WAVELET ESTIMATION FOR NONPARAMETRIC REGRESSION: BEYOND GAUSSIAN NOISE II

by

Yuhai Wu
Purdue University

Technical Report #99-03

Department of Statistics
Purdue University
West Lafayette, IN USA

April 1999

*

WAVELET ESTIMATION FOR NONPARAMETRIC REGRESSION: BEYOND GAUSSIAN NOISE II

by

Yuhai Wu
Purdue University

Abstract

Wavelet shrinkage estimation for the nonparametric regression problem with given noise that is not necessarily Gaussian, is discussed. The procedure provides a method to shrink wavelet coefficients based on a threshold value which depends on both the noise distribution and the underlying mother wavelet. Point process theory and domain of attraction theory play a crucial role in the derivation. This method extends DJ's wavelet method for the case when the regression noise follows i.i.d. Gaussian distribution. The risk of the new procedure is compared with the ideal risk. It is shown that for any noise which belongs to the maximum domain of attraction of the Gumbel distribution, the risk of the new constructed estimator is within a log term of the ideal risk. Therefore, wavelet shrinkage method is successfully applied to a wide family of Non-Gaussian noise distributions. The constructed estimator is proved to be "noise-free", and has the ability to adapt to broad function classes. Further more, its risk is closed to "ideal" risk. This procedure can also be similarly extended from the Gumbel distribution to other heavy tailed noise distributions.

WAVELET ESTIMATION FOR NONPARAMETRIC REGRESSION BEYOND GAUSSIAN NOISE II

YUHAI WU
DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY

ABSTRACT. Wavelet shrinkage estimation for the nonparametric regression problem with given noise that is not necessarily Gaussian, is discussed. The procedure provides a method to shrink wavelet coefficients based on a threshold value which depends on both the noise distribution and the underlying mother wavelet. Point process theory and domain of attraction theory play a crucial role in the derivation. This method extends DJ's wavelet method for the case when the regression noise follows i.i.d. Gaussian distribution. The risk of the new procedure is compared with the ideal risk. It is shown that for any noise which belongs to the maximum domain of attraction of the Gumbel distribution, the risk of the new constructed estimator is within a log term of the ideal risk. Therefore, wavelet shrinkage method is successfully applied to a wide family of Non-Gaussian noise distributions. The constructed estimator is proved to be "noise-free", and has the ability to adapt to broad function classes. Further more, its risk is closed to the "ideal" risk. This procedure can also be similarly extended from the Gumbel distribution to other heavy tailed noise distributions.

1. INTRODUCTION

Suppose we are given data

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n\ (n = 2^J), \tag{1}$$

where $x_i = \frac{i}{n+1}$, $\epsilon_1, \ldots, \epsilon_n$ have mean zero, variance σ^2, and f is the function to be estimated from the data. A variety of nonparametric methods have been proposed in the literature. Most of them are based on smoothing techniques, such as kernel estimation, spline smoothing, or Fourier series expansion. Although these methods use different smoothing techniques, most of them estimate the function f through

a linear combination of the observation data $\{y_i\}$. In recent years, wavelets have been successfully used for function estimation. These methods have been quickly developed and are becoming more and more important in nonparametric estimation. The method based on the wavelet transform can be described as follow:

(1) Transform the observation data y_i into the wavelet domain.
(2) Keep or discard the resulting wavelet coefficients by hard or soft thresholding.
(3) Transform the estimated wavelet coefficients back into the original domain to form the estimator.

This approach differs significantly from the linear smoothing methods mentioned before in that it uses the data nonlinearly. It has been proved that this approach has various optimal or near optimal properties in comparison to the linear methods when the noise $\{\epsilon_i\}$ are i.i.d. and normally distributed [6, 9, 7]. Due to the special multiresolution structure of the orthogonal wavelet basis, after the orthogonal wavelet transform on the observations with i.i.d. Gaussian noise, the wavelet coefficients are still i.i.d. and Gaussian. Therefore, the optimal properties of wavelet shrinkage method have been proved through normal related minimax theory, Bayes theory, and decision theory [6]. However, when the noise is i.i.d. and non-Gaussian, we lose the i.i.d. property of the noise coefficients in the wavelet domain. Even worse is that the wavelet basis itself is involved in the transformed noise distributions. Therefore, the distributions of the empirical wavelet coefficents now become very complicated, and this makes the selection of threshold

values much harder, Wu(1998a). Through computing the wavelet coefficients' distributions, Wu(1998a) constructed the wavelet shrinkage estimator for the i.i.d. shifted exponentially distributed noise, and he showed that the new estimator has some noise-free visual advantages, it can adapt to broad function classes, and its risk is close to the ideal risk. Hence, the wavelet shrinkage method can be effectively used to estimate regression curves for non-Gaussian noise, and the new shrinkage estimators share similar properties of those when the noise is i.i.d. Gaussian.

In this paper, we will show how to construct wavelet shrinkage estimators when the regression noise has any kind of distribution that belongs to a large distribution family. Our method takes the advantage of extreme value theory, Poisson processes, and combines them with the theory of optimal recovery. Since this method only requires the information of the domain of attraction of the noise distribution, it works for a wide distribution family which contains almost all of the interesting distribution in statistics. To give a clear idea about this method and its power, we will demonstrate it through discussing the following special family:

the distribution function $F(x)$ of the noise $\{\epsilon_i\}$ has the following tail expression

$$(2) \qquad \mathcal{F} = \{F : P(\epsilon_i > x) = Kx^{\alpha}e^{-Ax^{\beta}} \quad \alpha \geq 0,\ \beta, K, A > 0\}.$$

This is a rich family that includes most of the interesting distribution functions in statistics, such as normal, gamma, weibull, double exponential and so on. We will prove that any noise distribution belonging to this family can be effectively estimated by the wavelet shrinkage

method through selecting a proper threshold level. And it works for a wide range of smoothness spaces that the regression curve belongs to, such as the Besov class, or the Triebel class, only if the wavelet basis is an unconditional basis [6]. These cover most of the traditional functional classes such as Holder classes and Sobolev classes [5].

Our first result is summarized in the following theorem.

Theorem 1. *If the noise in (1) belongs to the tail distribution family $\mathcal{F}$, then the optimal threshold values for the wavelet shrinkage procedure are equal to*

$$\lambda_k = 2^{\frac{k-J}{2}} \max_{1 \le i \le S} \{a_i\} \left(\frac{log(n)}{A} \right)^{\frac{1}{\beta}} (1 + o(1)), \tag{3}$$

where a_i is the maximum absolute value of the mother wavelet function and S is the vector size of the mother wavelet.

When the noise distribution is Gaussian, the above formula will reduce to $\sqrt{2log(n)}$ except for the constant term from the mother wavelet function. The soft thresholding estimators are constructed by applying the expression $T_{\lambda_k}(y) = sgn(y)(|y| - \lambda_k)_+$ to the empirical wavelet coefficients. We recover the regression function $\hat{f}(t)$ by using the inverted wavelet transformation. In this paper, we will focus on using the soft threshold method. The corresponding hard threshold method can also be similarly used.

The constructed estimators have two properties similar to those of Gaussian noise as described in [5].

1. With high probability, $\hat{f}(t)$ is at least as smooth as f, with smoothness measured by any of a wide range of smoothness measures.

2. $\hat{f}(t)$ achieves almost the idealized mean square error over a wide range of smoothness classes.

To precisely express the properties of the constructed estimators, we first need to specify the range of the smoothness space. For an orthogonal wavelet basis of $L^2[0,1]$ with Dth highest derivative and M vanishing moments, let S represent the scale of all the Besov spaces $B^{\nu}_{p,q}[0,1]$ and all the Triebel spaces $F^{\nu}_{p,q}[0,1]$ which embed continuously in $C[0,1]$, so that $\nu > 1/p$, and for which the wavelet basis is an unconditional basis, so that $\nu < \min(D, M)$. The smoothnesses of these function classes are measured by the norms $\|.\|_{B^{\nu}_{p,q}}$ or $\|.\|_{F^{\nu}_{p,q}}$. This covers a large number of function classes, the traditional Holder classes and Sobolev classes are the special cases $B^{\nu}_{\infty,\infty}$ and $F^{\nu}_{p,2}$ respectively.

Theorem 2. *Let $\hat{f}(t)$ be the estimated function constructed by our procedure on $[0,1]$. There exist universal constants (π_n) with $\pi_n \to 1$ as $n \to \infty$ so that*

$$Pr\left\{\|\hat{f}\|_F \leq \|f\|_F \ \forall F \in S\right\} \geq \pi_n.$$

In words, $\hat{f}(t)$ is simultaneously as smooth as f in every smoothness class F taken from the scale S asymptotically.

This theorem says that if the function f is identically equal to zero on [0,1], which means that the signal is a pure noise, then with high probability, our estimator is zero since $\|0\|_F = 0$. From the visual or denoising point of view, our estimator is "noise-free". For the estimation of an one dimension parameter θ, a noise-free estimator is similarly expressed as $|\hat{\theta}| \leq |\theta|$.

The second property is expressed in term of adaptivity. Let $F[0,1]$ be a function class and let F_C denote the ball of functions $\{f : \|f\|_F \leq C\}$. Since the error $E\|\hat{f} - f\|_{l_n^2}^2$ depends on f, we use the worst behavior of our estimator to evaluate the performance, namely

$$\sup_{F_C} n^{-1} E\|\hat{f} - f\|_{l_n^2}^2;$$

and we try to do as well as we can to achieve to the minimax mean square error

$$\inf_{\hat{f}} \sup_{F_C} n^{-1} E\|\hat{f} - f\|_{l_n^2}^2,$$

or at least we try to get close to the optimal rate. According to the Parseval relation

$$E\|\hat{f} - f\|_{L_2}^2 = E\|\hat{\theta}_i - \theta_i\|_{l_2}^2,$$

we can equivalently consider the estimation of the sequence θ_i instead of discussing the estimation of the function f directly.

The risk performance can also be evaluated by using the "ideal risk" as a benchmark, see [7]. Assume $\omega_i = \theta_i + \epsilon z_i \quad (i = 1, \ldots, n)$. The ideal estimator is defined as

$$T_{DP}(\omega, \theta) = (\delta_i \omega_i)_{i=1}^n, \quad \delta_i \in \{0, 1\},$$

where $\delta_i = 1_{(|\theta_i| > \epsilon)}$. This estimator pretends that the unknown parameters θ_i are given, and then it retains or throws away ω_i by comparing the signals with the noise level. It yields the ideal risk

$$R_\epsilon(DP, \theta) = \sum_{i=1}^n min(\theta_i^2, \epsilon^2). \tag{4}$$

As $(\theta_i)_{i=1}^n$ is what we want to estimate, we do not have this kind of "oracle" information in practice. Hence, in general the ideal risk $R_\epsilon(DP, \theta)$

cannot be attained for all θ by any estimator. However, we can compare the risk of an estimator with the ideal risk to assess how well the estimator does. The following theorem shows that our threshold estimators are always within a log term of the sample size to the ideal risk for a broad class of functions.

Theorem 3. *For each ball F_C arising from an $F \in S$, there is a constant $C_1(F_C, \psi)$ which does not depend on n, such that $\forall f \in F_C$*

$$\sup_{F_C} n^{-1} E\|\hat{f} - f\|_{l_n^2}^2 \leq C_1 log(n)^{2/\beta} R_\epsilon(DP, \theta).$$

The estimator $\hat{f}$ which is constructed through observations only performs close to the ideal risk over every Besov, Triebel class with the scale S.

For practical problems, we can modify the threshold values according to the following formula.

Corollary 1. *If the noise in (1) belongs to $\mathcal{F}$, then the threshold values given in (3), for the wavelet shrinkage procedure, can be replaced by*

$$\lambda_k = 2^{\frac{k-J}{2}} \max_{1 \leq i \leq S} \{a_i\} \left(\frac{log(n_k)}{A} \right)^{\frac{1}{\beta}} (1 + o(1)), \tag{5}$$

where a_i is the maximum absolute value of the mother wavelet function and n_k is the sample size on the kth resolution level of the wavelet domain.

In the next two sections, we introduce the extreme value theory and the optimal recovery theory, respectively. These are then applied in Section 4 to prove our main results. Section 5 contains some discussion of several related issues. We leave some proofs to the appendix.

2. Extreme value theory

2.1. **Domain of atrractions.** The asymptotic distribution of the properly normalized maximum term of a sequence of random variables has been widely studied in the literature see [16, 22]. Gendenko(1943) has completely characterized the possible nondegenerate limiting distributions and their respective domains of attraction for i.i.d. sequences.

Lemma 1. *Suppose* $\{X_n,\ n \geq 1\}$ *is an i.i.d. sequence of random variables with common distribution* $F(x)$. *Set* $M_n = V_{i=1}^n X_i = \max\{X_1, X_2, \ldots, X_n\}$. *The distribution function of* M_n *is* $F^n(x)$. *If there exist* $a_n > 0$, $b_n \in (-\infty, \infty)$, $n \geq 1$ *such that*

$$P(\frac{M_n - b_n}{a_n} \leq x) = F^n(a_n x + b_n) \to G(x),$$

then the normalized maximum term $\frac{M_n - b_n}{a_n}$ *converges weakly to an nondegenerate distribution* $G(x)$ *and* G *belongs to one of the following three classes:*

(i)

$$\Phi_\alpha(x) = \begin{cases} 0 & x < 0, \\ \exp\{-x^{-\alpha}\} & x \geq 0, \end{cases}$$

for some $\alpha > 0$;

(ii)

$$\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{-\alpha}\} & x < 0, \\ 1 & x \geq 0, \end{cases}$$

for some $\alpha > 0$;

(iii) $\Lambda = \exp\{-e^{-x}\} \quad x \in R.$

Φ_α, Ψ_α, *and* Λ *are called the extreme value distributions.*

The independence requirement of the stochastic sequence $\{X_n\}$ can be weaken by assuming that $\{X_n\}$ is strictly stationary and that the dependence between X_i and X_j decreases in some fashion as $|i-j|$ increases. The simplest generalization of Gnedenko's results is that of m dependence, which requires that X_i and X_j be actually independent if $|i-j| > m$ see Watson(1954). Loyes(1965) considered the same problem under the strong mixing assumption for stationary sequences. The sequence $\{X_n\}$ is said to satisfy the strong mixing assumption if there is a mixing function $g(k)$ tending to zero as $k \to \infty$, and such that

$$|P(A \cap B) - P(A)P(B)| < g(k)$$

when $A \in \mathcal{F}(X_1, \ldots, X_p)$ and $B \in \mathcal{F}(X_{p+k+1}, X_{p+k+2}, \ldots)$ for any p and k; $\mathcal{F}(.)$ denotes the σ-field generated by the indicated random variables. Leadbetter(1974) proposed condition D, a distributional mixing condition that is weaker than most of the classical forms of dependence restrictions, involving only sets of the form $\{X_1 \leq c, X_2 \leq c, \ldots, X_n \leq c\}$, and generalizes the mixing function to a mixing sequence.

The condition $D(u_n)$ will be said to hold for a given real sequence $\{u_n\}$ if for any integers

$$1 \leq i_1 < \cdots < i_p < j_1 < \cdots < j_{p'} \leq n$$

for which $j_1 - i_p \geq l$, we have

$$|F_{i_1,\ldots,i_p,j_1,\ldots,j_{p'}}(u_n) - F_{i_1,\ldots,i_p}(u_n)F_{j_1,\ldots,j_{p'}}(u_n)| \leq \alpha_{n,l},$$

where $\alpha_{n,l_n} \to 0$ as $n \to 0$ for some sequence $l_n = o(n)$.

Besides the condition D, the following condition is also necessary for generalizing the i.i.d. results. Various forms of such a condition were

used in the literature, for example, Watson(1954), Loynes(1965), and Leadbetter(1974) .

The condition $D'(u_n)$ will be said to hold for the stationary sequence $\{X_n\}$ and sequence $\{u_n\}$ of contants if

$$\limsup_{n\to\infty} n\sum_{j=2}^{[n/k]} P\{X_1 > u_n, X_j > u_n\} \to 0 \quad as\ k \to \infty$$

To better understand the domain of attraction for dependent sequences, we introduce an i.i.d. sequence $\{\hat{X}_n\}$ having the same common distribution function F as each member of the stationary sequence $\{X_n\}$. The sequence $\{\hat{X}_n\}$ is called the independent sequence associated with $\{X_n\}$. Set $\hat{M}_n = \max\{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n\}$. The following result and its proof can be found in [16, 22].

Suppose that $D(u_n)$, $D'(u_n)$ are satisfied for the stationary sequence $\{X_n\}$, when $u_n = x/a_n + b_n$ for each x ($\{a_n > 0\}, \{b_n\}$ being given sequences of constants). Then $P\{a_n(M_n - b_n) \leq x\} \to G(x)$ for some non-degenerate G if and only if $P\{a_n(\hat{M}_n - b_n) \leq x\} \to G(x)$. The problem of the extreme value of a stationary sequence is then reduced to a problem of the extreme of an i.i.d. sequence according to the lemma. Intuitively, from the view point of point processes, the conditions $D(u_n)$ and $D'(u_n)$ guarantee that the exceedances of the level u_n by $\{X_n\}$ follow a Poisson process when n is large. In this, the condition $D(u_n)$ provides the independence associated with the occurrence of events in a Poisson process; the condition $D'(u_n)$ limits the possibility of clustering of exceedances so that multiple events are excluded in the limit. In summary, the conditions $D(u_n)$ and $D'(u_n)$ ensure that the extremes

of the stationary sequence $\{X_n\}$ have the same qualitative behaviour as those of the associated i.i.d. sequence.

2.2. **Extreme value theory for moving average processes.** The domain of attraction of a stationary sequence can be easily solved if the conditions $D(u_n)$ and $D'(u_n)$ are satisfied. But it is in general tedious and not obvious to verify these conditions. However, if a strictly stationary sequence $\{X_n\}$ has representation as a moving average process:

$$X_n = \sum \psi_{j-n} Z_j,$$

where the noise sequence $\{Z_n\}$ are i.i.d. and ψ_j are constants, the problem will be changed dramatically. Now the tail behavior of $\{Z_n\}$ and the coefficents ψ_j of the moving average completely determine the limit behaviour and properties of the extreme value. This has been widely studied in the literature. Rootzen(1978) studied the extremes of moving average of stable processes; Davis and Resnick(1985) discussed the extreme of linear processes with regularly varying tails; Rootzen(1986) showed the extreme results when the tail distribution of the noise sequences belongs to a special exponential family. Extremes of moving averages of exponential and subexponential noise were investgated by Davis and Resnick(1988), Goldie and Resnick(1988), or [11]. Some of their results are expressed in the following two lemmas when the noise distribution belongs to the maximum domain of attraction of the Gumbel distribution $\Lambda = \exp\{-e^{-x}\}$.

Before we introduce the lemmas, we need to define a function class and conditions on the tail of the distribution F and the coefficents ψ_j. First, a distribution F with support $(0, \infty)$ is subexponential, if for all

$n \geq 2$,

$$\lim_{x\to\infty} \frac{1-F^{n*}(x)}{1-F(x)} = n, \tag{6}$$

where F^{n*} is the $n-$th convolution of F. The class of subexponential distribution functions will be denoted by S. Second, we need the following tail balance condition:

$$\lim_{x\to\infty} \frac{P(Z>x)}{P(|Z|>x)} = p, \quad \lim_{x\to\infty} \frac{P(Z\leq -x)}{P(|Z|>x)} = q, \tag{7}$$

where $0 < p \leq 1$, $p+q=1$. Finally, for the coefficients ψ_j, let $k^+ = \text{card}\{j : \psi_j = 1\}$, $k^- = \text{card}\{j : \psi_j = -1\}$, and assume that

$$\sum_{j=-\infty}^{\infty} |\psi_j|^\delta < \infty \text{ for some } \delta \in (0,1). \tag{8}$$

Without loss of generality we assume that $\max_j |\psi_j| = 1$, since otherwise we can rescale the process X_n to $X_n / \max_j |\psi_j|$. The lemma below is proved in Davis and Resnick(1988) in a more general situation.

Lemma 2. *Assume that F_Z belongs to both the maximum domain of attraction of the Gumbel distribution and the subexponential distribution class S. Then there exist constants $c_n > 0$ and $d_n \in R$ such that*

$$n(1-F_Z(c_n x + d_n)) \to -ln\Lambda(x), \quad x \in R.$$

Furthermore, assume that conditions (6)-(8) hold. Then the point process

$$\sum_{k=1}^{\infty} \epsilon_{(n^{-1}k, c_n^{-1}(X_k - d_n))} \to k^+ N_1 + k^- N_2$$

in $M_p(R_+ \times E)$ with $E = (-\infty, \infty]$. Here

$$N_i = \sum_{k=1}^{\infty} \epsilon_{(t_{ki}, j_{ki})}, \quad i = 1, 2,$$

are two independent Poisson random measure $PRM(|.|\mu_i)$ *on* $R_+ \times E$, μ_1 *has density* $f_1(x) = \exp\{-x\}$ *and* μ_2 *has density* $f_2(x) = (q/p)\exp\{-x\}$, *both with respect to Lebesgue measure.*

The two independent processes k^+N_1 and k^-N_2 are due to the contributions of those Z_n for which $\psi_n = 1$ or $\psi_n = -1$. Based on the above lemma, we obtain the following result for the point process of exceedances of $c_n x + d_n$ by the linear process $\{X_k\}$:

Lemma 3. *Under the conditions of above lemma, the point process of exceedances of* $c_n x + d_n$ *by the linear process* $\{X_k\}$ *converge weakly in* $M_p(R_+)$ *as* $n \to \infty$*:*

$$\sum_{k=1}^{\infty} \epsilon_{n^{-1}k} I_{\{c_n^{-1}(X_k - d_n) > x\}} \to \sum_{k=1}^{\infty} (k^+ \epsilon_{T_k^+} + k^- \epsilon_{T_k^-}),$$

where $\{T_k^+\}$ *and* $\{T_k^-\}$ *are the sequences of the points of two independent homogeneous Poisson processes on* R_+ *with corresponding intensities* $\exp\{-x\}$ *and* $(q/p)\exp\{-x\}$.

Therefore, the limit process of the point process of exceedances is the sum of two independent compound Poisson processes with the cluster sizes z^+, z^-.

3. Risk Estimation in Optimal Recovery

Consider the regression model in the wavelet domain

$$Y_i = \theta_i + v Z_i, \quad i = 1, \ldots, n, \tag{9}$$

where the noise $\{Z_i\}$ have mean zero and variance one and v is the noise level. In general, $\{Z_i\}$ have the same distribution on the same resolution level but they are not independent of each other, and their

distributions are determined by the distribution of the regression noise and the mother wavelet. It is usually very difficult to find the exact distributions because of the complicated form of the distributions except in two special cases which give the simpler i.i.d. distributions. The first case is when the regression noise is Gaussian which leads to all Z_i i.i.d.. The other case is when the mother wavelet is the Haar function which results in the i.i.d. distribution for those Z_i on the same resolution level, see Wu(1998a). In order to estimate the risk of our threshold estimator and at the same time to escape the situation of the complicated distributions of Z_i, we now turn to a simpler abstract model (called optimal recovery problem) in which the noise is deterministic. We will explain its relation with the regression problem (9) later. Let

$$y_I = \theta_I + \delta u_I, \tag{10}$$

where I belongs to an index set $\mathcal{I}$, $\delta > 0$ is the known noise level and u_I is a nuisance term known only to satisfy $|u_I| \leq 1, \ \forall I \in \mathcal{I}$. We want to estimate (θ_I) based on the observations (y_I). The performance will be evaluated by the worst case error

$$M_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \leq 1} \|\hat{\theta}(y) - \theta\|_{l_n^2}^2$$

under the side condition that the estimator is noise-free. Then

$$|\hat{\theta}_I| \leq |\theta_I|, \ \forall I \in \mathcal{I}. \tag{11}$$

The best performance is defined by the minimax error

$$M_\delta^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} M_\delta(\hat{\theta}, \theta),$$

where Θ is the set of all possible θ. This is the smallest risk an estimator can be uniformly over $\theta \in \Theta$.

We now investigate the soft threshold estimator

$$\hat{\theta}_I^{(\delta)}(y_I) = sgn(y_I)(|y_I| - \delta)_+, \ \forall I \in \mathcal{I}.$$

Lemma 4. *The estimator $\hat{\theta}_I^{(\delta)}$ automatically satisfies the noise-free side condition (11), and its risk has the following upper bound*

$$M_\delta(\hat{\theta}^{(\delta)}, \theta) \leq \sum_I \min(\theta_I^2, 4\delta^2).$$

Proof. If $\hat{\theta}_I^{(\delta)}(y_I) = 0$, then it is obvious that (11) holds. For those $I \in \mathcal{I}$ that $\hat{\theta}_I^{(\delta)}(y_I) \neq 0$, $|\hat{\theta}_I^{(\delta)}(y_I)| = |y_I| - \delta$. According to (10), $|y_I - \theta_I| \leq \delta$ and this immediately leads to $|\theta_I| \geq |y_I| - \delta = |\hat{\theta}_I^{(\delta)}(y_I)|$.

To prove the risk inequality, we first verify that $sgn(\hat{\theta}_I^{(\delta)}(y_I)) = sgn(\theta_I)$. If $\hat{\theta}_I^{(\delta)}(y_I) > 0$, then $\theta_I = y_I - \delta u_I \geq y_I - \delta > 0$. Similarly, If $\hat{\theta}_I^{(\delta)}(y_I) < 0$, then $\theta_I = y_I - \delta u_I \leq -\delta - \delta u_I < 0$. Because $\hat{\theta}_I^{(\delta)}$ and θ_I have the same sign, then

$$|\hat{\theta}_I^{(\delta)}(y_I) - \theta_I| \leq ||\hat{\theta}_I^{(\delta)}(y_I)| - |\theta_I|| \leq |\theta_I|.$$

From the equation $\hat{\theta}_I^{(\delta)}(y_I) = y_I - sgn(y_I)\delta$, we have the inequality $|\hat{\theta}_I^{(\delta)}(y_I) - y_I| \leq \delta$ which immediately leads to

$$|\hat{\theta}_I^{(\delta)}(y_I) - \theta_I| \leq |\hat{\theta}_I^{(\delta)}(y_I) - y_I| + |y_I - \theta_I| \leq 2\delta.$$

The above two inequalities then give

$$|\hat{\theta}_I^{(\delta)}(y_I) - \theta_I| \leq \min(|\theta_I|, 2\delta).$$

Squaring and summing across $I \in \mathcal{I}$ gives the needed inequality. □

Lemma 5. *If $\hat{\theta}$ is any estimator satisfying the side condition (11), then $M_\delta(\hat{\theta}, \theta) \geq M_\delta(\hat{\theta}^{(\delta)}, \theta)$ $\forall \theta$. If the equality holds for all θ, then $\hat{\theta} = \hat{\theta}^{(\delta)}$.*

The soft threshold estimator $\hat{\theta}^{(\delta)}$ is therefore the unique optimal estimator under the noise-free restriction.

How is the performance of the risk $M_\delta(\hat{\theta}^{(\delta)}, \theta)$ when it is compared with the best risk $M_\delta^*(\Theta)$? It turns out that the error of $\hat{\theta}^{(\delta)}$ simultaneously approaches this minimum over a large class of Θ. A set Θ is called **solid and orthosymmetric** if $\theta \in \Theta$ implies $(s_I \theta_I) \in \Theta$ for all sequences (s_I) with $|s_I| \leq 1, \forall I$.

Lemma 6. *If Θ is solid and orthosymmetric, then $\hat{\theta}^{(\delta)}$ is near minimax*

$$M_\delta(\hat{\theta}^{(\delta)}, \theta) \leq 4 M_\delta^*(\Theta), \forall \theta \in \Theta.$$

We omit the proof of the above two lemmas, interested readers can find the proof in [5].

In summary, the soft threshold estimator does a great job for the optimal recovery problem (10). It is noise-free and approaches simultaneously to the minimax risk for a large class of functions. In the next section, we will combine the results in Section 2 and Section 3 to solve the regression model (9).

4. Risk Estimation in Statistical Estimation

The problem (9) is the transformed version in the sequence space of the regression problem (1) through a wavelet basis. It is easy to see that the noise $\{Z_i\}$ in (9) is the moving average of the noise (ϵ_i) in (1). The regression problem (9) and the optimal recovery problem (10)

are connected together through the following lemma with any mother wavelet, such as coiflets, daublets, or symmlets.

Lemma 7. *Assume $\{\epsilon_i\}$ are i.i.d. and belong to the family $\mathcal{F}$ in (2). Then the noise $(Z_{i,k})$ has the following property on each resolution level $k = 1, 2, \ldots$.*

$$\pi_n = Pr\{\|(Z_{i,k})\|_{l^\infty_{n_k}} \leq \lambda_k \ \forall k\} \to 1, \ n_k \to \infty \tag{12}$$

where λ_k is determined by the expression (3).

In other words, with probability 1, the statistical problem (9) is asymptotically equivalent to the optimal recovery problem (10) with noise level $\delta_{n_k} = \lambda_k$. Therefore, we could solve the statistical problem (9) through solving the deterministic problem (10).

Proof. Let $V_n = \max\{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$. According to the distribution assumption on ϵ_i and Lemma 1, there exist constants $c_n > 0$ and $d_n \in R$ such that

$$c_n^{-1}(V_n - d_n) \to \Lambda \quad n \to \infty, \text{ or}$$

$$P\left(c_n^{-1}(V_n - d_n) \leq x\right) = P(V_n \leq u_n) = \Lambda(x),$$

where $u_n = c_n x + d_n$. First, we need to compute the constants c_n and d_n from the distribution of ϵ. From the tail distribution of ϵ_i and $n\overline{F} = \exp\{-x\}$, we have

$$Ku_n^\alpha \exp\{-Au_n^\beta\} = \frac{\exp\{-x\}}{n}.$$

After taking natural log transformation and solving the linear equation,

$$u_n = \left(\frac{\log(n) + x + \alpha \log(u_n) + \log(K)}{A}\right)^{\frac{1}{\beta}}$$

and

$$log(u_n) = (loglog(n) - log(A))/\beta + o(1).$$

Then

$$\begin{aligned} u_n &= \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} \left(1 + \frac{x + \frac{\alpha}{\beta}(loglog(n) - log(A)) + \log(K)}{\log(n)}\right)^{\frac{1}{\beta}} \\ &\approx \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} \left(1 + \frac{x + \frac{\alpha}{\beta}(loglog(n) - log(A)) + \log(K)}{\beta\log(n)}\right) \\ &= \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} \frac{x}{\beta\log(n)} + \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} (1 + o(1)). \end{aligned}$$

It is easy to see from the equation $u_n = c_n x + d_n$ that

$$c_n = \frac{1}{\beta\log(n)} \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} \text{ and}$$

$$d_n = \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} (1 + o(1)).$$

Next we will compute the probability π_n. For this purpose, let $Z^0_{i,k} = \frac{Z_{i,k}}{2^{\frac{k-J}{2}} \max_{1\le i\le S}\{a_i\}}$ and let $N(x)$ be the number of $Z^0_{i,k}$ upcrossing the level $c_n x + d_n$. Then, by using Lemma 3,

$$\pi^0_n = Pr\{\|Z^0_{i,k}\|_{l^\infty_n} \le c_n x + d_n\}$$

$$= Pr(N(x) = 0) = \exp\{-e^{-x} - \frac{q}{p}e^{-x}\} = \exp\{-\frac{1}{p}e^{-x}\}.$$

If we choose $x = loglog(n)$ and let $n \to \infty$, then $\pi^0_n \to 1$. Now if we substitute $Z^0_{i,k}$ by $Z_{i,k}$ and notice that

$$c_n loglog(n) + d_n = \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} (1 + \frac{loglog(n)}{\beta\log(n)}) + o(1),$$

then we know that $\pi^0_n = \pi_n$. That is the end of the proof of the Lemma.

□

Proof of theorem 1: Let $u_n = \left(\frac{log(n)}{A}\right)^{\frac{1}{\beta}} \left(1 - \frac{loglog(n)}{\beta \log(n)}\right)$. According to Lemma 3, with $x = -loglog(n)$,

$$\pi_n^0 = Pr\{\|Z_{i,k}^0\|_{l_n^\infty} \leq c_n x + d_n\}$$

$$= Pr(N(x) = 0) = \exp\{-\frac{1}{p}e^{-x}\}$$

$$= \exp\{-\frac{1}{p}e^{loglog(n)}\} = n^{-\frac{1}{p}} \to 0$$

since $0 < p \leq 1$. Hence

$$(13) \quad Pr\{\|Z_{i,k}\|_{l_n^\infty} > 2^{\frac{k-J}{2}} \max_{1 \leq i \leq S}\{a_i\}u_n\} = Pr\{\|Z_{i,k}^0\|_{l_n^\infty} \geq u_n\} \to 1.$$

From the above lemma, we have

$$(14) \quad Pr\{\|Z_{i,k}\|_{l_n^\infty} \leq 2^{\frac{k-J}{2}} \max_{1 \leq i \leq S}\{a_i\}(\frac{log(n)}{A})^{\frac{1}{\beta}}(1 + \frac{loglog(n)}{\beta \log(n)})\} \to 1.$$

The theorem is proved by combining equation (13) and (14). □

Now we use $\delta_{n_k} = \lambda_k$ as the threshold value to construct the threshold estimators on the kth resolution level

$$(15) \qquad \hat{\theta}_{i,k} = \hat{\theta}_i^{\delta_{n_k}}(y_i), \ 1 \leq i \leq n_k, \ k = 1, 2, \ldots.$$

Then the results of optimal recovery in Section 3 will be applied in here now.

Theorem 4. *With π_n defined by (12) and for $\forall \theta_{i,k}$*

$$Pr\{|\hat{\theta}_{i,k}| \leq |\theta_{i,k}| \ 1 \leq i \leq n_k, \ k = 1, 2, \ldots\} \geq \pi_n.$$

Proof. Let $\Omega_0 = \{\omega : \|(Z_{i,k})\|_{l_{n_k}^\infty} \leq \lambda_k, \ k = 1, 2, \ldots\}$. For all $\omega \in \Omega_0$, the statistical problem $y_{ik} = \theta_{ik} + vZ_{ik}$ is equivalent to the optimal recovery problem $y_{ik} = \theta_{ik} + \lambda_k u_i$. By Lemma 4, we have $|\hat{\theta}_{ik}| \leq |\theta_{ik}|$. Therefore, $\Omega_0 \subseteq \Omega_1$ where

$$\Omega_1 = \{\omega : |\hat{\theta_{ik}}| \leq |\theta_{ik}| \ 1 \leq i \leq n_k\}.$$

This means $Pr(\Omega_1) \geq Pr(\Omega_0) \geq \pi_n$ and thus the proof is finished. □

The mean squared error $M_n(\hat{\theta}, \theta) = E\|\hat{\theta} - \theta\|_{l_n^2}^2$ of (15) will be evaluated and compared with the ideal risk $R_\epsilon(DP, \theta)$ in (4). The following theorem gives an upper bound of the risk for our estimator. We realize that this bound is not sharp, however it does tell us that our estimator's risk is within a log term of the ideal risk simultaneously for a large class of functions, such as the Besov spaces.

Theorem 5. *Let Θ be solid and orthosymmetric. Then the estimator $\hat{\theta}^{(n)}$ of (15) asymptotically approaches the ideal risk; i.e.*

$$M_n(\hat{\theta}^{(n)}, \theta) \leq \sum_k 4\lambda_k^2 \sum_j \min(\theta_{jk}^2, v^2) \leq 2 \max_{1 \leq i \leq S}\{a_i\}(\frac{log(n)}{A})^{\frac{2}{\beta}} R_\epsilon(DP, \theta)$$

for $\forall \theta \in \Theta$ and $\epsilon = v$.

Proof. We use the same argument as that used in the proof of the above theorem; the statistical problem is asymptotically equivalent to the optimal recovery problem with $\delta = \lambda_k$. Therefore, Lemma 4 applies here and

$$\begin{aligned}
M_n(\hat{\theta}^{(n)}, \theta) &= E\|\hat{\theta}^{(n)} - \theta\|_{l_n^2}^2 \\
&\leq \sum_k \sum_j \min(\theta_{jk}^2, 4v^2\lambda_k^2) \\
&\leq \sum_k 4\lambda_k^2 \sum_j \min(\theta_{jk}^2, v^2) \\
&\leq 2 \max_{1 \leq i \leq S}\{a_i\}(\frac{log(n)}{A})^{\frac{2}{\beta}} \sum_k 2^{k+1-J} \sum_j \min(\theta_{jk}^2, v^2) \\
&\leq 2 \max_{1 \leq i \leq S}\{a_i\}(\frac{log(n)}{A})^{\frac{2}{\beta}} \sum_k \sum_j \min(\theta_{jk}^2, v^2) \\
&= 2 \max_{1 \leq i \leq S}\{a_i\}(\frac{log(n)}{A})^{\frac{2}{\beta}} R_v(DP, \theta). \square
\end{aligned}$$

5. Discussion

5.1. **Data driven threshold value selection.** Donoho & Johnstone(1995) proposed a data driven threshold selecting method, called the SureShrink method, which is based on Stein's unbiased estimate of the loss function. They search for the optimal threshold to achieve the best mean squared error in the range between 0 and $\sqrt{2logn}$. Nason(1996) used the cross validation method with the quantity $\sqrt{2logn}$ with possibly a level dependent adjustion. We see that for the Gaussian noise problem, the universal threshold, $\sqrt{2logn}$ plays a fundamental role in wavelet thresholding estimation. Our new threshold value plays a similar role for the non-Gaussian noise problem. We can use it to determine the range for which the data driven methods can start to search for the optimal threshold value that obtains the best balance between bias and variance.

5.2. **Correlated regression noise.** Wang(1995) used a fractional Gaussian noise model to approximate long dependent data. Through the wavelet transformation, he converted the fractional Gaussian noise model to a sequence series in the wavelet domain and found the level-dependent threshold values. A thresholding estimator then can be constructed by using the three steps precedure which we gave in the introduction. Johnstone and Silverman(1994) also discussed the correlated noise problem and showed that although the correlation coefficients are involved, the methods for an i.i.d. Gaussian noise can be similarly extended to the correlated Gaussian noise. We can also apply our method to solve the correlated noise problems. Since our method

requires very limit information about the noise distribution, we can establish the noise-free threshold values for the correlated noise and then construct a wavelet estimator. It is very interesting to compare these methods. We will report the results of the comparison later.

5.3. **Minimax risk.** The relationship among the risk of the wavelet estimate, the ideal risk, and the minimax risk is well known for the Gaussian noise problem; see Donoho(1995), Donoho and Johnstone(1994), and Donoho, Liu, and MacGibbon(1990). We cannot give similar bounds in this study. The relationship between our new risk and the minimax risk is still unclear since it is very difficult to find the minimax risk for an arbitrary noise distribution. The possible solution is first to consider a noise which has a spherically symmetric distribution, and try to find a lower bound to the minimax risk based on the ideal risk.

5.4. **Density estimation.** We are now applying this method to density estimation. Since our method requires very limited information about the underlying density function, our result is very promising; see Wu(1998b). Also our method can explain the comparison as discussed in [9](pp 327). The key point is that the best threshold value depends on the underlying density itself. Different density functions require different threshold values.

6. Appendix

Proof of theorem 2:

We will prove the theorem for an arbitrary Besov space $B^{\nu}_{p,q}[0,1]$; all the Triebel spaces $F^{\nu}_{p,q}[0,1]$ follow similar arguments. The threshold

wavelet estimator has the following expression

$$\hat{f}(t) = \sum_{k=0}^{2^{j_0}-1} \hat{\beta}_{j_0,k}\varphi_{j_0,k}(t) + \sum_{j\geq j_0}^{J-1}\sum_{k=0}^{2^j-1} \hat{\alpha}_{j,k}\psi_{j,k}(t),$$

while the corresponding true function is

$$f(t) = \sum_{k=0}^{2^{j_0}-1} \beta_{j_0,k}\varphi_{j_0,k}(t) + \sum_{j\geq j_0}^{\infty}\sum_{k=0}^{2^j-1} \alpha_{j,k}\psi_{j,k}(t).$$

The norms of the estimator and the true function have the following expression in the Besov space

$$\|\hat{f}\|_F = \|\hat{\beta}_{j_0,k}\|_{l_p} + \left(\sum_{j\geq j_0}^{J-1}(2^{js}(\sum_k |\hat{\alpha}_{j,k}|^p)^{\frac{1}{p}})^q\right)^{\frac{1}{q}}$$

$$\|f\|_F = \|\beta_{j_0,k}\|_{l_p} + \left(\sum_{j\geq j_0}^{\infty}(2^{js}(\sum_k |\alpha_{j,k}|^p)^{\frac{1}{p}})^q\right)^{\frac{1}{q}}$$

According to Theorem 4, $|\hat{\beta}_{j_0,k}| \leq |\beta_{j_0,k}|$ and $|\hat{\alpha}_{j,k}| \leq |\alpha_{j,k}|$ for $\forall k, j$ with a probabilty that tends to 1 as $n \to \infty$. Therefore, $\|\hat{f}\|_F \leq \|f\|_F$. □

Proof of theorem 3:

By the quasi-orthogonality property [Donoho(1995)]:

$$n^{-1}E\|\hat{f} - f\|_{l_n^2}^2 \leq \gamma_1^2 E\|\hat{\theta}^n - \theta\|_{l_n^2}^2,$$

where γ_1 is a constant which does not depend on n and f. Using Theorem 5 we have the upper bound. □

References

[1] Cohen, A., Daubechies, I., Jawerth, B., and Vial, P. (1993). Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris A*,**316**, 417-421.

[2] Daubechies, I. (1991) Ten Lectures on Wavelets. SIAM: Philadelphia.

[3] Davis, R. and Resnick, S. (1985) Limit theory for moving averages of random variables with regularly varying tail probabilities. *The Annals of Probability* **13** 179-195.

[4] Davis, R. and Resnick, S. (1988) Extremes of moving averages of random variables from the domain of attraction of the double exponential distribution. *Stochastic Processes and their Applications* Ser **30**, 41-68.

[5] Donoho, D.L. (1995) De-Noising by Soft Thresholding. *IEEE Trans. Info. Thry.*, **41**, pp. 613-627.

[6] Donoho,D.L. (1993) Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, **1**, 100-115.

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

[8] Donoho, D. L. and Johnstone, I. M (1995) Adapting to unknown smoothness via Wavelet shrinkage. *J. Amer. Statist. Assoc.*,**432** 1200-1224.

[9] Donoho, D.L. Johnstone, I.M. Kerkyacharian, G. and Picard, D. (1995) Wavelet Shrinkage: Asymptopia?. *Journ. Roy. Stat. Soc.* Ser **B**, **57**, 301-369.

[10] Donoho,D.L. , Liu,R.C. and MacGibbon,K.B. (1990) Minimax risk over Hyperrectangles and implications. *Ann. Statist* **18** , 1416-1437.

[11] Embrechts, P., Kluppelberg, C., and Mikosch, T.(1997) Modelling extreme events. Springer

[12] Gnedenko, B.V. (1943) Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.* **44**, 423-453.

[13] Gao, H. (1993) Choice of thresholds for wavelet estimation of the log spectrum. Technical Report, Department of Statistics, Stanford University.

[14] Goldie, C. and Resnick, S. (1988) Distributions that are both subexponential and in the domain of attraction of an extreme-value distribution. *Adv. Appl. Prob.* Ser **20**, 706-718.

[15] Johnstone, I.M. and Silverman, B.W. (1994) Wavelet threshold estimators for data with correlatd noise. Technical Report, Department of Statistics, Stanford University.

[16] Leadbetter, M. R. (1974) On extreme values in stationary sequences. it Z. Wahrsh.verw.Gebiete. **28**, 289-303.

[17] Leadbetter, M. R., Lindgren, G., Rootzen, Holger (1983) Extremes and Related Properties of Random Sequences and Processes. New York: Springer-Verlag.

[18] Loynes, R.M. (1965) Extreme values in uniformly mixing stationary stochastic process. *Ann. Math. Statist.* **36** 993-999.

[19] Mallat, S. (1989b) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

[20] Nason, G.P. (1996) Wavelet Shrinkage using Crossing-validation. *J. R. Statist. Soc.* Ser **B**, **58**, 463-479.

[21] Neumann, M.H. and Sachs R.V. (1995) Wavelet Thresholding: Beyond the Gaussian I.I.D. Situation. *Lect. Notes Statist.*, **103**, 301-329.

[22] Resnick, S. (1987) Extreme values, regular variation, and point processes. New York: Springer-Verlag.

[23] Rootzen, H. (1978) Extremes of moving averages of stable processes. *The Annals of Probability* **6** 847-869.

[24] Rootzen, H. (1986) Extreme value theory for moving average processes. *The Annals of Probability* **14** 612-652.

[25] Wang, Y. (1996) Function estimation via wavelet shrinkage for long memory data. *Ann. Statist.* **24** 466-484.

[26] Watson, G.S. (1954) Extreme Values in samples from m-dependent stationary stochastic processes. *Ann. Math. Statist.* **25** 798-800.

[27] Wu, Y. (1998a) Wavelet Estimation for Nonparametric Regression: Beyond Gaussian Noise I. Technical report, Purdue University.

[28] Wu, Y. (1998b) Wavelet Estimation for density functions. In progress.