MINIMAX ESTIMATION WITH THRESHOLDING AND ITS APPLICATION TO WAVELET ANALYSIS

Harrison H. Zhou*
and
J. T. Gene Hwang**

Cornell University

May 1, 2003

Abstract. Many statistical practices involve selecting a model (a reduced model from the full model) and then use it to do estimation with possible thresholding. Is it possible to do so and still come up with an estimator always better than the naive estimator without model selection? The James-Stein estimator allows us to do so. However, the James-Stein estimator considers only one reduced model, the origin. What should be more desirable is to select a data chosen reduced model (of an arbitrary dimension) and then do estimation with possible thresholding. In this paper, we construct such estimators. We apply the estimators to the wavelet analysis. In the finite sample settings, these estimators are minimax and perform the best among the well-known estimators trying to do model selection and estimation at the same time. Some of our estimators are also shown to be asymptotically optimal.

Key words and phrases: James–Stein estimator, model selection, VisuShrink, Sureshrink, BlockJS

AMS 2000 Subject Classification: Primary 62G05, 62J07; Secondary 62C10, 62H25

1. Introduction.

In virtually all statistical activities, one constructs a model to summarize the data. Not only could the model provide a good and effective way of summarizing the data, the model if correct often provides more accurate prediction. This point has been argued forcefully in Gauch (1993). Is there a way to use the data to

*Also known as Huibin Zhou.

**Also known as Jiunn T. Hwang.

select a reduced model so that if the reduced model is correct the model based estimator will improve on the naive estimator (constructed using a full model) and yet never do worse than the naive estimator even if the full model is actually the only correct model? James–Stein estimation (1961) provide such a striking result under normality assumption. Any estimator such as the James-Stein estimator that does no worse than the naive estimator is said to be minimax. See the precise discussion right before Lemma 1 of Section 2. The problem with the James–Stein positive part estimator is however that it selects only between two models: the origin and the full model. It is possible to construct estimators similar to James–Stein positive part to select between the full model and another linear subspace. However it always chooses between the two. The nice idea of George (1986a,b) in multiple shrinkage does allow the data to choose among several models; it however does not do thresholding as is the aim of the paper.

In many applications, wavelets is a very important model in statistics. To use the model, it involves model selection among the full model or the models with smaller dimensions where some of the wavelet coefficients are zero. Is there a way to select a reduced model so that the estimator based on it does no worse in any case than the naive estimator based on the full model, but improves substantially upon the naive estimator when the reduced model is correct? Again, the James–Stein estimator provides such a solution. However it selects either the origin or the full model. Furthermore, the ideal estimator should do thresholding, namely it should truncate the components which are small and preserves (or shrinks) the other components. However, to the best knowledge of the authors, no such minimax estimators have been constructed. In this paper, we provide minimax estimators which perform thresholding simultaneously.

Section 1 through Section 3 develop the new estimator for the canonical form of the model by solving Stein's differential inequality. Sections 4 and 5 provide an approximate Bayesian justification and an empirical Bayes interpretation. Sections 7 and 8 apply the result to the wavelet analysis. The proposed method outperforms several prominent procedures in the statistical wavelet literature.

2. New Estimators for a Canonical Model.

In this section, we shall consider the canonical form of the problem of a multinormal mean estimation problem under the squared error loss. Hence we shall assume that our observation

$$Z = (Z_1, \ldots, Z_d) \sim N(\theta, I)$$

is a d–dimensional vector consisting of normal random variable with mean $\theta = (\theta_1, \ldots, \theta_d)$, and a known covariance identity matrix I. The case when the variance of Z_i is not known will be discussed in Section 7.

The connection of this problem with wavelet analysis will be pointed out in Sections 7 and 8. In short Z_i and θ_i represent the wavelet coefficients of the data and the true curve in the same resolution, respectively. Furthermore d is the dimension of a resolution. For now, we shall seek for an estimator of θ based on Z. We shall, without loss of generality, consider an estimator of this form $\delta(Z) = (\delta_1(Z), \ldots, \delta_d(Z))$, where

$$\delta_i(Z) = Z_i + g_i(Z)$$

where $g(Z) : R^d \to R$ and search for $g(Z) = (g_1(Z), \ldots, g_d(Z))$. To insure that the new estimator (perhaps with some thresholding) do better than Z (which does no thresholding), we shall compare the *risk* of $\delta(Z)$ to the risk of Z with respect to the ℓ_2 norm. Namely

$$E\|\delta(Z)-\theta\|^2 = E\sum_{i=1}^{d}(\delta_i(Z)-\theta_i)^2.$$

It is obvious that the risk of Z is then d. We shall say that an estimator *strictly dominates* the other if the former has a smaller risk for every θ. We shall say one *dominates* the other if the former has a risk no greater than the latter for every θ, but has smaller risk for some θ. Note that Z is a minimax estimator, i.e., it minimizes $\sup_\theta E|\delta^0(Z)-\theta|^2$ among all $\delta^0(Z)$. Consequently any $\delta(Z)$ that dominates Z is also minimax.

To construct estimator dominates Z, we use the following lemma.

Lemma 1. *(Stein 1981) Suppose that $g : R^d \to R^d$ is a measurable function with $g_i(\cdot)$ as the ith component. If for every i, $g_i(\cdot)$ is almost differentiable with respect to ith component. If*

$$E\Big(\Big|\frac{\partial}{\partial Z_i}g_i(Z)\Big|\Big) < \infty, \ \textit{for } i = 1, \dots, d$$

then

$$E_\theta\|Z+g(Z)-\theta\|^2 = E_\theta\{d + 2\nabla\cdot g(Z) + \|g(Z)\|^2\},$$

where $\nabla\cdot g(Z) = \sum_{i=1}^{d}\frac{\partial g_i(Z)}{\partial Z_i}$. Hence if $g(Z)$ solves the differential inequality

$$2\nabla\cdot g(Z) + \|g(Z)\|^2 < 0, \tag{0}$$

the estimator $Z+g(Z)$ strictly dominates Z.

Remark: $g_i(z)$ is said to be almost differentiable with respect to z_i, if for almost all z_j, $j \neq i$, $g_i(z)$ can be written as a one dimensional integral of a function with respect to z_i. For such z_j's, $j \neq i$, using Berger's (1980) terminology, one calls $g_i(Z)$ to be absolutely continuous with respect to z_i.

To motivate the proposed estimator, note that the James–Stein positive estimator has the form

$$\theta_i^{JS} = \left(1 - \frac{a}{\|Z\|^2}\right)_+ Z_i$$

when $c_+ = \max(c, 0)$ for any number c. This estimator, however, truncates independently of the magnitude of $|Z_i|$. Indeed, it truncates all or none of the coordinates. To construct an estimator that truncates only the coordinate with small $|Z_i|$'s, it seems necessary to replace a by a decreasing function $h(|Z_i|)$ of $|Z_i|$ and consider

$$\widehat{\theta}_i^+ = \left(1 - \frac{h(|Z_i|)}{D}\right)_+ Z_i$$

where D, independently of i, is yet to be determined. (In a somewhat different approach, Beran and Dümbgen (1998) constructs a modulation estimator corresponding to a monotonic shrinkage factor.) With such a form, $\widehat{\theta}_i^+ = 0$ if $h(|Z_i|) \geq D$, which has a better chance of being satisfied when $|Z_i|$ is small.

We consider a simple choice $h(|Z_i|) = |Z_i|^{-2/3}$, and find a $D = \Sigma |Z_i|^{4/3}$ to solve the differential inequality (0). This leads to the untruncated version $\widehat{\theta}$ with the ith component

$$\widehat{\theta}_i(Z) = Z_i + g_i(Z) \text{ where } g_i(Z) = -aD^{-1} sign(Z_i)|Z_i|^{1/3}. \tag{1}$$

Here and later $sign(Z_i)$ denotes the sign of Z_i. It is possible to use other decreasing functions $h(|Z_i|)$ and other D.

In general, we consider, for a fixed $\beta \leq 2$, an estimator of the form

$$\widehat{\theta}_i = Z_i + g(Z), \tag{2}$$

where

$$g_i(Z) = -a\frac{sign(Z_i)|Z_i|^{\beta-1}}{D} \quad \text{and} \quad D = \sum_{i=1}^{d} |Z_i|^{\beta}. \tag{3}$$

Although at first glance, it may seem hard to justify this estimator, it has a Bayesian and Empricial Bayes justification in Sections 4 and 5. It is also a class of estimators which include, as a special case, the James-Stein estimator corresponding to $\beta = 2$. Now we have

Theorem 2. *For $d \geq 2$ and $1 < \beta \leq 2$, $\widehat{\theta}(Z)$ dominates Z if and only if*

$$0 < a \leq 2(\beta - 1)\inf_{\theta} \frac{E_\theta\big(D^{-1}\sum_{i=1}^{p}|Z_i|^{\beta-2}\big)}{E_\theta(D^{-2}\sum_{i=1}^{p}|Z_i|^{(2\beta-2)})} - 2\beta.$$

Proof: Obviously for $Z_j \neq 0$, $\forall\ j \neq i$, $g_i(Z)$ can be writen as the one–dimensional integral of

$$\frac{\partial}{\partial Z_i} g_i(Z) = \beta(-a)(-1)D^{-2}|Z_i|^{(2\beta-2)} + (\beta-1)(-a)D^{-1}(|Z_i|^{\beta-2})$$

with respect to Z_i. (The only concern is at $Z_i = 0$.) Consider only nonzero Z_j's, $j \neq i$. Since $\beta > 1$, this function however is integrable with respect to Z_i even over an integral including zero.) It takes some effort to prove that $E(|\frac{\partial}{\partial Z_i} g_i(Z)|) < \infty$. However one only needs to focus on Z_j close to zero. Using the spherical–like transformation $r^2 = \sum |Z_i|^\beta$, we may show that if $d > 2$ and $\beta > 1$ both terms in the above displayed expression is integrable.

Now

$$\|g(Z)\|^2 = a^2 D^{-2}\sum_{i=1}^{d}|Z_i|^{2\beta-2}.$$

Hence

$$E_\theta\|Z + g(Z) - \theta\|^2 \leq d, \text{ for every } \theta,$$

if and only if,

$$E_\theta\{2\nabla \cdot g(Z) + \|g(Z)\|^2\} \le 0, \text{ for every } \theta,$$

i.e.,

$$E_\theta\Big(a\Big((2\beta)D^{-2}\sum_{i=1}^{d}|Z_i|^{(2\beta-2)} - (2\beta-2)D^{-1}\sum_{i=1}^{d}|Z_i|^{\beta-2}\Big) + a^2D^{-2}\sum_{i=1}^{d}|Z_i|^{2\beta-2}\Big) \le 0, \quad \text{for every } \theta, \tag{4}$$

which is equivalent to the condition stated in the Theorem. □

Theorem 3. *The estimator $\widehat{\theta}(Z)$ with the ith component given in (2) and (3) dominates Z provided $0 < a \le 2(\beta-1)d - 2\beta$ and $1 < \beta \le 2$.*

Proof: By the correlation inequality

$$d\Big(\sum_{i=1}^{d}|Z_i|^{2\beta-2}\Big) < \Big(\sum_{i=1}^{d}|Z_i|^{(\beta-2)}\Big)\Big(\sum_{i=1}^{d}|Z_i|^{\beta}\Big).$$

Hence

$$\frac{E_\theta\big(D^{-1}\sum_{i=1}^{d}|Z_i|^{\beta-2}\big)}{E_\theta(D^{-2}\sum_{i=1}^{d}|Z_i|^{2\beta-2})} > \frac{E_\theta D^{-1}\sum|Z_j|^{\beta-2}}{\frac{1}{d}E_\theta D^{-1}\sum|Z_i|^{\beta-2}} = d.$$

Hence if $0 < a \le 2(\beta-1)d - 2\beta$, then the condition in Theorem 2 is satisfied, implying domination of $\widehat{\theta}(Z)$ over Z.

The following theorem is a generalization of Theorem 6.2 on page 302 of Lehmann (1983) and Theorem 5.4 on page 356 of Lehmann and Casella (1998). It shows that taking the positive part will improve componentwise. Specifically for an estimator $(\widetilde{\theta}_1(Z), \ldots, \widetilde{\theta}_d(Z))$ where

$$\widetilde{\theta}_i(Z) = (1 - h_i(Z))Z_i,$$

the positive part estimator of $\widetilde{\theta}_i(Z)$ is denoted as

$$\widetilde{\theta}_i^+(Z) = (1 - h_i(Z))_+Z_i.$$

Theorem 4. *Assume that $h_i(Z)$ is symmetric with respect to the ith coordinate, then*

$$E_\theta(\theta_i - \widetilde{\theta}_i^+)^2 \leq E_\theta(\theta_i - \widetilde{\theta}_i)^2.$$

Furthermore, if

$$P_\theta(h_i(Z) > 1) > 0, \tag{5}$$

then

$$E_\theta(\theta_i - \widetilde{\theta}_i^+)^2 < E_\theta(\theta_i - \widetilde{\theta}_i)^2.$$

Proof: Simple calculation shows that

$$E_\theta(\theta_i - \widetilde{\theta}_i^+)^2 - E_\theta(\theta_i - \widetilde{\theta}_i)^2 = E_\theta((\widetilde{\theta}_i^+)^2 - \widetilde{\theta}_i^2) - 2\theta_i E_\theta(\widetilde{\theta}_i^+ - \widetilde{\theta}_i). \tag{6}$$

Let's calculate the expectation by conditioning on $h_i(Z)$. For $h_i(Z) \leq 1$, $\widetilde{\theta}_i^+ = \widetilde{\theta}_i$. Hence it is sufficient to condition on $h_i(z) = b$ where $b > 1$ and show that

$$E_\theta((\widetilde{\theta}_i^+)^2 - \widetilde{\theta}_i^2 \mid h_i(Z) = b) - 2\theta_i E_\theta(\widetilde{\theta}_i^+ - \widetilde{\theta}_i \mid h_i(Z) = b) \leq 0,$$

or equivalently,

$$-E_\theta(\widetilde{\theta}_i^2 \mid h_i(Z) = b) + 2\theta_i E_\theta(\widetilde{\theta}_i \mid h_i(Z) = b) \leq 0.$$

Obviously, the last inequality is satisfied if we can show

$$\theta_i E_\theta(\widetilde{\theta}_i \mid h_i(Z) = b) = \theta_i(1-b)E_\theta(Z_i \mid h_i(Z) = b) \leq 0,$$

or equivalently

$$\theta_i E_\theta(Z_i \mid h_i(Z) = b) \geq 0.$$

We may further condition on $Z_j = z_j$ for $j \neq i$ and it suffices to establish

$$\theta_i E_\theta(Z_i \mid h_i(Z) = b, Z_j = z_j, j \neq i) \geq 0. \tag{7}$$

Given that $Z_i = z_j$, $j \neq i$, consider only the case where $h_i(Z) = b$ has solutions. Due to symmetry of $h_i(Z)$, these solutions are in pairs. Let $\pm y_k$, $k \in K$, denote the solutions. Hence the left hand side of (7) equals

$$\begin{aligned}&\theta_i E_\theta(Z_i \mid Z_i = \pm y_k, k \in K)\\ &\quad = \sum_{k \in K} \theta_i E_\theta(Z_i \mid Z_i = \pm y_k) P_\theta(Z_i = \pm y_k \mid Z_i = \pm y_k, k \in K).\end{aligned}$$

Note that

$$\theta_i E_\theta(Z_i \mid Z_i = \pm y_k) = \frac{\theta_i y_k e^{y_k \theta_i} - \theta_i y_k e^{-y_k \theta_i}}{e^{y_k \theta_i} + e^{-y_k \theta_i}}, \tag{8}$$

which is symmetric in $\theta_i y_k$ and is increasing for $\theta_i y_k > 0$. Hence (8) is bounded below by zero, a bound obtained by substituting $\theta_i y_k = 0$ in (8). Consequently we establish that (6) is nonpositive, implying the domination of $\widetilde{\theta}^+$ over $\widetilde{\theta}$.

The strict inequality of the theorem can be established by noting that the right hand side of (6) is bounded above by $E_\theta[(\widetilde{\theta}_i^+)^2 - \widetilde{\theta}_i^2]$ which by (5) is strictly negative.

Theorem 4 implies the following Corollary.

Corollary 5. *Under the assumption of Theorem 3, Z is dominated by $\widehat{\theta}$ which in turn, is strictly dominated by its positive part $\widehat{\theta}^+$ with ith component*

$$\widehat{\theta}_i^+ = (1 - aD^{-1}|Z_i|^{\beta-2})_+ Z_i. \tag{9}$$

It is interesting to note that estimator (9), for $\beta < 2$, does give zero as the estimator when $|Z_i|$ are small. When applied to the wavelet analysis, it truncates the small wavelet coefficients and shrinks the large wavelet coefficients. The estimator lies in a data chosen reduced model.

Moreover, for $\beta = 2$, Theorem 3 reduces to the classical result of Stein (1981) and (9) to the positive part James-Stein estimator. The above bound of a for domination stated in Theorem 3 works only if $\beta > 1$ and $d > \beta/(\beta - 1)$. Although we cannot provide a domination result for $\beta < 1$, it does not mean that such a result is impossible. We are particularly interested in $\beta > \frac{1}{2}$, since in our experiences with wavelet analysis, β may sometimes be below 1 and is usually large than $\frac{1}{2}$. The asymptotic result in Section 8 only assumes that $\beta > 0$.

Section 3. What is the Largest Possible a?

In wavelet analysis, for a reasonable smooth function, a vast majority of the wavelet coefficients are zero. Based on such information, it seems reasonable to choose an estimator that shrinks the most as long as it does not overshrink. Overshrinking can be prevented as long as the resultant estimator dominates Z. Hence in this section we shall set out to find the largest possible a. The pursuit also yields domination result for $\frac{1}{2} < \beta < 1$. Since ultimately we will recommend the positive part estimator, the reduction in risk will be maximized for small θ_i's, a situation that happens often in the wavelets analysis.

To investigate the largest possible shrinkage, we evaluate the Bayes risk of $\widehat{\theta}$ in (2) and (3), assuming that θ_i are i.i.d. $N(0, \tau^2)$. Note that the difference of the Bayes risk of $\widehat{\theta}$ and Z equals $E\mathcal{D}$, where

$$\mathcal{D} = \sum_{i=1}^{d} \left((Z_i + g_i(Z) - \theta_i)^2 - (Z_i - \theta_i)^2\right) = \sum_{i=1}^{d} (2(Z_i - \theta_i)g_i(Z) + g_i^2(Z)),$$

and

$$g_i(Z) = -a\ sign(Z_i) D^{-1} |Z_i|^{\beta-1}.$$

To calculate the expectation with respect to Z_i and θ_i, we first calculate the conditional expectation given Z_i. Since $E(\theta_i \mid Z_i) = \frac{\tau^2 Z_i}{1+\tau^2}$, we obtain

$$\begin{aligned} E\mathcal{D} = EE(\mathcal{D} \mid Z_1, \ldots, Z_p) &= E\Big(\sum_{i=1}^{d} \Big[2\frac{Z_i}{1+\tau^2} g_i(Z) + g_i^2(Z)\Big]\Big) \\ &= E\Big(\sum_{i=1}^{d} \Big[\frac{-2a|Z_i|^\beta}{(1+\tau^2)D} + \frac{a^2|Z_i|^{2\beta-2}}{D^2}\Big]\Big) \\ &= E\Big(\frac{a^2}{D^2}\sum_{i=1}^{d} |Z_i|^{2\beta-2} - \frac{2a}{(1+\tau^2)}\Big). \end{aligned}$$

Note that $\mathcal{D} \leq 0$ if

$$0 \leq a \leq \frac{\frac{2}{(1+\tau^2)}}{E\big(\frac{1}{D^2}\sum_{i=1}^{d} |Z_i|^{2\beta-2}\big)} \tag{10}$$

where the expectation is taken over Z_i which are i.i.d. and

$$Z_i \sim N(0, 1+\tau^2).$$

Let $\xi_i = Z_i/\sqrt{1+\tau^2}$ and consequently $\xi_i \sim N(0,1)$. We see that condition (10) is equivalent to

$$0 \leq a \leq a_B = 2/\Big(\frac{E\sum_{i=1}^{p} |\xi_i|^{2\beta-2}}{(\sum |\xi_i|^\beta)^2}\Big). \tag{11}$$

Hence we have the following theorem.

Theorem 6. *Assume the prior distribution that θ_i are i.i.d. $N(0,\tau^2)$. Then the Bayes risk of $\widehat{\theta}$ is no greater than $Z = (Z_1, \ldots, Z_p)$ for every τ^2 if and only if $0 \leq a \leq a_B$ where a_B is defined in (11).*

Obviously the bound a_B is a necessary bound for $\widehat{\theta}$ to dominate Z. Our numerical studies not reported here, however, show that it is sufficient for the domination of $\widehat{\theta}$ and hence $\widehat{\theta}_+$ over Z by Theorem 4.

There is a good reason for the domination result of $\widehat{\theta}_+$ when $a = a_B$ intuitively. Note that for every τ^2, and in particular for $\tau^2 \to \infty$, Theorem 6 implies that $\widehat{\theta}$

has Bayes risk no greater than Z. Since $\widehat{\theta}^{+}$ dominates $\widehat{\theta}$, this implies that $\widehat{\theta}^{+}$ tend to have smaller risk than Z for large θ. However $\widehat{\theta}^{+}$ shrinks Z toward the origin, it seems intuitively reasonable that it should have smaller risk than Z for small θ. Consequently its risk should have a good chance to be no greater than Z for all θ. This is similar to the argument of tail minimaxity of Berger (1976).

The normal assumption of θ_i seems limited. However, the domination result of Theorem 6 holds for many other distributions. Indeed for any variance mixture of normal, i.e., taking τ^2 to be random with an arbitrary distribution, Theorem 6 holds. A special case of variance mixture of normal is the multivariate t distribution. That is, θ_i has the same distribution as ξ_i/S. Here, as before, ξ_i are i.i.d. standard normal and S independent of ξ_i's, has the same distribution as $\sqrt{\chi^2_N/N}$ where χ^2_N is a chi–squared random variable with N degrees of freedom.

What is the bound a_B? It is easy to numerically calculate the bound a_B by simulating ξ_i ten thousand times and evaluate a_B. Figure 1 below shows that, for $\beta = 4/3$, a_B is at least as big as $\frac{5}{3}(d-2)$ for virtually all d, since the ratio of a_B to the latter, which is plotted in Figure 1, is always larger than one. This bound $\frac{5}{3}(d-2)$ is more than twice as big as the sufficient bound for $\beta = \frac{4}{3}$ given in Theorem 3.

Putting all these together, we come to the conclusion that the estimator $\widehat{\theta}^{+}$, with ith component

$$\widehat{\theta}_i^{+} = \left(1 - \frac{\frac{5}{3}(d-2)Z_i^{-2/3}}{\sum_{i=1}^{d} Z_i^{4/3}}\right)_{+} Z_i, \tag{12}$$

should have risk smaller than d. For $d = 50$, it is shown that $\widehat{\theta}^{+}$ dominates Z in Figure 2. This estimator when applied to wavelet examples in Section 7 usually produce risks smaller than $\widetilde{\theta}^+$ with $a = \frac{2}{3}(d-4)$, the bound given in Theorem 3 and Corollary 5 for $\beta = 4/3$. Also $\widetilde{\theta}^+$ with a larger shrinkage factor $a = \frac{6}{3}(d-2) = 2(d-2)$ does not do as well for the examples of Section 5 either. This seems to have overshrunk Z. It is interesting that the criterion of dominating Z does provide a very useful guidance in choosing a. Also using the largest possible a for domination leads to the best choice especially in the situation that most of θ_i's are zero as in the wavelet case.

It would be convenient to have an approximate formula for the upper bound a_β for every β. It seems tempting to derive the asymptotic limit of a_β/d as $d \to \infty$, which, for $\frac{1}{2} < \beta \le 2$, equals

$$C_\beta = 2/(E|\xi_i|^{2\beta-2}/(E|\xi_i|^\beta)^2) = \frac{4\left(\Gamma\left(\frac{\beta+1}{2}\right)\right)^2}{\sqrt{\pi}\Gamma\left(\frac{2\beta-1}{2}\right)}. \tag{13}$$

It may seem tempting to use $C_\beta(d-2)$. For the case of $\beta = 4/3$, this is about $(5.17)/[3(d-2)]$ rather than $5/[3(d-2)]$ as suggested by (12). Note that 97% of $(5.17)/3$ is approximately $5/3$. Hence we end up with the suggested formula

$$a = 0.97C_\beta(d-2). \tag{14}$$

Although this formula is suggested by $\beta = 4/3$, further numerical investigation not reported here shows that using (14) for a in (9) leads to a $\widehat{\theta}$ that dominates Z.

4. Approximate Bayesian Justification.

It would seem interesting to justify the proposed estimation from a Bayesian's point of view. To do so, we consider a prior of the form

$$\begin{aligned} \pi(\theta) &= 1 \qquad \|\theta\|_\beta \le 1 \\ &= 1/(\|\theta\|_\beta)^{\beta c},\ \|\theta\|_\beta > 1 \end{aligned}$$

where $\|\theta\|_\beta = (\sum \|\theta_i\|^\beta)^{1/\beta}$, and c is a positive constant which can be specified to match the constant a in (9). In general the Bayes estimator is given by

$$Z + \nabla \log m(Z)$$

where $m(Z)$ is the marginal probability density function of Z. Namely,

$$m(Z) = \int \cdots \int \frac{e^{-\frac{1}{2}\|Z-\theta\|^2}}{(\sqrt{2\pi})^d} \pi(\theta) d\theta.$$

The following approximation follows from Brown (1971), which asserts that $\nabla \log m(Z)$ can be approximated by $\nabla \log \pi(Z)$. The proof is given in the Appendix.

Theorem 7. *With $\pi(\theta)$ and $m(X)$ given above,*

$$\lim_{|Z_i| \to +\infty} \frac{\nabla_i \log m(Z)}{\nabla_i \log \pi(Z)} = 1.$$

Hence by Theorem 7, the ith component of the Bayes estimator equals approximately

$$Z_i + \nabla_i \log \pi(Z) = Z_i - \frac{c\beta |Z_i|^{\beta-1} sign(Z_i)}{\sum |Z_i|^\beta}.$$

This is similar to the untruncated version of $\widehat{\theta}$ in (2) and (3).

5. Empirical Bayes Justification.

Based on several signals and images, Mallat (1989) proposed a prior for the wavelelet coefficients θ_i as the exponential power distribution with the probability density function (p.d.f.) of the form

$$f(\theta_i) = ke^{-|\frac{\theta_i}{\alpha}|^\beta} \tag{15}$$

where α and $\beta < 2$ are positive constants and

$$k = \beta/(2\alpha\Gamma(1/\beta))$$

is the normalization constant. See also Vidakovic (1999, p.194). Using method of moments, Mallat estimated value of α and β to be 1.39 and 1.14 for a particular graph. However, α and β are typical unknown.

It seems reasonable to derive an Empirical Bayes estimator based on this class of prior distributions. First we assume that α is known. Then the Bayes estimator of θ_i is

$$Z_i + \frac{\partial}{\partial Z_i} \log m(Z).$$

Similar to the argument in Theorem 7 and noting that for $\beta < 2$,

$$e^{-|\theta_i + Z_i|^\beta / \alpha^\beta} / e^{-|\theta_i|^\beta/\alpha^\beta} \to 1 \text{ as } \theta_i \to \infty,$$

the Bayes estimator can be approximated by

$$Z_i + \frac{\partial}{\partial Z_i} \log \pi(Z_i) = Z_i - \frac{\beta}{\alpha^\beta} |Z_i|^{\beta-1} sign(Z_i). \tag{16}$$

Note that, under the assumption that α is known, the above expression is also the asymptotic expression of the maximum likelihood estimator of θ_i by maximizing the joint p.d.f. of (Z_i, θ_i). See Proposition 1 of Antoniadis, Leporini and Desquet (2002) as well as (8.23) of Vidakovic (1999). In the latter reference, the sign of Z_i of (16) is missing due to a minor typographic error.

Since α is unknown, it seems reasonable to replace α in (16) by an estimator. Assume that θ_i's are observable. Then by (15) the joint density of $(\theta_1, \ldots, \theta_d)$ is

$$\Big[\frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})}\Big]^d e^{-\Sigma\left(\frac{|\theta_i|^\beta}{\alpha^\beta}\right)}\Big].$$

Differentiating this p.d.f. with respect to α gives the maximum likelihood estimator of α^β as

$$(\beta\Sigma|\theta_i|^\beta)/d. \tag{17}$$

However since θ_i is unknown and hence the above expression can be further estimated by (16). For $\beta < 2$, the second term in (16) has a smaller order than the first when $|Z_i|$ is large. Replacing θ_i by the dominating first term Z_i in (16) leads to an estimator of α^β as $(\beta \Sigma |Z_i|^\beta)/p$.

Substituting this into (16) gives

$$Z_i - \frac{d}{\Sigma |Z_i|^\beta} |Z_i|^{\beta-1} sign(Z_i)$$

which is exactly estimator $\widehat{\theta}_i$ in (2) and (3) with $a = d$. Hence we have succeeded deriving $\widehat{\theta}_i$ as an Empirical Bayes estimator when Z_i is large.

6. Data Estimated β.

Which β should one use in the estimator (9)? It seems reasonable to let the data choose. To do so, let us rewrite (9) as

$$\widehat{\theta}_i^{(\beta)} = \Big(1 - \frac{a_\beta |Z_i|^{\beta-2}}{D_\beta}\Big)_+ Z_i$$

where, to emphasize their dependence on β, we use D_β and a_β to denote D and a which are specified in (3), (13) and (14). One may then calculate SURE, namely the Stein's unbiased estimate of the risk of $\widehat{\theta}^{(\beta)} = (\widehat{\theta}_1^\beta, \ldots, \widehat{\theta}_p^\beta)'$ as

$$SURE = d + \sum_{i=1}^{d} (Z_i^2 - 2) I_i + a_\beta \Big(\frac{a_\beta |Z_i|^{2\beta-2}}{D_\beta^2} - 2(\beta - 1) \frac{|Z_i|^{\beta-2}}{D_\beta} + 2 \frac{\beta |Z_i|^{2\beta-2}}{D_\beta^2} \Big) I_i^c,$$

where $I_i^c = 1 - I_i$ and I_i is one or zero according to whether $a_\beta |Z_i|^{\beta-2} > D_\beta$ or not.

Now $\widehat{\beta}$, the minimizer of SURE, can be used to estimate β in $\widehat{\theta}^{(\beta)}$. The resultant estimator with a data estimated β is denoted as $\widehat{\theta}^S$. Hence

$$\widehat{\theta}^S = \widehat{\theta}^{(\widehat{\beta})} \tag{18}$$

where $\widehat{\beta}$ minimizes SURE. This estimator turns out to have the smallest risk function as will be discussed in Section 7.

7. Connection to the Wavelet Analysis and the Numerical Results.

Wavelets have become a very important tool in many areas including Mathematics, Applied Mathematics, Statistics, and signal processing. It is also applied to numerous other areas of science such as chemometrics and genetics.

In statistics, wavelets have been applied to function estimation with amazing results of being able to catch the sharp change of a function. Celebrated contributions by Donoho and Johnstone (1994 and 1995) focus on developing thresholding techniques and asymptotic theories. In the 1994 paper, relative to the oracle risk, their VisuShrink was shown to be asymptotically optimal. Further in 1995's paper, the expected squared error loss of their SureShrink is shown to nearly achieve the asymptotic minimax rate over Besov spaces. Cai (1999) improved on their result by establishing that the Block James–Stein (BlockJS) thresholding achieve exactly the asymptotic global or local minimax rate over a class of Besov spaces.

Now specifically let $y = (Y_1, \ldots, Y_n)'$ be samples of a function f, satisfying

$$Y_i = f(t_i) + \varepsilon_i \tag{19}$$

where $t_i = (i-1)/n$ and ε_i are independently identically distributed (i.i.d.) $N(0, \sigma^2)$. Here σ^2 is assumed to be known and is taken to be one without loss of generality. See a comment at the end of the paper regarding the unknown σ case. One wishes to choose an estimate $\widehat{f} = (\widehat{f}(t_1), \ldots, \widehat{f}(t_n))$ so that its risk function

$$E\|\widehat{f} - f\|^2 = E\sum_{i=1}^{n}(\widehat{f}(t_i) - f(t_i))^2,$$

is as small as possible. Many discrete wavelet transformations are orthogonal transformations. See Donoho and Johnstone (1995). Consequently, there exists an orthogonal matrix W, such that the wavelet coefficients of Y and f are $Z = WY$ and $\theta = Wf$. Obviously the components Z_i of Z are independent, having a normal distribution with mean θ_i and standard deviation 1. Hence previous sections apply and exhibit many good estimators of θ. Note that, by orthogonality of W, for any estimator $\delta(Z)$ of θ, its risk function is identical to $W'\delta(Z)$ as an estimator of $f = W'\theta$. Hence the good estimators in previous sections can be inversely transformed to estimate f well.

In all the applications to wavelets discussed in this paper, the estimators (including our proposed estimator) apply separately to the wavelet coefficients of the same resolution. Hence in (12), for example, d is taken to be the number of coefficients of a resolution when applied to the resolution. In all the literature that we are aware of, this has been the case as well. Figure 3 gives six true curves (made famous by Donoho and Johnstone) from which the data are generated. For these six cases, Figure 4 plots the ratios of the risks of the aforementioned estimator to n, the risk of Y. Since most relative risks are less than one, this indicates that most estimators perform better than the raw data Y. Our estimators $\widehat{\theta}^{+}$ in (12) and $\widehat{\theta}^{S}$ in (18), however, are the ones that are consistently better than Y. Furthermore, our estimators $\widehat{\theta}^{+}$ and $\widehat{\theta}^{S}$ virtually dominate all the other estimators in risk. Generally, $\widehat{\theta}^{S}$ performs better than $\widehat{\theta}^{+}$ virtually in all cases.

As shown in Figure 4, the difference in risks between $\widehat{\theta}^{+}$ and $\widehat{\theta}^{S}$ are quite minor. Since $\widehat{\theta}^{+}$ is computationally less intensive, we focus on $\widehat{\theta}^{+}$ for the rest of the numerical studies.

Picturewise, our estimator does slightly better than other estimators. See Figure 5 for an example. Note that the picture corresponding to $\widehat{\theta}^{+}$ distinguishes most clearly the first and second bumps from the right.

Based on asymptotic calculation, the next section also recommends a choice of a in (20). It would seem interesting to comment on its numerical performance. The difference between the a's defined in (14) and (20) are very small when $64 \leq p \leq 8192$ and when β is estimated by minimizing SURE. Consequently, for such β, the risk functions of the two estimators with different a's are very similar, with a difference virtually bounded by 0.02. The finite sample estimator (where a is defined in (14)) has a smaller risk about 75% of the times.

James–Stein estimator produces very attractive risk functions, sometimes as good as the proposed estimator (12). However, it does not seem to produce good graphs. Compare Figures 6 and 7.

In the simulation studies, we use the procedures MultiVisu and MultiHybrid which are VisuShrink and SureShrink in WaveLab802. See http://playfair.stanford.edu/∼wavelab. We use Symmlet 8 to do wavelet transformation. In Figure 4, signal to noise ratio (SNR) is taken to be 3. Results are similar for other SNR's. To include block thresholding result of Cai (1999), we choose the lowest integer resolution level $j \geq \log_2(\log n) + 1$.

8. What if the Variance is not Known to be One?

So far, we have been focusing on the case where Z_i has the standard deviation σ known to be one. When σ is known and is not equal to one, a simple transformation applied to the problem suggest that (9) be modified as

$$(1 - a\sigma^2 D^{-1}|Z_i|^{\beta-2})_+ Z_i. \tag{20}$$

Namely, a is replaced by $a\sigma^2$.

In real applications, however, σ^2 is typically unknown. One could then estimate σ by $\widehat{\sigma}$, the proposed estimator for σ in Donoho and Johnston (1995, page 1218) with this modification in (12) and (18), the resultant estimators are not minimax according to some numerical simulations. However, they still perform the best or nearly the best among all the estimators studied in Figure 4.

9. Asymptotic Optimality.

To study the asymptotic rate of a wavelet analysis estimator, it is customary to assume the model

$$Y_i = f(t_i) + \sigma\varepsilon_i, \quad i = 1, \dots, n \tag{21}$$

where $t_i = (i-1)/n$, $\sigma = 1/\sqrt{n}$ and ε_i are assumed to be i.i.d. $N(0,1)$. The estimator $\widehat{f}$ for $f(\cdot)$ that can be proved asymptotically optimal applies estimator (20) with

$$a = (2\ln d)^{(2-\beta)/2} m_\beta, \ 0 \le \beta \le 2, \tag{22}$$

and

$$m_\beta = E|\varepsilon_i|^\beta = 2^{\beta/2} B((\beta+2)/2)\sqrt{\pi},$$

to the wavelet coefficients Z_i of each resolution with dimensionality d of the wavelet transformation of Y_i's. After applying the estimator to each resolution one at a time to come up with the new wavelet coefficient estimators, one then uses the wavelet base function to obtain one function $\widehat{f}$ in the usual way.

To state the theorem, we use $\beta^{\alpha}_{p,q}$ to denote the Besov's space with smoothness α and shape parameters p and q. The definition of the Besov's spce with respect to the wavelet coefficients are given in (A.19). Now the asymptotic theorem is given below.

Theorem 8. *Let α, p, q and $p > \max(\beta, \frac{1}{\alpha})$, then there exists a constant C independent of n and f such that*

$$\sup_{\theta \in B^{\alpha}_{p,q}} E \int_0^1 |f(t) - \widehat{f}(t)|^2 dt \le C(\ln n)^{1-\beta/2} n^{-2\alpha/(2\alpha+1)}. \tag{23}$$

The asymptotic optimality stated in (22) is as good as what has been established for hard and soft thresholding estimators in Donoho and Johnstone (1994), the Garrott method in Gao (1998) and Theorem 4 in Cai (1999) and SCAD method in Antoniadis and Fan (2001). However, the real advantage of our estimator is in the finite sample risk as reported in Section 7. Also our estimators are constructed to be minimax and hence have finite risk functions uniformly smaller than the risk of Z. This estimator $\widehat{\theta}^A$ for $\beta = 4/3$ however has a risk very similar to (12). See Section 7.

References

Antoniadis and Fan (2001), *Regularized wavelet approximations* (with discussion), J. Am. Statist. Ass. **96**, 939–967.

Antoniadis, Leporini and Desquet (2002), Statistica Neerlandica (to appear).

Beran, R. and Dümbgen, L. (1998), *Modulation of estimators and confidence set*, Ann. Statist. **26**, 1826–1856.

Berger, J. (1976), *Tail minimaxity in location vector problems and its applications*, Ann. Statist. **4**, 33–50.

Berger, J. (1980), *Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters*, Annals of Statist. **8**, 545–571.

Brown, L. D. (1971), *Admissible estimators, recurrent diffusions, and insoluble boundary value problems*, Ann. Math. Statist. **42**, 855–903.

Cai, T. (1999), *Adaptive wavelet estimation: A block thresholding and oracle inequality approach*, Ann. of Stat. **27, 3**, 898–924.

Donoho, D. L. and Johnstone, I. (1994), *Ideal spatial adaption via wavelet shrinkage*, Biometrika **81**, 425–455.

Donoho, D. L. and Johnstone, I. (1995), *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Stat. Assoc. **90**, 1200–1224.

Gao, H. Y. (1998), *Wavelet shrinkage denoising using non–negative garrote*, J. Comput. Graph. Statist. **7**, 469–488.

Gauch, H. (1993), *Prediction, parsimony and noise*, American Scientist **81**, 468–478.

George, E. I. (1986a), *Minimax multiple shrinkage estimation*, Ann. Statist. Ass. **14**, 188–205.

George, E. I. (1986b), *Combining minimax shrinkage estimation*, J. Am. Statist. Ass. **81**, 437–445.

James, W. and Stein, C. (1961), *Estimation with quadratic loss*, Proc. Fourth Berkeley Symp. Math. Statist. Probab. **1**, 311–319.

Lehmann, E. L. (1983), *Theory of Point Estimation*, Wiley, New York.

Lehmann, E. L. and Casella, G. C. (1998), *Theory of Point Estimation*, Second edition, Springer-Verlag, New York.

Mallat, S. G. (1989), *A theory for multiresolution signal decomposition: The wavelet representation*, IEEE Trans. on Patt. Anal. Mach. Intell. **11(7)**, 674–693.

Stein, C. (1981), *Estimation of the mean of a multivariate normal distribution*, Ann. Statist. **9**, 1135–1151.

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, John Wiley & Sons, Inc., New York.

Appendix.

Proof of Theorem 8. Before relating to model (21), we shall work on the canonical form:

$$Z_i = \theta_i + \sigma\varepsilon_i,\ \ i = 1, 2, \ldots, d$$

where $\sigma > 0$, and ε_i are independently identically distribution standard normal random errors. Here $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_d)$ denotes the estimator in (20) with a defined in (22). For the rest of the paper C denotes a generic quantity independent of d and the unknown parameters. Hence the C's below are not necessarily identical. We shall first prove Lemma A.1 below. Inequality (A.1) will be applied to the lower resolutions in the wavelet regression. The other two inequalities (A.2) and (A.3) are for higher resolutions.

Lemma A.1. *For any* $0 \le \beta < 2$, $0 < \delta < 1$, *and some* $C > 0$, *independent of* d *and* θ_i*'s, we have*

$$\sum_{i=1}^{d} E(\widetilde{\theta}_i - \theta_i)^2 \le C\sigma^2 d(\ln d)^{(2-\beta)/2}, \tag{A.1}$$

and

$$E(\widehat{\theta}_i - \theta_i)^2 \le C(\theta_i^2 + \sigma^2 d^{\delta-1}(\ln d)^{-1/2})\ \textit{if}\ \sum_{1}^{d} |\theta_i|^\beta \le \sigma^\beta \Big(\frac{2-\beta}{2\beta}\Big)^\beta \delta^2 m_\beta d. \tag{A.2}$$

Here and below, m_β denotes the expectation of $|\varepsilon_i|^\beta$, defined right above the statement of Theorem 8. Furthermore, for any $0 \le \beta < 1$, there exists $C > 0$ such that

$$E(\widehat{\theta}_i^A - \theta_i)^2 \le C\sigma^2 \ln d. \tag{A.3}$$

Proof: Without loss of generality, we will prove the theorem for the case $\sigma = 1$.

By Stein's identity,

$$\begin{aligned} &E(\widehat{\theta}_i - \theta_i)^2 \qquad\qquad (A.4)\\ &= E\Big[1 + (Z_i - 2)I_i + \Big(\frac{a^2|Z_i|^{2\beta-2}}{D} - 2a(\beta-1)\frac{|Z_i|^{\beta-2}}{D} + 2a\beta\frac{|Z_i|^{2\beta-2}}{D^2}\Big)I_i^c\Big]. \end{aligned}$$

Here I_i denotes the indicator function $I(a|Z_i|^{\beta-2} > D)$ and $I_i^c = 1 - I_i$. Consequently

$$I_i = 1 \quad \text{if} \quad |Z_i|^{2-\beta} < a/D, \tag{A.5}$$

and

$$I_i^c = 1 \quad \text{if} \quad a|Z_i|^{\beta-2}/D \leq 1. \tag{A.6}$$

From (A.4), and after some straightforward calculations,

$$\begin{aligned} &E\sum_{i=1}^{d}(\widehat{\theta}_i - \theta_i)^2 \qquad\qquad (A.7)\\ &= d + E\Big[\sum_{i=1}^{d}(|Z_i|^{2-\beta}|Z_i|^{\beta} - 2)I_i + \frac{a|Z_i|^{\beta-2}}{D}\Big(\frac{a|Z_i|^{\beta}}{D} - 2(\beta-1) - 2\beta\frac{|Z_i|^{\beta}}{D}\Big)I_i^c\Big]. \end{aligned}$$

Using this and the upper bounds in (A.5) and (A.6), we conclude that (A.7) is bounded above by

$$d + E\Big[\sum_{i=1}^{d}\frac{a|Z_i|^{\beta}}{D} + \frac{a|Z_i|^{\beta}}{D} + 2\beta\frac{|Z_i|^{\beta}}{D}\Big] + 2|\beta-1|d \leq C(\ln d)^{(2-\beta)/2}d,$$

completing the proof of (A.1).

To derive (A.2) for $1 < \beta < 2$, note that

$$E(1 + (Z_i^2 - 2)I_i) = \theta_i^2 + E(-Z_i^2 + 2)I_i^c.$$

This and (A.4) imply that

$$\begin{aligned} E(\widehat{\theta}_i - \theta_i)^2 = \theta_i^2 &+ E\Big\{\Big[\Big(\frac{a|Z_i|^{\beta-2}}{D}\Big)^2 Z_i^2 - Z_i^2\Big]I_i^c\Big\}\\ &+ E\Big\{\Big[-2(\beta-1)\frac{a|Z_i|^{\beta-2}}{D} + 2\Big]I_i^c\Big\} + E\Big[\Big(2\beta a\frac{|Z_i|^{\beta-2}}{D}\ \frac{|Z_i|^{\beta}}{D}\Big)I_i^c\Big]. \end{aligned}$$

Using (A.7), one can establish that the last expression is bounded above by

$$\theta_i^2 + E[(-2(\beta-1)+2)I_i^c] + E2\beta\frac{|Z_i|^\beta}{D}I_i^c \le \theta_i^2 + E[(4+2\beta)I_i^c] \le \theta_i^2 + 8EI_i^c. \quad \text{(A.8)}$$

We shall show, under the condition in (A.2), that

$$EI_i^c \le C(|\theta_i|^2 + d^{\delta-1}(\log d)^{-1/2}). \quad \text{(A.9)}$$

This and (A.8) obviously establish (A.2). To prove (A.9), we shall consider two cases: (i) $0 \le \beta \le 1$ and (ii) $1 < \beta < 2$. For case (i), note that, for any $\delta > 0$, EI_i^c equals

$$\begin{aligned} P(a|Z_i|^{\beta-2} \le D) = P(D \ge a|Z_i|^{\beta-2}, |Z_i| \le (2\ln d)^{1/2}/(1+\delta)) \\ + P(D \ge a|Z_i|^{\beta-2}, |Z_i| \ge (2\ln d)^{1/2}/(1+\delta)). \end{aligned}$$

Obviously, the last expression is bounded above by

$$P(D \ge (1+\delta)^{2-\beta}dm_\beta) + P(|Z_i| \ge (2\ln d)^{1/2}/(1+\delta)). \quad \text{(A.10)}$$

Now the second term is bounded above by

$$C(|\theta_i|^2 + (d^{1-\delta}\sqrt{\ln d})^{-1}) \quad \text{(A.11)}$$

by a result in Donoho and Johnstone (1994). To find an upper bound for the first term in (A.10), note that by a simple calculus

$$|Z_i|^\beta \le |\varepsilon_i|^\beta + |\theta_i|^\beta$$

due to $0 \le \beta \le 1$. Hence the first term of (A.10) is bounded above by

$$P\Big(\sum_1^d |\varepsilon_i|^\beta \ge (1+\delta)^{2-\beta}dm_\beta - \sum|\theta_i|^\beta\Big).$$

Replacing $\sum |\theta_i|^\beta$ by the assumed upper bound in (A.2), the last displayed expression is bounded above by

$$P\Big(\sum_1^d |\varepsilon_i|^\beta \geq dm_\beta[(1+\delta)^{2-\beta} - (2-\beta)\delta^2]\Big). \tag{A.12}$$

Using the inequality

$$(1+\delta)^{2-\beta} > 1 + (2-\beta)\delta,$$

one concludes that the quantity inside the bracket, is bounded below by

$$1 + (2-\beta)(\delta - \delta^2) > 1.$$

Hence the probability (A.12) decays exponentially fast. This and (A.11) then establish (A.9) for $0 \leq \beta \leq 1$.

To complete the proof for (A.2), all we need to do is to prove (A.9) for case (ii), $1 < \beta < 2$.

Similar to the argument for case (i), all we need to do is to show that the first term in (4.10) is bounded by (4.11). Now applying the triangle inequality

$$D^{1/\beta} \leq \Big(\sum |\varepsilon_i|^\beta\Big)^{1/\beta} + \Big(\sum |\theta_i|^\beta\Big)^{1/\beta}$$

to the first term of (A.10) and using some straightforward algebraic manipulation, we obtain

$$\begin{aligned} &P(D \geq (1+\delta)^{2-\beta} dm_\beta) \\ &\quad \leq P\Big(\sum_1^d |\varepsilon_i|^\beta \geq dm_\beta\Big[\Big\{(1+\delta)^{(2-\beta)/\beta} - \Big(\frac{2-\beta}{2\beta}\Big)\delta^{2/\beta}\Big\}^\beta\Big]\Big). \end{aligned} \tag{A.13}$$

Note that

$$(1+\delta)^{(2-\beta)/\beta} \geq 1 + \frac{(2-\beta)\delta}{2\beta}$$

and consequently the quantity inside the bracket is bounded below by

$$\Big[1+\frac{2-\beta}{2\beta}(\delta-\delta^{2/\beta})\Big]^{\beta}\geq 1+(2-\beta)(\delta-\delta^{2/\beta})/2>1.$$

Now this shows that the probability on the right hand side decreases exponentially fast. Hence inequality (A.9) is established for case (ii) and the proof for (A.2) is now completed.

To prove (A.3) for $0\leq\beta\leq 1$, we may rewrite (A.4) as

$$\begin{aligned}E(\widehat{\theta}_i-\theta_i)^2&=1+E(Z_i^2-2)I_i+E\Big(|Z_i|^{2\beta-2}\Big(\frac{a^2}{D^2}+\frac{2\beta a}{aD^2}\Big)I_i^c\Big)\\&\quad+2(1-\beta)E\Big[\frac{|Z_i|^{\beta-2}a}{D}I_i^c\Big].\end{aligned}\qquad\text{(A.14)}$$

The inequality (A.3), sharper than (A.1), can be possibly established due to the critical assumption $\beta\leq 1$, which implies that

$$|Z_i|^{2\beta-2}<\Big(\frac{a}{D}\Big)^{-(2-2\beta)/(2-\beta)}\quad\text{if}\quad I_i^c=1.\qquad\text{(A.15)}$$

Note that the last term in (A.14) is obviously bounded above by $2(1-\beta)$. Furthermore, replace $|Z_i|^{2\beta-2}$ in the third term on the right hand side of (A.14) by the upper bound in (A.15) and replace Z_i^2 in the second term by the upper bound below

$$|Z_i|^2<(a/D)^{2/(2-\beta)}\text{ when }I_i=1,$$

which follows easily for (A.5). We then obtain an upper bound for (A.14)

$$\begin{aligned}1+E(a/D)^{2/(2-\beta)}&+E\Big[(a/D)^{(2\beta-2)/(2-\beta)}\Big(\frac{a^2}{D^2}+2\frac{\beta a}{D^2}\Big)I_i^c\Big]+2(1-\beta)\\&\leq(3-2\beta)+CE(a/D)^{2/(2-\beta)}.\end{aligned}$$

Here, in the last inequality, $2\beta a/D^2$ was replaced by $2\beta a^2/D^2$. To establish (A.3), obviously the only thing left to do is

$$E(a/D)^{2/(2-\beta)}\leq C\ln(d).\qquad\text{(A.16)}$$

This inequality can be established if we can show that

$$E(d/D)^{2/(2-\beta)} \leq C \tag{A.17}$$

since the definition of a and a simple calculation show that

$$a^{2/(2-\beta)} = Ca^{2/(2-\beta)} \ln(d).$$

To prove (A.17), we apply Anderson's theorem (Anderson 1955) which implies that $|Z_i|$ is stochastically larger than $|\varepsilon_i|$. Hence

$$E(d/D)^{2/(2-\beta)} \leq E\Big[d/\Big(\sum |\varepsilon_i|^\beta\Big)\Big]^{2/(2-\beta)},$$

which is bounded by $A + B$. Here

$$A = E\Big[d/\Big(\sum |\varepsilon_i|^\beta\Big)\Big]^{2/(2-\beta)} I\Big(\sum_1^d |\varepsilon_i|^\beta \leq dm_\beta/2\Big)$$

and

$$B = E\Big[d/\Big(\sum |\varepsilon_i|^\beta\Big)\Big]^{2/(2-\beta)} I\Big(\sum_1^d |\varepsilon_i|^\beta > dm_\beta/2\Big)$$

and as before $I(\cdot)$ denotes the indicator function.

Now B is obviously bounded above by

$$(2/m_\beta)^{2/(2-\beta)} < C.$$

Also by Cauchy–Schwartz inequality

$$A^2 \leq E\Big[d/\big(\sum |\varepsilon_i|^\beta\big)\Big]^{4/(2-\beta)} P\Big(\sum_1^d |\varepsilon_i|^\beta \leq dm_\beta/2\Big) < C.$$

Here the last inequality holds since the probabiity decays exponentially fast. This completes the proof for (A.17) and consequently for (A.3).

Now we apply Lemma A.1 to the wavelet regression. Applying a discrete wavelet transformation to model (21) gives a double index data

$$Z_{jk} = \theta_{jk} + \varepsilon_{jk}/\sqrt{n}, \; k = 1, \dots , 2^j, \tag{A.18}$$

where ε_{jk}'s are i.i.d. standard normal random variables. Also assume that θ's live in the Besov's space with smoothness α and shape parameters p and q, i.e.,

$$\sum_j 2^{jq(\alpha+1/2-1/p)} \Big(\sum_k |\theta_{jk}|^p \Big)^{q/p} \le M^q \tag{A.19}$$

for some positive constants α, p, q and M. The estimator $\widehat{\theta}$ below for model (A.18) refers to (20) with a defined in (22) and $\sigma^2 = 1/n$. For such a $\widehat{\theta}$, the total risk can be decomposed into the sum of the following three quantities:

$$R_1 = \sum_{j<j_0} \sum_k E(\widehat{\theta}_{jk} - \theta_{jk})^2,$$

$$R_2 = \sum_{J>j\ge j_0} \sum_k E(\widehat{\theta}_{jk} - \theta_{jk})^2$$

and

$$R_3 = \sum_{j\ge J} \sum_k E(\widehat{\theta}_{jk} - \theta_{jk})^2$$

where $j_0 = [\log_2(C_\delta n^{1/(2\alpha+1)})]$, and C_δ is a positive constant to be specified later. Applying (A.1) to R_1, which corresponds to the risk of low resolutions, we establish some simple calculation

$$R_1 \le C(\ln n)^{(2-\beta)/2} n^{-2\alpha/(2\alpha+1)}. \tag{A.20}$$

For $j \ge j_0$, (A.19) implies

$$\sum_k |\theta_{jk}|^p \le M^p 2^{-jp(\alpha+1/2-1/p)} = M^p 2^j 2^{-jp(\alpha+1/2)}. \tag{A.21}$$

Furthermore, for $p \geq \beta$

$$2^{-jp(\alpha+1/2)} \leq 2^{-j\beta(\alpha+1/2)} \leq 2^{-j_0\beta(\alpha+1/2)} = (C_\delta)^{-\beta(\alpha+1/2)}\sigma^\beta.$$

Choose $C_\delta > 0$ such that

$$M^p / C_\delta^{(1/2+\alpha)\beta} = \Big(\frac{2-\beta}{2\beta}\Big)^\beta \Big(\frac{1}{2\alpha+1}\Big)^2 m_\beta.$$

This then implies that

$$\begin{aligned}\sum_k |\theta_{jk}|^p &\leq \frac{M^p}{C_\delta^{(1/2+\alpha)\beta}} 2^j \sigma^\beta \\ &\leq \Big(\frac{2-\beta}{2\beta}\Big)^\beta \Big(\frac{1}{2\alpha+1}\Big) m_\beta 2^j \sigma^\beta,\end{aligned}$$

satisfying the condition in (A.2) for $d = 2^j$ and $\delta = (2\alpha+1)^{-1}$.

Now for $p \geq 2$ we give an upper bound for the total risk.

From (A.2), we obtain

$$R_2 + R_3 \leq C \sum_{j \geq j_0} \sum_k \theta_{jk}^2 + o(n^{-2\alpha/(2\alpha+1)})$$

and from Holder inequality the first term is bounded above by

$$\sum_{j \geq j_0} 2^{j(1-2/p)} \Big(\sum_k |\theta_{jk}|^p\Big)^{2/p}.$$

Then inequality (A.21) gives

$$\begin{aligned}R_2 + R_3 &\leq C \sum_{j \geq j_0} 2^{j(1-2/p)} 2^{-j2(\alpha+1/2-1/p)} + o(n^{-2\alpha/(2\alpha+1)}) \\ &= C \sum_{j \geq j_0} 2^{-j2\alpha} + o(n^{-2\alpha/(2\alpha+1)}) \\ &\leq C n^{-2\alpha/(2\alpha+1)}.\end{aligned}$$

This and (A.20) imply (23) for $0 \leq \beta \leq 2$ and $p \geq 2$.

Note that for $\beta = 2$, the proof can be found in Donoho and Johnstone (1995). For $\beta \neq 2$, our proof is very different and much more involved.

To complete the proof of the theorem, we now focus on the case $0 \leq \beta \leq 1$, and $2 > p \geq \max\{1/\alpha, \beta\}$ and establish (23). We similarly decompose the risk of $\widehat{\theta}$ as the sum of R_1, R_2 and R_3. Note that the bound for R_1 in (A.20) is still valid. Inequalities (A.2) and (A.3) imply

$$R_2 \leq \sum_{J \geq j \geq j_0} \sum_k \theta_{jk}^2 \wedge \frac{\log n}{n} + o\Big(\frac{1}{n^{1-\delta}}\Big)$$

for some constants $C > 0$. Furthermore, the following inequality

$$\sum x_i \wedge A \leq A^{1-t} \sum x_i^t, \; x_i \geq 0, \; A > 0, \; 1 \geq t > 0$$

implies

$$\sum_{J \geq j \geq j_0} \sum_k \theta_{jk}^2 \wedge \frac{\log n}{n} \leq \Big(\frac{\log n}{n}\Big)^{1-p/2} \sum_{J > j \geq j_0} \sum_k |\theta_{jk}|^p.$$

Some simple calculations, using (A.21), establish

$$\begin{aligned} R_2 &\leq C\Big(\frac{\log n}{n}\Big)^{1-p/2} \sum_{J > j \geq j_0} 2^{-jp(\alpha+1/2-1/p)} + o(n^{-2\alpha/(2\alpha+1)}) \\ &\leq C(\log n)^{1-p/2} n^{-2\alpha/(2\alpha+1)}. \end{aligned} \tag{A.22}$$

From Holder inequality, it can be seen that R_3 is bounded above by

$$\sum_{j \geq j_0} \Big(\sum_k |\theta_{jk}|^p\Big)^{2/p}.$$

Similar to (A.22), we obtain the upper bound of R_3,

$$R_3 \leq C \sum_{j \geq J} 2^{-j2(\alpha+1/2-1/p)} = o(n^{-2\alpha/(2\alpha+1)}),$$

where J is taken to be $\log_2 n$. Thus for $0 \leq \beta \leq 1$ nd $2 \geq p \geq \max\{1/\alpha, \beta\}$, we have

$$\sup_{f \in B_{p,q}^\alpha} E\|\widehat{\theta} - \theta\|^2 \leq C(\log n)^{1-\beta/2} n^{-2\alpha/(2\alpha+1)}.$$

Acknowledgment. The authors wish to thank Professor Martin Wells at Cornell for his interesting discussions and suggestions which led to a better version of this paper. They also wish to thank Professors L. D. Brown, Dipak Dey, Ed George, Mark Low and William Strawderman, for their encouraging comments. The authors would like to thank Arletta Havlik for tirelessly typing this manuscript.

Figure 1. Ratio of a_B in (11) to 5(d-2)/3

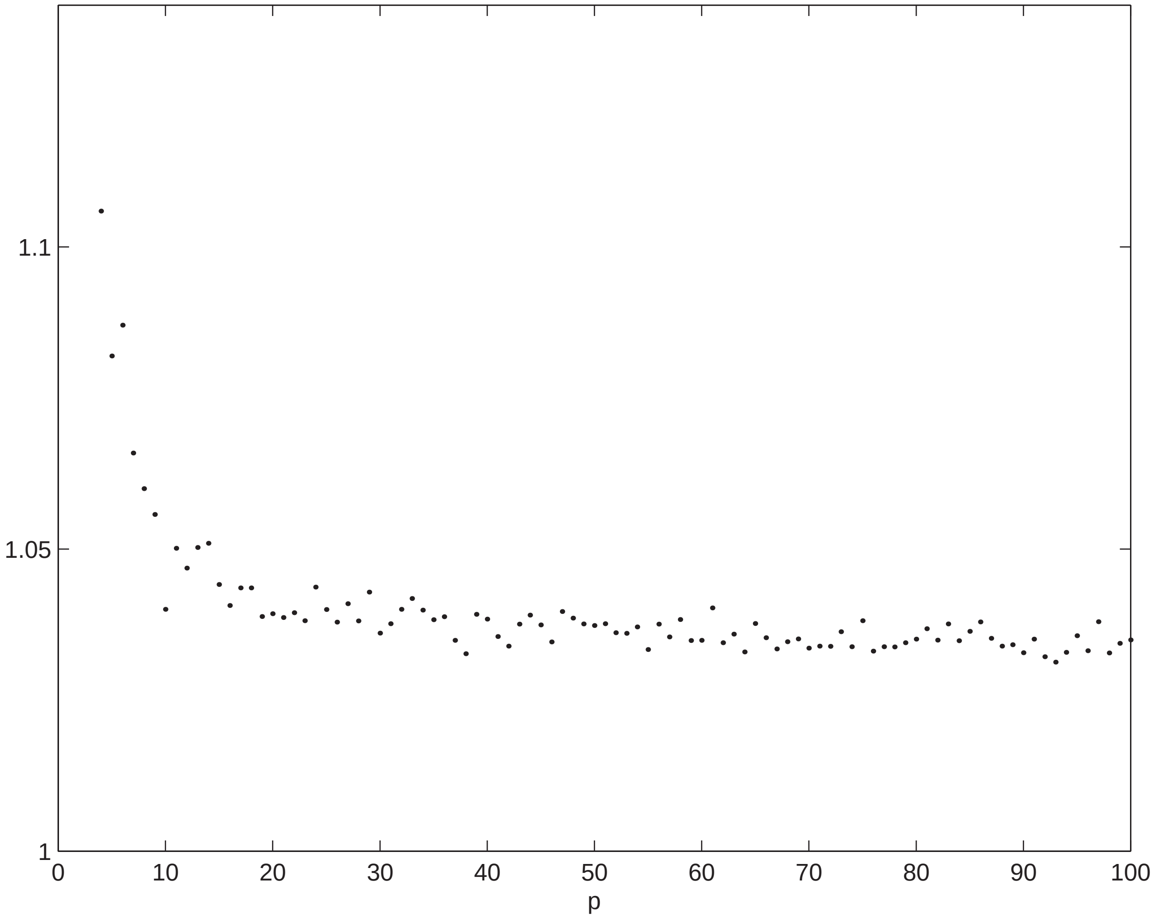


Figure 2. Relative Frequentist risk of the proposed estimator (12) to d=50

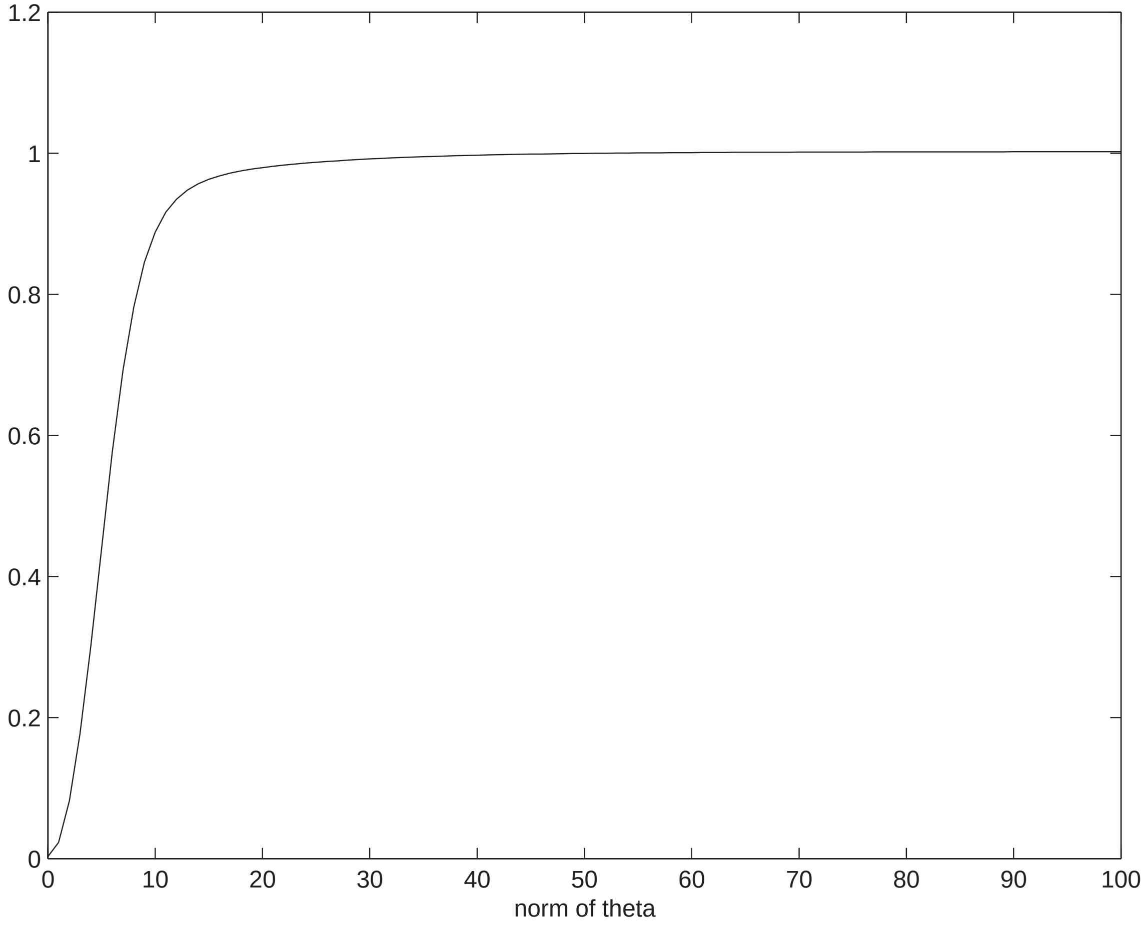


Figure 3. The curves represent the true curves f(t) in (19).

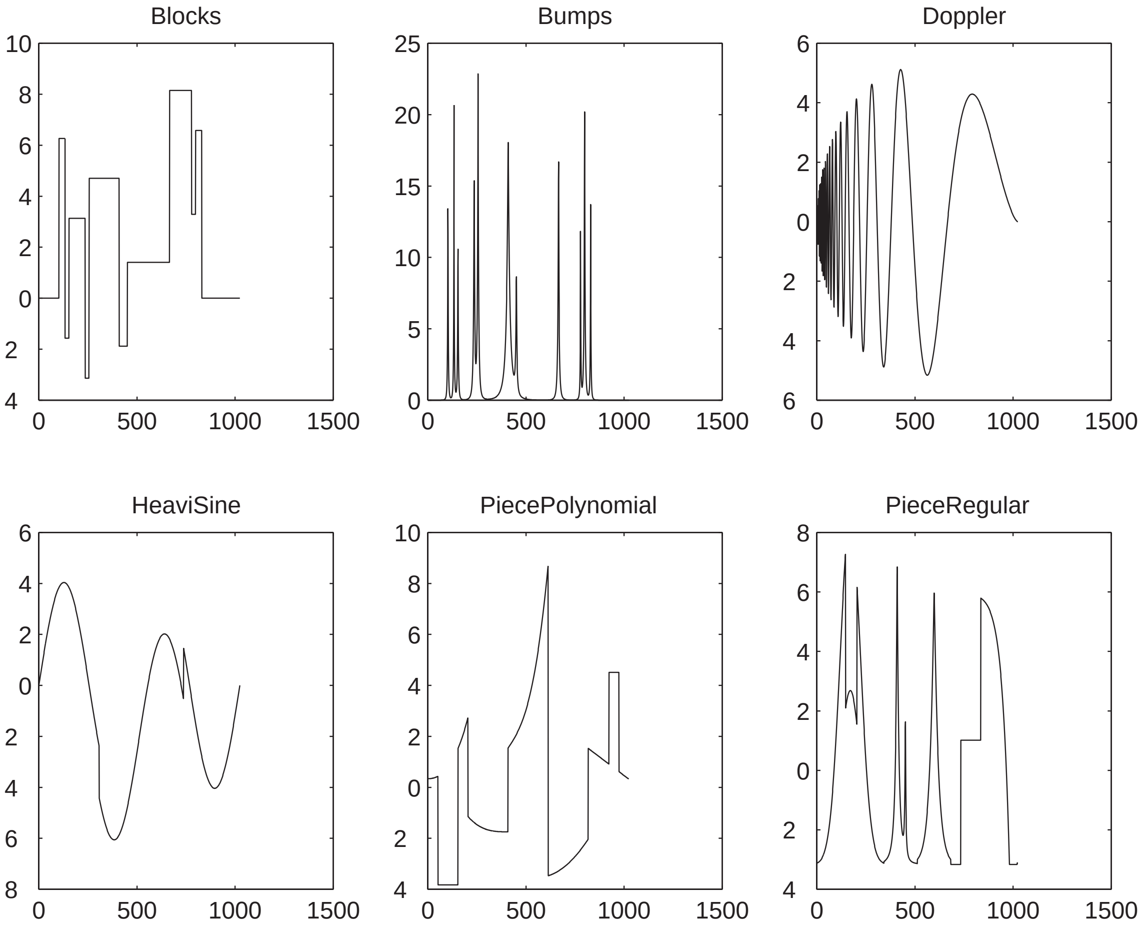


Figure 4. In each of the six cases corresponding to Blocks, Bumps, etc., the eight curves plot the risk function, from top to the bottom, when $n = 64, 128, \ldots, 8192$. For each curve (see for example, the top curve on the left), the circles "o" from left to the right give, with respect to n, the relative risks of VisuShrink, Block James–Stein, SureShrink, and the proposed methods (12) and (18).

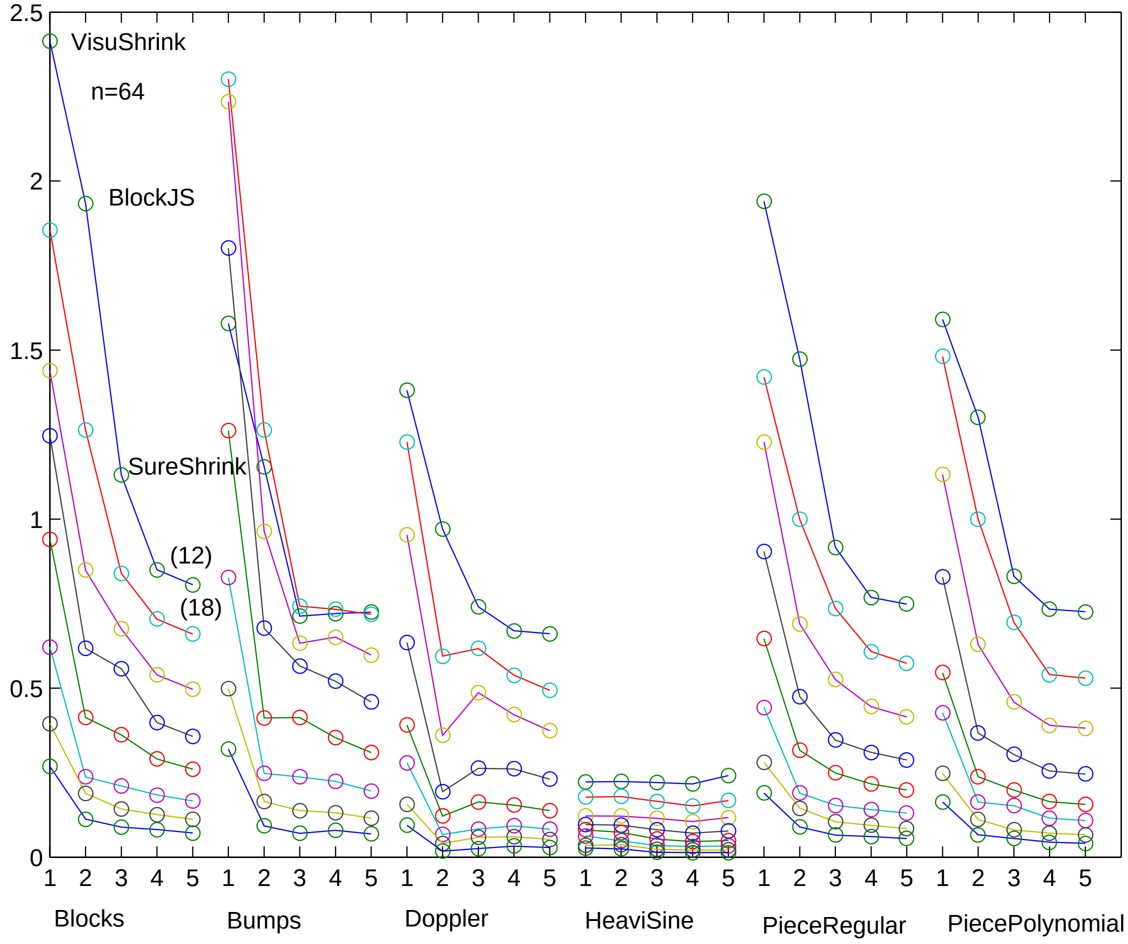
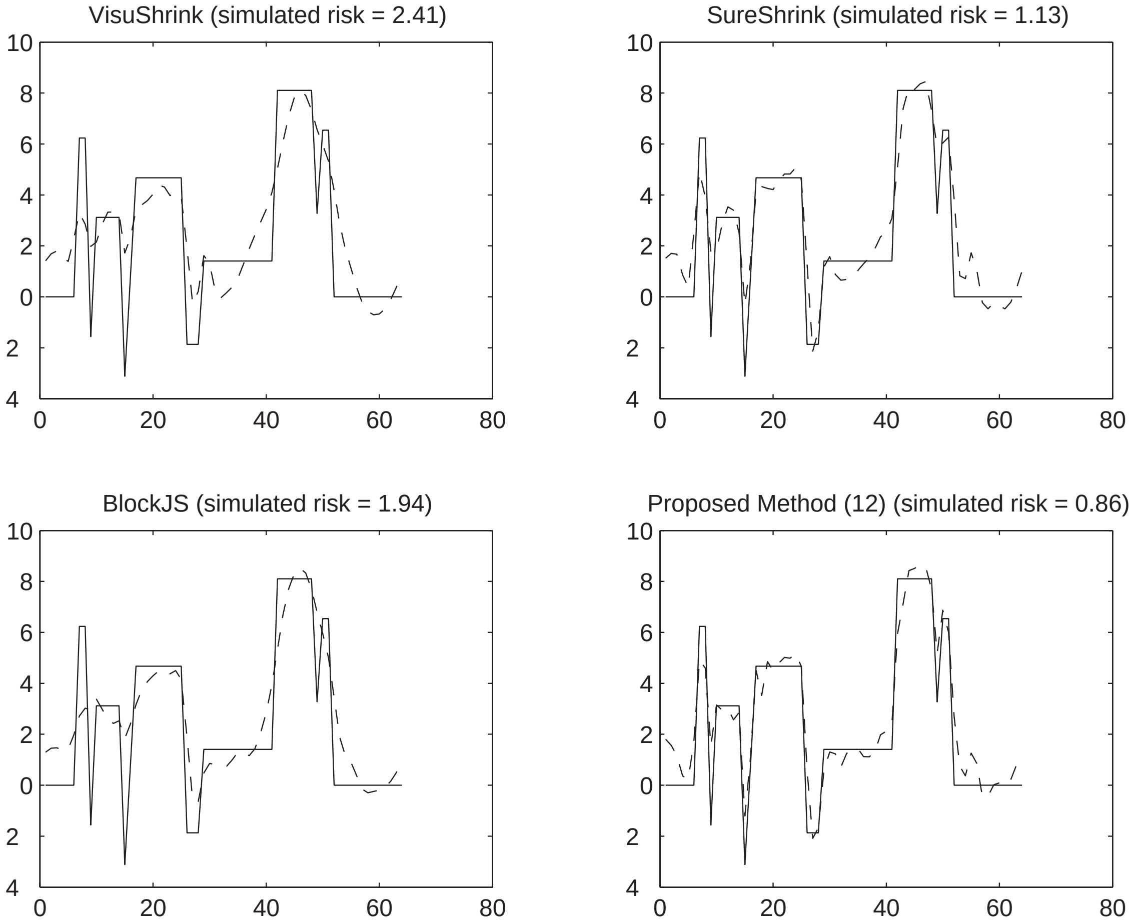


Figure 5. Solid lines represent the true curves, where dotted lines represent the curves corresponding to various estimators. The simulated risk is based on 500 simulations.



VisuShrink (simulated risk = 2.41)

SureShrink (simulated risk = 1.13)

BlockJS (simulated risk = 1.94)

Proposed Method (12) (simulated risk = 0.86)

Figure 6. Proposed Estimator (12) Applied to Reconstruct Figure 3.

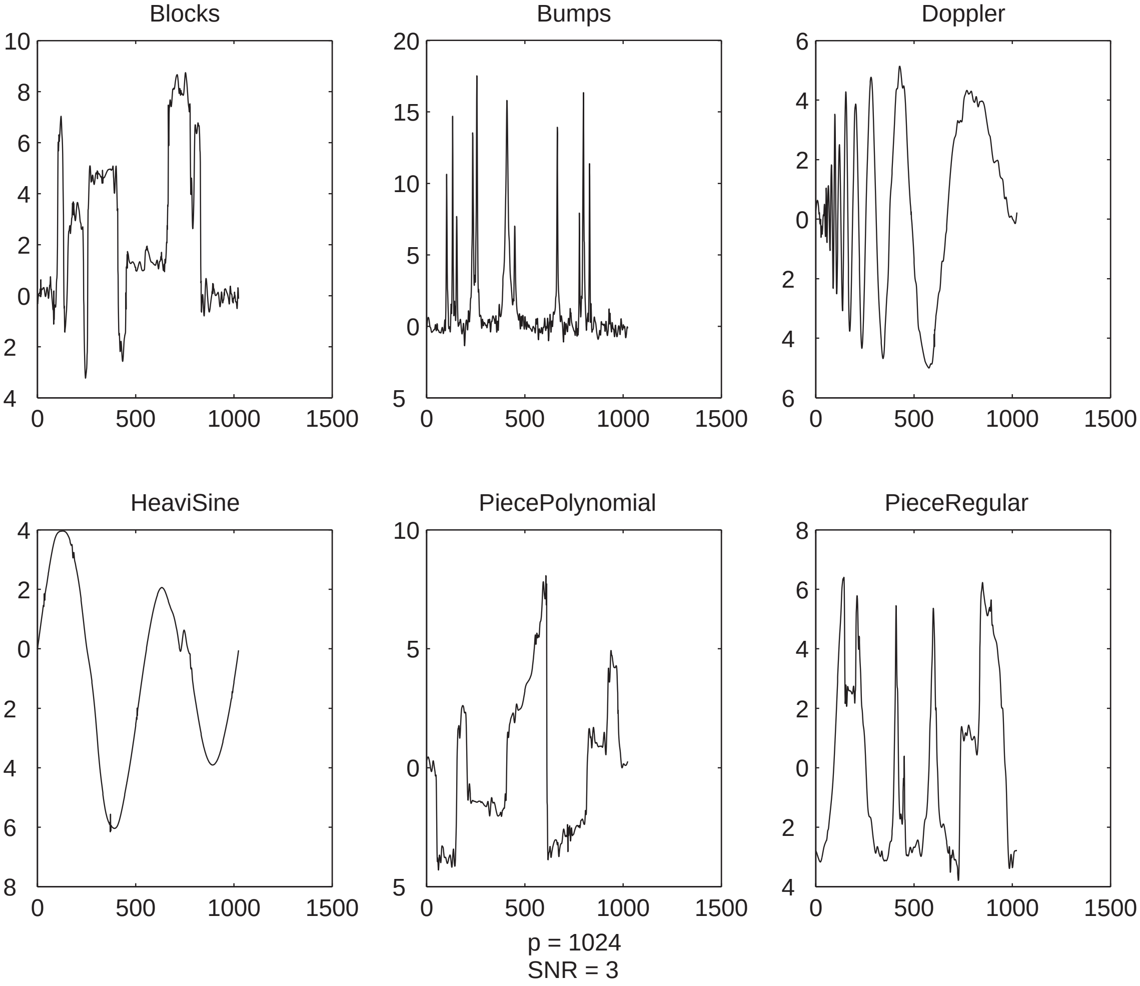
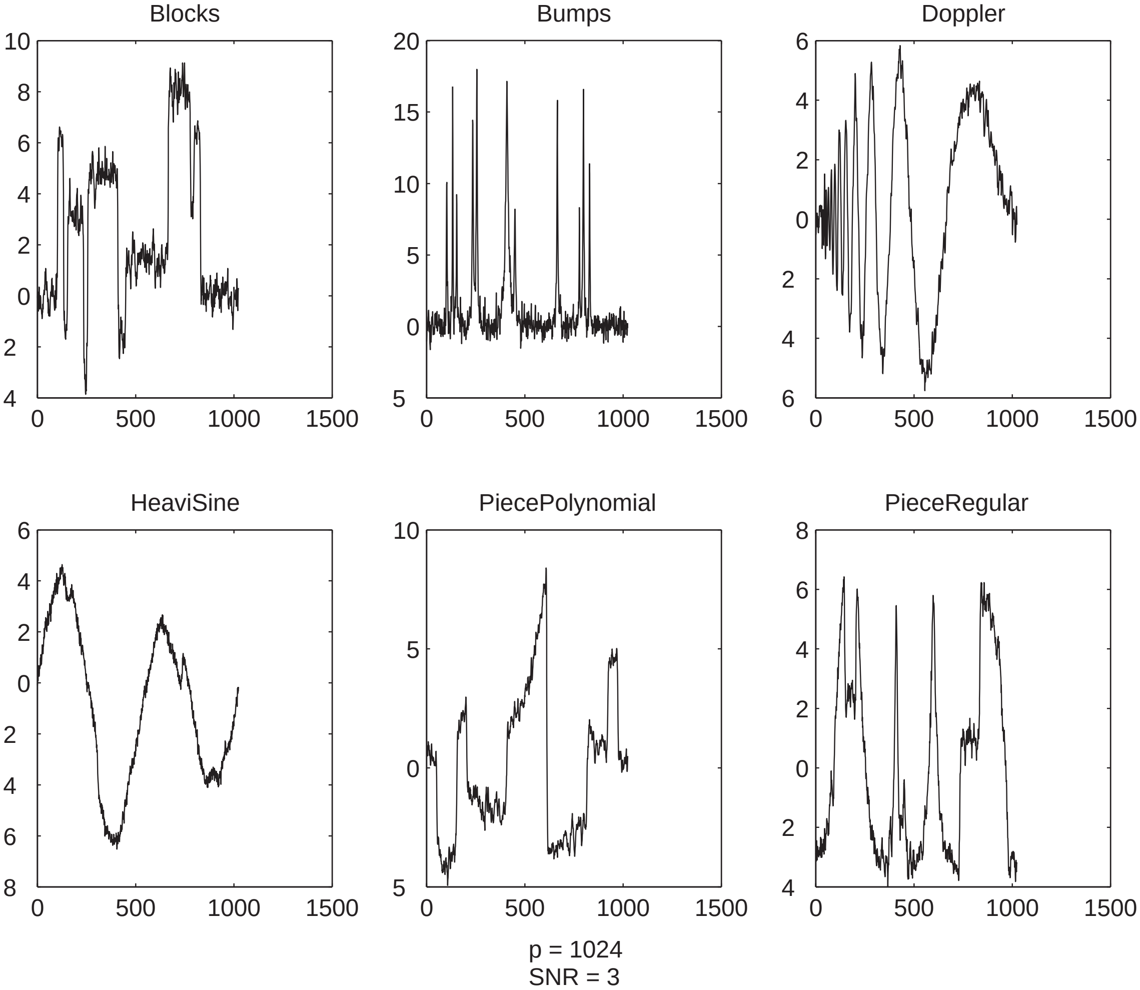


Figure 7. JamesStein Positive Part Applied to Reconstruct Figure 3.


Blocks
Bumps
Doppler
HeaviSine
PiecePolynomial
PieceRegular
p = 1024
SNR = 3