Notes from Nonparametric Statistical Theory

Cambridge Part III Mathematical Tripos 2012-2013

Lecturer: Adam Bull

Vivak Patel

April 13, 2013

Contents

1 Introduction

1.1 Convergence of Random Variables

Definition 1.1. *Let $(\Omega, A, \boldsymbol{P}[])$ be a probability space. Let $X_1, X_2, \ldots, X : \Omega \to (S, d)$ be random variables taking values in metric space (S, d). Let $\boldsymbol{E}[]$ denote the expectation with respect to the probability measure.*

1. *If $\boldsymbol{P}[X_n \to X] = 1$ then $X_n \xrightarrow{a.s.} X$*
2. *If $\forall \epsilon > 0$, $\boldsymbol{P}[d(X_n, X) > \epsilon] \to 0$ then $X_n \xrightarrow{\boldsymbol{P}[]} X$*
3. *IF $\forall f \in C(S)$ and bounded for which $\boldsymbol{E}[f(X_n)] \to \boldsymbol{E}[f(X)]$ then $X_n \xrightarrow{d} X$*

Proposition 1.1. *For random variables $X_1, X_2, \ldots, X$:*

1. *$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\boldsymbol{P}[]} X \implies X_n \xrightarrow{d} X$*
2. *$X \implies X_n \xrightarrow{d} X$ if and only if the distribution function F_n convergest pointwise to the distribution function F.*

Theorem 1.1. *(Strong Law of Large Numbers) Suppose $X_1, X_2, \ldots, X$ are i.i.d R.V. taking values in $(\Re^p, \|\cdot\|)$. If X has mean μ then the empirical mean $\boldsymbol{E}_n X = \frac{1}{n}\sum_{i=1}^n X_i$ converges a.s. to μ.*

Theorem 1.2. *(Central Limit Theorem) Suppose $X_1, X_2, \ldots, X$ are i.i.d R.V. taking values in $(\Re^p, \|\cdot\|)$. If X has mean μ and positive definite covariance matrix Σ then the error rate of estimation $\sqrt{n}(\boldsymbol{E}_n X - \mu) \xrightarrow{d} N(0, \Sigma)$.*

Theorem 1.3. *(Hoeffding's Inequality) Suppose $X_1, X_2, \ldots X_n$ are zero-mean R.V. taking values in $[b_i, c_i]$. Then for every natural number n and $u > 0$,*

$$\boldsymbol{P}\left[\left|\sum_{i=1}^n X_i\right| > u\right] \leq 2\exp\left(\frac{-2u^2}{\sum_{i=1}^n (c_i - b_i)^2}\right)$$

1.2 Uniform Law of Large Numbers

1. Motivation: In the nonparametric context, we want the LLN to hold uniformly over a class of functions taking values in a measure space T.
2. Notations and Conventions
   (a) Let H be the class of functions for which $\mathbf{E}[|h(X)|] < \infty$, $\forall h \in H$
   (b) The empirical mean is denoted $\mathbf{E}_n[h] = \frac{1}{n}\sum_{i=1}^n h(X_i)$
   (c) Suppose $l, u : T \to \Re$ are measurable. The bracket $[l, u] = \{f : T \to \Re | l(x) \leq f(x) \leq u(x)\}$
3. The Uniform Law of Large Numbers:

   Proposition 1.2. *Suppose for $\epsilon > 0$, there exists a finite set of brackets $[l_i, u_i]$ with $l_i, u_i \in L^1$ and $\boldsymbol{E}[u_j] - \boldsymbol{E}[l_j] < \epsilon$. If $\forall \epsilon > 0, \forall h \in H$, $\exists j$ such that $h \in [l_j, u_j]$ then $\sup_{h \in H} |\boldsymbol{E}_n[h] - \boldsymbol{E}[h]| \xrightarrow{a.s.} 0$.*

Proof. We want to show that $\forall \epsilon > 0 \ \exists n_0 \in N$ such that for $n \geq n_0$

$$\sup_{h \in H} |\mathbf{E}_n [h] - \mathbf{E} [h] | \leq 2\epsilon \quad a.s.$$

(a) Let $\epsilon > 0$ and $[l_i, u_i]$ be a bracketing as above. By the SLLN, $\forall j, \ \exists n_j$ such that for $n \geq n_j$, $|\mathbf{E}_n [l_j] - \mathbf{E} [l_j] | \leq \epsilon$ and similarly for u_j. Since j is finite, let $n_0 = \max(n_j)$.

(b) Let $h \in H$. Then $\exists j$ such that $h \in [l_j, u_j]$.

$$\mathbf{E}_n [h] \geq \mathbf{E}_n [l_j] \geq \mathbf{E} [l_j] - \epsilon \geq \mathbf{E} [u_j] - 2\epsilon \geq \mathbf{E} [h] - 2\epsilon$$
$$\mathbf{E}_n [h] \leq \mathbf{E}_n [u_j] \leq \mathbf{E} [u_j] + \epsilon \geq \mathbf{E} [l_j] + 2\epsilon \geq \mathbf{E} [h] + 2\epsilon$$

(c) Therefore, $|\mathbf{E}_n [h] - \mathbf{E} [h] | \leq 2\epsilon$. Since h is arbitrary, we can simply take the supremum.

□

2 Classical Empirical Process Theory

2.1 Empirical Distribution Function and Properties

Definition 2.1. *The empirical distribution function* $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} [x_i \leq t]$.

2.1.1 Convergence (LLN)

Theorem 2.1. *(Uniform Convergence, Glivenko-Cantelli) Let* $X_1, X_2, \ldots, X_n$ *be i.i.d R.V. with an arbitrary distribution function* F*. Then*

$$\sup_{t \in \Re} |F_n(t) - F(t)| \xrightarrow{a.s.} 0$$

Proof. We want to apply the ULLN to the set of functions $H = \{\mathbf{1} [x \leq t] \, | t \in \Re\}$. So, we simply need to find brackets of arbitrary size. There are two cases: F is continuous, and F has discontinuities.

1. When F is continuous, it takes on every value from $[0, 1]$. So for any $\epsilon > 0$ we can find a k such that $\frac{1}{k} < \epsilon$. Let t_j be such that $F(t_j) = \frac{j}{k}$ for $j = 1, \ldots, k-1$. Then H is covered by the brackets

   $$[0, \mathbf{1} [x \leq t_1]], [\mathbf{1} [x \leq t_1], \mathbf{1} [x \leq t_2]], \ldots, [\mathbf{1} [x \leq t_{k-1}], 1]$$

   for which

   $$|\mathbf{E} [\mathbf{1} [x \leq t_j]] - \mathbf{E} [\mathbf{1} [x \leq t_{j+1}]] | \leq \frac{1}{k}$$

2. When F has discontinuities, it may skip certain values of $\frac{j}{k}$. So $F(t_j^-) < \frac{j}{k}$ and $F(t_j^+) > \frac{j}{k}$. To deal with this, we add the brackets $[\mathbf{1} [x \leq t_{j-1}], \mathbf{1} \left[x \leq t_j^-\right]]$ and $[\mathbf{1} \left[x \leq t_j^+\right], \mathbf{1} [x \leq t_{j+1}]]$. We still have a finite bracket for which the expected distance between the upper and lower limits are $\leq \frac{1}{k}$.

□

Theorem 2.2. *(Multi-variate Glivenko-Cantelli) Let $X_1, X_2, \ldots, X_n$ be i.i.d in $\Re^d$ with distribution $F(t)$. Let $F_n(t)$ be the empirical distribution function. Then*

$$\sup_{t \in \Re^d} |F_n(t) - F(t)| \xrightarrow{a.s.} 0$$

2.1.2 Rate of Convergence (CLT)

Proposition 2.1. *For a fixed $t \in \Re$.*

$$\sqrt{n}(F_n(t) - F(t) \xrightarrow{d} N(0, F(t)(1 - F(t))$$

Theorem 2.3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with distribution function F. Then $\sqrt{n}(F_n - F)$ converges to the F-Brownian bridge process in L^∞.*

Definition 2.2. *The F-Brownian bridge process G_F is the zero-mean Gaussian process indexed by $\Re$ with covariance $\boldsymbol{E}[G_F(t_i), G_F(t_j)] = F(t_j \wedge t_i) - F(t_i)F(t_j)$*

2.1.3 Finite Sample Error Bound (Hoeffding)

The following theorem is the empirical distribution process equivalent to Hoeffding's inequality.

Theorem 2.4. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. R.V. with distribution F. Then $\forall n \in N$ and $\forall \lambda > 0$,*

$$\boldsymbol{P}\left[\|\sqrt{n}\,(F_n - F)\,\|_\infty > \lambda\right] \leq 2\exp[-2\lambda^2]$$

2.1.4 Optimality

Given no prior information on F, the empirical distribution function is the asymptotically optimal choice for estimating F.

Theorem 2.5. *Let $X_1, \ldots, X_n$ be i.i.d. R.V. with distribution F. Let T_n be an estimator for F based on the R.V. X_i. Then:*

$$\lim_{n \to \infty} \frac{\sup_F \boldsymbol{E}[\|F_n - F\|_\infty]}{\inf_{T_n} \sup_F \boldsymbol{E}[\|T_n - F\|_\infty]} = 1$$

Note 2.1. $\inf_{T_n}$ *gives us the best estimator for* $\sup_F$*, the worst distribution function available.*

2.2 Kolmogorov-Smirnov Test

Motivation: We know that $\sqrt{n}(F_n - F)$ behaves like G_F. However, we often do not know F, so determining G_F is difficult. The next result gives us a way around this.

Theorem 2.6. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. R.V. with continuous distribution F. Then*

$$\sqrt{n}\|F_n - F\|_\infty \xrightarrow{d} \|G\|_\infty$$

the standard brownian bridge. Moreover, for $\lambda > 0$,

$$\boldsymbol{P}[\|G\|_\infty > \lambda] = 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp(2j^2\lambda^2)$$

Proof. We prove the first part

1. Let $Y_i = F(X_i)$. Then Y_i are uniformly distributed with distribution $U(x) = x$.

2. Note that $U(F(t)) = F(t)$. And $U_n(F(t)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[Y_i \leq F(t)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[X_i \leq t\right] = F_n(t)$. Therefore,

$$\|U_n - U\|_\infty = \|U_n(F) - U(F)\|_\infty = \|F_n - F\|_\infty$$

3. By the rate of convergence (CLT) for empirical distributions, we have that $\sqrt{n}(U_n - U)$ converges in L^∞ to the standard brownian bridge process. Since $f \mapsto \|f\|_\infty$ is continuous, the continuous mapping theorem yields the desired first result.

□

3 Minimax Lower Bounds

3.1 Introduction

1. In parametric theory and distribution function estimation, the rate of convergence to a limiting distribution is $\frac{1}{\sqrt{n}}$. However, in other nonparametric cases, our rate of convergence is strictly worse than this.

2. Minimax Risk

   Definition 3.1. *Let* F *be a collection of densities or regression functions. Suppose under some distribution function* $f \in \mathrm{F}$*, the data have probability measure* $\boldsymbol{P}_f[]$ *and expectation* $\boldsymbol{E}_f[]$*. Let* $d(f, g)$ *be a metric on* F*. The best performance estimator under loss* d *for an estimator* $\hat{f}_n$ *is the minimax risk:*

$$R_n(\mathrm{F}, d) = \inf_{\hat{f}_n} \sup_{f \in \mathrm{F}} \boldsymbol{E}_f\left[d(f, \hat{f}_n)\right]$$

3. Minimax Rate of Convergence

   Definition 3.2. *If the risk is bounded above and below by a deterministic sequence* r_n *(i.e. for some* $C, D > 0$*,* $Cr_n \leq R_n(\mathrm{F}, d) \leq Dr_n$*) then* r_n *is the minimax rate of convergence.*

3.2 Reduction to Testing Problem

We will determine lower bounds by framing this as a testing problem, allowing us to lower bound the minimax risk. Then, we will lower bound this to get an estimate for the lower bound constant.

1. We reduce to a testing problem as follows:

   (a) Suppose we want to test the hypotheses $H_0 : f = f_0$ and $H_1 : f = f_1$ where $f_0, f_1 \in \mathrm{F}$ such that $d(f_0, f_1) > 2r_n$.

(b) Given an estimator $\hat{f}_n$, we construct the statistic:

$$T_n = \begin{cases} 0 & \text{if } d(f_0, \hat{f}_n) \leq d(f_1, \hat{f}_n) \\ 1 & \text{if } d(f_0, \hat{f}_n) > d(f_1, \hat{f}_n) \end{cases}$$

(c) It follows that

$$R(\mathrm{F}, d)/r_n \geq \inf_{T_n} \max_{j \in \{0,1\}} \mathbf{P}_{f_j} [T_n \neq j]$$

Derivation 3.1. *By the Markov Property:*

$$\frac{R(\mathrm{F}, d)}{r_n} = \inf_{\hat{f}_n} \sup_{f \in \mathrm{F}} \frac{\boldsymbol{E}_f \left[d(\hat{f}_n, f) \right]}{r_n} \geq \inf_{\hat{f}_n} \sup_{f \in \mathrm{F}} \boldsymbol{P}_f \left[d(\hat{f}_n, f) > r_n \right]$$

Also note that if $T_n = 1$ then $d(f_0, \hat{f}_n) > d(f_1, \hat{f}_n)$. If $T_n = 0$ then $d(f_0, \hat{f}_n) \leq d(f_1, \hat{f}_n)$. So we have by triangle inequality when T_n is 1 and 0 respectively:

$$d(f_0, \hat{f}_n) \geq d(f_0, f_1) - d(f_1, \hat{f}_n) \geq 2r_n - d(f_1, \hat{f}_n) \geq 2r_n - d(f_0, \hat{f}_n)$$
$$d(f_1, \hat{f}_n) \geq d(f_0, f_1) - d(f_0, \hat{f}_n) \geq 2r_n - d(f_0, \hat{f}_n) \geq 2r_n - d(f_1, \hat{f}_n)$$

We now recast the probability from the beginning in terms of T_n:

$$\begin{aligned} \inf_{\hat{f}_n} \sup_{f \in \mathrm{F}} \boldsymbol{P}_f \left[d(\hat{f}_n, f) > r_n \right] &\geq \inf_{\hat{f}_n} \max_{j \in \{0,1\}} \boldsymbol{P}_{f_j} \left[d(f_j, \hat{f}_n) > r_n \right] \\ &= \inf_{T_n} \max_{j \in \{0,1\}} \boldsymbol{P}_{f_j} [T_n \neq j] \end{aligned}$$

2. To lower bound this probability, we have the following Lemma:

Lemma 3.1. *Let $Z = \frac{d\boldsymbol{P}_{f_1}[]}{d\boldsymbol{P}_{f_0}[]}$ be the likelihood ratio. Then,*

$$\inf_{T_n} \max_{j \in \{0,1\}} \boldsymbol{P}_{f_j} [T_n \neq j] \geq \frac{1}{2} \left(1 - \sqrt{\boldsymbol{E}_{f_0} [(Z-1)^2]} \right)$$

Proof. The penultimate line follows from Cauchy-Schwarz:

$$\begin{aligned} 2 \max \mathbf{P}_{f_j} [T_n \neq j] &\geq \mathbf{P}_{f_0} [T_n = 1] + \mathbf{P}_{f_1} [T_n = 0] \\ &= 1 - \mathbf{P}_{f_0} [T_n = 0] + \mathbf{E}_{f_1} [\mathbf{1} [T_n = 0]] \\ &= \mathbf{E}_{f_0} [1 - \mathbf{1} [T_n = 0]] + \int \mathbf{1} [T_n = 0] \, d\mathbf{P}_{f_1} \\ &\geq \mathbf{E}_{f_0} [1 - \mathbf{1} [T_n = 0]] + \int \mathbf{1} [T_n = 0] \frac{d\mathbf{P}_{f_1}}{d\mathbf{P}_{f_0}} d\mathbf{P}_{f_0} \\ &= 1 - \mathbf{E}_{f_0} [(1 - Z) \mathbf{1} [T_n = 0]] \\ &\geq 1 - \sqrt{\mathbf{E}_{f_0} [(1-Z)^2] \, \mathbf{E}_{f_0} \left[\mathbf{1} [T_n = 0]^2 \right]} \\ &\geq 1 - \sqrt{\mathbf{E}_{f_0} [(Z-1)^2]} \end{aligned}$$

□

3. When the two densities for f_1 and f_0 are close, the test is impossible.

3.3 Lower Bounds for Differentiable Densities

The minimax rate of convergence depends on the number of derivatives the density has and a maximum over the derivatives. We define the following accordingly:

Definition 3.3. *For the class of s-times differentiable densities, let:*

1. $\|f\|_{s,\infty} = \max_{k \le s} \left\|f^{(k)}\right\|_\infty$ *be a norm*
2. $C^s(M) = \{f | f$ *is a density,* $\|f\|_{s,\infty} \le M\}$ *is the subset of differentiable densities with bounded derivatives.*

Theorem 3.1. *Let* $s \in N$, $M > 1$ *and* $x_0 \in [0,1]$. *Let* $X_1, \ldots, X_n$ *be i.i.d. R.V. with density* f *on* $[0,1]$ *and* $\hat{f}_n$ *be an estimator depending only on* $X_1, \ldots, X_n$. *Then for some constant* $C > 0$ *and* n *large enough,*

$$\inf_{\hat{f}_n} \sup_{f \in C^s(M)} \boldsymbol{E}_f \left[\left| \hat{f}_n(x_0) - f(x_0) \right| \right] > C n^{\frac{-s}{2s+1}}$$

Proof. PROOF GOES HERE. □

Remark 3.1. *Similar results can be proven with different loss functions such as* L^p *or* L^∞ *loss. They can also be proven for non-integer smoothness or in regression, but the rates will be no better than* $n^{-s/(2s+1)}$ *or worse than* $\left(\frac{n}{\log(n)}\right)^{-s/(2s+1)}$.

4 Approximation of Functions

4.1 Introduction

1. When estimating density or regression function f, there is no natural estimator $\hat{f}_n$ as there is for the distribution function. The "best" (maximum likelihood) natural estimators either:
   (a) Place peaks at every data point in density estimation
   (b) Interpolate between data points in regression estimation
2. Although such MLE have low bias, they have extremely high variance. So we exchange some bias for lower variance by estimating a nicer function near f, which does have a natural estimator.
3. Notation:
   (a) L^∞ is the space of $f : \Re \to \Re$ with $\|f\|_\infty = \sup_{x \in \Re} |f(x)| \le \infty$
   (b) For $p \in [1, \infty)$, L^p is the space of $f : \Re \to \Re$ with $\|f\|_p = \left(\int_\Re |f|^p\right)^{1/p} \le \infty$
   (c) f is locally integrable if for all bounded borel sets $C \subset \Re$, $\int_C |f| < \infty$

4.2 Regularisation by Convolution

1. Convolution

   Definition 4.1. *Let $f, g : \Re \to \Re$. Their convolution, if the integral exists, is $(f * g)(x) = \int_{\Re} f(x-y)g(y)dy = \int_{\Re} f(y)g(x-y)dy = (g * f)(x)$.*

2. Convolution with K_h

   (a) Let $K : \Re \to \Re$ such that $\int_{\Re} K(x)dx = 1$. For $h > 0$ we have the localised kernel $K_h(x) = \frac{1}{h}K(\frac{x}{h})$

   (b) h is called the bandwidth and controls how localised the effects of K_h are.

3. "Consistency" and "Error" of kernel convolution of a function with respect to the original function.

   Proposition 4.1. *Let $K : \Re \to \Re$ be a bounded kernel (integrates to 1) with compact support. Let $f : \Re \to \Re$ be locally integrable. Let $p \in [1, \infty]$. Then as $h \to 0$:*

   (a) If f is continuous at x then $|(K_h * f)(x) - f(x)| \to 0$

   (b) If $p = \infty$ and f is uniformly continuous, OR $p < \infty$ and $f \in L^p$ then $\|K_h * f - f\|_p \to 0$

   (c) Let $\int_{\Re} u^r K(u)du = 0$, $r = 1, \ldots, s-1$, and $\kappa(s) = \int_{\Re} |u^s K(u)|du$. Suppose for $M > 0$, f is s-times differentiable and $\left\|f^{(s)}\right\|_p \leq M$. If $p = \infty$ OR $f^{(s-1)}$ is absolutely continuous then

   $$\|K_h * f - f\|_p \leq \frac{M\kappa(s)}{s!}h^s$$

   Proof. By assumptions on the Kernel, K, we know that $\exists C > 0$ such that $K(x) \leq C$ for all $x \in \Re$. Moreover, by compactness, $\exists R > 0$ such that $supp(K) \subset [-R, R]$. Finally, since f is locally integrable, $K_h * f$ exists.

   (a) The last step follows from continuity at x.

   $$\begin{aligned}
   |(K_h * f)(x) - f(x)| &= \left| \int h^{-1}K(\frac{x-y}{h})f(y)dy - f(x) \right| \\
   &= \left| \int K(u)f(x-uh)du - f(x)\int K(u) \right| \\
   &= \left| \int K(u)[f(x-uh) - f(x)]du \right| \\
   &\leq \int |K(u)||f(x-uh) - f(x)|du \\
   &\leq \left(\int |K(u)| \right) \left(\sup_{|u| \leq R} |f(x-uh) - f(x)| \right) \\
   &\leq 2RC \left(\sup_{|u| \leq R} |f(x-uh) - f(x)| \right) \to 0
   \end{aligned}$$

(b) For $p = \infty$, the result follows from (a) since f is uniformly continuous.

For $p \leq \infty$, we start with $p = 1$ The remaining p follow from Minkowski's integral inequality.

$$\begin{aligned}
\int_{\Re} |(K_h * f)(x) - f(x)|dx &= \int_{\Re} \left| \int_{-R}^{R} K(u)[f(x-uh) - f(x)]du \right| dx \\
&\leq \int_{\Re} \int_{-R}^{R} |K(u)[f(x-uh) - f(x)]|dudx \\
&\leq \int_{\Re} \left(\int |K(u)| \right) \left(\sup_{|u| \leq R} |f(x-uh) - f(x)| \right) dx \\
&= \left(\int |K(u)| \right) \int_{\Re} \left(\sup_{|u| \leq R} |f(x-uh) - f(x)| \right) dx \\
&\leq 2RC \int_{\Re} \left(\sup_{|u| \leq R} |f(x-uh) - f(x)| \right) dx \to 0
\end{aligned}$$

The last integral tends to 0. See the example sheet.

(c) We first prove $p = \infty$ and then $p = 1$ (use Minkowski for the remaining ones). Taylor's theorem is required.

i. Taylor's Theorem(s):
For $y \in (x, x-uh)$:

$$f(x-uh) - f(x) = \sum_{r=1}^{s-1} \frac{(-h)^r u^r}{r!} f^{(r)}(x) + \frac{(-uh)^s}{s!} f^{(s)}(y)$$

For f absolutely continuous:

$$\begin{aligned}
f(x-uh) - f(x) = &\sum_{r=1}^{s-1} \frac{(-h)^r u^r}{r!} f^{(r)}(x) \\
&+ \int_0^{-uh} \frac{t^{s-1}}{(s-1)!} f^{(s)}(x-t)dt
\end{aligned}$$

ii. The result follows for $p = \infty$ since for all x we have:

$$\begin{aligned}
&|(K_h * f)(x) - f(x)| \\
&\quad = \left| \int_{-R}^{R} K(u) \left(\sum_{r=1}^{s-1} \frac{(-h)^r u^r}{r!} f^{(r)}(x) + \frac{(-uh)^s}{s!} f^{(s)}(y) \right) du \right| \\
&\quad = \left| \int_{-R}^{R} K(u) \frac{(-uh)^s}{s!} f^{(s)}(y) du \right| \\
&\quad \leq \frac{h^s}{s!} \left\| f^{(s)} \right\|_{\infty} \kappa(s)
\end{aligned}$$

iii. By absolute continuity we have:

$$\begin{aligned}
&\int_{\Re} |(K_h * f)(x) - f(x)| dx \\
&\quad = \int_{\Re} \left| \int_{-R}^{R} K(u) \left(\int_{0}^{-uh} \frac{t^{s-1}}{(s-1)!} f^{(s)}(x-t) dt \right) du \right| dx \\
&\quad \leq \int_{\Re} \int_{-R}^{R} \int_{0}^{uh} |K(u)| \left| \frac{t^{s-1}}{(s-1)!} \right| \left| f^{(s)}(x-t) \right| dt du dx \\
&\quad \leq \int_{-R}^{R} |K(u)| \left[\int_{0}^{uh} \frac{t^{s-1}}{(s-1)!} \left(\int_{\Re} \left| f^{(s)}(x-t) \right| dx \right) dt \right] du \\
&\quad \leq \left\| f^{(s)} \right\|_1 \frac{h^s}{s!} \int_{-R}^{R} |K(u) u^s| du \\
&\quad \leq \frac{h^s}{s!} \kappa(s) \left\| f^{(s)} \right\|_1
\end{aligned}$$

□

4. If the density is symmetric, it satisfies our conditions for $s = 1, 2$.

4.3 Approximation by Basis Functions

4.3.1 Haar Basis

1. Suppose we want to approximate a function in L^2. The fourier series is optimal, but may not converge pointwise. Instead, we consider the piecewise-constant Haar Basis.

2. Haar Basis and Approximation

   Definition 4.2. *Suppose f is a locally integrable function. It's Haar Approximation of resolution level j is*

   $$H_j(f)(x) = \sum_{k \in Z} c_k \varphi_{j,k}(x)$$

   where:

   (a) $\varphi_{j,k}(x) = 2^{j/2} \mathbf{1}\left[2^{-j}k, 2^{-j}(k+1)\right]$

   (b) $c_k = \langle f, \varphi_{j,k} \rangle$

3. An equivalent expression, which approximates f in terms of a basis not dependent on j is:

   $$H_j(f)(x) = \sum_{k \in Z} a_k \varphi_{0,k}(x) + \sum_{l=0}^{j-1} \sum_{k \in Z} b_{l,k} \psi_{l,k}(x)$$

   where:

   (a) $\varphi_{0,k}(x) = \mathbf{1}\left[k, k+1\right]$ and $a_k = \langle f, \varphi_{0,k} \rangle$

(b) $\psi_{l,k}(x) =$
$2^{\frac{l}{2}}\left(\mathbf{1}\left[2^{-l}k \le x \le 2^{-l}(\frac{1}{2}+k)\right] - \mathbf{1}\left[2^{-l}(\frac{1}{2}+k) \le x \le 2^{-l}(1+k)\right]\right)$

(c) $b_{l,k} = \langle f, \psi_{l,k}\rangle$

Derivation 4.1. *Let $V_j = span(\varphi_{j,k})$. Then V_j is the space of all L^2 functions which are constant on the intervals $(2^{-j}k, 2^{-j}(k+1))$, $\forall k \in Z$. Then we have that $V_j \supset V_{j-1}$. Letting $W_{j-1} = V_{j-1}^{\perp} \cap V_j$, we have $V_j = V_{j-1} \oplus W_{j-1}$. Note that $W_{j-1} = span(\psi_{j-1,k})$. Continuing in this fashion, $V_j = V_0 \oplus W_0 \oplus \cdots \oplus W_{j-1}$.*

4. "Consistency" and "Error" of Haar Basis

Proposition 4.2. *Let f be locally integrable function. Let $p \in [1, \infty]$. As $j \to \infty$:*

(a) If f is continuous at x, then $|H_j(f)(x) - f(x)| \to 0$

(b) If $p = \infty$ and f is uniformly continuous OR $p < \infty$ and $f \in L^p$ then $\|H_j(f) - f\|_p \to 0$

(c) Suppose f is differentiable. If $p = \infty$ OR $p < \infty$ and f is absolutely continuous, then $\|H_j(f) - f\|_p \le \frac{1}{2}\|f'\|_p 2^{-j}$

5. The Haar basis cannot make use of higher-order derivatives, so a basis which can, can be used to approximate functions with higher-order derivatives.

4.3.2 B-Spline Basis on Dyadic Break Points

1. For the B-spline Basis

Definition 4.3. *We have:*

(a) The B-spline scaling function of order 1: $N^1(x) = \mathbf{1}[0,1] * \mathbf{1}[0,1] = x\mathbf{1}[0,1] + (1-x)\mathbf{1}[1,2]$

(b) The B-spline basis of order 1: $\{\varphi_{j,k}(x) = N^1(2^j x - k) | k \in Z, j \in N_0\}$

(c) The B-spline scaling function of order r: $N^r(x) = \mathbf{1}[0,1] * \mathbf{1}[0,1] * \cdots * \mathbf{1}[0,1]$ (r times).

(d) The B-spline basis of order r: $\{\varphi_(j,k)(x) = N^r(2^j x - k) | k \in Z, j \in N_0\}$

2. "Consistency" and "Error" for B-spline

Proposition 4.3. *For $s \in N_0$, define the B-spline basis of order $s-1$ on dyadic break points. If $f \in L^\infty$ and s-times differentiable, then for $c_k \in \Re$ we have for some constant $\kappa_s > 0$:*

$$\left\| \sum_{k \in Z} c_k \varphi_{j,k} - f \right\|_\infty \le \kappa_s \left\| f^{(s)} \right\| 2^{-j}$$

3. Note that determining c_k is difficult since the basis is not orthonormal.

4. Note that the B-spline basis of order r is $r-1$ times differentiable.

4.4 Approximation by Orthonormal Wavelet Basis

1. The Ideal Basis
   (a) Localised, (unlike the fourier basis), so that we can approximate f pointwise
   (b) Smooth, (unlike Haar), so that we can approximate a smooth f
   (c) Orthogonal, (unlike B-spline), so that we can control variance of estimators for the constants
2. Ideal Father Wavelet (Scaling Function):
   (a) Bounded and compactly supported
   (b) Has orthonormal translates ($\varphi_{0,k} = \varphi(x-k)$) for all $k \in Z$
   (c) The linear space $V_j = span(\varphi_{j,k})$ are nested ($V_j \supset V_{j-1}$)
   (d) $\bigcup_{j=0}^{\infty} V_j$ is dense in L^2
3. Given the ideal father wavelet forming the ideal basis, we can do the following:
   (a) Decompose V_j so that $V_j = V_0 \oplus W_0 \oplus \cdots \oplus W_{j-1}$
   (b) Determine the mother wavelet for the spaces W_i, $\psi : \Re \to \Re$
   (c) Write the approximation as

$$W_j(f)(x) = \sum_{k \in Z} a_k \varphi_{0,k}(x) + \sum_{l=1}^{j-1} b_{l,k} \psi_{l,k}(x)$$

   where a_k and $b_{l,k}$ are determined by orthogonal projections of f onto V_0 and W_l.
4. "Consistency" and "Error" of the Orthonormal Wavelet Basis

Proposition 4.4. *Let $\{\varphi_{0,k}, \psi_{l,k} | k \in Z\}$ be a compactly supported orthonormal wavelet basis, which is either the Haar basis or has a continuous father wavelet. Also let $f : \Re \to \Re$ be locally integrable, $p \in [1, \infty]$, $s \in N_0$, and $\int_{\Re} u^r \psi(u) du = 0$ for $r = 0, \ldots, s-1$. As $j \to \infty$:*

(a) If f is continuous at x then $|W_j(f)(x) - f(x)| \to 0$

(b) If either $p = \infty$ and f is uniformly continuous OR $p < \infty$ and $f \in L^p$ then $\|W_j(f) - f\|_p \to 0$

(c) Suppose f is s-times differentiable. If $p = \infty$ OR $f^{(s-1)}$ is absolutely continuous then $\|W_j(f) - f\|_p \leq \frac{\kappa_s}{s!} \left\|f^{(s)}\right\|_p 2^{-js}$

Proof. The wavelet estimator is essentially a moving kernel estimator. Let $K(x,y) = \sum_{k \in Z} \varphi(x-k)\varphi(y-k)$. Since $\varphi(x)$ are bounded and compactly supported, so is $K(x,y)$. Moreover, $\int_{\Re} K(x,y) dy = 1$ (see example sheet). Therefore:

$$W_j(f)(x) = \int_{\Re} 2^j f(2^j x, 2^j y) f(y) dy$$

(a) Since f is continuous at x, and K is bounded and compactly supported: (equivalent proof for convolution with kernel)

$$\begin{aligned}|W_j(f)(x) - f(x)| &= \left| \int_{\Re} K(2^j x, x - 2^{-j}u) \left(f(x - 2^{-j}u) - f(x) \right) du \right| \\ &\leq \sup_{u \in [-R,R]} |f(x - 2^{-j}u) - f(x)| C \to 0\end{aligned}$$

(b) This also holds based on the equivalent proof for convolution with a stationary kernel.

(c) We need several things to use the proof for convolution with a stationary kernel:

i. $\int_{\Re} K(x,y)(y-x)^r dy = 0$ for $r = 1, \ldots, s-1$. For the Haar basis, this holds vacuously. For higher-order orthonormal wavelet bases, see the example sheet.

ii. Also, since K is periodic, we let $\kappa_s = \sup_{x \in (0,1)} \int_{\Re} |u^s K(x, x-u)| du$, which is finite.

iii. Note for the Haar Basis:

$$\begin{aligned}\kappa_s &= \sup_{x \in (0,1)} \int_{\Re} \sum_{k \in Z} u \mathbf{1}\left[x \in [k, k+1]\right] \mathbf{1}\left[x - u \in [k, k+1]\right] du \\ &= \sup_{x \in (0,1)} \int_{\Re} u \mathbf{1}\left[x \in [0,1]\right] \mathbf{1}\left[x - u \in [0,1]\right] du, \text{ so when x=1:} \\ &= \int_{\Re} u \mathbf{1}\left[1 - u \in [0,1]\right] du \\ &= \int_0^1 u du = \frac{1}{2}\end{aligned}$$

□

5. An example of such wavelets is Daubechies' wavelets.

5 Density Estimation

5.1 Motivation

There is no natural estimator for densities given $X_1, X_2, \ldots, X_n$ independent samples with density f. So we approximate using Kernels and wavelets. This gives us many approximation choices, smoothness choices (i.e. resolution level for wavelets or bandwidth for kernels), and loss measurements (e.g. pointwise, L^p, etc.).

5.2 Kernel Density Estimation

5.2.1 Consistency and Error

1. Framework: notice that

$$(K_h * f)(x) = \int_{\Re} K_h(x-y) f(y) dy = \mathbf{E}_f \left[K_h(x - X) \right]$$

Therefore, we have the natural kernel density estimator, based on the law of large numbers:

$$\hat{f}^k_{n,h} = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

2. "Consistency" of Kernel Density Estimator

Theorem 5.1. *Let $X_1, \ldots, X_n$ be I.i.d. R.V. taking values in $\Re$ with density f, and let $\hat{f}^k_{n,h}$ be the Kernel Density Estimator. Suppose:*

(a) h is chosen in terms of n, so that $h \to 0$, $nh \to \infty$ as $n\infty$

(b) K_h is non-negative, bounded, compactly supported and integrates to 1

Then as $n \to \infty$, $\boldsymbol{E}_f\left[\left\|\hat{f}^k_{n,h} - f\right\|_1\right] \to 0$

Proof. We will first split the term of interest into variance and bias.

(a) $\mathbf{E}\left[\left\|\hat{f}^k_{n,h} - f\right\|_1\right] \leq \mathbf{E}\left[\left\|\hat{f}^k_{n,h} - (K_h * f)\right\|_1\right] + \mathbf{E}\left[\|(K_h * f) - f\|_1\right]$

(b) The bias term is deterministic, and by Proposition 4.1(b), as $n \to \infty$, $h \to 0$ and so $\|(K_h * f) - f\|_1 \to 0$

(c) For the variance term, we first need to demonstrate several properties:

i. $\mathbf{E}\left[\hat{f}^k_{n,h}(x)\right] = \frac{1}{n}\sum_{i=1}^{n}\int_{\Re} K_h(x-y)f(y)dy = (K_h * f)(x)$

ii. By Fubini's Theorem and assumption (b)

$$\begin{aligned}\int_{\Re} \hat{f}^k_{n,h}(x) - (K_h * f)(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{\Re} K_h(x - X_i)dx \\ &\quad - \int_{\Re}\int_{\Re} K_h(x-y)f(y)dydx \\ &= 1 - \int\int K_h(x-y)dx f(y)dy \\ &= 1 - \int f(y)dy = 0\end{aligned}$$

Therefore,

$$\int_{\Re}\left(\hat{f}^k_{n,h}(x) - (K_h * f)(x)\right)_+ = \int_{\Re}\left(\hat{f}^k_{n,h}(x) - (K_h * f)(x)\right)_-$$

iii. By Jensen's Inequality:

$$\mathbf{E}\left[\left(\hat{f}^k_{n,h}(x) - (K_h * f)(x)\right)_+\right]^2 \leq \mathbf{E}\left[\left(\hat{f}^k_{n,h}(x) - (K_h * f)(x)\right)^2\right]$$

iv. Since K is bounded by some $\sqrt{C}$ with compact support in some $[-R, R]$:

$$\begin{aligned}
\mathbf{E}\left[\left(\hat{f}_{n,h}^K\right)^2\right] &= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[K_h^2(x - X)\right] \\
&= \frac{1}{n}\int_{\Re} K_h^2(x-y)f(y)dy \\
&= \frac{1}{nh}\int_{\Re} K^2(u)f(x-uh)du \\
&\leq \frac{C}{nh}\int_{-R}^{R} f(x-uh)du
\end{aligned}$$

By Lebesgue differentiation theorem, the integral ends to $f(x)$ as $h \to 0$. Therefore, as $nh \to \infty$, the entire integral tends to 0.

(d) We have by Fubini's Theorem:

$$\begin{aligned}
\mathbf{E}\left[\left\|\hat{f}_{n,h}^K - (K_h * f)\right\|_1\right] &= \int_{\Re} \mathbf{E}\left[\left|\hat{f}_{n,h}^K - (K_h * f)\right|\right] dx \\
&= 2\int_{\Re} \mathbf{E}\left[\left(\hat{f}_{n,h}^k(x) - (K_h * f)(x)\right)_+\right] dx
\end{aligned}$$

By Property (ii) and (iii)

$$\begin{aligned}
&\leq 2\int_{\Re} \min\left((K_h * f)(x), \mathbf{E}\left[\left(\hat{f}_{n,h}^K(x) - (K_h * f)(x)\right)_+\right]\right) dx \\
&\leq 2\int_{\Re} \min\left((K_h * f)(x), \sqrt{\mathbf{E}\left[\left(\hat{f}_{n,h}^K(x) - (K_h * f)(x)\right)^2\right]}\right) dx \\
&\leq 2\int_{\Re} \min\left(f(x), \sqrt{\mathbf{E}\left[\left(\hat{f}_{n,h}^K(x)\right)^2\right]}\right) + 2\left\|K_h * f - f\right\|_1
\end{aligned}$$

By Proposition 4.1(b) the term on the right tends to 0. By Property (iv) the integrand will tend pointwise to 0. Since the integrand is dominated by $f(x)$, the dominated convergence theorem gives that the integral tends to 0.

□

3. Although the estimator converges, it is not "optimal," in the sense that it has the worst possible error over the class of all possible densities.

Proposition 5.1. *Let $X_1, \ldots, X_n$ be i.i.d. R.V. with density f. Let $\hat{f}_{n,h}^K$ be the Kernel Density Estimator. Suppose K is non-negative and integrates to 1, then:*

(a) $\sup_f \inf_{h>0} \boldsymbol{E}_f \left[\left\| \hat{f}^K_{n,h} - f \right\|_1 \right] = 2$

4. "Error" of Kernel Density Estimator over class of smooth functions with bounded derivatives

Proposition 5.2. *Let $X_1, \ldots, X_n$ be i.i.d. R.V. with density f. Let $\hat{f}^K_{n,h}$ be the Kernel Density Estimator. Let $s \in N$. Suppose that K satisfies the conditions of Proposition 4.1(c) and f is s-times differentiable. Then:*

(a) If $x \in \Re$ then

$$\boldsymbol{E}\left[|\hat{f}^K_{n,h}(x) - f(x)|\right] \leq \|K\|_2 \sqrt{\frac{\|f\|_\infty}{nh}} + \frac{\kappa_s}{s!} \left\|f^{(s)}\right\|_\infty h^s$$

(b) If $f^{(s-1)}$ is absolutely continuous then

$$\boldsymbol{E}\left[\left\|\hat{f}^K_{n,h} - f\right\|_2\right] \leq \|K\|_2 \sqrt{\frac{1}{nh}} + \frac{\kappa_s}{s!} \left\|f^{(s)}\right\|_\infty h^s$$

Proof. Again we split the terms into bias and variance components. The bias can be bounded by Proposition 4.1(c), since all the conditions are satisfied.

(a) We use the Proof of Theorem 5.1, Property (iv):

$$\begin{aligned}
\mathbf{E}\left[\left|\hat{f}^K_{n,h}(x) - (K_h * f)(x)\right|\right]^2 &\leq \mathbf{E}\left[\left(\hat{f}^K_{n,h}(x) - (K_h * f)(x)\right)^2\right] \\
&\leq \mathbf{E}\left[\left(\hat{f}^K_{n,h}\right)^2\right] \\
&\leq \frac{1}{nh}\int_{\Re} K^2(u) f(x-uh)du \\
&\leq \frac{\|f\|_\infty}{nh}\|K\|_2^2
\end{aligned}$$

(b) Again, we use the Proof of Theorem 5.1, Property (iv) along with Fubini:

$$\begin{aligned}
\mathbf{E}\left[\left\|\hat{f}^K_{n,h} - (K_h * f)\right\|_2\right]^2 &\leq \mathbf{E}\left[\left\|\hat{f}^K_{n,h} - (K_h * f)\right\|_2^2\right] \\
&= \int_{\Re} \mathbf{E}\left[\left(\hat{f}^K_{n,h}(x) - (K_h * f)(x)\right)^2\right] dx \\
&\leq \int_{\Re} \frac{1}{nh}\int_{\Re} K^2(u) f(x-uh)dudx \\
&= \frac{1}{nh}\int_{\Re} K^2(u) \int_{\Re} f(x-uh)dxdu \\
&= \frac{1}{nh}\|K\|_2^2
\end{aligned}$$

□

5.2.2 Asymptotic Behaviour of Kernel Density Estimator

1. By Proposition 5.2, the error in the Kernel Density Estimator is bounded by two competing terms, and is optimised when the terms are equal:

$$h \sim \left(\frac{\|f\|_\infty \|K\|_2^2}{n \left\|f^{(s)}\right\|_\infty^2} \right)^{1/(2s+1)}$$

However, we cannot estimate the smoothness s from the data. So, if we want to do inference, we must look at the behaviour of $\hat{f}_{n,h}^K$ as n gets large in distribution.

2. Asymptotically, the Kernel Density Estimator is normally distributed.

Proposition 5.3. *Let $X_1, \ldots, X_n$ be i.i.d. R.V. with density f. Let f be the Kernel Density Estimator. Suppose that:*

(a) h is chosen so that as $n \to \infty$, $nh^{2s+1} \to 0$ and $nh \to \infty$

(b) K satisfies the conditions of Proposition 4.1(c).

(c) f is s-times differentiable with $f, f^{(s)} \in L^\infty$

Then, $\sqrt{nh}\left(\hat{f}_{n,h}^K(x) - f(x)\right) \xrightarrow{d} N\left(0, f(x)\|K\|_2^2\right)$

Proof. We first do our standard bias-variance decomposition. We use Proposition 4.1(c) to show that the bias term tends to 0. Then we use Lindeberg-Feller CLT to determine the distribution of the remaining variance term.

(a) $\sqrt{nh}\left(\hat{f}_{n,h}^K(x) - f(x)\right)$
$= \sqrt{nh}\left(\hat{f}_{n,h}^K(x) - (K_h * f)(x)\right) + \sqrt{nh}\left((K_h * f)(x) - f(x)\right)$

(b) By Proposition 4.1(a) and assumption (a) the bias term

$$\sqrt{nh}\left((K_h * f)(x) - f(x)\right) \leq \frac{M\kappa_s}{s!}\sqrt{nh^{2s+1}} \to 0$$

(c) To use Lindeberg-Feller CLT, we need to first choose a Y_{ni}, demonstrate that $\sum_{i=1}^n Y_{ni} - \mathbf{E}\left[Y_{ni}\right] = \sqrt{nh}(\hat{f}_{n,h}^K(x) - f(x))$, determine it's variance (hence property (b)), and demonstrate property (a) of the Lindeberg-Feller CLT.

i. Let $Y_{ni} = \frac{1}{\sqrt{nh}} K\left(\frac{x - X_i}{h}\right)$

ii. Its sum, centred about its mean is:

$$\begin{aligned}
&\sum_{i=1}^n \frac{1}{\sqrt{nh}} K\left(\frac{x - X_i}{h}\right) - \mathbf{E}\left[\frac{1}{\sqrt{nh}} K\left(\frac{x - X_i}{h}\right)\right] \\
&= \sqrt{nh}\frac{1}{n}\sum_{i=1}^n K_h(x - X_i) - \sqrt{nh}\int \frac{1}{h} K(\frac{x-y}{h}) f(y) dy \\
&= \sqrt{nh}\left(\hat{f}_{n,h}^K(x) - (K_h * f)(x)\right)
\end{aligned}$$

iii. The variance, $n\mathbf{Cov}\left[Y_{n,i}\right] = n\mathbf{E}\left[Y_{ni}^2\right] - n\left(\mathbf{E}\left[Y_{ni}\right]\right)^2$

A. For the first term, as $h \to 0$, will by continuity and dominated convergence:

$$\begin{aligned} n\mathbf{E}\left[Y_{ni}^2\right] &= \frac{1}{h}\int K^2\left(\frac{x-y}{h}\right) f(y)dy \\ &= \int K^2(u) f(x-uh)du \\ &\to f(x)\left\|K\right\|_2^2 \end{aligned}$$

B. For the second term, by Cauchy-Schwarz, as $h \to 0$

$$\begin{aligned} n(\mathbf{E}\left[Y_{ni}\right])^2 &= h\left(\int_{\Re} K(u) f(x-uh)du\right)^2 \\ &\leq h\left\|f\right\|_\infty^2 \left\|K\right\|_1^2 \\ &\to 0 \end{aligned}$$

C. Therefore, the variance term is $n\mathbf{Cov}\left[Y_{n,i}\right] = f(x)\left\|K\right\|_2^2$

iv. To prove property (a) of the Lindeberg-Feller CLT, we note that $|Y_{n,i}| \leq \frac{1}{\sqrt{nh}}\left\|K\right\|_\infty \to 0$ as $n \to \infty$.

(d) The result follows by the Lindeberg-Feller CLT.

□

3. Lindeberg-Feller Central Limit Theorem

Theorem 5.2. *For each n, let $Y_{n,1}, \ldots, Y_{n,n}$ be i.i.d. R.V. with finite variance. As $n \to \infty$ suppose that:*

(a) $\forall \epsilon > 0$, $n\boldsymbol{E}\left[Y_{n,i}^2 \boldsymbol{1}\left[|Y_{n,i}| > \epsilon\right]\right] \to 0$

(b) $n\boldsymbol{Cov}[Y_{n,i}] \to \sigma^2$

Then $\sum_{i=1}^n \left(Y_{n,i} - \boldsymbol{E}[Y_{n,i}]\right) \xrightarrow{d} N(0, \sigma^2)$

4. We can use the result of Proposition 5.3 to construction point-wise confidence intervals.

5.3 Histogram Density Estimation

1. A simple estimator of density divides the real line into bins $I_k = [x_k, x_{k+1}]$ and estimates the density f by the proportion of observations in each I_k

Definition 5.1. *The Histogram Density Estimator is:*

$$\hat{f}^H(x) = \sum_{k \in Z}\left(\frac{1}{n(x_{k+1} - x_k)}\sum_i \boldsymbol{1}\left[X_i \in I_k\right]\right)\boldsymbol{1}\left[x \in I_k\right]$$

2. We can use dyadic break points, of resolution j, giving the following estimator instead:

$$\begin{aligned}
&\hat{f}^H_{n,j}(x) \\
&= \sum_{k\in Z}\left(2^j\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[X_i\in\left[2^{-j}k,2^{-j}(k+1)\right]\right]\right)\mathbf{1}\left[x\in\left[2^{-j}k,2^{-j}(k+1)\right]\right] \\
&= \sum_{k}\left(2^{\frac{j}{2}}\frac{1}{n}\sum_{i}\mathbf{1}\left[X_i\in\left[2^{-j}k,2^{-j}(k+1)\right]\right]\right)2^{\frac{j}{2}}\mathbf{1}\left[x\in\left[2^{-j}k,2^{-j}(k+1)\right]\right]
\end{aligned}$$

3. Notice that the histogram density estimator with dyadic break points is the Haar approximation to a function with the c_k estimated by

$$\hat{c}_k = 2^{\frac{j}{2}}\frac{1}{n}\sum_{i}\mathbf{1}\left[X_i\in\left[2^{-j}k,2^{-j}(k+1)\right]\right]$$

Therefore, its expectation is the Haar approximation to the function f.

4. "Error" of the Kernel Density Estimator

Proposition 5.4. *Let $X_1,\ldots,X_n$ be i.i.d. R.V. with density f. Let $\hat{f}^H_{n,j}$ be the histogram estimator. Suppose f is once differentiable. Then:*

(a) If $x\in\Re$ then

$$\boldsymbol{E}\left[\left|(x)-\hat{f}^H_{n,j}\right|\right]\le\sqrt{\frac{\|f\|_\infty}{n2^{-j}}}+\frac{1}{2}\|f'\|_\infty 2^{-j}$$

(b) If f is absolutely continuous, then

$$\boldsymbol{E}\left[\left\|f-\hat{f}^H_{n,j}\right\|_2\right]\le\sqrt{\frac{1}{n2^{-j}}}+\frac{1}{2}\|f'\|_2 2^{-j}$$

5.4 Wavelet Density Estimation

1. Motivation: as in histogram density estiamtion, we can use smoother wavelets to approximate smoother functions. To do so, we must approximate the coefficients of $W_j(f)(x)=\sum_{k\in Z}a_{j,k}\varphi_{j,k}(x)$.

Definition 5.2. *Noting that $a_j,k=\langle f,\varphi_{j,k}\rangle=\boldsymbol{E}[\varphi_j,k(x)]$, the estimators for our coefficients are $\hat{a}_{j,k}=\frac{1}{n}\sum_{i=1}^{n}\varphi_{j,k}(X_i)$, and the wavelet density estimator is:*

$$\hat{f}^W_{n,j}(x)=\sum_{k\in Z}\hat{a}_{j,k}\varphi_{j,k}(x)$$

2. "Error" of Wavelet Density Estimators

Proposition 5.5. *Let $X_1,\ldots,X_n$ be i.i.d. R.V. with density f. Let $\hat{f}^W_{n,j}$ be the wavelet density estimator. Let $s\in N$. Suppose f is s-times differentiable, and the wavelet basis satisfies the conditions of Proposition 4.4.*

(a) If $x \in \Re$ then

$$\boldsymbol{E}\left[\left|\hat{f}_{n,j}^{W}(x) - f(x)\right|\right] \leq \kappa_s \left[\sqrt{\frac{\|f\|_\infty}{n2^{-j}}} + \frac{1}{2}\left\|f^{(s)}\right\|_\infty 2^{-js}\right]$$

(b) If $f^{(s-1)}$ is absolutely continuous then

$$\boldsymbol{E}\left[\left\|\hat{f}_{n,j}^{W} - f\right\|_2\right] \leq \kappa_s \left[\sqrt{\frac{1}{n2^{-j}}} + \frac{1}{2}\left\|f^{(s)}\right\|_2 2^{-js}\right]$$

Proof. Again we split the terms into variance and bias. The bias term $W_j(f)(x) - f(x)$ is bounded by Proposition 4.4(c).

(a) Notice that $\hat{f}_{n,j}^{W}(x) = \frac{1}{n}\sum_i Z_i = \frac{1}{n}\sum_i 2^j K(2^j x, 2^j X_i)$ where

$$K(x,y) = \sum_{k \in Z} \varphi(x-k)\varphi(y-k)$$

(b) Then, by i.i.d. of Z_i and boundedness/compactness of K:

$$\begin{aligned}
\mathbf{E}\left[\left(\hat{f}_{n,j}^{W}(x) - W_j(f)(x)\right)^2\right] &\leq \mathbf{E}\left[\left(\hat{f}_{n,j}^{W}(x)\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}\left[Z_i^2\right] \\
&= \frac{1}{n}\mathbf{E}\left[Z^2\right] \\
&= \frac{1}{n}\int_{-R}^{R} 2^{2j} K^2(2^j x, 2^j y) f(y) dy \\
&= \frac{2^j}{n}\int_{-R}^{R} K^2(2^j x, 2^j x - u) f(x - 2^{-j}u) du \\
&\leq \frac{1}{n2^{-j}}\kappa_s
\end{aligned}$$

(c) For the L^2 norm, we will have a double integral. But this is also bounded by some constant since K is non-zero on a a compact support and bounded.

□

6 Nonparametric Regression

6.1 Introduction

In typical regression we are given $(X_1, Y_1), \ldots, (X_n, Y_n)$ and are asked to estimate the relationship between X_i and Y_i. There are two approaches, of which we mainly consider fixed design.

1. Fixed Design: We assume that the design points X_i are fixed, and known in advance, denoted x_i. We want to estimate the function m where $Y_i = m(x_i) + \epsilon_i$, where ϵ_i are independent, zero-mean random variables.

2. Random Design: We assume that the pairs (X_i, Y_i) are i.i.d. and want to estimate the mean function $m(x) = \mathbf{E}[Y|X_i = x]$.

6.2 Kernel Estimation

1. One way of estimating m is by local averaging (e.g. kernel convolution) of response variables around a given x. Formally, this is the Nadaraya-Watson Estimator.

2. Consider the form of the mean estimator:

$$m(x) = \frac{\int yf(x,y)dy}{\int f(x,y)dy}$$

Then a natural estimator (the Nadayara-Watson Estimator) for f is the Nadayara-Watson Estimator:

$$\hat{m}_{n,h}^K = \frac{\sum_i K_h(x - x_i)Y_i}{\sum_i K_h(x - x_i)}$$

where if the denominator is 0, we set the estimator to 0.

3. "Error" of the Nadayara-Watson Estimator.

Theorem 6.1. *Let $Y_1, \ldots, Y_n$ be independent R.V. with $\boldsymbol{E}[Y_i] = m(x_i)$, and variance $v(x_i)$. Let $x_i = \frac{i}{n}$ for $i = 1, \ldots, n$. Let $\hat{m}_{n,h}^K$ be the Nadayara-Watson Estimator. Suppose*

(a) As $n \to \infty$, $h \to 0$

(b) K satisfies the conditions of Proposition 4.1(c), K is absolutely continuous, K is differentiable and $K' \in L^1$.

(c) Suppose m is s-times differentiable

(d) Suppose $v(x_i)$ are all abounded by $\sigma^2 > 0$.

Then for $x \in (0, 1)$,

$$\boldsymbol{E}\left[|\hat{m}_{n,h}^K - m(x)|\right] \leq \frac{\sigma \|K\|_2}{\sqrt{nh}} + \frac{\kappa_s}{s!}\left\|m^{(s)}\right\|_\infty h^s + O\left(\frac{1}{nh}\right)$$

Proof. Denote the numerator of the estimator as $ng_{n,h}(x)$ and the denominator as $nf_{n,h}(x)$). We decompose the error into three terms, variance, bias and denominator control. The key to many of the bounds is first extending the sum to an integral, using Reimann Sums,, incurring an error, and then from an integral from 0 to 1 to all of $\Re$, incurring another error.

(a) $|\hat{m}_{n,h}^K(x) - m(x)| \leq |\hat{m}_{n,h}^K(x) - g_{n,h}(x)| + |g_{n,h}(x) - \mathbf{E}[g_{n,h}(x)]| + |\mathbf{E}[g_{n,h}(x)] - m(x)|$

(b) We first look at $g_{n,h}$ and $f_{n,h}$ individually:

$$\begin{aligned}
\mathbf{E}\left[g_{n,h}(x)\right] &= \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)\mathbf{E}\left[Y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)m(x_i) \\
&= \int_0^1 h^{-1}K\left(\frac{x-y}{h}\right)m(y)dy + O\left(\frac{1}{nh}\right) \\
&= \int_{(x-1)/h}^{x/h} K\left(u\right)m(x-uh)du + O\left(\frac{1}{nh}\right) \\
\text{As } h\to 0 &= \int_{-\infty}^{\infty} K(u)m(x-uh)du + O\left(\frac{1}{nh}\right) \\
&= (K_h * m)(x) + O\left(\frac{1}{nh}\right)
\end{aligned}$$

$$\begin{aligned}
\mathbf{Cov}\left[g_{n,h}(x)\right] &= \mathbf{E}\left[\left(\frac{1}{n}\sum_i K_h(x-x_i)(Y_i - m(x_i))\right)^2\right] \\
\text{independence} &= \frac{1}{n^2}\sum_i K^2(x-x_i)\mathbf{Cov}\left[Y_i\right] \\
&= \frac{1}{n^2}\sum_i K^2(x-x_i)v(x_i) \\
&\le \frac{\sigma^2}{(nh)^2}\sum_i K^2\left(\frac{x-x_i}{h}\right) \\
&= \frac{\sigma^2}{nh}\int_0^1 \frac{1}{h}K^2\left(\frac{x-y}{h}\right)dy + O\left(\frac{1}{(nh)^2}\right) \\
&= \frac{\sigma^2}{nh}\int_{\Re} K^2(u)du + O((nh)^{-2}) \\
&= \frac{\sigma^2}{nh}\left\|K\right\|_2^2 + O((nh)^{-2})
\end{aligned}$$

$$\begin{aligned}
f_{n,h}(x) = \frac{1}{n}\sum_i K_h(x-x_i) &= \int_0^1 K_h(x-y)dy + O((nh)^{-1}) \\
&= \int_{\Re} K(u)du + O((nh)^{-1}) \\
&= 1 + O((nh)^{-1})
\end{aligned}$$

(c) Now we look at each term:

i. Bias: $\left|\mathbf{E}\left[g_{n,h}(x)\right] - m(x)\right| = \left|(K_h * m)(x) - m(x)\right| + O((nh)^{-})$ which is bounded by Proposition 4.1(c).

ii. Variance: By Jensen, $\mathbf{E}\left[\left|g_{n,h}(x) - \mathbf{E}\left[g_{n,h}\right]\right|\right] \le \sqrt{\mathbf{Cov}\left[g_{n,h}\right]} \le \frac{\sigma}{\sqrt{nh}}\left\|K\right\|_2 + O((nh)^{-})$

iii. Denominator: Since $m(x)$ is continuous on a compact set, it is bounded. And we have bounded the other terms in the penultimate line, therefore:

$$\begin{aligned}\mathbf{E}\left[|\hat{m}(x) - g_{n,h}(x)|\right] &= \mathbf{E}\left[\left|g_{n,h}(x)\frac{1 - f_{n,h}(x)}{f_{n,h}(x)}\right|\right] \\ &= O((nh)^{-})\mathbf{E}\left[|g_{n,h}(x)|\right] \\ &\leq O((nh)^{-})\{\mathbf{E}\left[|g_{n,h}(x) - \mathbf{E}\left[g_{n,h}(x)\right]|\right] \\ &+ \mathbf{E}\left[|\mathbf{E}\left[g_{n,h}(x)\right] - m(x)|\right] + |m(x)|\} \\ &= O((nh)^{-})\end{aligned}$$

□

4. If we choose a bandwidth $h \sim Cn^{-1/(2s+1)}$, we attain the optimal convergence rate $n^{-s/(2s+1)}$ just as we did for Kernel density estimation.

5. The estimator does not perform well at the boundaries. We can fix this within the Nadaraya-Watson framework, but we will instead consider local polynomials.

6.3 Local Polynomials

1. Framework: Let $s \in N$ and $Y = m(x) + \epsilon$, where m is s-times differentiable

   (a) Let:

$$U(t) = \left[1, t, \ldots, \frac{t^{s-1}}{(s-1)!}\right]^T$$

$$M(x) = \left[m(x), hm'(x), \ldots, h^{s-1}m^{(s-1)}(x)\right]^T$$

   (b) By Taylor's Theorem, we can compute $m(y)$, where y is in the neighbourhood of x by:

$$m(y) = U^T\left(\frac{y-x}{h}\right)M(x) + O\left(\left\|m^{(s)}\right\|h^s\right)$$

2. Estimator: For $K \in L^1$, non-negative and $\int K = 1$, and fixed effects (x_i, Y_i):

   (a) We begin by estimating M by

$$\hat{M}(x) \in \arg\min_{M \in \Re^s} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\left[Y_i - U^T\left(\frac{x_i - x}{h}\right)M\right]^2$$

   (b) Local polynomial estimator of the mean function:

   Definition 6.1. *Given $\hat{M}(x)$ and $U(t)$ as defined above, and $s \in N$, the local polynomial estimator of order $s-1$ for the mean function $m(x)$ is:*

$$\hat{m}^P_{n,h}(x) = U^T(0)\hat{M}(x)$$

   And when $s > 1$, we can estimate the other derivatives as well.

Remark 6.1. *When $s = 1$, this is the Nadaraya-Watson estimator with kernel K.*

3. We can rewrite (see example sheet) the estimator as

$$\hat{m}^P_{n,h}(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i$$

where:

$$W_{n,i}(x) = \frac{1}{nh} K\left(\frac{x_i - x}{h}\right) U^T(0) B^{-1} U\left(\frac{x_i - x}{h}\right)$$

$$B = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) U\left(\frac{x_i - x}{h}\right) U^T\left(\frac{x_i - x}{h}\right)$$

   (a) Note that B must be invertible. This is asymptotically guaranteed when the design is uniformly spaced. In fact B will be positive definite.

   (b) For $s > 1$ the weights $(W_{n,i})$ can be thought of as adapting to the boundaries, removing any boundary difficulties.

4. "Error" of the Local Polynomial Estimator

Proposition 6.1. *Let $Y_1, Y_2, \ldots, Y_n$ be independent R.V. with Y_i having a mean $m(x_i)$ and variance $v(x_i)$ for $x_i = \frac{i}{n}$. Let $\hat{m}^P_{n,h}$ be the local polynomial estimator of order $S - 1$ for $S \in N$. Suppose that:*

(a) The smallest eigenvalue of B is $\lambda > 0$

(b) K is bounded and compactly supported

(c) m is s-times differentiable for $s \leq S$.

(d) $v(x)$ are bounded by $\sigma^2 > 0$.

Then for $x \in [0, 1]$ and $n \geq \frac{1}{h}$, we have:

$$\boldsymbol{E}\left[\left|\hat{m}^P_{n,h}(x) - m(x)\right|\right] \leq \kappa_s \left(\frac{\sigma}{\sqrt{nh}} + \left\|m^{(s)}\right\|_\infty h^s\right)$$

Proof. We again decompose into bias and variance, and bound each component separately:

$$\left|\hat{m}^P_{n,h}(x) - m(x)\right| \leq \left|\hat{m}^P_{n,h}(x) - \mathbf{E}\left[\hat{m}^P_{n,h}(x)\right]\right| + \left|\mathbf{E}\left[\hat{m}^P_{n,h}(x)\right] - m(x)\right|$$

   (a) First we need some properties of $W_{n,i}(x)$

      i. (Second Example Sheet) For $r = 1, \ldots, s - 1$

$$\sum_i W_{n,i}(x)(x_i - x)^r = \mathbf{1}\left[r = 0\right]$$

ii. Since $supp(K) \subset [-R, R]$:

$$\begin{aligned}
|W_{n,i}(x)| &\leq \frac{1}{nh} \left\|B^{-1}\right\| \left\|U\left(\frac{x_i - x}{h}\right)\right\| \|K\|_\infty \mathbf{1}\left[\left|\frac{x_i - x}{h}\right| \leq R\right] \\
&\leq \frac{1}{nh\lambda} \left\|U\left(\frac{x_i - x}{h}\right)\right\| \|K\|_\infty \mathbf{1}\left[\left|\frac{x_i - x}{h}\right| \leq R\right] \\
&\leq \frac{C}{nh} \mathbf{1}\left[|x_i - x| < Rh\right]
\end{aligned}$$

By assumption, $nh > 1$, so

$$\sum_i |W_{n,i}(x)| < \frac{C}{nh} \sum_i \mathbf{1}\left[|x_i - x| < Rh\right] < C'$$

(b) The bias term, we use Taylor Expansion, and the first property about $W_{n,i}(x)$ to remove all terms $(x_i - x)^r$ for $r = 1, \ldots, s-1$. Then, we use the second property. Then, we use $nh > 1$:

$$\begin{aligned}
|\mathbf{E}\left[\hat{m}\right] - m| &= \left|\sum_i W_{n,i}(x) m(x_i) - m(x)\right| \\
&= \left|\sum_i W_{n,i}(x) m(x_i) - \left[\sum_i W_{n,i}(x)\right] m(x)\right| \\
&= \left|\sum_i W_{n,i}(x) \left[m(x_i) - m(x)\right]\right| \\
&\leq \left|\sum_i W_{n,i}(x) \frac{C' \left\|m^{(s)}\right\|_\infty}{s!} |x_i - x|^s\right| \\
&\leq \frac{C'C}{nh} \frac{\left\|m^{(s)}\right\|_\infty}{s!} \sum_i \mathbf{1}\left[|x_i - x| \leq Rh\right] |x_i - x|^s \\
&\leq C'C \frac{\left\|m^{(s)}\right\|_\infty}{s!} \sum_i \mathbf{1}\left[|x_i - x| \leq Rh\right] |x_i - x|^s \\
&\leq C'' \left\|m^{(s)}\right\|_\infty h^s
\end{aligned}$$

(c) For the variance term, we use Jensen's and bound $\mathbf{Cov}\left[\hat{m}\right]$:

$$\begin{aligned}
\mathbf{E}\left[\left(\hat{m}(x) - \mathbf{E}\left[\hat{m}(x)\right]\right)^2\right] &= \mathbf{E}\left[\left(\sum_i W_{n,i}(x) \left[Y_i - m(x_i)\right]\right)\right] \\
&= \sum_i W_{n,i}^2(x) \mathbf{Cov}\left[Y_i\right] \text{ by independence} \\
&\leq \sigma^2 \sum_i W_{n,i}^2(x) \\
&\leq \sigma^2 \max_i |W_{n,i}(x)| \sum_i |W_{n,i}(x)| \\
&\leq \sigma^2 \frac{C}{nh} C'
\end{aligned}$$

□

5. Local polynomials have the same convergence rates as Nadaraya-Watson estimators, except it is uniform over $[0,1]$. Therefore, we can prove L^p convergence results.

6.4 Smoothing Splines

1. Smoothing Spline Estimator

   Definition 6.2. *The smooth spline estimator is:*

   $$\hat{m}^S_{n,\lambda} \in \arg\min_m \left[\sum_i (Y_i - m(x_i))\right] + \lambda \left\|m^{(s)}\right\|_2^2$$

2. General Splines:

   (a) A spline g of order $2s-1$ on $[0,1]$ with breakpoints at $0 < x_1 < x_2 < \cdots < x_n < 1$ is $2s-2$ times continuously differnetiable on each interval $(0,x_1),\ldots,(x_n,1)$, and is given by polynomials of degree $2s-1$ on each interval

   (b) A spline g is called natural if for $r = s,\ldots,2s-1$, $g^{(r)}(0) = g^{(r)}(1) = 0$

3. Properties of Spline Estimators:

   (a) (Example Sheet) $\hat{m}^S_{n,\lambda}$ is a natural spline of order $2s-1$

   (b) Given B-splines, N_k^{2s-1} for $k = 1,\ldots,n+2s$, we can rewrite the Spline Estimator as:

   $$\hat{m}^S_{n,\lambda}(x) = \sum_{k=1}^{n+2s} \hat{c}_k N_k^{2s-1}(x)$$

   (c) To estimate constants c_k, we can reformulate the original minimisation problem as follows:

      i. Let N be an $n \times n+2s$ matrix with entries $N_{i,k} = N_k^{2s-1}(x_i)$
      ii. Let C be the vector of unknown coefficients $\hat{c}_k$
      iii. Then:

      $$\hat{C} = \arg\min_C (Y - NC)^T (Y - NC) + \lambda C^T \Omega C$$

      where

      $$\Omega_{k,l} = \int_{\Re} \left(N_k^{2s-1}\right)^{(s)}(x) \left(N_l^{2s-1}\right)^{(s)}(x) dx$$

6.5 Wavelet Regression

1. Recall that for all $x \in \Re$ we can approximate a function m using father and mother wavelets:

$$W_j(m)(x) = \sum_{k \in Z} a_k \varphi_{0,k}(x) + \sum_{l=0}^{j-1} \sum_{k \in Z} b_{l,k} \psi_{l,k}(x)$$

2. In estimation, we focus on $x \in [0,1]$. A simplification is to assume $m = 0$ on $[0,1]^C$. We approximate the coefficients, then, as follows:

$$\begin{aligned}
\hat{a}_k &= \frac{1}{n} \sum_{i=1}^{n} \varphi_{0,k}(x_i) Y_i \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \varphi_{0,k}(x_i) m(x_i) \\
&\approx \int_0^1 \varphi_{0,k}(x) m(x) dx \\
&= \int_{\Re} \varphi(x) m(x) dx \text{ since } m = 0 \text{ outside of } [0,1] \\
\hat{b}_{l,k} &= \frac{1}{n} \sum_{i=1}^{n} \psi_{l,k}(x_i) Y_i \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \psi_{l,k}(x_i) m(x_i) \\
&\approx \int_0^1 \psi_{l,k}(x) m(x) dx \\
&= \int_{\Re} \psi_{l,k}(x) m(x) dx \text{ since } m = 0 \text{ outside of } [0,1]
\end{aligned}$$

3. This approximation scheme is similar to the Nadaraya-Watson estimator. It has the same boundary problems, which can be fixed using L^2 wavelets such as the Cohen-Daubechies-Vial basis.

7 Parameter Selection

7.1 Introduction

1. Motivation: The methods in previous sections required choosing parameters such as bandwidth h or resolution j. Optimal choices depend on unknown quantities, specifically, the smoothness of the function being estimated. Therefore, how do we choose the parameter correctly? Additionally, if smoothness of the function changes from region to region, how can we allow our smoothing parameter to vary in space?

2. Solutions:

   (a) Practical Approach: Implementing procedures which have intuitive appeal, but have no theoretical guarantees. These techniques tend to perform well.

(b) Adaptation: obtaining parameters which have asymptotic performance that is as good as if we knew the smoothness a priori. Adaptation can be used to select global or local parameters, but may not improve practical performance.

7.2 Global Bandwidth Choices

We consider methods which give a global parameter choice, and we discuss it in the context of Kernel Density Estimation, although they are easily applied to other methods.

1. **Plug-in Methods:** Guess a value for the smoothness s. Determine h_0 dependent on this value s. Calculate the estimator $\hat{f}^K_{n,h}$. Plug this back into the optimal bandwidth equation to get h_1. Continue iterating until h_n converges.

   (a) This method depends heavily on the starting point s

   (b) Theoretical results often assume that f has smoothness larger than the initial guess s, and, so, h will always be sub-optimal

2. **Cross-Validation:** Let $\hat{f}^{K,i}_{n,h}$ denote the Kernel Density Estimator from the observations after leaving out the i^{th} data point. Then for density or regression estimation (respectively), we determine h by minimising:

$$\int_{\Re} \left[\hat{f}^K_{n,h}(x)\right]^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}^{K,i}_{n,h}(x_i)$$

$$\frac{1}{n}\sum_{i=1}^{n} \left[\hat{m}^{K,i}_{n,h}(x_i) - Y_i\right]^2$$

   Derivation 7.1. *Consider the L^2 error, which we can approximately estimate as follows:*

$$\begin{aligned}\int_{\Re} \left[\hat{f}^K_{n,h} - f\right]^2 dx &= \int_{\Re} \left[\hat{f}^K_{n,h}(x)\right]^2 dx - 2\int_{\Re} \hat{f}^K_{n,h}(x) f(x) dx + C \\ &\approx \int_{\Re} \left[\hat{f}^K_{n,h}(x)\right]^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}^K_{n,h}(x_i) + C\end{aligned}$$

   We can ignore C since we do not know it, and because we estimate $\hat{f}^K_{n,h}$ using the data points, we cannot estimate its error using the same data. Therefore, we replace the last term with $\hat{f}^{K,i}_{n,h}(x)$.

   (a) Generalisations of this model often include weights to account for boundary problems

   (b) In general, bandwidths will only be good for L^2 or whatever loss function is being estimated.

3. **Model Selection:** Given an appropriate penalty function λ, we determine h by minimising:

$$\frac{1}{n}\sum_{i=1}^{n} \left(\hat{m}^K_{n,h}(x_i) - Y_i\right)^2 + \lambda(h)$$

(a) λ should be chosen to penalise small values in h, which would result in a simple interpolation of the data in the case of regression.

(b) This method is again tied to the choice of loss function, and only performs well in the context of that specific loss function.

7.3 Lepski's Method

1. Motivation: We need parameter selection methods which are good over all error metrics.

2. Lepski's Method (for Kernel Density Estimation)

   (a) Start with a finite collection of bandwidths $H = \{h_0 > h_1 > \cdots h_k\}$

   (b) We want to compare the biases for each $\hat{f}^K_{n,h}$ where $h \in H$, and select the one with the smallest bias. We do this by starting with $\hat{f}^K_{n,h_0}$ and comparing it with $\hat{f}^K_{n,h_i}$ for $i > 0$ since the smaller the smoothing parameter, the smaller the bias.

   (c) If the differences between the estimators with h_0 and h_i are significant (greater than some threshold), then we will reject h_0 in favour of the h_i that had the significant difference.

   (d) We repeat this process for all h_i with respect to h_j for $j > i$, until we find the largest bandwidth for which no significant comparison exists between this bandwidth and any smaller bandwidths. We use this bandwidth for Lepski's estimator $\hat{f}^L_n$

   (e) Summary: Given some threshold function $\tau(h)$, we want

$$\hat{h}_n = \max\{h \in H| \ \left|\hat{f}^K_{n,h} - \hat{f}^K_{n,h'}\right| \leq \tau(h'), \ \forall h' \in H \text{ s.t. } h' < h\}$$

3. Rate of Convergence for Lepski's Estimator

Theorem 7.1. *Let $X_1, \ldots, X_n$ be i.i.d. real-valued R.V. with a bounded density f. Let $\hat{f}^L_n$ be the Lepski estimator. Suppose that:*

(a) K satisfies the conditions of Proposition 4.1(c) with order $S \in N$

(b) The set H has bandwidths: $h_0 \sim n^{-1/(2S+1)}$, $h_i = \alpha^i h_0$ for $\alpha \in (0,1)$, and $h_k \sim \left(\frac{n}{\log n}\right)^{-1}$

(c) $\tau(h) = \kappa\sqrt{\log(n)/nh}$ for $\kappa > 0$ depending on $\|f\|_\infty$ and K

(d) f is s-times differentiable for $s \leq S$, and $f^{(s)} \in L^\infty$

Then,

$$\boldsymbol{E}\left[\left|\hat{f}^L_n(x) - f(x)\right|\right] = O\left(\left[\frac{n}{\log n}\right]^{-s/(2s+1)}\right)$$

Proof. We want to compare $\hat{h}_n$'s performance to the nearly-optimal $h^*_n \in H$, and study the behaviour of $\hat{h}_n$ when it is greater than and less than the nearly-optimal choice.

(a) Define

$$h_n^* = \max\{h \in H | \forall h' \in H \text{ s.t. } h' \leq h,\ |K_h * f - f| \leq \frac{1}{4}\tau(h')\}$$

By Proposition 4.1(c), we know that $|K_h * f - f| = O(h^s)$. So h_n^* is well defined, and we can solve for it by setting $O([h_n^*]^s) = \tau(h_n^*)$, yielding:

$$\frac{1}{h_n^*} = O\left(\left[\frac{n}{\log n}\right]^{1/(2s+1)}\right)$$

(b) When $\hat{h}_n > h_n^*$: (By definition of $\hat{h}_n$, then by Proposition 5.2(a), and by definition of h_n^*)

$$\begin{aligned}
&\mathbf{E}\left[\left|f_n^L(x) - f(x)\right| \mathbf{1}\left[\hat{h}_n > h_n^*\right]\right] \\
&\leq \mathbf{E}\left[\left|\hat{f}_n^L(x) - \hat{f}_{n,h_n^*}^K\right|\right] + \mathbf{E}\left[\left|\hat{f}_{n,h_n^*}^K - \left(K_{h_n^*} * f\right)\right|\right] + \left|\left(K_{h_n^*} * f\right) - f\right| \\
&\leq \tau(h_n^*) + \mathbf{E}\left[\left|\hat{f}_{n,h_n^*}^K - \left(K_{h_n^*} * f\right)\right|\right] + \left|\left(K_{h_n^*} * f\right) - f\right| \\
&\leq \tau(h_n^*) + \|K\|_2 \sqrt{\frac{\|f\|_\infty}{nh_n^*}} + \left|\left(K_{h_n^*} * f\right) - f\right| \\
&\leq \tau(h_n^*) + \|K\|_2 \sqrt{\frac{\|f\|_\infty}{nh_n^*}} + \frac{1}{4}\tau(h_n^*) \\
&= O\left(\left[\frac{n}{\log n}\right]^{-s/(2s+1)}\right)
\end{aligned}$$

(c) When $\hat{h}_n < h_n^*$, we have by assumption (b) that $Ch_n^* = \hat{h}_n$ for $C \in (0,1)$. Therefore,

$$\frac{1}{\hat{h}_n} = O\left(\left[\frac{n}{\log n}\right]^{1/(2s+1)}\right)$$

Therefore, since $\hat{f}_n^L = \hat{f}_{n,\hat{h}_n}^K$ by definition, and by definition of h_n^* and then Proposition 5.2(a):

$$\begin{aligned}
&\mathbf{E}\left[\left|\hat{f}_{n,\hat{h}_n}^K - f\right| \mathbf{1}\left[\hat{h}_n < h_n^*\right]\right] \\
&\leq \mathbf{E}\left[\left|\hat{f}_{n,\hat{h}_n}^K - \left(K_{\hat{h}_n} * f\right)\right|\right] + \left|\left(K_{\hat{h}_n} * f\right) - f\right| \\
&\leq \frac{1}{4}\tau(\hat{h}_n) + \left|\left(K_{\hat{h}_n} * f\right) - f\right| \\
&\leq \frac{1}{4}\tau(\hat{h}_n) + \|K\|_2 \sqrt{\frac{\|f\|_\infty}{n\hat{h}_n}} \\
&= O\left(\left[\frac{n}{\log n}\right]^{-s/(2s+1)}\right)
\end{aligned}$$

□

4. Similar results can be proven for other loss metrics, and for other regression and density estimators

5. Results hold locally: pointwise rate of convergence holds if f is s-times differentiable in a neighbourhood of the point

6. Choosing κ depends on $\|f\|_\infty$ which we can estimate using $\left\|\hat{f}^K_{n,h}\right\|_\infty$ for an h that gives a consistent estimate of f.

7.4 Wavelet Thresholding

1. Overview: We let $j_0 < j_1 \in N$, where j_0 is a minimum resolution and j_1 is a maximum resolution that we want to use in estimating our function. Our estimator is then:

$$\hat{f}^W_{n,j_1}(x) = \sum_{k\in Z} \hat{a}_k \varphi_{j_0,k}(x) + \sum_{l=j_0}^{j_1-1}\sum_{k\in Z} \hat{b}_{l,k}\psi_{l,k}(x)$$

   (a) If $|\hat{b}_{l,k}| \le \tau$, we will believe $b_{l,k}$ is 0. If $|\hat{b}_{l,k}| > \tau$, we believe it is non-zero. Therefore, we have the following threshold wavelet coefficients:

$$\hat{b}^T_{l,k} = \hat{b}_{l,k}\mathbf{1}\left[|\hat{b}_{l,k}| > \tau\right]$$

   (b) The wavelet thresholding estimate is then:

$$\hat{f}^T_n(x) = \sum_{k\in Z} \hat{a}_k \varphi_{j_0,k}(x) + \sum_{l=j_0}^{j_1-1}\sum_{k\in Z} \hat{b}^T_{l,k}\psi_{l,k}(x)$$

2. Rate of Convergence for the Wavelet Thresholding Estimator

Theorem 7.2. *Let $X_1, \dots, X_n$ be real valued, i.i.d. R.V with bounded density function f. Let $\hat{f}^T_n$ be the wavelet thresholding estimator of f. Suppose:*

(a) The wavelet basis satisfies the conditions of Proposition 4.4 with S vanishing moments ($r = 0, 1, \dots, S-1$)

(b) j_0 and j_1 are chosen in terms of n so that $2^{j_0} \sim n^{1/(2S+1)}$ and $2^{j_1} \sim \frac{n}{\log n}$

(c) $\tau = \kappa\sqrt{\frac{\log n}{n}}$ for κ depending only on $\|f\|_\infty$ and φ

(d) f is s-times differentiable for $s \le S$ and $f^{(s)} \in L^\infty$

Then:

$$\boldsymbol{E}\left[\left|\hat{f}^T_n(x) - f(x)\right|\right] = O\left(\left[\frac{n}{\log n}\right]^{-s/(2s+1)}\right)$$

Proof. This proof relies on splitting terms into different levels of resolution, then splitting terms into bias and variance, and again into regions of high and lower probability, which we can bound "easily"

(a) We first divide the terms into high and low resolution parts:

$$\left|\hat{f}_n^T(x) - f(x)\right| = \left|\hat{f}_{n,j_0}^W + \sum_{l=j_0}^{j_1-1}\sum_{k\in Z}\hat{b}_{l,k}^T\psi_{l,k}(x) - f\right|$$

$$\leq \left|\hat{f}_{n,j_0}^W - W_{j_0}(f)\right| + \left|\sum_{l=j_0}^{j_1-1}\sum_{k\in Z}\left[\hat{b}_{l,k}^T - b_{l,k}\right]\psi_{l,k}(x) - f\right| + |W_{j_1}(f) - f|$$

(b) The low and high resolution terms:

i. $\mathbf{E}\left[\left|\hat{f}_{n,j_0}^W - W_{j_0}(f)\right|\right]$ is bounded by Proposition 5.5(a) and assumption (b)

ii. $|W_{j_1}(f) - f| = O\left(\left[\frac{n}{\log n}\right]^{-s/(2s+1)}\right)$ by Proposition 4.4(c) and assumption (b)

(c) Variance-Bias decomposition of the remaining term:

$$\left|\sum_{l=j_0}^{j_1-1}\sum_{k\in Z}\left[\hat{b}_{l,k}^T - b_{l,k}\right]\psi_{l,k}(x) - f\right|$$
$$\leq \left|\sum\sum\left(\hat{b}_{l,k} - b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| > \tau\right]\right|$$
$$+ \left|\sum\sum\left(b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| \leq \tau\right]\right|$$

(d) Low-High Probability Decomposition of Variance:

$$\left|\sum\sum\left(\hat{b}_{l,k} - b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| > \tau\right]\right|$$
$$\leq \left|\sum\sum\left(\hat{b}_{l,k} - b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| > \tau, |b_{l,k}| < \frac{\tau}{2}\right]\right|$$
$$+ \left|\sum\sum\left(\hat{b}_{l,k} - b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| > \tau, |b_{l,k}| > \frac{\tau}{2}\right]\right|$$

(e) Low-High Probability Decomposition of Bias:

$$\left|\sum\sum\left(b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| \leq \tau\right]\right|$$
$$\leq \left|\sum\sum\left(b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| \leq \tau, |b_{l,k}| > 2\tau\right]\right|$$
$$\leq \left|\sum\sum\left(b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right| \leq \tau, |b_{l,k}| \leq 2\tau\right]\right|$$

(f) Notice that we want to bound the infinite sum over $k \in Z$, the term $|b_{l,k}|$ and the variance $|\hat{b}_{l,k} - b_{l,k}$

i. Because ψ is compactly supported in $[-R, R]$ for $R \in N$, for any $x \in \Re$, $\sum_{k\in Z}\psi_{l,k}$ can have at most $2R+1$ terms. So for a fixed $x \in \Re$, $\sum_{k\in Z}\psi_{l,k}(x) \leq (2R+1)\left\|\psi\right\|_\infty 2^{l/2}$

ii. To bound $|b_{l,k}|$ we use Taylor's Theorem and the vanishing moments assumption. For some fixed number $f(2^{-l}k)$:

$$\begin{aligned}
|b_{l,k}| =& \left| \int_{\Re} \psi_{l,k}(x) f(x) dx - 0 \right| \\
=& \left| \int_{\Re} \psi_{l,k}(x) \left[f(x) - f(2^{-l}k) \right] dx \right| \\
=& 2^{-l/2} \left| \int_{\Re} \psi(u) \left[f(2^{-l}u + 2^{-l}k) - f(2^{-l}k) \right] du \right| \\
=& 2^{-l/2} \left| \int_{\Re} \psi(u) \left[\sum_{k=1}^{s-1} \frac{f^{(k)}(2^{-l}k)}{k!} \left(2^{-l}u\right)^k + \frac{f^{(s)}(y)}{k!} 2^{-ls}(u)^s \right] \right| \\
=& 2^{-l/2} \left| \int_{\Re} \psi(u) \left[\frac{f^{(s)}(y)}{k!} 2^{-ls}(u)^s \right] du \right| \\
\leq& 2^{-l(s+1/2)} \left\| f^{(s)} \right\|_\infty \frac{1}{k!} \left| \int_{\Re} \psi()u) u^s du \right| \\
\leq& C 2^{-l(s+1/2)}
\end{aligned}$$

iii. To bound the variance term, we use Jensen's inequality and bound the variance:

$$\begin{aligned}
\mathbf{E}\left[\left(\hat{b}_{l,k} - b_{l,k} \right)^2 \right] &\leq \mathbf{E}\left[\left(\hat{b}_{l,k} \right) \right] \\
&= \mathbf{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} \psi_{l,k} X_i \right)^2 \right] \\
&= \frac{1}{n} \mathbf{E}\left[\psi_{l,k}^2(X) \right] \\
&\leq \frac{C}{n} \left\| f \right\|_\infty
\end{aligned}$$

(g) To bound the Low-Probability Bias Term, we use the Bernstein Inequality, if $\kappa \geq 1$ and for some $D > 0$:

i. Notice that:

$$\begin{aligned}
\mathbf{P}\left[\left| \hat{b}_{l,k} \right| \leq \tau, |b_{l,k}| > 2\tau \right] &\leq \mathbf{P}\left[\left| b_{l,k} - \hat{b}_{l,k} \right| > \tau \right] \\
&\leq 2 \exp\left[-\kappa \log(n) D \right] \\
&= 2 n^{-D\kappa}
\end{aligned}$$

ii. Therefore, for sufficiently large κ:

$$\begin{aligned}
&\mathbf{E}\left[\left|\sum\sum\left(b_{l,k}\right)\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right|\leq\tau,|b_{l,k}|>2\tau\right]\right|\right]\\
&\leq(2R+1)\left\|\psi\right\|_{\infty}2^{l/2}\mathbf{P}\left[\left|\hat{b}_{l,k}\right|\leq\tau,|b_{l,k}|>2\tau\right]\sum_{l=j_0}^{j_1-1}C2^{-l(s+1/2)}\\
&\leq(2R+1)\left\|\psi\right\|_{\infty}2Cn^{-\kappa D}\sum_{j_0}^{j_1-1}2^{-ls}\\
&=O\left(\left[\frac{n}{\log(n)}^{-S/(2S+1)}\right]\right)
\end{aligned}$$

(h) To bound the High-Probability Bias Term:

i. We have that:

$$\begin{aligned}
&\left|\sum\sum b_{l,k}\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right|\leq\tau,|b_{l,k}|\leq2\tau\right]\right|\\
&\leq(2R+1)\left\|\psi\right\|_{\infty}\sum_{l=j_0}^{j_1-1}2^{l/2}\min(2\tau,C2^{-l(s+1/2)})\\
&\leq(2R+1)\left\|\psi\right\|_{\infty}\left(2\kappa\sum_{l=j_0}^{j-1}\sqrt{\frac{\log n}{n}}2^{l/2}+C\sum_{j=l}^{j_1-1}2^{-ls}\right)\\
&=O\left(\left[\frac{n}{\log(n)}^{-s/(2s+1)}\right]\right)
\end{aligned}$$

ii. Where: $j\in[j_0,j_1-1]$, is such that $2^j\sim\left(\frac{\log n}{n}\right)^{1/(2s+1)}$ so that we have:

$$\begin{aligned}
2\kappa\sqrt{\frac{\log n}{n}}2^{j/2}&\sim C2^{-js}\\
\left(\frac{\log n}{n}\right)^{\frac{1/2}{(2s+1)}-1/2}&\sim\left(\frac{\log n}{n}\right)^{-s/(2s+1)}\\
\left(\frac{\log n}{n}\right)^{\frac{1/2-s-1/2}{(2s+1)}}&\sim\left(\frac{\log n}{n}\right)^{-s/(2s+1)}
\end{aligned}$$

(i) To bound the High-Probability Variance Term:

i. Note that if $|b_{l,k}|>\frac{\tau}{2}$ then $\min(2|b_{l,k}|,\tau)=\tau$, so $l<j$ for j s.t.

$$2^j\sim\left(\frac{\log n}{n}\right)^{1/(2s+1)}$$

ii. Therefore we have:

$$\begin{aligned}
&\mathbf{E}\left[\left|\sum_{l=j_0}^{j_1-1}\sum(\hat{b}_{l,k}-b_{l,k})\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right|>\tau,|b_{l,k}|>\frac{\tau}{2}\right]\right|\right]\\
&\leq(2R+1)\left\|\psi\right\|_\infty\sqrt{\mathbf{E}\left[(\hat{b}_{l,k}-b_{l,k})^2\right]}\sum_{l=j_0}^{j-1}2^{l/2}\\
&\leq(2R+1)\left\|\psi\right\|_\infty\sqrt{\left\|f\right\|_\infty n^{-1}}\sum_{l=j_0}^{j-1}2^{l/2}\\
&=O\left(\left[\frac{n}{\log(n)}^{-s/(2s+1)}\right]\right)
\end{aligned}$$

(j) To bound the Low-Probability Variance Term:

i. By Cauchy-Schwarz, and Bernstein's Inequality:

$$\begin{aligned}
&\mathbf{E}\left[\left|\sum_{l=j_0}^{j_1-1}\sum(\hat{b}_{l,k}-b_{l,k})\psi_{l,k}\mathbf{1}\left[\left|\hat{b}_{l,k}\right|>\tau,|b_{l,k}|\leq\frac{\tau}{2}\right]\right|\right]\\
&\leq(2R+1)\left\|\psi\right\|_\infty\sum_{l=j_0}^{j_1-1}2^{l/2}\mathbf{E}\left[|\hat{b}_{l,k}-b_{l,k}|\mathbf{1}\left[\left|\hat{b}_{l,k}\right|>\tau,|b_{l,k}|\leq\frac{\tau}{2}\right]\right]\\
&\leq(2R+1)\left\|\psi\right\|_\infty\times\\
&\quad\sum_{l=j_0}^{j_1-1}2^{l/2}\sqrt{\mathbf{E}\left[|\hat{b}_{l,k}-b_{l,k}|^2\right]\mathbf{P}\left[\left|\hat{b}_{l,k}\right|>\tau,|b_{l,k}|\leq\frac{\tau}{2}\right]}\\
&\leq(2R+1)\left\|\psi\right\|_\infty\sqrt{\left\|f\right\|_\infty n^{-1}}2n^{-D'\kappa}\sum_{l=j_0}^{j_1}2^{l/2}\\
&=O\left(\left[\frac{n}{\log(n)}^{-s/(2s+1)}\right]\right)
\end{aligned}$$

□