CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: A Detailed Analysis of Some Properties of Nonparametric Estimators

AUTHOR: Joshua Miller

SUPERVISOR: Dr. Natalia Stepanova

DATE: May 1st, 2020

Contents

List of Notation

$\mathbf{R}$:	The set of real numbers
$\mathbf{R}_+$:	The set of positive real numbers
$\mathbf{N}$:	The set of positive integers
$A^\top$:	Transpose of a matrix A
$\|\| \cdot \|\|$:	The Euclidean norm of a vector
$\mathrm{E}[\cdot]$:	Expected value of a random variable
$\mathrm{Var}(\cdot)$:	Variance of a random variable
MSE :	Mean square error
MISE :	Mean integrated square error
$\mu(A)$:	The Lebesgue measure of a set $A \subseteq \mathbf{R}$
$\lfloor x \rfloor$:	Greatest integer less than the real number x
$L_p(A)$:	The class of functions for which $\int_A \|f(x)\|^p dx < \infty$, where $A \subseteq \mathbf{R}$
$L_\infty(A)$:	The class of functions for which the essential supremum is finite on a set $A \subseteq \mathbf{R}$
$\hat{K}(\omega)$:	The Fourier transform of a kernel K at the point $\omega \in \mathbf{R}$
$N(\mu, \sigma^2)$:	A normal distribution with mean $\mu \in \mathbf{R}$ and variance $\sigma^2 \in \mathbf{R}_+$
$I(A)$:	The indicator function of a set A
$a_n = o(b_n)$:	$\lim_{n\to\infty} a_n/b_n = 0$
$a_n = O(b_n)$:	$\lim_{n\to\infty} a_n/b_n = C \neq 0$
$\tilde{W}(\beta, L)$:	Sobolev class of smooth functions, defined on page 16
$P(\beta, L)$:	The Hölder class of densities, defined on page 17
$P_S(\beta, L)$:	The Sobolev class of densities, defined on page 39
$\Sigma(\beta, L)$:	The Hölder class of smooth functions, defined on page 10
$F(x)$:	The cumulative distribution function (CDF)
$F_n(x)$:	The empirical distribution function (EDF)
δ_{ij} :	The Kronecker delta function
$\limsup a_n$:	$\limsup a_n = \lim_{n\to\infty} \left(\sup_{m \geq n} a_n \right)$

Abstract

In a classical framework, the problem of drawing statistical inference from a population can often be reduced to estimating the distribution function of some assumed probability distribution indexed by a finite-dimensional parameter, say, $\theta \in \Theta \subseteq \mathbf{R}^m, m \geq 1$. For example, if one assumes that a random sample $X_1, \ldots, X_n$ is drawn from $N(\mu, \sigma^2), \mu \in \mathbf{R}, \sigma^2 \in \mathbf{R}_+$, then the problem reduces to obtaining estimators of μ and σ^2. A collection of independent random variables $X_1, \ldots, X_n$ with a common distribution function $F \in \{F_\theta : \theta \in \Theta\}$, where F_θ is of known functional form depending on an unknown parameter θ is known as a parametric model. A model with a distribution function that assumes no functional form is known as a nonparametric model. Typically, parametric estimation is relatively simpler and more convenient than nonparametric estimation. If the assumptions made on a parametric model are valid, then the estimators typically have more power then their nonparametric counterparts. Unfortunately, in realistic scenarios, it is rarely justified to make strong assumptions about the data in question, and the parametric assumptions that are commonly made may not be appropriate. In this project, we investigate mathematical properties of various nonparametric estimation procedures under a variety of assumptions. A particular emphasis is placed on deriving upper bounds on the mean square error (MSE) and mean integrated square error (MISE) of the studied estimators. Three estimation procedures and their properties will be studied, namely: local polynomial regression estimation, orthogonal series estimation, and kernel density estimation. The project begins by presenting the models and estimators in more detail, followed by the solutions to various problems posed in the text Introduction to Nonparametric Estimation, by Tsybakov [1]. In addition, a few applied examples will be presented to demonstrate the estimation procedures in practice.

1 Introduction

The work of this project extends material covered in STAT 4506, an introductory course in nonparametric statistics that I had the pleasure of taking in the winter of 2019. In this section, we shall present two nonparametric models from the literature, introduce estimators under the model based assumptions, and provide some justification as to why the estimators are good.

In general, the quality of the presented estimators will be studied in terms of their mean square error (MSE) and mean integrated squared error (MISE). Consider a collection of random variables $X_1, \ldots, X_n$, taking values in $\mathbf{R}$, and an estimator $\hat{f}_n(x) = \hat{f}_n(x, X_1, \ldots, X_n)$ of a function $f : \mathbf{R} \to \mathbf{R}$. For a fixed point $x_0 \in \mathbf{R}$, the quantity

$$\mathrm{MSE}(\hat{f}_n(x_0)) = \mathrm{E}\left[\left(\hat{f}_n(x_0) - f(x_0)\right)^2\right], \tag{1.1}$$

where expectation is taken with respect to the distribution of $(X_1, \ldots, X_n)$, is referred to as the MSE of the estimator $\hat{f}_n(x)$ at x_0. It is often convenient to express the MSE in the form

$$\mathrm{MSE}(\hat{f}_n(x_0)) = \mathrm{Var}(\hat{f}_n(x_0)) + Bias^2(\hat{f}_n(x_0)), \tag{1.2}$$

where $Bias(\hat{f}_n(x_0))$ is the bias of $\hat{f}_n(x_0)$ and $\mathrm{Var}(\hat{f}_n(x_0))$ is its variance. Note that expressions (1.1) and (1.2) only assess the quality of the estimator $\hat{f}_n(x)$ at a fixed point in the domain of f. To assess the global quality of an estimator, one may define the MISE of an estimator $\hat{f}_n(x)$ of $f(x)$. The quantity

$$\mathrm{MISE} = \mathrm{E}\left[\int_{-\infty}^{\infty} (\hat{f}_n(x) - f(x))^2 dx\right], \tag{1.3}$$

where expectation is taken with respect to $(X_1, \ldots, X_n)$, is referred to as the MISE of the estimator $\hat{f}_n(x)$ of $f(x)$. As was the case for MSE, it is often more convenient to express MISE in the equivalent form

$$\text{MISE} = \int_{-\infty}^{\infty} \text{Var}(\hat{f}_n(x))dx + \int_{-\infty}^{\infty} Bias^2(\hat{f}_n(x))dx. \tag{1.4}$$

Three nonparametric estimators will be studied, namely: kernel density estimators, local polynomial regression estimators, and orthogonal series estimators. After introducing the models and estimators, we shall prove certain mathematical properties of the studied estimators, or extensions of the estimators, as suggested in Introduction to Nonparametric Estimation, by Tsybakov [1]. In addition, we shall present a few applied examples where the estimation procedures are implemented.

1.1 Kernel Density Estimation

Consider a random sample $X_1, \ldots, X_n$, where X_1 is a random variable taking its values in $\mathbf{R}$, with cumulative distribution function (CDF) $F(x) = \int_{-\infty}^{x} p(y)dy$. Suppose one wishes to derive an estimator of $p(x)$. Consider first the empirical distribution function (EDF) evaluated at $x \in \mathbf{R}$, defined as

$$F_n(x) \coloneqq \sum_{i=1}^{n} \frac{I(X_i \leq x)}{n}. \tag{1.5}$$

Note that

$$\text{E}[F_n(x)] = \text{E}[I(X_1 \leq x)] = P(X_1 \leq x) = F(x). \tag{1.6}$$

Thus the EDF at any $x \in \mathbf{R}$ is an unbiased estimator of the CDF $F(x)$. Also, by independence, we have

$$\text{Var}[F_n(x)] = \frac{\text{Var}(I(X_1 \leq x))}{n} = \frac{F(x)(1 - F(x))}{n}. \tag{1.7}$$

By the law of large numbers, and using relations (1.6) and (1.7), it can be shown that $F_n(x)$ is a consistent estimator of the distribution function $F(x)$, at any point x such that $F(x) \neq 0$ and $F(x) \neq 1$. Now, note that

$$\lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = \frac{dF(x)}{dx} = p(x).$$

Thus, if $h = h_n$ tends to zero as n tends to infinity, we have that

$$\hat{p}_n(x) \coloneqq \frac{F_n(x+h) - F_n(x-h)}{2h} \xrightarrow{P} p(x), \quad x \in \mathbf{R}.$$

We now introduce the notion of a kernel function. Consider the estimator above and observe that for any $x \in \mathbf{R}$ and $h > 0$

$$\frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{2nh} \sum_{i=1}^{n} I(x-h \leq X_i \leq x+h). \tag{1.8}$$

Setting $K(x) \coloneqq \frac{1}{2} I(|x| \leq 1)$, we can consider the above relation in the form

$$\frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{1.9}$$

More generally, it has been shown by Parzen [2] that for any bounded even function $K(x), x \in \mathbf{R}$, satisfying $\int_{-\infty}^{\infty} |K(x)|dx < \infty, \int_{-\infty}^{\infty} K(x)dx = 1, \lim_{x\to\infty} |xK(x)| = 0$ with bandwidth $h = h_n > 0$ such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$ the estimator

$$\hat{p}_n(x) \coloneqq \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{1.10}$$

is a *consistent* estimator of a density $p(x)$. In Part A of Exercise 1.2 of the project, the notion of a kernel density estimator will be extended to include estimators of the derivative of a density $p(x)$, denoted $p^{(m)}(x), m \geq 1$, provided that the derivative exists and belongs to a suitable class of functions defined later. In Exercise 1.5, we shall study two kernel functions, and show that they are *inadmissible* with respect to MISE. An inadmissible estimator will be defined rigorously later in the project. In Exercise 1.6, we shall study a sufficient condition for which the MISE of a kernel density estimator is $O(n^{\frac{-2\beta}{2\beta+1}})$, where n is the sample size, and the density $p(x)$ belongs to a Sobolev class of densities, with smoothness parameter $\beta > 0$, which will be defined later. In Exercise 1.8, we shall derive an upper bound on the MISE when the sinc kernel is used. Consider the visual below, found in The Elements of Statistical Learning, by Hastie et al. [3] as an example of kernel density estimation.

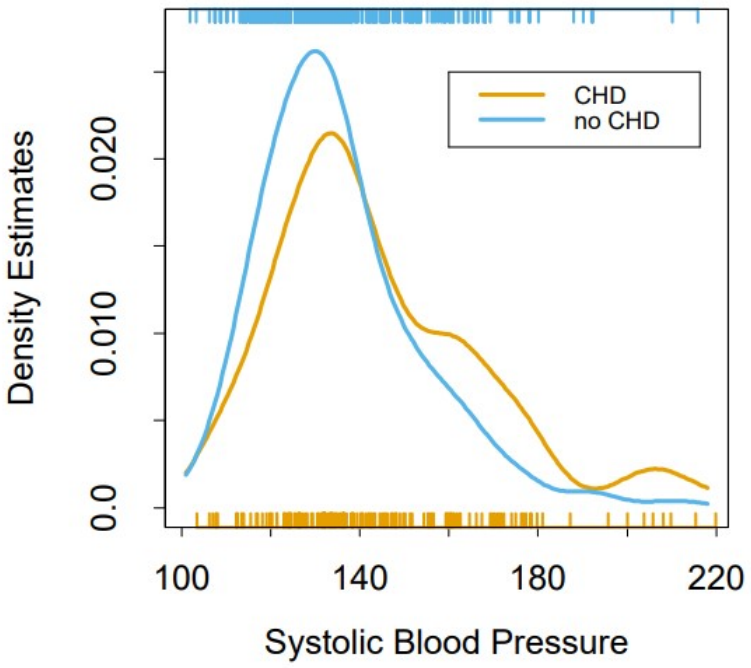


Figure 1: *Density estimates of the systolic blood pressure for an individual with or without coronary heart disease (CHD). The vertical dashes represent realized values from the random sample, and the curve represents the estimated density for different values of systolic blood pressure.*

We now shift our focus to nonparametric regression estimation.

1.2 Nonparametric Regression Estimation

Consider a collection of n independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. We shall focus our attention on a nonparametric regression model of the form

$$Y_i = f(X_i) + \xi_i, \quad X_i \in [0,1], \quad i = 1, \ldots, n, \tag{1.11}$$

where $\xi_1, \ldots, \xi_n$ are independent random variables with $\mathrm{E}[\xi_i] = 0$ and $\mathrm{E}[\xi_i^2] < \infty, i = 1, \ldots, n$. Suppose that interest lies in deriving an estimator of the unknown regression function $f : [0,1] \to \mathbf{R}$. The nonparametric regression problem can be considered under two cases:

1. Nonparametric regression with fixed design: In this scenario, we consider $X_1, \ldots, X_n$ as fixed and deterministic quantities. In particular, we shall study the case where $X_i = i/n, i = 1, \ldots, n$.

2. Nonparametric regression with random design: In this case, the design points $X_1, \ldots, X_n$ are random variables.

In this project, local polynomial regression estimators and orthogonal series estimators will both be investigated in the case of nonparametric regression with fixed design.

1.2.1 Local Polynomial Regression Estimation

Suppose the assumptions of model (1.11) are satisfied and such that the design points $X_1, \ldots, X_n$ are fixed with $X_i = i/n, i = 1, \ldots, n$. Assume further that the regression function f belongs to the Hölder class of functions, defined by

$$\Sigma(\beta, L) = \big\{ f : [0,1] \to \mathbf{R}, f \text{ is } \ell = \lfloor \beta \rfloor \text{ times differentiable,} \\ |f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'|^{\beta - \ell}, \forall x, x' \in [0,1] \big\}, \quad (1.12)$$

where $\beta > 0$, $L > 0$. Let x_0 be some point in the domain of f. Suppose we wish to derive an estimator of the regression function f at a fixed point x in the domain of f. Note that

$$f(x_0) = f(x) + \frac{f'(x)}{1!}(x_0 - x) + \frac{f''(x)}{2!}(x_0 - x)^2 + \ldots + \frac{f^{(\ell)}(x')}{\ell!}(x_0 - x)^\ell,$$

where x' is between x_0 and x. If we assume that x is very close to x_0, then the Hölder condition implies that the remainder term $\frac{f^{(\ell)}(x')}{\ell!}(x_0 - x)^\ell$ is approximately the same as $\frac{f^{(\ell)}(x)}{\ell!}(x_0 - x)^\ell$, giving us

$$\begin{aligned} f(x_0) &\approx f(x) + \frac{f'(x)}{1!}(x_0 - x) + \frac{f''(x)}{2!}(x_0 - x)^2 + \ldots + \frac{f^{(\ell)}(x)}{\ell!}(x_0 - x)^\ell \\ &= \theta^\top(x) U\left(\frac{x_0 - x}{h}\right), \end{aligned} \quad (1.13)$$

where

$$U(u) \coloneqq \left(1, u, u^2/2!, \ldots, u^\ell/\ell!\right)^\top, \tag{1.14}$$

$$\theta(x) \coloneqq \left(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}(x)h^\ell\right)^\top, \tag{1.15}$$

and $h > 0$ is a bandwidth. The local polynomial estimator will now be presented as a weighted least square estimator. Let $K : \mathbf{R} \to \mathbf{R}$ be a kernel, and let $\ell \geq 0$ be an integer. A vector $\hat{\theta}_n(x) \in \mathbf{R}^{\ell+1}$ defined by

$$\hat{\theta}_n(x) = \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right), \tag{1.16}$$

is called a *local polynomial estimator of order ℓ of $\theta(x)$* or $LP(\ell)$ estimator of $\theta(x)$ for short. The statistic $\hat{f}_n(x) = U^\top(0)\hat{\theta}_n(x)$ is called a *local polynomial estimator of order ℓ of $f(x)$* or $LP(\ell)$ estimator of $f(x)$ for short. Note that the Taylor series approximation in relation (1.13) is only approximately valid for x_0 close to x. The kernel weights $K\left(\frac{X_i - x}{h}\right), i = 1, \ldots, n$, will downplay the observations for which X_i is not close to x. This is why the fitting procedure is deemed "local". Now, note that

$$\begin{aligned}
\hat{\theta}_n(x) &= \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left[Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right) \\
&= \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left(Y_i^2 K\left(\frac{X_i - x}{h}\right) - 2Y_i\theta^\top U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right. \\
&\qquad \left. + \theta^\top U\left(\frac{X_i - x}{h}\right)\theta^\top U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right).
\end{aligned}$$

Since $\sum_{i=1}^{n} Y_i^2 K\left(\frac{X_i - x}{h}\right)$ does not depend on θ, we can write

$$\begin{aligned}
\hat{\theta}_n(x) = \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} &\left(-2Y_i\theta^\top U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right. \\
&\left. + \theta^\top U\left(\frac{X_i - x}{h}\right)\theta^\top U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right)
\end{aligned}$$

$$= \arg \min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \Big(-2Y_i \theta^\top U \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right) + \theta^\top U \left(\frac{X_i - x}{h} \right) U^\top \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right) \theta \Big)$$

$$= \arg \min_{\theta \in \mathbf{R}^{\ell+1}} \left(-2\theta^\top a_{nx} + \theta^\top B_{nx} \theta \right) \tag{1.17}$$

$$= \arg \min_{\theta \in \mathbf{R}^{\ell+1}} R(\theta), \tag{1.18}$$

where

$$a_{nx} = \frac{1}{nh} \sum_{i=1}^{n} Y_i U \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right), \tag{1.19}$$

$$B_{nx} = \frac{1}{nh} \sum_{i=1}^{n} U \left(\frac{X_i - x}{h} \right) U^\top \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right), \tag{1.20}$$

$$R(\theta) = \left(-2\theta^\top a_{nx} + \theta^\top B_{nx} \theta \right). \tag{1.21}$$

Define the derivative of $R(\theta)$ with respect to $\theta = (\theta_1, \ldots, \theta_{\ell+1})^\top$ by the vector $\frac{dR(\theta)}{d\theta} = \left(\frac{\partial R(\theta)}{\partial \theta_1}, \ldots, \frac{\partial R(\theta)}{\partial \theta_{\ell+1}} \right)^\top$. Taking the derivative of $R(\theta)$ with respect to θ, we obtain

$$\frac{dR(\theta)}{d\theta} = -2 \frac{d}{d\theta} (\theta^\top a_{nx}) + \frac{d}{d\theta} (\theta^\top B_{nx} \theta). \tag{1.22}$$

It can be shown that

$$\frac{d}{d\theta} (\theta^\top a_{nx}) = a_{nx}, \text{ and } \frac{d}{d\theta} (\theta^\top B_{nx} \theta) = 2B_{nx} \theta,$$

hence

$$\frac{dR(\theta)}{d\theta} = -2a_{nx} + 2B_{nx} \theta. \tag{1.23}$$

Setting (1.23) equal to the zero vector, it can be seen that if B_{nx} is positive semi-definite, then the minimizing vector $\hat{\theta}_n(x)$ will satisfy the normal equations

$$B_{nx} \hat{\theta}_n(x) = a_{nx}.$$

If in addition, the matrix B_{nx} is positive definite, the unique solution $\hat{\theta}_n(x)$ of the normal equations will be

$$\hat{\theta}_n(x) = B_{nx}^{-1} a_{nx}, \tag{1.24}$$

giving us an $LP(\ell)$ estimator of $f(x)$ of the form

$$\hat{f}_n(x) = (U(0))^\top B_{nx}^{-1} a_{nx}. \tag{1.25}$$

We will also make the following assumptions in our setup (see page 37 of Tsybakov [1]):

(a) there exists a real number $\lambda_0 > 0$ such that the smallest eigenvalue $\lambda_{min}(B_{nx})$ of B_{nx} satisfies $\lambda_{min}(B_{nx}) \geq \lambda_0$;

(b) there exists a real number $a_0 > 0$ such that for any interval $A \subseteq [0,1]$ and all $n \geq 1$, $\frac{1}{n}\sum_{i=1}^{n} I\left(X_i \in A\right) \leq a_0 \max\left(\mu(A); \frac{1}{n}\right)$, where $\mu(A)$ denotes the Lebesgue measure of A,

(c) the kernel K has compact support belonging to $[-1,1]$ and there exists a number $K_{max} < \infty$ such that $|K(u)| \leq K_{max}, \forall u \in [-1,1]$.

In Theorem 1.6 of Tsybakov [1], it is shown that under assumptions (a)-(c), and for a suitably chosen bandwidth $h = h_n$, we have that

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathrm{E}\left[n^{\frac{2\beta}{2\beta+1}} |\hat{f}_n(x_0) - f(x_0)|^2\right] \leq C(\beta, L, \lambda_0, a_0, \sigma^2_{max}, K_{max}) < \infty, \tag{1.26}$$

where $\hat{f}_n(x_0)$ is the $LP(\ell)$ estimator of $f(x)$ of order ℓ evaluated at x_0, and $C(\beta, L, \lambda_0, a_0, \sigma^2_{max}, K_{max})$ is a positive constant depending on the constants within the parentheses. Hence, for any $x_0 \in [0,1]$, the estimator $\hat{f}_n(x_0)$ converges in probability to $f(x_0)$. In Exercise 1.4, the notion of a $LP(\ell)$ estimator will be extended to include estimators of the derivatives of $f(x)$, and suitable bounds analogous to that of (1.26) will be derived.

Consider the visual below that demonstrates the use of local polynomial regression estimation, found in Eubank [4]. Observing the below fit, we can see that the fitted regression function reasonably describes the behavior of the data. If one were to assume a functional form for the regression function f, it would be unclear how to correctly specify such a function as the responses behave differently for different values of t. From $t = 0$ to $t = 10$, y is roughly constant, it then exhibits a U shape between 10 and 30, and then slowly decreases between $t = 35$ and $t = 60$.

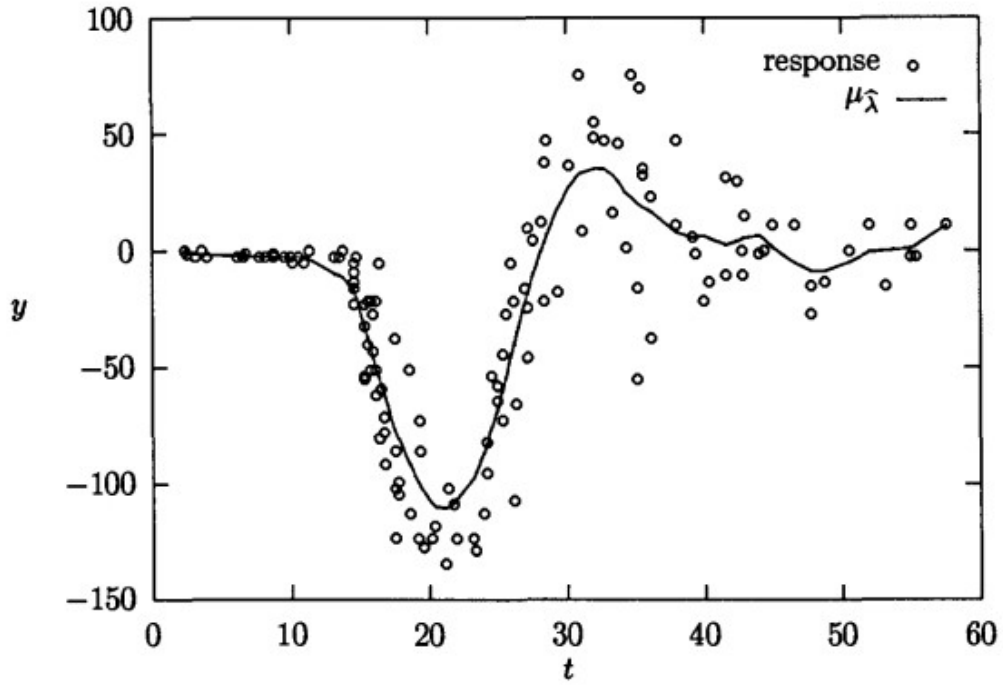


Figure 2: *The above plot is a local polynomial regression function fitted to simulated motorcycle crash data. $\mu_{\hat{\lambda}}$ represents the fitted regression function. In our notation, $\mu_{\hat{\lambda}}$ represents the estimate $\hat{f}_n(x)$. The responses $y_i, i = 1, \ldots, n$, represent the acceleration of a human head at time t milliseconds after a crash.*

We now shift our attention to orthogonal series estimators.

1.2.2 Orthogonal Series Estimators

Suppose that the assumptions of model (1.11) are valid and the design points $X_1, \ldots, X_n$ are fixed and such that $X_i = i/n, i = 1 \ldots, n$. We also assume that the regression function $f \in L_2[0,1]$, finally, we also assume that the regression function $f \in L_2[0,1]$ admits a Fourier series representation defined as

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x), \tag{1.27}$$

where $\{\phi_j(x)\}_{j=1}^{\infty}$ is an orthonormal basis in $L_2[0,1]$ and the Fourier coefficients $\{\theta_j\}_{j=1}^{\infty}$ are such that $\sum_{j=1}^{\infty} |\theta_j|^2 < \infty$. When we speak of two functions ϕ_i and ϕ_j being orthonormal in $L_2[0,1]$, we mean that $\int_0^1 \phi_i(x)\phi_j(x)dx = \delta_{ij}$, where δ_{ij} is the Kronecker delta function. The idea behind orthogonal series estimation is to assume that $f(x)$ can be well approximated by the first N basis functions in (1.27). The orthogonal series estimator, also called the projection estimator, with tuning parameter $N \in \mathbf{N}$ is given by

$$\hat{f}_{nN}(x) = \sum_{j=1}^{N} \hat{\theta}_j \phi_j(x), \quad x \in \mathbf{R}, \tag{1.28}$$

where

$$\hat{\theta}_j = \sum_{i=1}^{n} \frac{Y_i \phi_j(X_i)}{n}, \quad j = 1, \ldots, N, \tag{1.29}$$

is typically an unbiased estimator of θ_j under certain conditions on $\{\phi_j(x)\}_{j=1}^{\infty}$ and X_i.

Orthogonal series estimators can be shown to perform quite well under very weak assumptions. Assume that the orthonormal basis $\{\phi_j(x)\}_{j=1}^{\infty}$ in (1.27) and (1.28) is the trigonometric basis, defined by

$$\phi_j(x) = \begin{cases} 1, & j = 1, \\ \sqrt{2}\sin(\pi(j-1)x), & j \text{ is odd}, \\ \sqrt{2}\cos(\pi j x), & j \text{ is even}. \end{cases} \tag{1.30}$$

Assuming representation (1.27) holds, define the Sobolev class of smooth functions as

$$\tilde{W}(\beta, L) = \left\{ f \in L_2[0,1] : \theta \in \ell^2(\mathbf{N}), \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q \right\}, \tag{1.31}$$

where $\theta = \{\theta_j\}_{j=1}^{\infty}$, and $\ell^2(\mathbf{N}) = \{c : \sum_{j=1}^{\infty} c_j^2 < \infty\}$, $Q = L^2/\pi^{2\beta}, L > 0, \beta > 0$, and

$$a_j = \begin{cases} (j-1)^{\beta}, & j \text{ is odd}, \\ j^{\beta}, & j \text{ is even}. \end{cases} \tag{1.32}$$

Under the assumptions that

1. the orthonormal basis is given by (1.30),
2. the regression function f belongs to $\tilde{W}(\beta, L)$, $\beta \geq 1$, $L > 0$,
3. $\sum_{j=1}^{\infty} |\theta_j| < \infty$,

it can be shown quite easily that if $N = N_n$ is selected such that $N = \lfloor \alpha n^{\frac{1}{2\beta+1}} \rfloor$, where $\alpha > 0$, then as $n \to \infty$, the maximum MISE of the orthogonal series estimator $\hat{f}_{nN}(x)$ is of order $O(n^{\frac{-2\beta}{2\beta+1}})$, implying that orthogonal series estimators are consistent estimators. Refer to Section 1.7 of Tsybakov [1] for further details. In Exercise 1.9 below, we shall show that orthogonal series estimators can also be used to estimate a density function $p \in L_2[0,1]$, and derive some special properties of such an estimator. In Exercise 1.11, we shall present the weighted orthogonal series estimator, and show that it performs uniformly better then the projection estimator defined by (1.28), in terms of MISE.

Consider the data visualization given in Figure 3 as an example of Orthogonal Series Estimation, found in Nonparametric Regression and Spline Smoothing, by Eubank [4]. Observing Figure 3, we can see that the regression estimator $\hat{f}_{nN}(t)$ almost perfectly models the data everywhere in t.

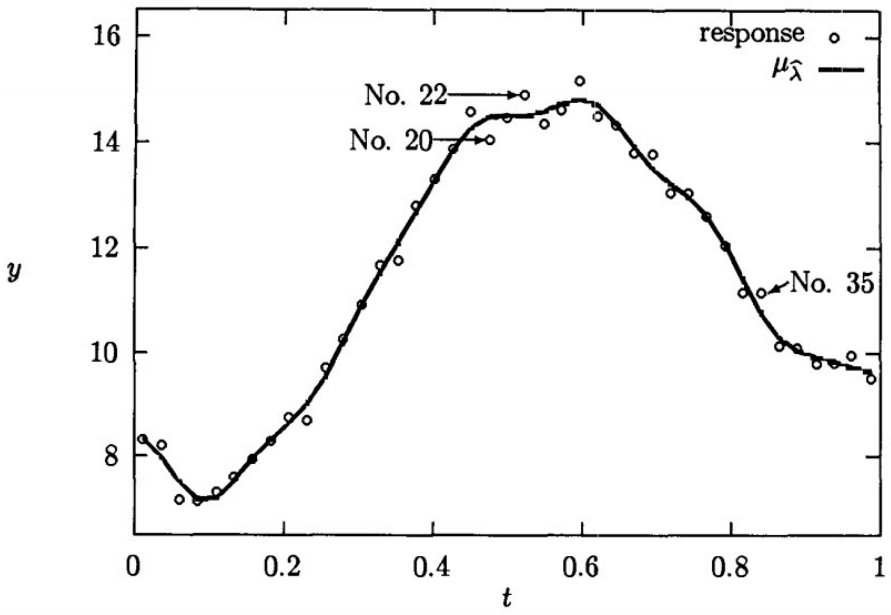


Figure 3: *The plot of a regression function fitted via orthogonal series estimation.* $\mu_{\hat{\lambda}}$ *represents the fitted regression function, which is* $\hat{f}_{nN}(x)$ *in our notation. The data represents the voltage drop of a guided missile battery at time* t*, where the number of basis functions is* $N = 14$*. The basis* $\{\phi_j(t)\}_{j=1}^{\infty}$ *that is used in this problem is given by* $\phi_1(t) = 1, \phi_j(t) = \sqrt{2}\cos((j-1)\pi t), j = 2, \ldots, 14$*.*

The remainder of the project will be devoted to solving selected problems from Introduction to Nonparametric Estimation, by Tsybakov [1], and demonstrating a few applications.

2 Exercise 1.2 (Tsybakov Page 73)

Suppose that $P(\beta, L)$ is the Hölder class of densities given by

$$P(\beta, L) = \Big\{ p(x) : p(x) \geq 0, \int_{-\infty}^{\infty} p(x)dx = 1, \\ |p^{(\ell)}(x) - p^{(\ell)}(x')| \leq L|x - x'|^{\beta-\ell}, \forall x, x' \in \mathbf{R} \Big\}, \quad (2.1)$$

where p is an ℓ-times differentiable probability density function, $L > 0, \beta > 0, \ell = \lfloor \beta \rfloor$.

A kernel estimator of the sth derivative $p^{(s)}$ of density $p \in P(\beta, L), s < \beta$, can be defined as follows:

$$\hat{p}_{n,s}(x) = \frac{1}{nh^{s+1}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{2.2}$$

Here $h > 0$ is a bandwidth and $K : \mathbf{R} \to \mathbf{R}$ is a bounded kernel of support $[-1, 1]$ satisfying for some $0 < s < \ell$:

$$\int_{-1}^{1} u^j K(u) du = 0, j = 0, 1, \ldots, s-1, s+1, \ldots, \ell, \tag{2.3}$$

$$\int_{-1}^{1} u^s K(u) du = s!. \tag{2.4}$$

2.1 Part 1

Problem: Consider observing a sequence $X_1, X_2, \ldots$ of independent identically distributed random variables with density $p \in P(\beta, L)$. Prove that, uniformly over the class of probability densities $p \in P(\beta, L)$, the bias of $\hat{p}_{n,s}(x_0)$ is bounded by $ch^{\beta-s}$ and the variance of $\hat{p}_{n,s}(x_0)$ is bounded by $\frac{c'}{nh^{2s-1}}$ where $c, c' > 0$ are appropriate constants and x_0 is a given point in $\mathbf{R}$.

Solution: The bias of $\hat{p}_{n,s}(x_0)$ equals

$$\begin{aligned}
Bias\left(\hat{p}_{n,s}(x_0)\right) &= \mathrm{E}\left[\hat{p}_{n,s}(x_0)\right] - p^{(s)}\left(x_0\right) \\
&= \mathrm{E}\left[\frac{1}{nh^{s+1}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\right] - p^{(s)}(x_0) \\
&= \mathrm{E}\left[\frac{1}{h^{s+1}} K\left(\frac{X_1 - x_0}{h}\right)\right] - p^{(s)}(x_0) \\
&= \int_{-1}^{1} \frac{1}{h^s} K(u) p\left(x_0 + uh\right) du - p^{(s)}\left(x_0\right).
\end{aligned}$$

Taking a Taylor expansion of $p(x_0 + uh)$ at the point x_0, we get

$$Bias\left(\hat{p}_{n,s}(x_0)\right) = \frac{1}{h^s}\int_{-1}^{1} K(u)\left(p(x_0) + \ldots + \frac{(uh)^s}{s!}p^{(s)}\left(x_0\right) + \right.$$
$$\left. \ldots + \frac{(uh)^\ell}{\ell!}p^{(\ell)}\left(x_0 + \tau uh\right)\right) du - p^{(s)}(x_0),$$

where $0 \leq \tau \leq 1, 1 \leq s \leq \ell$. By (2.3) and (2.4), we have

$$Bias\left(\hat{p}_{n,s}(x_0)\right)$$
$$= 0 + \ldots + \frac{h^s s!}{h^s s!}p^{(s)}(x_0) + \int_{-1}^{1}\frac{(uh)^\ell K(u)p^{(\ell)}(x_0 + \tau uh)}{h^s \ell!}du - p^{(s)}(x_0)$$

$$= \int_{-1}^{1}\frac{(uh)^\ell K\left(u\right)p^{(\ell)}\left(x_0 + \tau uh\right)}{h^s \ell!}du.$$

Recall from (2.3) that $\int_{-1}^{1} u^\ell K(u)du = 0$, therefore

$$\int_{-1}^{1}\frac{(uh)^\ell K(u)p^{(\ell)}(x_0)}{h^s \ell!}du = 0.$$

We may thus write

$$\left|Bias\left(\hat{p}_{n,s}(x_0)\right)\right|$$
$$= \left|\int_{-1}^{1}\frac{(uh)^\ell K(u)p^{(\ell)}\left(x_0 + \tau uh\right)}{h^s \ell!}du - \int_{-1}^{1}\frac{(uh)^\ell K(u)p^{(\ell)}(x_0)}{h^s \ell!}du\right|.$$

Assuming that $p \in P(\beta, L)$, we have from the above relation that

$$\left|Bias\left(\hat{p}_{n,s}(x_0)\right)\right| \leq \frac{h^{\ell-s}}{\ell!}\int_{-1}^{1}\left|u^\ell K(u)\left(p^{(\ell)}\left(x_0 + \tau uh\right) - p^{(\ell)}(x_0)\right)\right| du$$
$$\leq \frac{h^{\ell-s}}{\ell!}\int_{-1}^{1}\left|u^\ell LK(u)(\tau uh)^{\beta-\ell}\right| du.$$

By the assumption that the kernel K is a bounded function, we can define

$$K_{max} = \sup_{t\in[-1,1]} |K(t)| < \infty.$$

Then for some $c = c(\beta, L) > 0$,

$$\begin{aligned}|Bias(\hat{p}_{n,x}(x_0))| &\le \frac{h^{\ell-s}}{\ell!}\int_{-1}^{1}\left|u^{\ell}LK(u)(\tau uh)^{\beta-\ell}\right| du \\ &\le \frac{h^{\beta-s}\tau^{\beta-\ell}LK_{max}}{\ell!}\int_{-1}^{1}|u|^{\beta}\,du \le ch^{\beta-s}.\end{aligned}$$

Thus, for any given $\beta > 0$ and $L > 0$,

$$\sup_{p\in P(\beta,L)} |Bias(\hat{p}_{n,s}(x_0))| \le ch^{\beta-s}, \tag{2.5}$$

with some constant $c = c(\beta, L) > 0$. Next,

$$\mathrm{Var}\left(\hat{p}_{n,s}(x_0)\right) = \mathrm{Cov}\left(\frac{1}{nh^{s+1}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right), \frac{1}{nh^{s+1}}\sum_{j=1}^{n} K\left(\frac{X_j - x_0}{h}\right)\right).$$

Recall that if a random variable X is independent of some other random variable Y, then $g(X)$ is independent of $g(Y)$, implying $\mathrm{Cov}(g(X), g(Y)) = 0$. Thus

$$\begin{aligned}\mathrm{Var}(\hat{p}_{n,s}(x_0)) &= \mathrm{Cov}\left(\frac{1}{nh^{s+1}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right), \frac{1}{nh^{s+1}}\sum_{j=1}^{n} K\left(\frac{X_j - x_0}{h}\right)\right) \\ &= \frac{1}{n^2h^{2s+2}}\sum_{i=1}^{n}\mathrm{Cov}\left(K\left(\frac{X_i - x_0}{h}\right), K\left(\frac{X_i - x_0}{h}\right)\right) \\ &= \frac{1}{nh^{2s+2}}\mathrm{Var}\left(K\left(\frac{X_1 - x_0}{h}\right)\right) \le \frac{1}{nh^{2s+2}}\mathrm{E}\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right] \\ &= \frac{1}{nh^{2s+2}}\int_{x_0-h}^{x_0+h} K^2\left(\frac{x - x_0}{h}\right)p(x)dx = \frac{1}{nh^{2s+1}}\int_{-1}^{1} K^2(u)p(x_0 + uh)du \\ &\le \frac{K_{max}^2}{nh^{2s+1}}\int_{-1}^{1} p(x_0 + uh)du = \frac{c'}{nh^{2s+1}}.\end{aligned}$$

Thus, for any given $\beta > 0$ and $L > 0$,

$$\sup_{p\in P(\beta,L)} |\mathrm{Var}(\hat{p}_{n,s}(x_0)| \le \frac{c'}{nh^{2s+1}}, \tag{2.6}$$

with some constant $c' = c'(\beta, L) > 0$.

2.2 Part 2

Problem: Prove that the maximum of the MSE of $\hat{p}_{n,s}(x_0)$ over $P(\beta, L)$ is of order $O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right)$ as $n \to \infty$ if the bandwidth $h = h_n$ is chosen optimally.

Solution: Recall that the MSE of an estimator $\hat{\theta}$ of θ is equal to $\text{Var}(\hat{\theta}) + Bias^2(\hat{\theta})$. With this in mind, using bounds (2.5) and (2.6), we have that

$$\begin{aligned} \text{MSE}(h) &= Bias^2(\hat{p}_{n,s}(x_0)) + \text{Var}(\hat{p}_{n,s}(x_0)) \\ &\leq c^2 h^{2\beta-2s} + c' n^{-1} h^{-2s-1} =: g(h). \end{aligned} \tag{2.7}$$

Then

$$g'(h) = c^2(2\beta - 2s)h^{2\beta-2s-1} - c'n^{-1}(2s+1)h^{-2s-2},$$

and also

$$\begin{aligned} g''(h) = c^2 \left(2\beta - 2s\right)\left(2\beta - 2s - 1\right) h^{2\beta-2s-2} \\ + cn^{-1}(2s+1)(2s+2)h^{-2s-3} > 0, \quad \forall h > 0. \end{aligned}$$

Therefore the root of $g'(h)$ will correspond to a global minimizer of $g(h)$. Setting $g'(h) = 0$ and rearranging, we have that $h = h_n$ must satisfy

$$\begin{aligned} h^{2\beta+1} &= \frac{c'(2s+1)}{nc^2(2\beta-2s)} \\ h &= \left(\frac{c'(2s+1)}{n(2\beta-2s)\left(c'\right)^2}\right)^{\frac{1}{2\beta+1}}. \end{aligned}$$

Therefore, as $n \to \infty$

$$h_n = O(n^{-\frac{1}{2\beta+1}}).$$

Substituting h_n into our expression for the MSE given by (2.7), we get that as $n \to \infty$

$$\begin{aligned}\text{MSE}(h_n) &= \left(O\left(n^{-\frac{1}{2\beta+1}}\right)\right)^{2\beta-2s} + O\left(n^{-1}\right)\left(O\left(n^{-\frac{1}{2\beta+1}}\right)\right)^{-2s-1} \\ &= O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right) + O\left(n^{\frac{2s+1}{2\beta+1}-\frac{2\beta+1}{2\beta+1}}\right) \\ &= O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right).\end{aligned}$$

Therefore we have shown that if $h = h_n$ is selected such that $h = O\left(n^{-\frac{1}{2\beta+1}}\right)$, then $\text{MSE}(h_n) = O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right)$ as $n \to \infty$.

2.3 Part 3

Problem: Let $\{\phi_m(x)\}_{m=0}^{\infty}$ be the orthonormal Legendre basis on $[-1, 1]$, defined as

$$\phi_m(x) = \begin{cases} \frac{1}{\sqrt{2}}, & m = 0, \\ \sqrt{\frac{2m+1}{2}}\frac{1}{2^m m!}\frac{d^m}{dx^m}\left[(x^2-1)^m\right] & m = 1, 2, \ldots, \end{cases}$$

for $x \in [-1, 1]$. Show that the kernel

$$K(u) = \sum_{m=0}^{\ell} \phi_m^{(s)}(0)\phi_m(u) I\left(|u| \leq 1\right) \tag{2.8}$$

satisfies (2.3) and (2.4).

Solution: Note that u^j can be represented by a linear combination of the first $(j+1)$ Legendre polynomials. Therefore

$$\begin{aligned}\int_{-1}^{1} u^j K(u)du &= \int_{-1}^{1} \sum_{q=0}^{j} b_{qj}\phi_q(u) \sum_{m=0}^{\ell} \phi_m^{(s)}(0)\phi_m(u) I\left(|u| \leq 1\right) du \\ &= \int_{-1}^{1} \sum_{q=0}^{j} b_{qj}\phi_q(u) \sum_{m=0}^{\ell} \phi_m^{(s)}(0)\phi_m(u) du.\end{aligned}$$

By the property that Legendre polynomials form an orthornormal basis in $L_2[-1,1]$, we have

$$\int_{-1}^{1} \sum_{q=0}^{j} b_{qj}\phi_q(u) \sum_{m=0}^{\ell} \phi_m^{(s)}(0)\phi_m(u)du$$
$$= \sum_{q=0}^{j} b_{qj}\phi_q^{(s)}(0) = \frac{d^s}{du^s}\left(\sum_{q=0}^{j} b_{qj}\phi_q(u)\right)\Big|_{u=0} = \frac{d^s}{du^s}\left(u^j\right)\Big|_{u=0} = (s!)\delta_{js},$$

where δ_{js} is the Kronecker delta function. We have therefore shown that

$$\int_{-1}^{1} u^j K(u)du = \begin{cases} s!, & j = s, \\ 0, & j \neq s. \end{cases}$$

3 Example of Kernel Density Estimator

Consider a collection of independent identically distributed random variables $X_1, \ldots, X_{1000}$ belonging to the truncated standard normal distribution on the interval $[-1,1]$, with density $p \in L_2[-1,1]$. Suppose we wish to estimate the second derivative of the density $p(x)$ using (2.2). Let us show that there exists $\beta > 0, L > 0$ such that $p \in P(\beta, L)$, where $P(\beta, L)$ is defined by (2.1), and construct an estimator $\hat{p}_{1000,2}(x)$ of $p''(x)$. First, note that $p(x)$ coincides with the Gaussian standard normal density on $[-1,1]$, which is an infinitely continuously differentiable function with compact support $[-1,1]$. It follows that there exists some $L > 0$ such that $|p^{(\ell)}(b) - p^{(\ell)}(a)| \leq L|b-a|$ for any $a \in [-1,1], b \in [-1,1], \ell \in \mathbf{N}$. Now, note that

$$L|b-a| = L|b-a|^{\beta-\ell+\ell-\beta+1} \leq L \sup_{(a,b)\in[-1,1]^2} (|b-a|)^{\ell-\beta+1}|b-a|^{\beta-\ell}$$
$$\leq 2L|b-a|^{\beta-\ell}.$$

Therefore $|p^{(\ell)}(b) - p^{(\ell)}(a)| \leq L|b-a|^{\beta-\ell}$ for any $\beta > 0, b \in [-1,1], a \in [-1,1]$, and $\ell = \lfloor \beta \rfloor$. We conclude that $p \in P(\beta, L)$ for any $L > 0, \beta > 0$.

Note that the true second derivative of $p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ is $p''(x) = \frac{d}{dx}p'(x) = \frac{d}{dx}(-xp(x)) = (x^2-1)p(x)$. Let us construct an estimator of

$p''(x)$ on $[-1, 1]$ and compare it to the true second derivative of the density. We shall construct the estimator $\hat{p}_{n,s}(x)$ using a kernel function $K(x)$ such that (2.3) and (2.4) are satisfied with $\ell = 3$, namely, the kernel defined by (2.8). Comparing a variety of bandwidths, the optimal bandwidth is found to be $h_{min} = 1.6$ using the method of leave-one-out cross-validation. Figure 4 shows a plot which illustrates the quality of the estimator $\hat{p}_{1000,2}(x)$ of $p''(x)$ as a function of x on $[-1, 1]$.

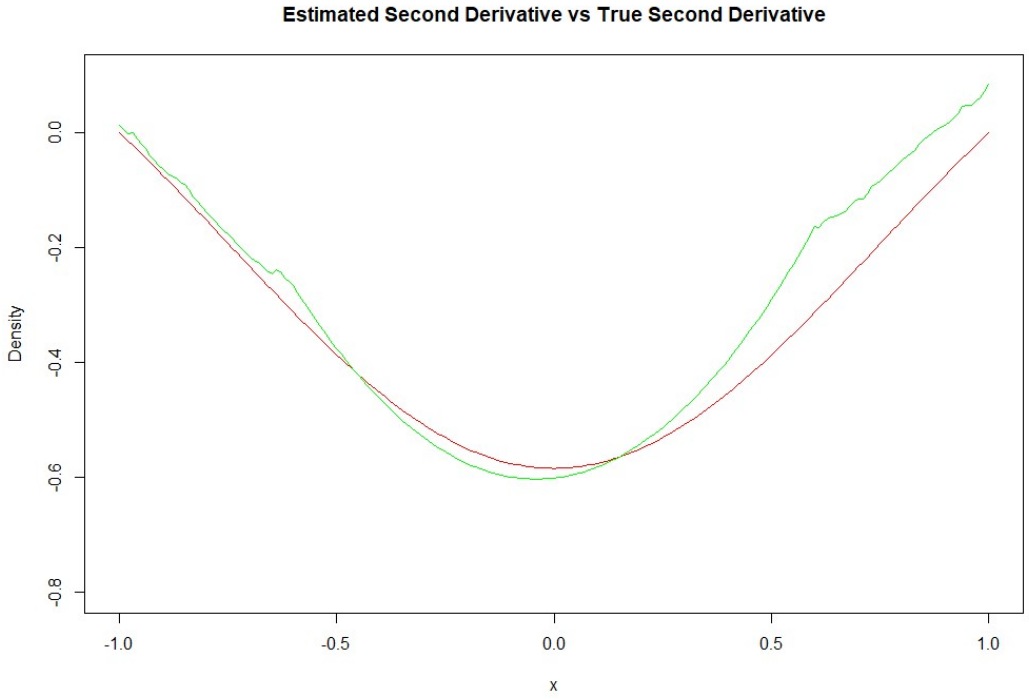


Figure 4: *Estimated second derivative of density p. Red represents the true second derivative of the density and green represents the estimated second derivative of the density.*

Observing Figure 4, we can see that the estimate of the second derivative $p(x)$ over $[-1, 1]$ is reasonably close to the true $p''(x)$ for all values of x. Note however, that our estimate is slightly asymmetrical. Between the values

$x = 0.5$ and $x = 1$, there appears to be a systematic bias in the estimated second derivative of $p(x)$, namely, the estimates are too high.

4 Exercise 1.4 (Tsybakov Page 73)

Suppose we observe the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ where the design points are fixed with $X_i = i/n, i = 1, \ldots, n$. Assume further that

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \ldots, n,$$

where $\xi_1, \ldots, \xi_n$ are independent random variables, with $\mathrm{E}[\xi_i] = 0, \mathrm{Var}(\xi_i) < \infty, i = 1, \ldots, n$. Define the $LP(\ell)$ estimator of the derivative $f^{(s)}(x), s = 1, \ldots, \ell$, by

$$\hat{f}_{ns}(x) = \left(U^{(s)}(0)\right)^\top \hat{\theta}_n(x) h^{-s}, \tag{4.1}$$

where $U^{(s)}(u)$ is the vector whose coordinates are the sth derivative of the corresponding coordinates of $U(u)$ as in (1.14). Assume that assumptions (a)-(c) of Section 1.2.1 are valid.

4.1 Part 1

Problem: Prove that if B_{nx} given by (1.20) is positive definite, then the estimator $\hat{f}_{ns}(x)$ is linear and it reproduces a polynomials of degree $\leq \ell - s$.

Solution: Proving $\hat{f}_{ns}(x)$ is a linear estimator reduces to showing that $\hat{f}_{ns}(x)$ can be written as $\hat{f}_{ns}(x) = \sum_{i=1}^n Y_i W_{is}(x)$, where $W_{is}(x)$ is some weight function depending on x. Since B_{nx} is positive definite, it is also invertible, implying (1.25) is valid. We can therefore substitute (1.19) into (1.25), giving us

$$\begin{aligned}
\hat{f}_{ns}(x) &= \left(U^{(s)}(0)\right)^\top \hat{\theta}_n(x) h^{-s} = \left(U^{(s)}(0)\right)^\top B_{nx}^{-1} a_{nx} h^{-s} \\
&= \frac{1}{nh^{s+1}} \left(U^{(s)}(0)\right)^\top B_{nx}^{-1} \sum_{i=1}^n Y_i U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)
\end{aligned}$$

$$= \sum_{i=1}^{n} Y_i W_{is}(x), \tag{4.2}$$

where

$$W_{is}(x) = \frac{1}{nh^{s+1}} \left(U^{(s)}(0)\right)^\top B_{nx}^{-1} U \left(\frac{X_i - x}{h}\right) K \left(\frac{X_i - x}{h}\right). \tag{4.3}$$

Therefore $\hat{f}_{ns}(x)$ is a linear estimator.

We will now show that the estimator $\hat{f}_{ns}(x)$ reproduces a polynomial of degree $\leq \ell - s$. Recall Proposition 1.12 on page 36 from Tsybakov [1], where it was proven that the $LP(\ell)$ estimator $\hat{f}_n(x)$ of $f(x)$ reproduces a polynomial $Q(x)$ of degree at most ℓ, with $Y_i \coloneqq Q(X_i)$. In the proof, it was shown that the $LP(\ell)$ estimator $\hat{\theta}_n(x)$ of θ takes the form $\hat{\theta}_n(x) = \left(Q(x), Q'(x)h, \ldots, Q^{(\ell)}(x)h^\ell\right)^\top$. Thus for $0 \leq s \leq \ell$

$$\begin{aligned}
\hat{f}_{ns}(x) &= \left(U^{(s)}(0)\right)^\top \hat{\theta}_n(x) h^{-s} \\
&= (0, 0, \ldots, 0, 1, 0, \ldots, 0) \left(Q(x), Q'(x)h, \ldots, Q^{(\ell)}(x)h^\ell\right)^\top h^{-s} \\
&= Q^{(s)}(x).
\end{aligned} \tag{4.4}$$

Since $Y_i = Q(X_i)$, by means of (4.2) and (4.4), we can write $\hat{f}_{ns}(x)$ as

$$\hat{f}_{ns}(x) = \sum_{i=1}^{n} Q(X_i) W_{is}(x) = Q^{(s)}(x). \tag{4.5}$$

Since $Q(x)$ is a polynomial of degree $\leq \ell$, we have that the *sth* derivative $Q^{(s)}(x)$ of $Q(x)$ is a polynomial of degree $\leq \ell - s$. Therefore the $LP(\ell)$ estimator $\hat{f}_{ns}(x)$ of $f^{(s)}(x)$ reproduces polynomials of degree $\leq \ell - s$.

4.2 Part 2

Recall from the Introduction (see Section 1.2.1) that the following assumptions are made on the model:

(a) there exists a real number $\lambda_0 > 0$ such that the smallest eigenvalue $\lambda_{min}(B_{nx})$ of B_{nx} satisfies $\lambda_{min}(B_{nx}) \geq \lambda_0$;

(b) there exists a real number $a_0 > 0$ such that for any interval $A \subseteq [0,1]$ and all $n \geq 1$, $\frac{1}{n}\sum_{i=1}^{n} I\left(X_i \in A\right) \leq a_0 \max\left(\mu(A), \frac{1}{n}\right)$ where $\mu(A)$ denotes the Lebesgue measure of A;

(c) the kernel K has compact support belonging to $[-1,1]$ and there exists a number $K_{max} < \infty$ such that $|K(u)| \leq K_{max}, \forall u \in [-1,1]$.

Problem: Under assumptions (a)-(c), prove that the maximum MSE of $\hat{f}_{ns}(x)$ over the Hölder class $\Sigma(\beta, L)$ is of order $O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right)$ as $n \to \infty$ if the bandwidth $h = h_n$ is chosen optimally.

Solution: We shall choose a bandwidth $h = h_n$ in such a way that $h \geq 1/2n$. Prior to deriving the bias and variance of $\hat{f}_{ns}(x)$, we prove the following properties for the weight functions $W_{is}(x), i = 1, \ldots, n$, given by (4.3):

(A') $\sup_{i,x} |W_{is}(x)| \leq \frac{C}{nh^{s+1}}$, where $C > 0$ is a constant depending on λ, a_0, and K_{max}.

(A'') $\sum_{i=1}^{n} |W_{is}(x)| \leq \frac{C'}{h^s}$, where $C' > 0$ is some constant depending on n, λ, and K_{max}.

(A''') $W_{is}(x) = 0$, for $|X_i - x| > h$.

Proof of (A')*:* Note that by relation (4.3), we have that

$$|W_{is}(x)| \leq \frac{1}{nh^{s+1}} \left|\left|(U^{(s)}(0))^{\top} B_{nx}^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)\right|\right|.$$

Note that the Euclidean norm of $(U^{(s)}(0))$ is 1. Observe that if assumption (a) holds, then applying Results 1 and 2 from the Appendix, we have that for any $v \in \mathbf{R}^{\ell+1}$, $||B_{nx}^{-1} v|| \leq ||v||/\lambda_0$. Hence

$$|W_{is}(x)| \leq \frac{1}{nh^{s+1}\lambda_0} \left|\left|U\left(\frac{X_i - x}{h}\right)\right|\right| \left|K\left(\frac{X_i - x}{h}\right)\right|.$$

By assumption (c), we can write

$$|W_{is}(x)| \leq \frac{K_{max}}{nh^{s+1}\lambda_0} \left\| U\left(\frac{X_i - x}{h}\right) \right\| I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right).$$

Since K is nonzero when $\left|\frac{X_i - x}{h}\right| \leq 1$, we have by definition of $U(u)$ from (1.14) that

$$\begin{aligned}
|W_{is}(x)| &\leq \frac{K_{max}}{nh^{s+1}\lambda_0} \sqrt{1 + \frac{1}{(1!)^2} + \ldots + \frac{1}{(\ell!)^2}} \\
&\leq \frac{K_{max}}{nh^{s+1}\lambda_0} \sqrt{1 + \frac{1}{1!} + \ldots + \frac{1}{\ell!}} \\
&\leq \frac{K_{max}}{nh^{s+1}\lambda_0} \sqrt{e} \\
&\leq \frac{2K_{max}}{nh^{s+1}\lambda_0} =: \frac{C}{nh^{s+1}}.
\end{aligned}$$

We have thus obtained the desired result.

Proof of (A'')*:* We have already shown that

$$|W_{is}(x)| \leq \frac{1}{nh^{s+1}\lambda_0} \left\| U\left(\frac{X_i - x}{h}\right) \right\| \left| K\left(\frac{X_i - x}{h}\right) \right|,$$

therefore

$$\sum_{i=1}^{n} |W_{is}(x)| \leq \sum_{i=1}^{n} \frac{1}{nh^{s+1}\lambda_0} \left\| U\left(\frac{X_i - x}{h}\right) \right\| \left| K\left(\frac{X_i - x}{h}\right) \right|.$$

By assumption (c), we can write

$$\begin{aligned}
\sum_{i=1}^{n} |W_{is}(x)| &\leq \frac{K_{max}}{nh^{s+1}\lambda_0} \sum_{i=1}^{n} \left\| U\left(\frac{X_i - x}{h}\right) \right\| I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right) \\
&\leq \frac{K_{max}}{nh^{s+1}\lambda_0} \sqrt{1 + \frac{1}{(1!)^2} + \ldots + \frac{1}{(\ell!)^2}} \sum_{i=1}^{n} I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right) \\
&\leq \frac{2K_{max}}{nh^{s+1}\lambda_0} \sum_{i=1}^{n} I\left(x - h \leq X_i \leq x + h\right).
\end{aligned}$$

Now, applying assumption (b), we have

$$\sum_{i=1}^{n} |W_{is}(x)| \leq \frac{2K_{max}a_0}{h^{s+1}\lambda_0} \max\left(\mu\left(x-h, x+h\right)\right), \frac{1}{n}\Big) = \frac{2K_{max}a_0}{h^{s+1}\lambda_0} \max\left(2h, \frac{1}{n}\right).$$

Since we take $h \geq 1/2n$, we have that $\max\left(2h, 1/n\right) = 2h$, hence

$$\sum_{i=1}^{n} |W_{is}(x)| \leq \frac{4K_{max}a_0}{h^s\lambda_0} =: \frac{C'}{h^s}.$$

Proof of (A''')*:* Recall that

$$|W_{is}(x)| \leq \frac{K_{max}}{nh^{s+1}\lambda_0} \left|\left|U\left(\frac{X_i - x}{h}\right)\right|\right| I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right).$$

If $|X_i - x| > h$, then $I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right) = 0$, hence we would have $W_{is}(x) = 0$, thus $W_{is}(x) = 0$ for $|X_i - x| > h$.

We are now ready to derive the bias and variance of $\hat{f}_{ns}(x)$ for any $x \in \mathbf{R}$. The bias of $\hat{f}_{ns}(x)$ equals

$$\begin{aligned} Bias(\hat{f}_{ns}(x)) \coloneqq \mathrm{E}[\hat{f}_{ns}(x)] - f^{(s)}(x) &= \mathrm{E}\left[\sum_{i=1}^{n} Y_i W_{is}(x)\right] - f^{(s)}(x) \\ &= \sum_{i=1}^{n} f(X_i) W_{is}(x) - f^{(s)}(x). \end{aligned} \tag{4.6}$$

Taking a Taylor series expansion of $f(X_i)$, we have

$$\begin{aligned} f(X_i) = f(x) + f'(x)(X_i - x) + \ldots + f^{(s)}(x)(X_i - x)^s + \\ \ldots + \frac{f^{(\ell)}(x + \tau(X_i - x))}{\ell!}(X_i - x)^\ell, \quad 0 \leq \tau \leq 1. \end{aligned}$$

We can therefore rewrite (4.6) as

$$Bias(\hat{f}_{ns}(x)) = \sum_{i=1}^{n} \left(f(x) + f'(x)(X_i - x) + \ldots + \frac{f^{(s)}(x)}{s!}(X_i - x)^s + \right.$$
$$\left. \ldots + \frac{f^{(\ell)}(x + \tau(X_i - x))}{\ell!}(X_i - x)^{\ell} \right) W_{is}(x) - f^{(s)}(x). \quad (4.7)$$

Recall relation (4.5) and note that

$$\sum_{i=1}^{n} \left(Q(x) + Q'(x)(X_i - x) + \ldots + \frac{Q^{(\ell)}(x)}{\ell!}(X_i - x)^{\ell} \right) W_{is}(x) = Q^{(s)}(x), \quad (4.8)$$

where ℓ is the order of the local polynomial estimator, and s is the *sth* derivative of $Q(x)$. Inspecting the Taylor coefficients in the above relation, we must have that

$$\sum_{i=1}^{n} (X_i - x)^k W_{is}(x) = s!\delta_{ks}, \quad k = 0, 1, \ldots, \ell, \quad (4.9)$$

where δ_{ks} is the Kronecker delta function.

We shall use relation (4.9) to derive an upper bound on the bias of $\hat{f}_{ns}(x)$. There are two cases for simplifying (4.7) depending on the value of s.

Case 1: $1 \leq s < \ell$. By relation (4.9),

$$\sum_{i=1}^{n} f^{(k)}(x)(X_i - x)^k W_{is}(x)$$
$$= f^{(k)}(x) \sum_{i=1}^{n} (X_i - x)^k W_{is}(x) = f^{(k)}(x) s! \delta_{ks}, \quad k = 0, \ldots, \ell.$$

Relation (4.7) therefore reduces to

$$Bias(\hat{f}_{ns}(x)) = \sum_{i=1}^{n} \frac{f^{(\ell)}(x + \tau(X_i - x))}{\ell!}(X_i - x)^{\ell} W_{is}(x).$$

Since (4.9) implies $f^{(\ell)}(x)\sum_{i=1}^n \frac{(X_i-x)^\ell}{\ell!}W_{is}(x) = 0$, we can subtract the term $\sum_{i=1}^n f^{(\ell)}(x)\frac{(X_i-x)^\ell}{\ell!}W_{is}(x)$ from the bias, giving us

$$Bias(\hat{f}_{ns}(x)) = \sum_{i=1}^n \left(\frac{f^{(\ell)}(x+\tau(X_i-x)) - f^{(\ell)}(x)}{\ell!}\right)(X_i-x)^\ell W_{is}(x).$$

Case 2: $s = \ell$. Again using the property that

$$\sum_{i=1}^n f^{(k)}(x)(X_i-x)^k W_{is}(x)$$
$$= f^{(k)}(x)\sum_{i=1}^n (X_i-x)^k W_{is}(x) = f^{(k)}(x)s!\delta_{ks}, \quad k = 0,\ldots,\ell,$$

we have

$$Bias(\hat{f}_{ns}(x)) = \sum_{i=1}^n \left(\frac{f^{(\ell)}(x+\tau(X_i-x))}{\ell!}\right)(X_i-x)^\ell W_{is}(x) - f^{(\ell)}(x).$$

Since (4.9) implies $f^{(\ell)}(x) = f^{(\ell)}(x)\sum_{i=1}^n \frac{(X_i-x)^\ell W_{is}(x)}{\ell!}$, we can substitute $f^{(\ell)}(x)\sum_{i=1}^n \frac{(X_i-x)^\ell W_{is}(x)}{\ell!}$ into the above expression, giving us

$$Bias(\hat{f}_{ns}(x)) = \sum_{i=1}^n \left(\frac{f^{(\ell)}(x+\tau(X_i-x)) - f^{(\ell)}(x)}{\ell!}\right)(X_i-x)^\ell W_{is}(x).$$

Thus, regardless of s, we have that

$$Bias(\hat{f}_{ns}(x)) = \sum_{i=1}^n \left(\frac{f^{(\ell)}(x+\tau(X_i-x)) - f^{(\ell)}(x)}{\ell!}\right)(X_i-x)^\ell W_{is}(x).$$

By the assumption that $f \in \Sigma(\beta, L)$, we then have

$$|Bias(\hat{f}_{ns}(x))| \leq \sum_{i=1}^n \frac{L|X_i-x|^\beta}{\ell!}|W_{is}(x)|.$$

By (A'''), for nonzero $W_{is}(x)$, we have that $|X_i-x|^\beta \leq h^\beta$, hence

$$|Bias(\hat{f}_{ns}(x))| \leq \frac{Lh^\beta \sum_{i=1}^n |W_{is}(x)|}{\ell!}.$$

Using (A''), we get from the above inequality

$$|Bias(\hat{f}_{ns}(x))| \leq \frac{LC'}{\ell!} h^{\beta-s} =: q h^{\beta-s}, \tag{4.10}$$

where $q > 0$ is a constant depending on C', ℓ, and L.

We now seek to bound the variance. We have

$$\begin{aligned}
\mathrm{Var}(\hat{f}_{ns}(x)) &= \mathrm{Var}\left(\left(\sum_{i=1}^{n} Y_i W_{is}(x)\right)\right) \\
&= \mathrm{Var}\left(\left(\sum_{i=1}^{n} \left(f(X_i) + \xi_i\right) W_{is}(x)\right)\right) \\
&= \mathrm{Var}\left(\left(\sum_{i=1}^{n} \xi_i W_{is}(x)\right)\right) \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n} \xi_i W_{ni}(x)\right)^2\right] - \left(\mathrm{E}\left[\sum_{i=1}^{n} \xi_i W_{ni}(x)\right]\right)^2.
\end{aligned}$$

Since $\mathrm{E}[\xi_i] = 0, i = 1, \ldots, n$, the above relation reduces to

$$\mathrm{Var}(\hat{f}_{ns}(x)) = \mathrm{E}\left[\left(\sum_{i=1}^{n} \xi_i W_{ni}(x)\right)^2\right]. \tag{4.11}$$

Since $\xi_i, i = 1, \ldots, n$, are zero mean independent, we have

$$\mathrm{E}\left[\left(\sum_{i=1}^{n} \xi_i W_{ni}(x)\right)^2\right] = \sum_{i=1}^{n} (W_{is}(x))^2 \mathrm{E}[\xi_i^2]. \tag{4.12}$$

Defining now $\sigma_{max}^2 \coloneqq \max\left(\mathrm{E}[\xi_1^2], \ldots, \mathrm{E}[\xi_n^2]\right)$, we get

$$\sum_{i=1}^{n} (W_{is}(x))^2 \mathrm{E}[\xi_i^2] \leq \sigma_{max}^2 \sup_{i,x} |W_{is}(x)| \sum_{i=1}^{n} |W_{is}(x)|.$$

From this, using (4.11) and (4.12), and taking into account (A') and (A''), we have

$$\mathrm{Var}(\hat{f}_{ns}(x)) \leq \frac{\sigma_{max}^2 CC'}{n h^{2s+1}} =: \frac{C^*}{n h^{2s+1}}. \tag{4.13}$$

Where $C^* = CC'$. Recall that $\text{MSE}(\hat{f}_{n,s}(x)) = Bias^2(\hat{f}_{n,s}(x)) + \text{Var}(\hat{f}_{n,s}(x))$. Substituting equalities (4.10) and (4.13) into $\text{MSE}(\hat{f}_{ns}(x))$ gives us

$$\begin{aligned}\text{MSE}(\hat{f}_{ns}(x)) &= Bias^2(\hat{f}_{ns}(x)) + \text{Var}(\hat{f}_{ns}(x)) \\ &\leq q^2 h^{2(\beta-s)} + \frac{C^*}{nh^{2s+1}} =: g(h).\end{aligned}$$

Our strategy for finding the optimal $h = h_n$ will involve showing that $g(h)$ is convex, and seeking the root of $g'(h)$:

$$\begin{aligned}g'(h) &= 2(\beta-s)q^2 h^{2(\beta-s)-1} - (2s+1)\frac{C^*}{nh^{2s+2}}, \\ g''(h) &= 2(\beta-s)(2(\beta-s)-1)q^2 h^{2(\beta-s)-2} \\ &\quad + (2s+1)(2s+2)\frac{C^*}{nh^{2s+3}} > 0, \forall h > 0.\end{aligned}$$

Therefore the root of $g'(h)$ will correspond to a global minimizer of $g(h)$. Setting $g'(h) = 0$ and continuing, we get

$$\begin{aligned}g'(h) = 0 &= 2(\beta-s)q^2 h^{2(\beta-s)-1} - (2s+1)\frac{C^*}{nh^{2s+2}}, \\ (2s+1)\frac{C^*}{nh^{2s+2}} &= 2(\beta-s)q^2 h^{2(\beta-s)-1}, \\ h^{2\beta+1} &= O\left(n^{-1}\right), \\ h_n &= O\left(n^{-\frac{1}{2\beta+1}}\right).\end{aligned}$$

Plugging h_n into the above expression for the MSE, we get

$$\begin{aligned}\text{MSE}(\hat{f}_{ns}(x)) &= O\left(n^{-\frac{2(\beta-s)}{2\beta+2}}\right) + \frac{O(1)}{n\left(n^{-\frac{1}{2\beta+1}}\right)^{2s+1}} \\ &= O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right) + \frac{O(1)}{n^{\frac{2\beta+1}{2\beta+1} - \frac{2s+1}{2\beta+1}}} \\ &= O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right) + O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right) \\ &= O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right).\end{aligned}$$

Therefore we have shown that, for an appropriately chosen bandwidth, namely $h_n = O\left(n^{-\frac{1}{2\beta+1}}\right)$, the maximum risk of $\hat{f}_{ns}(x)$ satisfies

$$\sup_{f\in\Sigma(\beta,L)} \sup_{x\in[-1,1]} \mathrm{MSE}(\hat{f}_{ns}(x)) = O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right), \quad \text{as } n\to\infty.$$

5 Exercise 1.5 (Tsybakov Page 73)

Define the Fourier transform of a kernel $K \in L_1(\mathbf{R})$ as follows:

$$\hat{K}(\omega) = \int_{-\infty}^{\infty} e^{it\omega} K(t)dt, \quad \omega \in \mathbf{R}.$$

Suppose that $p \in L_2(\mathbf{R})$ is a probability density, and that $K \in L_2(\mathbf{R})$ is symmetric. In Theorem 1.4 of Tsybakov [1], it was shown that under these assumptions, we may express the mean integrated square error (MISE) of a kernel density estimator $\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ with bandwidth $h > 0$ as follows:

$$\mathrm{MISE} = \frac{1}{2\pi}\left[\int_{-\infty}^{\infty} \left|1 - \hat{K}(ht)\right|^2 |\phi(t)|^2 \, dt + \frac{1}{n}\int_{-\infty}^{\infty} \left|\hat{K}(ht)\right|^2 dt\right]$$

$$- \frac{1}{2\pi n}\int_{-\infty}^{\infty} |\phi(t)|^2 \left|\hat{K}(ht)\right|^2 dt =: J_n(K, h, \omega),$$

where $n \geq 1$, and $\phi(t)$ is the characteristic function of the density p.

In Tsybakov [1], the following definition is given for an *inadmissible* kernel. A symmetric kernel $K \in L_2(\mathbf{R})$ is said to be *inadmissible* if there exists some other kernel $K_0 \in L_2(\mathbf{R})$ such that the following two conditions hold:

For all characteristic functions $\phi \in L_2(\mathbf{R})$,

$$J_n(K_0, h, \phi) \leq J_n(K, h, \phi), \quad \forall h > 0, n \geq 1. \tag{5.1}$$

There exists a characteristic function $\phi_0 \in L_2(\mathbf{R})$ such that

$$J_n(K_0, h, \phi_0) < J_n(K, h, \phi_0), \quad \forall h > 0, n \geq 1. \tag{5.2}$$

Furthermore, referring to Proposition 1.8 from Tsybakov [1], if

$$\mu\left(\{t : \hat{K}(t) \notin [0,1]\}\right) > 0,$$

where μ denotes the Lebesgue measure on $\mathbf{R}$, then the kernel K is inadmissible.

5.1 Part 1(Page 73)

Problem: Show that the following two kernels are inadmissible:

$$K(u) = \frac{1}{2} I\left(|u| \leq 1\right), \qquad \text{(the rectangular kernel)}$$
$$K(u) = \frac{15}{16}(1-u^2)^2 I\left(|u| \leq 1\right). \qquad \text{(the biweight kernel)}$$

Solution: We shall show these two kernels are inadmissible by taking their respective Fourier transforms, and showing that there exists a set $A \subseteq \{t : \hat{K}(t) \notin [0,1]\}$ such that $\mu(A) > 0$, where μ is a real-valued set function that assigns Lebesque measure.

For the rectangular kernel, we have that

$$\begin{aligned}\hat{K}(t) &= \int_{-\infty}^{\infty} \frac{e^{itu} I(|u| \leq 1)}{2} du = \int_{-1}^{1} \frac{e^{itu}}{2} du \\ &= \frac{1}{2}\int_{-1}^{1} \left(\cos(tu) + i\sin(tu)\right) du = \frac{\sin(t)}{t},\end{aligned}$$

where, by continuity, $\frac{\sin(t)}{t} = 1$ when $t = 0$. Consider the following graphical plot of the Fourier transform of the rectangular kernel, given by Figure 5 on the following page. Viewing the below plot, we can see that the Fourier transform of K is below zero at some points. Let t_0 be some point such that $\hat{K}(t_0) < 0$. By the continuity of $\hat{K}(t)$, there exists a neighbourhood about

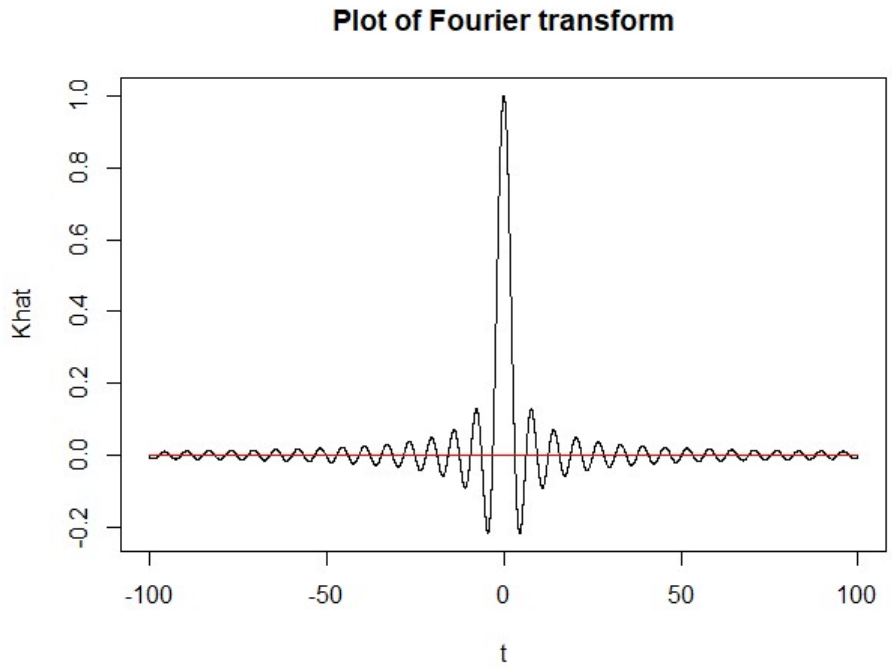


Figure 5: *The Fourier transform applied to the rectangular kernel, for* $t \in [-100, 100]$. *Red represents the* x *axis.*

t_0 such that $\hat{K}(t) < 0$ for all t in the neighbourhood. We conclude that the rectangular kernel is inadmissible.

We shall now show that the biweight kernel is inadmissible. Taking the Fourier transform of K, we have

$$\hat{K}(t) = \frac{15}{16}\int_{-\infty}^{\infty} e^{itu}(1-u^2)^2 I(|u| \leq 1)du = \frac{15}{16}\int_{-1}^{1} e^{itu}(1-2u^2+u^4)du.$$

The function $(1-2u^2+u^4)$ is symmetric about zero, therefore

$$\begin{aligned}\hat{K}(t) &= \int_{-1}^{1}(1-2u^2+u^4)\cos(ut)du \\ &= \int_{-1}^{1}\cos(ut)du - 2\int_{-1}^{1}u^2\cos(ut)du + \int_{-1}^{1}u^4\cos(ut)du.\end{aligned}$$

Our strategy will be to individually evaluate the integrals $\int_{-1}^{1} \cos(ut)du$, $\int_{-1}^{1} u^2 \cos(ut)du$, and $\int_{-1}^{1} u^4 \cos(ut)du$. First,

$$\int_{-1}^{1} \cos(ut)du = \frac{\sin(ut)}{t}\bigg|_{u=-1}^{u=1} = \frac{2\sin(t)}{t}.$$

Next, for $t \neq 0$,

$$\begin{aligned}\int_{-1}^{1} u^2 \cos(tu)du &= \frac{u^2 \sin(ut)}{t}\bigg|_{u=-1}^{u=1} - \frac{2}{t}\int_{-1}^{1} u\sin(tu)du \\ &= \frac{2\sin(t)}{t} + \frac{2u\cos(t)}{t^2}\bigg|_{u=-1}^{u=1} - \frac{2}{t^2}\int_{-1}^{1} \cos(ut)du \\ &= \frac{2\sin(t)}{t} + \frac{4\cos(t)}{t^2} - \frac{4\sin(t)}{t^3}. \end{aligned} \tag{5.3}$$

Lastly,

$$\begin{aligned}\int_{-1}^{1} u^4 \cos(tu)du &= \frac{u^4 \sin(ut)}{t}\bigg|_{u=-1}^{u=1} - \frac{4}{t}\int_{-1}^{1} u^3\sin(tu)du \\ &= \frac{2\sin(t)}{t} + \frac{4u^3\cos(t)}{t^2}\bigg|_{u=-1}^{u=1} - \frac{12}{t^2}\int_{-1}^{1} u^2\cos(ut)du. \end{aligned} \tag{5.4}$$

From (5.3) and (5.4), we obtain for $t \neq 0$,

$$\begin{aligned}\int_{-1}^{1} u^4 \cos(tu)du &= \frac{2\sin(t)}{t} + \frac{8\cos(t)}{t^2} - \frac{12}{t^2}\left(\frac{2\sin(t)}{t} + \frac{4\cos(t)}{t^2} - \frac{4\sin(t)}{t^3}\right) \\ &= \frac{2\sin(t)}{t} + \frac{8\cos(t)}{t^2} - \frac{24\sin(t)}{t^3} - \frac{48\cos(t)}{t^4} + \frac{48\sin(t)}{t^5}. \end{aligned}$$

Putting this together, we obtain for $t \neq 0$,

$$\begin{aligned}\hat{K}(t) &= \frac{15}{16}\left(\int_{-1}^{1}\cos(ut)dt - 2\int_{-1}^{1}u^2\cos(ut)dt + \int_{-1}^{1}u^4\cos(ut)dt\right) \\ &= \frac{15}{16}\left(\frac{2\sin(t)}{t} - 2\left(\frac{2\sin(t)}{t} + \frac{4\cos(t)}{t^2} - \frac{4\sin(t)}{t^3}\right)\right. \\ &\quad \left. + \left(\frac{2\sin(t)}{t} + \frac{8\cos(t)}{t^2} - \frac{24\sin(t)}{t^3} - \frac{48\cos(t)}{t^4} + \frac{48\sin(t)}{t^5}\right)\right).\end{aligned}$$

This reduces to

$$\hat{K}(t) = \begin{cases} -\dfrac{15\sin(t)}{t^3} + \dfrac{45\sin(t)}{t^5} - \dfrac{45\cos(t)}{t^4}, & t \neq 0. \\ 1, & t = 0. \end{cases}$$

Consider the attached plot of the Fourier transform of the biweight kernel, given by Figure 6. Viewing Figure 6, it is clear that there are points such that $\hat{K}(t) < 0$. By the same argument we applied for the rectangular kernel, we can conclude that the biweight kernel is inadmissible.

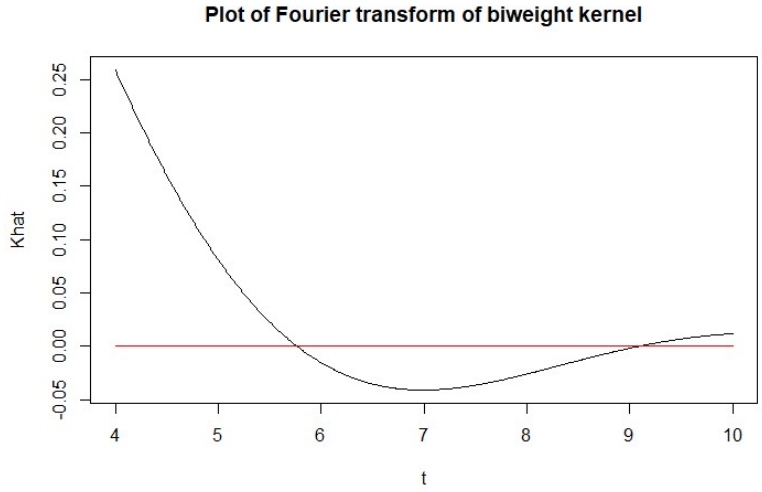


Figure 6: *The Fourier transform applied to the biweight kernel, for* $t \in [4, 10]$. *Red represents the* x *axis.*

6 Exercise 1.6 (Tsybakov Page 74)

Define a class of densities belonging to the Sobolev class as follows:

$$\mathcal{P}_S(\beta, L) = \left\{ p : p \geq 0, \int_{\mathbf{R}} p(x)dx, \int_{\mathbf{R}} |\omega|^{2\beta} |\phi(\omega)|^2 d\omega \leq 2\pi L^2 \right\}, \tag{6.1}$$

where $\beta > 0, L > 0$, and $\phi(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} p(x)dx$ is the characteristic function of p evaluated at $\omega \in \mathbf{R}$. In Theorem 1.5 on page 26 Tsybakov [1], it is proven that for a density $p(x)$ belonging to the Sobolev class as above, the condition that

$$\text{there exists } A, \text{ where } 0 < A < \infty \text{ is such that } \operatorname*{ess\,sup}_{t \in \mathbf{R} \setminus \{0\}} \frac{|1 - \hat{K}(t)|}{|t|^\beta} = A \tag{6.2}$$

implies that for a suitably chosen bandwidth $h = h_n$,

$$\sup_{p \in \mathcal{P}_s(\beta, L)} \mathrm{E} \left[\int_{\mathbf{R}} (\hat{p}_n(x) - p(x))dx \right] \leq C n^{\frac{-2\beta}{2\beta+1}}. \tag{6.3}$$

Note that the above relation is an upper bound on the maximum MISE of $\hat{p}_n(x)$. In Part 1 of this Exercise, we shall show that there is another condition equivalent to (6.2) under mild assumptions. In Part 2 of the Exercise, we shall prove that condition (6.2) is satisfied for a given scenario.

6.1 Part 1

Problem: Let $K \in L_2(\mathbf{R})$ be a symmetric kernel and such that $\hat{K} \in L_\infty(\mathbf{R})$. For $\beta > 0$, show that the following two conditions are equivalent:

(1) There exists A, where $0 < A < \infty$ such that $\operatorname*{ess\,sup}_{t \in \mathbf{R} \setminus \{0\}} \frac{|1 - \hat{K}(t)|}{|t|^\beta} = A$.

(2) There exists $t_0 < \infty, A_0 < \infty$ such that $\operatorname*{ess\,sup}_{0 < |t| \leq t_0} \frac{|1 - \hat{K}(t)|}{|t|^\beta} = A_0$.

Solution: We shall first show that (1) implies (2). Consider the set $A_{t_0} = \{t : 0 < |t| \leq t_0\}$. Clearly, $A_{t_0} \subseteq \{t : t \in \mathbf{R} \setminus \{0\}\}$. Since $A_{t_0} \subseteq \{t : t \in \mathbf{R} \setminus \{0\}\}$, it follows immediately that

$$A_0 = \operatorname*{ess\,sup}_{t \in A_{t_0}} \frac{|1 - \hat{K}(t)|}{|t|^\beta} \leq \operatorname*{ess\,sup}_{t \in \mathbf{R} \setminus \{0\}} \frac{|1 - \hat{K}(t)|}{|t|^\beta} = A.$$

Therefore for any $t_0 > 0$, there exists $A_0 < \infty$ such that $A_0 = \operatorname*{ess\,sup}_{t \in A_{t_0}} \frac{|1-\hat{K}(t)|}{|t|^\beta}$. We shall now show that (2) implies (1). By the assumption that $\hat{K} \in L_\infty(\mathbf{R})$, we have $0 < \hat{K}_{max} \coloneqq \operatorname*{ess\,sup}_{t \in \mathbf{R} \setminus \{\mathbf{0}\}} |\hat{K}(t)| < \infty$. By the triangle inequality, an upper bound on $|1 - \hat{K}(t)|$ will be

$$B \coloneqq 1 + |\hat{K}_{max}| < \infty,$$

therefore

$$\frac{|1 - \hat{K}(t)|}{|t|^\beta} \leq \frac{B}{|t|^\beta}, \quad \text{for any } t \neq 0.$$

For all points, except perhaps on a set of Lebesgue measure zero, we will have that for $|t| > t_0$, $\frac{B}{|t|^\beta} \leq \frac{B}{|t_0|^\beta}$. Taking $C = \max\left(\frac{B}{|t_0|^\beta}, A_0\right)$, it is clear that

$$\operatorname*{ess\,sup}_{t \in \mathbf{R} \setminus \{0\}} \frac{|1 - \hat{K}(t)|}{|t|^\beta} \leq C.$$

Hence, (2) implies (1), in which one can take A equal to C.

6.2 Part 2

For a kernel function $K : \mathbf{R} \to \mathbf{R}$ as above, such a kernel is called a kernel of order $\ell, \ell \in \mathbf{N}$, if the following conditions are satisfied:

$$\int_{-\infty}^{\infty} K(u)du = 1, \tag{6.4}$$

$$\int_{-\infty}^{\infty} u^j K(u)du = 0, \quad j = 1, \ldots, \ell. \tag{6.5}$$

Problem: Show that for integer β, condition (1) is satisfied if K is a kernel of order $\beta - 1$ and $\int_{-\infty}^{\infty} |u|^\beta |K(u)|du < \infty$.

Solution: We have

$$|1 - \hat{K}(t)| = \left|1 - \int_{-\infty}^{\infty} e^{itu} K(u) du\right|.$$

By the assumption that the kernel K is real-valued and symmetric, the Fourier transform must be real-valued, giving us

$$|1 - \hat{K}(t)| = |1 - \int_{-\infty}^{\infty} \cos(tu) K(u) du|.$$

Case 1: Suppose β is odd. Taking a Taylor expansion of $\cos(tu)$ about 0, we get

$$\begin{aligned} |1 - \hat{K}(t)| = \Bigg|1 - \int_{-\infty}^{\infty} \Bigg(1 - \frac{t^2}{2!}u^2 + \frac{t^4}{4!}u^4 + \\ \ldots + \frac{(-1)^{\frac{\beta-1}{2}} t^{\beta-1}}{(\beta-1)!}u^{\beta-1} + \frac{(-1)^{\frac{\beta-1}{2}+1} \sin(\epsilon tu) t^{\beta}}{\beta!}u^{\beta}\Bigg) K(u) du\Bigg|, \end{aligned}$$

where $0 \leq \epsilon \leq 1$. By the assumption that $K(u)$ is a kernel of order $\beta - 1$, we get

$$\begin{aligned} |1 - \hat{K}(t)| &= \left|\int_{-\infty}^{\infty} \frac{(-1)^{\frac{\beta-1}{2}+1} \sin(\epsilon tu) t^{\beta}}{\beta!} u^{\beta} K(u) du\right| \\ \frac{|1 - \hat{K}(t)|}{|t|^{\beta}} &= \left|\int_{-\infty}^{\infty} \sin(\epsilon tu) u^{\beta} K(u) du\right| \\ &\leq \int_{-\infty}^{\infty} |u|^{\beta} |K(u)| du. \end{aligned}$$

Hence, by the assumption that $\int_{-\infty}^{\infty} |u|^{\beta} |K(u)| du < \infty$, we have

$$\frac{|1 - \hat{K}(t)|}{|t|^{\beta}} \leq A, \quad t \in \mathbf{R},$$

where A is some finite positive constant.

Case 2: Suppose β is even. Again taking a Taylor expansion of $\cos(tu)$ about 0, we get

$$|1-\hat{K}(t)| = \left|1 - \int_{-\infty}^{\infty} \left(1 - \frac{t^2}{2!}u^2 + \frac{t^4}{4!}u^4 + \ldots + \frac{(-1)^{\frac{\beta}{2}}\cos(\epsilon tu)t^\beta}{\beta!}u^\beta\right) K(u)du\right|,$$

where $0 \leq \epsilon \leq 1$. By the assumption that K is a kernel of order $\beta - 1$, we have

$$\begin{aligned} |1 - \hat{K}(t)| &= \left|\int_{-\infty}^{\infty} (-1)^{-\frac{\beta}{2}} \cos(\epsilon tu)t^\beta u^\beta K(u)du\right| \\ \frac{|1 - \hat{K}(t)|}{|t|^\beta} &= \left|\int_{-\infty}^{\infty} \cos(\epsilon tu)u^\beta K(u)du\right| \\ &\leq \int_{-\infty}^{\infty} |u|^\beta |K(u)|du. \end{aligned}$$

Again assuming that $\int_{-\infty}^{\infty} |u|^\beta |K(u)|du < \infty$, this gives us

$$\frac{|1 - \hat{K}(t)|}{|t|^\beta} \leq A,$$

where A is the same constant as in *Case 1*. We have thus shown that if K is a kernel of order $\beta - 1$ and $K \in L_2(\mathbf{R}), \hat{K} \in L_\infty(\mathbf{R}), \int_{-\infty}^{\infty} |u|^\beta |K(u)|du < \infty$, then $\operatorname*{ess\,sup}_{t\in\mathbf{R}\setminus\{0\}} \frac{|1-\hat{K(t)}|}{|t|^\beta} = A$.

7 Exercise 1.8 (Tsybakov Page 74)

Consider a random sample $X_1, \ldots, X_n$. Let $\mathcal{P}_a$, where $a > 0$, be the class of all probability densities $p(x)$ on $\mathbf{R}$ such that the support of the characteristic function $\phi(t) = \mathrm{E}[e^{itX}]$ is included in a given interval $t \in [-a, a]$. Let $K(u)$ be the sinc kernel defined as

$$K(u) = \begin{cases} \dfrac{\sin u}{\pi u}, & u \neq 0, \\ \\ \dfrac{1}{\pi}, & u = 0. \end{cases} \tag{7.1}$$

Problem: Show that for any $n \geq 1$ and for any $p \in \mathcal{P}_a$, the estimator $\hat{p}_n(x) = \sum_{i=1}^n \frac{1}{nh} K\left(\frac{X_i - x}{h}\right)$ with appropriately chosen bandwidth $h > 0$ has the following upper bound on its mean integrated squared error (MISE):

$$\begin{aligned}\text{MISE} &= \mathrm{E}\left[\int_{-\infty}^{\infty} \left(\hat{p}_n(x) - p(x)\right)^2 dx\right] \leq \sup_{p \in \mathcal{P}_a} \mathrm{E}\left[\int_{-\infty}^{\infty} \left(\hat{p}_n(x) - p(x)\right)^2 dx\right] \\ &\leq \frac{a}{\pi n}.\end{aligned}$$

Solution: In Theorem 1.4 of Tsybakov [1], it has been shown that the MISE of a kernel density estimator $\hat{p}_n$ given by (1.10) is

$$\begin{aligned}\text{MISE} = \frac{1}{2\pi}&\left[\int_{-\infty}^{\infty} \left|1 - \hat{K}(ht)\right|^2 |\phi(t)|^2 \, dt + \frac{1}{n}\int_{-\infty}^{\infty} \left|\hat{K}(ht)\right|^2 dt\right] \\ &- \frac{1}{2\pi n}\int_{-\infty}^{\infty} |\phi(t)|^2 \left|\hat{K}(ht)\right|^2 dt,\end{aligned}$$

Referring to page 20 of Tsybakov [1], the Fourier transform of the sinc kernel is $\hat{K}(t) = I(|t| \leq 1)$. Since $\hat{K}(t) = I(|t| \leq 1)$, we can express the MISE of an estimator $\hat{p}_n(x)$ with density $p \in \mathcal{P}_a$ as

$$\begin{aligned}\text{MISE} = \frac{1}{2\pi}&\left[\int_{-\infty}^{\infty} |1 - I(|ht| \leq 1)|^2 |\phi(t)|^2 \, dt + \frac{1}{n}\int_{-\infty}^{\infty} |I(|ht| \leq 1)|^2 dt\right] \\ &- \frac{1}{2\pi n}\int_{-\infty}^{\infty} |\phi(t)|^2 |I(|ht| \leq 1)|^2 \, dt \\ \leq &\frac{1}{2\pi}\left[\int_{-\infty}^{\infty} |1 - I(|ht| \leq 1)| |\phi(t)|^2 \, dt + \frac{1}{n}\int_{-\infty}^{\infty} |I(|ht| \leq 1)| \, dt\right].\end{aligned}$$

Since by assumption the support of the characteristic function of all densities $p \in \mathcal{P}_a$ is contained in $[-a, a]$, and $|\phi(t)| \leq 1$, for all $t \in \mathbf{R}$, we have

$$\text{MISE} \leq \frac{1}{2\pi}\left[\int_{-a}^{a} |1 - I(|ht| \leq 1)| \, dt + \frac{1}{n}\int_{-\infty}^{\infty} |I(|ht| \leq 1)| \, dt\right].$$

We shall decompose $[-a, a]$ into two sets. Let $A = \{t : |ht| \leq 1, |t| \leq a\}$ and let $B = \{t : |ht| > 1, |t| \leq a\}$. We then get from above that

$$
\begin{aligned}
\text{MISE} \leq & \frac{1}{2\pi}\int_A |1 - I(|ht| \leq 1)|dt + \frac{1}{2\pi}\int_B |1 - I(|ht| \leq 1)|dt \\
& + \frac{1}{2\pi n}\int_{-\infty}^{\infty} |I(|ht| \leq 1)|\, dt \\
\leq & \frac{1}{2\pi}\int_A \max_{t\in A}\left(|1 - I(|ht| \leq 1)|\right) dt \\
& + \frac{1}{2\pi}\int_B \max_{t\in B}\left(|1 - I(|ht| \leq 1)|\right) dt + \frac{1}{2\pi n}\int_{-\infty}^{\infty} |I(|ht| \leq 1)|\, dt \\
\leq & \frac{1}{2\pi}\int_B dt + \frac{1}{2\pi n}\int_{-\infty}^{\infty} |I(|ht| \leq 1)|dt \\
\leq & \frac{1}{2\pi}\int_B dt + \frac{1}{\pi n}\int_0^{\infty} I(0 \leq ht \leq 1)dt \\
\leq & \frac{1}{2\pi}\mu(B) + \frac{1}{\pi n}\mu(0 \leq ht \leq 1), \qquad (7.2)
\end{aligned}
$$

where μ denotes the Lebesgue measure on $\mathbf{R}$. We now evaluate $\mu(B)$ and $\mu(0 \leq ht \leq 1)$.

$$
\mu(B) = \mu(\{t : |ht| > 1, |t| \leq a\}) = \mu(-a, a) - \mu\left(-\frac{1}{h}, \frac{1}{h}\right) = 2a - \frac{2}{h},
$$

$$
\mu(0 \leq ht \leq 1) = \frac{1}{h}.
$$

Substituting these expressions into (7.2) , we obtain

$$
\begin{aligned}
\text{MISE} &\leq \frac{1}{\pi n h} + \frac{a}{\pi} - \frac{1}{\pi h} \\
&= \frac{1}{h}\left(\frac{1}{\pi n} - \frac{1}{\pi}\right) + \frac{a}{\pi}. \qquad (7.3)
\end{aligned}
$$

We shall select h such that

$$
\frac{1}{h}\left(\frac{1}{\pi n} - \frac{1}{\pi}\right) + \frac{a}{\pi} = \frac{a}{\pi n},
$$

or

$$
\frac{1}{h\pi}\left(\frac{1-n}{n}\right) = \frac{a}{\pi}\left(\frac{1-n}{n}\right).
$$

Rearranging, we get

$$h = \frac{1}{a}.$$

Substituting $h = \dfrac{1}{a}$ into (7.3), we obtain

$$\text{MISE} \leq \frac{a}{\pi n}.$$

We have therefore shown that

$$\sup_{p \in \mathcal{P}_a} \mathrm{E}\left[\int_{-\infty}^{\infty} \left(\hat{p}_n(x) - p(x)\right)^2 dx\right] \leq \frac{a}{\pi n}.$$

8 Exercise 1.9 (Tsybakov Page 74)

Let $X_1, \ldots, X_n$ be a random sample from a density $p \in L_2[0,1]$. Consider the following estimator $\hat{p}_{nN}$ of p:

$$\hat{p}_{nN}(x) = \sum_{j=1}^{N} \hat{c}_j \phi_j(x), \quad x \in [0,1], \tag{8.1}$$

where $N \geq 1, \hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)$, and $\{\phi_j(x)\}_{j=1}^{\infty}$ is an orthonormal basis in $L_2[0,1]$. Suppose that the basis $\{\phi_j(x)\}_{j=1}^{\infty}$ is given by (1.30).

8.1 Part 1

Problem: Show that $\hat{c}_j$ is an unbiased estimator of the Fourier coefficients $c_j = \int_0^1 p(x)\phi_j(x)dx, j = 1, 2, \ldots,$ and find the variance of $\hat{c}_j$.

Solution: First,

$$\begin{aligned}\mathrm{E}[\hat{c}_j] &= \mathrm{E}\left[\sum_{i=1}^{n} \frac{\phi_j(X_i)}{n}\right] = \mathrm{E}\left[\phi_j(X_1)\right] \\ &= \int_0^1 \phi_j(x)p(x)dx = c_j.\end{aligned}$$

Next, using the independence of $X_1, \ldots, X_n$,

$$\begin{aligned}
\mathrm{Var}(\hat{c}_j) = \mathrm{E}\left[\hat{c}_j^2\right] - c_j^2 &= \mathrm{E}\left[\left(\sum_{i=1}^{n} \frac{\phi_j(X_i)}{n}\right)^2\right] - c_j^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \frac{\phi_j^2(X_i)}{n^2} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n} (\phi_j(X_i)\phi_j(X_k))\right] - c_j^2 \\
&= \frac{\mathrm{E}\left[\phi_j^2(X_1)\right]}{n} + \frac{n(n-1)}{n^2}c_j^2 - c_j^2 \\
&= \frac{\mathrm{E}\left[\phi_j^2(X_1)\right]}{n} - \frac{c_j^2}{n}, \qquad (8.2)
\end{aligned}$$

where for the trigonometric basis given by (1.30), the expression for $\mathrm{E}\left[\phi_j^2(X_1)\right]$ is provided below.

Case 1: Suppose $j = 1$, then $\mathrm{E}\left[\phi_1^2(x)\right] = \int_0^1 p(x)dx = 1$.

Case 2: Suppose j is odd and greater than 1. Then
$\phi_j(x) = \sqrt{2}\sin\left(\pi\left(j-1\right)x\right)$ and we have

$$\mathrm{E}\left[\phi_j^2(X_1)\right] = \int_0^1 \phi_j^2(x)p(x)dx = \int_0^1 2\sin^2\left(\pi\left(j-1\right)x\right)p(x)dx.$$

Using the trigonometric identity $\sin^2(x) = \dfrac{1}{2} - \dfrac{\cos(2x)}{2}$, we get

$$\int_0^1 2\sin^2\left(\pi\left(j-1\right)x\right)p(x)dx = \int_0^1 p(x)dx - \int_0^1 \cos\left(\pi\left(2j-2\right)x\right)p(x)dx.$$

Note that since $p \in L_2[0,1]$, $p(x)$ has the representation $p(x) = \sum_{i=1}^{\infty} c_j\phi_j(x)$, therefore

$$\mathrm{E}\left[\phi_j^2(X_1)\right] = 1 - \int_0^1 \cos\left(\pi\left(2j-2\right)x\right)\sum_{j=1}^{\infty} c_j\phi_j(x)dx.$$

Using the fact that the trigonometric basis is orthonormal, and noting that $\cos(\pi(2j-2)x) = \frac{\phi_{2j-2}(x)}{\sqrt{2}}$, we have

$$\begin{aligned}\mathrm{E}\left[\phi_j^2(X_1)\right] &= 1 - \int_0^1 \frac{\phi_{2j-2}(x)}{\sqrt{2}} \sum_{j=1}^{\infty} c_j \phi_j(x) dx \\ &= 1 - \frac{c_{2j-2}}{\sqrt{2}}.\end{aligned}$$

Case 3: Suppose j is even. Then $\phi_j(x) = \sqrt{2}\cos(\pi j x)$, implying

$$\mathrm{E}\left[\phi_j^2(X_1)\right] = \int_0^1 \phi_j^2(x) p(x) dx = \int_0^1 2\cos^2(\pi j x)\, p(x) dx.$$

Using the identity $\cos^2(x) = \dfrac{1}{2} + \dfrac{\cos(2x)}{2}$, we have

$$\begin{aligned}\int_0^1 2\cos^2(\pi j x)\, p(x) dx &= \int_0^1 p(x) dx + \int_0^1 \cos(\pi(2j)x)\, p(x) dx \\ &= 1 + \int_0^1 \cos(\pi(2j)x) \sum_{j=1}^{\infty} c_j \phi_j(x) dx.\end{aligned}$$

Noting that $\phi_{2j}(x) = \sqrt{2}\cos(\pi(2j)x)$, we have that

$$\begin{aligned}2\int_0^1 \cos^2(\pi j x) p(x) dx &= 1 + \int_0^1 \cos(\pi(2j)x) \sum_{j=1}^{\infty} c_j \phi_j(x) dx \\ &= 1 + \int_0^1 \frac{\phi_{2j}(x)}{\sqrt{2}} \sum_{j=1}^{\infty} c_j \phi_j(x) dx.\end{aligned}$$

Recalling that the trigonometric basis is orthonormal gives us

$$\mathrm{E}\left[\phi_j^2(X_1)\right] = 1 + \int_0^1 \frac{\phi_{2j}(x)}{\sqrt{2}} \sum_{j=1}^{\infty} c_j \phi_j(x) dx = 1 + \frac{c_{2j}}{\sqrt{2}}.$$

Therefore

$$\mathrm{E}\left[\phi_j^2(X_1)\right] = \begin{cases} 1, & j = 1, \\ 1 - \frac{c_{2j-2}}{\sqrt{2}}, & j \text{ is odd}, \\ 1 + \frac{c_{2j}}{\sqrt{2}}, & j \text{ is even}. \end{cases} \tag{8.3}$$

8.2 Part 2

Problem: Express the mean integrated squared error (MISE) of the estimator $\hat{p}_{nN}$ given by (8.1) as a function of p and the trigonometric basis $\{\phi_j(x)\}_{j=1}^{\infty}$. Denote it by MISE($N$).

Solution: Using the definition of $\hat{p}_{nN}$ and the fact that $\mathrm{E}[\hat{c}_j] = c_j$, we can write

$$\begin{aligned}
\mathrm{MISE}(N) &= \int_0^1 \mathrm{E}\left[(\hat{p}_{nN}(x) - p(x))^2\right] dx \\
&= \int_0^1 \mathrm{E}\left[(\hat{p}_{nN}(x) - \mathrm{E}[\hat{p}_{nN}(x)] + \mathrm{E}[\hat{p}_{nN}(x)] - p(x))^2\right] dx \\
&= \int_0^1 \left(\mathrm{E}\left[\hat{p}_{nN}(x)\right] - p(x)\right)^2 dx + \int_0^1 \mathrm{Var}\left(\hat{p}_{nN}(x)\right) dx \\
&= \int_0^1 \left(\sum_{j=1}^{N} c_j\phi_j(x) - \sum_{j=1}^{\infty} c_j\phi_j(x)\right)^2 dx + \int_0^1 \mathrm{Var}\left(\hat{p}_{nN}(x)\right) dx \\
&= \int_0^1 \left(\sum_{j=N+1}^{\infty} c_j\phi_j(x)\right)^2 + \int_0^1 \mathrm{Var}\left(\sum_{j=1}^{N} \hat{c}_j\phi_j(x)\right) dx \\
&= \int_0^1 \left(\sum_{j=N+1}^{\infty} c_j^2\phi_j^2(x) + \sum_{N+1\le i\neq j<\infty}\sum c_i c_j \phi_i(x)\phi_j(x)\right) dx \\
&+ \int_0^1 \sum_{i=1}^{N}\sum_{j=1}^{N} \phi_i(x)\phi_j(x)\mathrm{Cov}\left(\hat{c}_i, \hat{c}_j\right) dx.
\end{aligned}$$

Using the property that the trigonometric basis is orthonormal in $L_2[0,1]$, we get

$$\begin{aligned}
\mathrm{MISE}(N) &= \sum_{j=N+1}^{\infty} c_j^2 + \sum_{j=1}^{N} \mathrm{Cov}\left(\hat{c}_j, \hat{c}_j\right) \\
&= \sum_{j=N+1}^{\infty} c_j^2 + \sum_{j=1}^{N} \mathrm{Var}\left(\hat{c}_j\right). \qquad (8.4)
\end{aligned}$$

From this, noting that $c_j = \int_0^1 \phi_j(x)p(x)dx$, and using formula (8.2), we can write

$$\text{MISE}(N) = \sum_{j=N+1}^{\infty} \left(\int_0^1 p(x)\phi_j(x)dx \right)^2 + \sum_{j=1}^{N} \left(\frac{\text{E}\left[\phi_j^2(X_1)\right]}{n} - \frac{\left(\int_0^1 p(x)\phi_j(x)dx \right)^2}{n} \right).$$

8.3 Part 3

Problem: Derive an unbiased risk estimation method. Show that

$$\text{E}\left(\hat{J}(N)\right) = \text{MISE}(N) - \int_0^1 p^2(x)dx,$$

where

$$\hat{J}(N) = \frac{1}{n-1} \sum_{j=1}^{N} \left[\frac{2}{n} \sum_{i=1}^{n} \phi_j^2(X_i) - (n+1)\hat{c}_j^2 \right]. \tag{8.5}$$

Solution: In view of (8.2), $\text{E}\left[\hat{c}_j^2\right] = \frac{1}{n}\text{E}\left[\phi_j^2(X_1)\right] + \frac{n-1}{n}c_j^2$. Therefore

$$\begin{aligned}
\text{E}\left[\hat{J}(N)\right] &= \text{E}\left[\frac{1}{n-1} \sum_{j=1}^{N} \left(\frac{2}{n} \sum_{i=1}^{n} \phi_j^2(X_i) - (n+1)\hat{c}_j^2 \right) \right] \\
&= \frac{1}{n-1} \sum_{j=1}^{N} \text{E}\left[\frac{2\sum_{i=1}^{n} \phi_j^2(X_i)}{n} - (n+1)\hat{c}_j^2 \right] \\
&= \frac{1}{n-1} \sum_{j=1}^{N} \text{E}\left[2\phi_j^2(X_1) - (n+1)\hat{c}_j^2 \right] \\
&= \frac{1}{n-1} \sum_{j=1}^{N} \left(2\text{E}\left[\phi_j^2(X_1)\right] - (n+1)\left(\frac{1}{n}\text{E}\left[\phi_j^2(X_1)\right] + \frac{n-1}{n}c_j^2 \right) \right) \\
&= \frac{1}{n-1} \sum_{j=1}^{N} \left(2\text{E}\left[\phi_j^2(X_1)\right] - \frac{n+1}{n}\text{E}\left[\phi_j^2(X_1)\right] - \frac{(n+1)(n-1)}{n}c_j^2 \right) \\
&= \frac{1}{n-1} \sum_{j=1}^{N} \left(\frac{n-1}{n}\text{E}\left[\phi_j^2(X_1)\right] - \frac{(n+1)(n-1)}{n}c_j^2 \right) \\
&= \sum_{j=1}^{N} \left(\frac{1}{n}\text{E}\left[\phi_j^2(X_1)\right] - \left(c_j^2 + \frac{c_j^2}{n} \right) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \mathrm{Var}\left[\hat{c}_j\right] - \sum_{j=1}^{N} c_j^2 \\
&= \sum_{j=1}^{N} \mathrm{Var}\left[\hat{c}_j\right] - \sum_{j=1}^{N} c_j^2 - \sum_{j=N+1}^{\infty} c_j^2 + \sum_{j=N+1}^{\infty} c_j^2 \\
&= \mathrm{MISE}(N) - \sum_{j=1}^{\infty} c_j^2,
\end{aligned}
$$

where the last equality is due to relation (8.4). It remains to show that $\sum_{j=1}^{\infty} c_j^2 = \int_0^1 p^2(x)dx$. To this end, we write

$$
\begin{aligned}
p(x) &= \sum_{j=1}^{\infty} c_j \phi_j(x), \\
p^2(x) &= \left(\sum_{j=1}^{\infty} c_j \phi_j(x)\right)^2 = \sum_{j=1}^{\infty} c_j^2 \phi_j^2(x) + \sum_{1 \le i \ne j < \infty} \sum \sum c_i c_j \phi_i(x) \phi_j(x), \\
\int_0^1 p^2(x)dx &= \int_0^1 \sum_{j=1}^{\infty} c_j^2 \phi_j^2(x)dx + \int_0^1 \sum_{1 \le i \ne j < \infty} \sum \sum c_i c_j \phi_i(x) \phi_j(x)dx = \sum_{j=1}^{\infty} c_j^2,
\end{aligned}
$$

where again we have used the property that the trigonometric basis is orthonormal in $L_2[0,1]$. Therefore

$$
\mathrm{E}\left[\hat{J}(N)\right] = \mathrm{MISE}(N) - \int_0^1 p^2(x)dx.
$$

Observe that $\hat{J}(N)$ is an unbiased estimator of the MISE up to a shift of $\int_0^1 p^2(x)dx$. The shift $\int_0^1 p^2(x)dx$ does not depend on N, thus the value of N that minimizes the estimator $\hat{J}(N)$ is a reasonable data-driven choice for determining an optimal N.

8.4 Part 4

Consider the following class of densities:

$$
W^{per}(\beta, L) = \left\{ p(x) : p(x) = \sum_{i=1}^{\infty} c_j \phi_j(x), c_j \in \ell^2(\mathbf{N}), \sum_{j=1}^{\infty} a_j^2 c_j^2 \le Q \right\}, \quad (8.6)
$$

where
$\ell^2(\mathbf{N}) = \{c : \sum_{j=1}^{\infty} c_j^2 < \infty\}$, $Q = \frac{L^2}{\pi^{2\beta}}, L > 0, \beta > 0$, a_j is given by (1.32), and $\{\phi_j\}_{j=1}^{\infty}$ is the trigonometric basis given by (1.30).

Problem: Show that the MISE of $\hat{p}_{nN}$ as in (8.1) is bounded by $\frac{N+1}{n} + \rho_n$, where $\rho_n = \sum_{i=N+1}^{\infty} c_i^2$. Use this bound to prove that, uniformly over the class of densities p belonging to $W^{per}(\beta, L), \beta > 0, L > 0$, the MISE of $\hat{p}_{nN}$ is of order $O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$ for an appropriate choice of $N = N_n$.

Solution: Using (8.2), we have

$$\begin{aligned} \text{MISE}(N) &= \sum_{j=1}^{N} \text{Var}\left(\hat{c}_j\right) + \sum_{j=N+1}^{\infty} c_j^2 \\ &\leq \sum_{j=1}^{N} \frac{\text{E}\left[\phi_j^2(X_i)\right]}{n} + \rho_n. \end{aligned} \tag{8.7}$$

Recalling the expression for $\text{E}[\phi_j^2(X_1)]$ given by (8.3), we can write

$$\frac{\text{E}\left[\phi_j^2(X_i)\right]}{n} + \rho_n = \frac{1}{n} + \frac{\sum_{j_{odd}, j>1}^{N} \left(1 - \frac{c_{2j-2}}{\sqrt{2}}\right) + \sum_{j_{even}, j>1}^{N} \left(1 + \frac{c_{2j}}{\sqrt{2}}\right)}{n} + \rho_n.$$

In the above expression for the bound on MISE(N), the coefficients $-\frac{c_{2j-2}}{\sqrt{2}}$ in the first sum cancel with the coefficients $\frac{c_{2j}}{\sqrt{2}}$ in the second sum, yielding

$$\frac{\text{E}\left[\phi_j^2(X_i)\right]}{n} + \rho_n = \frac{N + \frac{c_{2N} I(\frac{N}{2} \in \mathbf{N})}{\sqrt{2}}}{n} + \rho_n.$$

Note by definition that $c_j = \int_0^1 \phi_j(x) p(x) dx$, therefore

$$c_{2N} = \int_0^1 \sqrt{2} \cos(\pi(2N)) p(x) dx \leq \sqrt{2} \int_0^1 p(x) dx = \sqrt{2}.$$

Implying

$$\frac{c_{2N}}{\sqrt{2}} \leq 1.$$

Therefore

$$\frac{N + \frac{c_{2N} I(\frac{N}{2} \in \mathbf{N})}{\sqrt{2}}}{n} + \rho_n \leq \frac{N+1}{n} + \rho_n = \frac{N+1}{n} + \sum_{j=N+1}^{\infty} c_j^2. \quad (8.8)$$

Now, if $p \in W^{per}(\beta, L)$, then

$$\rho_n = \sum_{j=N+1}^{\infty} c_j^2 = \sum_{j=N+1}^{\infty} \frac{a_j^2 c_j^2}{a_j^2},$$

where the sequence $\{a_j\}_{j=1}^{\infty}$, with a_j given by (1.32), is monotonically increasing. Hence, we can write

$$\rho_n \leq \frac{1}{\min(\{a_j^2\}_{j=N+1}^{\infty})} \sum_{j=1}^{\infty} a_j^2 c_j^2 = \sum_{j=1}^{\infty} \frac{a_j^2 c_j^2}{a_{N+1}^2} \leq \sum_{j=1}^{\infty} \frac{a_j^2 c_j^2}{(N+1-1)^{2\beta}} \leq \frac{Q}{N^{2\beta}}.$$

Thus, using (8.7) and (8.8),

$$\text{MISE}(N) \leq \frac{N+1}{n} + QN^{-2\beta} =: g(N),$$

giving us

$$g'(N) = \frac{1}{n} - 2\beta Q N^{-2\beta-1}$$
$$g''(N) = 2(2\beta+1) Q N^{-2\beta-2} > 0, \quad \forall N > 0.$$

Since $g''(N)$ is positive for any N, the root of $g'(N)$ is a global minimizer of $g(N)$. Letting N_n represent the unique root of $g'(N)$, we get

$$2\beta Q (N_n)^{-2\beta-1} = \frac{1}{n},$$
$$N_n^{2\beta+1} = 2\beta Q n.$$

Rearranging, we have as $n \to \infty$,

$$N_n = O\left(n^{\frac{1}{2\beta+1}}\right).$$

Substituting $N_n = O\left(n^{\frac{1}{2\beta+1}}\right)$ into $g(N)$, we get

$$\begin{aligned} g(N_n) &= \frac{O\left(n^{\frac{1}{2\beta+1}}\right)}{n} + O\left(n^{\frac{1}{2\beta+1}}\right)^{-2\beta} \\ &= O\left(n^{\frac{1}{2\beta+1}-\frac{2\beta+1}{2\beta+1}}\right) + O\left(n^{-\frac{2\beta}{2\beta+1}}\right) \\ &= O\left(n^{-\frac{2\beta}{2\beta+1}}\right). \end{aligned}$$

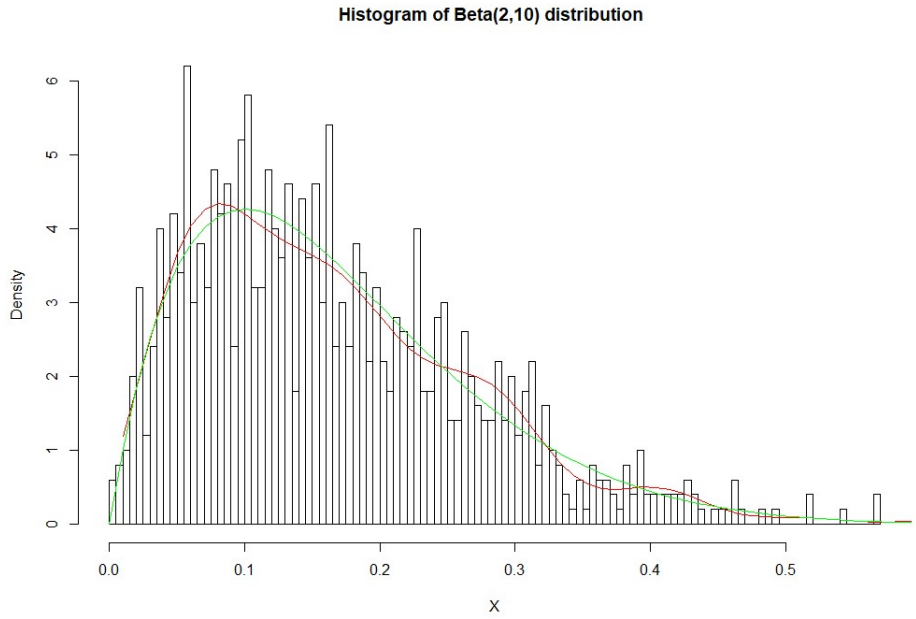
Thus, for any given $\beta > 0, L > 0$, and for a suitably chosen $N = N_n$, we have that as $n \to \infty$,

$$\sup_{p \in W^{per}(\beta,L)} \text{MISE}(N_n) = O\left(n^{-\frac{2\beta}{2\beta+1}}\right).$$

9 Example of Orthogonal Series Estimation

In connection with Exercise 1.9, consider a collection of independent identically distributed random variables $X_1, \ldots, X_{1000}$ from a $Beta(2, 10)$ distribution. Suppose we only know that the sample $X_1, \ldots, X_{1000}$ is taken from a distribution with density $p(x) \in L_2[0, 1]$. Let us estimate $p(x)$ using (8.1) and propose an optimal tuning parameter N, where N is the number of orthonormal basis functions used to estimate p. First, using the cross validation estimator $\hat{J}(N)$ given by (8.5), we can select the value $N_{min} = \underset{N \in N_k}{\arg\min} \hat{J}(N)$ as the optimal choice for N, where $N_k \subseteq \mathbf{N}$. We shall consider the sequence $N_k = \{1, 2, \ldots, 200\}$ as possible choices for N. Note that in theory we always choose N such that $N \ll n$, where n is the sample size. The parameter N is selected in this fashion because N/n should tend to zero for n large. Considering these choices, we determine the optimal N over N_k as $N_{min} = 18$. The estimated coefficients $\hat{c}_1, \ldots, \hat{c}_{18}$ given by $\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)$ are

$$\begin{aligned} &\hat{c}_1 = 1.00,\ \hat{c}_2 = 0.62,\ \hat{c}_3 = 0.97,\ \hat{c}_4 = -0.12,\ \hat{c}_5 = 0.65,\ \hat{c}_6 = -0.18, \\ &\hat{c}_7 = 0.30,\ \hat{c}_8 = -0.13,\ \hat{c}_9 = 0.18,\ \hat{c}_{10} = -0.14,\ \hat{c}_{11} = 0.13,\ \hat{c}_{12} = -0.12, \\ &\hat{c}_{13} = 0.04,\ \hat{c}_{14} = -0.04,\ \hat{c}_{15} = 0.04,\ \hat{c}_{16} = -0.08,\ \hat{c}_{17} = 0.06,\ \hat{c}_{18} = -0.07. \end{aligned}$$



centering

Figure 7: *Histogram based on a random sample of size* $n = 1000$ *from* $Beta(2, 10)$ *distribution. Red represents the estimated density and green represents the true density.*

Figure 7 displays a plot of the estimated beta density. Observing Figure 7, we can see that $\hat{p}_{nN}(x)$ almost perfectly estimates the density of $Beta(2, 10)$. Although the estimator in general performs quite well, the mode does appear to be slightly underestimated, and there do appear to be some wiggles in the estimated density as x is increased. Overall, the estimator $\hat{p}_{nN}(x)$ of $p(x)$ performs reasonable in this scenario.

Note that in theory and in applications, projection estimators perform much better than histograms for estimating unknown densities. In Nonparametric and Semiparametric Models, by Härdle et al. [5], it is shown that a histogram with optimally chosen binwidth $h = h_n$, has asymptotic MISE of order $O(n^{-2/3})$. Contrast this with an orthogonal series estimator of a density $p \in W^{per}(\beta, L)$, where we proved that the MISE is of order $O(n^{\frac{-2\beta}{2\beta+1}})$. For

$\beta > 1$, $n^{\frac{-2\beta}{2\beta+1}}$ is less than $n^{-2/3}$, therefore the projection estimator will perform better than a histogram for n large and density $p \in W^{per}(\beta, L), \beta > 1$.

10 Exercise 1.10 (Page 75)

Consider a nonparametric regression model of the form

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \ldots, n,$$

where

1. f is a function from $[0,1]$ to $\mathbf{R}$ and $f \in W^{per}(\beta, L)$, where $W^{per}(\beta, L)$ is given by (8.6) and $\beta > 0, L > 0$.

2. The random variables $\xi_i, i = 1, \ldots, n$, are independent with $\mathrm{E}[\xi_i] = 0, \mathrm{E}[\xi_i^2] = \sigma_{\xi^2} < \infty$, and $X_i = i/n$ for $i = 1, \ldots, n$.

We shall consider a weighted orthogonal series estimator of f of the form

$$f_{n,\lambda}(x) = \sum_{i=1}^{n} \lambda_j \hat{\theta}_j \phi_j(x), \tag{10.1}$$

where $\{\phi_j\}_{j=1}^{\infty}$ is the trigonometric basis defined by (1.30), $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(X_i)$, $\theta_j = \int_0^1 f(x)\phi_j(x)dx$, and $\sum_{j=1}^{\infty} |\theta_j| < \infty$.

10.1 Part 1

Problem: Prove that the MISE of $\hat{f}_{n,\lambda}(x)$ is minimized with respect to $\{\lambda_j\}_{j=1}^{n}$ at

$$\lambda_j^* = \frac{\theta_j(\theta_j + \alpha_j)}{\epsilon^2 + (\theta_j + \alpha_j)^2}, \quad j = 1, \ldots, n, \tag{10.2}$$

where $\epsilon^2 = \sigma_\xi^2/n$ (λ_j^* are the optimal weights corresponding to the estimator in (10.1)).

Solution: Note that from formula (1.100) of Tsybakov [1], the MISE of $f_{n,\lambda}(x)$ has the form

$$\mathrm{MISE}(\{\lambda_j\}_{j=1}^n) = \mathrm{E}\left[\sum_{j=1}^n (\lambda_j \hat{\theta}_j - \theta_j)^2\right] + \rho_n, \tag{10.3}$$

where $\rho_n = \sum_{j=n+1}^{\infty} \theta_j^2$. Differentiating with respect to $\lambda_j, j = 1, \ldots, n$, we get

$$\frac{\partial}{\partial \lambda_j}\mathrm{MISE}(\{\lambda_j\}_{j=1}^n) = 2\lambda_j \mathrm{E}[\hat{\theta}_j^2] - 2\theta_j \mathrm{E}[\hat{\theta}_j] =: g(\lambda_j).$$

Clearly, $g'(\lambda_j) > 0$ for any $\lambda_j > 0$, therefore the root of $g(\lambda_j)$ minimizes (10.3). Setting $g(\lambda_j)$ equal to zero, we get

$$\lambda_j \mathrm{E}[\hat{\theta}_j^2] - \theta_j \mathrm{E}[\hat{\theta}_j] = 0. \tag{10.4}$$

By Proposition 1.16 on page 53 of Tsybakov [1], we have that $\mathrm{E}[\hat{\theta}_j] = (\theta_j + \alpha_j)$, where $\alpha_j := Bias(\hat{\theta}_j)$ and $\mathrm{E}[\hat{\theta}_j^2] = \mathrm{Var}(\hat{\theta}_j) + (E[\hat{\theta}_j])^2$, with $\mathrm{Var}(\hat{\theta}_j) = \epsilon^2$. Applying the necessary substitutions to (10.4), we get

$$\lambda_j(\epsilon^2 + (\theta_j + \alpha_j)^2) - \theta_j(\theta_j + \alpha_j) = 0,$$

yielding the solution of relation (10.4) in the form

$$\lambda_j^* = \frac{\theta_j(\theta_j + \alpha_j)}{\epsilon^2 + (\theta_j + \alpha_j)^2}, \quad j = 1, \ldots, n,$$

as claimed.

10.2 Part 2

Problem: Check that the corresponding value of the risk is

$$\mathrm{MISE}(\{\lambda_j^*\}_{j=1}^n) = \sum_{j=1}^n \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + (\theta_j + \alpha_j)^2} + \rho_n, \tag{10.5}$$

where $\rho_n = \sum_{j=n+1}^{\infty} \theta_j^2$.

Solution: Again using equation (1.100) from Tsybakov [1], and substituting (10.2) into (10.3), we have

$$\begin{aligned}
\mathrm{MISE}(\{\lambda_j^*\}_{j=1}^n) &= \mathrm{E}\left[\sum_{j=1}^n \left(\frac{\theta_j(\theta_j+\alpha_j)}{\epsilon^2+(\theta_j+\alpha_j)^2}\hat{\theta}_j - \theta_j\right)^2\right] + \rho_n \\
&= \mathrm{E}\left[\sum_{j=1}^n \left(\left(\frac{\theta_j^2(\theta_j+\alpha_j)^2}{(\epsilon^2+(\theta_j+\alpha_j)^2)^2}\right)\hat{\theta}_j^2 \right.\right. \\
&\qquad\qquad \left.\left. - 2\left(\frac{\theta_j^2(\theta_j+\alpha_j)}{\epsilon^2+(\theta_j+\alpha_j)^2}\right)\hat{\theta}_j + \theta_j^2\right)\right] + \rho_n \\
&= \sum_{j=1}^n \left(\left(\frac{\theta_j^2(\theta_j+\alpha_j)^2}{(\epsilon^2+(\theta_j+\alpha_j)^2)^2}\right)(\epsilon^2+(\theta_j+\alpha_j)^2) \right. \\
&\qquad\qquad \left. - 2\left(\frac{\theta_j^2(\theta_j+\alpha_j)}{\epsilon^2+(\theta_j+\alpha_j)^2}\right)(\theta_j+\alpha_j) + \theta_j^2\right) + \rho_n \\
&= \sum_{j=1}^n \left(\theta_j^2 - \frac{\theta_j^2(\theta_j+\alpha_j)^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right) + \rho_n \\
&= \sum_{j=1}^n \left(\theta_j^2\left(\frac{\epsilon^2+(\theta_j+\alpha_j)^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right) - \frac{\theta_j^2(\theta_j+\alpha_j)^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right) + \rho_n \\
&= \sum_{j=1}^n \left(\frac{\theta_j^2\epsilon^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right) + \rho_n,
\end{aligned}$$

which is our desired result.

10.3 Part 3

Problem: Prove that

$$\sum_{j=1}^n \left(\frac{\theta_j^2\epsilon^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right) = (1+o(1))\sum_{j=1}^n \frac{\epsilon^2\theta_j^2}{\epsilon^2+\theta_j^2} \tag{10.6}$$

Solution: Firstly,

$$\begin{aligned}\sum_{j=1}^{n}\frac{\epsilon^2\theta_j^2}{\epsilon^2+(\theta_j+\alpha_j)^2} &= \sum_{j=1}^{n}\left(\frac{\epsilon^2\theta_j^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right)\left(\frac{\epsilon^2+\theta_j^2}{\epsilon^2+\theta_j^2}\right)\\ &= \sum_{j=1}^{n}\left(\frac{\epsilon^2+\theta_j^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right)\left(\frac{\epsilon^2\theta_j^2}{\epsilon^2+\theta_j^2}\right)\\ &\leq \sum_{j=1}^{n}\left(\frac{\epsilon^2+\theta_j^2}{\epsilon^2+(\theta_j-|\alpha_j|)^2}\right)\left(\frac{\epsilon^2\theta_j^2}{\epsilon^2+\theta_j^2}\right)\end{aligned}$$

Note that by Lemma 1.8 on page 54 of Tsybakov [1], $\max_{1\leq i\leq n}|\alpha_i| \leq 2\sum_{m=n+1}^{\infty}|\theta_m|$, for all $n \geq 1$. By the assumption that the Fourier coefficients are absolutely convergent, $\lim_{n\to\infty}\sup_{1\leq i\leq n}|\alpha_i| = 0$, thus the term $\frac{\epsilon^2+\theta_j^2}{\epsilon^2+(\theta_j-|\alpha_j|)^2}$ approaches 1 from above as n approaches infinity. We conclude that as $n \to \infty$

$$\sum_{j=1}^{n}\left(\frac{\epsilon^2\theta_j^2}{\epsilon^2+(\theta_j+\alpha_j)^2}\right) = (1+o(1))\sum_{j=1}^{n}\frac{\epsilon^2}{\epsilon^2+\theta_j^2}.$$

10.4 Part 4

Problem: Prove that as $n \to \infty$,

$$\rho_n = (1+o(1))\sum_{j=n+1}^{\infty}\frac{\epsilon^2\theta_j^2}{\epsilon^2+\theta_j^2}, \tag{10.7}$$

where $\rho_n = \sum_{j=n+1}^{\infty}\theta_j^2$.

Solution: We have

$$\begin{aligned}\rho_n &= \sum_{j=n+1}^{\infty}\left(\frac{\theta_j^2\epsilon^2}{\epsilon^2}\right)\left(\frac{\epsilon^2+\theta_j^2}{\epsilon^2+\theta_j^2}\right)\\ &= \sum_{j=n+1}^{\infty}\left(\frac{\theta_j^2\epsilon^2}{\epsilon^2+\theta_j^2}\right)\left(\frac{\epsilon^2+\theta_j^2}{\epsilon^2}\right)\\ &= \sum_{j=n+1}^{\infty}\frac{\theta_j^2\epsilon^2}{\epsilon^2+\theta_j^2} + \sum_{j=n+1}^{\infty}\frac{\theta_j^2}{\epsilon^2}\left(\frac{\theta_j^2\epsilon^2}{\epsilon^2+\theta_j^2}\right)\end{aligned}$$

Note that the condition $\sum_{j=1}^{\infty} |\theta_j| < \infty$ implies that $\lim_{n\to\infty} \sum_{j=n+1}^{\infty} |\theta_j|^2 = 0$, hence $\sum_{j=n+1}^{\infty} \theta_j^2/\epsilon^2 = o(1)$. Thus

$$\rho_n = (1 + o(1)) \sum_{j=1}^{n} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2}.$$

10.5 Part 5

Problem: Deduce from the above results that as $n \to \infty$,

$$\text{MISE}(\{\lambda_j^*\}_{j=1}^{\infty}) = \mathcal{A}_n^*(1 + o(1)),$$

where

$$\mathcal{A}_n^* := \sum_{j=1}^{\infty} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2}, \tag{10.8}$$

with $\epsilon^2 = \sigma_\xi^2/n$.

Solution: From Part 2, we have that

$$\text{MISE}(\{\lambda_j^*\}_{j=1}^{n}) = \sum_{j=1}^{n} \left(\frac{\theta_j^2 \epsilon^2}{\epsilon^2 + (\theta_j + \alpha_j)^2} \right) + \rho_n.$$

Using (10.6) and (10.7), observe that

$$\begin{aligned}
\text{MISE}(\{\lambda_j\}_{j=1}^{n}) &= (1 + o(1)) \sum_{j=1}^{n} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2} + (1 + o(1)) \sum_{j=n+1}^{\infty} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2} \\
&= (1 + o(1)) \sum_{j=1}^{\infty} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2} \\
&= (1 + o(1)) \mathcal{A}_n^*.
\end{aligned}$$

This result implies that as the sample size tends to infinity, the bound on the MISE approaches $\mathcal{A}_n^*$.

10.6 Part 6

Problem: Check that

$$\mathcal{A}_n^* < \min_{N \geq 1} \mathcal{A}_{nN},$$

where

$$\mathcal{A}_{nN} = \epsilon^2 N + \rho_N, \text{ with } \rho_N = \sum_{j=N+1}^{\infty} \theta_j^2.$$

Note that A_{nN} is a bound on the MISE of the simple projection estimator as $n \to \infty$.

Solution: Using (10.8),

$$\begin{aligned}
\mathcal{A}_n^* &= \sum_{j=1}^{N} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2} + \sum_{j=N+1}^{\infty} \frac{\epsilon^2 \theta_j^2}{\epsilon^2 + \theta_j^2} \\
&= \sum_{j=1}^{N} \left(\frac{\theta_j^2}{\epsilon^2 + \theta_j^2} \right) \epsilon^2 + \sum_{j=N+1}^{\infty} \left(\frac{\epsilon^2}{\epsilon^2 + \theta_j^2} \right) \theta_j^2 \\
&< \epsilon^2 N + \sum_{j=N+1}^{\infty} \theta_j^2 = \mathcal{A}_{nN}.
\end{aligned}$$

Therefore for suitably chosen weights, the weighted orthogonal series estimator is uniformly better than the simple orthogonal series estimator, with respect to MISE.

11 Conclusion

In this project, we have demonstrated a variety of interesting mathematical and statistical properties of nonparametric estimators, including kernel density estimators, local polynomial regression estimators, and orthogonal series estimators. We have presented suitable extensions to each of these estimators and shown that each estimator is quite flexible. We have also considered a few situations where the aformentioned estimators can be used in practice.

Many properties of such estimators and their extensions are unknown to the majority of statisticians. We hope that showcasing these estimators, and demonstrating to the reader their practical use will play a role in revitalizing interest in nonparametric estimation.

As was stated previously, nonparametric statistics enjoys a wider applicability than parametric statistics. It is rare that the statistician will know enough about their data to justify a parametric model, especially in the early stages of analysis. Even if one chooses to use a parametric model, a nonparametric model can often be used as a benchmark for determining the feasibility of a parametric model.

Nonparametric estimation continues to be an actively studied area of mathematical statistics. We chose mainly to focus on the quality of an estimator in terms of its quadratic risk. There are other possible estimators that one could consider that are optimal in other senses, for example, estimators that minimize the L_1-norm. Typically, however, dealing with L_2-risk is more mathematically convenient than with L_p-risk, $p \neq 2$. One could also choose to study minimax estimators, that is, estimators whose maximum risk is minimal among all possible estimators. We encourage an interested reader to further explore the plethora of topics in nonparametric estimation.

12 Appendix

Result 1: Let A be a positive definite matrix in $\mathbf{R}^{n\times n}$ that is symmetric, then for any vector $x \in \mathbf{R}^n$, $||Ax|| \leq \lambda_{max}||x||$, where λ_{max} is the largest eigenvalue of A.

Proof: Note that $||Ax|| = \sqrt{(Ax)^\top (Ax)}$. There exist eigenvectors $v_1, \ldots, v_n$ that form a orthonormal basis in $\mathbf{R}^n$, with corresponding positive eigenvalues $\lambda_1, \ldots, \lambda_n$. Since the eigenvectors form a basis in $\mathbf{R}^n$, there exist real numbers $c_1, \ldots, c_n$ such that $x = \sum_{i=1}^n c_i v_i$. Substituting the expression for x into

$||Ax||$, we obtain

$$||Ax|| = \sqrt{\left(A\sum_{i=1}^{n} c_i v_i\right)^\top \left(A\sum_{i=1}^{n} c_i v_i\right)} = \sqrt{\left(\sum_{i=1}^{n} \lambda_i c_i v_i\right)^\top \left(\sum_{i=1}^{n} \lambda_i c_i v_i\right)}$$
$$\leq \lambda_{max}\sqrt{\left(\sum_{i=1}^{n} c_i v_i\right)^\top \left(\sum_{i=1}^{n} c_i v_i\right)} = \lambda_{max}||x||.$$

Result 2: Let $A \in \mathbf{R}^{n\times n}$ be an invertible matrix that is positive definite. Suppose that the eigenvalues of the eigenvectors of A are such that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, then the eigenvalues of A^{-1} arranged from minimum to maximum are $1/\lambda_n \leq 1/\lambda_{n-1} \leq \ldots \leq 1/\lambda_1$.

Proof: Using the spectral decomposition of $A = PDP^\top$, with orthogonal matrix $P = (v_1, \ldots, v_n)$, and $D = diag(\lambda_1, \ldots, \lambda_n)$, we have $A^{-1} = (PDP^\top)^{-1} = PD^{-1}P^\top$. Noting that $D = diag(\lambda_1, \ldots, \lambda_n)$, we must have $D^{-1} = diag\,(1/\lambda_1, \ldots, 1/\lambda_n)$, thus the eigenvalues of A^{-1} are $1/\lambda_1, \ldots, 1/\lambda_n$.

References

[1] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics, Paris, France, 2009.

[2] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 1962.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics, New York, NY, 2017.

[4] R. Eubank. *Nonparametric Regression and Spline Smoothing.* Marcel Dekker, New York, NY, 1999.

[5] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models.* Springer Series in Statistics, Berlin, Germany, 2004.

13 R Code

```
######### EXERCISE 1.5 CODE ########################
#Plots of the Fourier transforms of the rectangular kernel and
biweight kernel.

t<-seq(-10,10,by=0.01)
zeroVector<-t*0
Khat<-sin(t)/t
plot(Khat~t,type="l",
main="Plot of Fourier transform of rectangular kernel",
ylab="Khat", xlab="t")
lines(zeroVector~t,col="red")

KhatBiweight<-(-15*sin(t)/(t^3)) + (45*sin(t)/t^5) -(45*cos(t)/t^4)
plot(KhatBiweight~t,type="l",
main="Plot of Fourier transform of biweight kernel",
ylab="Khat", xlab="t")
zeroVectorBiweight<-t*0
lines(zeroVectorBiweight~t,col="red")


######### EXAMPLE OF ORTHOGONAL SERIES ESTIMATION ##################

set.seed(431242)
x<-rbeta(1000, 2, 10)
hist(x,breaks=100,freq=FALSE)

#Construct trigonometric basis
```

```
basisfunction<-function(j,x)
{
  if (j==1){
    return(x/x)
  } else if (j %% 2 != 0){
    return(sqrt(2)*sin(pi*(j-1)*x))
  } else if ( j %% 2 ==0) {
    return(sqrt(2)*cos(pi*(j)*x))
  }
}

#Return \hat{c}_{j}, j=1,2,...N
CoefEstimate<-function(x,N){
CCoeffs<-rep(0,N)
for (j in 1:N){
  CCoeffs[j]<-mean(basisfunction(j,x))
}
return(CCoeffs)
}

#Sequence of points we shall estimate the density at

pofI<-seq(0,1,by=1/99)

#Function that returns density estimate at fixed point

phat<-function(a,Coefs){
basis<-rep(0,length(Coefs))
for (i in 1:length(Coefs)){
```

```
  basis[i]<-basisfunction(i,a)
}
return(sum(Coefs*basis))
}

#Data-driven estimator for determining optimal N with respect to MISE

Jhat<-function(x,N){
  Coefficients<-CoefEstimate(x,N)
  J<-rep(0,N)
  for(i in 1:N){
    J[i]<-(2/length(x))*sum(basisfunction(i,x)^2)
    -(length(x)+1)*Coefficients[i]^2
  }
  return((1/(length(x)-1))*sum(J))
}

#Cross-validation scores

CrossValidation<-function(data,NMin,NMax){
  CVScores<-seq(NMin,NMax,by=1)
for (i in NMin:NMax){
  CVScores[i]<-Jhat(data,i)
}
return(CVScores)
}

Scores<-CrossValidation(x,1,200)

#Select N that minimizes cross-validation score
```

```
Nmin<-which.min(Scores)

#Estimate density with optimal N

EstimateCoeff<-CoefEstimate(x,Nmin)
density<-rep(0,100)
for (i in 1:length(density)){
  density[i]<-max(phat(pofI[i],EstimateCoeff),0)
}

#Plot density estimate

hist(x,breaks=100,freq=FALSE,
main="Histogram of Beta(2,10) distribution",xlab="X")
lines(density~pofI,col="red")
lines(dbeta(pofI,2,10)~pofI,col="green")
#Return chat_j, j=1,2,...,N
round(EstimateCoeff,2)

######### END OF EXAMPLE ##############################################

######### EXAMPLE OF KERNEL DENSITY ESTIMATION ########################

#Evaluate Legendre polynomial of order m, unnormalized, at the point x.

LegendrePolynomialatXUnnormalized<-function(x,m){
  {
  if (m==0)
  {
```

```
    return(1)
  } else if (m==1)
  {
    return(x)
  } else if (m>=2)
  {
    return((((2*(m-1)+1)/m)*x*LegendrePolynomialatXUnnormalized(x,m-1)
    -((m-1)/m)*LegendrePolynomialatXUnnormalized(x,m-2)))
  }
}
}

#Evaluate Legendre polynomial of order m, normalized, at the point x.
LegendrePolynomialatXNormalized<-function(x,m)
{
  if (m==0)
  {
    return(1/sqrt(2))
  } else if (m>=1)
  {
    return(sqrt((2*m+1)/2)*LegendrePolynomialatXUnnormalized(x,m))
  }
}

#sth derivative of mth degree Legendre polynomial,unnormalized,
evaluated at x.

LegendreDerivative<-function(x,m,s)
{
  if (s==1)
```

```
  {
  if(m<=0){
    return(0)
  } else if (m==1)
  {
    return(1)
  } else if (m==2)
  {
    return(3*(x))
  } else if (m>=3)
    k<-m-1
    deriv<-0
    while (k>=0)
    {
    deriv<-deriv+(2*k+1)*LegendrePolynomialatXUnnormalized(x,k)
    k<-k-2
    }
  return(deriv)
} else if (s>=2)
{
if (m<s){
  return(0)
} else if (m>=s){
  k<-m-1
  deriv<-0
  while (k>=0)
  {
  deriv<-deriv +(2*k+1)*LegendreDerivative(x,k,s-1)
  k<-k-2
  }
```

```
  }
    return(deriv)
  }
}

#sth derivative of mth degree legendre polynomial, normalized,
evaluated at x.

DerivativeNormalised<-function(x,m,s)
{
  return(sqrt((2*m+1)/2)*LegendreDerivative(x,m,s))
}

#Construct kernel of order l.
KatX<-function(x,l,s)
{
Kernel<-0
    for (k in 0:l)
  {
Kernel<-Kernel+DerivativeNormalised(0,k,s)
*LegendrePolynomialatXNormalized(x,k)*as.numeric(abs(x)<=1)
  }
return(Kernel)
}

#Estimator of sth derivative of density p, evaluated at a set of
#points specified by the vector x.

pHatest<-function(x,data,h,l,s){
  pHat<-0
```

```
  for(i in 1:length(data))
  {
  pHat<-pHat+(1/(length(data)*(h^(s+1))))*KatX((data[i]-x)/h,l,s)
  }
  return(pHat)
}

#Generate sequence of numbers on interval [-1,1].
v<-seq(-1,1,0.01)
PhatNormal<-rep(0,length(v))
set.seed(4324)

#Generate random sample from truncated normal distribution on [-1,1]
NormalTrunc<-qnorm(pnorm(-1)+runif(1000)*(pnorm(1)-pnorm(-1)))

#Set of possible bandwidths.

h<-seq(0.1,4,0.1)

#Vector that will store absolute deviation between true
and estimated density.

dev<-rep(0,length(h))

#Actual density of second derivative of truncated normal distribution

densityActual<-(v^2-1)*(dnorm(v))/(pnorm(1)-pnorm(-1))

#Plot density estimates over [-1,1] for all considered bandwidths,
#and compare to actual density.
```

```
for (j in 1:length(h)){
for(i in 1:length(PhatNormal))
{
  PhatNormal[i]<-pHatest(v[i],NormalTrunc,h[j],3,2)
}
 #Take mean absolute deviation between true and
 estimated density
   dev[j]<-mean(abs(PhatNormal-densityActual))
}
#Select bandwidth with minimum mean absolute deviation

which.min(dev)
h[which.min(dev)]

#Compute density estimate for optimal h

for(i in 1:length(PhatNormal))
{
  PhatNormal[i]<-pHatest(v[i],NormalTrunc,h[which.min(dev)],3,2)
}
#Plot density estimate with optimal h and compare to true density

plot(densityActual~v,col="red",type='l',ylim=c(-0.8,0.1),main=
"Estimated Second Derivative vs True Second Derivative",
xlab="x",ylab="Density")
lines(PhatNormal~v,col="green")

######### END OF CODE ######################################
```