2\lambda^F(L-2)}{S_b^2} - S_b^2 + 2L\right) I\left(\lambda_1 \geq S_b^2 > \lambda_F\right)$$

Note that

$$\frac{\left(\lambda^F\right)^2 - 2\lambda^F(L-2)}{S_b^2} - S_b^2 + 2L \leq 0, \text{ when } S_b^2 \geq \lambda_F + 2$$

so

$$\begin{aligned} B &\leq E_{\underline{\theta}_b}\left(\frac{\left(\lambda^F\right)^2 - 2\lambda^F(L-2)}{S_b^2} - S_b^2 + 2L\right) I\left(\lambda_F + 2 \geq S_b^2 > \lambda_F\right) \\ &\leq 4P_{\underline{\theta}_b}\left(\lambda_F + 2 \geq S_b^2 > \lambda_F\right). \end{aligned}$$

Then

$$\begin{aligned} B &\leq 4P_{\underline{\theta}_b}\left(\lambda_F + 2 \geq S_b^2\right) \\ &\leq 4P_{\underline{\theta}_b}\left(\lambda_F + 2 \geq 1/2\|\theta_b\|_2^2 - \|z_b\|_2^2\right) = O\left(d^{-1}\right) \end{aligned}$$

when $\|\underline{\theta}_b\|_2^2 \geq 4\lambda_F$.

Now we consider the case $\|\underline{\theta}_b\|^2 \leq 4\lambda_F$. We have

$$\begin{aligned} r\left(\lambda^F, \underline{\theta}_b\right) - r\left(\lambda_1, \theta_b\right) &= -\int_{\lambda^F}^{\lambda_1} \frac{\partial}{\partial\lambda} r\left(\lambda, \underline{\theta}_b\right) d\lambda \\ &= -\int_{\lambda^F}^{\lambda_1} \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{y>\lambda} \frac{2\lambda - 2L + 4}{y} f_{L+2k}\left(y\right) dy - 4 f_{L+2k}\left(\lambda\right)\right) d\lambda \end{aligned}$$

and

$$\begin{aligned} f_{L+2k}\left(\lambda\right) &= -\int_{y>\lambda} f'_{L+2k}\left(y\right) dy \\ &= -\int \frac{\left(L+2k\right)/2 - 1 - y/2}{y} f_{L+2k}\left(y\right) dy \end{aligned}$$

which leads to

$$\begin{aligned} & r\left(\lambda^F, \underline{\theta}_b\right) - r\left(\lambda_1, \theta_b\right) \\ = & -2\int_{\lambda^F}^{\lambda_1} \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{y>\lambda} \frac{\lambda + 2k - y}{y} f_{L+2k}\left(y\right) dy\right) d\lambda \\ = & I_1 - I_2 \end{aligned}$$

where

$$\begin{aligned} I_1 &= 2\int_{\lambda^F}^{\lambda_1} \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{y>\lambda+2k} \frac{y - (\lambda + 2k)}{y} f_{L+2k}\left(y\right) dy\right) d\lambda \\ I_2 &= 2\int_{\lambda^F}^{\lambda_1} \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{\lambda+2k\geq y>\lambda} \frac{(\lambda + 2k) - y}{y} f_{L+2k}\left(y\right) dy\right) d\lambda \end{aligned}$$

Note that

$$\begin{aligned} I_1 &= 2\int_{\lambda^F}^{\lambda_1} \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{y>\lambda+4k} + \int_{\lambda+4k\geq y>\lambda+2k} \frac{y - (\lambda + 2k)}{y} f_{L+2k}\left(y\right) dy\right) d\lambda \\ &= I_{11} + I_{12} \end{aligned}$$

and

$$\begin{aligned} I_2 &= 2\int_{\lambda^F}^{\lambda_1} \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{\max\{\lambda, L+2k-2\}}^{\lambda+2k} + \int_{\min\{\lambda, L+2k-2\}}^{\max\{\lambda, L+2k-2\}} \frac{(\lambda + 2k) - y}{y} f_{L+2k}\left(y\right) dy\right) d\lambda \\ &= I_{21} + I_{22} \end{aligned}$$

where

$$I_{11} = o\left(d^{-1/2}\right),$$

since

$$\begin{aligned}\int_{y>\lambda+4k} \frac{y-(\lambda+2k)}{y} f_{L+2k}(y)\,dy &\leq P\left(\chi^2_{L+2k} \geq \lambda^F + 4k\right) \\ &\leq \frac{1}{2}\left[\left(\frac{\lambda^F+4k}{L+2k}\right)^{-1} \exp\left(-\frac{1}{2}\left(\lambda^F + 4k - L + 2k\right)\right)\right] \\ &= o\left(d^{-1/2}\right)\end{aligned}$$

by Lemma 2 in Cai (1999), and $I_{22} \geq 0$, then

$$I_1 - I_2 \leq o\left(d^{-1/2}\right) + I_{12} - I_{21}.$$

When $\lambda \geq L + 2k - 2$, then $f_{L+2k}(y)$ is increasing on $[\lambda, \infty)$, thus we have

$$\begin{aligned}&\int_{\lambda+2k}^{\lambda+4k} \frac{y-(\lambda+2k)}{y} f_{L+2k}(y)\,dy - \int_{\lambda}^{\lambda+2k} \frac{(\lambda+2k)-y}{y} f_{L+2k}(y)\,dy \\ =\ &\int_0^{2k}\left[\frac{z}{z+\lambda+2k} f_{L+2k}(z+\lambda+2k) - \frac{z}{\lambda+2k-z} f_{L+2k}(\lambda+2k-z)\right] dz \leq 0\end{aligned}$$

because the integrand is nonpositive. When $\lambda \leq L + 2k - 2$, then $k \geq \left(\lambda^F - L + 2\right)/2$. Let's assume $k \leq M\lambda^F$. In this case, we consider the following difference

$$\begin{aligned}&\int_{\lambda+2k}^{\lambda+4k} \frac{y-(\lambda+2k)}{y} f_{L+2k}(y)\,dy - \int_{L+2k-2}^{\lambda+2k} \frac{(\lambda+2k)-y}{y} f_{L+2k}(y)\,dy \\ \leq\ &\int_{\lambda+2k}^{\lambda+4k} \frac{y-(\lambda+2k)}{y} f_{L+2k}(y)\,dy - \int_{L+2k-2}^{(\lambda+L)/2+2k-2} \frac{(\lambda+2k)-y}{y} f_{L+2k}(y)\,dy\end{aligned}$$

Note that, for $c_1 > c_2 > 1$ and $m = L + 2k$ we have

$$\frac{f_m(c_1 m)}{f_m(c_2 m)} = \left(\frac{c_1}{c_2}\right)^{m/2-1} e^{(c_2-c_1)m/2} = \frac{c_2}{c_1} e^{m(c_2-\log c_2-(c_1-\log c_1))} \geq \frac{c_2}{c_1} e^{m(1-1/c_1)(c_2-c_1)}$$

because $(x - \log x)' = 1 - 1/x \geq 1 - 1/c_1$, for $c_2 \geq x \geq c_1$, as $d \to \infty$ we have $\frac{f_m(c_1 m)}{f_m(c_2 m)} \gg m$ which implies as $d \to \infty$ the following term is negative

$$\begin{aligned}&\int_{\lambda+2k}^{\lambda+4k} \frac{y-(\lambda+2k)}{y} f_{L+2k}(y)\,dy - \int_{L+2k-2}^{(\lambda+L)/2+2k-2} \frac{(\lambda+2k)-y}{y} f_{L+2k}(y)\,dy \\ \leq\ &2k f_{L+2k}(\lambda+2k) - \left(\frac{(\lambda+2k) - ((\lambda+L)/2+2k-2)}{(\lambda+L)/2+2k-2}\right) f_{L+2k}\left((\lambda+L)/2+2k-2\right)\end{aligned}$$

and further $I_{12} \leq I_{21}$, as $d \to \infty$. If $k \geq M\lambda^F$, it is easy to see $\sum \frac{\theta_*^k e^{-\theta_*}}{k!} = o\left(\frac{1}{d}\right)$ by Stirling formula when M is sufficient large, which implies

$$\left| 2 \int_{\lambda^F}^{\lambda_1} \sum_{k \geq M_2 L}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{y>\lambda} \frac{\lambda + 2k - y}{y} f_{L+2k}(y)\, dy \right) d\lambda \right| = o\left(\frac{1}{d}\right)$$

Putting all together, we obtain the lemma. ∎

7.2 Proof of Lemma 3

As in the proof of Theorem 4, set $T_d = d^{-1} \sum (x_i^2 - 1)$, $\gamma_d = d^{-\frac{1}{2}} \log_2^{\frac{3}{2}} d$ and define the event $A_d = \{T_d \leq \gamma_d\}$. Then it follows from the definition of the SureBlock estimator that

$$\|\hat{\theta}^* - \theta\|_2^2 = \sum_b \|\widehat{\underline{\theta}}_b(\lambda^*, L^*) - \underline{\theta}_b\|_2^2 I(A_d) + \|\hat{\theta}^F - \theta\|_2^2 I(A_d^c) \tag{49}$$

where θ^F denotes the non-negative garrote estimator given in (12).

Note that the James-Stein estimator satisfies

$$\begin{aligned} \|\widehat{\underline{\theta}}_b(\lambda, L) - \underline{\theta}_b\|_2^2 &= \|(1 - \frac{\lambda}{S_b^2})\underline{x}_b - \underline{\theta}_b\|_2^2 \leq 2\|\underline{z}_b\|_2^2 + 2\frac{\lambda^2}{S_b^2} \quad \text{if} \quad S_b^2 \geq \lambda \\ &= \|\underline{\theta}_b\|_2^2 \leq 2S_b^2 + 2\|\underline{z}_b\|_2^2 \quad \text{if} \quad S_b^2 < \lambda \end{aligned}$$

where $\underline{z}_b = \underline{x}_b - \underline{\theta}_b$. Hence

$$\|\widehat{\underline{\theta}}_b(\lambda, L) - \underline{\theta}_b\|_2^2 \leq 2\lambda + 2\|\underline{z}_b\|_2^2. \tag{50}$$

Noting $\lambda^* \leq \lambda^F$ and applying (50) to both terms on the RHS of (49), we have

$$\|\hat{\theta}^* - \theta\|_2^2 \leq 2\sum_b (\lambda^* + \|\underline{z}_b\|_2^2) I(A_d) + 2(\lambda^F d + \|z\|_2^2) I(A_d^c) \leq 2\lambda^F d + 2\|z\|_2^2$$

and this proves (30).

We now turn to the proof of (31). Note that

$$SURE(\underline{x}_b, \lambda, L) = \left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} - S_b^2 + 2L \right) I\left(S_b^2 > \lambda\right) + S_b^2 - L.$$

Then

$$\begin{aligned} r_b(\lambda, L) &= E_{\underline{\theta}_b}(SURE(\underline{x}_b, \lambda, L)) \\ &= \|\underline{\theta}_b\|_2^2 + E_{\underline{\theta}_b}\left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} - S_b^2 + 2L \right) I\left(S_b^2 > \lambda\right). \end{aligned}$$

Note that S_b^2 has noncentral chi-squared distribution with density function

$$h(y) = \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} f_{L+2k}(y),$$

where $\theta_* = \|\underline{\theta}_b\|_2^2/2$ and $f_m(y) = \frac{1}{2^{m/2}\Gamma(m/2)} y^{m/2-1} e^{-y/2}$. Straightforward calculations then yield

$$\frac{\partial}{\partial\lambda} r_b(\lambda, L) = \sum_{k=0}^{\infty} \frac{\theta_*^k e^{-\theta_*}}{k!} \left(\int_{y>\lambda} \frac{2\lambda - 2L + 4}{y} f_{L+2k}(y)\, dy - 4 f_{L+2k}(\lambda) \right).$$

It is easy to see $f_m(y)$ achieves maximum at $m-2$, then we have

$$f_m(m-2) = \left(\sqrt{\pi}\right)^{-1} \left(\frac{m-2}{m}\right)^{m/2} e^{1-\epsilon_{m/2}} m^{-1/2} (m-2)^{-1} < 1$$

from Stirling formula $j! = \sqrt{2\pi} j^{j+1/2} \exp(-j + \epsilon_j)$ with $1/(12j+1) < \epsilon_j < 1/(12j)$. Thus

$$\left| \frac{\partial}{\partial\lambda} r_b(\lambda, L) \right| \le 2 + 4h(\lambda) < 6. \quad \blacksquare$$

7.3 Proof of Proposition 2

The following lemma is a preparation to prove the key technical result, Proposition 2.

Lemma 4 *For $\lambda > \max\{L-2, 0\}$ and $L \ge 1$, we have*

$$\begin{aligned}
|SURE(\underline{x}_b, \lambda, L)| &\le \lambda + 4 && (51)\\
|r_b(\lambda, L)| &\le \lambda + 4 && (52)\\
\|\widehat{\theta}_b(\lambda, L) - \theta_b\|_2^2 &\le 2\lambda + 2\|z_b\|_2^2 && (53)\\
\left| \frac{\partial}{\partial\lambda} \|\widehat{\theta}_b(\lambda, L) - \underline{\theta}_b\|^2 \right| &\le 3 + \frac{1}{\lambda}\|z_b\|_2^2 && (54)
\end{aligned}$$

Proof : (i) From (8) we have

$$SURE(\underline{x}_b, \lambda, L) = L + \frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} I\left(S_b^2 > \lambda\right) + \left(S_b^2 - 2L\right) I\left(S_b^2 \le \lambda\right)$$

then

$$|SURE(\underline{x}_b, \lambda, L)| \le \max\left\{ \left| \left(\frac{\lambda^2 - 2\lambda(L-2) + S_b^2 L}{S_b^2} \right) I\left(S_b^2 > \lambda\right) \right|, \left| \left(S_b^2 - L\right) I\left(S_b^2 \le \lambda\right) \right| \right\}.$$

It is easy to see

$$\begin{aligned} 0 &\leq \frac{\lambda^2 - 2\lambda(L-2) + S_b^2 L}{S_b^2} = \frac{\lambda(\lambda - (L-4)) + (S_b^2 - \lambda)L}{S_b^2} \leq \lambda + 4, \text{ if } S_b^2 > \lambda \\ -(\lambda+2) &\leq -L \leq S_b^2 - L \leq \lambda - L \leq \lambda, \text{ if } S_b^2 < \lambda, \end{aligned}$$

then we have $|SURE(\underline{x}_b, \lambda)| \leq \lambda + 4$.

(ii) The inequality (51) implies $|r_b(\lambda, L)| = |E_{\theta_b}(SURE(\underline{x}_b, \lambda, L))| \leq \lambda + 4$.

(iii) Inequality (53) follows directly from (50) given in the proof of Lemma 3.

(iv) It is also easy to see almost surely

$$\left| \frac{\partial}{\partial \lambda} \|\widehat{\theta}_b(\lambda, L) - \underline{\theta}_b\|^2 \right| = \left| \sum_i \left(2\lambda \frac{x_i^2}{S_b^4} - 2z_i \frac{x_i}{S_b^2} \right) I\left(S_b^2 > \lambda\right) \right|.$$

where $z_b = \underline{x}_b - \underline{\theta}_b$. For $S_b^2 > \lambda$, we have $\sum_i 2\lambda \frac{x_i^2}{S_b^4} = \frac{2\lambda}{S_b^2} \leq 2$ and $\left|\sum_i 2z_i \frac{x_i}{S_b^2}\right| \leq \sum_i \frac{z_i^2 + x_i^2}{S_b^2} \leq \frac{1}{\lambda}\|z_b\|_2^2 + 1$. ∎

Now we are ready to prove the two propositions using the lemma above and the following exponential inequalities:

(i) Let $Y_1, \ldots, Y_m$ be independent, $|\max Y_i - \min Y_i| \leq M$, and $\bar{Y}_m = m^{-1}\sum_{i=1}^m Y_i$ and $\mu = E\bar{Y}_m$. For $t > 0$,

$$P\left(\left|\bar{Y}_m - \mu\right| \geq t\right) \leq 2\exp\left(-2mt^2/M^2\right). \tag{55}$$

(ii) Let $Z_1, \ldots, Z_m$ be i.i.d $N(0,1)$. For $t > 0$,

$$P\left(\left|m^{-1}\sum\left(Z_i^2 - 1\right)\right| > t\right) \leq 2\exp\left(-mt\left(t \wedge 1\right)/8\right). \tag{56}$$

Proof of Proposition 2: We first prove (33). Set

$$Z_d(\lambda, L) \triangleq U_d(\lambda, L) - r(\lambda, L) = \frac{1}{d}\sum_{b=1}^m \left(SURE(\underline{x}_b, \lambda, L) - r_b(\lambda, L)\right)$$

where $m = d/L$. For $r_d = L^{1/2}(\log d)^{1+\rho}$ with $\rho > 1/2$ to be specified later, (51) and (52) imply

$$|SURE(\underline{x}_b, \lambda, L) - r_b(\lambda, L)| \leq 2(\lambda + 4),$$

then the exponential inequality (55) gives

$$P\left(|Z_d(\lambda, L)| > r_d d^{-1/2}\right) = P\left(|Z_d(\lambda, L)| L > r_d d^{-1/2} L\right) \leq 2\exp\left\{-\frac{r_d^2 L}{2(\lambda+4)^2}\right\}.$$

For distinct $\lambda < \lambda'$ with $\lambda' - \lambda \leq \delta_d$, where δ_d will be specified later, let $N_d(\lambda, \lambda') = \#\{i : \lambda < S_b^2 \leq \lambda'\}$, then from (8)

$$|U(\lambda, L) - U(\lambda', L)| = \left| \frac{1}{d} \sum_{b=1}^{m} (SURE(\underline{x}_b, \lambda, L) - SURE(\underline{x}_b, \lambda', L)) \right| = U_1 + U_2$$

where

$$\begin{aligned} U_1 &= \left| \frac{1}{d} \sum_b \left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} - S_b^2 - 2L \right) I\left(\lambda < S_b^2 \leq \lambda'\right) \right| \\ U_2 &= \left| \frac{1}{d} \sum_b \left(\frac{\lambda^2 - 2\lambda(L-2)}{S_b^2} - \frac{(\lambda')^2 - 2\lambda'(L-2)}{S_b^2} \right) I\left(\lambda' < S_b^2\right) \right| \end{aligned}$$

Simple algebra gives

$$\begin{aligned} U_1 &= \left| \frac{1}{d} \sum_b \left(\frac{(\lambda + S_b^2 - 2(L-2))(\lambda - S_b^2) + 4S_b^2}{S_b^2} \right) I\left(\lambda < S_b^2 \leq \lambda'\right) \right| \\ &\leq \frac{4}{d}(1 + \delta_d/L) N_d(\lambda, \lambda') \\ U_2 &= \left| \frac{1}{d} \sum_b \left(\frac{(\lambda' + \lambda - 2(L-2))(\lambda' - \lambda)}{S_b^2} \right) I\left(\lambda' < S_b^2\right) \right| \leq 2\delta_d/L \end{aligned}$$

From (31) we have $|r(\lambda) - r(\lambda')| \leq 6\delta_d/L$ for $\lambda' - \lambda \leq \delta_d$. Then so long as $|\lambda' - \lambda| \leq \delta_d$

$$|Z_d(\lambda, L) - Z_d(\lambda', L)| \leq \frac{4}{d}(1 + \delta_d/L) N_d(\lambda, \lambda') + 8\delta_d/L$$

Now choose $\lambda_j = j\delta_d \in \left[\max\{L-2, 0\}, \lambda^F\right]$, clearly

$$A_d = \left\{ \sup_{[L-2, \lambda^F]} |Z_d(\lambda, L)| \geq 3 r_d d^{-1/2} \right\} \subset D_d \cup E_d$$

where

$$\begin{aligned} D_d &= \left\{ \sup_j |Z_d(\lambda_j, L)| \geq r_d d^{-1/2} \right\} \\ E_d &= \left\{ \sup_j \sup_{|\lambda - \lambda_j| \leq \delta_d} |Z_d(\lambda, L) - Z_d(\lambda_j, L)| \geq 2 r_d d^{-1/2} \right\} \end{aligned}$$

Choose $\delta_d/L = o\left(r_d d^{-1/2}\right)$; then E_d is contained in

$$\begin{aligned} E_d' &= \left\{ \sup_j \frac{4}{d} N_d(\lambda_j, \lambda_j - \delta_d) \geq r_d d^{-1/2} \right\} \\ &\subset \left\{ \sup_j \frac{1}{d} |N_d(\lambda_j, \lambda_j - \delta_d) - E N_d(\lambda_j, \lambda_j - \delta_d)| \geq \frac{1}{5} r_d d^{-1/2} \right\} \end{aligned}$$

say, for large d where we used

$$EN_d(\lambda_j, \lambda_j - \delta_d) = \sum_{b=1}^{m} P\left(\lambda_j - \delta_d < S_b^2 \leq \lambda_j\right) \leq c_0 \frac{d}{L}\delta_d = o\left(r_d d^{1/2}\right).$$

Again from exponential inequality (55)

$$\begin{aligned} & P\left(\frac{1}{d}\left|N_d(\lambda_j, \lambda_j - \delta_d) - EN_d(\lambda_j, \lambda_j - \delta_d)\right| \geq \frac{1}{5} r_d d^{-1/2}\right) \\ \leq\ & 2\exp\left(-2\left(\frac{d}{L}\right)\left(\frac{1}{5} r_d d^{-/12} L\right)^2\right) \\ =\ & 2\exp\left(-2r_d^2 L/25\right) \end{aligned}$$

Finally

$$\begin{aligned} P(A_d) &\leq P(D_d) + P(E_d') \\ &\leq \frac{\lambda^F}{\delta_d} \cdot \left(2\exp\left\{-\frac{r_d^2 L}{8(\lambda+4)^2}\right\} + 2\exp\left(-2r_d^2 L/25\right)\right), \end{aligned} \tag{57}$$

which decays faster than any polynomial of d when $\rho > \nu + 1/2$. Then

$$P\left\{\sup_{L-2\leq\lambda\leq\lambda^F, 1\leq L\leq d^v} |Z_d(\lambda, L)| \geq 3r_d d^{-1/2}\right\} = o\left(d^{-b+v}\right)$$

for any $b > 0$, which implies

$$\begin{aligned} E\sup_{L-2\leq\lambda\leq\lambda^F, 1\leq L\leq d^v} |Z(\lambda, L)| &\leq C d^{\eta-1/2+v/2} + \left(E\left(\sup_{\max\{L-2,1\}\leq\lambda\leq\lambda^F, 1\leq L\leq d^v} |Z(\lambda, L)|\right)^2\right)^{1/2} P(A_d) \\ &\leq C d^{\eta-1/2+v/2} + 2\left(\lambda^F + 4\right) P(A_d) = O\left(d^{\eta-1/2+v/2}\right) \end{aligned}$$

for any $\eta > 0$.

We now turn to the proof of (34). Let

$$W(\lambda, L) = D(\lambda, L) - r(\lambda, L) = \frac{1}{d}\sum_{b=1}^{m}\left(\|\widehat{\theta}_b - \underline{\theta}_b\|_2^2 - r(\lambda, L)\right) = \frac{1}{d}\sum_{b=1}^{m} Y_b$$

From (31) and (54) we have

$$\sup_{L-2\leq\lambda\leq\lambda^F} |W'(\lambda, L)| \leq \frac{1}{d\lambda}\|z\|_2^2 + \frac{9}{L}.$$

Let

$$\begin{aligned} S_1 &= \cap_{j\in J}\left\{W\left(t_j\right)\right\} \leq d^{\eta-1/2+v/2} \\ S_2 &= \left\{\sup_{L-2\leq\lambda\leq\lambda^F}\left|W'\left(\lambda\right)\right| \leq 11\log d\right\} \end{aligned}$$

Then so long as $11\left(\log d\right)\delta_d \leq d^{\delta-1/2-v/2}$, then

$$S_1 \cap S_2 \Longrightarrow \sup_{L-2\leq\lambda\leq\lambda^F}\left|W\left(\lambda\right)\right| \leq 2d^{\eta-1/2+v/2}$$

Now let's consider bounds for $P\left(S_1^c\right)$ and $P\left(S_2^c\right)$. We see $E\left|Y_b\right|^k \leq \left(c\lambda^F\right)^k$ for some $c>0$ from (52) and (53), then

$$E\left|W\left(\lambda,L\right)\right|^k \leq \frac{m^{k/2}}{d^k}\left(c\lambda^F\right)^k$$

for any integer $k>0$, and Chebyshev inequality gives

$$P\left(\left|W\left(\lambda\right)\right| > d^{\eta-1/2+v/2}\right) \leq \frac{E\left|W\left(\lambda,L\right)\right|^k}{d^{(\eta-1/2+v/2)k}} \leq \frac{\left(c\lambda^F/L\right)^{k/2}\left(c\lambda^F\right)^{k/2}}{d^{(2\eta+v)k/2}}$$

which decays faster than any polynomial of d when k is large enough, so

$$P\left(S_1^c\right) \leq \frac{\lambda^F}{\delta_d}\max P\left(\left|W\left(\lambda\right)\right| > d^{\eta-1/2+v/2}\right)$$

decays faster than any polynomial of d. Thus

$$P\left\{\sup_{L-2\leq\lambda\leq\lambda^F,1\leq L\leq d^v}\left|W\left(\lambda,L\right)\right| \geq d^{\eta-1/2+v/2}\right\} = o\left(d^{-b+v}\right)$$

for any $b>0$. It is easy to see that

$$P\left(S_2^c\right) \leq P\left(\frac{1}{d}\left(\left\|z\right\|_2^2-1\right)\geq 1\right) \leq 2e^{-d/8}$$

also decays faster than any polynomial of d by Hoeffding inequality (56). Thus

$$\begin{aligned} E\sup_{L-2\leq\lambda\leq\lambda^F,1\leq L\leq d^v}\left|W\left(\lambda\right)\right| &\leq 2d^{\eta-1/2} + \left(E\left(\sup_{L-2\leq\lambda\leq\lambda^F,1\leq L\leq d^v}\left|W\left(\lambda\right)\right|\right)^2\right)^{1/2} P\left(S_1^c\cup S_2^c\right) \\ &= O\left(d^{\eta-1/2+v/2}\right) \quad \blacksquare \end{aligned}$$