5-1-2012

Function estimation of irregularly spaced data with long memory dependence

Rosalie Michelle Wheeler
University of Missouri-St. Louis

Function estimation of irregularly spaced data with long memory dependence

Rosalie M. Wheeler

M.A., Mathematics, Univeristy of Missouri - St. Louis, 2007

B.S., Mathematics, University of Missouri - St. Louis, 2004

A Dissertation Submitted to The Graduate School at the University of Missouri - St. Louis in partial fulfillment of the requirements for the degree Doctor of Philosophy in Mathematics with an emphasis in Statistics.

May 2012

Advisory Committee

Haiyan Cai, Ph.D.

Chairperson

Ronald Dotzel, Ph. D.

Wenjie He, Ph. D.

Qingtang Jiang, Ph. D.

FUNCTION ESTIMATION OF IRREGULARLY SPACED DATA WITH LONG MEMORY DEPENDENCE

ROSE WHEELER

Abstract. We examine the problem of estimating an underlying function from collected data. The methods considered include parametric regression, density estimation, kernel estimation, wavelet regression, and specific results from when our underlying function $f(x)$ is a member of the Besov or the Triebel spaces. Then we consider the problem of long memory error in several settings, including data which is equally spaced, data which is unequally spaced, and data which is a member of the Holder class and several other spaces. Ultimately we focus on three different problems. The first involves using linear interpolation or local averaging to account for the problem of irregularly spaced data. The second involves using a function H to reorder the data in a more general space. The third involves solving the problem in the matrix setting and considers the use of penalty functions. This method leads to general equations which describe the Mean Square Error in terms of Oracle risk. All three of these problems attempt to bound the Mean Integrated Square Error when the data is subject to long memory error.

Key words and phrases. Parametric regression, density estimation, kernel estimation, wavelet regression, Besov spaces, Triebel spaces, irregularly spaced data, long memory error, matrix estimation, incomplete systems of equations.

Contents

Part 1. Introduction and Brief History.

In this dissertation we give an overview of estimation of functions. The methods considered include parametric regression, density estimation, kernel estimation, wavelet regression, and specific results from when our underlying function $f(x)$ is a member of the Besov or the Triebel spaces. We also consider advanced recent papers dealing with the problem of recovering an underlying function under the conditions of irregularly spaced data, long memory error, and in the matrix setting. The solutions to all of these problems begin with the same assumption below.

Suppose we are given data of the form

$$y_i = f(x_i) + \epsilon_i$$

with $\epsilon \sim N(0, \sigma^2)$. We wish to estimate the value of the function $f(x_i)$. This problem appears everywhere in statistics, and in almost any technical field.

The first part, Part 2, deals with many basic methods of function estimation as well as some advanced techniques from the work of Donoho and Johnstone in [8].

There are many ways to approach a problem like this. One way is to assume that $f(x)$ has some form and solve for the parameters of this form. For example, one of the first methods statistics students learn is linear regression. Students are told to look at a set of data, determine whether or not the points "look" linear, and then proceed to find a line of best fit according to the mean square error. Data which is periodic could perhaps be modeled with a sine curve. Similarly, one could do the same thing for probability density functions. These methods are examined in Section 2.

The problem with this is that the assumption that $f(x)$ has some pre-defined form is artificial. One cannot just look at data and decide what form the underlying function has. Error distorts data and even relationships which are truly linear do not produce data points which lie in a line. Often not much is known about the underlying relationship between two quantities being modeled, only rarely when data is already governed by some physical law is a predetermined form of $f(x)$ known.

To solve this problem, one could take a non-parametric approach. Section 3 of this paper gives an overview of methods for estimating $f(x)$ which do not make any assumptions about its true form. A good overview of this material is given in [23]. The simplest estimator for $f(x)$ is a histogram. This method computes $f(x)$ by averaging the data points in equally spaced intervals. Unfortunately, this estimator is discontinuous and the choice of intervals is arbitrary. One could fix this problem by defining the naive estimator. This estimator puts every data point at the center of an interval and averages the surrounding points. However, this function is still discontinuous.

One could make an estimator which was continuous by using a kernel estimator. Kernels can be thought of as smooth "bumps". There are many different ways to create an estimator out of kernels. The simplest way is to create a weighted bump at each data point. This allows $f(x)$ to inherit whatever smoothness the kernel possesses. The kernel estimator can be made more sophisticated by making bumps sharper in areas where data is denser and smoother where data is more sparse. That is the idea behind the nearest neighbor method. This makes the tails flatter and the estimate "look" better.

Kernels are very useful in creating an estimator for $f(x)$ which is smooth and has nice properties. However, our choice of kernel is still an artificial one. Suppose we have an orthonormal basis for the support of $f(x)$. Then we can express $f(x)$ as a linear combination of this basis. One could estimate the coefficients of this orthogonal series. We then either truncate the series or slowly decrease the later coefficients to zero. This defines our estimator.

There are other miscellaneous ways of estimating $f(x)$. Another way to estimate $f(x)$ would be to try and optimize the maximum likelihood of an estimator with a penalty for the function's "roughness". Or, one could define an estimator as a sum of generalized weight functions.

All of these estimators still cannot account for simple common problems with raw data. Suppose for instance that data is only available for positive x_i. For any of the methods of estimating $f(x)$ there would be errors near $x = 0$. Several methods of extending the data to account for this problem are discussed.

The next section of this first part deals with the specifics of choosing a kernel and choosing the window width for a function estimator. These are many methods for choosing this window width. One could minimize mean square error or make underlying assumptions about the structure of $f(x)$. Or, one could use automatic methods to choose the window width. These include minimizing the square error or maximizing the likelihood of an estimator. Many other technical ways of improving simple kernel estimators are discussed in Subsection 3.2. Finally, the ultimate effectiveness of kernel estimators is analyzed in several theorems.

All of these estimators except for the orthogonal series estimator require some assumptions about the structure of $f(x)$ or some choice of a kernel. There is an alternative to these choices. The multiresolution analysis MRA structure of wavelets combines the automatic method of orthogonal series with the adaptability of piecewise estimators. In a sense, any arbitrarily small interval is broken into its own orthogonal estimator. Also, because the wavelet coefficients are computed by multiplying data by an orthogonal matrix, normal noise in the data becomes normal noise in the coefficients' estimators. Section 4 gives a brief review of the structure of wavelets. Reviews of wavelets can be found in many places, including [6, 25]. We begin by examining the continuous wavelet transform and then discuss the discrete wavelet transform which would be applied to data.

In Section 5, we deal with the problem of recovering a true signal from noisy data. Suppose now that we estimate $f(x)$ a signal. We express this $f(x)$ as a vector of values. From this vector of values we derive the wavelet coefficients. The idea is that some of these wavelet coefficients express the underlying shape of the true signal and some of the wavelet coefficients are just due to the noise in the signal. Using a process called thresholding, we choose according to a rule which coefficients to keep and which coefficients to discard. Section 5 discusses how by using the soft or hard thresholding operators we can approach a "best case scenario" risk, called the oracle risk. We define the oracle risk to be the best choice of wavelet coefficients in the sense of minimizing the risk. In fact, the thresholding risk is within a factor of $2\log n$ of the oracle risk where n is the sample size. This method is explored in [8].

Lastly, in Section 6, we deal with this same problem only under the assumption that $f(x)$ is a member of either the Besov class of functions or the Triebel class of functions. This problem is explored in [9]. The coefficients are dealt with in a way similar to that in the previous section. Here, the thresholding risk is bounded by a constant factor times the ideal risk. Ultimately, the main theorem of this section shows that wavelet methods either perform as well as linear methods or surpass them depending on the initial conditions on $f(x)$.

Next, we examine function estimation in a more recent and complex context in Parts 3, 4, 5 and 6.

Suppose now that we estimate $f(x)$ a signal. We express this $f(x)$ as a vector of values. From this vector of values we derive the wavelet coefficients. The idea is that some of these wavelet coefficients express the underlying shape of the true signal and some of the wavelet coefficients are just due to the noise in the signal. Using a process called thresholding, we choose according to a rule which coefficients to keep and which coefficients to discard. In the work of Donoho and Johnstone by using the soft or hard thresholding operators we can approach a "best case scenario" risk, called the oracle risk. We define the oracle risk to be the best choice of wavelet coefficients in the sense of minimizing the risk. In fact, the thresholding risk is within a factor of $2\log n$ of the oracle risk where n is the sample size. This method is explored in [8].

Another thing that could be changed is the method of thresholding. One could use block thresholding. Instead of thresholding each wavelet coefficient, the numbers are divided into groups. These groups are then either discarded or kept according to some rule. This method of thresholding approaches the true function more quickly. The authors Hall and Picard in their paper [13] first examine this problem.

Within the structure of the wavelet problem, the integers j are divided among consecutive, nonoverlapping blocks of length l_i, say $\mathcal{B}_{ik} = \{j \ : \ (k-1)l_i + \nu + 1 \leq \ j \ \leq kl_i + \nu\}$. Let $\hat{\mathcal{B}}_{ik}$ be an estimator of the average value of β_{ij}^2 for $j \in \mathcal{B}_{ij}$. Then groups of coefficients are thresholded according to $I\left(\hat{\mathcal{B}}_{ij} > cn^{-1}\right)$. Errors are considered to be independent.

This problem was expanded several years later to the case of long memory error in the data. Before the error was independent, that is $E(\varepsilon_i \varepsilon_j) = 0$. With long memory error $E(\varepsilon_i \varepsilon_{i+j}) \sim C|j|^{-\alpha}$ for some $\alpha \in (0, 1]$. In the paper [18], the authors Li and Xiao consider an estimator based on block thresholding. They address the problem of long memory error with equally spaced data. We explore this method of function estimation in Part 3. The estimator the authors use is listed below.

$$\hat{g}(x) = \sum \hat{\alpha}_{i_0 j} \phi_{i_0 j}(x) + \sum \sum \left(\sum_{(ik)} \hat{\beta}_{ij} \psi_{ij}(x) \right) I\left(\hat{\mathcal{B}}_{ik} > \delta \right)$$

Here the $\hat{\mathcal{B}}_{ik}$ is as in [13] and the summation term (ik) means the sum over the defined blocks. The structure of the Mean Integrated Square Error (MISE) between the actual function and the estimator is divided up in the same way as in the work of Hall and Picard in [13]. This division is common in all of the advanced works. By using Parseval's identity, one can split the error due to the mother and father wavelets. This can be further split according to which wavelet coefficients are kept and which are "killed" by our thresholding rule. Coefficients which are not large enough are discarded.

Another entirely separate problem is that of data which is not equally spaced. When wavelet coefficients are computed from data it is required that the data be equally spaced. There are several methods of dealing with this problem.

One method is presented in [11]. We examine this problem in Part 4. Here data which is irregularly spaced is interpolated by a linear function. Let $\mathcal{Y} = \{(X_m, Y_m), 1 \leq m \leq n\}$ be the data, considered as points. These are interpolated by

$$Y(x) = \sum_m w_m(x) Y_m \quad \text{for } x \in (X_{-\nu_1}, X_{n-\nu_2}]. \tag{0.1}$$

for some weights $w(x)$. This function is then used to compute the wavelet coefficients $\hat{b}_j = \int_{\mathcal{J}} Y \phi_j$ and $\hat{b}_{ij} = \int_{\mathcal{J}} Y \psi_{ij}$. The authors, Hall and Turlach, then examine the Mean Integrated Square Error (MISE).

Still another solution to the problem of irregularly spaced data is presented in [1]. The authors Antoniadis and Fan avoid the method of interpolation altogether and use matrices. They suppose that the data set is incomplete and use matrices to solve the overdetermined system for the missing wavelet coefficients. We examine this problem in Part 5. They express the observed data as

$$\mathbf{Y}_n = \mathbf{A}\theta + \epsilon, \qquad \epsilon \sim N\left(0, \sigma^2 I_n\right) \tag{0.2}$$

where ϵ is the noise vector. They wish to minimize

$$2^{-1}\left\|\mathbf{Y}_n - \mathbf{A}\theta\right\|^2 + \lambda \sum_{i=1}^{N} p\left(|\theta_i|\right) \tag{0.3}$$

for a given penalty function p and a regularization parameter $\lambda > 0$. This penalty function is a marriage of the soft and hard thresholding rules of Donoho and Johnstone's work.

The solution (Rao 1973) is what is called the normalized method of frame whose solution is given by

$$\theta = \mathbf{D}\mathbf{A}^T\left(\mathbf{A}\mathbf{D}\mathbf{A}^T\right)^{-1}\mathbf{f}_n,$$

where $\mathbf{D} = \text{Diag}\left(2^{-2sj_i}\right)$ with j_i denoting the resolution level with which θ_i is associated.

Another method for dealing with this problem of irregularly spaced data is presented in [4]. Here the authors Cai and Brown assume that the data points $x_i = H^{-1}(i/n)$ for some cumulative density function H on $[0,1]$. We examine the solution to this problem in Part 6.

A rough outline of the procedure this paper describes is recorded below.

(1) Precondition the data by a sparse matrix.
(2) Transform the preconditioned data by the discrete wavelet transform.
(3) Denoise the noisy wavelet coefficients via thresholding.
(4) Apply the inverse transform to the denoised coefficients.
(5) Postcondition the data by a matrix to get the estimate at the sample points.

The wavelet coefficients are computed in the following way.

$$\tilde{\alpha}_{jk} = n^{-1/2}\sum_{i=1}^{n} y_i \left\langle \phi_{Ji} \circ H, \phi_{jk} \right\rangle, \qquad \tilde{\beta}_{jk} = n^{-1/2}\sum_{i=1}^{n} y_i \left\langle \phi_{Ji} \circ H, \psi_{jk} \right\rangle.$$

This solution is restricted to functions which are members of the Holder class.

In Part 7, we attempt to solve the problems of long memory dependence and irregularly spaced data points simultaneously. We consider a linear interpolation of the data and then analyze the MISE. Lastly, we find the wavelet coefficients and then threshold them.

In Part 8, we attempt to solve the problems of long memory dependence and irregularly spaced data points by using the methods utilized in [4]. We use a function H as mentioned in Part 6 and consider the MISE while considering the long memory error.

In Part 9, we compare and contrast the results from Parts 7 and 8, namely the different spaces that the results refer to as well as their convergence.

Next, we try to consider long memory error in a matrix context. In Part 10 we apply the work of Antoniadis and Fan to the long memory setting. We phrase the problem of incomplete data and long memory error in terms of wavelets. Also, we extend on the work of Donoho and Johnstone and introduce the necessary notation needed to consider oracle risk with long memory error.

We consider these problems because most real life data sets are not independent. Many data sets are not equally spaced. Long memory situations include: hydrology, econometrics, traffic modeling, spatial data (flooding, spread of disease, etc) and may other examples. Any of these could provide samples which were unequally spaced. We extend several kinds of research to accommodate long memory error.

Part 2. A Survey on Methods of Function Estimation.

1. Introduction.

In this part we consider many different methods of function estimation, including Parametric regression, density estimation, kernel estimation, wavelet regression, and estimating a function $f(x)$ when the function is a member of the Besov space or Triebel space. These early methods of estimation are surprisingly effective, and several theorems are studied which analyze their overall effectiveness.

2. Parametric Regression.

Let us consider the problem of determining an expression for an underlying function $f(x)$ parametrically. We choose a form for $f(x)$ and then determine using a variety of methods what the best parameters of $f(x)$ are. While the next part deals only with probability distributions, these methods can be used to determine a from for any $f(x)$. If a rescaling were applied to the data any methods which assume $f(x)$ is a probability density function could still be applied.

2.1. **Method of Maximum Likelihood.** Suppose $f = f(x, \theta)$ where f is a probability density function and θ is a parameter it is dependent on. Also, let

$$Y_i = f(x_i).$$

Definition 1. Let $Y_1, Y_2, ..., Y_n$ be a random sample from a function $f(x)$. The likelihood function is the product of the probability function $f(x, \theta)$ evaluated at n data points. That is

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta).$$

Furthermore, we have the following definition.

Definition 2. Let $Y_1, Y_2, ..., Y_n$ be a random sample from $f(x, \theta)$ and let $L(\theta)$ be the corresponding likelihood function. Suppose $L\left(\hat{\theta}\right) \geq L(\theta)$ for all possible values of θ. Then $\hat{\theta}$is called the maximum likelihood estimator or MLE for θ.

Note that maximizing the likelihood function is the same as maximizing the log-likelihood $\ln L(\theta)$. We now consider the following examples.

Example 3. Suppose $k_1, k_2, ..., k_n$ is a set of n observations representing the geometric probability model, $f(k_i, p) = (1-p)^{k_i - 1}p$ where $k_i = 1, 2, ...$ We wish to find the MLE for p.

$$L(p) = \prod_{i=1}^{n}(1-p)^{k_i-1}p = (1-p)^{\sum_{i=1}^{n} k_i - n}p^n.$$

Let $k = \sum_{i=1}^{n} k_i$. Take the ln of $L(p)$.

$$\ln L(p) = (k-n)\ln(1-p) + n\ln p$$

Now differentiate with respect to p to find the maximum.

$$\frac{n-k}{1-p} + \frac{n}{p} = 0$$

Solving this yields MLE $\hat{p} = \frac{n}{k}$.

This method can be applied to may different forms of probability model.

Example 4. Suppose $y_1, y_2, ..., y_n$ is a set of measurements representing an exponential probability density function with an unknown parameter θ. That is, $f(y_i, \theta) = e^{-(y_i - \theta)}$ for $y \geq \theta$ and $\theta > 0$. We find the MLE for θ.

$$L(\theta) = \prod_{i=1}^{n} e^{-(y_i - \theta)} = e^{-\sum_{i=1}^{n} y_i + n\theta}$$

We cannot use the log-likelihood because the derivative of this log-likelihood is n and never 0. We note that $L(\theta)$ is maximized when the exponent is as large as possible. This means that θ must be as large as possible. Because $y \geq \theta$, θ can only be as large as the smallest y_i. Thus, $\hat{\theta} = y_{\min}$.

An examination of MLEs might not be complete without examining the normal distribution.

Example 5. Suppose a random sample of size n is drawn from the two parameter normal probability distribution.

$$f\left(x, \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad -\infty < y < \infty, \quad -\infty < \mu < \infty, \quad \sigma^2 > 0$$

We find $L(\theta)$.

$$L\left(\mu,\sigma^2\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2}=\left(2\pi\sigma^2\right)^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2}$$

Then

$$\ln L\left(\mu,\sigma^2\right)=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2.$$

Also

$$\frac{\partial \ln L\left(\mu,\sigma^2\right)}{\partial\mu}=-\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)\left(-\frac{1}{\sigma}\right)$$

and

$$\frac{\partial \ln L\left(\mu,\sigma^2\right)}{\partial\sigma^2}=-\frac{n}{2}\cdot\frac{1}{2\pi\sigma^2}\cdot 2\pi-\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\left(\frac{-1}{\sigma^4}\right).$$

Setting each of these equations to 0 yields $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i$ and $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$.

Next we consider the method of moments.

2.2. **Method of Moments.** Let us now define $f=f\left(x,\theta_1,\theta_2,...,\theta_k\right)$ where the $\theta_1,\theta_2,...,\theta_k$ are parameters of f. Define the first k moments of f as below.

$$E\left(Y^j\right)=\int_{-\infty}^{\infty}y^j f\left(y,\theta_1,\theta_2,...,\theta_k\right)dy,\qquad j=1,2,...,k$$

We then find the k parameters by computing the first k moments and solving the resulting system of equations.

Example 6. Suppose that

$$f(y,\theta)=\theta y^{\theta-1},\quad 0\le y\le 1.$$

Then

$$E(Y)=\int_0^1 y\cdot\theta y^{\theta-1}dy=\theta\cdot\left.\frac{y^{\theta+1}}{\theta+1}\right|_0^1=\frac{\theta}{\theta+1}.$$

Setting $E(Y)=\bar{y}$ we obtain $\hat{\theta}=\frac{\bar{y}}{1-\bar{y}}$.

This method can be applied to many different distributions.

2.3. **Method of Least squares.** We now move away from regression on probability distributions and consider a different kind of problem. Suppose we have data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ and we wish to find a polynomial of degree m which is closest in the sense of least squares to the data. We write the polynomial $p(x)$below.

$$p(x) = \sum_{k=0}^{m} \beta_k x^k$$

The quantity we wish to minimize is

$$L = \sum_{i=1}^{n} \left[y_i - p\left(x_i\right) \right]^2 .$$

As an example, lets consider the case where $m = 2$.

Example 7. We wish to find the quadratic function $y = \beta_0 + \beta_1 x + \beta_2 x^2$ which minimizes L.

$$L = \sum_{i=1}^{n} \left[y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right]^2 .$$

We take the derivatives to optimize L.

$$\frac{\partial L}{\partial \beta_2} = \sum_{i=1}^{n} \left(-2\right) x_i^2 \left[y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right]$$

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^{n} (-2) x_i \left[y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right]$$

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} (-2) \left[y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right]$$

Setting these derivatives to 0 yields the following system.

$$\left(\beta_0\right)(n) + \left(\beta_1\right)\left(\sum_{i=1}^{n} x_i\right) + \left(\beta_2\right)\left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} y_i$$

$$\left(\beta_0\right)\left(\sum_{i=1}^{n} x_i\right) + \left(\beta_1\right)\left(\sum_{i=1}^{n} x_i^2\right) + \left(\beta_2\right)\left(\sum_{i=1}^{n} x_i^3\right) = \sum_{i=1}^{n} x_i y_i$$

$$\left(\beta_0\right)\left(\sum_{i=1}^{n} x_i^2\right) + \left(\beta_1\right)\left(\sum_{i=1}^{n} x_i^3\right) + \left(\beta_2\right)\left(\sum_{i=1}^{n} x_i^4\right) = \sum_{i=1}^{n} x_i^2 y_i$$

We can solve these equations for $\beta_0, \beta_1, \beta_2$.

This polynomial regression is somewhat limited but can be applied to any data which can be transformed into a polynomial form. Suppose we assume that our data is of the form $y = \beta_0 e^{\beta_1 x}$. We can transform this by taking the ln of both sides of this equation.

$$\ln y = \ln \beta_0 + \beta_1 x$$

We could solve for these parameters using the method of least squares and then put them back into the original equation.

This method can be more generally applied to many different forms. However, for every example considered within this part we have assumed that the true form of $f(x)$ is known and have proceeded to estimate the parameters of $f(x)$ based on that assumption. While this is useful in many situations, it is not practical in many others. The following part explores many different methods of estimating $f(x)$ nonparametrically based on the assumption that the function is a probability distribution.

Later we will return to the problem presented in this section in Parts 5 and 6.

3. Density Estimation.

In this part we examine many different methods of estimating a function $f(x)$. These methods are applicable to any set of data and do not make assumptions about the structure of $f(x)$ except in places where the text explicitly says so. Occasionally $f(x)$ will be assumed to be the normal distribution and further conclusions about particular problems will be drawn. The first section deals with several different methods of estimation.

The second section addresses the problem of optimizing our choice of estimation under different criteria. We consider choices of window width several different ways. We could chose window widths which minimize the Mean Square Error, maximize the Maximum likelihood, or by using least squares cross validation. The window width is only a part of the problem of estimating $f(x)$. One could choose kernels with many different properties according to whether the estimator needs to be smooth, differentiable, or any other number of conditions. Other estimators can be derived from the Fourier transform of $f(x)$.

These methods do not use wavelets, but are still practical and also easy to apply.

3.1. Probability Density Estimators.

The information in this section comes directly from [23]. We wish to consider the problem of estimating a probability density function from data. Recall that

$$\int_a^b f(x)dx = P(a < x < b)$$

for all $a < b$. We estimate $P(a < x < b)$ from the data to find $f(x)$.

Our approach is the parametric one. We assume $f(x)$ has a certain form, for example the form of a normal distribution, then estimate μ and σ^2. However, this approach makes assumptions on the data which may not be true.

Density estimates provide indication of features of the data, such as skewness and multimodality. (A preexisting assumption of data with a normal distribution would suppress these properties.) Also, density estimates are comparatively easy to understand.

We assume we have a sample of n real observations $x_1, ..., x_n$ whose underlying density we wish to estimate. $\hat{f}$ will denote this estimator.

3.1.1. *Histograms.* The oldest density estimator is the histogram. Given an origin x_0 and a bin width h, define the bins to be the intervals $[x_0 + mh, x_0 + (m+1)h)$ for $m \in \mathbb{Z}$. Then

$$\hat{f}(x) = \frac{1}{nh}(\text{number of } x_i \text{ in the same bin as } x).$$

We must choose the origin and the bin width. The bin width is the defining choice. Large widths conceal features of the graph, while small widths make the picture look too much like the data.

Histograms give us a general overview of things, but their discontinuities make them not very valuable if we need to use derivatives. Also, the choice of bin origin can change the picture substantially. Histograms are often hard to read in trivariate or multivariate data, and the problems with choosing h and x_0 are multiplied over a grid.

3.1.2. *The Naive Estimator.* One way of solving this problem is by using the naive estimator. Note that

$$f(x) = \lim_{h \to 0} P(x - h < x < x + h)/(2h).$$

We can estimate $P(x - h < x < x + h)$ by the proportion of the sample falling into $(x - h, x + h)$. Choose

$$\hat{f}(x) = \frac{1}{2hn} (\text{number of } x_i \text{ in } (x - h, x + h)).$$

We can also write this estimator in a form similar to that of kernels. Define

$$w(x) = \begin{cases} \frac{1}{2} \text{ if } |x|<1 & 0 \text{ otherwise.} \end{cases}$$

We can write

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - x_i}{h}\right).$$

We can see that the naive estimate can be seen to be an attempt to construct a histogram where every point is the center of a sampling interval, thus freeing the histogram from a choice of bin positions.

Note however that $\hat{f}(x)$ is not continuous but has jumps at the points $x_i \pm h$ and has zero derivative everywhere else.

3.1.3. *Kernels.* We now wish to generalize the naive estimator to fix the problems with discontinuity. Replace w with a kernel function k which satisfies

$$\int_{-\infty}^{\infty} k(x)dx = 1. \tag{3.1}$$

Usually, k is a symmetric probability density function. Now

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{n}\right).$$

The naive estimator can be considered as a sum of boxes, the kernel estimator as a sum of bumps. If the kernel is everywhere non-negative and satisfies 3.1, then $\hat{f}$ will be a probability density function. Furthermore, $\hat{f}$ inherits the continuity and differentiability of k. The only major disadvantage in using kernels to estimate density functions is that they sometimes produce noise in the tails of estimates.

3.1.4. *Nearest Neighbor Method.* This method attempts to account for the local density of data. Suppose the density at t is $f(t)$. Then with a sample size of n, we would expect $2rnf(t)$ observations in an interval $[t - r, t + r]$. Define $d(x - y) = |x - y|$ and for each t define

$$d_1(t) \leq ... \leq d_n(t)$$

to be the distance from t to the points in the sample. The kth nearest estimate is defined by

$$\hat{f}(t) = \frac{k - 1}{2nd_k(t)}.$$

This is obtained by letting $r = k-1$ in the interval $[t-d_k(t), t+d_k(t)]$. The nearest neighbor estimate is not continuous. Note that $d_k(t)$ is continuous, but its derivative is discontinuous at $\frac{1}{2}(x_j + x_{j+k})$. It does not provide a probability density since d does not integrate to 1.

Let $\int k(x) = 1$. The generalized kth nearest neighbor estimate is defined

$$\hat{f}(t) = \frac{1}{nd_k(t)} \sum_{i=1}^{n} k(\frac{t - x_i}{d_k(t)}).$$

3.1.5. *Variable Kernel Method.* Rather then the uniform heights of regular kernel estimation, the height of the bumps at each x varies from data point to point. Define $d_{j,k}$ to be the distance from x_j to the kth nearest point in the set of the other $n-1$ data points. Let h be the smoothing parameter.

$$\hat{f}(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{hd_{j,k}} k\left(\frac{t-x_j}{hd_{j,k}}\right).$$

There are many ways to chose this smoothing parameter. The idea is that data points in sparser regions will get flatter bumps.

3.1.6. *Orthogonal Series Estimators.* All of the previous estimation techniques have been related under the idea of kernels. Orthogonal series estimators would estimate a function by finding the coefficients of f with respect to an orthogonal basis. Let us consider the Fourier basis and estimate f on the interval $[0,1]$ by its Fourier coefficients. Define $\phi_v(x)$ by

$$\phi_0(x) = 1 \qquad \phi_{2r-1}(x) = \sqrt{2}\cos 2\pi r x \qquad \phi_{2r}(x) = \sqrt{2}\sin 2\pi r x$$

for $r = 1, 2, ...$ Then almost everywhere $f(x) = \sum_{v=0}^{\infty} f_v \phi_v$, where for each $v \geq 0$,

$$f_v = \int_0^1 f(x)\phi_v(x)dx.$$

Suppose X is a random variable with density f. Then

$$f_v = E\phi_v(x).$$

Hence a natural unbiased estimator of f_v based on a sample $X_1, ..., X_n$ from f is

$$\hat{f}_v = \frac{1}{n}\sum_{i=1}^{n} \phi_v(X_i).$$

However, the sum $\sum \hat{f}_v \phi_v$ will not be a good estimate of f, but will converge to a sum of delta functions of the observations. To see this, let

$$w(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i).$$

Then for each v,

$$\hat{f}_v = \int_0^1 w(x)\phi_v(x)dx.$$

These $\hat{f}_v$ are the Fourier coefficients of the function $w(x)$. We must somehow smooth this to get a useful estimate. One way to do this is by truncating the expansion. Choose an integer k. Then let

$$\hat{f}(x) = \sum_{v=0}^{k} \hat{f}_v \phi_v(x).$$

Another way would be to use a sequence of weights $\lambda_v$which satisfy $\lambda_v \to 0$ as $v \to \infty$.

$$\hat{f}(x) = \sum_{v=0}^{\infty} \lambda_v \hat{f}_v \phi_v(x).$$

We can use other orthonormal basis as well. Suppose $a(x)$ is a weighting function and $\{\psi_v\}$ is a series satisfying for $u, v \geq 0$

$$\int_{-\infty}^{\infty} \psi_u(x)\psi_v(x)a(x)dx = \begin{cases} 1 \text{ if } u = v \\ 0 \text{ otherwise.} \end{cases}$$

Then

$$\hat{f}_v = \frac{1}{n} \sum_i \psi_v(x_i)a(x_i)$$

and

$$\hat{f}(x) = \sum_{v=0}^{k} \hat{f}_v \psi_v(x) \text{ or } = \sum_{v=0}^{\infty} \lambda_v \hat{f}_v \psi_v(x).$$

The properties of these estimates will match the properties of whatever orthonormal series we use. $\hat{f}(x)$ inherits the continuity and differentiability of the functions $\{\psi_v\}$.

3.1.7. *Maximum Penalized Likelihood Estimators.* The likelihood of a curve g as a density underlying a set of identically independent distributed observations is given by

$$L(g|X_1, ..., X_n) = \prod_{i=1}^{n} g(x_i).$$

This quantity has no finite maximum over all densities. Let $\hat{f}_n$ be the naive estimator with window width $\frac{1}{2}h$. Then $\hat{f}_n(x_i) \geq \frac{1}{nh}$. So

$$\prod \hat{f}_n(x_i) \geq n^{-n}h^{-n} \to \infty \text{ as } h \to 0.$$

Our method is to define $R(g)$, a function which quantifies the roughness of g.

$$R(g) = \int_{-\infty}^{\infty} (g'')^2$$

Define the penalized loglikelihood by

$$l_\alpha(g) = \sum_{i=1}^{n} \log g(x_i) - \alpha R(g)$$

where α is a positive smoothing parameter. This function represents the conflict between smoothness and goodness of fit to the data. Here $R(g)$ is the smoothness, and $\sum \log g(x_i)$ measures the fit. We would wish to maximize the likelihood $l_\alpha(g)$ over $\int_{-\infty}^{\infty} g = 1$, $g(x) \geq 0$, and $R(g) < \infty$.

3.1.8. *General Weight Function Estimates.* Let $w(x, y)$ be our weight with

$$\int_{-\infty}^{\infty} w(x, y)dy = 1 \qquad w(x, y) \geq 0 \qquad \hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} w(x_i, t).$$

Many of the methods discussed earlier can be expressed in this more general framework, including the orthonormal function estimate. For example, letting

$$w(x, y) = \frac{1}{h} k \left(\frac{y - x}{n} \right)$$

yields the kernel estimator. Letting

$$w(x, y) = \sum_{v=0}^{k} \phi_v(x)\phi_v(y)$$

yields the orthogonal series estimator.

3.1.9. *Bounded Domains and Directional Data.* There are several situations where we may have extra conditions on our estimator $\hat{f}$. We may wish for our estimator to always be positive, or we may only wish to find an estimator for certain subsets of x. We will examine what may be done in some of these cases.

Sometimes we wish for our $\hat{f}$ to be zero for negative x. There are many practical situations where this may be important. We could only calculate $\hat{f}$ for positive x, and then set $\hat{f}(x) = 0$ for $x \leq 0$. This presents several problems. The estimator may not integrate to one afterwards. Also, data points near 0 will be overweighted. Another solution would be to adopt some orthonormal functions to the half-line.

One way to deal with the overweighted data points near 0 would be to transform the data for positive x by taking the logarithm of it. This maps our data to the real line. If the density estimated from the logarithms of the data is $\hat{g}$, then

$$\hat{f}(x) = \frac{1}{x}\hat{g}(\log x) \text{ for } x > 0.$$

Lastly, one could extend the data set by adding reflections. Suppose we perform an even extension $\{X_1, -X_1, ..., X_n, -X_n\}$. Construct a kernel estimate f^* for this. Then

$$\hat{f}(x) = \Big\{ 2f^*(x) \; x \geq 0 \quad 0 \; x < 0 \; .$$

This corresponds to a general weight function

$$w(x,y) = \frac{1}{h}\left[k\left(\frac{y-x}{h}\right) + k\left(\frac{y+x}{h}\right)\right]$$

or we could use a negative reflection, thus yielding the estimator

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[k\left(\frac{x-x_i}{h}\right) - k\left(\frac{x+x_i}{h}\right)\right].$$

One could also extend the data periodically. Later, when we study wavelet transforms, we will require that the data be of size 2^n where n is an integer. These last methods of extension, and also a method of extending data so that it is continuous are all commonly used methods of making data the right size to do wavelet decomposition and reconstruction.

3.2. **Properties of the Estimator Derived via the Kernel.** Suppose that $\{X_1, ..., X_n\}$ is an identically independent sample with a probability function f that we wish to estimate. $\hat{f}$ will be the kernel estimate with kernel k and window width h. There are many different criteria for choosing h.

Suppose we wish to estimate h by minimizing the Mean Square Error (MSE).

$$MSE_x(\hat{f}) = E\{\hat{f}(x) - f(x)\}^2 = \{E\hat{f}(x) - f(x)\}^2 + \text{var}\hat{f}(x)$$

We call the first term of the right side of the equation above the bias, and the second term the variance. One can see that there is a trade-off between minimizing the bias and the variance, which is adjusted by changing the smoothness of the estimate.

Consider the Mean Integrated Square Error (MISE).

$$MISE(\hat{f}) = E\int\{\hat{f}(x) - f(x)\}^2 dx = \int E\{\hat{f}(x) - f(x)\}^2 dx \tag{3.2}$$

$$= \int MSE_x(\hat{f})dx = \int \{E\hat{f}(x) - f(x)\}^2 dx + \int var\hat{f}(x)dx.$$

Suppose

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, t),$$

the general weight function estimate. Recall that this method encompasses the kernel estimate. Then

$$E\hat{f}(t) = \int w(x,t)f(x)dx$$

and since the X_i are independent,

$$\text{var}\hat{f}(t) = \frac{1}{n}\text{var}w(x_i,t) = \frac{1}{n}\left[\int w(x,t)^2 f(x)dx - \{\int w(x,t)f(x)dx\}^2\right].$$

Note that the bias does not depend on the sample size n, only on the weight function. Let $h = h(n)$ and

$$w(x,y) = \frac{1}{h}k\left(\frac{y-x}{h}\right)$$

and we have the kernel estimator where

$$E\hat{f}(x) = \int \frac{1}{h}k\left(\frac{x-y}{h}\right) f(y)dy.$$

When (3.2) can be computed, we can minimize it with respect to h to find the optimal window width.

3.2.1. *Approximate Properties of the Estimator from the Kernel.* Suppose the kernel k is a symmetric function satisfying

$$\int k(t)dt = 1 \quad \int tk(t)dt = 0 \quad \int t^2 k(t)dt = k_2 \neq 0 \tag{3.3}$$

and f has continuous derivatives of all required orders.

The bias in estimation of $f(x)$ does not depend directly on the sample size n, but it does depend on h, the window width. If $h = h(n)$, then the bias will depend on n.

$$bias_h(x) = E\hat{f}(x) - f(x) = \int h^{-1}k\left(\frac{x-y}{n}\right) f(y)dy - f(x).$$

Let $y = x - ht$. Then $dy = -hdt$. Then since $\int k(t)dt = 1$,

$$\int k(t)f(x-ht)dt - f(x) = \int k(t)\left(f(x-ht) - f(x)\right)dt.$$

We examine the Taylor series expansion of $f(x-ht)$.

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + ...$$

$$bias_h(x) = -hf'(x)\int tk(t)dt + \frac{1}{2}h^2f''(x)\int t^2k(t)dt + ...$$

Then by (3.3),

$$= \frac{1}{2}h^2f''(x)k_2 + O(h^3). \tag{3.4}$$

Then

$$\int bias_h(x)^2dx \approx \frac{1}{4}h^4k_2^2\int f''(x)^2dx. \tag{3.5}$$

A similar calculation yields

$$var\hat{f}(x) \approx n^{-1}h^{-1}f(x)\int k(t)^2dt$$

so that

$$\int var\hat{f}(x)dx \approx n^{-1}h^{-1}\int k(t)^2dt. \tag{3.6}$$

To minimize the MISE, we wish to choose an h which will make (3.5) and (3.6) small. One can now see more clearly the trade-off between these two errors.

3.2.2. *Ideal Window Width and kernel.* We wish to minimize

$$\frac{1}{4}h^4k_2^2\int f''(x)^2dx + n^{-1}h^{-1}\int k(t)^2dt \tag{3.7}$$

with respect to h. Then by using calculus

$$h_{opt} = k_2^{-\frac{2}{5}}\left\{\int k(t)^2dt\right\}^{\frac{1}{5}}\left\{\int f''(x)^2dx\right\}^{-\frac{1}{5}}n^{\frac{1}{5}}. \tag{3.8}$$

Note h does depend on f, but also that as n increases, h decreases. Putting h_{opt} into (3.7) shows

$$MISE \approx \frac{5}{4}C(k)\left\{\int f''(x)^2dx\right\}^{\frac{1}{5}}n^{-\frac{4}{5}}$$

$$C(k) = k_2^{\frac{2}{5}} \left\{ \int k(t)^2 dt \right\}^{\frac{4}{5}}.$$

So, to further decrease the MISE, we would like to choose a k with a small $C(k)$.

If $k_2 \neq 1$, replace the kernel with $k_2^{-\frac{1}{2}} k(k_2^{-\frac{1}{2}} t)$. Then minimizing $C(k)$ is reducing

$$\int k(t)^2 dt \text{ subject to } \int k(t)dt = \int t^2 k(t)dt = 1.$$

This is solved by the Epanechnikov kernel below.

$$k_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2) & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 \text{ otherwise} \end{cases}$$

Define the efficiency of k to be

$$eff(k) = \left\{ \frac{C(k_e)}{C(k)} \right\}^{\frac{5}{4}} = \frac{3}{5\sqrt{5}} \left\{ \int t^2 k(t)dt \right\}^{-\frac{1}{2}} \left\{ k(t)^2 dt \right\}^{-1}.$$

If we examine a table comparing efficiencies, we can see that most commonly used kernels are much the same in terms of efficiency. So, often kernels are chosen based on other criteria such as differentiability or the amount of computational effort required to implement them.

3.2.3. *Choosing the Smoothing Parameter using assumptions on $f(x)$.* There are many different ways to chose the smoothing parameter h. The first and most obvious way is to just plot the data with different window widths and choose the "best looking" h.

Another way to chose h is by making some assumption about the distribution of f and then choosing a value for $\int f''(x)^2 dx$ in the expression for window width (3.11) which is discussed later. For example, one could assume that $f(x)$ is normally distributed. In that case

$$\int f''(x)^2 dx = \sigma^{-5} \int \phi''(x)^2 dx = \frac{3}{8}\pi^{-\frac{1}{2}}\sigma^{-5}. \tag{3.9}$$

Using the Gaussian kernel, and putting (3.9) into (3.11) we obtain

$$h_{opt} = (4\pi)^{-\frac{1}{10}} \frac{3}{8}\pi^{-\frac{1}{2}}\sigma n^{-\frac{1}{5}}$$

$$= \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}} = 1.06\sigma n^{-\frac{1}{5}}.$$

Note that this only works if one assumes you have a normal distribution.

Alternately, using the innerquartile range R

$$h_{opt} = 0.79Rn^{-\frac{1}{5}}.$$

The best of both worlds can be obtained by using the minimum of these two things.

3.2.4. *Least-squares Cross Validation.* This method of choosing h is completely automatic.

$$\int \left(\hat{f} - f\right)^2 = \int \hat{f}^2 - 2\int \hat{f}f + \int f^2$$

We have no control over the $\int f^2$ term, so we minimize

$$R\left(\hat{f}\right) = \int \hat{f}^2 - 2\int \hat{f}f.$$

We construct $R\left(\hat{f}\right)$ from the data, and then minimize it over h to get the window width. Define $\hat{f}_{-i}$ to be the estimate of f from all data except X_i.

$$\hat{f}_{-i}(x) = (n-1)^{-1} h^{-1} \sum_{j \neq i} k\left(h^{-1}(x - X_j)\right).$$

Define

$$M_0(h) = \int \hat{f}^2 - 2n^{-1} \sum_i \hat{f}_{-i}(X_i).$$

Note

$$En^{-1} \sum_i \hat{f}_{-i}(X_i) = E\hat{f}_{-n}(X_n) = E\int \hat{f}_{-n}(x) f(x) dx = E\int \hat{f}(x) f(x) dx,$$

so $EM_0(h) = ER\left(\hat{f}\right)$, and minimizing $E(M_0)$ is close to minimizing M_0. Let $k^{(2)}$ be the convolution of the kernel with itself. We can write a simpler function $M_1(h)$ to minimize with $M_1(h) \approx M_0(h)$.

$$M_1(h) = n^{-2} h^{-1} \sum_i \sum_j k^*\left(h^{-1}(X_i - X_j)\right) + 2n^{-1}h^{-1}k(0)$$

where $k^*(t) = k^{(2)}(t) - 2k(t).\hat{}$

Stone's theorem from [24] gives us a strong large sample justification of cross-validation. This theorem says that if $I_{s\times v}(X_1, ...X_n)$ is the integrated square error of the density estimate constructed using the smoothing parameter that minimizes $M_1(h)$ and $I_{opt}(X_1, ...X_n)$ is the minimum of $\int \left(\hat{f} - f\right)^2$ over all h, and under mild conditions on the kernel with fixed data, then with probability 1,

$$\frac{I_{s\times v}}{I_{opt}} \to 1 \text{ as } n \to \infty.$$

So, asymptotically, least squares cross-validation achieves the best possible choice of smoothing parameter, in the sense of minimizing the MISE.

However, there are some errors in cross-validation that we must consider. We must note that rounding data can create serious errors in our estimates. In a data set $X_1, ...X_n$ let m be the number of pairs $i < j$ for which $X_i = X_j$. For example, if the data set is a histogram of counts k_r,

$$m = \sum_r \frac{1}{2} k_r \left(k_r - 1\right).$$

If a data set of size n is discretized to a grid of l points, then by Jensen's inequality

$$\frac{m}{n} \geq \frac{1}{2} \cdot \frac{n}{l} - 1.$$

In fact, if $\frac{m}{n}$ is larger than some threshold value β depending on k the kernel, then $M_1(h) \to -\infty$ as $h \to 0$. This means that our minimization technique will give us $h = 0$. One can compute $\beta = \frac{1}{2} k^{(2)}(0) / \left\{2k(0) - k^{(2)}(0)\right\}$. For the normal kernel $\beta = 0.55$.

It is dangerous to use least squares cross validation for discretized data. Small variations in the data mean a small choice of h would be troublesome.

3.2.5. *Likelihood Cross Validation.* Suppose in addition to the original data set, an independent observation Y from f were available. Then the likelihood of f as the density underlying the observation Y would be $\log f(Y)$, with h the variable, $X_1, ...X_n$ fixed. Note that $\log f(Y)$ would be the log likelihood of the smoothing parameter h. The likelihood cross-validation choice of h is the value of h which maximizes $CV(h)$.

$$CV(h) = n^{-1} \sum_{i=1}^{n} \log \hat{f}_{-i}(X_i)$$

$CV(h)$ yields a density estimate which is close to the true density in terms of the Kullback-Leibler information defined below:

$$I\left(f, \hat{f}\right) = \int f(x) \log \left(\frac{f(x)}{\hat{f}(x)}\right) dx.$$

Then we see

$$E\left\{CV(h)\right\} = E \log \hat{f}_{-n}(X_n) = E \int f(x) \log \hat{f}_{n-1}(x) dx$$

$$\approx E \int f(x) \log \hat{f}(x) dx = -E\left\{I\left(f, \hat{f}\right)\right\} + \int f \log f.$$

So up to a constant, this is an unbiased estimator of the Kullback-Leibler error. However, this only works for very specific choices of $f(x)$.

3.2.6. *Test Graph Method.* From [22] we have the following. Suppose that kernel k is symmetric and satisfies certain conditions, and $\int x^2 k(x)dx$ is nonzero. Suppose f is uniformly continuous and $|f''| < \infty$. Now choose $h = h(n)$ to ensure the most rapid possible convergence of $\sup|\hat{f} - f| \to 0$. Then using the same h with $n \to \infty$

$$\frac{\sup|\hat{f}'' - E\hat{f}''|}{\sup|E\hat{f}''|} \to m \tag{3.10}$$

where

$$m = \frac{1}{2}\int |x^2 k(x)dx| \left\{ \int (k'')^2 \, dx / \int k^2 dx \right\}^{\frac{1}{2}}.$$

This m is a constant which depends only on the kernel. If k is the Gaussian kernel then $m \approx 0.4$. We try different h's and pick the one that yields a $\hat{f}$ which corresponds to (3.10). We would then choose the h that gives us this ratio of noise to trend.

3.2.7. *Internal Estimation of Density Roughness.* Recall the formula for an optimal h.

$$h_{opt} = k_2^{\frac{-2}{5}} \left\{ \int k(t)^2 dt \right\}^{\frac{1}{5}} \left\{ \int f''(x)^2 dx \right\}^{-\frac{1}{5}} n^{\frac{1}{5}}. \tag{3.11}$$

Let

$$\alpha(k) = k_2^{-\frac{2}{5}} \left\{ \int k(t)^2 dt \right\}^{\frac{1}{5}} \qquad \beta(f) = \left\{ \int f''(x)^2 \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}}.$$

Then

$$h_{opt} = \alpha(k)\beta(f)n^{-\frac{1}{5}}.$$

Let the estimate of $\beta(f)$ be

$$\hat{\beta}(h_0) = \left(\int \hat{f}_0''^2 \right)^{-\frac{1}{5}} = \beta(\hat{f}_0).$$

Here $\hat{f}_0$ is the density estimate constructed with h_0. For our new estimate, we would use

$$h_1 = \alpha(k)\hat{\beta}(h_0)n^{-\frac{1}{5}}.$$

To avoid choosing an initial h_0, one could use an iterative approach.

$$h_i = \alpha(k)\hat{\beta}(h_{i-1})n^{-\frac{1}{5}}$$

In practice, one would solve

$$h = \alpha(k)\hat{\beta}(h)n^{-\frac{1}{5}}$$

using Newton's method.

3.2.8. *Finding estimators by using the Fourier Transform.* Recall that the definition of the Fourier transform is defined as

$$\tilde{g}(s) = (2\pi)^{-\frac{1}{2}} \int e^{ist} g(t)dt.$$

The discrete Fourier transform of the data is defined

$$u(s) = (2\pi)^{-\frac{1}{2}} n^{-1} \sum_{j=1}^{n} \exp\{isX_j\}.$$

Let $\tilde{f}_n(s)$ be the Fourier transform of the kernel density estimate.

$$\tilde{f}_n(s) = (2\pi)^{\frac{1}{2}} \tilde{k}(hs)u(s).$$

We used here that the Fourier transform of $h^{-1}k\left(h^{-1}t\right)$ is $\tilde{k}(hs)$. One could then use the fast Fourier transform to find u and then invert $\tilde{f}_n$ to find $\hat{f}$.

3.2.9. *Bias Reduction Technique.* Suppose we don't require that k is non-negative and choose a k that satisfies

$$\int k(t)dt = 1 \qquad \int t^2 k(t)dt = k_2 \neq 0 \tag{3.12}$$

$$\int t^4 k(t)dt = k_4 \neq 0 \tag{3.13}$$

We use the Taylor expansion of $f(x - ht)$ as before to get

$$bias_h(x) = \frac{1}{24} h^4 f^4(x) k_4 + o\left(h^5\right).$$

Here the h and h^3 terms drop out because of the symmetry of k. The h^2 term drops out because of (3.12). Then the estimated MISE is

$$\frac{1}{576}h^8k_4^2\int f^4(x)^2dx + n^{-1}h^{-1}\int k(t)^2dt. \tag{3.14}$$

Then minimizing (3.14) yields

$$h_{opt} = (72)^{\frac{1}{9}}k_4^{-\frac{2}{9}}\left\{k(t)^2dt\right\}\left\{\int f^4(x)^2dx\right\}^{-\frac{1}{9}}n^{-\frac{1}{9}}.$$

Substituting h_{opt} into the MISE yields a minimum of

$$= C_4(k)\left\{\int f^4(x)^2dx\right\}^{\frac{1}{9}}n^{-\frac{8}{9}}$$

with

$$C_4(k) = 9^{\frac{8}{9}}2^{-\frac{10}{3}}k_4^{\frac{2}{9}}\left\{\int k(t)^2dt\right\}^{\frac{8}{9}}.$$

We now wish to use a kernel which makes $C_4(k)$ as small as possible. One choice is

$$k(y) = \left\{\frac{3}{8}\left(3-5y^2\right) \text{ if } |y|<1 \quad 0 \text{ otherwise.}\right.$$

One could also make a more artificial choice. For any kernel $k_0(t)$, choose

$$k(t) = \frac{k_0(t) - c^{-3}k_0\left(c^{-1}t\right)}{1-c^{-2}}$$

where $k_0(t)$ is some positive kernel. The new kernel $k(t)$ satisfies (3.12) and (3.13).

One could further extend the requirements for the kernel to

$$\int t^j k(t)dt = 0 \text{ for } 0<j<2m\int t^{2m}k(t)dt \neq 0.$$

The bias for these requirements would be of order h^{2m}, and the optimal MISE would be $o\left(h^{-\frac{4m}{4m+1}}\right)$.

The problem with this is that the kernel may be negative in more places, and may give $\hat{f}$ odd properties. $\hat{f}$ would no longer be a probability density function as before.

3.2.10. *Asymptotic properties of $\hat{f}$.* Lastly, we report some properties of the convergence of $\hat{f}$ to f. From [21] we have the following.

Theorem 8. *If k is a bounded Borel function and if*

$$\int |k(t)|dt < \infty, \quad \int k(t)dt = 1 \tag{3.15}$$

and

$$|tk(t)| \to 0 \text{ as } t \to \infty,$$

and

$$h_n \to 0 \text{ and } nh_n \to \infty \text{ as } n \to \infty$$

then if f is continuous at x

$$\hat{f}(x) \to f(x) \text{ in probabilty as } n \to \infty. \tag{3.16}$$

We note that the conditions (3.15) are satisfied by almost any conceivable kernel.

In other words,

$$P(|\hat{f}(x) - f(x)| > \epsilon) \to 0.$$

We also have the following theorem.

Theorem 9. *Suppose k is bounded, has bounded variation, satisfies (3.15), and has Lebesque measure zero. Suppose f is uniformly continuous on $(-\infty, \infty)$ and*

$$h_n \to 0 \text{ and } nh_n (\log n)^{-1} \to \infty \text{ as } n \to \infty.$$

Then

$$\sup_x |\hat{f}(x) - f(x)| \to 0 \text{ as } n \to \infty. \tag{3.17}$$

Lastly, we have a theorem which requires no assumptions on $f(x)$.

Theorem 10. *Assume k is a non-negative Borel Function which integrates to one. Then if*

$$h_n \to 0 \text{ and } nh_n \to \infty \text{ as } n \to \infty,$$

$$\int |\hat{f}(x) - f(x)| dx \to 0$$

with probability 1 as $n \to \infty$.

In other words,

$$P\left(\int \left|\hat{f}(x) - f(x)\right| dx = 0\right) \to 1 \quad \text{as } n \to \infty.$$

3.2.11. *Conclusion.* We can see from the theorems above that under fairly lax conditions we can achieve a closeness between $\hat{f}(x)$ and $f(x)$ in several different senses. However, we do not have information about the Mean Square Error, and in the strongest theorem above we require that $f(x)$ be continuous. This requirement is not practical in many situations. We also have no information about the speed of converge to 0 as $n \to \infty$. We answer this question in Part 4. This motivates our study of wavelets.

4. A brief introduction to Wavelets.

Because of the spatial adaptability of wavelets and their ability to model arbitrarily small intervals they are a powerful tool in modeling data. As we will see in later sections, because of their structure wavelets perform better than linear methods in some situations, and always as well as linear methods. It is for this reason that we study the continuous wavelet transform and the discrete wavelet transform.

4.1. **A brief wavelet review.**

4.1.1. *The continuous wavelet transform and multiresolution analysis.* The material in this section was taken from [6, 25].

In very general terms, wavelet analysis on a signal means separating a signal into low and high frequency components. In discrete terms, this amounts to convolving our digital signal with a filter.

We can define wavelets in terms of a multiresolution analysis, or MRA. Suppose $f \in L_2(\mathbb{R})$, that is,

$$\left(\int_{-\infty}^{\infty} |f(x)|^2 dx\right)^{\frac{1}{2}} < \infty.$$

An MRA will define a sequence of spaces V_j, V_{j+1} such that the projections of f onto these spaces give finer and finer approximations (as $j \to \infty$) of f.

Define $T_n g(x) = g(x-n)$ and $D_{2^j} g(x) = 2^{\frac{j}{2}} g\left(2^j x\right)$. We present a framework for constructing functions $\psi(x) \in L^2(\mathbb{R})$ such that

$$\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}} = \left\{2^{\frac{j}{2}}\psi\left(2^j x - k\right)\right\}_{j,k\in\mathbb{Z}} = \{D_{2^j} T_k(\psi(x))\}_{j,k\in\mathbb{Z}}$$

is an orthonormal basis on $\mathbb{R}$.

Definition 11. The collection $\{T_n g(x)\}_{n\in\mathbb{Z}}$ is called an orthonormal system of translates.

This collection is an orthonormal basis for

$$\overline{\text{span}}\left\{T_n g(x)\right\}.$$

That is, $f \in \overline{\text{span}}\left\{T_n g(x)\right\}$ if and only if $f(x) = \sum_n \langle f, T_n g\rangle T_n g(x)$.

Within the set of translations, we see that a dilation performed on the set would make functions finer and finer until all of $\mathbb{R}$ was covered.

Definition 12. A MRA on $\mathbb{R}$ is a sequence of subspaces $\{V_j\}_{j\in\mathbb{Z}}$ of functions L^2 on $\mathbb{R}$ such that

(1) For all $j \in \mathbb{Z}$, $V_j \subseteq V_{j+1}$.

(2) If $f(x)$ is C_c^0 (that is, compact and continuous) on $\mathbb{R}$, then $f(x) \in \overline{\text{span}}\,\{V_j\}_{j\in\mathbb{Z}}$.

(3) $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$.

(4) $f \in V_0$ iff $D_{2^j} f(x) \in V_j$.

(5) There exists a function $\varphi(x) \in L^2(\mathbb{R})$ called the scaling function such that the collection $\{T_n\varphi(x)\}$ is an orthonormal system of translates and $V_0 = \overline{\text{span}}\,\{T_n\varphi(x)\}$.

We form an MRA by first defining V_0 and then letting

$$V_j = \{f(x) : f(x) = D_{2^j} g(x),\ g(x) \in V_0\}.$$

Definition 13. $\varphi_{j,k}(x) = 2^{\frac{j}{2}}\varphi\left(2^j x - k\right) = D_{2^j} T_k \varphi(x)$.

We also define the following operators.

Definition 14. Define the approximation and detail operators as follows.

$$P_j f(x) = \sum_k \langle f, \varphi_{j,k}\rangle\, \varphi_{j,k}(x)$$

$$Q_j f(x) = P_{j+1} f(x) - P_j f(x).$$

Note that $\{\varphi_{j,k}(x)\}_{j,k}$ is an orthonormal basis for V_j.

Lemma 15. *There exists an l^2 sequence of coefficients $\{h(k)\}$ such that*

$$\varphi(x) = \sum_k h(k) 2^{\frac{1}{2}} \varphi(2x - k).$$

This is true because $\{\varphi_{j,k}(x)\}$ is a basis.

$$\varphi(x) = \sum_k \langle \varphi, \varphi_{1,k}\rangle\, 2^{\frac{1}{2}} \varphi(2x - k)$$

Therefore, $h(k) = \langle \varphi, \varphi_{1,k}\rangle$.

We have discussed the scaling function, we can now construct a wavelet orthonormal basis given an MRA.

Theorem 16. *Let $\{V_j\}$ be an MRA and $\varphi(x)$ be a scaling function with filter $h(k)$. Define a wavelet filter*

$$g(k) = (-1)^k \overline{h(1-k)}.$$

Then we have wavelet

$$\psi(x) = \sum_k g(k) 2^{\frac{1}{2}} \varphi(2x-k).$$

Then $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$ is a wavelet orthonormal basis in $\mathbb{R}$. In other words, given any $J \in \mathbb{Z}$

$$\{\varphi_{J,k}(x)\}_{k\in\mathbb{Z}} \bigcup \{\psi_{j,k}(x)\}_{j\geq J,k\in\mathbb{Z}}$$

is an orthonormal basis on $L^2(\mathbb{R})$.

This is a very powerful theorem. It allows us to examine signals at certain levels of detail and perform wavelets transform to describe certain subsets of details.

Definition 17. For each $j \in \mathbb{Z}$ the wavelet subspace W_j is defined

$$W_j = \overline{\text{span}} \{\psi_{j,k}(x)\}_{k\in\mathbb{Z}}.$$

Lemma 18. *If there exists $\psi(x) \in V_1$ such that*

a) $\{T_n\psi(x)\}$ is orthonormal

b) $\{T_n\psi, T_m\varphi\} = 0$ for all $n, m \in \mathbb{Z}$

c) Given $f(x) \in C_c^0(\mathbb{R})$, $Q_0 f(x) \in \overline{span} \{T_n\psi(x)\} = W_0$.

Then $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}} = \left\{2^{\frac{1}{2}}\psi\left(2^j x - k\right)\right\}_{j,k\in\mathbb{Z}}$ is a wavelet orthonormal basis on $L^2(\mathbb{R})$.

Note that $V_{j+1} = V_j \oplus W_j$ and $L_2(\mathbb{R}) = \overline{V_{j_0} \oplus_{j=j_0}^{\infty} W_j}$. So if $f \in L_2(\mathbb{R})$,

$$f = P_{j_0} f + \sum_{j=j_0}^{\infty} Q_j f$$

$$= \sum_k \langle f, \varphi_{j_0,k}\rangle \varphi_{j_0,k} + \sum_{j=j_0}^{\infty} \sum_k \langle f, \psi_{j,k}\rangle \psi_{j,k}.$$

We require here that $\int \varphi(x) = 1$ and $\int \psi(x) = 0$.

One important property of wavelets is that of vanishing moments.

Definition 19. Suppose that $\psi(x)$ has N vanishing moments. Then

$$\int_{\mathbb{R}} x^p \psi(x) dx = 0$$

for $p = 0, ..., N-1$.

If $\psi(x)$ has N vanishing moments then any polynomial of degree $N-1$ can be reproduced exactly by the scaling function.

4.1.2. *The discrete wavelet transform.* In statistical papers dealing with discrete data, we wish to apply wavelet transforms to collections of numbers which are not continuous functions. We want to perform a discrete wavelet transform or DWT. The information in this section follows the lines of [25]. Suppose we have a signal or data $\{c_0(k)\}_{k\in\mathbb{Z}}$. In order to analyze the data, we assume

$$c_0(k) = \langle f, \varphi_{0,k}\rangle .$$

This allows the MRA construction to work, but we had to make an interesting (and incorrect) assumption. The idea is that the basis functions needed to represent $c_0(k)$ exactly have such small support that they can be considered to be the delta function. Thus, $c_o(k) \in V_0$.

If $h(n)$ and $g(n)$ are the refinement sequences of φ and ψ respectively.

$$c_j(k) = \langle f, \varphi_{-j,k}\rangle = \sum_n c_{j-1}(n)\overline{h(n-2k)} \tag{4.1}$$

$$d_j(k) = \langle f, \psi_{-j,k}\rangle = \sum_n c_{j-1}(n)\overline{g(n-2k)} \tag{4.2}$$

$$P_{-j}f(x) = \sum_n c_j(n)\varphi_{-j,n}(x)$$

$$Q_{-j}f(x) = \sum_n d_j(n)\psi_{-j,n}(x)$$

$$P_{-j}f(x) = P_{-j-1}f(x) + Q_{-j-1}f(x).$$

Thus,

$$c_j(k) = \sum_n c_{j+1}(n)h(k-2n) + \sum_n d_{j+1}(n)g(k-2n). \tag{4.3}$$

The key object in the DWT is the scaling filter $h(k)$ and not $\varphi(x)$ the scaling function. This scaling function must satisfy (4.1) and (4.2) to be inverted by (4.3). These are called the Quadrature Mirror Filter or QMF conditions.

Theorem 20. *Let $\{V_j\}$ be an MRA with scaling filter $h(k)$ and wavelet filter $g(k) = (-1)^k\overline{h(1-k)}$.*

a) $\sum_n h(n) = \sqrt{2}$

b) $\sum_n g(n) = 0$

c) $\sum_k h(k)\overline{h(k-2n)} = \sum_k g(k)\overline{g(k-2n)} = \delta(n)$

d) $\sum_k g(k)\overline{h(k-2n)} = 0$ *for all* $n \in \mathbb{Z}$ *and*

e) $\sum_k \overline{h(m-2k)}h(n-2k) + \sum_k \overline{g(m-2k)}g(n-2k) = \delta(n-m)$.

We note here that some authors divide out the factor of $\sqrt{2}$ from the refinement sequence, thus making a) into $\sum_n h(n) = 1$.

Definition 21. Let $c(n)$ be a signal

a) The downsampling operator $\downarrow$ is defined

$$(\downarrow c)(n) = c(2n).$$

b) The upsampling operator $\uparrow$ is defined

$$(\uparrow c)(n) = \begin{cases} c\left(\frac{n}{2}\right) & \text{n even} \\ 0 & \text{n odd.} \end{cases}$$

We have the following operators which are the analogues of the P and Q operators from the last section.

Definition 22. Define the approximation operator H and detail operator G by

$$(Hc)(k) = \sum_n c(n)\overline{h(n-2k)}$$

$$(Gc)(k) = \sum_n c(n)\overline{g(n-2k)}.$$

We define their adjoints by

$$(H * c)(k) = \sum_n c(n)h(k-2n)$$

$$(G * c)(k) = \sum_n c(n)g(k-2n).$$

Note that

(1) H and G can be thought of as convolution with the filters $\underline{h}(n) = \overline{h(-n)}$ and $\underline{g}(n) = \overline{g(-n)}$ followed by downsampling.

$$(Hc)(n) = \downarrow (c * \underline{h})(n)$$

$$(Gc)(n) = \downarrow (c * \underline{g})(n).$$

(2) H^* and G^* can be thought of as upsampling followed by convolution with h and g.

$$(H^*c)(n) = (\uparrow c) * h(n)$$

$$(G^*c)(n) = (\uparrow c) * g(n).$$

(3) H^* and G^* are formal adjoints of H and G.

Then the c, d and e conditions of Theorem 20 can be reformulated as

$$HH^* = GG^* = I$$

$$HG^* = GH^* = 0$$

$$H^*H + G^*G = I$$

respectively.

We can now define formally the DWT for signals.

Definition 23. Let $h(k)$ satisfy the QMF conditions. Define $g(k) = (-1)^k h(1-k)$ and let H, G, H^* and G^* be as just defined. Fix $J \in \mathbb{N}$. The DWT of a signal $c_0(n)$ is the collection of sequences

$$\{d_j(k) : 1 \leq j \leq J; k \in \mathbb{Z}\} \bigcup \{c_J(k) : k \in \mathbb{Z}\}$$

where

$$c_{j+1}(n) = (Hc_j)(n) \text{ and } d_{j+1}(n) = (Gc_j)(n).$$

The inverse transform is defined by

$$c_j(n) = (H^*c_{j+1})(n) + (G^*d_{j+1})(n).$$

We notice that the sums used to define the DWT are infinite despite the fact that we are working with a finite signal. There are two common methods used to deal with the problem. The first method is the zero-padding method. One pads the signal on either side with zero entries. This can create problems at the ends of the signal. The second method is periodization. This method of extension lessens the discontinuities at the ends, but also distorts the data.

Fact 24. *Suppose $c(n)$ is a 2^N periodic signal. Then $(Hc)(n)$ and $(Gc)(n)$ are 2^{N-1} periodic and $(H^*c)(n)$ and $(G^*c)(n)$ are 2^{N+1} periodic.*

We now note that the DWT of a period $M = 2^N$ signal is a transform taking the M vector

$$c_0 = \left[\begin{array}{cccc} c_0(0) & c_0(1) & \ldots & c_0(M-1) \end{array} \right]$$

into the M vector

$$d = [d_1|d_2|...|d_J|c_J]$$

where

$$d_j = \left[\begin{array}{cccc} d_j(0) & d_j(1) & \ldots & d_j\left(2^{-j}M - 1\right) \end{array} \right].$$

This linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^m$ can be represented by an $M \times M$ matrix W such that

$$Wc_0 = d.$$

In fact, for $M = p$,

$$W_p = \begin{pmatrix} H_p \\ G_p \end{pmatrix}$$

and

$$W_p^* W_p = \begin{pmatrix} H_p \\ G_p \end{pmatrix}^* \begin{pmatrix} H_p \\ G_p \end{pmatrix} = \begin{pmatrix} H_p^* & G_p^* \end{pmatrix} \begin{pmatrix} H_p \\ G_p \end{pmatrix}$$

$$= H_p^* H_p + G_p^* G_p = I_p$$

by the rewritten QMF conditions. Thus we see that W_p is an orthogonal matrix.

So, when performing a DWT, the first step is

$$W_M c_0 = \begin{pmatrix} d_1 \\ c_1 \end{pmatrix}$$

the second step is

$$\begin{pmatrix} I_{\frac{M}{2}} & 0 \\ 0 & W_{\frac{M}{2}} \end{pmatrix} \begin{pmatrix} d_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ c_2 \end{pmatrix}$$

and the jth step is

$$\begin{pmatrix} I_{(1-2^{-j})M} & 0 \\ 0 & W_{2^{-j}M} \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_{j-1} \\ c_{j-1} \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_{j-1} \\ d_j \\ c_j \end{pmatrix}.$$

Lastly, we note that the rows of W form an orthonormal basis for $\mathbb{R}^m$ called the discrete wavelet basis. These vectors can be calculated by taking the inverse DWT of e_i in $\mathbb{R}^m$. This concludes our review of the properties of wavelets.

5. Estimating $f(x)$ with no requirements on the underlying function.

Donoho and Johnstone in [8] introduce the concept of an oracle, which is an imaginary device which makes the best choice of estimator for a given f under certain circumstances. For instance, when one samples wavelets coefficients, one may decide just to keep these coefficients or filter them according to their size. An oracle would chose the "best" combination of coefficients to keep in terms of minimizing the risk. In [8] Donoho and Johnstone show that by using soft and hard thresholding with a derived "universal threshold", one can guarantee a risk which can be expressed in terms of the oracle risk. This estimator comes only from the data and performs better than piecewise polynomials in terms of convergence to f.

5.1. **Other methods of Estimation and Notation.**

5.1.1. *Notation.* Suppose the data we are working with is

$$y_i = f(t_i) + e_i \; i = 1, ..., n \tag{5.1}$$

where $t_i = \frac{i}{n}$ and the e_i are independently distributed as $N(0, \sigma^2)$. Let $f(\cdot)$ be the function we wish to estimate and $\hat{f}(\cdot)$ be the estimator. We measure the closeness of the estimator to $f(\cdot)$ in terms of quadratic loss $\left(\hat{f}(t) - f(t)\right)^2$. Specifically, let $f = (f(t_i))_{i=1}^n$ and $\hat{f} = \left(\hat{f}(t_i)\right)_{i=1}^n$ be the vectors of true and estimated values. Recall $\|x\|_{2,n}^2 = \sum_{i=1}^n x_i^2$ is the squared l_n^2 norm for vectors. Then the risk is

$$R\left(\hat{f}, f\right) = n^{-1} E \|\hat{f} - f\|_{2,n}^2. \tag{5.2}$$

We wish to minimize this risk.

5.1.2. *Other reconstruction Methods.* We define some other estimates for f with the notation

$$\hat{f}(\cdot) = T(y, d(y))(\cdot).$$

Here $T(y, \delta)$ is a reconstruction formula with "smoothing" parameter δ and $d(y)$ is a data-adaptive choice of that δ.

5.1.3. *Piecewise Constant Reconstruction.* Here δ is a list of L numbers defining a partition $(I_1, ..., I_L)$ of $[0, 1]$ such that

$$I_1 = [0, \delta_1), \; I_2 = [\delta_1, \delta_1 + \delta_2), \; ..., \; I_L = [\delta_1 + ... + \delta_{L-1}, \delta_1 + ... + \delta_L]$$

so that $\sum_{i=1}^L \delta_i = 1$. Here L is variable.

$$T_{PC}(y, \delta)(t) = \sum_{i=1}^{L} Ave(y_i : t_i \in I_i) \, 1_{I_i(t)}$$

Recall that $1_{(condition(t))}$ means

$$1_{(condition(t))} = \begin{cases} 1 \text{ if the condition is met} \\ 0 \text{ otherwise.} \end{cases}$$

5.1.4. *Piecewise Polynomial Reconstruction.* Here δ is the same as in (5.1.3), only reconstruction uses polynomials of degree D.

$$T_{PP(D)}(y, \delta)(t) = \sum_{i=1}^{L} \hat{p}_l(t) 1_{I_i(t)}$$

where $\hat{p}_l(t) = \sum_{k=0}^{D} a_k t^k$ is determined by minimizing the least square error below.

$$\sum_{t_i \in I_i} \left\{ \hat{p}_l \left(t_i \right) - y_i \right\}^2$$

5.1.5. *Variable-knot splines.* This estimator is the same as (5.1.4), except we require that the estimator be continuous and have continuous derivatives up to order $D-1$. If t_l is the left endpoint of I_l ($l = 1, ..., L$),

$$\left(\frac{d^k}{dt^k} s \right) \left(t_l^- \right) = \left(\frac{d^k}{dt^k} s \right) \left(t_l^+ \right).$$

Subject to this constraint minimize the quantity below.

$$\sum_{i=1}^{n} \left\{ s \left(t_i \right) - y_i \right\}^2$$

5.1.6. *Ideal Adaption with Oracles.* We wish to study ideal adaption. This is the performance which can be achieved from smoothing with the aid of an oracle. This oracle does not tell us what the true f is, but it does give us the best choice of δ for the true f. In (5.1.3), (5.1.4), and (5.1.5) this would be the true division of the partition. We define the ideal risk as

$$\mathcal{R}_{n,\sigma}(T, f) = \inf_{\delta} R(T(y, \delta), f).$$

For a kernel estimator, this would be the best smoothing parameter. It is dependent on a selection of $\Delta(f)$ which satisfies

$$R(T(y, \Delta(f)) = \mathcal{R}_{n,\sigma}(T, f).$$

This ideal is unattainable.

5.1.7. *Wavelet Reconstruction Preliminaries.* Suppose we have data $y = (y_i)_{i=1}^n$ with $n = 2^{J+1}$. Let M be the number of vanishing moments, S be the support width, and j_0 be the low-resolution cutoff. We may construct an $n \times n$ matrix W which is the finite wavelet transform matrix.

This matrix gives a vector w of wavelet coefficients with $w = Wy$. Since W is an orthogonal matrix, $y = W^T w$.

The vector w has 2^{J+1} elements. We index them dyadically $n - 1 = 2^{J+1} - 1$ following

$$w_{j,k} \ \left(j = 0, ..., J \ k = 0, ..., 2^j - 1 \right)$$

and the remaining element we label $w_{-1,0}$. Let $W_{j,k}$ denote the (j,k)th row of W. Then

$$y_i = \sum_{j,k} w_{j,k} W_{j,k}(i).$$

We call the $W_{j,k}$ wavelets. The plot of the vector $W_{j,k}$ looks like a localized wiggle. This is where the name 'wavelet' comes from.

For j and k with $j_0 \leq j < J - j_1$ and $S < k < 2^j - S$,

$$n^{\frac{1}{2}} W_{j,k}(i) \approx 2^{\frac{j}{2}} \psi\left(2^j t - k\right) \text{ for } t = \frac{i}{n}$$

where ψ is a fixed wavelet in the sense of the usual wavelet transform on $\mathbb{R}$. Information on this can be found in [19, 5]. This relation improves as n and j increase. Therefore, $w_{j,k}$ is localized to spatial positions near $t = k2^{-j}$ and frequencies near 2^j. There are two important properties that we need.

1) $W_{j,k}$ has vanishing moments up to order M if $j \geq j_0$.

$$\sum_{i=0}^{n-1} i^l W_{j,k}(i) = 0 \qquad \left(l = 0, ..., M;\ j \geq j_0;\ k = 0, ..., 2^j - 1\right)$$

2) $W_{j,k}$ is supported in $\left[2^{J-j}(k-s), 2^{J-j}(k+s)\right]$ if $j \geq j_0$.

Because of the spatial localization of wavelet basis, the wavelet coefficients allow one to determine if there is a significant change near t by looking at the $w_{j,k}$ with $j = j_0, ..., J$ near k and $k2^{-j} \approx t$. If these are large, then there is a significant change. Given a finite list δ of (j,k) pairs define $T_{sw}(y, \delta)$ by

$$T_{sw}(y, \delta) = \hat{f} = \sum_{(j,k) \in \delta} w_{jk} W_{jk}. \tag{5.3}$$

We reconstruct by choosing only a subset of the empirical wavelet coefficients. The 'sw' stands for selective wavelet reconstruction. We do this because every empirical wavelet coefficient contributes noise of variance σ^2, but only a few wavelet coefficients contribute signal.

For risk (5.2), the ideal risk is defined

$$\mathcal{R}_{n,\sigma}(sw, f) = \inf_{\delta} R_{n,\sigma}\left(T_{sw}(y, \delta), f\right)$$

with optimal spatial parameter $\delta = \triangle(f)$, namely a list of indexes attaining

$$R_{n,\sigma}\left(T_{sw}(y, \triangle(f), f\right) = \mathcal{R}_{n,\sigma}(sw, f).$$

Suppose we have $f = \sum_{i=1}^{n} p_i(t) 1_{I_i(t)}$ where f is a piecewise polynomial of degree D, and we have a wavelet basis with parameter $M \geq D$. Then (5.1.7) and (5.1.7) imply that the wavelet coefficients for f are all zero except for

i) coefficients at the coarse levels $0 \leq j \leq j_0$

ii) coefficients at $j_0 \leq j \leq J$ whose associated interval $\left[2^{-j}(k-s), 2^{-j}(k+s)\right]$ contains a breakpoint of f.

The number of coefficients that satisfy i) is fixed, and at each resolution level j $\left(\theta_{jk}, k = 0, ..., 2^j - 1\right)$ contains at most (# breakpoints)$\times(2s+1)$ which satisfy ii). If L is the number of pieces then

$$\#\left\{(j,k) : \theta_{jk} \neq 0\right\} \leq 2^{j_0} + (J + 1 - j_0)(2s+1)L.$$

Let $\delta^* = \left\{(j,k) : \theta_{jk} \neq 0\right\}$. Then because of the orthogonality of the (w_{jk}), $\sum_{(j,k)\in\delta^*} w_{jk}W_{jk}$ is the least squares estimate of f and

$$R\left(T\left(y, \delta^*\right), f\right) = n^{-1}\left\{\#\left(\delta^*\right)\right\}\sigma^2 \leq \left(C_1 + C_2 J\right) L\sigma^2/n \tag{5.4}$$

for all $n = 2^{J+1}$ with C_1 and C_2 depending linearly on s but not on f. Also note that

$$\mathcal{R}_{n\sigma}(sw, f) = o\left(\frac{\sigma^2 \log n}{n}\right).$$

5.1.8. *Results about coefficient estimation.* Suppose we are given observations $w = (w_i)_{i=1}^n$ with

$$w_i = \theta_i + \epsilon z_i \tag{5.5}$$

where $i = 1, ..., n$, and the z_i are iid as $N(0,1)$. Here $\epsilon > 0$ is the known noise level and $\theta = (\theta_i)$ is the true value. We define the risk

$$R\left(\hat{\theta}, \theta\right) = E\|\hat{\theta} - \theta\|_{2,n}^2. \tag{5.6}$$

Our oracle is based on a family of diagonal linear projections

$$T_{DP}(w, \delta) = (\delta_i w_i)_{i=1}^n$$

where $\delta_i \in \{0, 1\}$. These estimators either keep or discard a coordinate based on whether or not it is big enough. Suppose we knew the perfect choice of coordinates to keep to minimize the risk. These ideal coefficients are $\delta_i = 1_{(|\theta_i| > \epsilon)}$. This yields the ideal risk

$$\mathcal{R}_{n\sigma}(DP, \theta) = \sum_{i=1}^n \min\left(|\theta_i|^2, \epsilon^2\right).$$

Though this ideal risk cannot be attained, with a simple choice of estimator we can approximate it. Define the soft and hard thresholding operators by

$$\eta_H(w, \lambda) = w \cdot 1\left\{|w| > \lambda\right\}$$

$$\eta_S(w, \lambda) = sgn(w)\left(|w| - \lambda\right)_+$$

respectively. For the soft thresholding operator we have the following.

Theorem 25. *Assume (5.5) and (5.6). The estimator*

$$\hat{\theta}_i^{\mu} = \eta_S\left(w_i, \epsilon\left(2\log n\right)^{\frac{1}{2}}\right) \;\; i = 1, ..., n$$

satisfies

$$E\|\hat{\theta}^{\mu} - \theta\|_{2,n}^2 \leq (2\log n + 1)\left\{\epsilon^2 + \sum_{i=1}^{n} \min\left(\theta_i^2, \epsilon^2\right)\right\}$$

or

$$R\left(\hat{\theta}^{\mu}, \theta\right) \leq (2\log n + 1)\left\{\epsilon^2 + \mathcal{R}_{\epsilon}(DP, \theta)\right\}$$

for all $\theta \in \mathbb{R}^n$.

We can gain a better understanding of the choice of λ_n by examining the aysmptotics below.

Theorem 26. *Assume (5.5) and (5.6). Let*

$$\Lambda_n^{\#} = \inf_{\lambda} \sup_{\mu} \frac{\rho_{ST}(\lambda, \mu)}{n^{-1} + \min\left(\mu^2, 1\right)} \tag{5.7}$$

$$\lambda_n^* = \text{ the largest } \lambda \text{ attaining } \Lambda_n^* \text{ above.} \tag{5.8}$$

Then

$$\hat{\theta}_i^* = \eta_S\left(w_i, \lambda_n^* \epsilon\right) \;\; i = 1, ..., n \tag{5.9}$$

satisfies

$$E\|\hat{\theta} - \theta\|_{2,n}^2 \leq \Lambda_n^* \left\{\epsilon^2 + \sum_{i=1}^{n} \min\left(\theta_i^2, \epsilon^2\right)\right\}. \tag{5.10}$$

Furthermore

$$\Lambda_n^* \sim 2\log n, \;\; \lambda_n^* \sim \left(2\log n\right)^{\frac{1}{2}}.$$

We have a similar result for the hard thresholding operator.

Theorem 27. *With* (l_n) *a thresholding sequence sufficiently close to* $\left(2\log n\right)^{\frac{1}{2}}$ *in the following sense*

$$(1 - \gamma)\log\log n \leq l_n^2 - 2\log n \leq o\left(\log n\right)$$

the hard thresholding estimator

$$\hat{\theta}_i^{+} = w_i \cdot 1\left\{|w_i| > l_n \epsilon\right\}$$

satisfies for $l_n \sim 2\log n$ *the inequality*

$$R\left(\hat{\theta}^{+}, \theta\right) \leq l_n \left\{\epsilon^2 + \sum_{i=1}^{n} \min\left(\theta_i^2, \epsilon^2\right)\right\}$$

for all $\theta \in \mathbb{R}^n$ *and* $\gamma > 0$.

The proofs of these theorems will be given later.

5.1.9. *Results applied to function estimation.* We apply the previous results to function estimation. Let $n = 2^{J+1}$ and W be the wavelet coefficient matrix. This is an orthogonal transformation of $\mathbb{R}^n$ into $\mathbb{R}^n$. If $f = (f_i)$ and $\hat{f} = \left(\hat{f}_i\right)$ are two n-vectors and (θ_{jk}) and $\left(\hat{\theta}_{jk}\right)$ are their W transforms, we have the Parseval relation

$$\|f - \hat{f}\|_{2,n} = \|\theta - \hat{\theta}\|_{2,n}. \tag{5.11}$$

If (y_i) is the data as in model (5.1) and $w = Wy$ is its discrete wavelet transform then with $\epsilon = \sigma$

$$w_{jk} = \theta_{jk} + \epsilon z_{jk} \qquad \left(j = 0, ..., J; k = 0, ..., 2^j - 1\right).$$

Recall the selective wavelet reconstruction $T_{SW}(y, \delta)$ via (5.3) discussed earlier.

$$T_{SW} = W^T \circ T_{DP} \circ W$$

Because of (5.11)

$$E\|T_{SW}(y, \delta) - f\|_{2,n}^2 = E\|T_{DP}(w, \delta) - \theta\|_{2,n}^2.$$

If $\hat{\theta}^*$ denotes the nonlinear estimator from (5.9) and $\hat{f}^* = W^T \hat{\theta}^* W$ then

$$E\|\hat{f}^* - f\|_{2,n}^2 = E\|\hat{\theta}^* - \theta\|_{2,n}^2.$$

This gives us the following result.

Corollary 28. *For all* f *and all* $n = 2^{J+1}$

$$R\left(\hat{f}^*, f\right) \leq \Lambda_n^* \left\{\frac{\sigma^2}{n} + \mathcal{R}_{n,\sigma}(sw, f)\right\}.$$

Moreover no estimator can satisfy a better inequality then this for all f and all n in the sense that Λ_n^* cannot be replaced by $\{2 - \epsilon + o(1)\} \log n$. A similar inequality holds for the hard thresholding operator.

Lastly we note that piecewise polynomials are not more powerful than wavelets.

Theorem 29. *Let* $D \leq M$ *and* $n = 2^{J+1}$. *Then with constants* C_1 *and* C_2 *depending on the wavelet transform alone,*

$$\mathcal{R}_{n,\sigma}(sw, f) \leq (C_1 + C_2 J)\, \mathcal{R}_{n,\sigma}(PP(D), f)$$

for all f *and all* $\sigma > 0$.

So, the ideal risk of wavelets is better than that of polynomials.

5.1.10. *Variations on the choice of oracle.* An alternative to the keep or kill T_{DP} estimators is given by the diagonal shrinkers

$$\Gamma_{DS}(w,\delta) = (\delta_i w_i)_{i=1}^n \qquad \delta_i \in [0,1].$$

Here different coordinates are shrunk differently. An oracle $\Delta_{DS}(\theta)$ for this family of estimators provides the ideal coefficients $(\delta_i) = \left(\frac{\theta_i^2}{\theta_i^2+\epsilon^2}\right)_{i=1}^n$ and would yield an ideal risk

$$\mathcal{R}_\epsilon(DS,\theta) = \sum_{i=1}^n \frac{\theta_i^2\epsilon^2}{\theta_i^2+\epsilon^2}.$$

This gives us another oracle inequality.

Theorem 30. *The soft thresholding estimator $\hat{\theta}^*$ with threshold λ_n^* satisfies*

$$R\left(\hat{\theta}^*,\theta\right) \le \tilde{\Lambda}_n\left\{\epsilon^2 + \sum_{i=1}^n \frac{\theta_i^2\epsilon^2}{\theta_i^2+\epsilon^2}\right\}$$

for all $\theta \in \mathbb{R}^n$, with $\tilde{\Lambda}_n \sim 2\log n$.

5.1.11. *Proofs of the theorems.*

1.8.1 Proof of Theorem 25.

Proof. We consider the univariate case. Let $X \sim N(\mu,1)$ and $\eta_t(x) = sgn(x)\left(|x|-t\right)_+$. [8] shows that for all $\delta \le \frac{1}{2}$ and with $t = \left(2\log\delta^{-1}\right)^{\frac{1}{2}}$

$$E\left\{\eta_t(x)-\mu\right\}^2 \le \left(2\log\delta^{-1}+1\right)\left(\delta+\mu^2\wedge 1\right). \tag{5.12}$$

This follows directly from the fact that

$$E\left\{\eta_t(x)-\mu\right\}^2 = 1 - 2pr_\mu\left(|x|<t\right) + E_\mu x^2\wedge t^2$$

$$\le 1+t^2 \le \left(2log\delta^{-1}+1\right)(\delta+1)$$

$$E\left\{\eta_t(x)-\mu\right\}^2 \le 2pr_\mu\left(|x|\ge t\right)+\mu^2.$$

We need to verify

$$g(\mu) = 2pr_\mu(|x|\ge t) \le \delta\left(2\log\delta^{-1}+1\right) + \left(2\log\delta^{-1}\right)\mu^2.$$

Since g is symmetric about 0, by examining the Taylor series expansion we know

$$g(\mu) \le g(0) + \frac{1}{2}\left(\sup|g''|\right)\mu^2. \tag{5.13}$$

Using calculus,

$$g(0) = 2pr_0\left(|x| \geq t\right) = 4pr(x > t) = 4\phi(-t) \leq \delta\left(2\log\delta^{-1} + 1\right).$$

$$\sup|g''| \leq 4\sup|x\phi(x)| \leq 4\log\delta^{-1}$$

for $\delta \leq \frac{1}{2}$. Together with (5.13) this gives us (5.12). We see that putting our new information into (5.13) yields

$$\delta\left(2\log\delta^{-1} + 1\right) + \frac{1}{2}4\log d^{-1}\mu^2 + \mu^2 = \delta\left(2\log\delta^{-1}\right) + \delta + \mu^2(2\log\delta^{-1}) + \mu^2$$

$$\leq \left(2\log\delta^{-1} + 1\right)\left(\delta + \mu^2 \wedge 1\right).$$

□

1.8.2 Proof of Theorem 26.

Proof. Suppose we have a single observation $Y \sim N(\mu, 1)$. Define $\rho_{ST}(\lambda, \mu) = E\left\{\eta_S(Y, \lambda) - \mu\right\}^2$. We will list properties of this quantity as they are needed. Recall (5.7) and (5.8). The inequality

$$E\|\hat{\theta}^{\mu} - \theta\|_{2,n}^2 \leq (2\log n + 1)\left\{\varepsilon^2 + \sum_{i=1}^{n}\min\left(\theta_i^2, \varepsilon^2\right)\right\}$$

follows below. Set $\varepsilon = 1$ and $\hat{\theta}_i^* = \eta_{ST}\left(w_i, \lambda_n^*\right)$. Then

$$E\|\hat{\theta}^* - \theta\|_2^2 = \sum_{i=1}^{n}\rho_{ST}\left(\lambda_n^*, \theta_i\right) \leq \sum_{i=1}^{n}\Lambda_n^*\left\{n^{-1} + \min\left(\theta_i^2, 1\right)\right\}$$

$$= \Lambda_n^*\left\{1 + \sum_{i=1}^{n}\min\left(\theta_i^2, 1\right)\right\}.$$

If $\varepsilon \neq 1$ then for $\hat{\theta}_i^* = \eta_{ST}\left(w_i, \lambda_n^*\varepsilon\right)$ we get

$$E\|\hat{\theta}^* - \theta\|_2^2 = \sum\rho_{ST}\left(\lambda_n^*, \frac{\theta_i}{\varepsilon}\right)\varepsilon^2 \leq \Lambda_n^*\varepsilon^2\sum_{i=1}^{n}\left\{n^{-1} + \min\left(\frac{\theta_i^2}{\varepsilon^2}, 1\right)\right\}$$

$$= \Lambda_n^*\varepsilon^2 + \sum_{i=1}^{n}\min\left(\theta_i^2, \varepsilon^2\right).$$

This gives us (5.10). Now we must consider the asymptotics. We wish to analyze

$$\inf_{\lambda}\sup_{\mu}\frac{\rho_{ST}(\lambda, \mu)}{n^{-1} + \min\left(\mu^2, 1\right)}.$$

We consider

$$\Lambda_n^0 = \inf_{\lambda} \sup_{\mu \in \{0,\infty\}} \frac{\rho_{ST}(\lambda,\mu)}{n^{-1} + \min(\mu^2, 1)} \tag{5.14}$$

where λ_n^0 is the largest λ attaining Λ_n^0.

We would like to show that the quantity

$$L\left(\lambda_n^0, \mu\right) = \sup_{\mu} \frac{\rho_{ST}\left(\lambda_n^0, \mu\right)}{n^{-1} + \min(1, \mu^2)}$$

attains it maximum at either $\mu = 0$ or $\mu = \infty$. Note

$$\rho_{ST}(\lambda,\mu) =$$

$$1 + \lambda^2 + \left(\mu^2 - \lambda^2 - 1\right)\{\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)\} - (\lambda - \mu)\phi(\lambda + \mu) - (\lambda + \mu)\phi(\lambda - \mu) \tag{5.15}$$

For $\mu \in [1, \infty]$, $\rho_{ST}\left(\lambda_n^0, \mu\right)$ is monotone increasing. As $\mu \to \infty$, $\rho_{ST}\left(\lambda_n^0, \mu\right) \leq 1 + \lambda^2$. Note that

$$\rho_{ST}(\lambda, 0) = 1 + \lambda^2 - \left(1 + \lambda^2\right)(\Phi(\lambda) - \Phi(-\lambda)) - 2\lambda\phi(\lambda) \tag{5.16}$$

$$\rho_{ST}(\lambda, \infty) = 1 + \lambda^2. \tag{5.17}$$

We define a quantity

$$\rho_n(\lambda) = (n+1)\rho_{ST}(\lambda, 0) - \rho_{ST}(\lambda, \infty). \tag{5.18}$$

As we will show later, this quantity will define our λ_n^0. For now, we take this assumption on faith and note that $\rho_n\left(n^{-\frac{1}{2}}\right) \geq 0$ for n=3. This implies $\lambda_n^0 \geq n^{-\frac{1}{2}}$. Then

$$n\rho_{ST}\left(\lambda_n^0, 0\right) = \frac{\left\{1 + \left(\lambda_n^0\right)^2\right\}}{1 + n^{-1}} \geq \frac{1 + \left(n^{-\frac{1}{2}}\right)^2}{1 + n^{-1}} = 1.$$

Applying (5.19) yields

$$L\left(\lambda_n^0, \mu\right) \leq \frac{\rho_{ST}\left(\lambda_n^0, 0\right) + \mu^2}{n^{-1} + \mu^2} \leq n\rho_{ST}\left(\lambda_n^0, 0\right).$$

Thus L attains its maximum over $\mu \in [0, 1]$ at 0. This establishes what we needed to show.

Now examine (5.17) and (5.18). Note that $\rho_{ST}(\lambda, \infty)$ is increasing in λ and $\rho_{ST}(\lambda, 0)$ is decreasing in λ. This means that

$$L(\lambda, 0) = L(\lambda, \infty)$$

$$\frac{\rho_{ST}(\lambda, 0)}{n^{-1}} = \frac{\rho_{ST}(\lambda, \infty)}{n^{-1} + 1}$$

$$(n+1)\rho_{ST}(\lambda,0)=\rho_{ST}(\lambda,\infty).$$

λ_n^0 is the root of $\rho_n(\lambda)=(n+1)\rho_{ST}(\lambda,0)-\rho_{ST}(\lambda,\infty)$. We can see that this function $\rho_n(\lambda)$ is continuous. It has one zero on $[0,\infty)$ in λ.

$$\rho_n(\lambda)=\left(1+\lambda^2\right)\{2(n+1)\Phi(-\lambda)-1\}-2(n+1)\lambda\phi(\lambda)$$

for $\lambda\geq 0$. Note that if the quantity inside the brackets is negative, then the entire expression is negative on $[\lambda,\infty)$. Apply $\Phi(-\lambda)\leq\lambda^{-1}\phi(\lambda)$. This yields $\lambda=(2\log n)^{\frac{1}{2}}$. Consider

$$2(n+1)\lambda^{-1}\phi(\lambda)-1=2(n+1)\frac{1}{(2\log n)^{\frac{1}{2}}}e^{-\frac{2\log n}{2}}-1=\frac{2(n+1)}{n}\frac{1}{(2\log n)^{\frac{1}{2}}}-1.$$

Note that

$$\frac{2(n+1)}{n(2\log n)^{\frac{1}{2}}}\leq 1$$

when $n\geq 3$. This implies the zero λ_n^0 of ρ_n is less than $(2\log n)^{\frac{1}{2}}$.

Now we define $\lambda_{\eta,n}$ for large n via

$$\lambda_{\eta,n}^2=2\log(n+1)-4\log\log(n+1)-\log 2\pi+\eta.$$

Then $\rho_n\left(\lambda_{\eta,n}\right)$ converges to ∞ or $-\infty$ according to $\eta>0$ or $\eta<0$ respectively. We are concerned with whether

$$2(n+1)\lambda^{-1}\phi(\lambda)\leq 1.$$

$$2(n+1)\frac{1}{(2\log n-4\log\log(n+1)-\log 2\pi+\eta)^{\frac{1}{2}}}e^{-\frac{1}{2}(2\log(n+1)-4\log\log(n+1)-\log 2\pi+\eta)}$$

$$2(n+1)\frac{1}{(2\log n-4\log\log(n+1)-\log 2\pi+\eta)^{\frac{1}{2}}}\left(\frac{1}{n+1}(\log(n+1))^2(2\pi)^{\frac{1}{2}}\right)e^{-\frac{\eta}{2}}$$

$$\frac{2(\log(n+1))^2(2\pi)^{\frac{1}{2}}e^{-\frac{\eta}{2}}}{(2\log n-4\log\log(n+1)-\log 2\pi+\eta)^{\frac{1}{2}}}$$

If $\eta<0$ this quantity is less than 1, and $\rho_n(\lambda_{\eta,n})$ is negative. If $\eta>0$, then $\rho_n(\lambda_{\eta,n})$ is positive.

Lastly, $\rho_{ST}(\lambda_n^0,\infty)=1+(\lambda_n^0)^2$ yields

$$\Lambda_n^0=\frac{\left(\lambda_n^0\right)^2+1}{1+n^{-1}}\sim 2\log n$$

as $n\to\infty$. □

1.8.3 Proof of Theorem 27.

Proof. Let

$$L(\lambda,\mu)=\frac{\rho(\lambda,\mu)}{n^{-1}+\frac{\mu^2}{\mu^2+1}}$$

with ρ either ρ_{ST} or ρ_{HT}. We show

$$L(\lambda,\mu)\le(2\log n)\,(1+\delta_n)$$

uniformly in μ as long as

$$c\log\log n\le\lambda^2-2\log n\le\varepsilon_n\log n.$$

Here $\delta_n\to 0$ and depends on ε_n and c. For ρ_{ST}, $c<5$ and for ρ_{HT}, $c<1$. It has been shown by [2] that

$$\rho(\lambda,\mu)\le\begin{cases}\lambda^2+1 & \mu\in\mathbb{R},\lambda>c_1\\ \mu^2+1 & \mu\in\mathbb{R}\\ \rho(\lambda,0)+c_2\mu^2 & 0<\mu<c_3.\end{cases}\tag{5.19}$$

For soft thresholding $(c_1,c_2,c_3)=(0,1,\infty)$. For hard thresholding $(1,1.2,\infty)$. For $\mu=0$

$$\rho_{ST}(\lambda,0)\le 4\lambda^{-3}\phi(\lambda)\left(1+1.5\lambda^{-2}\right)\tag{5.20}$$

$$\rho_{HT}(\lambda,0)\le 2\phi(\lambda)(\lambda+1)\text{ if }\lambda>1.\tag{5.21}$$

For $\mu\in\left[(2\log n)^{\frac12},\infty\right]$, the numerator is bounded via (5.19) by $1+\lambda^2$.

$$\frac{\rho(\lambda,\mu)}{n^{-1}+\frac{\mu^2}{\mu^2+1}}\le\frac{1+\lambda^2}{\frac{\mu^2}{\mu^2+1}}\le\frac{1+\lambda^2}{\frac{2\log n}{2\log n+1}}\le(2\log n)(1+o(1)).$$

For $\mu\in\left[1,(2\log n)^{\frac12}\right]$ apply (5.19)

$$\frac{\rho(\lambda,\mu)}{n^{-1}+\frac{\mu^2}{\mu^2+1}}\le\frac{\mu^2+1}{\frac{\mu^2}{\mu^2+1}}=\mu^{-2}\left(1+\mu^2\right)^2=\frac{1}{2\log n}(1+2\log n)^2\le(2\log n)(1+o(1)).$$

For $\mu\in[0,1]$ apply (5.19)

$$\frac{\rho(\lambda,\mu)}{n^{-1}+\frac{\mu^2}{\mu^2+1}}=\frac{\rho(\lambda,0)+\rho(\lambda,\mu)-\rho(\lambda,0)}{n^{-1}+\frac{\mu^2}{\mu^2+1}}\le\frac{\rho(\lambda,0)}{n^{-1}}+\frac{\rho(\lambda,\mu)-\rho(\lambda,0)}{\frac{\mu^2}{\mu^2+1}}$$

$$\le n\rho(\lambda,0)+\frac{c_2\mu^2}{\frac{\mu^2}{\mu^2+1}}\le n\rho(\lambda,0)+2c_2.$$

If $\lambda_n(c) = (2\log n - c\log\log n)^{\frac{1}{2}}$ then $n\phi(\lambda_n(c)) = \phi(0)(\log n)^{\frac{c}{2}}$.

$$ne^{-\frac{1}{2}(2\log n - c\log\log n)} = ne^{\log\frac{1}{n}}e^{\log(\log n)^{\frac{c}{2}}} = n\frac{1}{n}(\log n)^{\frac{c}{2}} = (\log n)^{\frac{c}{2}}\cdot 1 = \phi(0)(\log n)^{\frac{c}{2}}.$$

Then by (5.20) and (5.21) $n\rho(\lambda, 0)$ is $o(\log n)$. Thus the theorem is proved. □

1.8.4 Proof of Theorem 29.

Proof. Suppose Δf is the partition supplied by an oracle for piecewise polynomial reconstruction. Suppose this partition has L elements. Let s be the least squares fit using this partition to the noiseless data. Then the ideal risk is the bias squared plus the variance.

$$R\left(T_{PP(D)}(y, \Delta f), f\right) = n^{-1}\|f - s\|_{2,n}^2 + (D+1)L\sigma^2/n. \tag{5.22}$$

Let $\theta = Ws$ be the wavelet transform of s. Then as s is a piecewise polynomial, most of the wavelet coefficients are zero. Let $\delta^* = \{(j,k) : \theta_{j,k} \neq 0\}$, the set of coefficients which do not vanish. Then

$$\#\left(\delta^*\right) \leq \left(C_1 + C_2 J\right) L$$

as in (5.4). Thus

$$R\left(T_{SW}\left(y, \delta^*\right), f\right) \leq n^{-1}\|f - s\|_{2,n}^2 + \#\left(\delta^*\right)\sigma^2/n.$$

Compare this with (5.22) to get

$$R\left(T_{SW}\left(y, \delta^*\right), f\right) \leq \left\{1 + \left(C_1 + C_2 J\right)/(D+1)\right\} R\left(T_{PP(D)}(y, \Delta f), f\right).$$

The theorem follows from the assumption

$$\mathcal{R}_{n,\sigma}(PP(D), f) = R\left(T_{PP(D)}(y, \Delta f), f\right)$$

and the definition

$$\mathcal{R}_{n,\sigma}(SW, f) \leq R\left(T_{SW}\left(y, \delta^*\right), f\right).$$

Now lets consider splines. Let $\tilde{s}$ be the optimal variable knot spline $\tilde{s}$ of order D. Then

$$\|f - s\|^2 \leq \|f - \tilde{s}\|^2.$$

The risk associated with splines depends on L unknown parameters and has variance $\frac{1}{D+1}$ times that of (5.22). Thus

$$\mathcal{R}_{n,\sigma}(PP(D), f) \leq (D+1)\mathcal{R}_{n,\sigma}(spl(D), f)$$

and we have the theorem. □

6. Estimating $f(x)$ where the function is a member of the Besov or Triebel space.

In this last part, we will see that if $f(x)$ is bounded in a certain special way then the risk associated with wavelet shrinkage is within a constant factor of the minimax risk. This is a very powerful result in comparison to the last section, where the difference was a factor of $2\log n$. However, it does limit our choices of $f(x)$ to the spaces mentioned.

6.1. Estimating $f(x)$ where the function is a member of the Besov or Triebel space, details.

In the first Donoho and Johnstone paper we examined, we only used the data and the assumption that the noise was normal to derive a bound dependent on n for the ideal risks. With just one additional assumption about the function underlying the data, namely that the function is a member of either the Besov or Triebel spaces, one can fine a bound which is not dependent on n, and is in fact a real number.

Suppose we have n samples of a function f.

$$y_i = f(t_i) + z_i \tag{6.1}$$

where $i = 1, ..., n$, $t_i = \frac{i}{n}$, and the z_i are identically independent distributed as $N\left(0, \sigma^2\right)$. We want to estimate f depending on $y_1, ..., y_n$ with risk $R\left(\hat{f}, f\right) = E\|\hat{f} - f\|_2^2 = E\int_0^1 \left(\hat{f}(t) - f(t)\right)^2$.

We begin by assuming f is a member of the Besov space. This represents a very large collection of spaces, for example, the Bump Algebra and the L^2Sobolev Space. The method derived in this paper [9] performs better that any linear method. Below is the important result of the paper.

Corollary 31. *Let* $\mathcal{F}$ be a ball in the Besov space $B^{\sigma}_{p,q}$ with $\sigma > \frac{1}{p}$ and $1 \leq p, q \leq \infty$. Let $R(n, \mathcal{F})$ denote the minimax risk from (6.1) and let $R_L(n, \mathcal{F})$ denote minimax risk when estimators are linear in data (y_i). Then as $n \to \infty$

$$R(n, \mathcal{F}) \sim n^{-r}$$

$$R_L(n, \mathcal{F}) \sim n^{-r'}$$

with

$$r = \frac{2\sigma}{2\sigma + 1} \text{ and } r' = \frac{\sigma + \left(\frac{1}{\hat{p}} - \frac{1}{p}\right)}{\sigma + \frac{1}{2} + \left(\frac{1}{\hat{p}} - \frac{1}{p}\right)}$$

where $\hat{p} = \max(p, 2)$.

We can see that if $p < 2$, the minimax risk approaches 0 faster than the linear estimator. If $p \geq 2$ the performance is the same.

6.1.1. *Some Notations.* Consider the interval $[0,1]$. Define the dyadic subintervals

$$I_{j,k} = \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]$$

for $j \geq 0$ and $k = 0, ..., 2^j - 1$. Let $\mathcal{I}$ denote the collection of all $I_{j,k}$ and I_j denote the collection of all 2^j intervals with length 2^{-j}. Individual subintervals may be denoted by I, I' or $I_{j,k}$. Then the wavelet ψ is denoted

$$\psi_I(t) = 2^{\frac{j}{2}}\psi\left(2^j t - k\right)$$

and ψ_I is supported in $I = \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]$. The scaling function is written $\phi_I(t)$. Suppose $f \in L^2[0,1]$. Then

$$\beta_{l,k} = \int_0^1 f(t)\phi_I(t)dt$$

$$\alpha_I = \int_0^1 f(t)\psi_I(t)dt$$

$$f = \sum_{k\in K} \beta_{l,k}\phi_{l,k} + \sum_{I\in\mathcal{J}} \alpha_I\psi_I.$$

Here K is all k from $-\infty$ to ∞. Here $\mathcal{J}$ is the collection of all dyadic intervals of length $|I| < 2^{-l}$. By Parseval's relation

$$\|\hat{f} - f\|^2_{L^2[0,1]} = \sum_{k\in K}\left(\hat{\beta}_{l,k} - \beta_{l,k}\right)^2 + \sum_{I\in\mathcal{J}}\left(\hat{\alpha}_I - \alpha\right)^2.$$

The functions $\phi_{l,k}$ describe the gross structure behavior of f and the functions ψ_I describe the details localized to I.

We say that such a wavelet analysis has a regularity of r if the functions in the analysis are of compact support and have r continuous derivatives. The coefficients of a regular wavelet analysis can measure function smoothness very precisely if $r > 1$.

From [16] we know that if f is locally Holderian at x_0 with exponent δ then $\alpha_I = o\left(2^{-\left(\frac{1}{2}+\delta\right)j}\right)$ for every sequence (I) with $|I| \to 0$ for $x_0 \in I$ From [19] we have that if f is differentiable at x_0 then $\alpha_I = O\left(2^{-\frac{3}{2}j}\right)$ for every sequence (I) with $|I| \to 0$ for $x_0 \in I$. These results have near converses. This motivates our study of the wavelet coefficients.

Now we wish to define the Besov seminorm in terms of wavelet coefficients. Define the rth difference

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

Define the rth modulus of smoothness as

$$w_{r,p}(f;h) = \|\Delta_n^{(r)} f\|_{L^p[0,1-rh]}.$$

Then the Besov seminorm of index (σ, p, q) defined for $r > \sigma$

$$|f|_{B^\sigma_{p,q}} = \left(\int_0^1 \left(\frac{w_{r,p}(f;h)}{h^\sigma} \right)^q \frac{dh}{h} \right)^{\frac{1}{q}}$$

if $q < \infty$ and

$$|f|_{B^\sigma_{p,q}} = \sup_{0<h<1} \frac{w_{r,p}(f;h)}{h^\sigma}$$

if $q = \infty$. Then a Besov space $B^\sigma_{p,q}$ is the set of all functions $f : [0,1] \to \mathbb{R}$ with $f \in L^p$, that is,

$$\left(\int_0^1 |f(x)|^p dx \right)^{\frac{1}{p}} < \infty$$

and $|f|_{B^\sigma_{p,q}} < \infty$. Define the following norm in terms of the wavelet coefficients.

$$|\alpha|_{\tilde{b}^s_{p,q}} = \left(\sum_{j \geq l}^{\infty} \left(2^{js} \left(\sum_{I_j} |\alpha_I|^p \right)^{\frac{1}{p}} \right)^q \right)^{\frac{1}{q}}$$

Theorem 32. *From* [20] *we have for* $\alpha = \alpha(f)$ *and* $\beta = \beta(f)$

$$\left(\|f\|_p + \|f\|_{B^\sigma_{p,q}} \right) \fallingdotseq \left(\|\beta_k\|_{l^p} + |\alpha|_{\tilde{b}^s_{p,q}} \right)$$

for every $f \in L^p[0,1]$ *where* $s = \sigma + \frac{1}{2} - \frac{1}{p}$. *Here the* $\fallingdotseq$ *means that the ratios of the two sides are bounded by a* c *and a* C *depending on* $(\psi, \phi, p, q, r, \sigma)$ *but not on* f.

So by using a wavelet analysis we have a transformation from the continuous function space to a sequence space with two important properties.

(1) If $\hat{f}$ and f are two functions,

$$\|\hat{f} - f\|^2 = \sum_{k \in K} \left(\hat{\beta}_k - \beta_k \right)^2 + \sum_{j \geq l} \sum_{I_j} \left(\hat{\alpha}_I - \alpha_I \right)^2.$$

(2) Let Θ denote the collection of all wavelet expansions $\left((\beta_k)_{k \in K}, (\alpha_I)_{I \in \mathcal{J}} \right)$ of functions in the ball $\mathcal{F}$ defined by $\|f\|_{B^\sigma_{p,q}} \leq 1$. Let $\Theta_0 = \mathbb{R}^{\#(K)}$ and let $\tilde{\Theta}^s_{p,q}(C)$ denote the collection of $(\alpha_I)_{I \in \mathcal{J}}$ satisfying $|\alpha|_{\tilde{b}^s_{p,q}} \leq C$. Then for positive constants c and C,

$$\{0\} \times \tilde{\Theta}^s_{p,q}(c) \subset \Theta \subset \Theta_0 \times \tilde{\Theta}^s_{p,q}(C).$$

6.1.2. *Estimation in the Sequence Space.* Our data is of the form

$$y_I = \theta_I + z_I \tag{6.2}$$

for $I \in \mathcal{I}$. Here the z_I are identically independent and distributed as $N\left(0, \sigma^2\right)$. The unknown quantity $\theta = (\theta_I)$ is the one we wish to estimate. We assume that $\|\theta\|_{b^s_{p,q}} \leq C$ where

$$\|\theta\|_{b^s_{p,q}} = \left(\sum_{j \geq 0} \left(2^{js} \left(\sum_{I_j} |\theta_I|^p \right)^{\frac{1}{p}} \right)^q \right)^{\frac{1}{q}}.$$

Denote this convex set as $\Theta^s_{p,q}(C) = \left\{ \theta : \|\theta\|_{b^s_{p,q}} \leq C \right\}$ We call this set a Besov Body and often abbreviate it as $\Theta^s_{p,q} = \Theta^s_{p,q}(C)$.

Define the minimax risk

$$R^*\left(\epsilon, \Theta^s_{p,q}\right) = \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}} E\|\hat{\theta} - \theta\|_2^2.$$

Define the minimax linear risk

$$R^*_L\left(\epsilon, \Theta^s_{p,q}\right) = \inf_{\hat{\theta} \text{ linear}} \sup_{\Theta^s_{p,q}} E\|\hat{\theta} - \theta\|_2^2.$$

The '$*$' always denotes minimax risk.

We can connect estimation in the sequence space with the regression model (6.1). If $\mathcal{F}$ is a class of functions on the interval and Θ is the set in the sequence space of the wavelets coefficients of functions in $\mathcal{F}$, (1) and (2) have the following consequences.

(1) The minimax risk from sampled data is asymptotically equivalent to the risk in the sequence space.

$$R(n, \mathcal{F}) \sim R^*\left(\sigma/\sqrt{n}, \Theta\right), \ n \to \infty$$

$$R_L(n, \mathcal{F}) \sim R^*_L(\sigma/\sqrt{n}, \Theta), \ n \to \infty$$

(2) If $\mathcal{F}$ is a Besov class, the body Θ is risk-equivalent to a Besov Body.

$$R^*(\epsilon, \Theta) \eqsim R^*\left(\epsilon, \Theta^s_{p,q}\right), \ \epsilon \to 0$$

$$R^*_L(\epsilon, \Theta) \eqsim R^*_L(\epsilon, \Theta^s_{p,q}), \ \epsilon \to 0$$

Thus, an understanding of minimax estimation in the sequence model will allow us to understand the function model.

6.1.3. *Minimax Bayes Estimation.* To solve the minimax problem for a Besov Body, we relax the constraint of boundedness that we have imposed so far and consider a weaker condition which will allow us to analyze the risk differently. We assume our observed (θ_I) are random variables which are arbitrary except for the constraint that

$$\|\tau\|_{b^s_{p,q}} \leq C$$

where τ is a moment sequence defined by

$$\tau_I = (E|\theta_I|^{p\wedge q})^{\frac{1}{p\wedge q}}$$

for $I \in \mathcal{I}$. (Note that if $p \wedge q = \infty$ we put $\tau_I = \sup |\theta_I|$.) Define the minimax Bayes risk

$$\mathcal{B}^* \left(\epsilon; \Theta^s_{p,q}\right) = \inf_{\hat{\theta}} \sup_{\tau \in \Theta^s_{p,q}} E\|\hat{\theta} - \theta\|^2. \tag{6.3}$$

Since this constraint is weaker than $\|\theta\|_{b^s_{p,q}} \leq C$, $\mathcal{B}^* \geq \mathcal{R}^*$.

We present the main results of the paper. First, we see that the minimax estimator for $\mathcal{B}^*$ are separable nonlinearities.

Theorem 33. *A minimax estimator for $\mathcal{B}^*(\epsilon)$ has the form*

$$\hat{\theta}^*_I = \delta^*_j (y_I)$$

for $I \in \mathcal{I}$, where $\delta^*_j(y)$ is a scalar nonlinear function of the scalar y. In fact there is a 3-parameter family $\delta_{(\tau,\epsilon,p)}$ of nonlinear functions of y from which the minimax estimator is built:

$$\delta^*_j = \delta_{\left(t^*_j, \epsilon, p\wedge q\right)}$$

for $j = 0, 1, ...$ for a sequence $\left(t^*_j\right)_{j=0}^{\infty}$ which depends on s, p, q, C and ϵ.

Next we have a result about the asymptotics of $\mathcal{B}^*$. Below $p \wedge q$ means $\min\{p, q\}$.

Theorem 34. *Let $p, q > 0$ and $s + 1/p > 1/(2 \wedge p \wedge q)$; then $\mathcal{B}^*(\epsilon) < \infty$ and as $\epsilon \to 0$*

$$\mathcal{B}^* \left(\epsilon, \Theta^s_{p,q}\right) \sim \gamma(\epsilon) C^{2(1-r)} \epsilon^{2r},$$

where

$$r = \frac{s + \frac{1}{p} - \frac{1}{2}}{s + \frac{1}{p}},$$

and $\gamma(\epsilon) = \gamma(\epsilon; C, s + 1/p, p \wedge q, q)$ is a continuous, periodic function of $\log_2(\epsilon/C)$ defined later on.

Last, we demonstrate the asymptotic equivalence of R^* and $\mathcal{B}^*$.

Theorem 35. *For $s + 1/p > 1/2$, $s, p, q > 0$ and $\epsilon > 0$*

$$R^*\left(\epsilon; \Theta^s_{p,q}\right) \geq \tilde{\gamma}(\epsilon) C^{2(1-r)} \epsilon^{2r} - \epsilon^2 \tag{6.4}$$

where r is as above, and $\gamma(\epsilon) = \gamma(\epsilon; C, s + 1/p, \infty, q)$ is a continuous, periodic function of $\log_2(\epsilon/C)$. If $q \geq p$, then

$$R^*\left(\epsilon; \Theta^s_{p,q}\right) = \mathcal{B}^*(\epsilon)\left(1 + o(1)\right), \tag{6.5}$$

$\epsilon \to 0$.

Combining these theorems together we have for $p \leq q$ that the estimator $\hat{\theta}^*$ which will be defined later is asymptotically minimax with respect to R^* for $\epsilon \to 0$. If $p > q$, the Bayes-Minimax estimator is within a constant factor of minimax.

6.1.4. *Properties of the Risk Function of a Single Variable.* Let us briefly consider the scalar problem. Observe

$$v = \xi + z. \tag{6.6}$$

Here ξ is a random variable and z is independent of ξ and distributed as $N\left(0, \epsilon^2\right)$. We assume that $(E_\pi |\xi|^p)^{\frac{1}{p}} \leq \tau$. We estimate ξ relative to square-error loss. Define the minimax Bayes risk

$$\rho_p(\tau, \epsilon) = \inf_{\delta} \sup_{(E_\pi |\xi|^p)^{\frac{1}{p}} \leq \tau} E_\pi E_\xi \left(\delta(y) - \xi\right)^2. \tag{6.7}$$

We report several properties of this quantity analyzed in [7].

$$\rho_p(\tau, \epsilon) = \epsilon^2 \rho_p(\tau/\epsilon, 1) \tag{6.8}$$

$$\rho_p(a\tau, \epsilon) \leq a^2 \rho_p(\tau, \epsilon), \ a > 1 \tag{6.9}$$

$$\rho_p(\tau, 1) \sim \begin{cases} \tau^2 & p \geq 2 \\ \tau^p \left(2 \log\left(\tau^{-p}\right)\right)^{\frac{2-p}{2}} & p < 2 \end{cases} \tag{6.10}$$

as $\tau \to 0$. The function ρ_p is continuous, is monotone increasing in τ, is concave in τ^p and has $\rho_p(\tau, \epsilon) \to \epsilon^2$ as $\tau\epsilon \to \infty$.

6.1.5. *Properties of the Besov Space, Separable Rules are Minimax.* Here are two structural facts about Besov Bodies which allow us eventually to find a prior distribution and begin to analyze the minimax Bayes problem.

(1) For $q < \infty$, $J^s_{p,q}(\tau) = \|\tau\|^q_{b^s_{p,q}}$ is a convex functional of the moment sequence $\tau = \left(\tau_I^{p\wedge q}\right)$. For $q = \infty$, the functional $J^s_{p,\infty}(\tau) = \|\tau\|_{b^s_{p,\infty}}$ has nested convex level sets. (Recall a convex function has the property that for any $t \in [0,1]$, $f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$.)

(2) If (τ_I) is an arbitrary positive sequence, and we set $\bar{\tau}_I^{p\wedge q} = Ave_{I\in I_j}\left(\tau_I^{p\wedge q}\right)$, then

$$\|\bar{\tau}\|_{b^s_{p,q}} \le \|\tau\|_{b^s_{p,q}}. \tag{6.11}$$

We prove Theorem 33 by working out the statistical implications of these facts.

Proof. Proof of Theorem 33.

Let $\mathcal{M}^s_{p,q} = \left\{\mu : J^s_{p,q}(\tau(\mu)) \le C^q\right\}$ denote the set of prior measures μ which would be feasible in (6.3). By (1), $\mathcal{M}^s_{p,q}$ is convex. It is also weakly compact for weak convergence of probability measures; and the l^2 loss yields lower-semicontinuous risk functions. Thus we can apply the Minimax Theorem of Statistical Decision Theory from [17]. This theorem implies that the Bayes rule of a least favorable prior is a minimax rule. This least favorable prior is what we hope to find.

Let $\beta(\mu)$ denote the Bayes risk for estimating (θ_I) with squared l^2 loss from data (6.2). If μ^* is least favorable then by definition

$$\beta(\mu^*) = \sup\left\{\beta(\mu) : \mu \in \mathcal{M}^s_{p,q}\right\}. \tag{6.12}$$

We will now use (2) to show that a least favorable distribution makes the coordinates of (θ_I) independent. Suppose μ is an arbitrary distribution for the vector (θ_I). Let μ_I denote the prior distribution of each scalar component θ_I. From this prior we will derive another prior $\bar{\mu}$ which makes the coordinates (θ_I) independent random variables. Set $\bar{\mu}_j = Ave_{I_j}(\mu_I)$. Then this quantity is the average of all the μ_I on one level j. This prior makes the θ_I identically independent within one resolution level, with j fixed.

This prior $\bar{\mu}$ is less favorable than μ. The Bayes risk of μ is the sum of coordinate-wise risks:

$$\beta(\mu) = \sum_{I\in\mathcal{I}} E_\mu\left(E\left(\theta_I|\left(y_{I'}\right)_{I'\in\mathcal{I}}\right) - \theta_I\right)^2. \tag{6.13}$$

Note that $(y_{I'})_{I'\in\mathcal{I}}$ gives us more information about θ_I than just y_I. Thus

$$E_\mu\left(E\left(\theta_I|\left(y_{I'}\right)_{I'\in\mathcal{I}}\right) - \theta_I\right)^2 \le E_\mu\left(E\left(\theta_I|y_I\right) - \theta_I\right)^2.$$

Let $b(\pi)$ denote the Bayes risk in the scalar problem of estimating ξ from data. Here $v = \xi + z$ with $z \sim N\left(0,\sigma^2\right)$ and $\xi \sim \pi$. Then the right side of (6.13) is $b(\mu_I)$ and

$$B(\mu) \le \sum_{I\in\mathcal{I}} b(\mu_I). \tag{6.14}$$

We can examine (6.7) and consider it as a function of the distribution π to confirm that this scalar Bayes risk is concave. This means

$$Ave_{I\in I_j}\left(b\left(\mu_I\right)\right) \leq b\left(Ave_{I\in I_j}\left(\mu_I\right)\right).$$

We conclude

$$\beta(\mu) \leq \sum_j 2^j b\left(\bar{\mu}_j\right) = \beta(\bar{\mu}).$$

This means exactly that $\bar{\mu}$ is less favorable than μ because its Bayes risk is bigger. Lastly, we note that the moment sequence of $\bar{\mu}$ is given by:

$$E_{\bar{\mu}_j}|\theta_{I_{j,k}}|^{p\wedge q} = Ave_{I\in I_j}\left(E_{\mu_I}|\theta_I|^{p\wedge q}\right) = Ave_{I\in I_j}\left(\tau_I^{p\wedge q}\right) = \bar{\tau}_I^{p\wedge q}.$$

Hence we can apply (6.11) and

$$\mu \in \mathcal{M}_{p,q}^s \implies \bar{\mu} \in \mathcal{M}_{p,q}^s.$$

So from any given candidate μ for a least favorable prior we derive $\bar{\mu}$ which is less favorable but still satisfies (6.12). In short, (2) implies we may find a least favorable measure within the subclass of measures having coordinates which are identically independent within each resolution level.

For any given prior π on the scalar ξ obeying $E_\pi|\xi|^{p\wedge q} \leq \tau^{p\wedge q}$ we have by (6.7) (that is, the fact that this risk is minimax)

$$b(\pi) \leq \rho_{p\wedge q}(\tau, \epsilon).$$

Then by (6.14)

$$\beta(\mu) \leq \sum_I \rho_{p\wedge q}\left(\tau_I, \epsilon\right). \tag{6.15}$$

This means that no prior in $\mathcal{M}_{p,q}^s$ can obtain a larger Bayes risk than

$$\sup\left\{\sum_I \rho_{p\wedge q}\left(\tau_I, \epsilon\right)\right\} \text{ subject to } \tau \in \Theta_{p,q}^s. \tag{6.16}$$

We show later that this supremum is finite when $s + p^{-1} > (2\wedge p\wedge q)^{-1}$ or when $p = q = 2$, $s = 0$; the supremum is attained by a sequence called τ^*. We have equality in (6.15) if the prior on coordinate I is chosen to be least-favorable for $\rho_{p\wedge q}\left(\tau_I, \epsilon\right)$. Choosing coordinate priors in this way from the sequence τ^* yields a sequence prior μ^*which is least favorable.

The Bayes rule for u^* is

$$\hat{\theta}_I^* = \delta_{\left(\tau_I^*, \epsilon, p\wedge q\right)}\left(y_I\right),\ I \in \mathcal{I}.$$

These τ_I^* are all equal within one resolution level by construction and Theorem 26 is proved. □

6.1.6. *A Dyadic Renormalization.*

Proof. Proof of Theorem 34.

We now need to prove 34. By formula (6.16) we have $\mathcal{B}^*\left(\epsilon, \Theta^s_{p,q}\right) = val\left(P_{\epsilon,C}\right)$ where we define $(P_{\epsilon,C})$ to be the optimization problem

$$(P_{\epsilon,C}) \;\; \sup \sum_{j=0}^{\infty} 2^j \rho\left(t_j, \epsilon\right) \;\; \text{subject to} \;\; \sum_{j=0}^{\infty} \left(2^{sj}\left(2^j t_j^p\right)^{\frac{1}{p}}\right)^q,$$

with just boundedness if $p = \infty$ or $q = \infty$. Here $\rho = \rho_{p \wedge q}$.

Define the optimization problem $(Q_{\epsilon,C})$ on the space of bilateral sequences $T = \left\{(t_j)_{j=-\infty}^{j=\infty}\right\}$

$$(Q_{\epsilon,C}) \;\; \sup \sum_{j=-\infty}^{\infty} 2^j \rho\left(t_j, \epsilon\right) \;\; \text{subject to} \;\; \sum_{j=-\infty}^{\infty} \left(2^{\beta j} t_j\right)^q \leq C^q. \tag{6.17}$$

Set $\beta = s + \frac{1}{p}$. We can see that this problem is very similar to $(P_{\epsilon,C})$. If a unilateral sequence $(t_j)_{j=0}^{\infty}$ is solution for the unilateral problem $(P_{\epsilon,C})$ then the extension to a sequence which is a solution to $(Q_{\epsilon,C})$, call it $\tilde{t}_j$ is defined by setting $\tilde{t}_j = 0$ for $j < 0$ and $\tilde{t}_j = t_j$ for $j > 0$. We conclude that

$$val\left(P_{\epsilon,C}\right) \leq val\left(Q_{\epsilon,C}\right)$$

for all $\epsilon > 0$ and $C > 0$. On the other hand, if a sequence (t_j) is a solution to $(Q_{\epsilon,C})$, then the sequence $\tilde{t}_j$ formed by dropping the $j < 0$ portion from (t_j) is a solution to $(P_{\epsilon,C})$. Moreover, the part of the objective function lost in dropping the negative indexes is bounded by ϵ^2. This is because $\rho_p\left(t_j, \epsilon\right) \leq \epsilon^2$ implies $\sum_{j<0} 2^j \rho\left(t_j, \epsilon\right) \leq \sum_{j<0} 2^j \epsilon^2 = 1 \cdot \epsilon^2 = \epsilon^2$. Hence

$$val\left(Q_{\epsilon,C}\right) \leq val\left(P_{\epsilon,C}\right) + \epsilon^2$$

for all $\epsilon > 0$ and $C > 0$. This ϵ^2 is asymptotically negligible. Thus $val\left(P_{\epsilon,C}\right) \sim val\left(Q_{\epsilon,C}\right)$ as $\epsilon \to 0$. We must show the following to prove the theorem. □

Fact 36. *If $\beta > 1/(2 \wedge p \wedge q)$ then*

$$val\left(Q_{\epsilon,C}\right) = \gamma(\epsilon, C) C^{2(1-r)} \epsilon^{2r} \tag{6.18}$$

for $\epsilon > 0$, where $r = \frac{\beta - 1/2}{\beta} > 0$ and $\gamma(\epsilon, C)$ is a continuous, periodic function of $\log_2(\epsilon, C)$.

We set

$$J_{\rho,\epsilon}(t) = \epsilon^2 \sum_{-\infty}^{\infty} 2^j \rho\left(t_j/\epsilon, 1\right)$$

and

$$J_{\rho,\beta}(t) = \left(\sum_{-\infty}^{\infty} 2^{j\beta q} t_j^q \right)^{\frac{1}{q}}.$$

Because of the properties of the risk (6.8), we have

$$val\left(Q_{\epsilon,C}\right) = \sup J_{\rho,\epsilon}(t) \text{ subject to } J_{q,\beta}(t) \leq C.$$

Let $\left(\mathcal{U}_{a,h}t\right)_j = at_{j-h}$. Note that

$$J_{\rho,\epsilon}\left(\mathcal{U}_{\epsilon,h}t\right) = \epsilon^2 2^h J_{\rho,1}(t),$$

and also,

$$J_{q,\beta}\left(\mathcal{U}_{\epsilon,h}t\right) = \epsilon 2^{\beta h} J_{q,\beta}(t).$$

These scaling relations imply that if $\epsilon = \epsilon_h = 2^{-\beta h}$ for h an integer, and if (t_j) is a solution to the noise-level 1 problem $(Q_{1,C})$, then the sequence $\tilde{t} = \mathcal{U}_{\epsilon,h}t$ is a solution to the noise-level ϵ problem $(Q_{\epsilon,C})$. Furthermore,

$$val\left(Q_{\epsilon_h,C}\right) = J_{\rho,\epsilon}\left(\tilde{t}\right) = \epsilon_h^2 2^h J_{\rho,1}(t) = \left(\epsilon_h^2\right)^r val\left(Q_{1,C}\right).$$

This is because $\epsilon_h^2 2^h = \left(\epsilon_h^2\right)^r$. Recall $r = \frac{\beta - 1/2}{\beta}$, then

$$\left(\epsilon_h^2\right)^r = \left(2^{-2\beta h}\right)^{\frac{\beta-1/2}{\beta}} = \left(2^{-2(\beta-1/2)h}\right) = \left(2^{-2\beta h}2^h\right) = \epsilon_h^2 2^h.$$

More generally for any $\epsilon > 0$ and h an integer,

$$val\left(Q_{\epsilon,C}\right) = \epsilon^2 2^h val\left(Q_{1,\frac{C}{\epsilon}2^{-\beta h}}\right). \tag{6.19}$$

Choose the integer $h = h(\epsilon, C)$ so that $\frac{C}{\epsilon}2^{-\beta h}$ exceeds 1 by as little as possible:

$$\frac{C}{\epsilon}2^{-\beta h} = 2^{\beta\eta} \in \left[1, 2^\beta\right).$$

Solving this for h one gets $h = \left\lfloor \beta^{-1}\log_2(C/\epsilon) \right\rfloor$, where η is the corresponding fractional part and $\epsilon^2 2^h = \epsilon^2 (C/\epsilon)^{\beta^{-1}} 2^{-1}$. Combining this with (6.19) and noting that $2 - \beta^{-1} = 2r$ yields

$$val\left(Q_{\epsilon,C}\right) = \epsilon^{2r} C^{2(1-r)} 2^{-\eta} val\left(Q_{1,2^{\eta\beta}}\right).$$

Now (6.19) shows that $2^{-x} val\left(Q_{1,2^{\beta x}}\right) = 2^{k-x} val\left(Q_{1,2^{\beta(x-k)}}\right)$ for each integer k, so

$$\gamma(\epsilon; C, \beta, p, q) = 2^{-\eta(\epsilon,C)} val\left(Q_{1,2^{\eta\beta}}\right)$$

is a periodic function of η and hence of $\log_2(C/\epsilon)$ for fixed β. To prove Theorem (34) we now only need continuity. Finiteness and continuity follow from the following fact.

Fact 37. *Let T_C denote the class of bilateral sequences (t_j) such that $J_{q,\beta}(t) \leq C$. If $\beta \cdot (2 \wedge p \wedge q) > 1$, then the class of sequences $\left\{\left(2^j \rho(t_j)\right) : t \in T_C\right\}$ is a compact subset of l_1; the maximum $\sum_{-\infty}^{\infty} 2^j \rho(t_j)$ over $t \in T_C$ is finite, and the maximum is attained by some $t \in T_C$. The maximum value of $J_{1,\rho}$ over T_C is continuous in C.*

This is proved by using the asymptotics (6.9) and (6.10). Because $J_{q,\beta}(t) \leq C$, the sum $\sum_{-\infty}^{\infty} 2^j \rho(t_j)$ is bounded in l_1. Let's examine the case where $p \geq 2$. Consider

$$\sum_{-\infty}^{\infty} 2^j \rho(t_j) \sim \sum_{-\infty}^{\infty} 2^j t_j^2 = \sum_{-\infty}^{\infty} \left|2^j t_j^2\right| \leq \sum_{-\infty}^{\infty} 2^{j\beta q} t_j^q \leq C^q.$$

Clearly the set of sequences is closed and totally bounded. The maximum of this sequence is bounded by C^q. This quantity is continuous in C.

6.1.7. *Asymptotic Equivalence.* Now we prove Theorem (35). Firstly, note that we have not considered that case where either $p = \infty$ or $q = \infty$. By the definition of minimax risk, $R^*\left(\epsilon; \Theta_{p,q}^s\right)$ is the supremum of Bayes risks for the priors supported in $\Theta_{p,q}^s$. Let $\tau \in \Theta_{p,q}^s$. Consider the prior defined by letting each coordinate of the prior be the one which attains the minimax risk $\rho_\infty(\tau_I, \epsilon)$ in the scalar bounded normal mean problem. This prior is supported in $\Theta_{p,q}^s$ and has Bayes risk $\sum_I \rho_\infty(\tau_I, \epsilon)$. This coordinate-wise minimax risk is a lower bound on the minimax risk. We find the best bound of this form by solving the optimization problem

$$\sup\left\{\sum_I \rho_\infty(\tau_I, \epsilon) : \tau \in \Theta_{p,q}^s\right\}.$$

This problem is the same as (6.16) except we have ∞ instead of $p \wedge q$. Thus we have the risk bound of Theorem (35), equation (6.4) by the same arguments as last section. We must now prove (6.5). This assertion is the same as saying that there exist priors which are almost least favorable for the enlarged minimax problem. The phrase 'almost least favorable' is embodied in (6.20). We will show that for each $\eta > 0$ we may construct a sequence of priors $v^{(h)}$, $h = 1, 2, ...$ such that along specially dyadically generated sequences with

$$\epsilon_h = 2^{-h(s+1/p)},$$

$h = 1, 2, ...$ we have for large enough h that

$$B\left(v^{(h)}\right) \geq \mathcal{B}^*(\epsilon_h; C)(1 - \eta). \tag{6.20}$$

These priors will be supported in $\Theta^s_{p,q}(C\cdot(1+\eta))$. We can conclude that

$$R^*(\epsilon_h; C\cdot(1+\eta)) \geq \mathcal{B}^*(\epsilon_h; C)(1-\eta),$$

as $h\to\infty$. Recall that in the requirements of the theorem $s+1/p>1/2$. Then since $r=\frac{s+1/p-1/2}{s+1/p}$, $r>0$ always. Also, $r<1$. Then the quantity $r(1-r)>0$, and $(1+\eta)^{r(1-r)}>1$. Then as $h\to\infty$ note

$$(1+\eta)^{r(1-r)}\mathcal{R}^*(\epsilon_h, C) \sim (1+\eta)^{r(1-r)}\left(\tilde{\gamma}(\epsilon_h)C^{r(1-r)}\epsilon_h^2 - \epsilon_h^2\right)$$

$$\geq \tilde{\gamma}(\epsilon_h)C^{r(1-r)}(1+\eta)^{r(1-r)}\epsilon_h^2 - \epsilon_h^2$$

$$= R^*(\epsilon_h, C(1+\eta)) \geq \mathcal{B}^*(\epsilon_h, C)(1-\eta) \sim \gamma(\epsilon_h)C^{r(1-r)}\epsilon_h^{2r}(1-\eta).$$

Divide this inequality on both sides by $(1+\eta)^{r(1-r)}$ to get

$$R^*(\epsilon_h, C) \geq \mathcal{B}^*(\epsilon_h, C)\frac{1-\eta}{(1+\eta)^{r(1-r)}}$$

or

$$R^*(\epsilon_h, C) \geq \mathcal{B}^*(\epsilon_h, C)(1+o(1)).$$

The argument for dyadic sequences $c\cdot 2^{-h(s+1/p)}$ where $c\neq 1$ is similar to the above argument. Thus we have Theorem 35. We know that

$$val(Q_{\epsilon_h,C}) = val(Q_{1,C})\left(\epsilon_h^2\right)^r.$$

Consider $(Q_{1,C})$. The last section implicitly defined a countable sequence of prior distributions $\bar{\mu}_j$ which satisfies $\sum_{-\infty}^{\infty}2^j b_1(\bar{\mu}_j) = val(Q_{1,C})$, where b_1 stood for the Bayes risk in the $\epsilon=1$ scalar problem $v=\xi+z$ with z standard normal. We can renormalize the constant which attains $(Q_{\epsilon_h,C})$ for $h=1,2,....$

We must now establish the earlier assertions of the section. For $\eta>0$ we can find a near solution to $(Q_{1,C})$ with certain additional support properties. We can find finite positive integers J and M so that

(1) For $-J\leq j\leq J$ there is a prior distribution μ_j for a scalar random variable ξ.

(2) Each μ_j is supported in $[-M,M]$.

(3) The moment sequence $t_j^{p\wedge q} = E_{\mu_j}|\xi|^{p\wedge q}$ obeys $\sum_{-J}^{J}2^{j\beta q}t_j^q \leq C^q$.

(4) The coordinate-wise Bayes risks obey $\sum_{-J}^{J}2^j b_1(\mu_j) \geq val(Q_{1,C})\cdot(1-\eta)$.

Define for $-J\leq j\leq J$ an infinite sequence of random variables $(X_{j,k})_{k=0}^{\infty}$ with $X_{j,k}$ identically independent distributed as μ_j. Suppose $h>J$ and define random variables (θ_I) by

$$\theta_I = \epsilon_h\cdot X_{j,k},$$

with $I \in I_{j+h}$, for $-J \leq j \leq J$, and $\theta_I = 0$ otherwise. Denote $\mu^{(h)}$ as the distribution of the sequence (θ_I) just defined.

When estimating (θ_I) from the sequence data, the independence of θ_I and z_I makes the Bayes risk add coordinate-wise.

$$B\left(\mu^{(h)}\right) = \epsilon_h^2 \sum_{-J}^{J} 2^{j+h} b_1\left(\mu_j\right) = \left(\epsilon_h^2\right)^r \sum_{-J}^{J} 2^j b_1\left(\mu_j\right) \geq \left(\epsilon_h^2\right)^r \cdot val\left(Q_{1,C}\right)(1-\eta) \tag{6.21}$$

Here we have applied (4). This prior above is then almost least favorable. Also, this prior is almost supported in $\theta_{p,q}^s\left(C \cdot (1+\eta)\right)$ in the following sense.

Fact 38. *Define the event*

$$A_\eta = \left\{ \|\theta\|_{b_{p,q}^s} \leq C\left(1+\eta\right) \right\}.$$

Then

$$\mu^{(h)}\left(A_\eta\right) \to 1, \ h \to \infty.$$

Proof. We prove this for $p, q < \infty$. Define random variables

$$L_{j,h} = 2^{(j+h)s} \left(\sum_{k=0}^{2^{j+h}-1} |\theta_{j+h,k}|^p \right)^{\frac{1}{p}}.$$

Thus, the event A_η is equivalent to $\left\{ \left(\sum_j L_{j,h}^q \right) \leq C \cdot (1+\eta) \right\}$ by the properties of the Besov norm. Note that $\epsilon_h 2^{hs} = 2^{-\frac{h}{p}}$. Then let

$$V_{j,h}^p = Ave_{0 \leq k < 2^{j+h}} |X_{j,k}|^p = \frac{1}{2^{j+h}} \sum_{k=0}^{2^{j+h}} |X_{j,k}|^p.$$

Thus, $L_{j,h} = 2^{j(s+1/p)} V_{j,h}$. Now the $X_{j,k}$ are bounded random variables and therefore $V_{j,h}$ is the mean of bounded random variables, thus

$$Prob\left\{ V_{j,h}^p > E\left(V_{j,h}^p\right) + \eta_j \right\} \to 0, \ h \to \infty$$

for any positive constant $\eta_j > 0$. Note $E\left(V_{j,h}^p\right) = E_{\mu_j} |X_{j,k}|^p$, and one of our properties of (μ_j), the distribution of $X_{j,k}$, was that

$$\sum_{j=-J}^{J} 2^{j(s+1/p)q} \left(E_{\mu_j} |X_{j,k}|^p \right)^{\frac{q}{p}} = C^q.$$

Here we have just used the assumption $p \leq q$ to set $p \wedge q = p$. Then by setting η_j small, we conclude

$$Prob\left\{\left(\sum_j L_{j,h}^q\right)^{\frac{1}{q}} \leq C(1+\eta)\right\} \to 1,\ h \to \infty.$$

□

We now show that this implies the theorem. We let $\nu(\cdot) = \mu(\cdot|A)$. Then provided $\mu(A^c)$ is small, ν and μ have almost the same Bayes risks. The phrase almost the same is clarified in the Fact below.

For the rest of the section, let π be a prior distribution for the vector parameter $\xi = (\xi_0, \xi_1, ...)$ and let $\beta(\pi)$ denote the Bayes risk for the problem of estimating ξ_0 with squared error loss from data $v_i = \xi_i + z_i$, $i = 0, 1, 2, 3, ...$ where $z_i \sim N(0,1)$ and are identically independent.

Fact 39. *Let ξ_0 be a bounded Random Variable with $|\xi_0| \leq M$. Let ω be the conditional prior distribution*

$$\omega(\cdot) = \pi(\cdot|A)$$

where A is an event. Then

$$|\beta(\omega) - \beta(\pi)| \leq 8M^2 \cdot \pi(A^c).$$

Proof. What we have here are two different Bayes risks for a quantity X both based on different priors. Note that the Bayes rules (estimators) are bounded almost everywhere by M and that their squared errors are bounded almost everywhere by $(2M)^2$. The Bayes risks are thus expectations of squared errors that are bounded almost everywhere by $(2M)^2$. The L^1 distance between π and ω is $P(A^c)$. The expectation of an almost everywhere bounded random variable under two different measures has a difference that is controlled by L^1 distance between the measures, times the bound on the random variable. □

We can now apply these Facts to prove the theorem. Let $\nu^{(h)} = \mu^{(h)}(\cdot|A_\eta)$. Then by the definition of A_η, $\nu^{(h)}$ is supported in $\Theta_{p,q}^s(C(1+\eta))$. The Bayes risk is

$$B\left(\nu^{(h)}\right) = \sum_{-J}^{J} \sum_{k=0}^{2^{J+h}} \tilde{b}_{j,k}$$

where

$$\tilde{b}_{j,k} = \inf_{\hat{\theta}} E_{\nu^{(h)}}\left(\hat{\theta}(y) - \theta_{I_{j,k}}\right)^2.$$

We now do a renumbering of the dyadic intervals in order to apply the Fact above. Let $J_{j,k}(i)$, $i = 0, 1, 2, ...$ be an enumeration of the dyadic intervals beginning with $J_{j,k}(0) = I_{j,k}$. We can do this

because the dyadic intervals are a countably infinite set. Let $\xi_0 = \theta_{I_{j,k}}/\epsilon$ and $\xi_i = \theta_{J_{j,k}(i)}/\epsilon$. It does not matter which ξ_i goes with which $\theta_{J_{j,k}(i)}$, it only matters that there is a one to one correspondence between the two. Let $\pi_{j,k}$ be the prior induced on ξ by the prior μ on θ; and let $\omega_{j,k}$ be the prior induced on ξ by $\nu^{(h)}$. Then since $\theta_I = \epsilon_h \cdot X_{j,k}$,

$$\tilde{b}_{j,k} = \epsilon_h^2 \beta\left(\omega_{j,k}\right)$$

where

$$\omega_{j,k}(\cdot) = \pi_{j,k}\left(\cdot | \theta \in \Theta_{p,q}^s\right).$$

Apply Fact 39.

$$\beta\left(\omega_{j,k}\right) \geq \beta\left(\pi_{j,k}\right) - 8M^2\mu^{(h)}\left(A_\eta^c\right)$$

Now since coordinates are independent and identically independent within one level of the prior μ,

$$\beta\left(\pi_{j,k}\right) = b_1\left(\mu_j\right),$$

for $0 \leq k < 2^{j+h}$. It follows that

$$\beta\left(\omega_{j,k}\right) \to b_1\left(\mu_j\right),$$

as $h \to \infty$ uniformly in $0 \leq k < 2^{j+h}$. Combining the above with $\epsilon_h^2 2^h = \left(\epsilon_h^2\right)^r$ and $\eta_h \to 0$, (6.21) gives

$$B\left(\nu^{(h)}\right) \geq \epsilon_h^2 \sum_{-J}^{J} 2^{j+h} b_1\left(\mu_j\right)\left(1 + o(1)\right) = \left(\epsilon_h^2\right)^r \sum_{-J}^{J} 2^j b_1\left(\mu_j\right)\left(1 + o(1)\right)$$

$$\geq \left(\epsilon_h^2\right)^r \cdot \left(val\left(Q_{1,C}\right)\left(1 - \eta\right)\right)\left(1 + o(1)\right).$$

This is true for each $\eta > 0$, so we get (6.20).

6.1.8. *Near-Minimax Threshold Estimates.* Donoho and Johnstone have now derived an asymptotically minimax estimator for $\Theta_{p,q}^s$ built out of a coordinate-wise nonlinear minimax estimator. Unfortunately, like the oracle risks of [8], these nonlinearities are not available to us in a closed form. We must approximate them by using soft or hard thresholding. Let

$$\delta_\lambda(y) = sgn(y)\left(|y| - \lambda\right)_+$$

denote the soft thresholding operator for this section. Recall that this operator is continuous and Lipschitz. Also define

$$\delta_\mu(y) = y 1_{\{|y| \geq \mu\}}.$$

This operator is discontinuous.

Suppose our data are of the form $y_I = \theta_I + z_I$. Here the θ_I satisfy the moment constraint $\tau \in \Theta^s_{p,q}$. We threshold coordinate-wise. Set $\lambda = (\lambda_I)$ and

$$\hat{\theta}^{\lambda}_I = \delta_{\lambda_I}(y_I), \ I \in \mathcal{I}.$$

The minimax risk for soft-threshold estimates is defined

$$\mathcal{B}^*_{\lambda}(\epsilon, \Theta) = \inf_{(\lambda_I)} \sup_{\tau \in \Theta} E \left\| \hat{\theta}^{\lambda} - \theta \right\|_2^2.$$

For hard thresholds $\hat{\theta}^{\mu}_I = \delta_{\mu_I}(y_I)$ we define the minimax risk $\mathcal{B}^*_{\mu}(\epsilon, \Theta)$ similarly. Here we establish the following theorem.

Theorem 40. *There are constants $\Lambda(p)$ and $M(p)$ both finite with*

$$\mathcal{B}^*_{\lambda}\left(\epsilon, \Theta^s_{p,q}\right) \leq \Lambda(p \wedge q) \mathcal{B}^*\left(\epsilon, \Theta^s_{p,q}\right)$$

$$\mathcal{B}^*_{\mu}\left(\epsilon, \Theta^s_{p,q}\right) \leq M(p \wedge q) \mathcal{B}^*\left(\epsilon, \Theta^s_{p,q}\right).$$

The thresholds which achieve these performances have the form

$$\lambda_I = \epsilon \cdot l\left(t^{\lambda}_j, \epsilon, p\right)$$

$$\mu_I = \epsilon \cdot m\left(t^{\mu}_j, \epsilon, p\right)$$

for $I \in \mathcal{I}$ with certain functions l and m and sequences t^{λ} and t^{μ}.

In order to show this, we must reconsider the sequence experiment where we found a thresholding sequence component-wise. We are estimating θ when the measure μ is known to lie in $\mathcal{M}^s_{p,q}$. Suppose we are using thresholds $\lambda = (\lambda_I)$. Denote $r(\lambda, \pi) = E_{\pi}\left(\delta_{\lambda}(\nu) - \xi\right)^2$ of the estimator δ_{ν} in the scalar problem $\nu = \xi + z$ with $\xi \sim \pi$ and $z \sim N\left(0, \epsilon^2\right)$. We can then rewrite the risk of the threshold estimator as

$$L(\lambda, \mu) = \sum_l r(\lambda_I, \mu_I),$$

and the minimax threshold risk is

$$\mathcal{B}^*_{\lambda}\left(\epsilon; \Theta^s_{p,q}\right) = \inf_{\lambda} \sup_{\mu \in \mathcal{M}^s_{p,q}} L(\lambda, \mu).$$

In order to calculate this, we need the following theorem.

Theorem 41. *We have*

$$\inf_{\lambda} \sup_{\mu \in \mathcal{M}^s_{p,q}} L(\lambda, \mu) = \sup_{\mu \in \mathcal{M}^s_{p,q}} \inf_{\lambda} L(\lambda, \mu). \tag{6.22}$$

Proof. We set the $\inf_\lambda r(\lambda,\mu) = \rho_*(\pi)$ and letting $\lambda_*(\pi)$ denote the minimizing threshold λ. Hence $\inf_\lambda L(\lambda,\mu) = \sum_I \rho_*(\mu_I)$ and the right side of (6.22) is equal to

$$\sup\left\{\sum_I \rho_*(\mu_I) : \mu \in \mathcal{M}^s_{p,q}\right\}.$$

This supremum is attained by some measure μ^*, which is a least favorable prior. There is a corresponding sequence $\lambda^* = (\lambda_*(\mu_I^*))$ of thresholds which are optimal if μ^* is the real prior.

We claim (λ^*, μ^*) is a saddlepoint of L. Let $l_I(\mu_I) = r(\lambda_I^*, \mu_I)$. Then

$$L(\lambda^*, \mu) = \sum_I l_I(\mu_I).$$

We have fixed λ_I^*, thus μ_I only plays a role in the expected value part of $l_I(\mu_I)$. Thus, l_I is affine in μ_I.

$$l_I(\mu_I) = l_I(\mu_I^*) + \dot{l}_I(\mu_I - \mu_I^*)$$

where $\dot{l}_I$ is a linear function. We have

$$L(\lambda^*, \mu) = L(\lambda^*, \mu^*) + \sum_I \dot{l}_I(\mu_I - \mu_I^*). \tag{6.23}$$

Let $m(\mu_I) = \rho_*(\mu_I)$. Then $\inf_\lambda L(\lambda,\mu) = \sum_I m(\mu_I)$. Because μ^* is least favorable for thresholds, we may write

$$\sum_I m(\mu_I^*) = \sup_{\mu \in \mathcal{M}^s_{p,q}} \sum_I m(\mu_I).$$

If we follow a path along $(1-t)\mu^* + t\mu$ away from μ^* towards $\mu \in \mathcal{M}^s_{p,q}$, the quantity $m(\mu_I)$ must decrease. This means that the pathwise derivative must decrease. With $\dot{m}_I$ the directional derivative of m at μ_I^*,

$$\sum_I \dot{m}_I(\mu_I - \mu_I^*) \le 0. \tag{6.24}$$

Comparing (6.23) with (6.24) shows that if

$$\dot{l}_I(\mu_I - \mu_I^*) \le \dot{m}_I(\mu_I - \mu_I^*)$$

for $I \in \mathcal{I}$, we would have

$$L(\lambda^*, \mu) \le L(\lambda^*, \mu^*).$$

The authors prove this by considering the quantity $\Gamma(\pi_1, \pi_2) = r(\lambda_*(\pi_1), \pi_2)$ and its directional derivatives. □

6.1.9. *Near Minimaxity among all estimates.* Define the following quantities in terms of the scalar situation. Let

$$\rho_{\lambda,p}(\tau,\epsilon) = \inf_{\lambda\in[0,\infty]} \sup_{(E|\xi|^p)^{\frac{1}{p}}\leq\tau} E\left(\delta_\lambda(\nu)-\xi\right)^2$$

and

$$\rho_{\mu,p}(\tau,\epsilon) = \inf_{\mu\in[0,\infty]} \sup_{(E|\xi|^p)^{\frac{1}{p}}\leq\tau} E\left(\delta_\mu(\nu)-\xi\right)^2$$

for the soft (λ) and hard (μ) thresholding operators respectively. To compare these estimates with the Bayes Minimax estimates we define

$$\Lambda(p) = \sup_{\tau,\epsilon} \frac{\rho_{\lambda,p}(\tau,\epsilon)}{\rho_p(\tau,\epsilon)}, \quad M(p) = \sup_{\tau,\epsilon} \frac{\rho_{\mu,p}(\tau,\epsilon)}{\rho_p(\tau,\epsilon)}. \tag{6.25}$$

Donoho and Johnstone show in [7] that for $p\in(0,\infty]$, $\Lambda(p)<\infty$ and $M(p)<\infty$. Briefly summarizing the proof of this result, the quantities $\rho_{\lambda,p}(\tau,\epsilon)$ and $\rho_{\mu,p}(\tau,\epsilon)$ are the same continuous quantities from [8]. These quantities are bounded. Also, $\Lambda(p), M(p)>1$. In [7], the following theorem is proved.

Theorem 42. *We have* $\frac{\rho_{\lambda,p}(\tau,\epsilon)}{\rho_p(\tau,\epsilon)}$ *and* $\frac{\rho_{\mu,p}(\tau,\epsilon)}{\rho_p(\tau,\epsilon)} \to 1$ *as* $n\to 0$ *and* ∞.

Because all quantities involved are bounded, these ratios are bounded. In fact, numerical experiments performed by the authors indicate these quantities are smaller than 2.22 for all $p\geq 2$.

Denote

$$r_{\lambda,p}(\lambda,\tau;\epsilon) = \sup_{E|\theta|^p\leq\tau^p} E\left(\delta_\lambda(\nu)-\theta\right)^2.$$

This denotes the worst case risk of using threshold λ under our given conditions. [7] shows this function to be concave in τ^p for each fixed λ and ϵ. Also, let

$$l(\tau,\epsilon,p) = \arg\min_{\lambda} r_{\lambda,p}(\lambda,\tau;\epsilon)$$

stand for the minimax threshold. The quantities $r_{\mu,p}$ and $m(\tau,\epsilon,p)$ from Theorem 40 are defined similarly.

Combining this section and the last we can derive the near-minimaxity of thresholds among all estimates. Let $\tau^* = (\tau_I^*)$ be the moment sequence associated with μ^*. As $\mu^*\in\mathcal{M}^s_{p,q}$, $\tau^*\in\theta^s_{p,q}$. By construction, it must be true that

$$\rho_*\left(\mu_I^*\right) \leq r_{\lambda,p\wedge q}\left(\tau_I^*,\epsilon\right).$$

Because μ_I^* is the optimizing quantity, equality holds. Hence

$$\mathcal{B}^*_\lambda\left(\epsilon,\Theta^s_{p,q}\right) = \sum_I \rho_*\left(\mu_I^*\right) \text{ by (6.22)}$$

$$= \sum_I \rho_{\lambda,p\wedge q}\left(\tau_I^*,\epsilon\right) \leq \Lambda(p\wedge q)\sum_I \rho_{p\wedge q}\left(\tau_I^*,\epsilon\right) \text{ by (6.25)}$$

$$\leq \Lambda(p\wedge q)\mathcal{B}^*\left(\epsilon;\Theta_{p,q}^s\right) \text{ by (6.16)}.$$

This is the proof for soft thresholds, the proof for hard thresholds is similar to the above.

6.1.10. *Minimax Linear Risk.* We now wish to show that for $p<2$, linear methods are unable to perform as well as the minimax rate of convergence described above.

We follow the lines of [10]. We will need the definition of a quadratic hull. Let Θ be a set of sequences. Let Θ_+^2 be the set of sequences $\theta^2 = \left(\theta_I^2\right)_{I\in\mathcal{I}}$ arising from $\theta\in\Theta$. Then

$$QHull(\Theta) = \left\{\theta : \theta^2 \in Hull\left(\Theta_+^2\right)\right\}.$$

[10] has shown that

$$QHull\left(\Theta_{p,q}^s\right) = \Theta^s_{\max(p,2),\max(q,2)}$$

$$R_L^*\left(\epsilon;\Theta\right) = R_L^*\left(\epsilon, QHull(\Theta)\right)$$

and

$$R^*\left(\epsilon;QHull(\Theta)\right) \leq R_L^*\left(\epsilon;QHull(\Theta)\right) \leq \frac{5}{4}R^*\left(\epsilon;QHull(\Theta)\right)$$

for a general class of sets Θ.

These facts mean that linear methods can attain only suboptimal rates of convergence when $p<2$. For example, suppose $p\leq q<2$. Then we have

$$R_L^*\left(\epsilon,\Theta_{p,q}^s\right) = R_L^*\left(\epsilon, QHull\left(\Theta_{p,q}^s\right)\right) = R_L^*\left(\epsilon,\Theta_{2,2}^s\right)$$

$$\doteq R^*\left(\epsilon,\Theta_{2,2}^s\right) \sim Const\left(\epsilon^2\right)^{r'}$$

as $\epsilon\to 0$. Here $r' = r'(s,p,q) = r(s,2,2)$. As $r(s,2,2) < r(s,p,q)$ for $p<2$, linear estimators cannot attain the optimal rate of convergence.

6.1.11. *Function Estimation in White Noise.* Up to now we have been considering minimax and near-minimax estimation in terms of the sequence model by using the wavelet coefficients. We can now consider the correspondence with Nonparametric Regression. We consider the problem of estimation in the white noise model. We observe the stochastic process $Y(t)$, $t\in[0,1]$ where

$$Y(dt) = f(t)dt + \epsilon W(dt) \tag{6.26}$$

with W a standard Wiener process, and f the function of interest. We estimate f on the basis of these data and the foreknowledge that $f \in \mathcal{F}$ a convex class of functions. We define the minimax risk

$$R(\epsilon; \mathcal{F}) = \inf_{\hat{f}} \sup_{\mathcal{F}} E \left\| \hat{f} - f \right\|_2^2$$

and the minimax linear risk

$$R_L(\epsilon; \mathcal{F}) = \inf_{\hat{f} \text{ linear}} \sup_{\mathcal{F}} E \left\| \hat{f} - f \right\|_2^2.$$

We relate this problem to the data later. Now we show the asymptotic equivalence of the function space risks with the sequence space risks.

Let's consider general classes of smooth functions, as in the Besov spaces.

Theorem 43. *Let the wavelet basis be of regularity $r > \sigma$. Let $\mathcal{F}$ denote the class of all functions with $|f|_{B^\sigma_{p,q}[0,1]} \leq 1$. There exist c and C so that*

$$R^* \left(\epsilon, \Theta^s_{p,q}(c)\right)(1 + o(1)) \leq R(\epsilon, \mathcal{F}) \leq R^* \left(\epsilon, \Theta^s_{p,q}(C)\right)(1 + o(1)). \tag{6.27}$$

Moreover, an estimator nearly attaining the minimax risk for the sequence problem yields an estimator nearly attaining the risk in the function problem.

We define

$$x_k = \int_0^1 \phi_{l,k} Y(dt)$$

$$y_I = \int_0^1 \psi_I Y(dt)$$

for $k \in K$ and $I \in \mathcal{J}$. Then

$$x_k = \beta_{l,k} + \epsilon z_k$$

and

$$y_I = \alpha_I + \epsilon z_I.$$

Let Θ denote the collection of inhomogeneous wavelet expansions $\left((\beta_{l,k})_{k \in K}, (\alpha_I)_{I \in \mathcal{J}}\right)$ arising from functions $f \in \mathcal{F}$. The Parseval relation gives us

$$R(\epsilon, \mathcal{F}) = R^*(\epsilon, \Theta).$$

We now apply (2). By additivity of coordinate risks and Independence of noise, if $\Theta_0 = \mathbb{R}^{\#(K)}$ then

$$R^* (\epsilon, \Theta_0 \times \Theta_1) = R^* (\epsilon, \Theta_0) + R^* (\epsilon, \Theta_1) = \#(K)\epsilon^2 + R^* (\epsilon, \Theta_1).$$

Thus

$$R^*\left(\epsilon, \tilde{\Theta}^s_{p,q}(c)\right) \leq R^*(\epsilon, \Theta) \leq \#(K)\epsilon^2 + R^*\left(\epsilon, \tilde{\Theta}^s_{p,q}(C)\right).$$

We have

$$\{0\} \times \tilde{\Theta}^s_{p,q}(C) \subset \Theta^s_{p,q}(C) \subset \mathbb{R}^{2^l} \times \tilde{\Theta}^s_{p,q}(C).$$

So

$$R^*\left(\epsilon, \tilde{\Theta}^s_{p,q}\right) \leq R^*\left(\epsilon, \Theta^s_{p,q}\right) \leq R^*\left(\epsilon, \tilde{\Theta}^s_{p,q}\right) + 2^l\epsilon^2.$$

Combining these inequalities and noting that the terms $O\left(\epsilon^2\right)$ are negligible asymptotically yields (6.27).

We attain this risk asymptotically by shrinking wavelet coefficients using the minimax Bayes estimator for the sequence model; specifically

$$\hat{\beta}_{l,k} = x_k$$

$$\hat{\alpha}_I = \delta^*_j\left(y_I\right)$$

for $k \in K$ and $I \in \mathcal{J}$.

6.1.12. *Triebel Spaces.* We note here that all the results derived for the Besov spaces can also be applied to the Triebel spaces. Let $\chi_{j,k}$ denote the indicator function of $\left[k/2^j, (k+1)/2^j\right]$. We define the norm for this space in terms of the wavelet coefficients.

$$|\alpha|_{\tilde{f}^s_{p,q}} = \left\| \left(\sum_{\mathcal{J}} \left(2^{js}|\alpha_I|\chi_I\right)^q \right)^{\frac{1}{q}} \right\|_{L^p[0,1]}$$

The seminorm on functions f with wavelet coefficients $\alpha = \alpha(f)$ via

$$|f|_{\tilde{F}^\sigma_{p,q}} = |\alpha|_{\tilde{f}^s_{p,q}}.$$

This class of functions includes the Sobolev spaces, which are not a subset of the Besov spaces.

All of the properties necessary for the proofs of this part are also properties of the Triebel spaces. The most interesting of these are the following.

(1) We know $J^s_{p,q}(\tau) = \|\tau\|^p_{\tilde{f}^s_{p,q}}$ is a convex functional of the moment sequence $\tau_I^{p\wedge q}$ $(p, q < \infty)$.

(2) If $(\tau_{j,k})$ is an arbitrary positive sequence, and we set $\bar{\tau}_I^{p\wedge q} = Ave_{I\in I_j}\left(\tau_I^{p\wedge q}\right)$, then

$$\|\tilde{\tau}\|_{\tilde{f}^s_{p,q}} \leq \|\tau\|_{\tilde{f}^s_{p,q}}.$$

These properties are the ones which allow us to find a least favorable prior and set the whole sequence of proofs in motion. The first property is evident by inspection. To prove the second we consider $p \leq q$ and $p \geq q$ separately.

In the case $p \leq q$, define $f_j = \sum_{I_j} 2^{jsp} \tau_I^p \chi_I$. Then with $r = q/p \geq 1$ we have

$$\|\tau\|^p_{\tilde{f}^s_{p,q}} = \int_0^1 \left(\sum_{j \geq 0} f_j^r \right)^{1/r} dt.$$

As $f_j \geq 0$ and t^r is convex,

$$\int_0^1 \left(\sum_{j \geq 0} f_j(t)^r \right)^{1/r} dt \geq \left(\sum_{j \geq 0} \left(\int_0^1 f_j(t) dt \right)^r \right)^{1/r}.$$

Now

$$\int_0^1 f_j(t) dt = 2^{jsp} Ave_{I \in I_j} \left(|\tau_I|^p \right) = 2^{jsp} t_j^p$$

say. The average measure $\bar{\mu}$ has moment sequence $\bar{\tau}_I = t_j$, so

$$\|\bar{\tau}\|^p_{\tilde{f}^s_{p,q}} = \left(\sum_{j \geq 0} \left(2^{jsp} t_j^p \right)^r \right)^{1/r}$$

and the second property follows by combining the inequalities.

In the case $q \leq p$ define $f_j = \sum_{I_j} 2^{jsp} \tau_I^q \chi_I$ and set $r = p/q \geq 1$. Then

$$\|\tau\|^p_{\tilde{f}^s_{p,q}} = \int_0^1 \left(\sum_{j \geq 0} f_j \right)^r dt.$$

As $f_j \geq 0$ and t^r is convex, Jensen's inequality gives

$$\int_0^1 \left(\sum_{j \geq 0} f_j(t) \right)^r dt \geq \left(\int_0^1 \sum_{j \geq 0} f_j(t) dt \right)^r.$$

Now

$$\int_0^1 f_j(t) dt = 2^{jsp} Ave_{I \in I_j} \left(|\tau_I|^q \right) = 2^{jsq} t_j^q$$

say. The average measure $\bar{\mu}$ has moment sequence $\bar{\tau}_I = t_j$, so

$$\|\bar{\tau}\|^p_{\tilde{f}^s_{p,q}} = \left(\sum_{j \geq 0} 2^{jsp} t_j^q \right)^r$$

and the second property follows by combining the inequalities.

This eventually leads us to a theorem parallel to the one for the Besov spaces. We let $\Phi^s_{p,q}$ denote the Triebel space in the sequence space, and $\Phi^s_{p,q}(C)$ denote a set of bounded sequences just as in the definitions of $\Theta^s_{p,q}(C)$ from before.

Theorem 44. *Let the wavelet basis be of regularity $r > m$ and let $1 < p < \infty$. Let $\mathcal{F}$ denote the class of f with $\left\|f^{(m)}\right\|_{L^p[0,1]} \leq 1$. There exist c and C, depending on the wavelet basis, so that*

$$R^* \left(\epsilon, \Phi^m_{p,2}(c)\right)(1 + o(1)) \leq R(\epsilon, \mathcal{F}) \leq R^* \left(\epsilon, \Phi^m_{p,2}(C)\right)(1 + o(1)).$$

Moreover, an estimator attaining the minimax risk for the sequence problem yields an estimator attaining the minimax risk in the function problem.

This particular subset of the Triebel space is the Sobolev space.

6.1.13. *Nonparametric Regression and White Noise.* We can now finally connect our results to the nonparametric regression model (6.1). Define the regression process $\{Y_n(t) : t \in [0,1]\}$ via $t_0 = 0$, $Y_n(0) = 0$ and

$$Y_n(t_i) = \frac{1}{n} \sum_{t \leq t_i} y_i,$$

for $i = 1, ..., n$ with interpolation between the t_i by independent Brownian Bridges $W_{0,i}$: for $t_i \leq t \leq t_{i+1}$ set

$$Y_n(t) = Y_n(t_i) + (t - t_i)\, y_{i+1} + \frac{\sigma}{n} W_{0,i}\left(n(t - t_i)\right).$$

Define the step function

$$f_n(t) = \sum_{i=1}^{n} f(t_i)\, 1_{\{t_{i-1} \leq t < t_i\}}.$$

Then

$$Y_n(dt) = f_n(t)dt + \epsilon W(dt)$$

where W is a Wiener process and $\epsilon = \frac{\sigma}{n}$. In a probability space the difference between Y and Y_n is the same as the distance between f and f_n.

In [3], Brown and Low study the degree of the approximation of the experiments $(Y_n, \mathcal{F})$ by $(Y, \mathcal{F})$. Set

$$D_n(\mathcal{F}) = \sup_{\mathcal{F}} \left\| \hat{f} - f_n \right\|_2^2.$$

They show that if $D_n = o(\frac{1}{n})$, then the experiments $(Y_n, \mathcal{F})$ and $(Y, \mathcal{F})$ are indistinguishable by any statistical tests. Consequently, if l is any bounded function and $\hat{f}$ is any measurable estimator the worst case risk

$$\sup_{\mathcal{F}} E\left(l\left(n^r \left\| \hat{f} - f \right\|_2^2 \right) \right)$$

has the same asymptotic limit in both experiments. In other words, results in the nonparametric regression model are identical to those in the white noise model. [9] presents two theorems without proof. First we give the result for the lower bound.

Theorem 45. *Let $\mathcal{F}$ consist of all function in a $B_{p,q}^s$ ball or an $F_{p,q}^\sigma$ ball, $\sigma > 0$. Then*

$$R(n, \mathcal{F}) \geq R^*(\frac{\sigma}{\sqrt{n}}, \mathcal{F})(1 + o(1))$$

as $n \to \infty$.

Now we present the result for the upper bound.

Theorem 46. *If $\mathcal{F}$ is a Besov or Triebel ball with either $\sigma > 1/p$, or with $\sigma = 1$, $p, q \leq 1$; or if $\mathcal{F}$ is a ball of functions of bounded variation, then*

$$R(n, \mathcal{F}) \leq R^*(\frac{\sigma}{\sqrt{n}}, \mathcal{F})(1 + o(1))$$

as $n \to \infty$.

Thus we see that the result are the same for the approximation and the white noise model and this method is applicable.

7. Conclusion.

We have now considered many different methods of approximating $f(x)$. These methods are powerful, but also somewhat limited. The methods of Part 2 are easy to apply, but we see that they give weak convergence to the true $f(x)$.

The most diverse method of estimation studied in this paper is the one in Part 4. This method works for any $f(x)$, but the bound of $2 \log n$ can still become large. The risk derived in Part 5 is bounded by a constant but requires that $f(x)$ be a member of a very special space. Also, the thresholding of Parts

4 and 5 is applied coordinate-wise. Deriving the appropriate threshold could take too much computing time.

These methods can be improved upon by something called block thresholding. In this method the wavelet coefficients are grouped into blocks and are thresholded as one object. A norm is defined for each block of values. This method requires less computing time, and the risk associated with block thresholding is comparable to the one associated with coordinate-wise thresholding.

Another problem with the methods above is that they all deal with data that has error which is distributed normally and is identically independent. In many real-life situations this is not practical. For instance, one might consider the phenomenon of long memory. Intuitively, one might consider a riverbank. Suppose every 1 foot square patch of ground by a river is assigned an x value. Then if a patch of ground x_j is next to a flooded patch of ground x_i, it is much more likely that x_j would be flooded. This is called long memory dependence. It is expressed by

$$E\left(\epsilon_i \epsilon_{i+j}\right) \sim C_0 |j|^{-\alpha}$$

where

$$y_i = f\left(x_i\right) + \epsilon_i$$

and the ϵ_i are long memory Gaussian errors and $\alpha \in (0, 1]$. Bounding thresholding operators with this type of error is discussed in [18]. This is only one example of the many different types of errors some of which have yet to be examined with respect to this method.

Another problem which we encounter is with the orthonormal wavelet basis we use to perform the wavelet transform. The problem is that if the wavelets have compact support, then their duals have unbounded support. This makes computation costly timewise and not practical. To understand this, let us recall $w = Wy$, where w is the vector of wavelet coefficients, y is the vector of data, and W is the transform matrix of Part 3. Since W is an orthogonal matrix, $y = W^T w$. Now suppose W is a bandlimited matrix as below.

$$W = \begin{bmatrix} \# & \# & \# & 0 & 0 & 0 & 0 \\ \# & \# & \# & \# & 0 & 0 & 0 \\ \# & \# & \# & \# & \# & 0 & 0 \\ 0 & \# & \# & \# & \# & \# & 0 \\ 0 & 0 & \# & \# & \# & \# & \# \\ 0 & 0 & 0 & \# & \# & \# & \# \\ 0 & 0 & 0 & 0 & \# & \# & \# \end{bmatrix}$$

This means that W will have zeros on the edges and have a band of nonzero entries in the center. The inverse of such a matrix would have global support, or very few zeros. This means that data reconstruction takes a very long time. We can address this problem by considering the use of tight frames. We relax the condition that our wavelets are orthogonal and allow redundancy in our basis for

the space. This means we may have linear dependence within our set of frames. Using these frames means our dual matrix will have compact support.

Another area which has not been discussed here is that of nonstationary data. All of the data we have considered has been equally spaced. This is rarely the case in real-life data. Data points are spread unevenly. Theory using nonstationary wavelets has not yet been applied to soft or hard thresholding. We may also use nonstationary tight frames to analyze data. Nonstationary tight frames have been constructed from B-splines in [14, 15]. The refinement matrices are based on adding knots to an initial sequence of knots.

Therefore, we can see that there are many potential areas of study left to explore within the topic of function estimation.

Part 3. Summary of the work of Li and Xiao in [18].

8. Preliminaries and Notations.

In this part we give a summary of the work of Li and Xiao in [18]. This paper analyzed data which is equally spaced but has long memory error. It also uses block thresholding. We consider the nonparametric regression

$$Y_m = g(x_m) + \epsilon_m, \qquad m = 1, 2, ..., n \tag{8.1}$$

where $x_m = m/n \in [0,1]$ and $\epsilon_1, ..., \epsilon_n$ are errors. The function g, which is supported in $[0,1]$ belongs to the class of functions $\mathcal{H}$ used in [11].

Definition 47. $\mathcal{H}$ is the class of functions g such that for any $i \geq 0$ there exists a set of integers S_i for which the following is true: $\text{card}(S_i) \leq C_3 2^{i\gamma}$ and

For each $j \in S_i$ there exist constants $a_0 = g\left(j/2^i\right)$, $a_1, ..., a_{N-1}$ such that

$$\left| g(x) - \sum_{i=0}^{N-1} a_l \left(x - 2^{-i}j\right)^l \right| \leq C_1 2^{-is_1} \text{ for all } x \in \left[j/2^i, (j+\nu)2^i\right].$$

For each $j \notin S_i$ there exist constants $a_0 = g\left(j/2^i\right)$, $a_1, ..., a_{N-1}$ such that

$$\left| g(x) - \sum_{i=0}^{N-1} a_l \left(x - 2^{-i}j\right)^l \right| \leq C_2 2^{-is_2} \text{ for all } x \in \left[j/2^i, (j+\nu)/2^i\right].$$

Here the errors are not normal and identically independent. Instead the errors ϵ_m are long memory dependent. There exist two constants $C_0 > 0$ and $\alpha \in (0,1]$ such that

$$r(j) = E(\epsilon_1 \epsilon_{1+j}) \sim C_0 |j|^{-\alpha} \tag{8.2}$$

where $a_j \sim b_j$ means that $a_j/b_j \to 1$ when $j \to \infty$.

We have the same standard wavelet properties. Let $\phi(x)$ and $\psi(x)$ be the mother and father wavelets. Also let

$$\phi_{ij}(x) = 2^{i/2}\phi\left(2^i x - j\right) \qquad \psi_{ij}(x) = 2^{i/2}\psi\left(2^i x - j\right).$$

And

$$\alpha_{ij} = \int f(x)\phi_{ij}(x)dx \qquad \beta_{ij} = \int f(x)\psi_{ij}(x)dx.$$

The work of Li and Xiao uses the Coiflets. It lists these wavelets as having the property that

$$\int x^k \phi(x)dx = \int x^k \psi(x)dx = 0 \qquad \text{for } k = 1, ..., N-1.$$

The wavelet expansion of $g(x)$ is

$$g(x) = \sum_{j=0}^{2^{i_0}-1} \alpha_{i_0}\phi_{i_0 j}(x) + \sum_{i \geq i_0}\sum_{j=0}^{2^i-1} \beta_{ij}\psi_{ij}(x) \tag{8.3}$$

where

$$\alpha_{ij} = \int_0^1 g(x)\phi_{ij}(x)dx \qquad \beta_{ij} = \int_0^1 g(x)\psi_{ij}(x)dx.$$

We assume the sample size is $n = 2^i$. We define a function $h(n)$ so that $2^{i(n)} \approx h(n)$ means $2^{i(n)} \leq h(n) < 2^{i(n)+1}$.

These coefficients are going to be block-thresholded. This means that they will be thresholded in groups. At each level i, the integers $\{0, 1, ..., 2^i - 1\}$ are divided into blocks of length l.

$$\Gamma_{ik} = \{j : (k-1)l + 1 \leq j \leq kl\}$$

Here $-\infty < k < \infty$.

The convergence rates of the Mean Integrated Square Error (MISE) are different for $\alpha \in (0, 1)$ and for $\alpha = 1$.

We define

$$\hat{G}_{i_1}(x) = n^{-1/2}\sum_{m=1}^{n} Y_m \phi_{i_1 m}(x).$$

The $n^{-1/2}$ is a part of the standard assumptions while performing wavelet decompositions. Also note

$$\text{Proj}_{V_{i_0}}\left(\hat{G}_{i_1}\right) = \sum_{j=0}^{2^{i_0}-1} \hat{\alpha}_{i_0 j}\phi_{i_0 j} \quad \text{and} \quad \text{Proj}_{W_i}\left(\hat{G}_{i_1}\right) = \sum_{j=0}^{2^i-1} \hat{\beta}_{ij}\psi_{ij}.$$

We put $\hat{B}_{ik} = l^{-1}\sum_{(ik)}\hat{\beta}_{ij}^2$ where $\sum_{(ik)}$ denotes summation over $j \in \Gamma_{ik}$ and l denotes the block length. For $\alpha \in (0,1)$ our estimator is defined

$$\hat{g}(x) = \sum_{j=0}^{2^{i_0}-1} \hat{\alpha}_{i_0 j}\phi_{i_0 j}(x) + \sum_{i=i_0}^{i_1-1}\sum_{k=-\infty}^{\infty}\left(\sum_{(ik)}\hat{\beta}_{ij}\psi_{ij}(x)\right) I\left(\hat{B}_{ik} > \delta_i\right) \tag{8.4}$$

where $2^{i_0} \simeq n^{\alpha/(2N+\alpha)}$, the block length is $l = (\log n)^{\theta}$ with $\theta > 1/\alpha$ and δ_i $(i_0 \le i < i_1)$ are the level-dependent thresholds satisfying $\delta_i = 48\tau_i^2$ with $\tau_i^2 = C_4 n^{-\alpha} 2^{-i(1-\alpha)}$ where $C_4 > 0$ is defined by

$$C_4 = C_0 \int_0^1\int_0^1 |x-y|^{-\alpha}\,\psi(x)\psi(y)dxdy. \tag{8.5}$$

Similarly, for $\alpha = 1$, the estimator is defined

$$\hat{g}_1(x) = \sum_{j=0}^{2^{i_0}-1} \hat{\alpha}_{i_0 j}\phi_{i_0 j}(x) + \sum_{i=i_0}^{i_1-1}\sum_{k=-\infty}^{\infty}\left(\sum_{(ik)}\hat{\beta}_{ij}\psi_{ij}(x)\right) I\left(\hat{B}_{ik} > \delta_i\right)$$

where the smoothing parameter is chosen to satisfy $2^{i_0} \simeq n^{1/(2N+1)}$, the block length $l = (\log n)^{\theta}$ with $\theta > 1$ and δ_i $(i_0 \le i < i_1)$ are the level dependent thresholds satisfying $\delta_i = 48\xi_i^2$ with $\xi_i^2 = 2C_0 n^{-1}\log\left(n2^{-i}e\right)$. The only difference for $\alpha = 1$ is in the threshold δ_i.

9. The main result of the paper.

Below is the main theorem of the paper.

Theorem 48. *Let the wavelets ϕ and ψ, and the estimators $\hat{g}$ and $\hat{g}_1$ be given as before. Then there exists a constant $C_5 = C(s_1, s_2, \gamma, C_1, C_2, C_3, N, \nu) > 0$ such that the following hold:*

(i) When $\alpha \in (0,1)$,

$$\sup_{g\in\mathcal{H}(s_1,s_2,\gamma,C_1,C_2,C_3,N,\nu)} E\int (\hat{g}-g)^2 \le C_5 2^{-2s_2\alpha/(2s_2+\alpha)}.$$

(ii) When $\alpha = 1$,

$$\sup_{g\in\mathcal{H}(s_1,s_2,\gamma,C_1,C_2,C_3,N,\nu)} E\int (\hat{g}-g)^2 \le C_5\left(\frac{\log n}{n}\right)^{2s_2/(2s_2+1)}.$$

Now we examine the proof.

10. Summary of the proof of the main result.

We split the mean square error into parts and bound each part. Observe that orthogonality implies

$$E\left\|\hat{g}-g\right\|_2^2 = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \sum_{i=i_1}^{\infty}\sum_{j=0}^{2^i-1}\beta_{ij}^2,$$

$$T_2 = \sum_{j=0}^{2^{i_0}-1} E\left(\hat{\alpha}_{i_0 j}-\alpha_{i_0 j}\right)^2 = E\left\|\mathrm{Proj}_{V_{i_0}}\left(\hat{G}_{i_1}-g\right)\right\|_2^2,$$

$$T_3 = \sum_{i=i_0}^{i_1-1}\sum_{k=-\infty}^{\infty} E\left\{I\left(\hat{B}_{ik}>\delta_i\right)\sum_{(ik)}\left(\hat{\beta}_{ij}-\beta_{ij}\right)^2\right\},$$

$$T_4 = \sum_{i=i_0}^{i_1-1}\sum_{k=-\infty}^{\infty} P\left(\hat{B}_{ik}\le\delta_i\right)\sum_{(ik)}\beta_{ij}^2.$$

We will need the following result.

Proposition 49. *For every* $g\in\mathcal{H}\left(s_1,s_2,\gamma,C_1,C_2,C_3,N,\nu\right)$ and our selected coiflets,

$$|\beta_{ij}|\le\|\psi\|_1 C_1 2^{-i(s_1+1/2)} \text{ if } j\in S_i$$

$$|\beta_{ij}|\le\|\psi\|_1 C_2 2^{-i(s_2+1/2)} \text{ if } j\notin S_i$$

$$\left|\alpha_{ij}-2^{-i/2}g\left(j/2^i\right)\right|\le\|\phi\|_1 C_1 2^{-i(s_1+1/2)} \text{ if } j\in S_i$$

$$\left|\alpha_{ij}-2^{-i/2}g\left(j/2^i\right)\right|\le\|\phi\|_1 C_2 2^{-i(s_2+1/2)} \text{ if } j\notin S_i \tag{10.1}$$

Note: these next lines are exactly the same in Hall 1999. As in Hall 1999 pg. 42, there exist real numbers $r_{i_1 m}$ $(m=1,...,n)$ which are small when n is large, such that

$$\alpha_{i_1 m} = \int g(x)\phi_{i_1 m}(x)dx = n^{-1/2}\int g\left(\frac{m}{n}+\frac{y}{n}\right)\phi(y)dy \equiv n^{-1/2}g\left(\frac{m}{n}\right) - r_{i_1 m}. \tag{10.2}$$

Thus we can write (8.4) as

$$\hat{G}_{i_1}(x) = \sum_{m=1}^{n}\left(\alpha_{i_1 m}+r_{i_1 m}\right)\phi_{i_1 m}(x) + n^{-1/2}\sum_{m=1}^{n}\epsilon_m\phi_{i_1 m}(x).$$

We can write then for any integer $0 \leq i < i_1$

$$\mathrm{Proj}_{W_i}\left(\hat{G}_{i_1}\right) = \sum_{j=0}^{2^i-1} \left(\beta_{ij} + u_{ij} + U_{ij}\right) \psi_{ij}(x),$$

$$\mathrm{Proj}_{V_{i_0}}\left(\hat{G}_{i_1}\right) = \sum_{j=0}^{2^i-1} \left(\alpha_{i_0 j} + \nu_{i_0 j} + V_{i_0 j}\right) \phi_{i_0 j}(x).$$

Since $\mathrm{Proj}_{V_{i_1}}(g) = \sum_j \alpha_{i_1 j}\phi_{i_1 j}(x)$ we have for $0 \leq i < i_1$, $\mathrm{Proj}_{W_i}(g) = \mathrm{Proj}_{W_i}\left(\mathrm{Proj}_{V_{i_1}}(g)\right)$. Now $\mathrm{Proj}_{W_i}(g) = \sum_j \beta_{ij}\psi_{ij}(x)$. Thus

$$\beta_{ij} = \sum_{m=1}^{n} \alpha_{i_1 m} \left\langle \phi_{i_1 m}, \psi_{ij} \right\rangle.$$

We examine these decompositions. We have

$$u_{ij} = \sum_{m=1}^{n} r_{i_1 m} \left\langle \phi_{i_1 m}, \psi_{ij} \right\rangle, \qquad \nu_{i_0 j} = \sum_{m=1}^{n} r_{i_1 m} \left\langle \phi_{i_1 m}, \phi_{i_0 j} \right\rangle, \tag{10.3}$$

and

$$U_{ij} = \frac{1}{\sqrt{n}} \sum_{m=1}^{n} \epsilon_m \left\langle \phi_{i_1 m}, \psi_{ij} \right\rangle, \qquad V_{i_0 j} = \frac{1}{\sqrt{n}} \sum_{m=1}^{n} \epsilon_m \left\langle \phi_{i_1 m}, \phi_{i_0 j} \right\rangle. \tag{10.4}$$

Recall that $\langle f, g\rangle = \int fg$, the inner product in $L^2([0,1])$. From Parseval's identity we have

$$\sum_{i=i_0}^{i_1-1} \sum_{j=0}^{2^i-1} u_{ij}^2 + \sum_{j=0}^{2^{i_0}-1} \nu_{i_0 j}^2 = \sum_{m=1}^{n} r_{i_1 m}^2.$$

From (10.1) and (10.2) and our choice of γ in (47), we have

$$\sum_{m=1}^{n} r_{i_1 m}^2 \leq C_1 C_3 n^{-(2s_1+1-\gamma)} + C_2 n^{-2s_2} \leq C n^{-2s_2/(2s_2+1)}. \tag{10.5}$$

Because our wavelets have compact support, there are at most 2^{i_1-i} non-zero terms of $\langle \phi_{i_1 m}, \psi_{ij}\rangle$, $m = 1, 2, ..., n$. Moreover,

$$\left|\langle \phi_{i_1 j}, \psi_{ij}\rangle\right| \leq 2^{i/2-i_1/2} \left\|\psi\right\|_\infty \left\|\phi\right\|_1 \quad \text{and} \quad |r_{i_1 l}| \leq (C_1 \vee C_2)\, 2^{-i_1(s_1+1/2)}.$$

Hence we have for all $i \geq i_0$

$$|u_{ij}| \leq C 2^{i_1-i} 2^{-i_1(s_1+1/2)} 2^{i/2-i_1/2} \leq C 2^{-i(s_1+1/2)}. \tag{10.6}$$

Now we calculate the variance of U_{ij}. Since $EU_{ij} = 0$ we have

$$\mathrm{Var}\left(U_{ij}^2\right) = \frac{1}{n}\sum_{m=1}^{n} E\left(\epsilon_m^2\right)\langle\phi_{i_1 m}, \psi_{ij}\rangle^2 + \frac{1}{n}\sum_{m=1}^{n}\sum_{k\neq m} E\left(\epsilon_m\epsilon_k\right)\langle\phi_{i_1 m}, \psi_{ij}\rangle\langle\phi_{i_1 k}, \psi_{ij}\rangle$$

$$= \frac{\sigma^2}{n} + \frac{1}{n}\sum_{m=1}^{n}\sum_{k\neq m} r(m-k)\langle\phi_{i_1 m}, \psi_{ij}\rangle\langle\phi_{i_1 k}, \psi_{ij}\rangle \equiv \frac{\sigma^2}{n} + I_1. \tag{10.7}$$

Recall $n = 2^{i_1}$. By a change of variables we may write

$$I_1 = 2^i \sum_{m=1}^{n}\sum_{k\neq m} r(m-k)\int\int \phi\left(2^{i_1}x - m\right)\phi\left(2^{i_1}y - k\right)\psi\left(2^i x - j\right)\psi\left(2^i y - j\right)dxdy$$

$$= \int\int \phi(u)\phi(v)\left\{\sum_{m=1}^{n}\sum_{k\neq m} r(m-k)\psi\left(2^i\frac{u+m}{n} - j\right)\psi\left(2^i\frac{v+k}{n} - j\right)\frac{1}{n^2}\right\}dudv. \tag{10.8}$$

We consider the case where $\alpha \in (0,1)$. From (8.2) and as $n \to \infty$,

$$\sum_{m=1}^{n}\sum_{k\neq m} r(m-k)\psi\left(2^i\frac{u+m}{n} - j\right)\psi\left(2^i\frac{v+k}{n} - j\right)\frac{1}{n^2}$$

$$\sim C_0 n^{-\alpha}\sum_{m=1}^{n}\sum_{k\neq m}\left|\frac{m}{n} - \frac{k}{n}\right|^{-\alpha}\psi\left(2^i\frac{u+m}{n} - j\right)\psi\left(2^i\frac{v+k}{n} - j\right)\frac{1}{n^2}$$

$$\sim C_0 n^{-\alpha}\int_0^1\int_0^1 |x-y|^{-\alpha}\psi\left(2^i x - j\right)\psi\left(2^i y - j\right)dxdy$$

$$= C_0 n^{-\alpha}2^{(\alpha-2)i}\int_0^1\int_0^1 |x-y|^{-\alpha}\psi(x)\psi(y)dxdy \tag{10.9}$$

uniformly for all u and v in the support ϕ. Combining (10.8) and (10.9) we have

$$I_1 \sim C_4 n^{-\alpha}2^{-i(1-\alpha)} \quad \text{as} \quad n \to \infty, \tag{10.10}$$

where C_4 is the constant defined by (8.5). Thus

$$\mathrm{Var}\left(U_{ij}\right) \sim C_4 n^{-\alpha}2^{-i(1-\alpha)} \quad \text{as} \quad n \to \infty. \tag{10.11}$$

Similarly we have $EV_{i_0 j} = 0$ and for any fixed integer i

$$\mathrm{Var}\left(V_{i_0 j}\right) \sim C_6 n^{-\alpha}2^{-i_0(1-\alpha)} \quad \text{as} \quad n \to \infty, \tag{10.12}$$

where $C_6 > 0$ is the constant given by

$$C_6 = C_0 \int_0^1 \int_0^1 |x-y|^{-\alpha} \phi(x)\phi(y)dxdy.$$

A similar proof appears for $\alpha = 1$. A log function appears because of the factor of $|x-y|^{-\alpha}$.

We are now in a position to bound the intervals T_1, T_2, T_3, T_4.

10.1. **Bound for T_1.** Since $g \in \mathcal{H}$, we use Proposition 49 to derive

$$T_1 = \sum_{i=i_1}^{\infty} \left(\sum_{j \in S_i} + \sum_{j \notin S_i} \right) \beta_{ij}^2 \leq C \sum_{i=i_1}^{\infty} 2^{i\gamma} 2^{-i(2s_1+1)} + C \sum_{i=i_1}^{\infty} 2^i 2^{-i(2s_2+1)}$$

$$\leq Cn^{-(2s_1+1-\gamma)} + Cn^{-2s_2} \leq Cn^{-2s_2/(2s_2+1)}.$$

We have $T_1 \leq Cn^{-2s_2\alpha/(2s_2+\alpha)}$ when $\alpha \in (0,1)$ and $T_1 \leq C\left(n^{-1}\log n\right)^{2s_2/(2s_2+1)}$ when $\alpha = 1$.

10.2. **Bound for T_2.** From the definition of $\hat{\alpha}_{i_0 j}$, (10.5) and (10.12), we have for $\alpha \in (0,1)$,

$$T_2 = \sum_{j=0}^{2^{i_0}-1} v_{i_0 j}^2 + \sum_{j=0}^{2^{i_0}-1} EV_{i_0 j}^2 \leq Cn^{-2s_2/(2s_2+1)} + C2^{i_0} n^{-\alpha} 2^{-i_0(1-\alpha)}$$

$$\leq Cn^{-2s_2\alpha/(2s_2+\alpha)}$$

where this last inequality follows from (8.4).

When $\alpha = 1$, $T_2 \leq C\left(\frac{\log n}{n}\right)^{2s_2/(2s_2+1)}$.

10.3. **Bound for T_3.** We will need the following lemma to bound T_3.

Proposition 50. *Let U_{ij} be the Gaussian random variables defined from (10.4). Let $\tau_i^2 = C_4 n^{-\alpha} 2^{-i(1-\alpha)}$ and $\xi_i^2 = 2C_0 n^{-1} \log\left(n2^{-i}e\right)$.*

(i) If $\alpha \in (0,1)$, then for all integers i, k, and for all real numbers $\lambda \geq 4l\tau_i^2$,

$$P\left\{ \sum_{(ik)} U_{ij}^2 \geq \lambda \right\} \leq \exp\left(-\frac{\lambda}{C_7 l^{1-\alpha} \tau_i^2} \right)$$

where $C_7 > 0$ is an absolute constant.

(ii) If $\alpha = 1$, then for all real numbers $\lambda \geq 4l\xi_i^2$

$$P\left\{ \sum_{(ik)} U_{ij}^2 \geq \lambda \right\} \leq \exp\left(-\frac{\lambda}{C_8 \xi_i^2 \log l} \right)$$

where $C_8 > 0$ is an absolute constant.

To bound T_3, we write it as

$$T_3 = \sum_{i=i_0}^{i_1-1} \sum_{k=-\infty}^{\infty} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) \sum_{(ik)} (u_{ij} + U_{ij})^2 \right\}$$

$$\leq 2 \sum_{i=i_0}^{i_1-1} \sum_{k} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) \sum_{(ik)} U_{ij}^2 \right\} + 2 \sum_{i=i_0}^{i_1-1} \sum_{k} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) \sum_{(ik)} u_{ij}^2 \right\}$$

$$\equiv 2T_3' + 2T_{3.}'' \tag{10.13}$$

From (10.5) we have

$$T_3'' \leq \sum_{i=i_0}^{i_1-1} \sum_{k} \sum_{(ik)} u_{ij}^2 \leq \sum_{i=i_0}^{i_1-1} \sum_{j} u_{ij}^2 \leq C n^{-2s_2/(2s_2+1)}.$$

We only need to concern ourselves with T_3'. Let

$$A_i = \{\text{blocks at level } i \text{ containing at least one coefficient } \beta_{ij} \text{ with indices in } S_i\};$$

$$A_i' = \{\text{blocks at level } i \text{ containing no coefficients } \beta_{ij} \text{ with indices in } S_i\}.$$

We split T_3' into several parts

$$T_3' = \sum_{i=i_0}^{i_s} \sum_{k} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) \sum_{(ik)} U_{ij}^2 \right\}$$

$$+ \sum_{i=i_s+1}^{i_1-1} \sum_{k \in A_i} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) I\left(B_{ik} > \delta_i/2\right) \sum_{(ik)} U_{ij}^2 \right\}$$

$$+ \sum_{i=i_s+1}^{i_1-1} \sum_{k \in A_i'} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) I\left(B_{ik} > \delta_i/2\right) \sum_{(ik)} U_{ij}^2 \right\}$$

$$+ \sum_{i=i_s+1}^{i_1-1} \sum_{k} E\left\{ I\left(\hat{B}_{ik} > \delta_i\right) I\left(B_{ik} \leq \delta_i/2\right) \sum_{(ik)} U_{ij}^2 \right\}$$

$$\equiv T_{31} + T_{32} + T_{33} + T_{34}.$$

Each of these is bounded by applying Lemma 49 and (10.11) and (10.12).

10.4. **Bound for T_4.** We now decompose T_4.

$$T_4 \leq \sum_{i=i_0}^{i_1-1} \sum_{k\in A_i} P\left(\hat{B}_{ik} \leq \delta_i \text{ and } B_{ik} \geq 2\delta_i\right) \sum_{(ik)} \beta_{ij}^2$$

$$+ \sum_{i=i_0}^{i_s} \sum_{k\in A_i'} P\left(\hat{B}_{ik} \leq \delta_i \text{ and } B_{ik} \geq 2\delta_i\right) \sum_{(ik)} \beta_{ij}^2$$

$$+ \sum_{i=i_0}^{i_s} \sum_{k} P\left(\hat{B}_{ik} \leq \delta_i \text{ and } B_{ik} < 2\delta_i\right) \sum_{(ik)} \beta_{ij}^2$$

$$+ \sum_{i=i_s+1}^{i_1-1} \sum_{k\in A_i} P\left(\hat{B}_{ik} \leq \delta_i \text{ and } B_{ik} < 2\delta_i\right) \sum_{(ik)} \beta_{ij}^2$$

$$+ \sum_{i=i_s+1}^{i_1-1} \sum_{k\in A_i'} P\left(\hat{B}_{ik} \leq \delta_i\right) \sum_{(ik)} \beta_{ij}^2$$

$$\equiv T_{41} + T_{42} + T_{43} + T_{44} + T_{45}.$$

Each of these is bounded by using Proposition 49 and Proposition 50.

11. Important notes about this paper.

This is a paper that describes a method for dealing with data that is long memory dependent and equispaced. To do this, it uses coiflets, a special kind of wavelets with high vanishing moments for the scaling function and the wavelets. This assumption is key to bounding the MISE and provides the bounds in Lemma 49. Also, this paper uses block thresholding, not component-wise.

Part 4. Summary of the work of Hall, Turlach and Berwin in [11].

12. Preliminaries and Notations.

In this paper we are dealing with irregularly spaced data. There are many different ways to deal with this problem. This paper reassigns the data set by performing a linear interpolation on the data. From this linear interpolation, new equispaced data points are drawn and used to interpolate the function.

Let $\mathcal{Y} = \{(X_m, Y_m), 1 \leq m \leq n\}$ be generated by the model $Y(X_m) = g(X_m) + \xi_m$ for $1 \leq m \leq n$, where the design sequence $\mathcal{X} = \{X_m, 1 \leq m \leq n\}$ represents the ordered values of a random sample from a distribution with density f having support $\mathcal{J} = [0, 1]$, and the ξ_m's are independent but not necessarily identically distributed random variables. We could reorder this data set according to the

X_m's. We may denote the reordered data as (X_{mn}, Y_{mn}) since the rank of the data point depends on how many points n there are in the set, but usually drop this notation and assume the data are in order.

Let w_m, $1 \le m \le n$, denote weight functions depending on the X_m's but not on the Y_m's and such that for integers ν_1, ν_2 satisfying $\nu_1 < 0 \le \nu_2$ we have $w_j = 0$ unless $\nu_1 \le j \le \nu_2$. Define the interpolant

$$Y(x) = \sum_m w_m(x) Y_m \quad \text{for } x \in (X_{-\nu_1}, X_{n-\nu_2}]\,. \tag{12.1}$$

This paper uses horizontal extrapolation on the other intervals, $Y(t) \equiv Y(X_{-\nu_1})$ on $[0, X_{-\nu_1}]$ and $Y(t) \equiv Y(X_{n-\nu_2})$ on $(X_{n-\nu_2}, 1]$. We consider two rules, local averaging and local linear interpolation. Assuming $x \in (X_l, X_{l+1}]$, in the first rule we define $w_m(x) = (2v)^{-1}$ if $-v+1 \le m-l \le v$, and $w_m(x) = 0$ otherwise. In the second rule

$$w_m(x) = \begin{cases} v^{-1}(X_{2l-m+1} - x)/(X_{2l-m+1} - X_m)\,, & \text{if } -v+1 \le m-1 \le 0, \\ v^{-1}(x - X_{2l-m+1})/(X_m - X_{2l-m+1})\,, & \text{if } 1 \le m-l \le v \\ 0, & \text{otherwise.} \end{cases}$$

Substituting these weights into (12.1) we obtain for $x \in (X_l, X_{l+1}]$

$$Y(x) = (2v)^{-1} \sum_{m=-v+1}^{v} Y_{l+m} \tag{12.2}$$

$$Y(x) = v^{-1} \sum_{m=1}^{v} \left(\frac{x - X_{l-m+1}}{X_{l+m} - X_{l-m+1}} Y_{l+m} + \frac{X_{l+m} - x}{X_{l+m} - X_{l-m+1}} Y_{l-m+1} \right). \tag{12.3}$$

We write ϕ and ψ for the mother and father wavelets respectively. Let $p = p(n)$ be the primary resolution level and define $p_i = 2^i p$ for $i \ge 0$ and let $\phi_j(x) = p^{1/2}\phi(px+j)$ and $\psi_{ij}(x) = p_i^{1/2}\psi(p_i x + j)$ be the functions that form the orthonormal basis of a wavelet expansion. Put $b_j = \int_{\mathcal{J}} g\phi_j$ and $b_{ij} = \int_{\mathcal{J}} g\psi_{ij}$. We assume ψ is of order r, meaning that $r \ge 1$ is the smallest integer such that $\int x^i \psi(x) dx$ is nonzero. Our estimators of b_j and b_{ij} are $\hat{b}_j = \int_{\mathcal{J}} Y\phi_j$ and $\hat{b}_{ij} = \int_{\mathcal{J}} Y\psi_{ij}$ respectively. This leads to the empirical wavelet transform

$$\hat{g} = \sum_j \hat{b}_j \phi_j + \sum_{i=0}^{q-1} \sum_j \hat{b}_{ij} I\left(\left|\hat{b}_{ij}\right| \ge \delta\right) \psi_{ij}. \tag{12.4}$$

We may approximate these to arbitrary accuracy on a dyadic grid. Taking $N = 2^k$ for an integer $k \geq 1$, we may define

$$\tilde{b}_j = N^{-1} \sum_{m=1}^{N} Y(m/N)\phi_j(m/N) \quad \text{and} \quad \tilde{b}_{ij} = N^{-1} \sum_{m=1}^{N} Y(m/N)\psi_{ij}(m/N).$$

The resulting estimator is $\tilde{g}$, which we obtain by replacing $\hat{b}_j$ and $\hat{b}_{ij}$ with $\tilde{b}_j$ and $\tilde{b}_{ij}$ respectively.

What we are doing is replacing the given data set $\mathcal{Y}$ with $\mathcal{Y}' = \{m/N, Y(m/N), 1 \leq m \leq N\}$. Provided that $N/n \to \infty$, the first-order asymptotics of $\tilde{g}$ are identical to $\hat{g}$.

13. Conditions (C).

Assume that g has r piecewise continuous derivatives, in the sense that there exist constants $0 = a_1 < a_2 < ... < a_k = 1$ such that g has r continuous derivatives on each interval (a_l, a_{l+1}) for $1 \leq l \leq k-1$, with left and right-hand limits at a_l and a_{l+1}, respectively. We assume of f that is is piecewise continuous in this sense, possibly with a different k and different a_i's, and it is bounded away from zero on $\mathcal{J} = [0, 1]$. We assume of ϕ and ψ that they are compactly supported and Holder continuous. Then for some $r \geq 1$, $\kappa \neq 0$, all integers $i \in [0, r]$ and $j \in (-\infty, \infty)$,

$$\int \psi^2 = 1, \qquad \int x^i \psi(x) dx = \kappa \left(r!\right)^{-1} \delta_{ir,}$$

$$\int \phi = 1, \qquad \int \phi(x)\phi(x + j)dx = \delta_{0j,}$$

where δ_{jk} is the Kronecker delta. Assume of the tuning parameters p, p_i and q in the definition of $\hat{g}$, that for some $u > 0$ and $\varepsilon > 0$

$$p^{-1} = o\left\{\left(n^{-1} \log n\right)^{1/(2r+1)}\right\}, \quad p_q^{-1} = o\left(n^{-2r/(2r+1)}\right), \quad p_q = O\left(n^{\min(u+1/(2r+1),1)-\varepsilon}\right); \tag{13.1}$$

and of errors $\xi_m = \xi_{nm}$ that they may be written as $\xi_{mn} = \sigma\left(X_{nm}\right)\xi'_{nm}$, where σ is a piecewise-continuous function on $\mathcal{J}$, $\xi'_{1m}, ..., \xi'_{nm}$ are stochastically independent of one another and of $X_1, ..., X_n$, and each ξ'_{nm} has for $1 \leq m \leq n < \infty$, the distribution of ξ', with $E\left(\xi'\right) = 0$, $E\left(\xi'^2\right) = 1$, $E\left|\xi'\right|^{2(1+u)} < \infty$ and u as in (13.1).

Note that these assumptions are equivalent to assuming what space the function g is contained in.

14. The main results.

Note that condition 13.1 and the assumption $E\left|\xi'\right|^{2(1+u)} < \infty$ are satisfied if we take p equal to a constant multiple of $n^{1/(2r+1)}$, q equal to the integer part of $(1-\varepsilon)\log_2 n$ for some $0 < \varepsilon < 2r/(2r+1)$, and $u = 2r/(2r+1)$; and if $E\left|\xi'\right|^4 < \infty$. This is clear by substitution. It will follow from the next theorem that p is optimal.

We assume the interpolation rule is given by either (12.2) or (12.3). Many other approaches could be used. In the case of the rule at (12.2), put $d_v \equiv 1 + (2v)^{-1}$, and for the rule at (12.3), let

$$d_v \equiv (2v)^{-2} E\left\{\sum_{l=-v}^{-1} Z_l \left(2\sum_{r=2l+1}^{l-1} Z_r + Z_l\right)\left(\sum_{r=2l+1}^{-1} Z_r\right)^{-1} + \sum_{l=0}^{v-1} Z_l \left(2\sum_{r=l+1}^{2l} Z_r + Z_l\right)\left(\sum_{r=0}^{2l} Z_r\right)^{-1}\right\}^2,$$

where $\{Z_r, -\infty < r < \infty\}$ are independent exponentially distributed random variables. For this definition, $d_1 = 3/2$ and $d_v = 1 + O\left(v^{-1}\right)$ as $v \to \infty$.

These bounds come directly from the fact that the Z_i are exponential variables.

We construct $\hat{g}$ using parameters p, q satisfying (13.1), and employing the threshold $\delta = \left(Dn^{-1}\log n\right)^{1/2}$, where the constant D satisfies

$$D > 2ud_v \sup\left(\sigma^2/f\right) \tag{14.1}$$

and u is as in (13.1). Define $D_1 = d_v \int \sigma^2 f^{-1}$ and $D_2 = \kappa^2\left(1-2^{-2v}\right)^{-1}\int\left(g^{(r)}\right)^2$.

Examining the proof closely we can see that these constants come from bounding each interval of the MISE.

Theorem 51. *Under conditions (C),*

$$\int E\left(\hat{g}-g\right)^2 = D_1 n^{-1}p + D_2 p^{-2r} + o\left(n^{-1}p + p^{-2r}\right). \tag{14.2}$$

15. Outline proof of Theorem 51.

The authors give an outline of the proof only in the case where f is uniformly continuous on $\mathcal{J} = [0,1]$ and the function σ^2 is a constant. Recall that f is the density function of the X_m's, the g is the function that we are trying to estimate. The authors say an additional argument would take care of the jump discontinuities in f and a varying σ^2. As in the Li and Xiao paper [18],

$$\int (\hat{g}-g)^2 = A_1 + A_2 + A_3 + A_4,$$

where

$$A_1 = \sum_j \left(\hat{b}_j - b_j\right)^2, \qquad A_2 = \sum_{i=0}^{q-1} \sum_j \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(\left|\hat{b}_{ij}\right| > \delta\right),$$

$$A_3 = \sum_{i=0}^{q-1} \sum_j b_{ij}^2 I\left(\left|\hat{b}_{ij}\right| \leq \delta\right), \qquad A_4 = \sum_{i=q}^{\infty} \sum_j b_{ij}^2.$$

15.1. **Moderate deviations.** Let $v_1, ..., v_n$ denote weights, which we shall take to be nonrandom, and suppose that for some $0 \leq \epsilon_1 < 1/20$ they satisfy

$$|v_n| \leq C_1 n^{\epsilon_1}, \qquad n^{-1} \sum_{m=1}^{n} v_m^2 \geq C_2 > 0. \tag{15.1}$$

Let $\xi_1, ..., \xi_n$ be independent and identically distributed random variables satisfying

$$E(\xi_1) = 0, \qquad E\left(\xi_1^2\right) = \sigma^2 > 0, \qquad E|\xi_1|^{C_3+2} \leq C_4 \tag{15.2}$$

for some $C_3 > 4\epsilon_1/(1-2\epsilon_1)$. Define $S_n = n^{-1/2} \sum_m v_m \xi_m$ and $\tau^2 = n^{-1}\sigma \sum_m v_m^2$.

We will need the following lemma.

Lemma 52. *Assume (15.1) and (15.2). Then for each $\epsilon_2 > 0$ there exist $C_5, C_6 > 0$ depending on ϵ_1, ϵ_2 and $C_1, ..., C_4$, such that*

$$E\left\{S_n^2 I(S_n \geq z)\right\} \leq C_5 \left[\exp\left\{-(1-\epsilon_2) z^2/\left(2\tau^2\right)\right\} + n^{2\epsilon_1+\epsilon_2-C_3(1/2-\epsilon_1)}\right]$$

uniformly in $0 \leq z \leq n^{C_6}$ for all n.

Proof. This is proved by citing Bernstien's and Bennett's inequalities, both of which of available in Hoeffding (1963). To state these results, let $Z_1, ..., Z_n$ denote independent random variables with zero means and satisfying $|Z_m| \leq b < \infty$ for each $1 \leq m \leq n$. Put $\tau'^2 \equiv n^{-1} \sum_m E\left(Z_m^2\right)$ and $T_n \equiv n^{-1/2} \sum_m Z_m$, and for $z > 0$ define $\eta = \eta(z) = bz/\left(n^{1/2}\tau'^2\right)$. Then

$$P(T_n > z) \leq \exp\left(-\frac{1}{2} b^{-2} z^2\right) \quad \text{for all } z \geq 0, \tag{15.3}$$

$$P(T_n > z) \leq \exp\left[-\left(n^{1/2} z/b\right)\left\{\left(1+\eta^{-1}\right)\log(1+\eta) - 1\right\}\right] \quad \text{for all } 0 \leq z \leq b. \tag{15.4}$$

By rearranging the constants we obtain the required result. □

15.2. **The wavelet coefficients.** Observe that $\hat{b}_{ij} = b_{ij} + B_{ij} + \hat{\xi}_{ij}$, where $B_{ij} = \int_J \Delta\psi_{ij}$, $\hat{\xi}_{ij} = n^{-1/2} S_{ij}$,

$$S_{ij} \equiv (p_i/n)^{1/2} \sum_m v_{ij;m}\xi_m, \qquad \tau_{ij}^2 \equiv p_i n^{-1} \sum_{m=v+1}^{n-v} v_{ij;m}^2,$$

$\Delta \equiv E(Y|\mathcal{X}) - g$ and $v_{ij;m} \equiv \left(n/p_i^{1/2}\right) \int w_m \psi_{ij}$. This expresses the relationship between the weights and the scaling properties of wavelet functions. If one observes the definition of $v_{ij;m}$, one can see that this is so. We are really examining the error ξ_m adjusted by a constant derived from the support of ψ_{ij}. Properties of spacings of order statistics are used to prove that

$$E\left(|B_{ij}|^k\right) = \begin{cases} O\left\{\left(p_i^{1/2}/n\right)^k n^\eta\right\}, & \text{uniformly in } j \in \mathcal{J}_i(\epsilon), \\ O\left(n^{\eta-k}\right), & \text{uniformly in } j \notin \mathcal{J}_i(\epsilon). \end{cases} \tag{15.5}$$

This set is just a small neighborhood of the integer i. One can apply Lemma 52 to show that if $E|\xi_l|^{2(1+t)+\eta} < \infty$ for some $t, \eta > 0$, then for each $\epsilon > 0$ and for the interpolation rules we have specified,

$$\sup_{i,j} E\left(S_{ij}^2 I\left[|S_{ij}| > \left\{2t(1+\epsilon)d_v\sigma^2\left(\sup f^{-1}\right)\log n\right\}^{1/2}\right]\right) = O\left(n^{-t}\right). \tag{15.6}$$

We can check this for each pair (i, j) by substitution.

15.3. **Calculation of $E(A_1)$.** Define $v_{j;m} \equiv \left(n/p^{1/2}\right) \int w_m \phi_j$ and also

$$B_j \equiv \int_{\mathcal{J}} \Delta\phi_j, \quad \hat{\xi}_j \equiv n^{-1/2} S_j, \quad S_j \equiv (p_i/n)^{1/2} \sum_m v_{j;m}\xi_m.$$

Then $\hat{b}_j = b_j + B_j + \hat{\xi}_j$. Let $[-c, c]$ be a compact interval containing the support of ψ, and let $\mathcal{J}(\epsilon)$ denote the set of indexes j such that, for some x that is a point of discontinuity of g, $px + j \in \left(-c - pn^{\epsilon-1}, c + pn^{\epsilon-1}\right)$. Then the analogue of (15.5) for B_j is, for all $\epsilon, \eta > 0$,

$$E\left(|B_j|^k\right) = \begin{cases} O\left\{\left(p^{1/2}/n\right)^k n^\eta\right\}, & \text{uniformly in } j \in \mathcal{J}(\epsilon), \\ O\left(n^{\eta-k}\right), & \text{uniformly in } j \notin \mathcal{J}(\epsilon). \end{cases}$$

Hence, for all $\epsilon, \eta > 0$,

$$E\left(\hat{b}_j - b_j\right)^2 = \begin{cases} O\left\{E\left(\hat{\xi}_j^2\right) + pn^{\eta-2}\right\}, & \text{uniformly in } j \in \mathcal{J}(\epsilon), \\ O\left(n^{\eta-k}\right), & \text{uniformly in } j \notin \mathcal{J}(\epsilon). \end{cases} \tag{15.7}$$

Take the following $\sup_{(1)}$ over j such that $\mathcal{J}_j = (-(c+j)/p, (c-j)/p) \subseteq \mathcal{J}$, and the $\sup_{(2)}$ over j such that $\mathcal{J}_j \cap \mathcal{J}$ is nonempty. It may be proved that

$$\sup_{(1)} \left| E\left(\hat{\xi}_j^2\right) - n^{-1}\sigma^2 d_v f(-j/p)^{-1} \right| = o\left(n^{-1}\right)$$

and

$$\limsup \sup_{(2)} E\left(\hat{\xi}_j^2\right) \leq n^{-1}\sigma^2 d_v \sup f^{-1}.$$

Combining the results from (15.7) down we conclude for all $\eta > 0$

$$E(A_1) = \sum_j E\left(\hat{b}_j - b_j\right)^2 \sim n^{-1} p \sigma^2 d_v \int f^{-1}. \tag{15.8}$$

15.4. **Bound for $E(A_2)$.** Let $\mathcal{K}_{i1}$ denote the set of indexes j that are contained in an interval $(p_i x - 2c, p_i x + 2c)$ for at least one of the discontinuity points x of at least one of the functions $g^{(0)}, ..., g^{(r)}$, and let $\mathcal{K}_{i2}$ be the set of all other j's. Write $A_2 = A_{21} + A_{22}$, where

$$A_{2k} = \sum_{i=0}^{q-1} \sum_{j \in \mathcal{K}_{ik}} \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(\left|\hat{b}_{ij}\right| > \delta\right).$$

From the formula $\hat{b}_{ij} = b_{ij} + B_{ij} + \hat{\xi}_{ij}$ we may prove that for $\eta > 0$,

$$E(A_{21}) = O\left\{ q \sup_{0 \leq i \leq q-1, j \in \mathcal{K}_{ik}} E\left(\hat{b}_{ij} - b_{ij}\right)^2 \right\} = O\left(q n^{\eta - 1}\right);$$

Note that there are q data points, and then apply (15.7).

By applying (15.9) to bound $E\left\{\hat{\xi}_{ij}^2 I\left(\left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right)\right\}$ and $P\left(\left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right)$, and assuming that the threshold satisfies

$$\delta > \left\{2t(1+\epsilon')d_v\sigma^2 \left(\sup f^{-1}\right) n^{-1} \log n\right\}^{1/2} \tag{15.9}$$

for some $\epsilon' > 0$, and that $E\left|\xi_1\right|^{2(1+t)+\epsilon''}$ for some $\epsilon'' > 0$; then

$$E(A_{22}) = O\left(\sum_{i=0}^{q-1} p_i n^{-(t+1)}\right) = O\left(p_q n^{-(t+1)}\right).$$

Combining these bounds we have that

$$E(A_2) = O\left(q n^{-1} + n^{\eta - 1} + p_q n^{-(t+1)}\right). \tag{15.10}$$

15.5. **Calculation of $E(A_3)$.** Write $E(A_3) = E(A_{31}) + E(A_{32})$, where

$$A_{3k} \equiv \sum_{i=0}^{q-1} \sum_{j \in \mathcal{K}_{ik}} b_{ij}^2 I\left(\left|\hat{b}_{ij}\right| \leq \delta\right).$$

Since $b_{ij}^2 \leq 2\left\{\left(\hat{b}_{ij} - b_{ij}\right)^2 + \hat{b}_{ij}^2\right\}$, and since the number of elements of $\mathcal{K}_{i1}$ is uniformly bounded, then we have for all $\eta > 0$,

$$E(A_{31}) = O\left[\sum_{i=0}^{q-1}\left\{\sup_{j \in \mathcal{K}_{i1}} E\left(\hat{b}_{ij} - b_{ij}\right)^2 + \delta^2\right\}\right] = O\left(qn^{\eta-1}\right).$$

Now, $b_{ij} = \kappa p_i^{-(2r+1)/2} g^{(r)}(-j/p_i) + o\left(p_i^{-(2r+1)/2}\right)$ since $g^{(r)}$ is piecewise continuous. Therefore, $E(A_{32}) = \kappa^2\left(1 - 2^{-2r}\right)^{-1} p^{-2r} \int \left(g^{(r)}\right)^2 + o\left(p^{-2r}\right)$. Combining these results we deduce that

$$E(A_3) = \kappa^2\left(1 - 2^{-2r}\right)^{-1} p^{-2r} \int \left(g^{(r)}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}\right). \tag{15.11}$$

15.6. **Bound for $E(A_4)$.** Divide the series into two portions, $E(A_4) = E(A_{41}) + E(A_{42})$, where

$$A_{4k} \equiv \sum_{i=q}^{\infty} \sum_{j \in \mathcal{K}_{ik}} b_{ij}^2.$$

Since $|b_{ij}| = O\left(p_i^{-1}\right)$ uniformly in $j \in \mathcal{K}_{i1}$, and the number of such j's is uniformly bounded, then $A_{41} = O\left(\sum_{i \geq q} p_i^{-1}\right) = O\left(p_q^{-1}\right)$. Furthermore, $|b_{ij}| = O\left(p_i^{-(2r+1)/2}\right)$ uniformly in $j \in \mathcal{K}_{i2}$, and the number of such j's for which b_{ij} does not vanish equals $O(p_i)$. Hence, $A_{42} = O\left(\sum_{i \geq q} p^{-2r}\right) = O\left(p_q^{-2r}\right)$. Combining these bounds we deduce that

$$A_4 = O\left(p_q^{-1}\right). \tag{15.12}$$

16. Conclusion.

Combining (15.8), (15.10), (15.11), and (15.12), we deduce that for all $\eta > 0$,

$$\int E(\hat{g} - g)^2 = D_1 n^{-1} p + D_2 p^{-2r} + o\left(n^{-1}p + p^{-2r}\right) \tag{16.1}$$

$$+ O\left(p_q n^{-(t+1)} + qn^{\eta-1} + p_q^{-1}\right).$$

By taking $t = u - \zeta$, where $u > 0$ is an in condition (13.1) and $\zeta > 0$ is sufficiently small, we see from conditions (C) imposed in Theorem 51 that $E|\xi_1|^{2(1+t)+\eta} < \infty$ for some $\eta > 0$ and that (15.9) holds

for some $\epsilon' > 0$. Furthermore, for such a t it follows from 13.1 that the $O(...)$ remainder term on the right-hand side of (16.1) equals $o\left(n^{-2r/(2r+1)}\right)$, and so may be incorporated into the $o\left(n^{-1}p + p^{-2r}\right)$ term. Result 14.2 follows immediately.

17. Important notes about this paper.

This paper uses a similar decomposition of the MISE as [18]. The bounds of these pieces follow directly from the assumptions about the space that the function g comes from, namely the Conditions (C) outlined in Section 13. The unique thing about this work is the linear interpolation used to deal with the irregularly spaced data. No long memory error is present in this paper. In Part 7 we will attempt to combine the analysis of long memory error from [18] with the linear interpolation of [11].

Part 5. Summary of the work of Antoniadis and Fan in [1].

18. Preliminaries and Notations.

In this part we study the work of Antoniadis and Fan in [1]. This paper deals with an incomplete set of dyadic data. A system of equations governing the wavelet coefficients can be found using this incomplete data. Suppose we have noisy data at irregular design points $\{t_1, ..., t_n\}$:

$$Y_i = f(t_i) + \epsilon_i, \qquad \epsilon_i \sim N\left(0, \sigma^2\right)$$

where the ϵ_i are identically and independently distributed and the f is an unknown regression to be estimated from the noisy sample. Assume f is defined on $[0, 1]$. Assume further that $t_i = n_i/2^J$ for some n_i and some resolution level J. Let $\mathbf{f}$ be the underlying regression function collected at all dyadic points $\left\{i/2^J,\ i = 0, ..., 2^J - 1\right\}$. Let $\mathbf{W}$ be a given wavelet transform and θ=$\mathbf{Wf}$ be the wavelet transform of $\mathbf{f}$. Because $\mathbf{W}$ is an orthogonal matrix, $\mathbf{f} = \mathbf{W}^T\theta$.

For $\mathbf{f}$ in the Besov space, the wavelet representation is sparse. The unknown signals are modeled by $N = 2^J$ parameters. This model is over-parametrized.

Denote the sampled data vector by $\mathbf{Y}_n$. Let $\mathbf{A}$ be $n \times N$ matrix whose ith row corresponds to the row of the matrix $\mathbf{W}^T$ for which the signal $f(t_i)$ is sampled with noise. We express the observed data as

$$\mathbf{Y}_n = \mathbf{A}\theta + \epsilon, \qquad \epsilon \sim N\left(0, \sigma^2 I_n\right) \tag{18.1}$$

where ϵ is the noise vector. We wish to minimize

$$2^{-1}\left\|\mathbf{Y}_n - \mathbf{A}\theta\right\|^2 + \lambda \sum_{i=1}^{N} p\left(|\theta_i|\right) \tag{18.2}$$

for a given penalty function p and a regularization parameter $\lambda > 0$. The penalty function is usually nonconvex on $[0, \infty)$ and irregular at point zero to produce sparse solutions.

19. Regularization of Wavelet Approximations.

19.1. **Regularized Wavelet Interpolations.** Assume for this section that the signals are observed with no noise $\epsilon = 0$. Being given signals only at the nonequispaced points $\{t_i, i = 1, ..., n\}$ necessarily means that we have no information at other dyadic points. Let

$$\mathbf{f}_n = \left(f\left(t_1\right), ..., f\left(t_n\right)\right)^T$$

be the observed signals. Then

$$\mathbf{f}_n = \mathbf{A}\theta. \tag{19.1}$$

This is an underdetermined system of equations, so there are many different solutions for θ. For the minimum Sobolev solution, we chose the $\mathbf{f}$ that interpolates the data and minimizes the weighted Sobolev norm of f.

$$\|\theta\|_s^2 = \sum_j 2^{2sj} \|\theta_{j\cdot}\|^2 \tag{19.2}$$

where $\theta_{j\cdot}$ is the vector of the wavelet coefficients at the resolution level j. We can then restate the problem as a wavelet-domain optimization problem: Minimize $\|\theta\|_s^2$ subject to the constraint (19.1). The solution (Rao 1973) is what is called the normalized method of frame whose solution is given by

$$\theta = \mathbf{D}\mathbf{A}^T \left(\mathbf{A}\mathbf{D}\mathbf{A}^T\right)^{-1} \mathbf{f}_n,$$

where $\mathbf{D} = \text{Diag}\left(2^{-2sj_i}\right)$ with j_i denoting the resolution level with which θ_i is associated.

19.2. **Regularized Wavelet Estimators.** The traditional regularization problem can be formulated in the wavelet domain as follows. Find the minimum of

$$2^{-1} \|\mathbf{Y}_n - \mathbf{A}\theta\|^2 + \lambda \|\theta\|_s^2. \tag{19.3}$$

We could replace the Sobolev norm with other penalty functions, leading to minimizing

$$l(\theta) = 2^{-1} \|\mathbf{Y}_n - \mathbf{A}\theta\|^2 + \lambda \sum_{i \geq i_0} p\left(|\theta_i|\right) \tag{19.4}$$

for a given penalty function $p(\cdot)$ and given value i_0. To facilitate the discussion, we change the notation θ_{jk} from a double array sequence into a single array sequence θ_i.

19.3. **Penalty Functions and Nonlinear Wavelet Estimators.** If $n = 2^J$, then $\mathbf{A}$ becomes the inverse wavelet transform matrix $\mathbf{W}^T$. In this case, (19.4) becomes

$$2^{-1}\sum_{i=1}^{n}(z_i-\theta_i)^2+\lambda\sum_{i\geq i_0}p\left(|\theta_i|\right), \tag{19.5}$$

where z_i is the ith component of the wavelet coefficient vector $\mathbf{z} = \mathbf{W}\mathbf{Y}_n$.

These regularized wavelet estimators are extensions of the soft and hard thresholding rules of Donoho and Johnstone. Let p_λ denote λp. For the L_1 penalty

$$p_\lambda\left(|\theta|\right)=\lambda|\theta|, \tag{19.6}$$

the solution is the soft-thresholding rule. When the penalty function is given by

$$p_\lambda\left(|\theta|\right)=\lambda^2-\left(|\theta|-\lambda\right)^2 I\left(|\theta|<\lambda\right), \tag{19.7}$$

the solution is the hard-thresholding rule. This can be confirmed by putting the penalty into $l(\theta)$ and then finding the derivative. Many other penalties are suggested.

20. Oracle Inequalities and Universal Thresholding.

20.1. **Characterization of Penalized Least Squares Estimators.** Let $p(\cdot)$ be a nonnegative, nondecreasing, and differentiable function on $(0,\infty)$. Minimize with respect to θ

$$l(\theta)=(z-\theta)^2/2+p_\lambda\left(|\theta|\right) \tag{20.1}$$

for a given penalty parameter λ. This is a component-wise minimization problem of (19.5). Note that the function above tends to infinity as $|\theta|\to\infty$. Thus, minimizers do exist. Let $\hat{\theta}(z)$ be a solution. Then we have the following.

Theorem 53. *Let $p_\lambda(\cdot)$ be a nonnegative, nondecreasing, and differentiable function in $(0,\infty)$. Further, assume that the function $-\theta-p'_\lambda(\theta)$ is strictly unimodal on $(0,\infty)$. Then we have the following results.*

1. The solution to the minimization problem (20.1) exists and is unique. It is antisymmetric: $\hat{\theta}(-z)=-\hat{\theta}(z)$.

2. The solution satisfies

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

where $p_0=\min_{\theta\geq 0}\{\theta+p'_\lambda(\theta)\}$. Moreover, $\left|\hat{\theta}(z)\right|\leq|z|$.

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

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

4. When $p'_\lambda(\theta)$ is continuous on $(0, \infty)$, the solution $\hat{\theta}(z)$ is continuous if and only if the minimum of $|\theta| + p'_\lambda(|\theta|)$ is attained at point zero.

5. If $p'_\lambda(|z|) \to 0$, as $|z| \to +\infty$, then

$$\hat{\theta}(z) = z - p'_\lambda(|z|) + o(p'_\lambda(|z|)).$$

We examine the implications of these results. When $p'_\lambda(0+) > 0$, $p_0 > 0$. Thus, for $|z| \leq p_0$, the estimate is thresholded to 0. For $|z| > p_0$, the solution has a shrinkage property. The amount of shrinkage is sandwiched between the soft and hard thresholding estimators, as we see from result 3. Recall that we are expanding the soft and hard thresholding operators to an entire family of penalty functions. Also note that a different estimator $\hat{\theta}$ may require a different p_0. The amount of shrinkage tapers off as $|z|$ gets large when $p'_\lambda(|z|)$ goes to zero.

20.2. **Risks of Penalized Least Squares Estimators.** We now study the risk function of the penalized least squares estimator $\hat{\theta}$ that minimizes (20.1). Assume $Z \sim N(\theta, 1)$. Let

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

The thresholding parameter p_0 is equivalent to the regularization parameter λ. We have the following theorem.

Theorem 54. *Suppose p satisfies conditions in Theorem 53 and $p'_\lambda(0+) > 0$. Then*

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

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

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

3. $R_p(0, p_0) \leq \sqrt{2/\pi}\left(p_0 + p_0^{-1}\right)\exp\left(-p_0^2/2\right)$.

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

These four properties are comparable with the properties of soft and hard thresholding rules given in Donoho and Johnstone (1994). The improvement here is that these results hold for a larger class of penalty functions.

20.3. **Oracle Inequalities and Universal Thresholding.** Following Donoho and Johnston (1994), we have an ideal oracle estimator $\hat{\theta}_o = ZI(|\theta| > 1)$, which attains the ideal L_2-risk $\min\left(\theta^2, 1\right)$. Let n be the sample size.

When $p_0 = \sqrt{2\log n}$, the universal threshold, by property 3 of Theorem 54, we need to add an amount cn^{-1} for some constant c to the risk of the oracle estimator, because it has no risk at point $\theta = 0$.

Define

$$\Lambda_{n,c,p_0}(p) = \sup_{\theta} \frac{R_p(\theta, p_0)}{cn^{-1} + \min(\theta^2, 1)}$$

and denote $\Lambda_{n,c,p_0}(p)$ by $\Lambda_{n,c}(p)$ for the universal thresholding $p_0 = \sqrt{2\log n}$. Then $\Lambda_{n,c,p_0}(p)$ is a sharp risk upper bound for using the universal thresholding parameter p_0.

$$R_p(\theta, p_0) \le \Lambda_{n,c,p_0}(p)\left\{cn^{-1} + \min\left(\theta^2, 1\right)\right\}. \tag{20.2}$$

As in Donoho and Johnstone, we also define

$$\Lambda^*_{n,c}(p) = \inf_{p_0}\sup_{\theta} \frac{R_p(\theta, p_0)}{cn^{-1} + \min(\theta^2, 1)}$$

and

$$p_n \;=\; \text{the largest constant attaining } \Lambda^*_{n,c}(p).$$

Then, the constant $\Lambda^*_{n,c}(p)$ is the sharp risk upper bound using the minimax optimal thresholding p_n. Necessarily,

$$R_p(\theta, p_n) \le \Lambda^*_{n,c}(p_n)\left\{cn^{-1} + \min\left(\theta^2, 1\right)\right\}. \tag{20.3}$$

By Theorem 54, property 2, if $p_0 \le \sqrt{2\log n}$ we have

$$R_p(\theta, p_0) \le 2\log n + \sqrt{4/\pi}\,(\log n)^{1/2} + 1 \tag{20.4}$$

The extra $\log n$ term is necessary because thresholding estimators create biases of order p_0 at $|\theta| \approx p_0$. The risk in $[0, 1]$ can be bounded by using this lemma.

Lemma 55. *If the penalty function satisfies conditions of Theorem 53 and $p'_\lambda(\cdot)$ is nonincreasing and $p'_\lambda(0+) > 0$, then*

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

for the universal thresholding

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

with $n \ge 4$, $c \ge 1$ and $p_0 > 1.14$.

The authors suggest using $d = 256$ and $c = 16$. A table is given which demonstrates the effectiveness of this constant. It performs very well when compared to initial constants given by Donoho and Johnstone.

20.4. **Performance of Regularized Wavelet Estimators.** These oracle inequalities can be directly applied to the estimators defined via (19.4) when the sampling points are equispaced and $n = 2^J$. Suppose the data are collected from model (18.1) and $\sigma = 1$. Then the wavelet coefficients $\mathbf{Z} = \mathbf{W}\mathbf{Y}_n \sim N(\theta, I_n)$. Let

$$R_p\left(\hat{f}_p, f\right) = n^{-1}\sum_{i=1}^{n}\left\{\hat{f}_p(t_i) - f(t_i)\right\}^2$$

be the risk function of the regularized wavelet estimator $\hat{f}_p$. Let $R\left(\hat{f}_o, f\right)$ be the risk of the oracle wavelet thresholding estimator, which selects a term to estimate depending on the value of unknown wavelet coefficients. So, $\hat{f}_o$ is the inverse wavelet transform of he ideally selected wavelet coefficients $\{Z_i I(|\theta_i| > 1)\}$. This is an ideal estimator and serves as a benchmark for our comparison. We assume $i_o = 1$.

Theorem 56. *With the universal thresholding $p_0 = \sqrt{2\log n}$, we have*

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

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

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

This risk can be computed. Assume that the signal f is in a Besov Ball. We can characterize this space by its wavelet coefficients. Define the Besov space ball $B^r_{p,q}(C)$ as

$$B^r_{p,q}(C) = \left\{f \in L_p : \sum_j \left(2^{j(r+1/2-1/p)}\left\|\theta_{j\cdot}\right\|_p\right)^q < C\right\}, \tag{20.5}$$

where $\theta_{j\cdot}$ is the vector of wavelet coefficients at the resolution level j. Here, r indicates the degree of smoothness of the underlying signal f. Note here that the reason why the Besov space is such a natural way to extend the results in [1] is because its norm is written in terms of the wavelet coefficients.

Theorem 57. *Suppose the penalty function satisfies the conditions of Lemma 55 and $r + 1/2 - 1/p > 0$. Then the maximum risk of the penalized least squares estimator $\hat{f}_p$ over the Besov ball $B^r_{p,q}(C)$ is of rate $O\left(n^{-2r/(2r+1)}\log n\right)$ when the universal thresholding $\sqrt{2n^{-1}\log n}$ is used. It achieves the rate of convergence $O\left(n^{-2r/(2r+1)}\log n\right)$ when the minimax thresholding $p_n/\sqrt{n}$ is used.*

Thus, we always arrive at a risk which if within a factor of logarithmic order.

21. Penalized Least Squares for Nonuniform Designs.

21.1. **Regularized One-Step Estimator.** We take advantage of the orthonormality of the wavelet matrix $\mathbf{W}$. Let us again consider (18.1) and let us collect the remaining rows of the matrix $\mathbf{W}^T$ that were not collected into the matrix $\mathbf{A}$ into matrix $\mathbf{B}$ of size $(N-n)\times N$. Then the penalized least squares (19.4) can be written as

$$l(\theta) = 2^{-1}\left\|\mathbf{Y}^* - \mathbf{W}^T\theta\right\|^2 + \sum_{i\geq i_0} p_\lambda\left(|\theta_i|\right),$$

where $\mathbf{Y}^* = \left(\mathbf{Y}_n^T, (\mathbf{B}\theta)^T\right)^T$. By orthonormality

$$l\left(\theta\right) = 2^{-1}\left\|\mathbf{W}\mathbf{Y}^* - \theta\right\|^2 + \sum_{i\geq i_0} p_\lambda\left(|\theta_i|\right). \tag{21.1}$$

We could solve this iteratively.

We could use our Sobolev wavelet interpolators to produce an initial estimate for θ and hence for $\mathbf{Y}^*$. Recall that $\theta = \mathbf{D}\mathbf{A}^T\left(\hat{\mathbf{A}\mathbf{D}\mathbf{A}^T}\right)^{-1}\mathbf{Y}_n$. Let

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

be the initial synthetic data. We see that

$$\hat{\theta}^* = \mathbf{W}\hat{\mathbf{Y}}_0^* \sim N\left(\theta^*, \sigma^2\mathbf{V}\right), \tag{21.2}$$

where

$$\mathbf{V} = \mathbf{D}\mathbf{A}^T\left(\mathbf{A}\mathbf{D}\mathbf{A}^T\right)^{-2}\mathbf{A}\mathbf{D} \quad \text{and} \quad \theta^* = \mathbf{D}\mathbf{A}^T\left(\mathbf{A}\mathbf{D}\mathbf{A}^T\right)^{-1}\mathbf{A}\theta$$

is the vector of wavelet coefficients. Call the components of $\mathbf{W}\hat{\mathbf{Y}}_0^*$ the empirical synthetic wavelet coefficients. Let $\hat{\theta}_1^*$ be a component-wise thresholded $\hat{\theta}^*$. Then we could use

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

Then one could minimize

$$l\left(\theta\right) = 2^{-1}\left\|\mathbf{W}\mathbf{Y}_1^* - \theta\right\|^2 + \sum_{i\geq i_0} p_\lambda\left(|\theta_i|\right) \tag{21.3}$$

component-wise. This procedure is called the ROSE.

21.2. **Thresholding for Nonstationary Noise.** The variances for the wavelet coefficients are no longer identical. However, we do know their covariance matrix $\mathbf{V}$ up to a constant. If v_i is the ith diagonal element of the matrix $\mathbf{V}$. Then the ith synthetic wavelet coefficient, denoted Z_i^* is distributed

$$Z_i^* \sim N\left(\theta_i^*, v_i\sigma^2\right). \tag{21.4}$$

Apply the threshold

$$p_i = \sqrt{2v_i \log n}\sigma \tag{21.5}$$

to each coefficient Z_i^* (also known as λ_i). Then

$$E\left(\hat{\theta}_i - \theta_i^*\right)^2 \leq \left(2\log n + 2\log^{1/2} n\right) \times \left[c\sigma^2 v_i/n + \min\left(\theta_i^{*2}, \sigma^2 v_i\right)\right]. \tag{21.6}$$

21.3. **Sampling Properties.** Define

$$R_p(f) = n^{-1}\sum_{i=1}^{n} E\left\{\hat{f}_p\left(t_i\right) - f\left(t_i\right)\right\}^2.$$

In fact,

$$R_p(f) = n^{-1}E\left\{\left\|\mathbf{A}\hat{\theta}_1 - \mathbf{A}\theta\right\|^2\right\} = n^{-1}E\left\{\left\|\mathbf{A}\hat{\theta}_1 - \mathbf{A}\theta^*\right\|^2\right\} \leq n^{-1}E\left\|\hat{\theta}_1 - \theta^*\right\|^2. \tag{21.7}$$

By (21.6), the mean square errors are bounded as follows.

Theorem 58. *Assume that the penalty function* p satisfies the condition in Lemma 55. Then, the NRSI with coefficient-dependent thresholding satisfies

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

where $tr(\mathbf{V})$ is the trace of the matrix $\mathbf{V}$.

Lastly, we have the following convergence result.

Theorem 59. *Suppose that the penalty function satisfies the conditions of Lemma 55 and* $r+1/2-1/p > 0$. Then, the maximum risk of the nonlinear regularized Sobolev interpolator over a Besov ball $B_{p,q}^r$ is of rate $O\left(n^{-2r/(2r+1)}\log n\right)$ when the universal thresholding rule is used. It achieves the rate of convergence $O\left(n^{-2r/(2r+1)}\log n\right)$ when the minimax thresholding $p_n/\sqrt{n}$ is used.

22. Important notes about this paper.

This paper translates bounding the MISE into the wavelet setting. Here the problem is not irregularly spaced data, but data which is dyadic but incomplete. The authors address the problem of nonstationary noise and pave the way for expansion into long memory error. Also, the authors relate their results to those of Donoho and Johnstone in [8]. They write the MISE in terms of oracle risk, which makes it easy to relate the error to more general spaces.

Part 6. Summary of the work of Cai and Brown in [4].

23. Preliminaries and Notations.

In this part we study the work of Cai and Brown in [4]. Here we have irregularly spaced data. We will use a function H, to be defined later to reorder the data and make it equally spaced. Eventually we analyze the Mean Integrated Square Error. Suppose we are given data:

$$y_i = f(t_i) + \epsilon z_i \tag{23.1}$$

where $i = 1, 2, ..., n$, $0 < t_1 < t_2 < ... < t_n = 1$ and the z_i are independently and identically distributed as $N(0, 1)$. These data points are not equally spaced.

We wish to construct an estimate $\hat{f}$ which minimizes the risk

$$R\left(\hat{f}, f\right) = E \int_0^1 \left(\hat{f}(t) - f(t)\right)^2 dt.$$

The authors formulate the data as follows:

$$y_i = f(t_i) + \epsilon z_i$$

with $i = 1, 2, ..., n$ where $n = 2^J$. Here $t_i = H^{-1}(i/n)$ for some cumulative density function H on $[0, 1]$. These design points are assumed to be fixed, not drawn randomly from H.

A rough outline of the procedure this paper describes is recorded below.

(1) Precondition the data by a sparse matrix.
(2) Transform the preconditioned data by the discrete wavelet transform.
(3) Denoise the noisy wavelet coefficients via thresholding.
(4) Apply the inverse transform to the denoised coefficients.
(5) Postcondition the data by a matrix to get the estimate at the sample points.

We also need some standard properties of wavelets. Suppose the father and mother wavelets ϕ and ψ are compactly supported. Assume $\text{supp}(\phi) = \text{supp}(\psi) = [0, N]$. Also assume $\int \phi = 1$. A wavelet ψ is

r-regular if it has r vanishing moments and r continuous derivatives. Let

$$\phi_{jk}(t) = 2^{j/2}\phi\left(2^j t - k\right), \qquad \psi_{jk}(t) = 2^{j/2}\psi\left(2^j t - k\right)$$

and denote the periodized wavelets

$$\phi^p_{jk}(t) = \sum_{l\in\mathbb{Z}} \phi_{jk}(t-l), \qquad \psi^p_{jk}(t) = \sum_{l\in\mathbb{Z}} \psi_{jk}(t-l) \text{ for } t \in [0,1].$$

This paper uses these periodized wavelets as the basis for the paper. The collection $\left\{\phi^p_{j_0k},\ k = 1, ..., 2^{j_0};\ \psi^p_{jk},\ j \geq j\right.$ constitutes an orthonormal basis of $L^2[0,1]$. From now on the p will be suppressed for convenience.

Recall that wavelets have an associated multiresolution analysis. Let V_j and W_j be the closed linear subspaces generated by $\left\{\phi_{jk},\ k = 1, ..., 2^j\right\}$ and $\left\{\psi_{jk},\ k = 1, ..., 2^j\right\}$ respectively. Then:

(1) We have $V_{j_0} \subset V_{j_0+1} \subset ... \subset V_j \subset ...$

(2) Also $\overline{\cup_{j=j_0}^{\infty} V_j} = L^2\left([0,1]\right)$.

(3) We have $V_{j+1} = V_j \oplus W_j$.

For a given square integrable function f on $[0,1]$, denote

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

This function can be expanded into a wavelet series:

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{jk}\psi_{jk}(x). \tag{23.2}$$

The functions in this work belong to the following class.

Definition 60. A piecewise Holder class $\Lambda^\alpha(M, B, m)$ on $[0,1]$ with at most m discontinuous jumps consists of functions f satisfies the following conditions

1. The function f is bounded by B, that is, $|f| \leq B$.

2. There exist $l \leq m$ points $0 \leq \alpha_1 < ... < \alpha_l \leq 1$ such that, for all $\alpha_i \leq x,\ y < \alpha_{i+1},\ i = 0, 1, ..., l$ (with $\alpha_0 = 0$ and $\alpha_{l+1} = 1$),

(i) $|f(x) - f(y)| \leq M|x-y|^\alpha$ if $\alpha > 1$.

(ii) $\left|f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(y)\right| \leq M|x-y|^{\alpha'}$ and $|f'(x)| \leq B$ if $\alpha > 1$.

where $\lfloor\alpha\rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$.

This function class contains all functions which are piecewise Holder with the number of discontinuities bounded by m.

24. Important Lemmas.

We have the following bounds for wavelet coefficients in the Holder class.

Lemma 61. *Let $f \in \Lambda^\alpha(M,B,m)$. Suppose that the wavelet function ψ is r-regular with $r \geq \alpha$. Then:*

(i) If supp (ψ_{jk}) does not contain any jump points of f, then

$$\theta_{jk} \equiv |\langle f, \psi_{jk}\rangle| \leq C 2^{-j(1/2+\alpha)}. \tag{24.1}$$

(ii) If supp (ψ_{jk}) contains as least one jump point of f, then

$$\theta_{jk} \equiv |\langle f, \psi_{jk}\rangle| \leq C 2^{-j/2}. \tag{24.2}$$

We now have the following Lemma.

Lemma 62. *Suppose $f \in \Lambda^\alpha(M,B,m)$. Let $\xi_{Jk} = \langle f, \phi_{Jk}\rangle$ and $s(\alpha) = \min(\alpha, 1)$. Then:*

(i) If supp (ϕ_{Jk}) does not contain any jump points of f, then

$$\left| n^{-1/2} f(k/n) - \xi_{Jk} \right| \leq C n^{-(1/2+s(\alpha))}. \tag{24.3}$$

(ii) If supp (ϕ_{Jk}) contains jump points of the function f, then

$$\left| n^{-1/2} f(k/n) - \xi_{Jk} \right| \leq C n^{-1/2}. \tag{24.4}$$

This means that we can use $f_n(t) = \sum_{k=1}^n n^{-1/2} f(k/n)\phi_{Jk}(t)$ as an approximation of the true function f.

25. The nonequispaced procedure.

Suppose we observe the data $\{y_i\}$ as in (23.1) and we wish to recover the function f. Let $\tilde{g}(t) = n^{-1/2}\sum_{i=1}^n y_i \phi_{J_i}(t)$ and let

$$\tilde{f}_J(t) = \mathrm{Proj}_{V_j}\tilde{g}\left(H(t)\right) = n^{-1/2}\sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\tilde{\theta}_{jk}\psi_{jk}(t),$$

where

$$\tilde{\xi}_{jk} = n^{-1/2}\sum_{i=1}^n y_i \left\langle \phi_{Ji}\circ H, \phi_{jk}\right\rangle, \qquad \tilde{\theta}_{jk} = n^{-1/2}\sum_{i=1}^n y_i \left\langle \phi_{Ji}\circ H, \psi_{jk}\right\rangle. \tag{25.1}$$

These coefficients can be regarded as estimators of the true coefficients ξ_{j_0k} and θ_{j_0k}. The function H maps whatever points $H^{-1}(i/n)$ that are given to us into the data. We let

$$\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k}, \qquad \hat{\theta}_{jk} = \mathrm{sgn}\left(\tilde{\theta}_{jk}\right)\left(\left|\tilde{\theta}_{jk}\right| - \lambda_{jk}\right)_+. \tag{25.2}$$

We obtain a soft threshold estimator for f by using these coefficients.

$$\hat{f}_n^*(t) = n^{-1/2} \sum_{k=1}^{2^{j_0}} \hat{\xi}_{j_0 k} \phi_{j_0 k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{jk} \psi_{jk}(t). \tag{25.3}$$

Similarly, one can obtain a hard-thresholding operator by setting the coefficients as

$$\hat{\xi}_{j_0 k} = \tilde{\xi}_{j_0 k} \qquad \hat{\theta}_{j_0 k} = I\left(\left|\tilde{\theta}_{jk}\right| > \lambda_{jk}\right), \tag{25.4}$$

with the same threshold λ_{jk} as in (25.2).

Note that the coefficients $\hat{\xi}_{j_o k}$ contain the gross structure of f and so we do not threshold these coefficients.

26. Approximation.

Now the authors explain why the estimation method makes sense. Denote by $\Lambda^1(h)$ the collection of Lipschitz functions f satisfying

$$|f(x) - f(y)| \leq h\,|x - y| \text{ for } x, y \in [0,1].$$

Suppose we are given a sampled function $\left\{f(t_i),\ i = 1, 2, ..., n\left(= 2^J\right)\right\}$ with $t_i = H^{-1}(i/n)$, where H is a strictly increasing cumulative density function on $[0,1]$ and $H^{-1} \in \Lambda^1(h)$ for some constant h.

If the t_i were equispaced, it follows from Lemma 61 and Lemma 62 that $f_n(t) = \sum_{k=1}^n n^{-1/2} f(t_k)\,\phi_{Jk}(t)$ is a good approximation with no extra work involved. When the t_i are nonequispaced, one can first approximate $f\left(H^{-1}(t)\right)$ by $g_n(t) = \sum_{k=1}^n n^{-1/2} f(t_k)\,\phi_{Jk}(t)$, then use the projection of $g_n(H(t))$ onto the multiresolution space V_J as the approximation of f. More specifically, let

$$\xi_{Ji}^{'} = n^{-1/2} \sum_{k=1}^{2^J} f(t_k)\,\langle \phi_{Jk} \circ H, \phi_{Ji} \rangle \tag{26.1}$$

and let

$$f_n(t) = \sum_{i=1}^{2^J} \xi_{Ji}^{'} \phi_{Ji}(t) \tag{26.2}$$

be an approximation of f. Note $f_n \in V_j$. We have the following bound on the approximation error.

Theorem 63. *Suppose that a sampled function $\left\{f(t_i),\ i = 1, 2, ..., n\left(= 2^J\right)\right\}$ is given with $t_i = H^{-1}(i/n)$, where H is a strictly increasing cumulative density function on $[0,1]$ and $H^{-1} \in \Lambda^1(h)$. Let the wavelet*

function ψ be r-regular with $r > \alpha$. Let $\xi_{Ji}^{'}$ and f_n be given as in (26.1) and (26.2) respectively. Then the approximation error $\|f_n - f\|^2$ satisfies

$$\sup_{f \in \Lambda^\alpha(M,B,m)} \|f_n - f\|_2^2 = o\left(n^{-2\alpha/(1+2\alpha)}\right), \tag{26.3}$$

where the maximum number of jump discontinuities $m = Cn^\gamma$ with constants $C > 0$ and $0 < \gamma < 1/(1+2\alpha)$.

27. The Threshold.

We must know the noise levels of the coefficients before we can threshold them. The function H^{-1} is strictly increasing, so H^{-1} is differentiable almost everywhere. Denote by $\tilde{h}(t)$ the derivative of $H^{-1}(t)$. Then

$$0 < \tilde{h}(t) \leq h \quad \text{for almost all } t \in [0,1].$$

We can see from (25.1) that

$$\sigma_{jk}^2 = \text{var}\left(\tilde{\theta}_{jk}\right) = n^{-1}\epsilon^2 \sum_{i=1}^{n} \left(\langle \phi_{Ji} \circ H, \psi_{jk} \rangle\right)^2$$

$$\leq n^{-1}\epsilon^2 \int \psi_{jk}^2(t)\tilde{h}(H(t))dt \equiv u_{jk}^2. \tag{27.1}$$

This inequality is asymptotically sharp, $\sigma_{jk} \to u_{jk}$ as $n \to \infty$. We set the threshold

$$\lambda_{jk} = u_{jk}\left(2\log n\right)^{1/2}. \tag{27.2}$$

This is our threshold.

28. Optimality results.

We have the following results. The theorem below describes global rates of convergence.

Theorem 64. *SSuppose we observe $\left\{(t_i, y_i),\ i = 1, 2, ...n\left(= 2^J\right)\right\}$ as in (23.1) with $t_i = H^{-1}(i/n)$, where H is a strictly increasing cumulative density function on $[0,1]$ and $H^{-1} \in \Lambda^1(h)$. Let $\hat{f}_n^*$ be either the soft-thresholded or hard-thresholded wavelet estimator of f given in (25.3) and (27.2). Suppose that the wavelet function ψ is r-regular. Then the estimator $\hat{f}_n^*$ is near optimal:*

$$\sup_{f \in \Lambda^\alpha(M,B,m)} E\left\|\hat{f}_n^* - f\right\|_2^2 \leq C(\log n/n)^{2\alpha/(1+2\alpha)}(1 + o(1)) \tag{28.1}$$

for all $0 < \alpha < r$ and all $m \leq Cn^\gamma$ with constants $C > 0$ and $0 < \gamma < 1/(1+2\alpha)$.

The next theorem describes convergence rates at a single point.

Theorem 65. *For any fixed $t_0 \in [0,1]$, let $\hat{f}_n^*(t)$ be given as in (25.3) and (27.2). Under the conditions given in Theorem 64, we have*

$$\sup_{f \in \Lambda^\alpha(M,B,0)} E\left(\hat{f}_n^*(t_0) - f(t_0)\right)^2 \leq C(\log n/n)^{2\alpha/(1+2\alpha)}(1+o(1)) \tag{28.2}$$

for all $0 < \alpha < r$.

This theorem apples as long as the jump points are away from a fixed neighborhood of t_0.

29. Discussion.

We can choose a more convenient threshold as follows. In (27.2), we set the threshold $\lambda_{jk} = u_{ij}(2\log n)^{1/2}$, where $u_{jk} = \left(n^{-1}\epsilon^2 \int \psi_{jk}^2(t)\tilde{h}\left(H(t)\right)dt\right)^{1/2}$. We see that

$$u_{jk}^2 \leq n^{-1}\epsilon^2 h_{jk}, \tag{29.1}$$

where $h_{jk} = \sup\left\{\tilde{h}(t) : \ t \in \left[H^{-1}\left(2^{-j}k\right), H^{-1}\left(2^{-j}\left(k+N\right)\right)\right]\right\}$.

We may replace the threshold λ_{jk} by

$$\lambda_{jk}^{'} = \epsilon\left(2h_{jk}n^{-1}\log n\right)^{1/2}. \tag{29.2}$$

The optimality results from before still hold with this new threshold, which is easier to compute.

30. Proofs.

Here we have the proofs of the results of the paper.

Proof. Proof of Theorem 63.

Let $g(t) = f\left(H^{-1}(t)\right)$. Denote $s(\alpha) = \min(\alpha, 1)$ and $M \vee B = \max(M, B)$. We see that $g \in \Lambda^{s(\alpha)}(h^{s(\alpha)}M \vee B, B, m)$. Now $f_n = \text{Proj}_{V_j} g_n \circ H$. It follows from Lemmas 61 and 62 that

$$\begin{aligned}\|f_n - f\|_2^2 &\leq \left\|\text{Proj}_{V_j}\left(g_n \circ H - g \circ H\right)\right\|_2^2 + \left\|\text{Proj}_{V_j} f - f\right\|_2^2 \\ &\leq Cn^{-2s(\alpha)} + Cmn^{-1} = o\left(n^{-2\alpha/(1+2\alpha)}\right).\end{aligned}$$

□

We now preliminaries for proving Theorems 64 and 65. Let $y \sim N\left(\theta, \sigma^2\right)$ be a normal variable with known variance σ^2. We estimate the true mean θ with the thresholding operator. Let $\lambda = a\sigma$ with $a \geq 1$. We label the hard and soft threshold operators $\hat{\theta}_\lambda^h$ and $\hat{\theta}_\lambda^s$ respectively. Recall the following lemma. (Note: These are from Nonparametric function estimation via wavelets. Do I need this paper?)

Lemma 66. *SSuppose* $y \sim N\left(\theta, \sigma^2\right)$. *Let* $\hat{\theta}_\lambda^h$ *and* $\hat{\theta}_\lambda^s$ *be the hard and soft thresholding operators, respectively. Let* $\lambda = a\sigma$ *with* $a \geq 1$. *Then*

$$(i) \quad E\left(\hat{\theta}_\lambda^s - \theta\right)^2 \leq \left(a^2+1\right)\sigma^2 \wedge \left(2\theta^2 + \exp\left(-a^2/2\right)\sigma^2\right), \tag{30.1}$$

$$(ii) \quad E\left(\hat{\theta}_\lambda^h - \theta\right)^2 \leq \left(2a^2+2\right)\sigma^2 \wedge \left(2\theta^2 + 2a\exp\left(-a^2/2\right)\sigma^2\right). \tag{30.2}$$

The proofs for the following theorems are only given for the soft thresholding operator.

Proof. Proof of Theorem 64.

Let $g(t) = f\left(H^{-1}(t)\right)$ and $\tilde{g}(t) = n^{-1/2}\sum_{i=1}^n y_i\phi_{Ji}(t)$ and let $\tilde{f}(t) = \tilde{g}(H(t))$. Then

$$\begin{aligned}\tilde{f}(t) &= n^{-1/2}\sum_{i=1}^n f\left(t_i\right)\phi_{Ji}(H(t)) + n^{-1/2}\epsilon\sum_{i=1}^n z_i\phi_{Ji}(H(t)) \\ &= f(t) + \Delta(t) + r(t),\end{aligned}$$

where $\Delta(t) = n^{-1/2}\sum_{i=1}^n f\left(t_i\right)\phi_{Ji}(H(t)) - f(t)$ is the approximation error and $r(t) = n^{-1/2}\epsilon\sum_{i=1}^n y_i\phi_{Ji}(H(t))$. Now project $\tilde{f}$ onto the multiresolution space V_J and decompose the orthogonal projection $\tilde{f}_J(t) = \mathrm{Proj}_{V_j}\tilde{f}(t)$ into three terms:

$$\tilde{f}_J(t) = f_J(t) + \Delta_J(t) + r_J(t), \tag{30.3}$$

where $f_J = \mathrm{Proj}_{V_J} f$, $\Delta_J = \mathrm{Proj}_{V_J}\Delta$ and $r_J = \mathrm{Proj}_{V_J} r$ respectively. Theorem 63 yields

$$\|\Delta_J\|_2^2 = o\left(n^{-2\alpha/(1+2\alpha)}\right). \tag{30.4}$$

(Note: This is because the noise in the data is neglected in this term.) Denote $\tilde{\theta}_{jk} = \left\langle \tilde{f}_J, \psi_{jk}\right\rangle$. Just as in (30.3), we decompose this into three parts.

$$\tilde{\theta}_{jk} = \theta_{jk} + d_{jk} + r_{jk} \quad \text{for } k = 1, ..., 2^j, \ j = j_0, ..., J-1,$$

where $\theta_{jk} = \langle f, \psi_{jk}\rangle$ is the true wavelet coefficient of f, $d_{jk} = \langle \Delta_J, \psi_{jk}\rangle$ is the approximation error and $r_{jk} = \langle r_J, \psi_{jk}\rangle$ is the noise. Similarly separate $\tilde{\xi}_{j_0k} = \left\langle \tilde{f}_J, \phi_{j_0k}\right\rangle$ into three terms:

$$\tilde{\xi}_{j_0k} = \xi_{j_0k} + d'_{j_0k} + r'_{j_0k} \quad \text{for } k = 1, .., 2^{j_0}.$$

Let $\hat{\xi}_{j_0k}$ and $\hat{\theta}_{jk}$ be given as in (25.2). Then because of Parseval's relation

$$\sum_{k=1}^{2^{j_0}}\left(d'_{j_0k}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} d_{jk}^2 = \|\Delta_J\|_2^2 = o\left(n^{-2\alpha/(1+2\alpha)}\right). \tag{30.5}$$

By the orthogonality of the wavelet basis, we have the isometry between the L^2 function norm and the l_2 wavelet sequence norm:

$$E\left\|\hat{f}_n^* - f\right\|^2 = \sum_{k=1}^{2^{j_0}} E\left(\hat{\xi}_{j_0k} - \xi_{j_0k}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 + \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}\theta_{jk}^2$$

$$\equiv S_1 + S_2 + S_3.$$

From (27.1) we can see

$$S_1 \le 2^{j_0} n^{-1}\epsilon^2 h + \sum_{k=1}^{2^{j_0}}\left(d'_{j_0k}\right)^2 = o\left(n^{-2\alpha/(1+2\alpha)}\right). \tag{30.6}$$

At each resolution level j denote

$$G_j \equiv \left\{k:\ \text{supp}\,(\psi_{jk}) = \left[2^{-j}k, 2^{-j}(N+k)\right] \text{ contains at least one jump point of } f\right\}.$$

Then $\text{card}\,(G_j) \le N(m+2)$ (counting two end points 0 and 1 as jump points as well). Lemma 61 yields

$$|\theta_{jk}| \le C2^{-j(1/2+\alpha)} \quad \text{for } k \notin G_j, \tag{30.7}$$

$$|\theta_{jk}| \le C2^{-j/2} \qquad \text{for } k \in G_j, \tag{30.8}$$

where C is a constant not depending on f. Therefore,

$$S_3 = \sum_{j=J}^{\infty}\sum_{k\in G_j}\theta_{jk}^2 + \sum_{j=J}^{\infty}\sum_{k\notin G_j}\theta_{jk}^2$$

$$\le \sum_{j=J}^{\infty} N(m+2)C^2 2^{-j} + \sum_{j=J}^{\infty}\sum_{k=1}^{2^j} C^2 2^{-j(1+2\alpha)}$$

$$= o\left(n^{-2\alpha/(1+2\alpha)}\right). \tag{30.9}$$

Now we consider S_2. Note from (27.2) that $\sigma_{jk} \le u_{jk}$ and $\lambda_{jk} = u_{jk}(2\log n)^{1/2}$, so $a_{ij} = \lambda_{jk}/\sigma_{jk} \ge (2\log n)^{1/2}$. From (30.1) it follows that

$$E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \le (4\log n + 2)h\epsilon^2 n^{-1} \wedge \left(8\theta_{jk}^2 + 2h\epsilon^2 n^{-2}\right) + 10d_{jk}^2 \tag{30.10}$$

Write

$$S_2 = \sum_{j=j_0}^{J-1}\sum_{k\in G_j} E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k\notin G_j} E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2$$

$$\equiv S_{21} + S_{22}.$$

Since card $(G_j) \leq N(m+2)$, it follows from (30.10) that

$$S_{21} \leq \sum_{j=j_0}^{J-1} N(m+2)\left[(4\log n + 2)h\epsilon^2 n^{-1} + 10d_{jk}^2\right] = o\left(n^{-2\alpha/(1+2\alpha)}\right). \tag{30.11}$$

Now let J_1 be an integer satisfying $2^{J_1(1+2\alpha)} = n/\log n$. (Note: If this integer doesn't exist, choose $J_1 = \lfloor 1/(1+2\alpha)\log_2(n/\log n)\rfloor$.) From (30.10) we have

$$E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \leq 5\epsilon^2 n^{-1}\log n + 10d_{jk}^2 \text{ for } j_0 \leq j \leq J_1 - 1,\ k \notin G_j, \tag{30.12}$$

$$E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \leq 8C^2 2^{-j(1+2\alpha)} + 2h\epsilon^2 n^{-2} + 10d_{jk}^2 \text{ for } J_1 \leq j \leq J - 1,\ k \notin G_j. \tag{30.13}$$

Therefore,

$$S_{22} \leq \sum_{j=j_0}^{J_1-1}\sum_{k\notin G_j} 5\epsilon^2 n^{-1}\log n + \sum_{j=J_1}^{J-1}\sum_{k\notin G_j}\left(8C^2 2^{-j(1+2\alpha)} + 2h\epsilon^2 n^{-2}\right)$$

$$+10\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} d_{jk}^2$$

$$= C(\log n/n)^{2\alpha/(1+2\alpha)}(1+o(1)). \tag{30.14}$$

Putting (30.6), (30.9), (30.11), and (30.14) together yields

$$E\left\|\hat{f}_n^* - f\right\|_2^2 \leq C(\log n/n)^{2\alpha/(1+2\alpha)}(1+o(1)). \tag{30.15}$$

□

I now give a brief summary of the proof of Theorem 65.

Proof. Proof of Theorem 65. We use the inequality below.

Let X_i be random variables, $i = 1, ..., n$. Then

$$E\left(\sum_{i=1}^n X_i\right)^2 \leq \left(\sum_{i=1}^n \left(EX_i^2\right)^{1/2}\right)^2. \tag{30.16}$$

Applying this inequality yields

$$E\left(f_n^*(t_0) - f(t_0)\right)^2 = E\left[\sum_{k=1}^{2^{j_0}}\left(\hat{\xi}_{j_0k} - \xi_{j_0k}\right)\phi_{j_0k}(t_0) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j}\left(\hat{\theta}_{jk} - \theta_{jk}\right)\psi_{jk}(t_0)\right]^2$$

$$\leq \left[\sum_{k=1}^{2^{j_0}}\left(E\left(\hat{\xi}_{j_0k}-\xi_{j_0k}\right)^2\phi_{j_0k}^2\left(t_0\right)\right)^{1/2}+\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\left(E\left(\hat{\theta}_{jk}-\theta_{jk}\right)^2\psi_{ij}^2\left(t_0\right)\right)^{1/2}+\sum_{j=J}^{\infty}\sum_{k=1}^{2^j}\left|\theta_{jk}\psi_{jk}\left(t_0\right)\right|\right]^2$$

$$\equiv \left(Q_1+Q_2+Q_3\right)^2.$$

Lastly, each of these terms is bounded using the same sort of properties and lemmas used in the last proof. □

30.1. **Important notes about this paper.** This paper uses the Holder class, which is the class of functions which are Holder continuous expect for m discontinuities. It deals with data which is identically independent and has no long memory error. It uses a function H to adjust the irregularly spaced data and find the wavelet coefficients.

Part 7. Using linear interpolation on irregularly spaced long memory data.

31. Introduction.

This part deals with the problem of function estimation from data. Very many variations of the problem are useful. Many real world problems which have been solved are very oversimplified. In most situations it is not reasonable to assume that data are independent. One example of this is the time series. Here we have data which are dependent. We could use this new research to compare two time series.

Another example of where this research is applicable is in the cause of spatially dependent data. For instance: flooding at certain points along the Nile river. Clearly if an area is flooded, a nearby area would be much more likely to be flooded. Also, your data points would very likely be unequally spaced. This is why we will try to address the problems of long memory data and unequally spaced data simultaneously.

In real world applications data could be equally spaced or unequally spaced. There many be more than one data point for the same value. Error may or may not be independent. We consider the specific variation where the data are irregularly spaced and the error is long memory dependent.

We organize the paper as follows. In Section 32 we give a basic presentation of the problem. In Section 33 we define a specific breakdown of the wavelet coefficients which will allow us to more easily find bounds for the MISE (mean integrated square error). In Section 34 we will separate this MISE and bound each piece.

32. Basic Notation: Preliminaries.

32.1. **Preliminaries.** We have

$$Y_m = g\left(X_m\right)+\epsilon_m \quad \text{for } 1\leq m\leq n \tag{32.1}$$

where $\mathcal{Y} = \{(X_m, Y_m), 1 \le m \le n\}$ and $\mathcal{X} = \{X_m, 1 \le m \le n\}$. These are the collected data. We will assume that the $\mathcal{X}$ has been put in order of size, thus these are ranked data. The data $\mathcal{X}$ are ordered values of a random sample from a probability distribution f having support $I = [0, 1]$. These X_i's are independent.

Also significant is that the fact that the sample size is not a factor of 2. We will define specific constants to deal with this later.

The ϵ_m are long memory dependent Gaussian variables with $E(\epsilon_m) = 0$ and $E(\epsilon_m^2) = \sigma^2 > 0$. Long memory means that the covariance has the following property. We define

$$r(j) = E(\epsilon_i \epsilon_{i+j}) \sim C_0 |j|^{-\alpha}$$

where $\alpha \in (0, 1]$. Here $a_j \sim b_j$ means that $a_j/b_j \to 1$ when $j \to \infty$. (We use this notation throughout the paper.)

32.2. **Interpolation Rules.** To deal with the problem of nonequispaced data, we interpolate the data as follows.

$$Y(x) = \sum_m w_m(x) Y_m \quad \text{for} \quad x \in (X_{-v_1}, X_{n-v_2}] \tag{32.2}$$

where $v_1 < 0 \le v_2$ and $w_j = 0$ unless $v_1 \le j \le v_2$. At the ends of the interval $(X_{-v_1}, X_{n-v_2}]$ Hall and Turlach advocate the use of horizontal expansion. That is, let $Y(t) \equiv Y(X_{-\nu_1}+)$ on $[0, X_{-\nu_1}]$, and $Y(t) \equiv Y(X_{n-\nu_2})$ on $(X_{n-\nu_2}, 1]$. The authors also note that one could use quadratic expansion but that this affects the error terms.

So we are interpolating among the data in $\mathcal{Y}$ to produce a process $Y = Y(x)$ which satisfies $E(Y) \approx g$. The values of $w_m(x)$ come from local averaging and linear interpolation. These w_m's are dependent on the X_m's but not on the Y_m's. For local averaging we let

$$w_m(x) = (2v)^{-1} \quad \text{if} \quad -v+1 \le m-l \le v, \quad 0 \text{ otherwise.} \tag{32.3}$$

For linear interpolation we let

$$w_m(x) = \begin{cases} v^{-1}(X_{2l-m+1} - x)/(X_{2l-m+1} - X_m) & -v+1 \le m-l \le 0 \\ v^{-1}(x - X_{2l-m+1})/(X_m - X_{2l-m+1}) & 1 \le m-l \le v \\ 0 & \text{otherwise.} \end{cases} \tag{32.4}$$

From these weights come two rules for $Y(x)$. For local averaging

$$Y(x) = (2v)^{-1} \sum_{m=-v+1}^{v} Y_{l+m}. \tag{32.5}$$

The second rule for linear interpolation is

$$Y(x) = v^{-1} \sum_{m=1}^{v} \left(\frac{x - X_{l-m+1}}{X_{l+m} - X_{l-m+1}} Y_{l+m} + \frac{X_{l+m} - x}{X_{l+m} - X_{l-m+1}} Y_{l-m+1} \right). \tag{32.6}$$

Both of these rules are for $x \in (X_l, X_{l+1}]$. We can see where this sum comes from by considering the definition $Y(x) = \sum w_m(x) Y_m$. For the first part of (49.4) we have

$$\sum_{m-l=1}^{v} \frac{x - X_{2l-m+1}}{X_m - X_{2l-m+1}} Y_m.$$

Consider the substitution $m = l + m'$. Then $m' = m - l$. Also $2l - m + 1 = l + l - m + 1 = l - m' + 1$. The sum then becomes

$$\sum_{m'=1}^{v} \frac{x - X_{l-m'+1}}{X_{l+m'} - X_{l-m'+1}} Y_{l+m'}$$

as required. For the second part of (49.4) we have

$$\sum_{m-l=0}^{-v+1} \frac{X_{2l-m+1} - x}{X_{2l-m+1} - X_m} Y_m.$$

Consider the substitution $m = l - m' + 1$. Then $m - l = 1 - m'$, and $m' = l + 1 - m$. Also $2l - m + 1 = l + l - m + 1 = l + m'$. We have

$$\sum_{m'=1}^{v} \frac{X_{l+m'} - x}{X_{l+m'} - X_{l-m'+1}} Y_{l-m'+1}$$

as required.

The authors Hall and Turlach point out that other interpolation rules might be usable but that the bounds for the weights require different computations. Particular weights may be more adept at dealing with the situation of long memory dependence. For instance, using the binomial coefficients would more heavily weight nearby points in the interpolation.

32.3. **Wavelet structure.** Now we discuss the structure of the wavelets. Write ϕ and ψ for the "father" and "mother" wavelets.

We let $p = p(n)$ be the resolution level of the wavelets. Denote $p_i = 2^i p$ for $i \geq 0$.

Let $\psi_{ij}(x) = p_i^{1/2} \psi(p_i x - j)$ and $\phi_j(x) = p^{1/2} \phi(px + j)$. These will form the multiresolution analysis. Here the collection $\{\phi_j, \psi_{ij}, i \geq i_0, j \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. We note here that our function g is not infinite, so the number of j's we estimate is finite in the summation (65.1) below.

We have the true wavelet coefficients

$$a_j = \int_I g\phi_j \tag{32.7}$$

$$b_{ij} = \int_I g\psi_{ij} \tag{32.8}$$

We will assume that ψ is of order r. That is $\int x^i\psi(x) = 0$ for $i = 0, 1, ..., r-1$.

Then

$$g = \sum_j a_j\phi_j + \sum_{i=1}^{\infty}\sum_j b_{ij}\psi_{ij}.$$

We estimate these true coefficients by computing

$$\hat{a}_j = \int_I Y\phi_j \qquad \hat{b}_{ij} = \int_I Y\psi_{ij}.$$

Then our new estimator is

$$\hat{g} = \sum_j \hat{a}_j\phi_j + \sum_{i=0}^{q-1}\sum_j \hat{b}_{ij} I\left(\left|\hat{b}_{ij}\right| \geq \delta\right)\psi_{ij}. \tag{32.9}$$

Here δ is a threshold parameter and q is the truncation point for the series. Note that this would need to be further approximated by

$$\tilde{a}_{i_0 j} = n^{-1}\sum_{m=1}^{n} Y\left(\frac{m}{n}\right)\phi_j\left(\frac{m}{n}\right) \quad \text{and} \quad \tilde{b}_{ij} = n^{-1}\sum_{m=1}^{n} Y\left(\frac{m}{n}\right)\psi_{ij}\left(\frac{m}{n}\right).$$

We will also require of our wavelets that the quantities

$$C_{6\star} = C_0\int_0^1\int_0^1 |x-y|^{-\alpha}\,\phi(x)\phi(y)dxdy$$

and

$$C_6 = C_0\int_0^1\int_0^1 |x-y|^{-\alpha}\,\psi(x)\psi(y)dxdy$$

are bounded. We consider this is in Section 33.8. The truth of these statements is a direct result of the functions ϕ and ψ being compactly supported.

32.4. **Other assumptions.** We must remember the intrinsic ordering of the $\mathcal{X} = \{X_m | 1 \leq m \leq n\}$. Strictly speaking they should be written X_{nm} to signify their dependence on n. A similar rule applies to ϵ_m. We drop this notation as in [11] and hope that this ordering is understood.

We assume of the function g that it has r piecewise continuous derivatives, in the sense that there exist constants $0 = a_1 < a_2 < ... < a_k = 1$ such that g has r continuous derivatives on each interval $[a_l, a_{l+1}]$ for $1 \leq l \leq k-1$. We assume the same of the density function f of the $\mathcal{X}_m$'s, possibly with different a_i's and a different k. Thus, the the function g and its derivatives have a bounded number of discontinuities.

We assume that the functions ϕ and ψ are compactly supported and bounded as expressed in Section 33.3. We also assume that for some $r \geq 1$ and $\kappa \neq 0$, and all integers $i \in [0, r]$ and $j \in (-\infty, \infty)$

$$\int \psi^2 = 1, \quad \int x^i \psi(x) dx = \kappa \left(r!\right)^{-1} \delta_{ir},$$

$$\int \phi = 1, \quad \int \phi(x)\phi(x+j)dx = \delta_{oj},$$

where δ_{jk} is the Kronecker delta and κ is some constant.

We will note here that these requirements do not affect our choice of weights $w_m(x)$ because those are piecewise polynomials, and therefore do not disappear in the integrals against ψ_{ij}.

33. Breaking Down Coefficients and Bounding Error terms.

33.1. **Initial breakdown.** We recall $I = [0, 1]$. First we will break the scaling function coefficients into pieces.

$$\hat{a}_j = \int Y \phi_j(x) = \int \left(\sum_m w_m(x) Y_m \right) \phi_j = \int \left(\sum_m w_m(x) \left(g\left(X_m\right) + \epsilon_m \right) \right) \phi_j$$

$$= \int \left(\sum_m w_m(x) g\left(X_m\right) \right) \phi_j + \int \left(\sum_m w_m(x) \epsilon_m \right) \phi_j. \tag{33.1}$$

For the first term, note that

$$E\left(\mathcal{Y}|\mathcal{X}\right) = \sum_m w_m\left(x\right) g(X_m).$$

This is because the error term of $Y = g\left(X_m\right) + \epsilon_m$ would be gone because of the expected value, the $g\left(X_m\right)$'s are constants, the only thing that varies is the $w_m(x)$ and it's linear, and $E\left(\epsilon_m\right) = 0$. Note

that

$$\int_I \left(\sum_m w_m(x) g\left(X_m\right) \right) \phi_j = \int_I E\left(\mathcal{Y}|\mathcal{X}\right) \phi_j - \int_I g(x)\phi_j(x) + \int_I g(x)\phi_j(x). \tag{33.2}$$

We can let

$$\Delta = E\left(\mathcal{Y}|\mathcal{X}\right) - g.$$

Then (33.2)

$$= \int_I \Delta\phi_j(x) + \int_I g(x)\phi_j(x) \tag{33.3}$$

$$\equiv A_j + a_j \tag{33.4}$$

For the second term, let

$$v_{j;m} = \left(n/p^{1/2}\right) \int_I w_m(x)\phi_j(x). \tag{33.5}$$

Then the second term is

$$R_j \equiv (p/n)^{1/2} \sum_m v_{j;m}\epsilon_m. \tag{33.6}$$

We define

$$\hat{\chi}_j \equiv n^{-1/2} R_j$$

Thus

$$\hat{a}_j = a_j + A_j + \hat{\chi}_j = a_j + A_j + n^{-1/2} R_j. \tag{33.7}$$

Now I will break down these estimators of the wavelet coefficients into pieces. Note that

$$\hat{b}_{ij} = \int Y(x)\psi_{ij}(x) = \int \left(\sum_m w_m(x) Y_m \right) \psi_{ij} = \int \left(\sum_m w_m(x) \left(g\left(X_m\right) + \epsilon_m \right) \right) \psi_{ij}$$

$$= \int \left(\sum_m w_m(x) g\left(X_m\right) \right) \psi_{ij} + \int \left(\sum_m w_m(x)\epsilon_m \right) \psi_{ij}. \tag{33.8}$$

Then the first term is via (33.8).

$$\int_I \left(\sum_m w_m(x) g\left(X_m\right) \right) \psi_{ij} = \int_I E\left(\mathcal{Y}|\mathcal{X}\right) \psi_{ij} - \int_I g(x)\psi_{ij}(x) + \int_I g(x)\psi_{ij}(x). \tag{33.9}$$

Then (33.2)

$$= \int_I \Delta\psi_{ij}(x) + \int_I g(x)\psi_{ij}(x) \tag{33.10}$$

$$\equiv B_{ij} + b_{ij} \tag{33.11}$$

For the second term, let

$$v_{ij;m} = \left(n/p_i^{1/2}\right) \int_I w_m(x)\psi_{ij}(x). \tag{33.12}$$

Then the second term is

$$S_{ij} \equiv (p_i/n)^{1/2} \sum_m v_{ij;m}\epsilon_m. \tag{33.13}$$

We define

$$\hat{\xi}_{ij} \equiv n^{-1/2} S_{ij}$$

Thus

$$\hat{b}_{ij} = b_{ij} + B_{ij} + \hat{\xi}_{ij} = b_{ij} + B_{ij} + n^{-1/2} S_{ij}. \tag{33.14}$$

33.2. **Bounds of A_j.** By an argument identical to the one in Section 33.3, we obtain the bound

$$E\left(|A_j|^k\right) = \begin{cases} O\left(\left(p^{1/2}/n\right)^k n^\eta\right) & \text{for } j \in J(\epsilon) \\ O\left(n^{\eta-k}\right) & \text{for } j \notin J(\epsilon) \end{cases} \tag{33.15}$$

Here the $J(\epsilon)$ contains points where $g(x)$ is discontinuous. Outside of $J(\epsilon)$ is where $g(x)$ is continuous. More specifically we let

$$x \in J(\epsilon) \implies px + j \in \left(-c - pn^{\epsilon-1}, c + pn^{\epsilon-1}\right).$$

33.3. **Bounds of B_{ij}.** We must show that

$$E\left(|B_{ij}|^k\right) = \begin{cases} O\left(\left(p_i^{1/2}/n\right)^k n^\eta\right) & \text{for } j \in J_i(\epsilon) \\ O\left(n^{\eta-k}\right) & \text{for } j \notin J_i(\epsilon) \end{cases} \tag{33.16}$$

Here the $J_i(\epsilon)$ is were $g(x)$ or one of its derivatives is discontinuous. We divide the entire support of $g(x)$ into intervals the size of $J_i(\epsilon)$. Outside of $J_i(\epsilon)$ is where these functions are continuous. We still can make the same assumption of closeness, the $J_i(\epsilon)$ only distinguishes between continuity and discontinuity. We let

$$x \in J_i(\epsilon) \implies p_i x + j \in \left(-c - p_i n^{\epsilon-1}, c + p_i n^{\epsilon-1}\right).$$

Suppose that $x, y \in J_i(\epsilon)$. Without loss of generality we assume that $y > x$. Then

$$-c - p_i n^{\epsilon-1} < p_i x + j < p_i y + j < c + p_i n^{\epsilon-1}$$

$$\frac{-c - p_i n^{\epsilon-1} - j}{p_i} < x < y < \frac{c + p_i n^{\epsilon-1} - j}{p_i}$$

We subtract these.

$$\frac{c + p_i n^{\epsilon-1} - j}{p_i} - \frac{-c - p_i n^{\epsilon-1} - j}{p_i} = \frac{c + p_i n^{\epsilon-1} - j + c + p_i n^{\epsilon-1} + j}{p_i}$$

$$= \frac{2c + 2p_i n^{\epsilon-1}}{p_i} = O\left(n^{\epsilon-1}\right).$$

$$\sup_{I(\epsilon)} E\left(|\Delta|^k\right) = O\left(n^{k\epsilon-k}\right) = O\left(n^{\eta-1}\right) \quad \text{for all } k \geq 1 \text{ and } \epsilon > 0,\ \eta = k\epsilon.$$

Noting that $|\psi_{ij}| = 1$ shows that we have found the bound for B_{ij}. Then $|B_{ij}|^k = O\left(p_i^{k/2} n^{\eta-k}\right)$ for $x, y \in J_i(\epsilon)$. Note that this factor of $p_i^{k/2}$ comes from the ψ_{ij} within the integral, $\psi_{ij}(x) = p_i^{1/2}\psi\left(p_i x - j\right)$.

Now suppose $x, y \notin J_i(\epsilon)$. In that case, $g(x)$ is continuous, as are all of its derivatives, specifically $g'(x)$ is continuous. Thus

$$\lim_{x \to y} \frac{|g(x) - g(y)|}{|x - y|} = C = g'(x).$$

$$|g(x) - g(y)| < C\,|x - y|$$

and $|x - y| < \delta$ implies $|g(x) - g(y)| < C\delta$. We assume that x and y are close in the same way and receive the result that $|B_{ij}|^k = O\left(n^{\eta-k}\right)$ for $x, y \notin J_i(\epsilon)$

33.4. **Bound of $v_{j;m}$.** Following the lines of the next section we obtain exactly the bounds of v_j. For some $0 < \epsilon_1 < 1/20$, or for some ϵ_1 which is close to zero for local averaging

$$|v_{j;m}| \leq n^{\epsilon_1} \sup |\phi|\,. \tag{33.17}$$

For linear interpolation

$$|v_{j;m}| \leq 2 \cdot n^{\epsilon_1} \sup |\phi|\,. \tag{33.18}$$

This means that for local averaging

$$E\left(v_{j;m}^2\right) \leq n^{2\epsilon_1} \sup |\psi|^2 \tag{33.19}$$

Also,

$$E\left(v_{j;m}^2\right) \leq 4n^{2\epsilon_1} \sup |\psi|^2 \tag{33.20}$$

for linear interpolation.

These bounds should be the same for $E\left(v_{j;k}v_{j;m}\right)$.

33.5. **Bound of $v_{ij;m}$.** We now examine

$$v_{ij;m} = \left(n/p_i^{1/2}\right) \int_I w_m \psi_{ij}(x)dx$$

The problem we must overcome is that the sample data points $\mathcal{X}_1, ..., \mathcal{X}_n$, which are later reordered according to size may become very close together, thus making the $v_{ij;m}$ large in the denominator.

Recall the weight formulas. We have the local averaging below.

$$w_m(x) = (2v)^{-1} \quad \text{if} \quad -v+1 \leq m-l \leq v, \quad 0 \text{ otherwise.} \tag{33.21}$$

Also we have the linear interpolation.

$$w_m(x) = \begin{cases} v^{-1}\left(X_{2l-m+1} - x\right)/\left(X_{2l-m+1} - X_m\right) & -v+1 \leq m-l \leq 0 \\ v^{-1}\left(x - X_{2l-m+1}\right)/\left(X_m - X_{2l-m+1}\right) & 1 \leq m-l \leq v \\ 0 & \text{otherwise.} \end{cases} \tag{33.22}$$

Define event $\mathcal{E}_1$ as the event where $\mathcal{X}_{l+1}$ and $\mathcal{X}_l$ are within $n^{\epsilon_1 - 1}$ of each other, where $0 < \epsilon_1 < 1$. In other words,

$$\mathcal{E}_1 : \quad \mathcal{X}_{l+1} - \mathcal{X}_l \leq n^{\epsilon_1 - 1}.$$

We wish to examine $P\left(\mathcal{E}_1\right)$. First we note the probability density functions of the ranked $\mathcal{X}_j$. Recall that $f(x)$ is the pdf of the $\mathcal{X}_1, ..., \mathcal{X}_n$, let $F(x)$ denote the cumulative density function. Then $P\left(\mathcal{E}_1\right) = P\left(\mathcal{X}_{l+1} \leq \mathcal{X}_l + n^{\epsilon_1 - 1}\right) =$

$$\int_{\mathcal{X}_l}^{\mathcal{X}_l + n^{\epsilon_1 - 1}} f_{\mathcal{X}_{l+1}}(x)dx \leq \int_{\mathcal{X}_l}^{\mathcal{X}_l + n^{\epsilon_1 - 1}} \sup_y f_{\mathcal{X}_{l+1}}(y)dx$$

$$= C\left[\mathcal{X}_l + n^{\epsilon_1 - 1} - \mathcal{X}_l\right] = Cn^{\epsilon_1 - 1} = O\left(n^{-\lambda}\right)$$

for $0 < \lambda < 1$. This means that

$$P\left(\mathcal{E}_1^c\right) = 1 - O\left(n^{-\lambda}\right). \tag{33.23}$$

Now for the local averaging,

$$v_{ij;m} = \left(n/p_i^{1/2}\right) \int_I w_m \psi(x) dx$$

$$\left| \int w_m \psi_{ij} \right| = \left| (2v)^{-1} \sum_{l=m-v}^{m+v+1} \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \psi_{ij} \right|$$

$$2v \cdot \frac{1}{2v} \left| \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \psi_{ij} \right| \leq p_i^{1/2} \sup\left(\mathcal{X}_{l+1} - \mathcal{X}_l\right) \sup|\psi| .$$

Therefore

$$|v_{ij;m}| \leq p_i^{1/2} \left(n/p_i^{1/2}\right) n^{\epsilon_1 - 1} \sup|\psi|$$

$$|v_{ij;m}| \leq p_i^{1/2} n^{\epsilon_1} / p_i^{1/2} \sup|\psi| = n^{\epsilon_1} \sup|\psi| . \tag{33.24}$$

Now let's consider the linear interpolation.

$$\left| \int w_m \psi \right| = \left| \int \frac{1}{v} \left(\sum_{l=-v+1}^{0} \frac{\mathcal{X}_{2l-m+1} - x}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} + \sum_{l=1}^{v} \frac{x - \mathcal{X}_{2l-m+1}}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \right) \psi_{ij}(x) dx \right|$$

$$\leq \left| \int \frac{1}{v} \sum_{l=-v+1}^{0} \frac{\mathcal{X}_{2l-m+1} - x}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \psi_{ij}(x) dx \right| + \left| \int \frac{1}{v} \sum_{l=1}^{v} \frac{x - \mathcal{X}_{2l-m+1}}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \psi_{ij}(x) dx \right|$$

Let's examine

$$\left| \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \frac{1}{v} \sum_{l=-v+1}^{0} \frac{\mathcal{X}_{2l-m+1} - x}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} dx \right| + \left| \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \frac{1}{v} \sum_{l=1}^{v} \frac{x - \mathcal{X}_{2l-m+1}}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} dx \right| .$$

Without loss of generality we assume that the first of these is greater than 0. Then

$$\left| \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \frac{x - \mathcal{X}_{2l-m+1}}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} dx \right| = \left| \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \frac{\mathcal{X}_{2l-m+1} - x}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} dx \right| = \int_{\mathcal{X}_l}^{\mathcal{X}_{l+1}} \frac{\mathcal{X}_{2l-m+1} - x}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} dx$$

$$= \frac{1}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \left[\mathcal{X}_{2l-m+1} \cdot x - \frac{x^2}{2} \right]_{\mathcal{X}_l}^{\mathcal{X}_{l+1}}$$

$$= \frac{1}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \left[\mathcal{X}_{2l-m+1} \left(\mathcal{X}_{l+1} - \mathcal{X}_l\right) - \frac{1}{2} \left(\mathcal{X}_{l+1}^2 - \mathcal{X}_l^2\right) \right]$$

$$= \frac{1}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \left[\mathcal{X}_{2l-m+1} \left(\mathcal{X}_{l+1} - \mathcal{X}_l\right) - \frac{1}{2} \left(\mathcal{X}_{l+1} - \mathcal{X}_l\right) \left(\mathcal{X}_{l+1} + \mathcal{X}_l\right) \right]$$

$$= \frac{\mathcal{X}_{l+1} - \mathcal{X}_l}{\mathcal{X}_{2l-m+1} - \mathcal{X}_m} \left[\mathcal{X}_{2l-m+1} - \frac{1}{2} \left(\mathcal{X}_{l+1} + \mathcal{X}_l \right) \right]$$

The rest of this is bounded by realizing that the difference of two variables is exponentially distributed as in [13]. Thus,

$$|v_{ij;m}| \leq 2 \cdot p_i^{1/2} \left(n/p_i^{1/2} \right) n^{\epsilon_1 - 1} \sup |\psi| = 2 \cdot n^{\epsilon_1} \sup |\psi| . \tag{33.25}$$

What does this mean for the quantities we need to estimate?

$$E\left(v_{ij;m}^2\right) = \int_I v_{ij;m}^2 f_{\mathcal{X}_{2l-m+1}}(x) f_{\mathcal{X}_m}(x) dx$$

$$\leq n^{2\epsilon_1} \sup |\psi|^2$$

for local averaging. Also,

$$\leq 4n^{2\epsilon_1} \sup |\psi|^2$$

for linear interpolation.

These bounds should be the same for $E\left(v_{ij;k} v_{ij;m}\right)$.

33.6. **Bounds of $\sum S_{ij}^2$, $\sum R_j^2$ and some probabilities.** Let $v_{ij;1}, ..., v_{ij;n}$ denote weights. Recall that $|v_{ij;m}| \leq C n^{\epsilon_1}$. We choose ϵ_1 such that $0 \leq \epsilon_1 < 1/20$. Suppose $n^{-1} \sum_{m=1}^n v_{ij;m}^2 \geq C_2 > 0$.

We also suppose $E\left(\epsilon_i\right) = 0$, $E\left(\epsilon_i^2\right) = \sigma^2 > 0$.

We need to bound

$$E\left\{ S_{ij}^2 I\left(S_{ij} \geq z\right) \right\}.$$

We define $A = \left\{ a \in \mathbb{R}^1 : a^2 = 1,\ a = \pm 1 \right\}$.

Note that

$$\left(S_{ij}^2\right)^{1/2} = |S_{ij}| = \sup_{a \in A} a S_{ij}.$$

So instead of considering $P\left(S_{ij}^2 \geq z\right)$, we can consider $P\left(\sup_{a \in A} a S_n \geq u\right)$. Here we let $z \geq 4\tau_{i*}^2$ and $u \geq 2\tau_{i*}$ where we will define these τ_{i*} later. Let

$$Z(a) = a S_{ij}.$$

We will need this $Z(a)$, a Gaussian process, to apply an inequality to later. Then by the Cauchy-Schwartz inequality and Jensen's inequality we have

$$E\left(\sup_{a \in A} Z(a) \right) \leq E\left\{ \left(S_{ij}^2\right)^{1/2} \right\} \leq \left\{ E\left(S_{ij}^2\right) \right\}^{1/2}.$$

We will need to bound the $E\left(Z(a)^2\right)$.

$$= \frac{p_i}{n}\sum_{m=1}^{n} a^2 E\left(\epsilon_m^2\right) E\left(v_{ij;m}^2\right) + \frac{p_i}{n}\sum_{m=1}^{n}\sum_{k\neq m} a^2 E\left(\epsilon_m\epsilon_k\right) E\left(v_{ij;m}v_{ij;k}\right)$$

$$= \frac{p_i}{n}\sum_{m=1}^{n} E\left(\epsilon_m^2\right) E\left(v_{ij;m}^2\right) + \frac{p_i}{n}\sum_{m=1}^{n}\sum_{k\neq m} E\left(\epsilon_m\epsilon_k\right) E\left(v_{ij;m}v_{ij;k}\right)$$

$$\equiv J_1 + J_2$$

Now

$$J_1 = p_i\frac{\sigma^2}{n}\sum_{m=1}^{n} E\left(v_{ij;m}^2\right) \leq p_i\frac{\sigma^2}{n}\cdot n^{2\epsilon_1}\sup|\psi|^2 = p_i\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2$$

for local averaging and

$$J_1 = p_i\frac{\sigma^2}{n}\sum_{m=1}^{n} E\left(v_{ij;m}^2\right) \leq p_i\frac{\sigma^2}{n}\cdot 4n^{2\epsilon_1}\sup|\psi|^2 = 4p_i\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2$$

for linear interpolation. So

$$J_1 = O\left(p_i\sigma^2 n^{2\epsilon_1-1}\right). \tag{33.26}$$

Now consider

$$J_2 = p_i\frac{1}{n}\sum_{m=1}^{n}\sum_{k\neq m} E\left(\epsilon_m\epsilon_k\right) E\left(v_{ij;m}v_{ij;k}\right)$$

$$= p_i\frac{1}{n}\sum_{m=1}^{n}\sum_{k\neq m} r(m-k)E\left(v_{ij;m}v_{ij;k}\right)$$

Note that E$(v_{ij;m}v_{ij;k})$is subject to the same upper bound as $E\left(v_{ij;m}^2\right)$.

$$\leq p_i\sum_{m=1}^{n}\sum_{k\neq m} C_1 n^{2\epsilon_1-1} r(m-k)\int\int p^{i/2}\psi\left(p^i x - j\right)p^{i/2}\psi\left(p^i y - j\right)dxdy \tag{33.27}$$

Now for a variable substitution. Let $x = \frac{u+m}{n}$ and $y = \frac{v+k}{n}$. Then $dx = \frac{1}{n}du$ and $dy = \frac{1}{n}dv$. Then

$$= \frac{p_i^2}{n^2}\sum_{m=1}^{n}\sum_{k\neq m} r(m-k)\int\int C_1 n^{2\epsilon_1-1}\psi\left(p_i\frac{u+m}{n} - j\right)\psi\left(p_i\frac{v+k}{n} - j\right)dudv.$$

Here $C_1 = 4\sup|\psi|$ for linear interpolation and $C_1 = \sup|\psi|$ for local averaging.

$$\leq \int\int C_1 n^{2\epsilon_1-1}\times$$

$$\left\{ \frac{p_i^2}{n^2} \sum_{m=1}^{n} \sum_{k \neq m} r(m-k) \psi \left(p_i \frac{u+m}{n} - j \right) \psi \left(p_i \frac{v+k}{n} - j \right) \right\} dudv.$$

We consider just the second part of this equation.

$$\frac{p_i^2}{n^2} \sum_{m=1}^{n} \sum_{k \neq m} r(m-k) \psi \left(p_i \frac{u+m}{n} - j \right) \psi \left(p_i \frac{v+k}{n} - j \right) \tag{33.28}$$

$$\sim C_0 \frac{p_i^2}{n^2} \sum_{m=1}^{n} \sum_{k \neq m} n^{-\alpha} \left| \frac{m}{n} - \frac{k}{n} \right|^{-\alpha} \psi \left(p_i \frac{u+m}{n} - j \right) \psi \left(p_i \frac{v+k}{n} - j \right)$$

Let $x = p_i \frac{u+m}{n} - j$ and $y = p_i \frac{v+k}{n} - j$. Here $dx = p_i du$ and $dy = p_i dv$. Then $x - y = p_i \frac{u+m}{n} - p_i \frac{v+k}{n}$ and $\frac{m}{n} - \frac{k}{n} = p_i^{-1}(x-y) + \frac{v-u}{n}$.

$$= C_0 \frac{1}{n^2} n^{-\alpha} \sum_{m=1}^{n} \sum_{k \neq m} \left| p_i^{-1}(x-y) + \frac{v-u}{n} \right|^{-\alpha} \psi(x)\psi(y) \tag{33.29}$$

Note that $\left| p_i^{-1}(x-y) + \frac{v-u}{n} \right| \leq \left| p_i^{-1}(x-y) \right| + \left| \frac{v-u}{n} \right|$. Therefore, (33.29)

$$\leq C_0 n^{-\alpha-2} p_i^{\alpha} \int \int \frac{\psi(x)\psi(y)}{|x-y|^{\alpha}} dxdy$$

$$\leq C_0 \tau_i^2 \tag{33.30}$$

where $\tau_i^2 = C_6 n^{-\alpha-2} p_i^{\alpha}$. Here

$$C_6 = C_0 \int_0^1 \int_0^1 |x-y|^{-\alpha} \psi(x)\psi(y) dxdy. \tag{33.31}$$

Let $\tau_{i\star}^2 = C_1 \cdot C_6 n^{2\epsilon_1 - 1 - \alpha - 2} p_i^{\alpha} = C_7 n^{2\epsilon_1 - \alpha - 3} p_i^{\alpha}$.

$$J_1 = O\left(p_i \sigma^2 n^{2\epsilon_1 - 1} \right)$$

Let $\tau_{i*}^2 = C_8 p_i n^{2\epsilon_1 - 1}$.

Thus, for $0 < \alpha < 1$ from (33.26) and (33.30) we have

$$D^2 \equiv \sup_{a \in A} E\left(Z(a)^2 \right) \leq \tau_{i*}^2 \tag{33.32}$$

or

$$E\left(S_{ij}^2 \right) \leq \sup_{a \in A} E\left(Z(a)^2 \right) \leq n^{2\epsilon_1} C_6 n^{-\alpha-2} p_i^{\alpha} \leq C_8 p_i n^{2\epsilon_1 - 1}. \tag{33.33}$$

More specifically

$$E\left(S_{ij}^2\right) \leq 4p_i\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2 + C_7 n^{2\epsilon_1-\alpha-3}p_i^\alpha.$$

Using a similar process we could derive that

$$E\left(R_j^2\right) \leq n^{2\epsilon_1}C_{7\star}n^{-\alpha-2}p^\alpha \leq C_{8\star}pn^{2\epsilon_1-1}. \tag{33.34}$$

where

$$C_{6\star} = C_0\int_0^1\int_0^1 |x-y|^{-\alpha}\,\phi(x)\phi(y)dxdy. \tag{33.35}$$

More specifically

$$E\left(R_{ij}^2\right) \leq 4p\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2 + C_{7\star}n^{2\epsilon_1-\alpha-3}p^\alpha.$$

Here $C_{7\star} = C_1 \cdot C_{6\star}$ and $C_{1\star} = 4\sup|\phi|$ for linear interpolation and $C_{1\star} = \sup|\phi|$ for local averaging. We summarize these results in the following theorem.

Theorem 67. *The bounds of S_{ij}^2 and R_j^2 are as follows.*

$$E\left(S_{ij}^2\right) \leq 4p_i\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2 + C_7 n^{2\epsilon_1-\alpha-3}p_i^\alpha.$$

$$E\left(R_j^2\right) \leq 4p\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2 + C_{7\star}n^{2\epsilon_1-\alpha-3}p^\alpha.$$

In the corresponding Li and Xiao notes here they reference Borell's inequality from Adler 1990 (p. 42).

Theorem 68. *Let X* be a centered Gaussian random variable with variance σ^2. Then choosing

$$\Psi(\lambda) = (2\pi)^{-\frac{1}{2}}\int_\lambda^\infty e^{-\frac{1}{2}x^2}dx$$

to denote the standard Gaussian distribution function,

$$P\{X > \lambda\} = \Psi(\lambda/\sigma) \leq \left(\sigma/\sqrt{2\pi}\right)\lambda^{-1}e^{-\frac{1}{2}\lambda^2/\sigma^2}.$$

A consequence of this is the following theorem (p. 43).

Theorem 69. *If we assume that* $\{X_t\}_{t\in T}$ is a centered Gaussian process and that $\{X_t\}_{t\in T}$ has bounded simple paths with probability one, then for all $\epsilon > 0$ and large enough λ

$$P\left\{\sup_{t\in T}X_t > \lambda\right\} \leq e^{\epsilon\lambda^2 - \frac{1}{2}\lambda^2/\sigma_T^2} \tag{33.36}$$

where

$$\sigma_T^2 \equiv \sup_{t\in T} EX_t^2.$$

This result is true for any arbitrary ϵ, so we may drop the $\epsilon\lambda^2$ term.

We can see that this correlates with exactly what we've been doing. Our Gaussian process $\sup_{a\in A} Z(a)$ is not centered. Let $\tilde{\mu} \equiv E\left(\sup_{a\in A} Z(a)\right)$. We do not know what this quantity is but we do know that it is bounded. Let $u \geq 2\tau_{i*}$ and $z \geq 4\tau_{i*}^2$. Then $u \geq 2\tilde{\mu}$ and $\tilde{\mu} \leq \frac{1}{2}\mu$

$$P\left\{\sup_{a\in A} aS_{ij} \geq u\right\} \leq \exp\left(-\frac{(u-\tilde{\mu})^2}{2D^2}\right)$$

Note that

$$-(\mu-\tilde{\mu})^2 \geq -\left(\mu - \frac{1}{2}\mu\right)^2 = -\frac{1}{4}\mu^2$$

and

$$D^2 \leq \tau_{i*}^2.$$

$$P\{S_{ij} \geq u\} \leq P\left\{\sup_{a\in A} aS_{ij} \geq u\right\} = P\left\{S_{ij}^2 \geq z\right\} \leq \exp\left(-\frac{u^2}{8\tau_{i*}^2}\right). \tag{33.37}$$

We must also consider the case where $\alpha = 1$. The only difference arises when we go to estimate J_2. We would need to rebound the second part of J_2. We begin from (33.28)

$$\frac{p_i}{n^2}\sum_{m=1}^{n}\sum_{k\neq m} r(m-k)\psi\left(2^i\frac{u+m}{n} - j\right)\psi\left(2^i\frac{v+k}{n} - j\right).$$

$$\sim C_0\frac{p_i^2}{n^2}\sum_{m=1}^{n}\sum_{k\neq m} n^{-1}\left|\frac{m}{n} - \frac{k}{n}\right|^{-1}\psi\left(p_i\frac{u+m}{n} - j\right)\psi\left(p_i\frac{v+k}{n} - j\right)$$

$$= C\frac{p_i}{n}\sum_{m=1}^{n}\sum_{k\neq m}\left|p_i\frac{m}{n} - p_i\frac{k}{n}\right|^{-1}\psi\left(p_i\frac{u+m}{n} - j\right)\psi\left(p_i\frac{v+k}{n} - j\right)$$

$$\leq C\frac{1}{n}\sum_{m=1}^{n}\sum_{k\neq m}\left|p_i\frac{m}{n} - p_i\frac{k}{n}\right|^{-1}\sup|\psi_{ij}|^2 \leq Cn^{-1}\left[\ln\left(np_i^{-1}\right) + 2\right] \leq Cn^{-1}\log\left(np_i^{-1}e\right). \tag{33.38}$$

We note that with a substitution this is similar to computing

$$\sum_{i=1}^{n}\left|\frac{1}{i/n}\right| \approx \int_{1/n}^{1}\frac{1}{x}dx$$

$$= \ln x|_{1/n}^{1} = \ln 1 - \ln\frac{1}{n} = \ln n.$$

This leads to another bound which is $O\left(p_i n^{2\epsilon_1-1}\log\left(np_i^{-1}e\right)\right)$. So, for $\alpha = 1$,

$$E\left(S_{ij}^2\right) \leq Cp_i n^{2\epsilon_1-1}\log\left(np_i^{-1}e\right). \tag{33.39}$$

Using a similar process we could derive that

$$E\left(R_j^2\right) \leq C_\star p_i n^{2\epsilon_1-1}\log\left(np_i^{-1}e\right). \tag{33.40}$$

33.7. **Bound of** $\sup_{ij} P\left(|B_{ij}| > C\right)$**.** We first note that

$$aI\left(|x| > a\right) \leq |x|$$

or

$$E\left(I\left(|x| \geq a\right)\right) = aP\left(|x| \geq a\right)$$

Thus

$$aP\left(|x| \geq a\right) \leq E\left(x\right).$$

So we now consider $P\left(|B_{ij}| > \epsilon n^{-1/2}\right)$. Because of (33.16), we know that for $j \in J_i(\epsilon)$

$$\epsilon n^{-1/2}P\left(|B_{ij}| > \epsilon n^{-1/2}\right) \leq Cn^{\eta-1}$$

$$P\left(|B_{ij}| > \epsilon n^{-1/2}\right) \leq \frac{C}{\epsilon}n^{\eta-1/2} = O\left(n^{-\lambda}\right) \tag{33.41}$$

for $0 < \lambda < 1/2$. Also, for $j \notin J_i(\epsilon)$

$$\epsilon n^{-1/2}P\left(|B_{ij}| > \epsilon n^{-1/2}\right) \leq Cp_i^{1/2}n^{\eta-1}$$

We must use the fact that $p_i = O\left(n^{1-\eta}\right)$.

$$P\left(|B_{ij}| > \epsilon n^{-1/2}\right) \leq \frac{C}{\epsilon}n^{-\eta/2} = O\left(n^{-\lambda}\right) \tag{33.42}$$

for $\lambda > 0$ and consequently, for $0 < \lambda < 1/2$.

33.8. **The bound of C_6 and $C_{6\star}$.** We spend a moment considering the bound of

$$C_6 = C_0 \int_0^1 \int_0^1 |x-y|^{-\alpha} \psi(x)\psi(y)dxdy. \tag{33.43}$$

We let $s = x - y$. Then $x = s + y$. We have

$$C_0 \int \int |s|^{-\alpha} \psi(s+y)\psi(y)dsdy.$$

We have a discontinuity at $s = 0$ which would require us to split this integral into two pieces and define limits.

$$C_0 I(s>0) \int \int s^{-\alpha}\psi(s+y)\psi(y)dsdy + C_0 I(s<0) \int \int (-s)^{-\alpha} \psi(s+y)\psi(s)dsdy$$

We consider the integral

$$\int \int s^{-\alpha}\psi(s+y)\psi(y)dsdy.$$

Let $z = \frac{s^{-\alpha+1}}{-\alpha+1}$. Then $dz = s^{-\alpha}ds$.

$$\int \int \psi\left(\left[(-\alpha+1)z\right]^{\frac{1}{-\alpha+1}} + y\right)\psi(y)\,dzdy.$$

This integral is infinite, but because the function ψ is compactly supported it converges . Thus C_6 is finite. In a similar way, $C_{6\star}$ is finite.

34. Bounding the Mean Square Error.

We are now in a position to bound the mean square error. We split the mean square error in the following way.

$$\int (g - \hat{g})^2 = A_1 + A_2 + A_3 + A_4 \tag{34.1}$$

where

$$A_1 \equiv \sum_j (\hat{a}_j - a_j)^2 \tag{34.2}$$

$$A_2 \equiv \sum_{i=0}^{q-1} \sum_j \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(\left|\hat{b}_{ij}\right| > \delta\right) \tag{34.3}$$

$$A_3 \equiv \sum_{i=0}^{q-1} \sum_j b_{ij}^2 I\left(\left|\hat{b}_{ij}\right| \leq \delta\right) \tag{34.4}$$

$$A_4 \equiv \sum_{i=q}^{\infty} \sum_j b_{ij}^2. \tag{34.5}$$

We will bound each of these in turn.

We note here that A_1 is the error in the ϕ coefficients, A_2 represents the $\hat{b}_{ij}$ which are large enough to keep, A_3 represents the $\hat{b}_{ij}$ which we throw away and A_4 represents the b_{ij} which were not estimated due to truncation.

34.1. **Bound for A_1.**

$$A_1 \equiv \sum_j \left(\hat{a}_j - a_j\right)^2 \tag{34.6}$$

Then

$$\left(\hat{a}_j - a_j\right)^2 = \left(\int_I \sum_m w_m(x) g\left(\mathcal{X}_m\right) \phi_j + \int_I \sum_m w_m(x) \epsilon_m \phi_j - \int_I g(x) \phi_j(x) \right)^2$$

$$= \left(\int_I \sum_m w_m(x) g\left(\mathcal{X}_m\right) \phi_j - \int_I g(x) \phi_j(x) + \int_I \sum_m w_m(x) \epsilon_m \phi_j \right)^2$$

$$= \left(A_j + n^{-1/2} R_j\right)^2 \leq 2A_j^2 + 2n^{-1} R_j^2.$$

Thus

$$E\left(\hat{a}_j - a_j\right)^2 \leq 2E\left(A_j^2\right) + 2n^{-1} E\left(R_j^2\right)$$

We also know that for $\alpha \in (0, 1]$

$$E\left(R_j^2\right) \leq C_\star p_i n^{2\epsilon_1 - 1} \log\left(n p_i^{-1} e\right).$$

Then if $j \in J_i(\epsilon)$,

$$E\left(\hat{a}_j^2 - a_j\right)^2 \leq C \left(\frac{p^{1/2}}{n}\right)^2 n^\eta + n^{-1} C_\star p_i n^{2\epsilon_1 - 1} \log\left(n p_i^{-1} e\right)$$

$$= O\left(p n^{\eta - 2} + n^{-1} E\left(R_j^2\right)\right).$$

We also have that if $j \notin J_i(\epsilon)$,

$$E\left(\hat{a}_j^2 - a_j\right)^2 \leq C n^{\eta - 2} + n^{-1} C_\star p n^{2\epsilon_1 - 1} \log\left(n p^{-1} e\right)$$

$$= O\left(n^{\eta-2} + n^{-1}E\left(R_j^2\right)\right).$$

We can conclude that

$$E\left(A_1\right) = \sum_j E\left(\hat{a}_j - a_j\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1-2}\log\left(np_i^{-1}e\right)\right). \tag{34.7}$$

34.2. **Bound for A_2.**

$$A_2 \equiv \sum_{i=0}^{q-1}\sum_j \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(\left|\hat{b}_{ij}\right| > \delta\right) \tag{34.8}$$

Recall that we have assumed that ψ is supported on $(-c, c)$. For the next two sections we let K_{i1} denote the set of indexes j that are contained in an interval $(p_i x - 2c, p_i x + 2c)$ for at least one of the discontinuity points x of at least one of the functions $g^{(0)}, ..., g^{(r)}$ and let K_{i2} be the set of all the other j's. We write $A_2 = A_{21} + A_{22}$ where

$$A_{2k} = \sum_{i=0}^{q}\sum_{j\in K_{ik}} \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(\left|\hat{b}_{ij}\right| > \delta\right).$$

We know that

$$\left(\hat{b}_{ij} - b_{ij}\right)^2 \leq 2\left(B_{ij}^2 + n^{-1}S_{ij}^2\right).$$

Now, for discontinuous indexes we have

$$\leq C\left(p_i n^{\eta-1} + p_i n^{2\epsilon_1-2}\log\left(np_i^{-1}e\right)\right) = O\left(p_i n^{\eta-1}\right). \tag{34.9}$$

Thus

$$E\left(A_{21}\right) = O\left\{q \sup_{0\leq i\leq q-1, j\in K_{i1}} E\left(\hat{b}_{ij} - b_{ij}\right)^2\right\} = O\left(qp_i n^{\eta-1}\right)$$

For continuous intervals we can come up with a better bound.

We note that $\left|\hat{b}_{ij}\right| \leq |b_{ij}| + |B_{ij}| + \left|\hat{\xi}_{ij}\right|$. Suppose that $|b_{ij}| \leq \epsilon\delta$. Note that

$$I\left(\left|\hat{b}_{ij}\right| > \delta\right) \leq I\left(|b_{ij}| + |B_{ij}| + \left|\hat{\xi}_{ij}\right| > \delta\right) \leq I\left(|B_{ij}| + \left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right).$$

Also note that $I\left(|a| + |b| > \epsilon\right) \leq I\left(|b| > \epsilon\right)$. So

$$I\left(|B_{ij}| + \left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right) \leq I\left(|B_{ij}| > (1-\epsilon)\delta\right)$$

$$I\left(|B_{ij}| + \left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right) \leq I\left(\left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right).$$

Now since ϵ is close to zero, we know that $1-3\epsilon < 1-\epsilon$ and also $\epsilon < 1-\epsilon$.

$$I\left(|B_{ij}| + \left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right) \leq I\left(|B_{ij}| > \epsilon\delta\right).$$

$$I\left(|B_{ij}| + \left|\hat{\xi}_{ij}\right| > (1-\epsilon)\delta\right) \leq I\left(\left|\hat{\xi}_{ij}\right| > (1-3\epsilon)\delta\right).$$

$$E\left\{\left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(\left|\hat{b}_{ij}\right| > \delta\right)\right\} \leq 2\left(B_{ij}^2 P\left(|B_{ij}| > \epsilon\delta\right) + n^{-1}S_{ij}^2 P\left(\left|\hat{\xi}_{ij}\right| > (1-3\epsilon)\delta\right)\right).$$

These quantities are all bounded. Let's suppose that $\delta = O\left(n^{-1/2}\right)$. (A reasonable assumption considering the bounds of b_{ij}. We will see in Section 34.4 that $|b_{ij}| \leq p_i^{-1}$.) We have

$$2\left(\frac{C}{\epsilon}n^{\eta-2}n^{-\lambda} + n^{-1}C_\star p_i n^{2\epsilon_1-1}\log\left(np_i^{-1}e\right)\exp\left(-\frac{u^2}{8\tau_{i*}^2}\right)\right)$$

for $0 < \lambda < 1/2$ and the τ_{i*}^2 from the earlier section. Here our bound is

$$E\left(A_{22}\right) = O\left(\frac{C}{\epsilon}n^{\eta-\lambda-2}\right).$$

Thus

$$E\left(A_2\right) = O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2}\right). \tag{34.10}$$

34.3. **Bound for A_3.** We divide this portion of the sum in a similar way. Let $A_3 = A_{31} + A_{32}$ where

$$A_{3k} = \sum_{i=0}^{q-1}\sum_{j\in K_{ik}} b_{ij}^2 I\left(\left|\hat{b}_{ij}\right| \leq \delta\right).$$

Note that $b_{ij}^2 \leq 2\left(\left(\hat{b}_{ij} - b_{ij}\right)^2 + \hat{b}_{ij}^2\right)$. Since the number of elements in K_{i1} is uniformly bounded (this is by assumption on the beginning), then from (33.16)

$$E\left(A_{31}\right) = O\left[\sum_{i=0}^{q-1}\left\{\sup_{j\in K_{i1}} E\left(\hat{b}_{ij} - b_{ij}\right)^2 + \delta^2\right\}\right]$$

$$= O\left(qn^{\eta-1}\right).$$

for all $\eta > 0$. By taking the Taylor expansion, we determine that

$$b_{ij} = p_i^{-(2r+1)/2}\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!} + o\left(p_i^{-(2r+1)/2}\right).$$

We will do this in more detail in Section 34.4. This is true uniformly in K_{i2} in (i,j). Thus

$$E\left(A_{32}\right)=\sum_{i=0}^{q-1}\sum_{j\in K_{i2}}b_{ij}^{2}P\left(\left|\hat{b}_{ij}\right|\leq\delta\right)$$

$$\sim\sum_{i=0}^{q}\sum_{j\in K_{i2}}b_{ij}^{2}$$

$$=\sum_{i=0}^{q}\sum_{j\in K_{i2}}\left\{p_{i}^{-(2r+1)}\left(\frac{g^{(r)}\left(\frac{j}{p_{i}}\right)}{r!}\right)^{2}+o\left(p_{i}^{-(2r+1)}\right)\right\}$$

$$=\left(1-2^{-2r}\right)^{-1}p_{i}^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_{i}}\right)}{r!}\right)^{2}+o\left(p^{-2r}\right).$$

Combining all of this yields

$$E\left(A_{3}\right)=\left(1-2^{-2r}\right)^{-1}p_{i}^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_{i}}\right)}{r!}\right)^{2}+o\left(p_{i}^{-2r}\right)+O\left(qn^{\eta-1}\right). \tag{34.11}$$

34.4. **Bound for A_4.** We divide this portion as before into two pieces. Let $A_4=A_{41}+A_{42}$ where

$$A_{4k}=\sum_{i=q}^{\infty}\sum_{j\in K_{ik}}b_{ij}^{2}.$$

Note that $|b_{ij}|=O\left(p_{i}^{-1}\right)$ uniformly in $j\in K_{i1}$. The number of these discontinuous points is bounded. This is because

$$b_{ij}=\int g(x)\psi_{ij}(x)dx.$$

Because ψ is bounded on $(-c,c)$, we have

$$-c\leq p_{i}x-j\leq c$$

$$\frac{-c+j}{p_{i}}\leq x\leq\frac{c+j}{p_{i}}.$$

So

$$|b_{ij}|\leq|g|\,|\psi_{ij}|\,\frac{2c}{p_{i}}=|g|\,\frac{2c}{p_{i}}=O\left(p_{i}^{-1}\right).$$

Therefore

$$A_{41} = O\left(\sum_{i=q}^{\infty} p_i^{-1}\right) = O\left(p_q^{-1}\right)$$

since $p_i = 2^i p$.

Now, also note that for $j \in K_{i2}$ where we have continuity, we can apply a Taylor expansion to g as follows. We expand about $\frac{j}{p_i}$.

$$g(x) = \frac{g^{(0)}(\frac{j}{p_i})}{0!} + \frac{g^{(1)}\left(\frac{j}{p_i}\right)}{1!}\left(x - \frac{j}{p_i}\right) + \frac{g^{(2)}\left(\frac{j}{p_i}\right)}{2!}\left(x - \frac{j}{p_i}\right)^2 + ... + \frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\left(x - \frac{j}{p_i}\right)^r + ...$$

Because of the vanishing moments of ψ, almost all of these terms disappear except for the ones that are at least $o(x^r)$.

$$b_{ij} = \int \left[x^r \frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!} + ...\right] = p_i^{-(2r+1)/2} \frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!} + o\left(p_i^{-(2r+1)/2}\right).$$

Thus for where we have continuity $|b_{ij}| = O\left(p_i^{-(2r+1)/2}\right)$. The number of such j's which don't vanish is $O(p_i)$. Thus

$$A_{42} = O\left(\sum_{i=q}^{\infty} p_i^{-2r}\right) = O\left(p_q^{-2r}\right).$$

Together this means that

$$A_4 = O\left(p_q^{-1}\right). \tag{34.12}$$

34.5. Final bound of $\int E(\hat{g} - g)^2$.

Theorem 70. *Suppose g is a function supported on $[0,1]$ with certain continuity properties established before. Suppose that the data generated by this function g is long memory and irregularly spaced. Long memory means that*

$$r(j) = E(\epsilon_i \epsilon_{i+j}) \sim C_0 |j|^{-\alpha}$$

where $\alpha \in (0,1]$. Suppose we let

$$\hat{g} = \sum_j \hat{a}_j \phi_j + \sum_{i=0}^{q-1} \sum_j \hat{b}_{ij} I\left(\left|\hat{b}_{ij}\right| \geq \delta\right) \psi_{ij}. \tag{34.13}$$

where

$$\hat{a}_j = \int_I Y\phi_j \qquad \hat{b}_{ij} = \int_I Y\psi_{ij}.$$

and Y is some interpolation rule. Then combing (34.7), (34.10), (34.11) and (34.12) yields for $\alpha = 1$

$$\int E\left(\hat{g}-g\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1-2}\log\left(np_i^{-1}e\right)\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2}\right)$$

$$+\left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}\right) + O\left(p_q^{-1}\right)$$

That is

$$= \left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2} + p_q^{-1}\right). \tag{34.14}$$

For $\alpha \in (0,1)$

$$\int E\left(\hat{g}-g\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1-1}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2}\right)$$

$$+\left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}\right) + O\left(p_q^{-1}\right)$$

That is

$$= \left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2} + p_q^{-1}\right). \tag{34.15}$$

35. Important notes about this paper.

This work has applied linear interpolation to irregularly spaced data which is normally distributed and has long memory error. The research could be used to compare two times series and be applied to a number of real-life data sets.

Part 8. Applying long memory error to the work of Cai and Brown in [4].

36. Introduction.

In this section we expand the work of Cai and Brown in [4] to include the error created by long memory error. As stated before, the problem of long memory has many real-life applications. Often data sets are not independent. In the previous section we addressed this problem with linear interpolation in the

space of functions g which were r piecewise continuous. Now we generalize that to the Holder class and use a function H to account for the irregularly spaced data. In Section 37 we examine some preliminaries and notations. In Section 38 we examine the initial definitions and the boundedness obtained from the coefficients of the mother wavelets. In Section 39 we bound the variance of the wavelet coefficients. In Section 41 we provide the initial separation of the wavelet coefficients into pieces. This is similar to the breakdown of wavelet coefficients in the Part 7 Section 33. Next, we provide the initial breakdown of the MISE in Section 42. In Sections 43, 44 and 45 each of the pieces of the MISE are bounded. Finally, we provide the overall bounds of the MISE in Section 46.

37. Preliminaries and Notations.

We are given data

$$y_i = f(t_i) + \epsilon z_i, \tag{37.1}$$

$i = 1, 2, ..., n$, $0 < t_1 < t_2 < ... < t_n = 1$, and z_i are long memory distributed as $N(0,1)$. These data are not equally spaced. We assume that $n = 2^{j_0}$.

Furthermore, long memory means that the covariance has the following property. We define

$$r(j) = E(\epsilon_i \epsilon_{i+j}) \sim C_0 |j|^{-\alpha} \tag{37.2}$$

for some constant C_0.

Assume $t_i = H^{-1}(i/n)$ for some cumulative density function H on [0,1]. Here ϵ is the noise level. We denote by $\Lambda^1(h)$ the collection of Lipschitz functions f satisfying

$$|f(x) - f(y)| \leq h|x - y| \quad \text{for } x, y \in [0,1]. \tag{37.3}$$

We assume that $H^{-1} \in \Lambda^1(h)$ for some constant h.

38. Preliminary information.

We are dealing with the following space in this work, just as in [4].

Definition 71. A piecewise Holder class $\Lambda^\alpha(M, B, m)$ on [0,1] with at most m discontinuous jumps consists of functions f satisfying the following conditions:

1. The function f is bounded by B, that is, $|f| \leq B$.

2. There exist $l \leq m$ points $0 \leq \alpha_1 < ... < \alpha_l \leq 1$ such that, for all $\alpha_i \leq x$, $y < \alpha_{i+1}$, $i = 0, 1, ..., l$ (with $\alpha_0 = 0$ and $\alpha_{l+1} = 1$),

(i) $|f(x) - f(y)| \leq M|x - y|^\alpha$ if $\alpha \leq 1$;

(ii) $\left|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)\right| \leq M|x - y|^{\alpha'}$ and $|f'(x)| \leq B$ if $\alpha > 1$

where $\lfloor \alpha \rfloor$ is the largest integer less that α and $\alpha' = \alpha - \lfloor \alpha \rfloor$.

We will assume for this work that $f \in \Lambda^{\beta}(M, B, m)$. This is because we have used the α constant to denote the boundedness that deals with long memory.

We assume of the wavelets that they have a multiresolution analysis which is set up in the usual way. Let $\psi_{jk}(x) = 2^{j/2}\psi\left(2^{j}x - k\right)$ and $\phi_{j_0k}(x) = 2^{j_0/2}\phi\left(2^{j_0}x - k\right)$. These will form the multiresolution analysis. Here the collection $\{\phi_{j_0k}, \psi_{jk}, j \geq j_0, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

We also have a theorem that we must cite from [4]. The proof does not change for long memory because this theorem does not take the error term ϵz_i into account.

Theorem 72. *Suppose that a sampled function* $\left\{f(t_i), i = 1, 2, ..., n\left(= 2^{J}\right)\right\}$ *is given with* $t_i = H^{-1}(i/n)$, *where* H *is a strictly increasing cumulative density function on [0,1] with* $H^{-1} \in \Lambda^1(h)$. *Let the wavelet function* ψ *be r-regular with* $r > \alpha$. *Let* ξ'_{Ji} *and* f_n *be given as previously. Then the approximation error* $\|f_n - f\|_2^2$ *satisfies*

$$\sup_{f \in \Lambda^{\beta}(M,B,m)} \|f_n - f\|_2^2 = o\left(n^{-2\beta/(1+2\beta)}\right),$$

where the maximum number of jump discontinuities $m = Cn^{\gamma}$ *with constants* $C > 0$ *and* $0 < \gamma < \beta/(1 + 2\beta)$.

Now some preliminary notation:

Let $\tilde{g}(t) = n^{-1/2}\sum_{i=1}^{n} y_i \phi_{Ji}(t)$ and let

$$\tilde{f}_J(t) = \mathrm{Proj}_{V_J}\tilde{g}(H(t)) = n^{-1/2}\sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^{j}} \tilde{\theta}_{jk}\psi_{jk}(t),$$

where

$$\tilde{\xi}_{jk} = n^{-1/2}\sum_{i=1}^{n} y_i \langle \phi_{Ji} \circ H, \phi_{jk}\rangle, \qquad \tilde{\theta}_{jk} = n^{-1/2}\sum_{i=1}^{n} y_i \langle \phi_{Ji} \circ H, \psi_{jk}\rangle.$$

Here $\tilde{\xi}_{j_0k}$ and $\tilde{\theta}_{jk}$are noisy observations of the true wavelet coefficients ξ_{j_0k} and θ_{jk}. We estimate θ_{jk} by thresholding $\tilde{\theta}_{jk}$. Let

$$\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k}, \qquad \hat{\theta}_{jk} = \mathrm{sgn}\left(\tilde{\theta}_{jk}\right)\left(\left|\tilde{\theta}_{jk}\right| - \lambda_{jk}\right)_{+} \tag{38.1}$$

where the threshold λ_{jk} is derived later. This is derived from an estimate of the variance of the wavelet coefficients. We will use the upper bounds of the variance and the universal bounds from Donoho and Johnstone as the thresholding coefficients.

We also have

$$f_n(t) = \sum_{i=1}^{2^J} \xi'_{Ji}\phi_{Ji}(t).$$

(We note here that when we examine this expression in light of Theorem 72, what the theorem actually gives us in the bound of just the wavelet coefficients in our expressions. This is what we will apply many times in the later sections.)

39. Bounds of Variance and error.

We suppose that H^{-1} is strictly increasing. Therefore H^{-1} is differentiable almost everywhere. Let

$$\tilde{h}(t) = \frac{d}{dt}H^{-1}(t).$$

Then

$$0 < \tilde{h}(t) \leq h$$

for almost all $t \in [0,1]$,by an assumption we make. We define

$$\sigma_{jk}^2 = var\left(\tilde{\theta}_{ij}\right) = \left(n^{-1/2}\epsilon \sum_{i=1}^{n} z_i \left\langle \phi_{Ji} \circ H, \psi_{jk} \right\rangle\right)^2$$

$$= \frac{\epsilon^2}{n}\sum_{i=1}^{n}\left(\left\langle \phi_{Ji} \circ H, \psi_{jk}\right\rangle\right)^2 + \frac{\epsilon^2}{n}\sum_{m=1}^{n}\sum_{m' \neq m} z_m z_{m'} \left\langle \phi_{Jm} \circ H, \psi_{jk}\right\rangle \left\langle \phi_{Jm'} \circ H, \psi_{jk}\right\rangle$$

$$\equiv J_1 + J_2.$$

We must bound each of there. Note that

$$E\left(J_1\right) \leq \frac{\epsilon^2}{n}\left[\int \phi_{Ji}\left(H\left(t\right)\right)\psi_{jk}\right]^2.$$

Note that $\|\phi_{Ji}\| = 1$.

$$E\left(J_1\right) \leq \frac{\epsilon^2}{n}\int \psi_{jk}^2\left(t\right)\tilde{h}\left(H\left(t\right)\right)dt \leq \frac{\epsilon^2}{n}\cdot h.$$

The factor of h appears because of the assumption of boundedness that we have made on H in (37.3). Now we must bound J_2.

$$E\left(J_2\right) = E\left(\frac{\epsilon^2}{n}\sum_{m=1}^{n}\sum_{m' \neq m} z_m z_{m'} \left\langle \phi_{Jm} \circ H, \psi_{jk}\right\rangle \left\langle \phi_{Jm'} \circ H, \psi_{jk}\right\rangle\right)$$

Recall that

$$r(j) = E\left(\epsilon_i \epsilon_{i+j}\right) \sim C_0 \left|j\right|^{-\alpha}.$$

$$E\left(J_2\right) \sim \frac{\epsilon^2}{n} \sum_{m=1}^{n} \sum_{m' \neq m} r\left(m - m'\right) \left\langle \phi_{Jm} \circ H, \psi_{jk} \right\rangle \left\langle \phi_{Jm'} \circ H, \psi_{jk} \right\rangle \equiv I_1. \tag{39.1}$$

Just as in Li and Xiao on page 2867 in [18] we have the following. We must bound 39.1).

(39.2)

$$I_1 = \frac{2^i \epsilon^2}{n} \sum_{m=1}^{n} \sum_{k \neq m} r\left(m - k\right) \int \int \phi\left(H\left(2^{j_0} x - m\right)\right) \phi\left(H\left(2^{j_0} y - k\right)\right) \psi\left(2^i x - j\right) \psi\left(2^i y - j\right) dxdy$$

We use the assumption that $n = 2^{j_0}$.

(39.3)

$$= \frac{2^i \epsilon^2}{n} \int \int \phi\left(H\left(u\right)\right) \phi\left(H\left(v\right)\right) \left\{ \sum_{m=1}^{n} \sum_{k \neq m} r\left(m - k\right) \psi\left(2^i \frac{u+m}{n} - j\right) \psi\left(2^i \frac{v+k}{n} - j\right) \frac{1}{n^2} \right\} dudv.$$

We first consider the case when $\alpha \in (0, 1)$. It follows from (50.2) that as $n \to \infty$,

$$\sum_{m=1}^{n} \sum_{k \neq m} r\left(m - k\right) \psi\left(2^i \frac{u+m}{n} - j\right) \psi\left(2^i \frac{v+k}{n} - j\right) \frac{1}{n^2}$$

$$\sim C_0 n^{-\alpha} \sum_{m=1}^{n} \sum_{k \neq m} \left|\frac{m}{n} - \frac{k}{n}\right|^{-\alpha} \psi\left(2^i \frac{u+m}{n} - j\right) \psi\left(2^i \frac{v+k}{n} - j\right) \frac{1}{n^2}$$

$$\sim C_0 n^{-\alpha} \int_0^1 \int_0^1 \left|x - y\right|^{-\alpha} \psi\left(2^i x - j\right) \psi\left(2^i y - j\right) dxdy$$

$$= C_0 n^{-\alpha} 2^{(\alpha-2)i} \int_0^1 \int_0^1 \left|x - y\right|^{-\alpha} \psi\left(x\right) \psi\left(y\right) dxdy. \tag{39.4}$$

uniformly for all u, v in the support of ϕ. Combining (39.2) and (39.4) we have

$$I_1 \sim C_4 n^{-\alpha} 2^{-i(1-\alpha)} \quad \text{as } n \to \infty. \tag{39.5}$$

We have the following bound.

$$\leq \int \int \epsilon^2 h \phi(w) \phi(v) \times$$

$$\left\{ \frac{2^j}{n} \sum_{m=1}^{n} \sum_{m'=m} r\left(\left|m - m'\right|\right) \psi\left(2^j \frac{u+m}{n} - k\right) \psi\left(2^j \frac{v+m'}{n} - k\right) \right\} dudv$$

We represent the second half of this equation with a I_2.

Now, just as in Li and Xiao's result in [18], we have

$$I_2 \leq C\tau_j^2$$

where

$$\tau_j^2 \leq C_4 n^{-\alpha} 2^{-j(1-\alpha)}$$

and

$$C_4 = C_0 \int_0^1 \int_0^1 |x-y|^{-\alpha} \psi(x)\psi(y)dxdy.$$

Thus, for all $\alpha \in (0,1)$,

$$E(J_2) \leq \epsilon^2 hC\tau_j^2.$$

Here we reproduce the Li and Xiao bounding in [18] for $\alpha = 1$. We begin from (39.3). We consider

$$2^i \sum_{m=1}^{n} \sum_{|m-k|>c} r(m-k) \int\int \phi\left(2^{j_0}x - m\right) \phi\left(2^{j_0}y - k\right) \psi\left(2^i x - j\right) \psi\left(2^i y - j\right) dxdy$$

$$\sim 2^i \int\int \phi(H(u)) \phi(H(v)) \left\{ \sum_{m=1}^{n} \sum_{|m-k|>c} r(m-k) \psi\left(2^i \frac{u+m}{n} - j\right) \psi\left(2^i \frac{v+k}{n} - j\right) \frac{1}{n^2} \right\} dudv.$$

They apply an argument from Hall and Hart, 1990, p.350. We have

$$\sum_{m=1}^{n} \sum_{|m-k|>c} r(m-k) \psi\left(2^i \frac{u+m}{n} - j\right) \psi\left(2^i \frac{v+k}{n} - j\right) \frac{1}{n^2}$$

$$\sim C_0 n^{-1} \int\int_{|x-y|>c/n} |x-y|^{-1} \psi\left(2^i x + 2^i u/n\right) \psi\left(2^i y + 2^i v/n\right) dxdy$$

$$\sim C_0 n^{-1} 2^{-1} \int\int_{|p-q|>c2^i/n} |p-q|^{-1} \psi(p) \psi(q) dpdq$$

$$= C_0 n^{-1} 2^{-i} \int \psi(q) \int_{|y|>c2^i/n} |y|^{-1} \psi(y+q) dydq$$

$$\sim C_0 n^{-1} 2^{-i} 2 \log\left(n2^{-i}e\right) \int \psi^2(q) dq$$

$$= 2C_0 n^{-1} 2^{-i} \log\left(n2^{-i}e\right). \tag{39.6}$$

Now for $\alpha = 1$, where we have in terms of the Li and Xiao paper, $j = j'$, as per equation (??),

$$I_2 \leq Cn^{-1} \log\left(n2^{-k}e\right).$$

Then

$$E\left(J_2\right) \leq \epsilon^2 hCn^{-1} \log\left(n2^{-k}e\right).$$

Then for $\alpha \in (0,1)$,

$$var\left(\tilde{\theta}_{ij}\right) \leq \frac{\epsilon^2}{n}h + \epsilon^2 hC\tau_j^2$$

$$= \frac{\epsilon^2}{n}h + \epsilon^2 hCC_4 n^{-\alpha} 2^{-k(1-\alpha)}$$

$$= O\left(n^{-\alpha}\right). \tag{39.7}$$

Now for $\alpha = 1$,

$$var\left(\tilde{\theta}_{ij}\right) \leq \frac{\epsilon^2}{n}h + \frac{\epsilon^2}{n}hC\log\left(n2^{-k}e\right)$$

$$= O\left(\frac{\log\left(n2^{-k}e\right)}{n}\right). \tag{39.8}$$

Thus, we have the bounds for the error terms. We summarize them in the theorem below.

Theorem 73. *The bounds of the variance* $var\left(\tilde{\theta}_{ij}\right)$ *are for* $\alpha \in (0,1)$,

$$= \frac{\epsilon^2}{n}h + \epsilon^2 hCC_4 n^{-\alpha} 2^{-k(1-\alpha)}$$

$$= O\left(n^{-\alpha}\right). \tag{39.9}$$

and for $\alpha = 1$,

$$\leq \frac{\epsilon^2}{n}h + \frac{\epsilon^2}{n}hC\log\left(n2^{-k}e\right)$$

$$= O\left(\frac{\log\left(n2^{-k}e\right)}{n}\right). \tag{39.10}$$

40. Other important notation.

We note that

$$\sigma_{jk}^2 = var\left(\tilde{\theta}_{ij}\right) = \left(n^{-1/2}\epsilon\sum_{i=1}^{n} z_i \left\langle \phi_{Ji} \circ H, \psi_{jk} \right\rangle\right)^2.$$

I will denote

$$\hat{\sigma}_{jk}^2 = var\left(\tilde{\xi}_{ij}\right) = \left(n^{-1/2}\epsilon\sum_{i=1}^{n} z_i \left\langle \phi_{Ji} \circ H, \phi_{j_0 k}\right\rangle\right)^2.$$

Note that the same bounds apply to this $\hat{\sigma}_{jk}$ as to σ_{jk}.

41. Breakdown of Wavelet Coefficients

Let $g(t) = f\left(H^{-1}(t)\right)$ and $\tilde{g}(t) = n^{-1/2}\sum_{i=1}^{n} y_i \phi_{Ji}(t)$ and let $\tilde{f}(t) = \tilde{g}\left(H(t)\right)$. Then

$$\tilde{f}(t) = n^{-1/2}\sum_{i=1}^{n} f(t_i)\,\phi_{Ji}\left(H(t)\right) + n^{-1/2}\epsilon\sum_{i=1}^{n} z_i \phi_{Ji}\left(H(t)\right)$$

$$= f(t) + \Delta(t) + r(t),$$

where $\Delta(t) = n^{-1/2}\sum_{i=1}^{n} f(t_i)\,\phi_{Ji}\left(H(t)\right) - f(t)$ is the approximation error and $r(t) = n^{-1/2}\epsilon\sum_{i=1}^{n} z_i \phi_{Ji}\left(H(t)\right)$ is the error due to the noise in the data. Now project $\tilde{f}$ onto the multiresolution space V_J and decompose the orthogonal projection $\tilde{f}_J(t) = \mathrm{Proj}_{V_J}\tilde{f}(t)$ into three terms:

$$\tilde{f}_J(t) = f_J(t) + \Delta_J(t) + r_J(t),$$

where $f_J = \mathrm{Proj}_{V_J} f$, $\Delta_J = \mathrm{Proj}_{V_J}\Delta$ and $r_J = \mathrm{Proj}_{V_J} r$, respectively. Theorem 72 yields

$$\|\Delta_J\|_2^2 = o\left(n^{-2\beta/(1+2\beta)}\right).$$

Then we finally have the decomposition. Denote $\tilde{\theta}_{jk} = \left\langle \tilde{f}_J, \psi_{jk}\right\rangle$. We decompose $\tilde{\theta}_{jk}$ into three parts:

$$\tilde{\theta}_{jk} = \theta_{jk} + d_{jk} + r_{jk} \text{ for } k = 1, ..., 2^j, \quad j = j_0, ..., J-1,$$

where $\theta_{jk} = \langle f, \psi_{jk}\rangle$ is the true wavelet coefficient of f, $d_{jk} = \langle \Delta_J, \psi_{jk}\rangle$ is the approximation error and $r_{jk} = \langle r_J, \psi_{jk}\rangle$ is the noise. Similarly separate $\tilde{\xi}_{j_0 k} = \left\langle \tilde{f}_J, \phi_{j_0 k}\right\rangle$ into three terms:

$$\tilde{\xi}_{j_0 k} = \xi_{j_0 k} + d'_{j_0 k} + r'_{j_0 k} \text{ for } k = 1, ..., 2^{j_0}.$$

Lemma 74. *Let $\hat{\xi}_{j_0 k}$ and $\hat{\theta}_{jk}$ be given as in (66.3). Then*

$$\sum_{k=1}^{2^{j_0}} \left(d'_{j_0 k}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} d_{jk}^2 = \|\Delta_J\|_2^2 = o\left(n^{-2\beta/(1+2\beta)}\right).$$

This Lemma will be very important in bounding each of the pieces of the MISE later.

42. Breakdown of the MISE.

Recall that the MISE is measured by the global squared L_2 norm risk:

$$E\int_0^1 \left(\hat{f}(t) - f(t)\right)^2 dt.$$

We break up the MISE as follows.

$$E\left\|\hat{f}_n^* - f\right\|^2 = \sum_{k=1}^{2^{j_0}} E\left(\hat{\xi}_{j_0k} - \xi_{j_0k}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} E\left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 + \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}\theta_{jk}^2$$

$$\equiv S_1 + S_2 + S_3.$$

We will bound each of these below.

43. Bound of S_1.

Recall that

$$S_1 = \sum_{k=1}^{2^{j_0}} \left(\hat{\xi}_{j_0k} - \xi_{j_0k}\right)^2$$

$$\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k} = \xi_{j_0k} + d'_{j_0k} + r'_{j_0k}$$

$$= \sum_{k=1}^{2^{j_0}} \left(d'_{j_0k} + r'_{j_0k}\right)^2$$

$$\leq 2^{j_0}\left[\text{bounds of } \hat{\sigma}^2_{j_0k}\right] + \sum_{k=1}^{2^{j_0}} \left(d'_{j_0k}\right)^2$$

$$= 2^{j_0}\left[\text{bounds of } \hat{\sigma}^2_{j_0k}\right] + O\left(n^{-2\beta/(1+2\beta)}\right).$$

This last line comes from Theorem 72 and the β comes from the definition of the space being examined.

44. Bounds of S_3.

Recall that

$$S_3 = \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}\theta_{jk}^2.$$

There are the wavelet coefficients that we never estimate because we must have a truncation point for the series. We define

$$G_j = \left\{k : supp\left[\psi_{jk}\right] = \left[2^{-j}k, 2^{-j}(N+k)\right] \text{ contains at least one jump point of } f\right\}.$$

Then $card\,(G_j) \leq N\,(m+2)$, here the m comes from the definition and the plus 2 comes from including the 0 and the 1, the endpoints of the support.

We have a lemma:

Lemma 75. *We have for the coefficients* θ_{jk}

$$|\theta_{jk}| \leq C2^{-j(-1/2+\beta)} \quad k \notin G_j$$

$$|\theta_{jk}| \leq C2^{-j/2} \qquad k \in G_j.$$

Then

$$S_3 = \sum_{j=J}^{\infty} \sum_{k \in G_j} \theta_{jk}^2 + \sum_{j=J}^{\infty} \sum_{k \notin G_j} \theta_{jk}^2.$$

$$\leq \sum_{j=J}^{\infty} N(m+2)C^2 2^{-j} + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} C^2 2^{-j(1+2\beta)}$$

$$= 2^{-J} \sum_{j=1}^{\infty} N(m+2)C^2 2^{-j} + \sum_{j=1}^{\infty} 2^j C^2 2^{-j} 2^{-j2\beta}$$

$$= 2^{-J} N(m+2)C^2 + 2^{-J} \sum_{j=1}^{\infty} C^2 2^{-j2\beta}$$

$$2^{-J} C^2 \left[N(m+2) + \frac{1}{1 - \frac{1}{2^{2\beta}}} \right] = 2^{-J} C^2 \left[N(m+2) + \frac{2^{2\beta}}{2^{2\beta} - 1} \right]$$

Recall that $n = o\left(2^J\right)$. Each of these is bounded above by the following.

$$S_3 = o\left(n^{-2\beta/(1+2\beta)}\right).$$

45. Bound of S_2.

Recall that

$$S_2 = \sum_{j=j_0}^{J-1} \left(\hat{\theta}_{jk} - \theta_{jk}\right)^2.$$

We also have

$$\hat{\theta}_{jk} = \theta_{jk} + d_{jk} + r_{jk}.$$

Recall that there are thresholded using the hard and soft thresholding operators. That means that we can divide this S_2 in four ways. By whether we are in or not in G_j, and by whether or not we keep or kill the coefficients.

Let's begin by considering hard thresholding.

$$S_2 = \sum_{j=j_o}^{J-1} \sum_{k\in G_j} \left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \left[\hat{\theta}_{jk} \text{ keep }\right] + \sum_{j=j_o}^{J-1} \sum_{k\in G_j} \left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \left[\hat{\theta}_{jk} \text{ kill }\right]$$

$$+ \sum_{j=j_o}^{J-1} \sum_{k\notin G_j} \left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \left[\hat{\theta}_{jk} \text{ keep }\right] + \sum_{j=j_o}^{J-1} \sum_{k\notin G_j} \left(\hat{\theta}_{jk} - \theta_{jk}\right)^2 \left[\hat{\theta}_{jk} \text{ kill }\right]$$

$$= \sum_{j=j_o}^{J-1} \sum_{k\in G_j} (d_{jk} + r_{jk})^2 \left[\hat{\theta}_{jk} \text{ keep }\right] + \sum_{j=j_o}^{J-1} \sum_{k\notin G_j} (d_{jk} + r_{jk})^2 \left[\hat{\theta}_{jk} \text{ keep }\right]$$

$$+ \sum_{j=j_o}^{J-1} \sum_{k\in G_j} \theta_{jk}^2 \left[\hat{\theta}_{jk} \text{ kill }\right] + \sum_{j=j_o}^{J-1} \sum_{k\notin G_j} \theta_{jk}^2 \left[\hat{\theta}_{jk} \text{ kill }\right].$$

$$S_{21} + S_{22} + S_{23} + S_{24}.$$

We consider $S_{21} + S_{22}$.

$$= \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} (d_{jk} + r_{jk})^2 = (J - j_0)\, 2^j \left[\text{bounds of } \sigma_{jk}^2\right] + o\left(n^{-2\beta/(1+2\beta)}\right).$$

Lastly we consider $S_{23} + S_{24}$.

$$\sum_{j=j_o}^{J-1} \sum_{k\in G_j} \theta_{jk}^2 + \sum_{j=j_o}^{J-1} \sum_{k\notin G_j} \theta_{jk}^2 \le N(m+2) \sum_{j=j_0}^{J-1} C2^{-j/2} + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} C2^{-j(1+2\beta)}$$

$$N(m+2) \sum_{j=j_0}^{J-1} C2^{-j/2} + \sum_{j=j_0}^{J-1} 2^j C2^{-j(1+2\beta)} = N(m+2) \sum_{j=j_0}^{J-1} C2^{-j/2} + \sum_{j=j_0}^{J-1} C2^{-j2\beta}$$

Here we use a substitution, $l = j - j_0 + 1$. Then if $j = j_0$, $l = 1$. Also, if $j = J - 1$, $l = J - j_0$. We have

$$= CN(m+2) \sum_{l=1}^{J-j_0} 2^{-(l+j_0-1)/2} + C \sum_{l=1}^{J-j_0} 2^{-2\beta(l+j_0-1)}$$

$$= CN(m+2) 2^{(1-j_0)/2} \sum_{l=1}^{J-j_0} 2^{-l/2} + 2^{2\beta(1-j_0)} \sum_{l=1}^{J-j_0} 2^{-2\beta l}$$

We will use the rule of infinite sums for $|x| < 1$.

$$\sum_{m=1}^{\infty} x^m = \frac{1}{1-x}.$$

$$\leq CN(m+2)2^{(1-j_0)/2}\frac{1}{1-2^{-1/2}} + 2^{2\beta(1-j_0)}\frac{1}{1-2^{-2\beta}}.$$

By recalling that $n = 2^J$, we can see that this is

$$o\left(n^{-2\beta/(1+2\beta)}\right).$$

46. Overall Bounds.

Now we can compute the overall abounds of the MISE.

For $\alpha \in (0,1)$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[n^{-\alpha}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[n^{-\alpha}\right]$$

$$= o\left(n^{-\alpha}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

For $\alpha = 1$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right).$$

$$= o\left(\frac{\log\left(n2^{-k}e\right)}{n}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

We state this as a theorem below.

Theorem 76. *Suppose f is a function supported on $[0,1]$ with $f \in \Lambda^\beta(M,B,m)$. Suppose that the data generated by this function f is long memory and irregularly spaced. Long memory means that*

$$r(j) = E\left(\epsilon_i\epsilon_{i+j}\right) \sim C_0\left|j\right|^{-\alpha}$$

where $\alpha \in (0,1]$. Let

$$\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k}, \qquad \hat{\theta}_{jk} = sgn\left(\tilde{\theta}_{jk}\right)\left(\left|\tilde{\theta}_{jk}\right| - \lambda_{jk}\right)_+ \tag{46.1}$$

where the threshold λ_{jk} is derived from an estimate of the variance of the wavelet coefficients.

For $\alpha \in (0,1)$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[n^{-\alpha}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[n^{-\alpha}\right]$$

$$= o\left(n^{-\alpha}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

For $\alpha = 1$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right).$$

$$= o\left(\frac{\log\left(n2^{-k}e\right)}{n}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

47. Important notes about this paper.

Here we have studied the MISE which arises from applying a function H to reorder irregularly spaced data which is normally distributed and has long memory error. The bounds are similar to those obtained in Part 7. We compare these two bounds and examine the implications in the following Part 9.

Part 9. Comparison of the results of Part 7 and Part 8.

48. Introduction.

In this section we compare the results in Parts 7 and 8. Ultimately, we compare the convergence of the MISE in the two different situations. Section 49 examines and summarizes the results from Part 7, and Section 50 examines and summarizes the results from Part 8. Finally, in Section 51 we give the comparison of the two.

49. First space and Theorem.

49.1. **Initial assumptions.** We have

$$Y_m = g\left(X_m\right) + \epsilon_m \quad \text{for } 1 \leq m \leq n \tag{49.1}$$

where $\mathcal{Y} = \{(X_m, Y_m), 1 \leq m \leq n\}$ and $\mathcal{X} = \{X_m, 1 \leq m \leq n\}$.

The ϵ_m are long memory dependent Gaussian variables with $E\left(\epsilon_m\right) = 0$ and $E\left(\epsilon_m^2\right) = \sigma^2 > 0$. Long memory means that the covariance has the following property. We define

$$r(j) = E\left(\epsilon_i \epsilon_{i+j}\right) \sim C_0\left|j\right|^{-\alpha}$$

where $\alpha \in (0,1]$. Here $a_j \sim b_j$ means that $a_j/b_j \to 1$ when $j \to \infty$. Here we interpolate the data with a function

$$Y(x) = \sum_m w_m(x) Y_m \quad \text{for} \quad x \in (X_{-v_1}, X_{n-v_2}]. \tag{49.2}$$

The weights are defined as follows.

For local averaging we let

$$w_m(x) = (2v)^{-1} \quad \text{if} \quad -v+1 \le m-l \le v, \quad 0 \text{ otherwise.} \tag{49.3}$$

For linear interpolation we let

$$w_m(x) = \begin{cases} v^{-1}\left(X_{2l-m+1} - x\right) / \left(X_{2l-m+1} - X_m\right) & -v+1 \le m-l \le 0 \\ v^{-1}\left(x - X_{2l-m+1}\right) / \left(X_m - X_{2l-m+1}\right) & 1 \le m-l \le v \\ 0 & \text{otherwise.} \end{cases} \tag{49.4}$$

We make the following assumptions about g.

49.2. **Assumptions on g.** We assume of the function g that it has r piecewise continuous derivatives, in the sense that there exist constants $0 = a_1 < a_2 < ... < a_k = 1$ such that g has r continuous derivatives on each interval $[a_l, a_{l+1}]$ for $1 \le l \le k-1$. We assume the same of the density function f of the $\mathcal{X}_m$'s, possibly with different a_i's and a different k. Thus, the the function g and its derivatives have a bounded number of discontinuities.

49.3. **Boundedness of our estimator of g.** We come to the following theorem regarding the boundedness for our estimator of g. Below, $\epsilon_1, \eta > 0$ and close to 0, and $0 < \lambda < 1/2$.

Theorem 77. *Suppose g is a function supported on $[0,1]$ with certain continuity properties established before. Recall that g has r piecewise continuous derivatives. Suppose that the data generated by this function g is long memory and irregularly spaced. Long memory means that*

$$r(j) = E\left(\epsilon_i \epsilon_{i+j}\right) \sim C_0 \left|j\right|^{-\alpha}$$

where $\alpha \in (0,1]$. Suppose we let

$$\hat{g} = \sum_j \hat{a}_j \phi_j + \sum_{i=0}^{q-1} \sum_j \hat{b}_{ij} I\left(\left|\hat{b}_{ij}\right| \ge \delta\right) \psi_{ij}. \tag{49.5}$$

where

$$\hat{a}_j = \int_I Y \phi_j \qquad \hat{b}_{ij} = \int_I Y \psi_{ij}.$$

and Y is some interpolation rule. Then for $\alpha = 1$

$$\int E\left(\hat{g}-g\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1-2}\log\left(np_i^{-1}e\right)\right)$$

$$+\left(1-2^{-2r}\right)^{-1} p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}+p_q^{-1}\right)$$

$$+O(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2}) \tag{49.6}$$

For $\alpha \in (0,1)$

$$\int E\left(\hat{g}-g\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1-1}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon}n^{\eta-\lambda-2}\right)$$

$$+\left(1-2^{-2r}\right)^{-1} p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}+p_q^{-1}\right) \tag{49.7}$$

Let $\epsilon = \max\{2\epsilon_1, \eta, \lambda\}$. We rewrite our bounds.

For $\alpha = 1$

$$\int E\left(\hat{g}-g\right)^2 = O\left(pn^{\epsilon-2} + p_i n^{\epsilon-2}\log\left(np_i^{-1}e\right)\right)$$

$$+\left(1-2^{-2r}\right)^{-1} p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\epsilon-1}+p_q^{-1}\right)$$

$$+O(qp_i n^{\epsilon-1} + \frac{C}{\epsilon}n^{\epsilon-2}).$$

For $\alpha \in (0,1)$

$$\int E\left(\hat{g}-g\right)^2 = O\left(pn^{\epsilon-2} + p_i n^{\epsilon-1}\right) + O\left(qp_i n^{\epsilon-1} + \frac{C}{\epsilon}n^{\epsilon-2}\right)$$

$$+\left(1-2^{-2r}\right)^{-1} p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\epsilon-1}+p_q^{-1}\right).$$

Now let's recall the properties of p, p_i and q. These were used to define the Multiresolution Analysis of the wavelets within Part 7.

$$p^{-1} = o\left\{\left(n^{-1}\log n\right)^{1/(2r+1)}\right\}, \qquad p_q^{-1} = o\left(n^{-2r/(2r+1)}\right),$$

$$p_q = O\left\{n^{\min\{\mu+1/(2r+1),1\}-\epsilon}\right\}$$

where $\mu > 0$ and $\epsilon > 0$.

$$p_i = 2^i p.$$

Here q is the truncation point for the series. We adjust what we have according to these parameters. Below is a reminder of big O and little o notations.

49.4. **Little o versus Big O.** We say that $f(x) = O(g(x))$ as $x \to \infty$ if and only if there is a positive constant M such that for all sufficiently large values of x,

$$|f(x)| \leq M|g(x)| \text{ for all } x > x_0.$$

We say that $f(x) = o(g(x))$ if for every constant ϵ there exists an x_0 such that

$$|f(x)| \leq \epsilon|g(x)| \text{ for all } x > x_0.$$

49.5. **Continuing to simplify.** We enter in the bounds of the p's and q's below.

For $\alpha = 1$

$$\int E(\hat{g}-g)^2 = O\left(pn^{\epsilon-2} + 2^i pn^{\epsilon-2}\log\left(n\left(2^i p\right)^{-1}e\right)\right)$$

$$+\left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(\left(\left(n^{-1}\log n\right)^{1/(2r+1)}\right)^{2r}\right) + O\left(qn^{\epsilon-1} + n^{-2r/(2r+1)}\right)$$

$$+O(q2^i pn^{\epsilon-1} + \frac{C}{\epsilon}n^{\epsilon-2}).$$

For $\alpha \in (0,1)$

$$\int E(\hat{g}-g)^2 = O\left(pn^{\epsilon-2} + 2^i pn^{\epsilon-1}\right) + O\left(q2^i pn^{\epsilon-1} + \frac{C}{\epsilon}n^{\epsilon-2}\right)$$

$$+\left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(\left(\left(n^{-1}\log n\right)^{1/(2r+1)}\right)^{2r}\right) + O\left(qn^{\epsilon-1} + n^{-2r/(2r+1)}\right).$$

We continue the simplifying by eliminating the lesser terms.

Theorem 78. *Our new simplified bounds are listed below. For* $\alpha = 1$

$$\int E(\hat{g}-g)^2 = O\left(2^i pn^{\epsilon-2}\log\left(n\left(2^i p\right)^{-1}e\right) + 2^i pqn^{\epsilon-1} + n^{-2r/(2r+1)}\right)$$

$$+\left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2+o\left(\left(\left(n^{-1}\log n\right)^{1/(2r+1)}\right)^{2r}\right). \tag{49.8}$$

For $\alpha\in(0,1)$

$$\int E\left(\hat{g}-g\right)^2=O\left(2^ipqn^{\epsilon-1}+n^{-2r/(2r+1)}\right)$$

$$+\left(1-2^{-2r}\right)^{-1}p^{-2r}\int\left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2+o\left(\left(\left(n^{-1}\log n\right)^{1/(2r+1)}\right)^{2r}\right). \tag{49.9}$$

This is the fully simplified theorem. I have left the factor of p^{-2r} in the term with the integral since it is an exact term.

50. Second Space and Theorem

50.1. **Initial assumptions.** We are given data

$$y_i=f\left(t_i\right)+\epsilon z_i, \tag{50.1}$$

$i=1,2,...,n$, $0<t_1<t_2<...<t_n=1$, and z_i are long memory distributed as $N\left(0,1\right)$. These data are not equally spaced. We assume that $n=2^{j_0}$.

Furthermore, long memory means that the covariance has the following property. We define

$$r(j)=E\left(\epsilon_i\epsilon_{i+j}\right)\sim C_0\left|j\right|^{-\alpha} \tag{50.2}$$

for some constant C_0.

Assume $t_i=H^{-1}\left(i/n\right)$ for some cumulative density function H on [0,1]. Here ϵ is the noise level. We denote by $\Lambda^1(h)$ the collection of Lipschitz functions f satisfying

$$\left|f\left(x\right)-f\left(y\right)\right|\leq h\left|x-y\right|\quad\text{for } x,y\in[0,1].$$

We assume that $H^{-1}\in\Lambda^1(h)$ for some constant h.

50.2. **Definition of the space.**

Definition 79. A piecewise Holder class $\Lambda^\alpha\left(M,B,m\right)$ on [0,1] with at most m discontinuous jumps consists of functions f satisfying the following conditions:

1. The function f is bounded by B, that is, $|f|\leq B$.

2. There exist $l\leq m$ points $0\leq\alpha_1<...<\alpha_l\leq1$ such that, for all $\alpha_i\leq x$, $y<\alpha_{i+1}$, $i=0,1,...,l$ (with $\alpha_0=0$ and $\alpha_{l+1}=1$),

(i) $|f(x)-f(y)|\leq M\left|x-y\right|^\alpha$ if $\alpha\leq1$;

(ii) $\left|f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(y)\right| \leq M|x-y|^{\alpha'}$ and $|f'(x)| \leq B$ if $\alpha > 1$

where $\lfloor\alpha\rfloor$ is the largest integer less that α and $\alpha' = \alpha - \lfloor\alpha\rfloor$.

We will assume for this work that $f \in \Lambda^{\beta}(M, B, m)$. We now examine the final theorem for this work.

50.3. **Preliminary notions.** We have the following theorem.

Theorem 80. *Suppose that a sampled function* $\left\{f(t_i), i = 1, 2, ..., n\left(= 2^J\right)\right\}$ *is given with* $t_i = H^{-1}(i/n)$, *where* H *is a strictly increasing cumulative density function on [0,1] with* $H^{-1} \in \Lambda^1(h)$. *Let the wavelet function* ψ *be r-regular with* $r > \alpha$. *Let* ξ'_{Ji} *and* f_n *be given as previously. Then the approximation error* $\|f_n - f\|_2^2$ *satisfies*

$$\sup_{f \in \Lambda^{\beta}(M,B,m)} \|f_n - f\|_2^2 = o\left(n^{-2\beta/(1+2\beta)}\right),$$

where the maximum number of jump discontinuities $m = Cn^{\gamma}$ *with constants* $C > 0$ *and* $0 < \gamma < \beta/(1+2\beta)$.

Let $\tilde{g}(t) = n^{-1/2}\sum_{i=1}^{n} y_i \phi_{Ji}(t)$ and let

$$\tilde{f}_J(t) = \mathrm{Proj}_{V_J}\tilde{g}(H(t)) = n^{-1/2}\sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \tilde{\theta}_{jk}\psi_{jk}(t),$$

where

$$\tilde{\xi}_{jk} = n^{-1/2}\sum_{i=1}^{n} y_i \langle \phi_{Ji} \circ H, \phi_{jk}\rangle, \qquad \tilde{\theta}_{jk} = n^{-1/2}\sum_{i=1}^{n} y_i \langle \phi_{Ji} \circ H, \psi_{jk}\rangle.$$

Here $\tilde{\xi}_{j_0 k}$ and $\tilde{\theta}_{jk}$are noisy observations of the true wavelet coefficients $\xi_{j_0 k}$ and θ_{jk}. We estimate θ_{jk} by thresholding $\tilde{\theta}_{jk}$. Let

$$\hat{\xi}_{j_0 k} = \tilde{\xi}_{j_0 k}, \qquad \hat{\theta}_{jk} = \mathrm{sgn}\left(\tilde{\theta}_{jk}\right)\left(\left|\tilde{\theta}_{jk}\right| - \lambda_{jk}\right)_{+} \tag{50.3}$$

where the threshold λ_{jk} is derived from an estimate of the variance of the wavelet coefficients.

50.4. **Boundedness for our estimator of** g**.** We have the following theorem.

Theorem 81. *Suppose* f *is a function supported on* $[0,1]$ *with* $f \in \Lambda^{\beta}(M, B, m)$. *Suppose that the data generated by this function* f *is long memory and irregularly spaced. Long memory means that*

$$r(j) = E(\epsilon_i \epsilon_{i+j}) \sim C_0 |j|^{-\alpha}$$

where $\alpha \in (0,1]$. *Let*

$$\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k}, \qquad \hat{\theta}_{jk} = sgn\left(\tilde{\theta}_{jk}\right)\left(\left|\tilde{\theta}_{jk}\right| - \lambda_{jk}\right)_+ \tag{50.4}$$

where the threshold λ_{jk} *is derived from an estimate of the variance of the wavelet coefficients.*

For $\alpha \in (0,1)$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[n^{-\alpha}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[n^{-\alpha}\right]$$

$$= o\left(n^{-\alpha}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

For $\alpha = 1$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right).$$

$$= o\left(\frac{\log\left(n2^{-k}e\right)}{n}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

We see that this boundedness is slightly different.

51. Comparison of the Two.

Notice that the first space is a subset of the second one. The second space allows for a finite number of discontinuities. The first space requires r piecewise continuous derivatives, the second only requires boundedness. Thus, the results for the first space are for a much more specific situation. Therefore, the bounds for the second space are much more general and could be applied in more situations.

The first theorem uses interpolation to correct the problem of unequally spaced data. We considered two specific rules, that of local averaging and linear interpolation. It uses tuning parameters p and q to define the multi-resolution analysis. Details can be found in Part 7 Section 32.3.

The second uses an inverse function H to correct the problem of unequally spaced data. The multiresolution analysis in this case is defined as it usually is. Details can be found in Part 8 Section 38.

We first consider the case where $\alpha \in (0,1)$.

The first error is either $O\left(2^i pqn^{\epsilon-1}\right)$ or $o\left(n^{-2r/(2r+1)}\right)$ where $\epsilon > 0$ and close to zero, whichever is larger.

The second error is either $o\left(\left[2^{j_0} + (J-j_0)\,2^j\right]n^{-\alpha}\right)$ or $o\left(n^{-2\beta/(1+2\beta)}\right)$, whichever is larger.

The main difference between the two of these is the factor of $n^{\epsilon-1}$ versus $n^{-\alpha}$. You can see minor differences due to the tuning parameters used to define the multiresolution analysis in the first paper. The convergence for the first error is better, which is what is expected. Recall that here ϵ was close to zero, which gives us the better convergence.

Next we consider the case where $\alpha = 1$.

The first error is either $O\left(2^i pn^{\epsilon-2} \log\left(n\left(2^i p\right)^{-1} e\right) + 2^i pqn^{\epsilon-1}\right)$ or $o\left(n^{-2r/(2r+1)}\right)$ where $\epsilon > 0$ and close to zero, whichever is larger.

The second error is either $o\left(2^{j_0} \frac{\log\left(n2^{-k}e\right)}{n}\right)$ or $o\left(n^{-2\beta/(1+2\beta)}\right)$, whichever is larger.

Again there are minor differences between these two bounds, partly because of the tuning parameters p and q. Again we see that the bound for the first error is better, which is what was expected given the circumstances. The first error has an extra factor of n^{-1} within the term with the logarithm.

Part 10. Writing long memory into a matrix context.

52. Introduction.

Here we consider data with long memory error viewed in terms of matrices. We assume that the data is dyadic and that we have incomplete data. Also, we later try to expand the work of Donoho and Johnstone to more readily incorporate long memory error.

53. Preliminaries and Notations.

Suppose that we have noisy data at irregular design points $\{t_1, t_2, ..., t_n\}$:

$$Y_i = f(t_i) + \epsilon_i, \qquad \epsilon_i \sim N\left(0, \sigma^2\right),$$

where f is an unknown regression to be estimated from the noisy sample. Assume f is defined on $[0, 1]$. Assume further that $t_i = n_i/2^J$ for some n_i and some fine resolution J that is determined by users. Usually $2^J \geq n$ so that the approximation errors by moving nondyadic points to dyadic points are negligible. Let f be the underlying regression function collected at all dyadic points $\left\{i/2^J, i = 0, ..., 2^{J-1}\right\}$. Let W be a given wavelet transform and $\theta = Wf$ be the wavelet transform of f. Because W is an orthogonal matrix, $f = W^T\theta$.

Denote the sampled data vector by Y_n. Let A be an $n \times N$ matrix whose ith row corresponds to the row of the matrix W^T for which the signal $f(t_i)$ is sampled with noise. Then the observed data can be expressed as a linear model

$$Y_n = A\theta + \epsilon \qquad \epsilon \sim N\left(0, \sigma^2 V\right) \tag{53.1}$$

where ϵ is the noise vector and V is the covariance matrix. The penalized least squares problem is to find θ to minimize

$$2^{-1}\left\|Y_n - A\theta\right\|^2 + \lambda\sum_{i=1}^{N} p\left(|\theta_i|\right) \tag{53.2}$$

for a given penalty function p and regularization parameter $\lambda > 0$.

When $n = 2^J$, the matrix A becomes a square orthogonal matrix W^T. This corresponds to the paper by Donoho and Johnstone [8]. We could then write (53.2) as

$$2^{-1}\left\|WY_n - \theta\right\|^2 + \lambda\sum_{i=1}^{N} p\left(|\theta_i|\right).$$

54. Solving the Problem with no noise.

Assume for this section that there is no noise, i. e. $\epsilon = 0$ in (53.1). We only have signal at the nonequispaced points $\{t_i, i = 1, ..., n\}$ which means we have no information at other dyadic points. Let

$$f_n = \left(f\left(t_1\right), ..., f\left(t_n\right)\right)^T$$

be the observed signals. Then

$$f_n = A\theta. \tag{54.1}$$

There are many solutions to this equation. We choose the one that provides the minimum Sobolev solution. We will use the double array sequence $\theta_{j,k}$ to denote the wavelet coefficient at the jth resolution level and the kth dyadic location $\left(k = 1, ..., 2^{j-1}\right)$. A Sobolev norm of f with degree of smoothness s can be expressed as

$$\left\|\theta\right\|_s^2 = \sum_j 2^{2sj}\left\|\theta_{j\cdot}\right\|^2,$$

where $\theta_{j\cdot}$ is the vector of the wavelet coefficients at the resolution level j.

We minimize $\left\|\theta\right\|_s^2$ subject to the constraint (54.1).

The solution (Rao 1973) is what is called the normalized method of frame whose solution is given by

$$\theta = DA^T\left(ADA^T\right)^{-1} f_n,$$

where $D = \text{Diag}\left(2^{-2sj_i}\right)$ with j_i denoting the resolution level with which θ_i is associated. When $s = 0$,$\theta = A^T f_n$ by orthogonality.

The traditional regularization problem can be formulated in the wavelet domain as follows. Find the minimum of

$$2^{-1}\left\|Y_n - A\theta\right\|^2 + \lambda\left\|\theta\right\|_s^2. \tag{54.2}$$

One can replace the Sobolev norm by other penalty functions, leading to minimizing

$$l(\theta) = 2^{-1}\left\|Y_n - A\theta\right\|^2 + \lambda\sum_{i\geq i_0} p(|\theta_i|). \tag{54.3}$$

for a given penalty function $p(\cdot)$ and given value i_0. This corresponds to penalizing wavelet coefficients above certain resolution level j_0.

When the sampling points are equally spaced and $n = 2^J$, the design matrix A in (54.1) becomes the inverse transform matrix W^T. In this case, (54.3) becomes

$$2^{-1}\sum_{i=1}^{n}(z_i - \theta_i)^2 + \lambda\sum_{i\geq i_0} p(|\theta_i|), \tag{54.4}$$

where z_i is the ith component of the wavelet coefficient vector $z = WY_n$. The solution to this problem is a component-wise minimization problem.

We will denote λp as p_λ. There are a variety of penalty functions to study. We give an outline of these function in Section 62.

55. Dealing with $p(\cdot)$.

Let $p(\cdot)$ be a nonnegative, nondecreasing and differentiable function on $(0,\infty)$. We wish to minimize with respect to θ

$$l(\theta) = (z-\theta)^2/2 + p_\lambda(|\theta|) \tag{55.1}$$

for a given penalty parameter λ. This is a component-wise minimization of (54.4). The function in (55.1) tend to infinity as $|\theta| \to \infty$. Thus, minimizers do exist. Let $\hat{\theta}(z)$ be a solution. The next theorem gives necessary and sufficient conditions for the solution to be thresholding, continuous, and approximately unbiased when $|z|$ is large.

Theorem 82. *Let $p_\lambda(\cdot)$ be a nonnegative, nondecreasing, and differentiable function in $(0,\infty)$. Further, assume that the function $-\theta - p'_\lambda(\theta)$ is strictly unimodal on $(0,\infty)$. Then we have the following results.*

(1) The solution to the minimization problem (55.1) exists and is unique. It is antisymmetric: $\hat{\theta}(-z) = -\hat{\theta}(z)$.

(2) The solution satisfies

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

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

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

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

(4) When $p'_\lambda(\theta)$ is continuous on $(0, \infty)$, the solution $\hat{\theta}(z)$ is continuous if and only if the minimum of $|\theta| + p'_\lambda(|\theta|)$ is attained at point zero.

(5) If $p'_\lambda(|z|) \to 0$, as $|z| \to +\infty$, then

$$\hat{\theta}(z) = z - p'_\lambda(|z|) + o\left(p'_\lambda(|z|)\right).$$

These results implicate that when $|z| \le p_0$, the estimate is thresholded to 0. For $|z| > p_0$, the solution has a shrinkage property. The amount of shrinkage is sandwiched between the soft-thresholding and hard-thresholding estimators, as is shown in result 3.

We now consider the risk functions. Assume that $E(Z) = \theta$ and $E\left(Z^2\right) \le 1$. Denote by

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

The thresholding parameter p_0 is equivalent to the regularization parameter λ.

In the work of Antoniadis and Fan we have the following theorem which will be extended later.

Theorem 83. *Suppose p satisfies conditions in Theorem 1 and $p'_\lambda(0+)$. Then*

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

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

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

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

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

These rules are improvements on the rules given by Donoho and Johnstone [8] which incorporate the new penalty functions.

From the work of Donoho and Johnstone and by Theorem 2, property 2, for any penalized least squares estimator, we have

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

if $p_0 \leq \sqrt{2\log n}$.

56. Cast of Characters.

Recall our estimator

$$\theta = DA^T\left(ADA^T\right)^{-1} f_n.$$

We have noisy data at irregular design points $\{t_1, ..., t_n\}$. We have

$$Y_i = f(t_i) + \epsilon_i, \qquad \epsilon_i \sim N\left(0, \sigma^2\right).$$

Here the ϵ_i's have the following property:

$$E(\epsilon_i \epsilon_{i+n}) \sim C_0 |n|^{-\alpha}.$$

Let A be the $n \times N$ matrix whose ith row corresponds to the row of the matrix W^T for which the signal $f(t_i)$ is sampled with noise. We express the observed data as the linear model

$$Y_n = A\theta + \epsilon \qquad \epsilon \sim N\left(0, \sigma^2 V\right).$$

where V is the covariance matrix. We wish to minimize

$$2^{-1}\left\|Y_n - A\theta\right\|^2 + \lambda \sum_{i=1}^{N} p(|\theta_i|)$$

for a given penalty function p and regularization parameter $\lambda > 0$. The function p has many properties which will be examined later.

Note that f_n is the vector of the signal values,

$$f_n = \{f(t_1), ..., f(t_n)\}^T.$$

Then f_n is an $n \times 1$ matrix. The matrix $D = \text{Diag}\left(2^{-2sj_i}\right)$ where j_i is the resolution level with which θ_i is associated. The parameter s is a smoothing parameter to be discussed later.

57. Dimensions of the matrix $DA^T\left(ADA^T\right)^{-1}$.

We now examine the dimensions of this matrix. A is an $n \times N$ matrix. A^T is an $N \times n$ matrix. D is an $N \times N$ matrix. So we have:

$$DA^T\left(ADA^T\right)^{-1}$$

$$N \times N \cdot N \times n\left(n \times N \cdot N \times N \cdot N \times n\right)^{-1}$$

$$N \times n \cdot \left(n \times N \cdot N \times n\right)^{-1}$$

$$N \times n \cdot \left(n \times n\right)^{-1}$$

$$N \times n \cdot n \times n$$

$$N \times n.$$

So, the result from these matrix multiplications is a $N \times n$ matrix. We multiply this times our vector, which is $n \times 1$ and get a result which is $N \times 1$, our θ. Note that all of these coefficients are bounded since they began as wavelet coefficients and the sums that the matrix multiplications produce are finite. We introduce a definition:

$$\bar{A} \equiv DA^T\left(ADA^T\right)^{-1} = \begin{bmatrix} \bar{a}_{11} & \bar{a}_{12} & ... & \bar{a}_{1n} \\ \bar{a}_{21} & \bar{a}_{22} & ... & \bar{a}_{2n} \\ ... & ... & ... & ... \\ \bar{a}_{N1} & \bar{a}_{N2} & ... & \bar{a}_{Nn} \end{bmatrix}.$$

This will allow us to examine the variance of of error vector ϵ more carefully.

58. Finding an Expression for the Variance

It is worth mentioning here that

$$\bar{A}\epsilon = \begin{bmatrix} \bar{a}_{11}\epsilon_1 + \bar{a}_{12}\epsilon_2 + ... + \bar{a}_{1n}\epsilon_n \\ \bar{a}_{21}\epsilon_1 + \bar{a}_{22}\epsilon_2 + ... + \bar{a}_{2n}\epsilon_n \\ ... \\ \bar{a}_{N1}\epsilon_1 + \bar{a}_{N2}\epsilon_2 + ... + \bar{a}_{Nn}\epsilon_n \end{bmatrix}.$$

We can see here clearly that $E\left(\bar{A}\epsilon\right) = 0$.

Recall that $\text{var}\left(A^T x\right) = A^T \text{var}\left(x\right) A$. Thus, $\text{var}\left(\bar{A}\epsilon\right) = \bar{A}\text{var}\left(\epsilon\right)\bar{A}^T$. Note that

$$\text{var}\left(\epsilon\right) = \begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & ... & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & ... & \epsilon_2\epsilon_n \\ ... & ... & ... & ... \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & ... & \epsilon_n^2 \end{bmatrix}.$$

This is an $n \times n$ matrix. We begin by computing $\bar{A}\text{var}(\epsilon)$.

$$\begin{bmatrix} \bar{a}_{11} & \bar{a}_{12} & ... & \bar{a}_{1n} \\ \bar{a}_{21} & \bar{a}_{22} & ... & \bar{a}_{2n} \\ ... & ... & ... & ... \\ \bar{a}_{N1} & \bar{a}_{N2} & ... & \bar{a}_{Nn} \end{bmatrix} \times \begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & ... & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & ... & \epsilon_2\epsilon_n \\ ... & ... & ... & ... \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & ... & \epsilon_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} \bar{a}_{11}\epsilon_1^2 + \bar{a}_{12}\epsilon_2\epsilon_1 + ... + \bar{a}_{1n}\epsilon_n\epsilon_1 & ... & \bar{a}_{11}\epsilon_1\epsilon_n + \bar{a}_{12}\epsilon_2\epsilon_n + ... + \bar{a}_{1n}\epsilon_n^2 \\ ... & ... & ... \\ \bar{a}_{N1}\epsilon_1^2 + \bar{a}_{N2}\epsilon_2\epsilon_1 + ... + \bar{a}_{Nn}\epsilon_n\epsilon_1 & ... & \bar{a}_{N1}\epsilon_1\epsilon_n + \bar{a}_{N2}\epsilon_2\epsilon_n + ... + \bar{a}_{Nn}\epsilon_n^2 \end{bmatrix}.$$

We write this in terms of summations below.

$$\begin{bmatrix} \sum_{i=1}^n \bar{a}_{1i}\epsilon_i\epsilon_1 & \sum_{i=1}^n \bar{a}_{1i}\epsilon_i\epsilon_2 & ... & \sum_{i=1}^n \bar{a}_{1i}\epsilon_i\epsilon_n \\ \sum_{i=1}^n \bar{a}_{2i}\epsilon_i\epsilon_1 & \sum_{i=1}^n \bar{a}_{2i}\epsilon_i\epsilon_2 & ... & \sum_{i=1}^n \bar{a}_{2i}\epsilon_i\epsilon_n \\ ... & ... & ... & ... \\ \sum_{i=1}^n \bar{a}_{Ni}\epsilon_i\epsilon_1 & \sum_{i=1}^n \bar{a}_{Ni}\epsilon_i\epsilon_2 & ... & \sum_{i=1}^n \bar{a}_{Ni}\epsilon_i\epsilon_n \end{bmatrix} = \bar{A}\text{var}(\epsilon).$$

We must simplify this a bit. Let $C_{jk} = \sum_{i=1}^n \bar{a}_{ji}\epsilon_i\epsilon_k$. Then we can further write

$$\bar{A}\text{var}(\epsilon) = \begin{bmatrix} C_{11} & C_{12} & ... & C_{1n} \\ C_{21} & C_{22} & ... & C_{2n} \\ ... & ... & ... & ... \\ C_{N1} & C_{N2} & ... & C_{Nn} \end{bmatrix}$$

We write explicitly

$$\bar{A}^T = \begin{bmatrix} \bar{a}_{11} & \bar{a}_{21} & ... & \bar{a}_{N1} \\ \bar{a}_{12} & \bar{a}_{22} & ... & \bar{a}_{N2} \\ ... & ... & ... & ... \\ \bar{a}_{1n} & \bar{a}_{2n} & ... & \bar{a}_{Nn} \end{bmatrix}.$$

We now find

$$\left(\bar{A}\text{var}(\epsilon)\right)\bar{A}^T = \begin{bmatrix} C_{11} & C_{12} & ... & C_{1n} \\ C_{21} & C_{22} & ... & C_{2n} \\ ... & ... & ... & ... \\ C_{N1} & C_{N2} & ... & C_{Nn} \end{bmatrix} \times \begin{bmatrix} \bar{a}_{11} & \bar{a}_{21} & ... & \bar{a}_{N1} \\ \bar{a}_{12} & \bar{a}_{22} & ... & \bar{a}_{N2} \\ ... & ... & ... & ... \\ \bar{a}_{1n} & \bar{a}_{2n} & ... & \bar{a}_{Nn} \end{bmatrix}.$$

$$= \begin{bmatrix} \bar{a}_{11}C_{11} + \bar{a}_{12}C_{12} + ... + \bar{a}_{1n}C_{1n} & ... & \bar{a}_{N1}C_{11} + \bar{a}_{N2}C_{12} + ... + \bar{a}_{Nn}C_{1n} \\ ... & ... & ... \\ \bar{a}_{11}C_{N1} + \bar{a}_{12}C_{N2} + ... + \bar{a}_{1n}C_{Nn} & ... & \bar{a}_{N1}C_{N1} + \bar{a}_{N2}C_{N2} + ... + \bar{a}_{Nn}C_{Nn} \end{bmatrix}.$$

We can write this as a summation.

$$= \begin{bmatrix} \sum_{k=1}^n \bar{a}_{1k}C_{1k} & \sum_{k=1}^n \bar{a}_{2k}C_{1k} & ... & \sum_{k=1}^n \bar{a}_{Nk}C_{1k} \\ \sum_{k=1}^n \bar{a}_{1k}C_{2k} & \sum_{k=1}^n \bar{a}_{2k}C_{2k} & ... & \sum_{k=1}^n \bar{a}_{Nk}C_{2k} \\ ... & ... & ... & ... \\ \sum_{k=1}^n \bar{a}_{1k}C_{Nk} & \sum_{k=1}^n \bar{a}_{2k}C_{Nk} & ... & \sum_{k=1}^n \bar{a}_{Nk}C_{Nk} \end{bmatrix}.$$

This is an $N \times N$ matrix. We will write

$$D_{lm} = \sum_{k=1}^{n} \bar{a}_{mk} C_{lk}.$$

Then our result can be rewritten as

$$= \begin{bmatrix} D_{11} & D_{12} & ... & D_{1N} \\ D_{21} & D_{22} & ... & D_{2N} \\ ... & ... & ... & ... \\ D_{N1} & D_{N2} & ... & D_{NN} \end{bmatrix}.$$

These D_{lm}'s are what we must bound. We begin by writing down exactly what D_{lm} stands for in terms of sums.

$$D_{lm} = \sum_{k=1}^{n} \bar{a}_{mk} C_{lk} = \sum_{k=1}^{n} \bar{a}_{mk} \sum_{i=1}^{n} \bar{a}_{li} \epsilon_i \epsilon_k = \sum_{k=1}^{n} \sum_{i=1}^{n} \bar{a}_{mk} \bar{a}_{li} \epsilon_i \epsilon_k = \sum_{k=1}^{n} \bar{a}_{mk} \bar{a}_{lk} \epsilon_k^2 + \sum_{k=1}^{n} \sum_{i \neq k}^{n} \bar{a}_{mk} \bar{a}_{li} \epsilon_i \epsilon_k.$$

We have bounded this kind of quantity many times in the course of our studies. The key difference between this and the other bounds that we have found are the $\bar{a}_{jk}$'s. If we can show that the $\bar{a}_{jk}$'s are the same "size" as regular wavelet coefficients, then we can apply the older results.

59. What size is $\bar{a}_{jk}$?

Recall that

$$\bar{A} \equiv DA^T \left(ADA^T\right)^{-1}.$$

Let

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1N} \\ a_{21} & a_{22} & ... & a_{2N} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nN} \end{bmatrix}.$$

Then

$$A^T = \begin{bmatrix} a_{11} & a_{21} & ... & a_{n1} \\ a_{12} & a_{22} & ... & a_{n2} \\ ... & ... & ... & ... \\ a_{1N} & a_{2N} & ... & a_{nN} \end{bmatrix}.$$

Recall $D = \text{Diag}\left(2^{-2sj_i}\right)$ and D is an $N \times N$ matrix.

$$D = \begin{bmatrix} 2^{-2sj_1} & 0 & 0 & 0 \\ 0 & 2^{-2sj_2} & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 2^{-2sj_N} \end{bmatrix}.$$

We calculate DA^T

$$= \begin{bmatrix} 2^{-2sj_1}a_{11} & 2^{-2sj_1}a_{21} & ... & 2^{-2sj_1}a_{n1} \\ 2^{-2sj_2}a_{12} & 2^{-2sj_2}a_{22} & ... & 2^{-2sj_2}a_{n2} \\ ... & ... & ... & ... \\ 2^{-2sj_N}a_{1N} & 2^{-2sj_N}a_{2N} & ... & 2^{-2sj_N}a_{nN} \end{bmatrix}.$$

Now we compute ADA^T

$$= \begin{bmatrix} a_{11} & a_{12} & ... & a_{1N} \\ a_{21} & a_{22} & ... & a_{2N} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nN} \end{bmatrix} \times \begin{bmatrix} 2^{-2sj_1}a_{11} & 2^{-2sj_1}a_{21} & ... & 2^{-2sj_1}a_{n1} \\ 2^{-2sj_2}a_{12} & 2^{-2sj_2}a_{22} & ... & 2^{-2sj_2}a_{n2} \\ ... & ... & ... & ... \\ 2^{-2sj_N}a_{1N} & 2^{-2sj_N}a_{2N} & ... & 2^{-2sj_N}a_{nN} \end{bmatrix}$$

$$= \begin{bmatrix} 2^{-2sj_1}a_{11}^2 + ... + 2^{-2sj_n}a_{1N}^2 & ... & 2^{-2sj_1}a_{11}a_{n1} + ... + 2^{-2sj_n}a_{1N}a_{nN} \\ ... & ... & ... \\ 2^{-2sj_1}a_{n1}a_{11} + ... + 2^{-2sj_n}a_{nN}a_{1N} & ... & 2^{-2sj_1}a_{n1}^2 + ... + 2^{-2sj_n}a_{nN}^2 \end{bmatrix}.$$

So, finally, we must examine $DA^T\left(ADA^T\right)^{-1}$.

$$= \begin{bmatrix} 2^{-2sj_1}a_{11} & 2^{-2sj_1}a_{21} & ... & 2^{-2sj_1}a_{n1} \\ 2^{-2sj_2}a_{12} & 2^{-2sj_2}a_{22} & ... & 2^{-2sj_2}a_{n2} \\ ... & ... & ... & ... \\ 2^{-2sj_N}a_{1N} & 2^{-2sj_N}a_{2N} & ... & 2^{-2sj_N}a_{nN} \end{bmatrix} \times$$

$$\left(\begin{bmatrix} 2^{-2sj_1}a_{11}^2 + ... + 2^{-2sj_n}a_{1N}^2 & ... & 2^{-2sj_1}a_{11}a_{n1} + ... + 2^{-2sj_n}a_{1N}a_{nN} \\ ... & ... & ... \\ 2^{-2sj_1}a_{n1}a_{11} + ... + 2^{-2sj_n}a_{nN}a_{1N} & ... & 2^{-2sj_1}a_{n1}^2 + ... + 2^{-2sj_n}a_{nN}^2 \end{bmatrix} \right)^{-1}.$$

We define $E_{ij} = \sum_{k=1}^{N} 2^{-2sj_k}a_{ik}a_{jk}$. Then

$$ADA^T = \begin{bmatrix} E_{11} & E_{12} & ... & E_{1n} \\ E_{21} & E_{22} & ... & E_{2n} \\ ... & ... & ... & ... \\ E_{n1} & E_{n2} & ... & E_{nn} \end{bmatrix}.$$

Recall That the determinant of an $n \times n$ matrix is defined as

$$\det(G) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} g_{i,\sigma_i}.$$

There are $n!$ permutations σ and n g_{i,σ_i}'s in each of the $n!$ terms in this sum.

Let

$$G^{-1} = det(G)\, H,$$

so

$$H = \frac{1}{\det(G)} G^{-1}.$$

For each entry of H, there will be a sum with $(n-1)!$ terms, and each of the terms in the sum will be a product of $n-1$ entries of the original matrix. Specifically, for an entry of H,

$$h_{ij} = (-1)^{i+j} G_{ij}$$

where G_{ij} is the minor of matrix G, which is defined as the determinant of the $(n-1) \times (n-1)$ matrix which results in deleting the ith row and jth column of G. This can be verified by examining the formulas for inverse matrices.

Now we have a very specific idea of what the entries of this matrix look like.

We know that $\det(W)$ is not zero, because it is a matrix of wavelet coefficients. Consequently, we know that the rows and columns of A are linearly independent, since A came from W. Therefore, we know that $\left(ADA^T\right)$is independent, since the a_{ij}'s are independent, no row of column of this matrix can be made from the others. Therefore, we know that an inverse does exist, and we can express it with the formulas above. We will denote

$$\left(ADA^T\right)^{-1} = \begin{bmatrix} F_{11} & F_{12} & ... & F_{1n} \\ F_{21} & F_{22} & ... & F_{2n} \\ ... & ... & ... & ... \\ F_{n1} & F_{n2} & ... & F_{nn} \end{bmatrix}.$$

Then we can write

$$DA^T\left(ADA^T\right)^{-1} = \begin{bmatrix} 2^{-2sj_1}a_{11} & 2^{-2sj_1}a_{21} & ... & 2^{-2sj_1}a_{n1} \\ 2^{-2sj_2}a_{12} & 2^{-2sj_2}a_{22} & ... & 2^{-2sj_2}a_{n2} \\ ... & ... & ... & ... \\ 2^{-2sj_N}a_{1N} & 2^{-2sj_N}a_{2N} & ... & 2^{-2sj_N}a_{nN} \end{bmatrix} \times \begin{bmatrix} F_{11} & F_{12} & ... & F_{1n} \\ F_{21} & F_{22} & ... & F_{2n} \\ ... & ... & ... & ... \\ F_{n1} & F_{n2} & ... & F_{nn} \end{bmatrix}.$$

The ijth entry of this matrix is $\sum_{k=1}^{n} 2^{-2sj_i} a_{ik} F_{kj}$. This is the quantity we must bound. We hope that these are the same "size" as the wavelet coefficients a_{jk}.

60. Bounding $\sum_{k=1}^{n} 2^{-2sj_i} a_{ik} F_{kj}$.

This sum is equal to

$$\sum_{k=1}^{n} 2^{-2sj_i} a_{ik} \left[\frac{(-1)^{i+j} \det\left(kj\text{th minor of } ADA^T\right)}{\det\left(ADA^T\right)} \right]$$

$$\sum_{k=1}^{n} 2^{-2sj_i} a_{ik} \left[\frac{n-1 \text{ terms with } n-1 \text{ items in each term.}}{n \text{ terms with } n \text{ items in each term.}} \right]$$

The items in each of these terms specifically are the $E_{ij} = \sum_{k=1}^{N} 2^{-2sj_k} a_{ik} a_{jk}$. These are a sum of N terms which are $o\left(a_{jk} a_{j_0 k_0}\right)$ with a constant in the front which is some negative power of 2.

$$\sum_{k=1}^{n} 2^{-2sj_i} a_{ik} \left[\frac{(n-1)! \text{ terms with } n-1 \text{ items with N terms which are the size of } E_{ij}}{n! \text{ terms with } n \text{ items with N terms which are the size of } E_{ij}} \right]$$

Note that

$$E_{ij} \leq N \max \left\{ 2^{-2sj_k} a_{ik} a_{jk} \right\} = b$$

Then our sum is

$$\leq \sum_{k=1}^{n} 2^{-2sj_i} a_{ik} \left[\frac{(n-1)! \text{ terms of } b^{n-1}}{n! \text{ terms with } n \text{ items with N terms which are the size of } E_{ij}} \right]$$

Now we must use the fact that the a_{jk} are wavelet coefficients. Many of these coefficients will be equal to zero. Let

$$C_1 = \text{ the number of wavelet coefficients which are nonzero}$$

and

$$C_2 = \text{the number of } E_{ij} \text{ which are nonzero.}$$

Also, note that the determinant has terms which are negative and terms which are positive. Further, let

$$a = \min_{\text{all } E_{ij}} \left\{ E_{ij} - E_{i_1 j_1} \right\}.$$

Then our sum is

$$\leq \sum_{k=1}^{n} 2^{-2sj_i} a_{ik} \left[\frac{(n-1)! \text{ terms of } b^{n-1}}{(n - C_2)\,(n-1)! \text{ terms of } a^n} \right]$$

$$\leq (C_1) \max_{k} \left\{ 2^{-2sj_i} a_{ik} \right\} \left[\frac{(n-1)! \text{ terms of } b^{n-1}}{(C_2)\,(n-1)! \text{ terms of } a^n} \right]$$

We can see that this is about the same "size" as our wavelet coefficients a_{jk}. Thus, we can apply the boundedness results from earlier work. We can see that these $\bar{a}_{jk}$ are linear combinations of the original a_{jk}, and thus inherit wavelet properties from those coefficients.

61. What does this mean for the work in [1]?

What we have here is a general beginning for bounding the wavelet coefficients found by this method. As we have seen in earlier sections, the rest of the boundedness of the space depends on the space which the function is associated with. For instance, if the function is within the Holder class, then the bounds for each coefficient are the same as those in Theorem 73.

If the function is bounded as in Part 7, then the bounds of the error are the same as those in Theorem 67. The same could be easily extended to other spaces, as the authors Antoniadis and Fan did with the Besov space in [1].

What we have found are the errors associated with the coefficients which were determined via

$$\theta = DA^T \left(ADA^T\right)^{-1} f_n.$$

This does not represent a full set of wavelet coefficients. The authors suggest the following process to remedy this situation.

The authors take advantage of the orthonormal of W. Recall that an orthonormal matrix has the property that

$$A^T = A^{-1}.$$

Also note that if W is orthonormal, so is A.

$$WW^T = I_N.$$

A term of WW^T has the form

$$\sum_{i=1}^{N} w_{ji} w_{ik} = \delta\left(j-k\right).$$

Recall that A is an $n \times N$ matrix which is made by taking rows of W and collecting them together as follows.

$$A = \begin{bmatrix} w_{i_1 1} & w_{i_1 2} & ... & w_{i_1 N} \\ w_{i_2 1} & w_{i_2 2} & ... & w_{i_2 N} \\ ... & ... & ... & ... \\ w_{i_n 1} & w_{i_n 2} & ... & w_{i_n N} \end{bmatrix}$$

A term of AA^T has the form

$$\sum_{h=1}^{N} w_{i_j h} w_{h i_k} = \delta\left(j-k\right).$$

Thus, A is orthonormal as well.

We collect the rows of W^T not put into the matrix A into a matrix B of size $(N-n)\times N$. Then the penalized least squares in (54.3) can be written as

$$l(\theta) = 2^{-1}\left\|Y^* - W^T\theta\right\|^2 + \sum_{i\geq i_0} p_\lambda(|\theta_i|),$$

where $Y^* = \left(Y_n^T, (B\theta)^T\right)^T$. By the orthonormaility of the wavelet transform,

$$l(\theta) = 2^{-1}\left\|WY^* - \theta\right\|^2 + \sum_{i\geq i_0} p_\lambda(|\theta_i|). \tag{61.1}$$

We can optimize this iteratively, though the authors Fan and Chen in (1999) point out that there is no guarantee of convergence, but if the initial estimators of θ are "reasonably good", then the one step method is as efficient as the fully iterative method. The key improvement here is in adding long memory error to the matrix setting which already incorporates penalty functions.

We must discuss good initial estimators of θ. The authors suggest using Sobolev wavelet interpolators to produce an initial estimate for θ and hence for Y^*. Recall that $\hat{\theta}^* = DA^T\left(ADA^T\right)Y_n$ was obtained via wavelet interpolation. Let

$$\hat{Y}_0^* = \left(Y_n^T, \left(B\hat{\theta}\right)^T\right)^T$$

be the initial synthetic data. Then

$$\hat{\theta}^* = W\hat{Y}_0^* \sim N\left(\theta^*, \epsilon^2 V\right)$$

where $V = DA^T\left(ADA^T\right)^{-2}AD$, as analyzed before and $\theta^* = DA^T\left(ADA^T\right)^{-1}A\theta$ is the vector of wavelet coefficients.

62. Examining the Penalty Functions.

We now examine some of the penalty functions which satisfy the requirements set forth in Theorem 84. We use p_λ to denote the penalty function λp in the following discussion.

For the L_1-penalty

$$p_\lambda(|\theta|) = \lambda|\theta|, \tag{62.1}$$

the solution is the soft-thresholding rule (Donoho 1992).

A clipped L_1-penalty as below

$$p(\theta) = \lambda\min(|\theta|, \lambda) \tag{62.2}$$

leads to a mixture of soft and hard thresholding rules (Fan 1997):

$$\hat{\theta}_j = (|z_j| - \lambda)_+ I\{|z_j| \leq 1.5\lambda\} + |z_j| I\{|z_j| > 1.5\lambda\}. \tag{62.3}$$

When the penalty function is given by

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda), \tag{62.4}$$

the solution is the hard-thresholding rule (Antoniadis 1997). This penalty function is smoother than $p_\lambda(|\theta|) = |\theta| I(|\theta| < \lambda) + \lambda/2 I(|\theta| \geq \lambda)$ suggested by Fan (1997) and the entropy penalty $p_\lambda(|\theta|) = 2^{-1}\lambda^2 I\{|\theta| \neq 0\}$, which lead to the same solution.

Recall that the hard thresholding rule is discontinuous which is not always desirable. The soft thresholding rule is continuous but it shifts the estimator by an amount of λeven when $|z_i|$ is far from the noise level, which creates unnecessary bias when θ is large.

To improve on these two drawbacks, Fan (1997) suggests using the quadratic spline penalty, called the smoothly clipped absolute deviate (SCAD) penalty

$$p'_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \tag{62.5}$$

for $\theta > 0$ and $a > 2$ leading to the piecewise linear thresholding

$$\hat{\theta}_j = \begin{cases} \operatorname{sgn}(z_j)(|z_j| - \lambda) & \text{when } |z_j| \leq 2\lambda, \\ +\frac{(a-1)z_j - a\lambda \operatorname{sgn}(z_j)}{a-2} & \text{when } 2\lambda < |z_j| < a\lambda, \\ z_j & \text{when } |z_j| > a\lambda. \end{cases} \tag{62.6}$$

This penalty function does not over-penalize large values of $|\theta|$and hence does not create large biases when the wavelet coefficients are large.

Nikolova suggests the following transformed L_1=penalty function

$$p_\lambda(|x|) = \lambda b |x| (1 + b|x|)^{-1} \qquad \text{for some } b > 0.$$

This function behaves similarly to SCAD. Other possible penalty functions include the L_p-penalty introduce by Bouman and Sauer:

$$p_\lambda(|\theta|) = \lambda |\theta|^p \qquad (p \geq 0) \tag{62.7}$$

Note that the choice $p \leq 1$ is necessary for the solution to be a thresholding operator, whereas $p \geq 1$ is a necessary condition for the solution to be continuous in z. Thus, the L_1-penalty function is the only member in this family that yields a continuous thresholding solution.

Lastly, we note that the regularization parameter λ for different penalty functions has a different scale.

63. Results.

Below is the theorem from [1].

Theorem 84. *Let $p_\lambda(\cdot)$ be a nonnegative, nondecreasing, and differentiable function in $(0,\infty)$. Further, assume that the function $-\theta - p'_\lambda(\theta)$ is strictly unimodal on $(0,\infty)$. Then we have the following results.*

(1) The solution to the minimization problem (55.1) exists and is unique. It is antisymmetric: $\hat{\theta}(-z) = -\hat{\theta}(z)$.

(2) The solution satisfies

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

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

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

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

(4) When $p'_\lambda(\theta)$ is continuous on $(0,\infty)$, the solution $\hat{\theta}(z)$ is continuous if and only if the minimum of $|\theta| + p'_\lambda(|\theta|)$ is attained at point zero.

(5) If $p'_\lambda(|z|) \to 0$, as $|z| \to +\infty$, then

$$\hat{\theta}(z) = z - p'_\lambda(|z|) + o\left(p'_\lambda(|z|)\right).$$

Proof. Recall that

$$l(\theta) = (z-\theta)^2/2 + p_\lambda(|\theta|)$$

for a given penalty parameter λ. We note that this function tends to infinity as $|\theta| \to \infty$, thus, minimizers do exist.

Also note that

$$p_0 = \min_{\theta\ge 0}\{\theta + p'_\lambda(\theta)\}.$$

When $z = 0$, it is clear that $\hat{\theta}(z) = 0$ is the unique minimizer. Without loss of generality we may assume that $z > 0$. Then, for all $\theta > 0$, $l(-\theta) > l(\theta)$. Hence, $\hat{\theta}(z) \ge 0$. Note that for $\theta > 0$,

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

When $z < p_0$, the function l is strictly increasing on $(0,\infty)$ because the derivative function is positive. Therefore, $\hat{\theta}(z) = 0$. When the function $l'(\theta)$ is strictly increasing, there is at most one zero-crossing, and hence the solution is unique.

Therefore, we only need to consider the case that $l'(\theta)$ has a valley on $(0,\infty)$ and $z > p_0$. In this case, there are two possible zero-crossings for the function l' on $(0,\infty)$. The larger one is the minimizer because the derivative function at that point is increasing. Hence, the solution is unique and satisfies

$$\hat{\theta}(z) = z - p'_\lambda\left(\hat{\theta}(z)\right) \le z. \tag{63.1}$$

Thus, $\hat{\theta}(z) \le z - p'_\lambda(z)$ when $p'_\lambda(\cdot)$ is nonincreasing. Let θ_0 be the minimizer of $\theta + p'_\lambda(\theta)$ over $[0,\infty)$. Then, from the preceding argument, $\hat{\theta}(z) > \theta_0$ for $z > p_0$. If $p'_\lambda(\cdot)$ is nonincreasing, then

$$p'_\lambda(\theta(z)) \le p'_\lambda(\theta_0) \le \theta_0 + p'_\lambda(\theta_0) = p_0.$$

This and (63.1) prove result 3. It is clear that the continuity of the solution $\hat{\theta}(z)$ at the point $z = p_0$ if and only if the minimum of the function $|\theta| + p'_\lambda(|\theta|)$ is attained at 0. The continuity at other locations follows directly from the monotonicity and continuity of the function $\theta + p'_\lambda(\theta)$ in the interval (θ,∞). The last conclusion follows directly from (63.1). □

Theorem 85. *Suppose p satisfies conditions in Theorem 84 and $p'_\lambda(0+)$. Then*

(1) $R_p(\theta, p_0) \le c_0^2 + \theta^2$.

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

$$R_p(\theta, p_0) \le p_0^2 + \sqrt{2/\pi} p_0 c_0 + c_0^2.$$

(3) $R_p(0, p_0) \le c_0^3 \sqrt{2/\pi}\left(\frac{p_0}{c_0} + \frac{c_0}{p_0}\right)\exp\left(-\frac{p_0^2}{2c_0^2}\right)$.

(4) $R_p(\theta, p_0) \le R_p(0, p_0) + \left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2$.

Proof. First, recall that

$$R_p(\theta, p_0) = E\left\{\hat{\theta}(z) - \theta\right\}^2$$

where z is normally distributed, $E(z) = \theta$, and $\mathrm{Var}(z^2) = c_0^2$ for some c_0.

Note that $R_p(\theta, p_0)$is symmetric about 0 by Theorem 84 Result 1. Thus, we can assume without loss of generality that $\theta \ge 0$. By Theorem 84, results 1 and 2,

$$E\left(\hat{\theta} - \theta\right)^2 \le E(z-\theta)^2 I\left(\hat{\theta} \notin [0,\theta]\right) + \theta^2 P\left(\hat{\theta} \in [0,\theta]\right) \le c_0^2 + \theta^2. \tag{63.2}$$

To prove result 2, we note that

$$E\left(\hat{\theta} - \theta\right)^2 = E\left(\left(\hat{\theta} - z\right) - (\theta - z)\right)^2 = E\left(\hat{\theta} - z\right)^2 - 2E\left(\left(\hat{\theta} - z\right)(\theta - z)\right) + E(\theta - z)^2$$

$$= E(\theta - z)^2 + 2E\left((z-\theta)\left(\hat{\theta} - z\right)\right) + E\left(\hat{\theta} - z\right)^2 = c_0^2 + 2E\left((z-\theta)\left(\hat{\theta} - z\right)\right) + E\left(\hat{\theta} - z\right)^2.$$

For $z > \theta$, we have $\hat{\theta} \leq z$ by Theorem 84, result 3, which implies that $(z-\theta)\left(\hat{\theta} - z\right) \leq 0$. Similarly, for $z < 0$, $(z-\theta)\left(\hat{\theta} - z\right) \leq 0$. Thus

$$c_0 + 2E\left((z-\theta)\left(\hat{\theta} - z\right)\right) + E\left(\hat{\theta} - z\right)^2 \leq c_0 + 2E\left((z-\theta)\left(\hat{\theta} - z\right)\right) I(0 \leq z \leq \theta) + E\left(\hat{\theta} - z\right)^2.$$

Here we have excluded the parts which are less than zero. By Theorem 84, result 3, $\left|\hat{\theta} - z\right| \leq p_0$. Thus,

$$E\left(\hat{\theta} - \theta\right)^2 \leq c_0^2 + 2p_0 E(\theta - z) I(z \leq \theta) + p_0^2 \leq c_0^2 + p_0 c_0 \sqrt{2/\pi} + p_0^2.$$

This factor of $\sqrt{2/\pi}$ comes from the centralized exact moment of the normal distribution, specifically if $X \sim N(\mu, \sigma)$,

$$E\left(|X-\mu|^p\right) = \sigma^p (p-1)!! \begin{cases} \sqrt{2/\pi} & \text{if } p \text{ is odd,} \\ 1 & \text{if } p \text{ is even.} \end{cases} \tag{63.3}$$

This establishes result 2.

Result 3 follows directly from the fact that

$$R_p(0, p_0) \leq E\left(z^2\right) I\{|z| \geq p_0\} \leq c_o^2 I\{|z| \geq p_0\}.$$

Note that for a standard normal variable $Z \sim N(0,1)$ (as in the work of fan),

$$P(|Z| \geq p_0) \leq \left(p_0 + p_0^{-1}\right) \exp\left\{-p_0^2/2\right\}.$$

Than for our variable z which has a variance of c_0 and for this property a mean of $\theta = 0$,

$$P(|z| \geq p_0) = P\left(\left|\frac{z}{c_0}\right| \geq \frac{p_0}{c_0}\right) = P\left(|Z| \geq \frac{p_0}{c_0}\right) \leq \left(\frac{p_0}{c_0} + \frac{c_0}{p_0}\right) \exp\left(-\frac{p_0^2}{2c_0^2}\right).$$

Further noting that

$$I(|z| \geq p_0) = E(|z|) P(|z| \geq p_0)$$

and applying (63.3) gives us result 3.

To show result 4, using the fact that $R_p'(0, p_0) = 0$ due to symmetry, we have by the Taylor expansion that

$$R_p(\theta, p_0) \leq R_p(0, p_o) + \frac{1}{2} \sup_{0<\eta<l} R_p''(\eta, p_0) \theta^2 \tag{63.4}$$

for $\theta \in [-1, 1]$.

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

$$R'_p(\theta, p_0) = \int_{-\infty}^{\infty} \left(\theta + z - 2\hat{\theta}\right) \phi(z-\theta)\, dz$$

$$= 2\theta - 2\int_{-\infty}^{\infty} \hat{\theta}\phi(z-\theta)\, dz$$

and $R''_p(\theta, p_0) = 2 + 2E\hat{\theta}(\theta - z)$. By using the same arguments as those in the proof of result 2, we have for $\theta > 0$

$$R''_p(\theta, p_0) \le 2 + 2E\hat{\theta}(\theta - z) I(0 \le z \le \theta).$$

Noting that $\hat{\theta} = 0$ for $|z| \le p_0$, we have for $p_0 \ge 1$ $R''_p(\theta, p_0) \le 2$. For the general case, using the fact that $\left|\hat{\theta}\right| \le |z|$, we have for $\theta \in [0,1]$

$$R''_p(\theta, p_0) \le 2 + 2\theta E(\theta - z) I(0 \le z \le \theta) = 2 + c_0\sqrt{2/\pi}\theta P(0 \le z \le \theta) \le 2 + c_0\sqrt{2/\pi}.$$

By (63.4), result 4 follows for $\theta \in [-1, 1]$. For θ outside this interval, 4 follows from (63.2). This completes the proof. □

We will need the following Lemma.

Lemma 86. *If the penalty function satisfies conditions of Theorem 84 and $p'_\lambda(\cdot)$ is nonincreasing and $p'_\lambda(0+) > 0$, then*

$$R_p(\theta, p_0) \le \left(2c_0^2 \log n + \left(c_0^2\sqrt{4/\pi} + c_o^2\right)\log^{1/2} n\right)\left\{c/n + \min\left(\frac{1}{2}\left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2, c_0^2\right)\right\}$$

or if c_0 is reasonably small, ie $c_0\sqrt{2/\pi}/2 \le 1$,

$$R_p(\theta, p_0) \le c_0^2\left(2\log n + \left(\sqrt{4/\pi} + 1\right)\log^{1/2} n\right)\left\{c/n + \min\left(\theta^2, c_0^2\right)\right\}$$

for the universal thresholding

$$p_0 = c_0\sqrt{2\log n - \log(1 + d\log n)}, \qquad 0 \le d \le c^2,$$

with $n \ge 4$ and $c \ge 1$ and $p_0 > 1.14$.

Proof. Note that by Theorem 85, property 2 we have

$$R_p(\theta, p_0) \le p_0^2 + \sqrt{2/\pi}p_0c_0 + c_0^2.$$

$$R_p(\theta, p_0) \le 2c_0^2\log n + c_0^2\sqrt{4/\pi}(\log n)^{1/2} + c_0^2 = c_0^2\left(2\log n + \sqrt{4/\pi}(\log n)^{1/2} + 1\right) \tag{63.5}$$

if $p_0 \leq c_0\sqrt{2\log n}$ where $c_0^2 = \text{Var}\left(z^2\right)$. For $|\theta| > c_0$, by (63.5) we have for $n \geq e$

$$R_p\left(\theta, p_0\right) \leq 2c_0^2 \log n + c_0^2\sqrt{4/\pi}\left(\log n\right)^{1/2} + c_0^2\left(\log n\right)^{1/2}$$

$$\leq 2c_0^2 \log n + \left(c_0^2\sqrt{4/\pi} + c_o^2\right)\left(\log n\right)^{1/2}.$$

Note that this is because $\left(\log n\right)^{1/2}$ is an increasing function if $n > 1$, as we can see from its derivative $1/\left[2n\left(\log n\right)^{1/2}\right] \geq 0$ for $n \geq e$.

Thus, we need to show that the inequality holds for $\theta \in [0, c_0]$. First, by Theorem 85, result 4,

$$R_p\left(\theta, p_0\right) \leq R_p\left(0, p_0\right) + \left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2.$$

Let $g\left(\theta\right) = \left(R_p\left(0, p_0\right) + \left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2\right) / \left(c/n + \frac{1}{2}\left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2\right)$. If $R_p\left(0, p_0\right) \leq 2c/n$, then $g\left(\theta\right) \leq 2 \leq 2\log n$. Hence, the result holds.

When $R_p\left(0, p_0\right) > 2c/n$, $g\left(\theta\right)$ is monotonically decreasing, as we can see from the derivative. Let $e = \left(1 + c_0\sqrt{2/\pi}/2\right)$. Then

$$g\left(\theta\right) = \frac{R_p\left(0, p_0\right) + e\theta^2}{c/n + \frac{1}{2}e\theta^2}.$$

$$g'\left(\theta\right) = \frac{2e\theta\left(c/n + \frac{1}{2}e\theta^2\right) - e\theta\left(R_p\left(0, p_0\right) + e\theta^2\right)}{\left(c/n + \frac{1}{2}e\theta^2\right)^2} = \frac{e\theta\left[2c/n + e\theta^2 - R_p\left(0, p_0\right) - e\theta^2\right]}{\left(c/n + \frac{1}{2}e\theta^2\right)^2}$$

$$= \frac{e\theta\left[2c/n - R_p\left(0, p_0\right)\right]}{\left(c/n + \frac{1}{2}e\theta^2\right)^2} \leq 0$$

If $R_p\left(0, p_0\right) > 2c/n$. Hence, $g\left(\theta\right) \leq g\left(0\right) = c^{-1}nR_p\left(0, p_0\right)$. By Theorem 85, result 3,

$$R_p\left(0, p_0\right) \leq c_0^3\sqrt{2/\pi}\left(\frac{p_0}{c_0} + \frac{c_0}{p_0}\right)\exp\left(-\frac{p_0^2}{2c_0^2}\right)$$

we have

$$g\left(\theta\right) \leq c^{-1}nc_0^3\sqrt{2/\pi}\left(\frac{p_0}{c_0}\right)\left(1 + \left(\frac{p_0}{c_0}\right)^{-2}\right)\exp\left(-\frac{p_0^2}{2c_0^2}\right)$$

$$\leq 2\pi^{-1/2}c^{-1}c_0^3\left(1 + \left(\frac{p_0}{c_0}\right)^{-2}\right)\left(\log n\right)^{1/2}\left(1 + d^{1/2}\left(\log n\right)^{1/2}\right). \tag{63.6}$$

This inequality a result of the following inequality in the work of [1].

$$nc^{-1}p_0\left(1+p_0^{-2}\right)\sqrt{2/\pi}\exp\left(-p_0^2/2\right)\leq 2\pi^{-1/2}c^{-1}\left(1+p_0^{-2}\right)(\log n)^{1/2}\left(1+d^{1/2}(\log n)^{1/2}\right). \tag{63.7}$$

By using the fact that for $\frac{p_0}{c_0}>1.14$, $\pi^{-1/2}\left(1+\left(\frac{p_0}{c_0}\right)^{-2}\right)\leq 1$, we conclude that $g(\theta)\leq 2c^{-1}c_0^3d^{1/2}(\log n)+2c^{-1}c_0^3(\log n)^{1/2}$. Thus, since $0<d\leq c^2$,

$$R_p(\theta,p_0)\leq R_p(0,p_0)+\left(1+c_0\sqrt{2/\pi}/2\right)\theta^2\leq 2c^{-1}c_0^3\left(c\log n+(\log n)^{1/2}\right)\left(\frac{c}{n}+\frac{\left(1+c_0\sqrt{2/\pi}/2\right)}{2}\theta^2\right).$$

□

Finally we have this last theorem. To bound the risk of the nonlinear estimator $\hat{\theta}(z)$ by that of the oracle estimator $\hat{\theta}_0$, we need to add an amount cn^{-1} for some constant c to the risk of the oracle estimator, because it has no risk at point $\theta=0$. We introduce the quantity

$$\Lambda_{n,c,p_0}(p)=\sup_{\theta}\frac{R_p(\theta,p_0)}{cn^{-1}+\min\left(\theta^2,c_0^2\right)}$$

and denote $\Lambda_{n,c,p_0}(p)$ by $\Lambda_{n,c}(p)$ for the universal thresholding $p_0=c_0\sqrt{2\log n}$. Then, $\Lambda_{n,c,p_0}(p)$ is a sharp risk upper bound for using the universal thresholding parameter p_0. That is,

$$R_p(\theta,p_0)\leq\Lambda_{n,c,p_0}(p)\left\{cn^{-1}+\min\left(\theta^2,c_0^2\right)\right\}. \tag{63.8}$$

Thus, the penalized least squares estimator $\hat{\theta}(z)$ performs comparably with the oracle estimator within a factor of $\Lambda_{n,c,p_0}(p)$. Likewise, let

$$\Lambda_{n,c}^*(p)=\inf_{p_0}\sup_{\theta}\frac{R_p(\theta,p_0)}{cn^{-1}+\min\left(\theta^2,c_0^2\right)}$$

and

$$p_n=\text{the largest constant attaining }\Lambda_{n,c}^*(p).$$

Then the constant $\Lambda_{n,c}^*(p)$ is the sharp risk upper bound using the minimax optimal thresholding p_n. Note then that

$$R_p(\theta,p_n)\leq\Lambda_{n,c}^*(p_n)\left\{cn^{-1}+\min\left(\theta^2,c_0^2\right)\right\}. \tag{63.9}$$

We have from Lemma 63.7

$$\Lambda_{n,c}^*(p)\leq\Lambda_{n,c}(p)\leq 2c_0^2\log n+\left(c_0^2\sqrt{4/\pi}+c_o^2\right)\log^{1/2}n=c_0^2\left(2\log n+\left(\sqrt{4/\pi}+1\right)\log^{1/2}n\right). \tag{63.10}$$

Theorem 87. *With the universal thresholding* $p_0 = c_0\sqrt{2\log n}$, we have

$$R_p\left(\hat{f}_p, f\right) \leq \Lambda_{n,c}(p)\left\{cn^{-1} + R\left(\hat{f}_0, f\right)\right\}.$$

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

$$R_p\left(\hat{f}_p, f\right) \leq \Lambda^*_{n,c}(p)\left\{cn^{-1} + R\left(\hat{f}_0, f\right)\right\}.$$

Further, $\Lambda_{n,c}(p)$ and $\Lambda^*_{n,c}(p)$ are bounded by (63.10).

Proof. This is a direct result of (63.9) and (63.10) where the oracle risk is the factor of $\min\left(\theta^2, c_0^2\right)$. □

64. Dealing with c_0.

We have used the constant c_0^2 to represent the variance of the wavelet coefficients z. Now we must include the boundedness of c_0 and see what that means in terms of the work of Donoho and Johnstone. The bounds of c_0 depend on what the space is. It will be different for every space and every interpolation method used for dealing with incomplete or irregularly spaced data.

64.1. **The Li and Xiao space of** [18]**.** The bounds of the variance in this setting where ϵ is the noise level in the original data, represented by D_{lm} are for $\alpha \in (0,1)$,

$$\leq \frac{\epsilon^2}{n} + \epsilon^2 C C_4 n^{-\alpha} 2^{-k(1-\alpha)} = O\left(n^{-\alpha}\right).$$

Now we report the boundedness for $\alpha = 1$.

$$\leq \frac{\epsilon^2}{n} + \frac{\epsilon^2}{n} C \log\left(n2^{-k}e\right) = O\left(\frac{\log\left(n2^{-k}e\right)}{n}\right).$$

Here C represents the number of discontinuities that are in the signal. $E(\varepsilon_i\varepsilon_{i+j}) \sim C_0|j|^{-\alpha}$ and

$$C_4 = C_0\int_0^1\int_0^1 |x-y|^{-\alpha}\psi(x)\psi(y)dxdy.$$

Using the fact that these wavelets are compactly supported and several changes of variable, we find

$$C_4 = C_0\int\int \psi\left(\left[(-\alpha+1)z\right]^{\frac{1}{-\alpha+1}} + y\right)\psi(y)\,dzdy.$$

We can see that both of these bounds for the variance are small.

64.2. **The bounds in Part 7.** The bounds of the variance in this setting where ϵ is the noise level in the original data and $\epsilon_1 > 0$ is close to 0, for $\alpha \in (0,1)$,

$$\leq 4p_i\epsilon^2 n^{2\epsilon_1-1}\sup|\psi|^2 + C_7 n^{2\epsilon_1-\alpha-3}p_i^{\alpha}.$$

For $\alpha = 1$,

$$\leq Cp_i n^{2\epsilon_1-1}\log\left(np_i^{-1}e\right). \tag{64.1}$$

Again, we see that the bounds for the variance are small.

64.3. **The bounds in Part 8.** The bounds of the variance in this setting where ϵ is the noise level in the original data, represented by D_{lm} are for $\alpha \in (0,1)$,

$$= \frac{\epsilon^2}{n}h + \epsilon^2 hCC_4 n^{-\alpha}2^{-k(1-\alpha)}$$

for $\alpha = 1$,

$$var\left(\tilde{\theta}_{ij}\right) \leq \frac{\epsilon^2}{n}h + \frac{\epsilon^2}{n}hC\log\left(n2^{-k}e\right)$$

Again, we see that the bounds for the variance are small.

64.4. **Important notes about this paper.** The bounds for the variance of the coefficients are going to vary for every space and every method that one uses. My work has examined two different spaces specifically which used two different methods to find the variance. With the generalized equation which stems from the work of Fan, Donoho and Johnstone, we have extended our ability to understand the boundedness of the risk in terms of the oracle risk. We have created the framework for generally bounding the risk in other spaces and settings.

Part 11. Summary of New Results and Theorems.

65. Theorems from Part 7.

Here we were considering long memory error with irregularly spaced data. We solve the problem of irregularly spaced data by performing an interpolation with a function Y. The variance of the wavelet coefficients, defined by S_{ij} and R_j in (33.13) and (33.6), is bounded as in the following theorem. A closer look at these theorems and the definitions of the constants can be found in Section 33.

Theorem 88. *The bounds of S_{ij}^2 and R_j^2 are as follows.*

$$E\left(S_{ij}^2\right) \leq 4p_i\sigma^2 n^{2\epsilon_1-1}\sup|\psi|^2 + C_7 n^{2\epsilon_1-\alpha-3}p_i^{\alpha}.$$

$$E\left(R_j^2\right) \leq 4p\sigma^2 n^{2\epsilon_1 - 1} \sup |\psi|^2 + C_{7\star} n^{2\epsilon_1 - \alpha - 3} p^{\alpha}.$$

Below we have the theorem which bounds the MISE when using the estimator $\hat{g}$ defined in (65.1).

Theorem 89. *Suppose g is a function supported on $[0,1]$ with certain continuity properties established before. Suppose that the data generated by this function g is long memory and irregularly spaced. Long memory means that*

$$r(j) = E\left(\epsilon_i \epsilon_{i+j}\right) \sim C_0 \left|j\right|^{-\alpha}$$

where $\alpha \in (0,1]$. Suppose we let

$$\hat{g} = \sum_j \hat{a}_j \phi_j + \sum_{i=0}^{q-1} \sum_j \hat{b}_{ij} I\left(\left|\hat{b}_{ij}\right| \geq \delta\right) \psi_{ij}. \tag{65.1}$$

where

$$\hat{a}_j = \int_I Y \phi_j \qquad \hat{b}_{ij} = \int_I Y \psi_{ij}.$$

and Y is some interpolation rule. Then combing (34.7), (34.10), (34.11) and (34.12) yields for $\alpha = 1$

$$\int E\left(\hat{g} - g\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1 - 2} \log\left(np_i^{-1} e\right)\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon} n^{\eta-\lambda-2}\right)$$

$$+ \left(1 - 2^{-2r}\right)^{-1} p^{-2r} \int \left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}\right) + O\left(p_q^{-1}\right)$$

That is

$$= \left(1 - 2^{-2r}\right)^{-1} p^{-2r} \int \left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon} n^{\eta-\lambda-2} + p_q^{-1}\right). \tag{65.2}$$

For $\alpha \in (0,1)$

$$\int E\left(\hat{g} - g\right)^2 = O\left(pn^{\eta-2} + p_i n^{2\epsilon_1 - 1}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon} n^{\eta-\lambda-2}\right)$$

$$+ \left(1 - 2^{-2r}\right)^{-1} p^{-2r} \int \left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!}\right)^2 + o\left(p^{-2r}\right) + O\left(qn^{\eta-1}\right) + O\left(p_q^{-1}\right)$$

That is

$$= \left(1-2^{-2r}\right)^{-1} p^{-2r} \int \left(\frac{g^{(r)}\left(\frac{j}{p_i}\right)}{r!} \right)^2 + o\left(p^{-2r}\right) + O\left(qp_i n^{\eta-1} + \frac{C}{\epsilon} n^{\eta-\lambda-2} + p_q^{-1} \right). \tag{65.3}$$

66. Theorems from Part 8.

Here we were considering long memory error with irregularly spaced data. We re-space the data using a function H. The variance of the wavelet coefficients, defined by $var\left(\tilde{\theta}_{ij}\right)$ in Section 39, is bounded as in the following theorem. A closer look at these theorems and the definitions of the constants can be found in Section 39.

Theorem 90. *The bounds of the variance* $var\left(\tilde{\theta}_{ij}\right)$ *are for* $\alpha \in (0,1)$,

$$= \frac{\epsilon^2}{n} h + \epsilon^2 hCC_4 n^{-\alpha} 2^{-k(1-\alpha)}$$

$$= O\left(n^{-\alpha}\right). \tag{66.1}$$

and for $\alpha = 1$,

$$\leq \frac{\epsilon^2}{n} h + \frac{\epsilon^2}{n} hC \log\left(n2^{-k}e\right)$$

$$= O\left(\frac{\log\left(n2^{-k}e\right)}{n} \right). \tag{66.2}$$

We have the following theorem which bounds the MISE when using our estimator $\hat{f}_n^*$ defined in (66.3).

Theorem 91. *Suppose* f *is a function supported on* $[0,1]$ *with* $f \in \Lambda^\beta(M,B,m)$. *Suppose that the data generated by this function* f *is long memory and irregularly spaced. Long memory means that*

$$r(j) = E\left(\epsilon_i \epsilon_{i+j}\right) \sim C_0 |j|^{-\alpha}$$

where $\alpha \in (0,1]$. *Let*

$$\hat{\xi}_{j_0 k} = \tilde{\xi}_{j_0 k}, \qquad \hat{\theta}_{jk} = sgn\left(\tilde{\theta}_{jk}\right)\left(\left|\tilde{\theta}_{jk}\right| - \lambda_{jk}\right)_+ \tag{66.3}$$

where the threshold λ_{jk} *is derived from an estimate of the variance of the wavelet coefficients. For* $\alpha \in (0,1)$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[n^{-\alpha}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J-j_0)\,2^j\left[n^{-\alpha}\right]$$

$$= o\left(n^{-\alpha}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

For $\alpha = 1$,

$$E\left\|\hat{f}_n^* - f\right\|^2 \leq 2^{j_0}\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right) + (J - j_0)\,2^j\left[\frac{\log\left(n2^{-k}e\right)}{n}\right] + o\left(n^{-2\beta/(1+2\beta)}\right).$$

$$= o\left(\frac{\log\left(n2^{-k}e\right)}{n}\right) + o\left(n^{-2\beta/(1+2\beta)}\right).$$

67. Theorems from Part 10.

Here we expanded the problem of long memory error to the matrix setting and also expanded the work of Donoho and Johnstone in [8].

The following theorem is directly from [1].

Theorem 92. *Let* $p_\lambda(\cdot)$ *be a nonnegative, nondecreasing, and differentiable function in* $(0,\infty)$. *Further, assume that the function* $-\theta - p'_\lambda(\theta)$ *is strictly unimodal on* $(0,\infty)$. *Then we have the following results.*

(1) The solution to the minimization problem (55.1) exists and is unique. It is antisymmetric: $\hat{\theta}(-z) = -\hat{\theta}(z)$.

(2) The solution satisfies

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

where $p_0 = \min_{\theta\geq 0}\{\theta + p'_\lambda(\theta)\}$. Moreover, $\left|\hat{\theta}(z)\right| \leq |z|$.

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

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

(4) When $p'_\lambda(\theta)$ is continuous on $(0,\infty)$, the solution $\hat{\theta}(z)$ is continuous if and only if the minimum of $|\theta| + p'_\lambda(|\theta|)$ is attained at point zero.

(5) If $p'_\lambda(|z|) \to 0$, as $|z| \to +\infty$, then

$$\hat{\theta}(z) = z - p'_\lambda(|z|) + o\left(p'_\lambda(|z|)\right).$$

We also have the following theorem which comes from extending the work in [1].

Theorem 93. *Suppose p satisfies conditions in Theorem 92 and $p'_\lambda(0+)$. Then*

(1) $R_p(\theta, p_0) \le c_0^2 + \theta^2$.

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

$$R_p(\theta, p_0) \le p_0^2 + \sqrt{2/\pi} p_0 c_0 + c_0^2.$$

(3) $R_p(0, p_0) \le c_0^3 \sqrt{2/\pi} \left(\frac{p_0}{c_0} + \frac{c_0}{p_0}\right) \exp\left(-\frac{p_0^2}{2c_0^2}\right)$.

(4) $R_p(\theta, p_0) \le R_p(0, p_0) + \left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2$.

We also have the following Lemma.

Lemma 94. *If the penalty function satisfies conditions of Theorem 92 and $p'_\lambda(\cdot)$ is nonincreasing and $p'_\lambda(0+) > 0$, then*

$$R_p(\theta, p_0) \le \left(2c_0^2 \log n + \left(c_0^2\sqrt{4/\pi} + c_o^2\right)\log^{1/2} n\right)\left\{c/n + \min\left(\frac{1}{2}\left(1 + c_0\sqrt{2/\pi}/2\right)\theta^2, c_0^2\right)\right\}$$

or if c_0 is reasonably small, ie $c_0\sqrt{2/\pi}/2 \le 1$,

$$R_p(\theta, p_0) \le c_0^2\left(2\log n + \left(\sqrt{4/\pi} + 1\right)\log^{1/2} n\right)\left\{c/n + \min\left(\theta^2, c_0^2\right)\right\}$$

for the universal thresholding

$$p_0 = c_0\sqrt{2\log n - \log(1 + d\log n)}, \qquad 0 \le d \le c^2,$$

with $n \ge 4$ and $c \ge 1$ and $p_0 > 1.14$.

The following theorem extends the work of Donoho and Johnstone to accommodate long memory in terms of oracle risk.

Theorem 95. *With the universal thresholding $p_0 = c_0\sqrt{2\log n}$,* we have

$$R_p\left(\hat{f}_p, f\right) \le \Lambda_{n,c}(p)\left\{cn^{-1} + R\left(\hat{f}_0, f\right)\right\}.$$

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

$$R_p\left(\hat{f}_p, f\right) \le \Lambda^*_{n,c}(p)\left\{cn^{-1} + R\left(\hat{f}_0, f\right)\right\}.$$

Further, $\Lambda_{n,c}(p)$ and $\Lambda^*_{n,c}(p)$ are bounded by (63.10).

These are all the significant new results from my work.

Part 12. Conclusion.

We have examined many different kinds of function estimation over the course of this dissertation.

We have made significant advances in dealing with irregularly spaced data with long memory error. In Part 7 we found the bounds associated with using a linear interpolation and local averaging interpolation on the data. In Part 8 we used the more general method of interpolating the data with a function H.

Very many variations of the problem are useful. Many real world problems which have been solved are very oversimplified. In most situations it is not reasonable to assume that data are independent. One example of this is the time series. Here we have data which are dependent. We could use this new research to compare two time series.

Another example of where this research is applicable is in the cause of spatially dependent data. Also, your data points would very likely be unequally spaced. This is why we will try to address the problems of long memory data and unequally spaced data simultaneously.

We have also generalized the problem of long memory to the wavelet setting. The analysis indicates that the method for dealing with incomplete data is still applicable in the case of long memory with a mean square error that is relatable to the oracle risk.

Lastly, we have generalized equations dealing with oracle risk which can be used more generally to find results about the Mean Integrated Square Error in different spaces. This means that we can find more results under different assumptions about the function $f(x)$.

Thanks: I would like to that Dr. Haiyan Cai for his help in my research. I would also like to thank Dr. Ron Dotzel, Dr. Qingtang Jiang and Dr. Wenjie He for their suggestions and a careful reading of my thesis.

References

[1] Antoniadis, Anestis and Fan, Jianqing (2001). Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96(455), pp. 939-967.

[2] Bickel, P. J. (1983). Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics*. New York: Academic Press.

[3] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, 24(6), pp. 2384-2398.

[4] Cai, T. Tony and Brown, Lawrence D. (1998). Wavelet Shrinkage for nonequispaced samples. *Annals of Statistics*, 26(5), pp. 1783-1799.

[5] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics*, 41, pp. 909-996.

[6] Delouille, V. (2009, September 14). An Introduction to Wavelet Analysis. Retrieved from the Connexions Web site: http://cnx.org/content/col10566/1.3/.

[7] Donoho, David L. and Johnstone, Iain M. (1990). Minimax risk over l_p-balls, Technical Report, Department of Statistics, University of California, Berkeley.

[8] Donoho, David L. and Johnstone, Iain M. (1994). Ideal spatial adaption by wavelet shrinkage, *Biometrika*, 81(3), pp. 425-55.

[9] Donoho, David L. and Johnstone, Iain M. (1998). Minimax Estimation via Wavelet Shrinkage, *Annals of Statistics*, 38(3), pp. 879-921.

[10] Donoho, David L., Liu, R. C. and MacGibbon, K. B. (1990). Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, 18, pp. 1416-1437.

[11] Hall, Peter and Turlach, Berwin (1997). Interpolation Methods for Nonlinear Wavelet Regression with Irregularly Spaced Design. *Annals of Statistics*, 25(5), pp. 1912-1925.

[12] Hall, Peter and Hart, Jeffery (1990). Nonparametric Regression with Long-range Dependence. *Stochastic Processes and their Applications*, 36, pp. 339-351.

[13] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the Minimax Optimality of Block Thresholded Wavelet Estimators. *Statistica Sinica*, 9, pp. 33-49.

[14] He, W., Chui, C. K. and Stoeckler, J. (2004). Nonstationary tight wavelet frames, I: bounded intervals. *Applied and Computational Harmonic Analysis*, 17, pp. 141-197.

[15] He, W., Chui, C. K. and Stoeckler, J. (2005). Nonstationary tight wavelet frames, II: unbounded intervals. *Applied and Computational Harmonic Analysis*, 18, pp. 25-66.

[16] Jaffard, S. (1989). Estimation Holderiennes Ponctuelle des fonctions an moyen des coefficients d'ondelettes, *Comptes Rendus Acad. Sciences Paris,* (A) 308(1), pp. 79-81.

[17] Le Cam, Lucien (1986). *Asymptotic Methods in Statistical Decision Theory.* New York: Springer-Verlag.

[18] Li, L and Xiao, Y. (2006). On the minimax optimality of block thresholded wavelet estimators with long memory data. *Journal of Statistical Planning and Inference*, 137, pp. 2850-2869.

[19] Meyer, Y. (1990) *Ondelettes.* Paris: Hermann.

[20] Meyer, Y. (1991). Ondelettes sur l'Intervalle. *Revista Mathematica Ibero-Americana.*

[21] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statis.*, 33, pp. 1065-1076.

[22] Silverman, B. W. (1978). Choosing the window width when estimating a density. *Biometrika*, 65, pp. 1-11.

[23] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.

[24] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, 12, pp. 1285-1297.

[25] Walnut, David F. (2004). *An Introduction to Wavelet Analysis.* Boston: Birkhauser.