NonParametricStatistics

Andrew Tulloch

Contents

Chapter 1. Introduction 5

1. Basic Concepts 5

1.1. Parametric vs Nonparametric models 5

1.2. Estimating an arbitrary distribution function 6

2. Density Estimators 8

3. Asymptotic MSE and MSIE approximation 9

4. Pointwise asymptotic distribution 12

5. Bandwidth Selection 13

5.1. Normal Scale rules 13

5.2. Least-squares Cross-Validation 13

5.3. Biased Cross-Validation 13

5.4. Solve-the-equation Rules 14

6. Other Topics 14

6.1. Choice of Kernel 14

6.2. Derivative Estimation 15

6.3. Higher Order Kernels 15

6.4. Local Bandwidths 15

6.5. Transformation Methods 16

6.6. Multi-Dimensional Density Estimation 16

Chapter 2. Nonparametric Regression 17

1. Introduction 17

2. Local polynomial estimator 17

3. MSE approximations 18

4. Splines 20

4.1. Motivation 20

4.2. Cubic Spline 21

4.3. Natural Cubic Smoothing Spline 22

4.4. Choice of λ 23

4.5. Regression Spline and Penalized Spline 24

4.6. Equivalent Kernel 25
5. Multivariate Regression and Additive Models 25

Chapter 3. Nearest Neighbor Classification 27

Chapter 4. Minimax Lower Bounds 30

Chapter 5. Extreme Value Theory 33
1. Preliminaries 33
2. Necessary and Sufficient Conditions for Convergence 35

Bibliography 38

CHAPTER 1

Introduction

1. Basic Concepts

THEOREM 1.1 (The Delta Method). *Let Y_n be a sequence of random vectors in $\mathbb{R}^d$ such that for some $\mu \in \mathbb{R}^d$ and a random vector Z, we have $n^{\frac{1}{2}}(Y_n - \mu) \xrightarrow{d} Z$. If $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable at μ, then $n^{\frac{1}{2}}(g(Y_n) - g(\mu)) \xrightarrow{d} \nabla g(\mu)^T Z$.*

PROOF. For $d = 1$. Let $g'(\mu) = \nabla g(\mu)$, and let $h : \mathbb{R} \to \mathbb{R}$, by

$$h(y) = \begin{cases} \frac{g(y)-g(\mu)}{y-\mu} & y \neq \mu \\ g'(\mu) & y = \mu \end{cases} \tag{1.1}$$

Then by the continuous mapping theorem and Slutsky's theorem, $n^{\frac{1}{2}}(g(Y_n) - g(\mu)) = h(Y_n)n^{\frac{1}{2}}(Y_n - \mu) \xrightarrow{d} g'(\mu)Z$. □

1.1. Parametric vs Nonparametric models. A statistical model postulates a family of possible data generating mechanisms. Examples include:

(i) Let $X_1, \ldots, X_n \sim T(m, \theta)$ IID, with m known and $\theta \in (0, \infty) = \Theta$ an unknown parameter.
(ii) Let $Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, \ldots, n$ where x_i are known and ϵ_i are IID with $\mathbb{E}(e_i) = 0, \mathbb{V}(\epsilon_i) = \sigma^2$. Here, the unknown parameter is $\theta = \begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix} \in \mathbb{R} \times \mathbb{R} \times (0, \infty) = \Theta$.

If the parameter space Θ is finite dimensional, we speak of a **parametric model**. In such situations, typically we can estimate θ using the MLE $\hat{\theta}_n$, and have $\hat{\theta}_n - \theta = O_p(n^{-\frac{1}{2}})$.[1]

This assumes the model contains the true data generating process, if not, inference can be misleading.

Examples of nonparametric models include:

(i) Let $X_1, \ldots, X_n, i = 1, \ldots, n$ be IID with arbitrary distribution function F.
(ii) Let $X_1, \ldots, X_n, i = 1, \ldots, n$ be IID with twice continuously differentiable density f.
(iii) Let $Y_i = m(x_i) + v(x_i)^{\frac{1}{2}}, i = 1, \ldots, n$ where m is twice continuously differentiable and s $\epsilon_1, \ldots, \epsilon_n$ are IID with $\mathbb{E}(\epsilon_i) = 0, \mathbb{V}(\epsilon_i) = 1$.

[1] Definition of O_p - TODO

Such infinite-dimensional models are much less vulnerable to model misspecification, typically, however we pay a price for our generality in terms of a slower convergence rate - e.g. $O_p(n^{-\frac{2}{3}})$ in problems (ii) and (iii) above.

1.2. Estimating an arbitrary distribution function. Let $X_1, \ldots, X_n$ be IID on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution function F. The **empirical distribution function** $\hat{F}_n$ is defined by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) . \tag{1.2}$$

THEOREM 1.2 (Glivenko-Cantelli (1933) - The Fundamental Theorem of Statistics).

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \stackrel{as}{\to} 0. \tag{1.3}$$

PROOF. Given $\epsilon > 0$, choose a partition $-\infty = x_0 < x_1 < \cdots < s_k = \infty$ such that, for each $i = 1, \ldots, k$, we have $F(x_i-) - F(x_{i-1}) \leq \epsilon$, where $F(x-) = \lim_{y \uparrow x} F(y)$.

Note that any point at which F jumps by more than ϵ must be in the partition. By the strong law of large numbers, there exists an event Ω_ϵ with $\mathbb{P}(\Omega_\epsilon) = 1$ such that for all $\omega \in \Omega_\epsilon$, there exists $n_0 = n_0(\omega, \epsilon)$ with

$$\left| \hat{F}_n(x_i) - F(x_i) \right| \leq \epsilon, i = 1, \ldots, k-1, n \geq n_0, \tag{1.4}$$

$$\left| \hat{F}_n(x_i-) - F(x_i-) \right| \leq \epsilon, 1 = i, \ldots, k-1, n \geq n_0. \tag{1.5}$$

Now, fix $x \in \mathbb{R}$, and find $i \in \{1, \ldots, k\}$ with $x \in [x_{i-1}, \ldots, x_i)$. Then for $\omega \in \Omega_\epsilon$ and $n \geq n_0$,

$$\hat{F}_n(x) - F(x) \leq \hat{F}_n(x_i-) - F(x_{i-1}) = \hat{F}_n(x_i-) - F(x_i-) + F(x_i-) - F(x_{i-1}) \leq \epsilon + \epsilon = 2\epsilon \tag{1.6}$$

Similarly, we have

$$F(x) - \hat{F}_n(x) \leq F(x_i-) - \hat{F}_n(x_{i-1}) = F(x_i-) - F(x_{i-1}) + F(x_{i-1}) - \hat{F}_n(x_{i-1}) \leq \epsilon + \epsilon = 2\epsilon \tag{1.7}$$

We deduce that

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \to 0 \right) = \mathbb{P}\left(\cap_{m=1}^{\infty} \cup_{n_0=1}^{\infty} \cap_{n=n_0}^{\infty} \{ \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \leq \frac{1}{m} \} \right) \tag{1.8}$$

$$= \lim_{m \to \infty} \mathbb{P}\left(\Omega_{\frac{1}{2m}} \right) = 1 \tag{1.9}$$

□

THEOREM 1.3 (Dvortesky-Kiefer-Wolfowitz). *Let $X_1, \ldots, X_n \sim F$ IID. Then for every $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon \right) \leq 2e^{-2n\epsilon^2}. \tag{1.10}$$

An application is to consider the problem of finding a confidence band for F at $1-\alpha$. Given $\alpha \in (0,1)$, set $\epsilon_n = (-\frac{1}{2n}\log\frac{\alpha}{2})^{\frac{1}{2}}$. Then by 1.3,

$$(\max(\hat{F}_n(x) - \epsilon_n, 0), \min(\hat{F}_n(x), 1)) \tag{1.11}$$

is a $1-\alpha$ confidence interval for F.

In fact, let $U_1, \dots, U_n \sim U(0,1)$ IID, and let $\hat{G}_n$ denote their empirical distribution function. Then

$$\hat{G}_n(F(x)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(U_i \leq F(x)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big(F^{-1}(u_i) \leq x\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) = \hat{F}_n(x) \tag{1.12}$$

It follows that

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{x\in R} |\hat{G}_n(F(x)) - F(x)| \leq \sup_{t\in(0,1)} |\hat{G}_n(t) - t| \tag{1.13}$$

with equality if F is continuous. We deduce that, if F is continuous, the distribution of $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|$ does not depend on F.

Other examples include Uniform Laws of Large Numbers (ULLN). Let $X, X_1, X_2, \dots$ be IID taking values in a measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{G}$ denote a class of measurable functions on $\mathcal{X}$. We say that $\mathcal{G}$ satisfies a ULLN if

$$\sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}(g(X))\Big| \xrightarrow{as} 0. \tag{1.14}$$

Thus Theorem 1 shows that the class $\mathcal{G} = \{\mathbb{I}(\cdot \leq x) : x \in \mathbb{R}\}$ satisfies a ULLN. In general, proving a ULLN amounts to controlling the **size** of $\mathcal{G}$, which can be done by using the idea of entropy (c.f. Statistical Theory).

Further results start with the observation that

$$n^{\frac{1}{2}}(\hat{F}_n - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))) \tag{1.15}$$

by the central limit theory. This result can be strengthened by studying $\{n^{\frac{1}{2}}(\hat{F}_n(x) - F(x)), x \in \mathbb{R}\}$ as a stochastic process.

PROPOSITION 1.4. *Let* $U_1, \dots, U_n \sim U(0,1)$ IID. *Let* $Y_1, \dots, Y_{n+1} \sim \text{EXP}(1)$ IID *and let* $S_j = \sum_{i=1}^{j} Y_i$ *for* $j = 1, \dots, n+1$. *Then*

$$U_j =^d \frac{S_j}{S_{n+1}} \sim \text{BETA}(j, n-j+1). \tag{1.16}$$

DEFINITION 1.5. For $p \in (0,1]$, the quartile function is defined by $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ and is left-continuous.

The sample quartile function is $\hat{F}_n^{-1}(p) = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\}$.

THEOREM 1.6. *Let* $U_1, U_2, \ldots, U_n \sim U(0,1)$ IID *and* $p \in (0,1)$. *Then*

$$\sqrt{n}(U_{\lceil np \rceil} - p) \xrightarrow{d} N(0, p(1-p)). \tag{1.17}$$

PROOF. Let $Y_1, \ldots, Y_n$ IID EXP(1), let $V_n = Y_1 + \cdots + Y_{\lceil np \rceil}$ and $W_n = Y_{\lceil np \rceil + 1}, \ldots, Y_{n+1}$. Note that V_n, W_n are independent and

$$\frac{V_n}{V_n + W_n} =^d U_{\lceil np \rceil} \tag{1.18}$$

by previous proposition. Then

$$\sqrt{n}(\frac{V_n}{n} - p) = \frac{\sqrt{\lceil np \rceil}}{\sqrt{n}}(\frac{V_n - \lceil np \rceil}{\sqrt{\lceil np \rceil}}) + \frac{\lceil np \rceil - np}{\sqrt{n}} \xrightarrow{d} N(0, p) \tag{1.19}$$

by the CLT and Slutsky's theorem.

Similarly, $\sqrt{n}(\frac{W_n}{n} - q) \xrightarrow{d} N(0, q)$, where $q = 1-p$, then by the delta method, with $g(x, y) = \frac{x}{x+y}$,

$$\sqrt{n}(U_{\lceil np \rceil} - p) =^d \sqrt{n}(g(\frac{V_n}{n}, \frac{W_n}{n}) - g(p, q)) \tag{1.20}$$

$$\xrightarrow{d} N(0, \nabla g(p,q)^T \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \nabla g(p, q)) \tag{1.21}$$

$$=^d n(0, p(1-p)) \tag{1.22}$$

□

THEOREM 1.7. *Let* $p \in (0,1)$ *and let* $X_1, \ldots, X_n \text{IID} F$ *where* F *is differentiable at* $F^{-1}(p)$ *with positive derivative* $f(F^{-1}(p))$. *Then*

$$\sqrt{n}(X_{\lceil np \rceil} - F^{-1}(p)) \xrightarrow{d} N(0, \frac{p(1-p)}{f(F^{-1}(p))^2}) \tag{1.23}$$

PROOF. Let $U_1, \ldots, U_n \text{IID} U(0,1)$ so that $F^{-1}(U_{\lceil np \rceil}) =^d X_{\lceil np \rceil}$. Then by the previous theorem and the delta method with $g = F^{-1}$,

$$\sqrt{n}(X_{\lceil np \rceil} - F^{-1}(p)) =^d \sqrt{n}(g(U_{\lceil np \rceil}) - g(p)) \tag{1.24}$$

$$\xrightarrow{d} N(0, \frac{p(1-p)}{f(F^{-1}(p))^2}) \tag{1.25}$$

□

2. Density Estimators

DEFINITION 1.8 (Histogram Estimator).

$$\tilde{f}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathbb{I}(X_i \in [x_k, x_{k+1})) \tag{1.26}$$

Definition 1.9 (Kernel Density Estimator).

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}). \tag{1.27}$$

where $K : \mathbb{R} \to \mathbb{R}$ satisfies $\int_{\mathbb{R}} K(x)dx = 1$ is called the kernel, $h > 0$ is the bandwidth.

Write $K_h(x) = \frac{1}{h}K(\frac{x}{h})$ so that

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i). \tag{1.28}$$

Definition 1.10 (MSE).

$$MSE(\hat{f}_h(x)) = \mathbb{E}\Big((\hat{f}_h(x) - f(x))^2\Big) \tag{1.29}$$

$$= \mathbb{E}\bigg((\hat{f}_h(x) - \mathbb{E}\Big(\hat{f}_h(x)\Big)^2)^2\bigg) + (\mathbb{E}\Big(\hat{f}_h(x)\Big) - f(x))^2. \tag{1.30}$$

Write $(f \star g)(x) = \int_{\mathbb{R}} f(x-y)g(y)dy$

Theorem 1.11. *For the KDE, we can write*

$$Bias(\hat{f}_h(x)) = \mathbb{E}(K_h(x - X_1)) - f(x) \tag{1.31}$$

$$= \int_{\mathbb{R}} K_h(x-y)f(y)dy - f(x) \tag{1.32}$$

$$= (K_h \star f)(x) - f(x) \tag{1.33}$$

Similarly,

$$\mathbb{V}(\hat{f}_h(x)) = \frac{1}{n}((K_h^2 \star f)(x) - (K_h \star f)(x)^2) \tag{1.34}$$

Usually, we prefer to choose h to minimize some expression measuring how well $\hat{f}_h$ estimates f as a function. We therefore define the Mean Integrated Squared Error ($MSIE$) as

$$MSIE(\hat{f}_h) = \mathbb{E}\Big(\int_{-\infty}^{\infty}\{\hat{f}_h(x) - f(x)\}^2 dx\Big) \tag{1.35}$$

$$= \int_{-\infty}^{\infty} MSE(\hat{f}_h(x))dx \tag{1.36}$$

$$= \int_{\infty}^{\infty}((K_h \star f)(x) - f(x))^2 + \frac{1}{h}((K_n^2 \star f)(x) - (K_h \star f)^2(x))dx \tag{1.37}$$

which is justified by Fubini's theorem as the integrand is non-negative. Although exact, this expression depends on h in a complicated way. We therefore seek asymptotic approximation to clarify this dependence and facilitate an asymptotically optimal choice of h.

3. Asymptotic MSE and MSIE approximation

We need the following conditions:

(i) f is twice differentiable, $f'$is bounded, and $R(f) = \int_{-\infty}^{\infty} f''(x)^2 dx < \infty$.
(ii) $h = h_n$ is a non-random sequence with $h \to 0$ and $nh \to \infty$ as $n \to \infty$.
(iii) K is non-negative, $\int_{-\infty}^{\infty} K(x)dx = 1$, $\int_{-\infty}^{\infty} xK(x)dx = 0$, $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x)dx < \infty$, and $R(x) < \infty$.

THEOREM 1.12. *Assume that the previous conditions hold. Then, for all $x \in \mathbb{R}$,*

$$MSE(\hat{f}_n(x)) = \frac{R(K)f(x)}{nh} + \frac{1}{4}h^4\mu_2^2(K)f''(x)^2 + o(\frac{1}{nh} + h^4) \tag{1.38}$$

as $n \to \infty$.

PROOF. We first claim that f is bounded. Otherwise, there would exists (x_n) such that $f(x_n) \geq n$. Since f is a density, the exists $x_{n,l} \in [x_n - \frac{2}{n}, x_n]$ such that $f(x_{n,l}) \leq \frac{n}{2}$ and there exists $x_{n,m} \in [x_n, x_n + \frac{2}{n}]$ such that $f(x_{n,m}) \leq \frac{n}{2}$. y the mean value theorem, there exists $x^\star_{n,l} \in [x_{n,l}, x_n]$ such that $f'(x^\star_{n,l}) \geq \frac{n^2}{4}$ and there exists $x^\star_{n,m} \in [x_n, x_{n,m}]$ such that $f'(x^\star_{n,m}) \leq -\frac{n^2}{4}$. By the mean value theorem again, we have that there exists $x_n^{\star\star} \in [x^\star_{n,l}, x^\star_{n,m}]$ such that $f''(x_n^{\star\star}) \leq -\frac{n^3}{8}$, contradicting boundedness of f''.

We can therefore define $C_0 = \sup_{x\in\mathbb{R}} f(x)$ and $C_2 = \sup_{x\in\mathbb{R}} |f''(x)|$.

Now,

$$\mathbb{E}\Big(\hat{f}_h(x)\Big) = \int_{-\infty}^{\infty} \frac{1}{h}K(\frac{x-y}{h})f(y)dy \tag{1.39}$$

$$= \int_{-\infty}^{\infty} K(z)f(x-hz)dz \tag{1.40}$$

$$= \int_{-\infty}^{\infty} K(z)(f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x))dz + REM_1 \tag{1.41}$$

$$= f(x) + \frac{1}{2}h^2\mu_2(K)f''(x) + REM_1. \tag{1.42}$$

To control the remainder, given $\epsilon > 0$, choose $\delta > 0$ such that

$$|f(x-hz) - (f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x))| \leq \epsilon h^2 z^2 \tag{1.43}$$

for all $|hz| \leq \delta$.

Then

$$|REM_1| = |\int_{-\infty}^{\infty} K(z)f(x-hz)dz - \int_{-\infty}^{\infty} K(x)(f(x) + \frac{1}{2}h^2z^2f''(x))dz| \tag{1.44}$$

$$\leq |\int_{|z|>\frac{\delta}{h}} K(z)f(x-hz)dz| + \int_{|z|\leq\frac{\delta}{h}} K(z)|f(x-hz) - (f(z) + \frac{1}{2}h^2z^2f''x)|dz \tag{1.45}$$

$$+ |\int_{|z|>\frac{\delta}{h}} K(z)(f(x) + \frac{1}{2}h^2z^2f''(x))dz| \tag{1.46}$$

$$\leq C_0\frac{h^2}{\delta^2}\int_{|z|>\frac{\delta}{h}} z^2K(x)dz \tag{1.47}$$

$$+ \epsilon h^2\int_{|z|\leq\frac{\delta}{h}} z^2K(z)dz + C_0\frac{h^2}{\delta^2}\int_{|z|>\frac{\delta}{h}} z^2K(z)dz + \frac{1}{2}h^2C_2\int_{|z|>\frac{\delta}{h}} z^2K(z)dz \tag{1.48}$$

$$\leq \epsilon h^2 1 + \mu_2(K) \tag{1.49}$$

since $\int_{-\infty}^{\infty} zK(z)dx = 0$, Markov's inequality, etc. Thus,

$$BIAS(\hat{f}_h(x)) = \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^4). \tag{1.50}$$

For the variance,

$$\mathbb{V}\Big(\hat{f}_h(x)\Big) = \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2(\frac{x-y}{h})f(y)dy - \frac{1}{n}\{\mathbb{E}\Big(\hat{f}_h(x)\Big)\}^2 \tag{1.51}$$

$$= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(z)f(x-hz)dz - \frac{1}{n}(f(x)+o(1))^2 \tag{1.52}$$

$$= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(z)f(x)dz + REM_2 + O(\frac{1}{n}) \tag{1.53}$$

$$= \frac{R(K)f(x)}{nh} + REM_2 + O(\frac{1}{n}) \tag{1.54}$$

To control REM_2, given $\epsilon > 0$, choose $y > 0$ such that $|f(x-hz) - f(x)| \leq \epsilon$ for $|hz| \leq y$. Then

$$nh|REM_2| = |\int_{-\infty}^{\infty} K^2(z)(f(x-hz) - f(x))dz| \tag{1.55}$$

$$\leq \epsilon\int_{|z|\leq\frac{y}{h}} K^2(z) + 2C_0\int_{|z|>\frac{y}{h}} K^2(z)dz \tag{1.56}$$

$$\leq \epsilon(R(K)+1) \tag{1.57}$$

for large n.

We deduce that $\mathbb{V}\Big(\hat{f}_h(x)\Big) = \frac{R(K)f(x)}{nh} + o(\frac{1}{nh})$ and

$$MSE(\hat{f}_h(x)) = \frac{R(K)f(x)}{nh} + \frac{1}{4}h^4\mu_2^2(K)f''(x)^2 + o(\frac{1}{nh} + h^2) \tag{1.58}$$

The hope is that to compute the MSIE, we can just integrate the MSE over range of the RV. We need to be careful - in general we cannot integrate asymptotic pointwise estimates - need to understand dependency on x.

With mild additional conditions and further work (see the example sheet), it can be shown that

$$MISE(\hat{f}_h) = \frac{R(K)}{nh} + \frac{1}{4}h^4\mu_2^2(K)R(f'') + o(\frac{1}{nh} + h^4) \tag{1.59}$$

We see that asymptotically the integrated variance term decreases with h while the integrated squared bias term increases with h. This is the **bias-variance tradeoff**.

This **bias-variance tradeoff** summarizes the critical role of the bandwidth. □

Consider now minimizing the asymptotic MISE (AMISE) $\frac{R(K)}{nh} + \frac{1}{4}h^4\mu_2^2(K)R(f'')$ with respect to h, yielding the asymptotically optimal bandwidth

$$h_{AMISE} = (\frac{R(K)}{\mu_2^2(K)R(f'')n})^{\frac{1}{5}} \tag{1.60}$$

Substituting back, we obtain

$$AMISE(\hat{f}_{AMISE}) = \frac{5}{4}R(K)^{\frac{4}{5}}\mu_2(K)^{\frac{2}{5}}R(f'')^{\frac{1}{5}}n^{-\frac{4}{5}}. \tag{1.61}$$

Notice the slower rate than the typical $O(n^{-1})$ parametric rate. Notice that for the "rough" densities, with larger $R(f'')$, we should use a smaller bandwidth, and these densities are harder to estimate.

4. Pointwise asymptotic distribution

THEOREM 1.13. *Assume the previous assumptions (i), (ii), (iii) and that K is bounded. Then, for all $x \in \mathbb{R}$,*

$$n^{\frac{2}{5}}(\hat{f}_{h_{AMISE}}(x) - f(x)) \xrightarrow{d} N(\frac{1}{2}\mu_2(K)f''(x), R(K)f(x)) \tag{1.62}$$

PROOF. First, observe that from the proof of the previous theorem,

$$n^{\frac{2}{5}}(\mathbb{E}\Big(\hat{f}_{h_{AMISE}}(x) - f(x)\Big)) \to \frac{1}{2}\mu_2 K f''(x) \tag{1.63}$$

For the stochastic term, let $Y_{ni} = \frac{1}{h^{\frac{1}{2}}}K(\frac{x-X_i}{h})$. We have

$$\mathbb{V}(Y_{ni}) = \frac{1}{h}\int_{-\infty}^{\infty} K^2(\frac{x-y}{n})f(y) - h(\mathbb{E}\Big(\hat{f}_h(x)\Big))^2 \tag{1.64}$$

$$= \int_{-\infty}^{\infty} K^2(z)f(x-hz)dz - h(f(x) + o(1))^2 \tag{1.65}$$

$$\to R(K)f(x) \tag{1.66}$$

as $n \to \infty$.

Moreover,

$$\mathbb{E}\Big(Y_{ni}^2\mathbb{I}\Big(|Y_{ni} \geq \epsilon n^{\frac{1}{2}}\Big)\Big) = \int_{-\infty}^{\infty} \frac{1}{n} K^2(\frac{x-y}{h}) f(y) \mathbb{I}\Big(K(\frac{x-y}{h}) \geq \epsilon(nh)^{\frac{1}{2}}\Big)\, dy \tag{1.67}$$

$$= 0 \tag{1.68}$$

for n large enough such that $\sup_{z\in R} K(z) < e(nh)^{\frac{1}{2}}$.

Thus by the Linderberg-Feller central limit theorem, we have our required result. □

5. Bandwidth Selection

Since h_{AMISE} depends on f through $R(f'')$, we still require practical bandwidth selection algorithms.

5.1. Normal Scale rules. If f is the $N(0, \sigma^2)$ density, then $R(f'') = \frac{3}{8\sqrt{\pi}}\sigma^{-5}$. The normal scale rate $\hat{h}_{NS}$ consists of replacing $R(f'')$ in h_{AMISE} with $\frac{3}{8\sqrt{\pi}}\hat{\sigma}^{-5}$, where $\hat{\sigma}$ is an estimate of σ. This tends to over-smooth.

5.2. Least-squares Cross-Validation. Recall that

$$MISE(\hat{f}_h) = \mathbb{E}\Big(\int_{-\infty}^{\infty} \hat{f}(x)^2 dx\Big) - 2\mathbb{E}\Big(\int_{-\infty}^{\infty} \hat{f}_h(x) f(x)\Big) + \int_{-\infty}^{\infty} f(x)^2 dx. \tag{1.69}$$

Observe that it suffices to minimize the sum of the first two terms. This depend on the unknown f, but an unbiased estimate is given by $LSCV(h)$, with

$$LSCV(h) = \int_{-\infty}^{\infty} \hat{f}_h(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} f_{-i,h}(x_i) \tag{1.70}$$

with

$$\hat{f}_{-i,h}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K(\frac{x-x_j}{h}) \tag{1.71}$$

Minimization of $LSCV(h)$ yields $\hat{h}_{LSCV}$.

5.3. Biased Cross-Validation. Under regularity conditions,

$$\mathbb{E}\Big(R(\hat{f}_h)\Big) = R(f'') + \frac{R(K'')}{nh^5} + O(h^2). \tag{1.72}$$

We can therefore define

$$BCV(h) = \frac{R(K)}{nh} + \frac{1}{4}\mu_2^2(K)\widetilde{R(f'')} \tag{1.73}$$

where

$$\widetilde{R(f'')} = R(\hat{f}_{h_1}) - \frac{R(K'')}{nh_1^5} \tag{1.74}$$

with h_1 a "pilot" bandwidth (c.f Ward and Jones, 1995). Minimization of $BCV(h)$ yields $\hat{h}_{BCV}$.

5.4. Solve-the-equation Rules. Under smoothness assumptions, we can integrate by parts to obtain

$$R(f'') = \int_{-\infty}^{\infty} f''''(x)f(x)dx = \mathbb{E}(f''''(X)) \tag{1.75}$$

We can therefore estimate $R(f'')$ by using

$$\hat{R}_{h_2} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}''''_{h_2}(x_i) \tag{1.76}$$

where again h_2 is a pilot bandwidth. By exploiting the relationship between h_{AMISE} and the $AMISE$-optimal bandwidth for estimating $R(f'')$ in this way, we obtain an equation which can be solved numerically to yield $\hat{h}_{SJE}$.

6. Other Topics

6.1. Choice of Kernel. The choice of kernel is coupled with the choice of bandwidth, because if we replace $K(x)$ by $\frac{1}{2}K(\frac{1}{2})$ and we halve the bandwidth, the estimate is unchanged. We therefore fix the scale by setting $\mu_2(K) = 1$. Minimizing $AMISE(\hat{f}_h)$ over K the amounts to minimizing $R(K)$ subject to

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{1.77}$$

$$\int_{-\infty}^{\infty} xK(x)dx = 0 \tag{1.78}$$

$$\mu_2(K) = 1 \tag{1.79}$$

$$K(x) \geq 0 \tag{1.80}$$

The solution is given by the Epanechnikov kernel (1969).

$$K_E(x) = \frac{3}{4\sqrt{5}}(1 - \frac{x^2}{5})\mathbb{I}\Big(|x| \leq \sqrt{5}\Big) \tag{1.81}$$

The ratio $\frac{R(K_E)}{R(K)}$ is called the **efficiency** of a kernel K, because it represents the ratio of the sample sizes needed to obtain the same $AMISE$ when using K_E compared with K.

Kernel	Efficiency
Epachnikov	1.0
Normal	0.951
Triangular	0.986
Uniform	0.930

6.2. Derivative Estimation. A natural estimator of the r-th derivative $f^{(r)}$ of f is given by

$$\hat{f}_h^{(r)}(x) = \frac{1}{nh^{r+1}} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{1.82}$$

obtained from differentiating the standard KDE for $\hat{f}$.

Under regularity conditions,

$$MSE(\hat{f}_h^{(r)}(x)) = \frac{R(K^{(r)})}{nh^{2r+1}} f(x) + \frac{1}{4} h^4 \mu_2^2 f^{(r-2)}(x)^2 + o(\frac{1}{nh} + h^4). \tag{1.83}$$

This leads to an optimal bandwidth of order $n^{-\frac{1}{2r+5}}$ and a rate of converge of $n^{-\frac{4}{2r+5}}$.

The intuition is that estimating derivatives of densities is harder than estimating densities themselves.

6.3. Higher Order Kernels. It is possible to make the dominant integrated squared bias term of $MISE(\hat{f}_h)$ vanish by choosing $\mu_2(K) = 0$. This means we have to allow the Kernel to take negative values, so the resulting estimate need not be a density.

We can set $\hat{f}_h(x) = \max(\hat{f}_h(x), 0)$ and then renormalize, but then we lose smoothness. Nevertheless, we define K to be a k-th order kernel if writing $\mu_j(K) = \int_{-\infty}^{\infty} x^j K(x) dx$, we have

$$\mu_0(K) = 1 \tag{1.84}$$

and $\mu_j(K) = 0$ for $j = 1, \ldots, k-1$, $\mu_k(K) \neq 0$, and

$$\int_{-\infty}^{\infty} |x|^k |K(x)| dx < \infty \tag{1.85}$$

If f has k continuous bounded derivatives with $R(f^{(k)}) < \infty$, then it is shown (example sheet) that $h_{AMISE} = cn^{-\frac{1}{2k+1}}$ and

$$AMISE(\hat{f}_{h_{AMISE}}) = O(n^{-\frac{2k}{2k+1}}) \tag{1.86}$$

Thus, under increasingly strong smoothness assumptions, convergence rates arbitrarily close to the parametric rate of $O(n^{-1})$ can be obtained.

The practical benefit of higher order kernels is not always apparent, and the negativity/smoothness/bandwidth selection problems mean that they are rarely used in practice.

6.4. Local Bandwidths. Choosing $h = h(x)$ is problematic, because the resulting estimate need not be a density. However, we can define

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h(x)}(x - X_i) \tag{1.87}$$

Theory suggests that we should choose $h(X_i) = h_0 f^{-\frac{1}{2}}(X_i)$, and, with four derivatives and a second order kernel, one can attain a "fourth-order kernel" rate of $O(n^{-\frac{8}{9}})$. There is no negativity problem, but we do require pilot bandwidth selection. Difficult to tune well and rarely used in practice.

6.5. Transformation Methods. It may be that f is difficult to estimate, but it may be that we can construct a strictly increasing, continuously differentiable function t on the support of f, such that, setting $Y_i = t(X_i)$, the density of $Y_1, \ldots, Y_n$ is easier to estimate. We "back transform" the estimate to obtain

$$\bar{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(t(x) - t(X_i))t'(x) \tag{1.88}$$

6.6. Multi-Dimensional Density Estimation. The general d-dimensional kernel estimator is of the form

$$\hat{f}_H(x) = \frac{1}{n}(\det H)^{\frac{1}{2}}\sum_{i=1}^{n} K(H^{-\frac{1}{2}}(x - X_i)) \tag{1.89}$$

where H is positive definite symmetric bandwidth matrix. The difficulties of choosing the $\frac{1}{2}d(d+1)$ independent entries mean that we often restrict attention to the diagonal H, or even $H = h^2 I$. In this latter case,

$$AMISE(\hat{f}_{h^2 I}) = \frac{R(K)}{nh^d} + \frac{1}{4}h^4\mu_2^2(K)\int_{\mathbb{R}^d}\{\Delta_f(x)\}^2 dx \tag{1.90}$$

where $\Delta_f(x) = \sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_j^2}(x)$ is the Laplacian of f at x. This leads to an

$$AMISE(\hat{f}_{h^2_{AMISE} I}) = O(n^{-\frac{4}{d+4}}) \tag{1.91}$$

Thus the "curse of dimensionality", together with bandwidth selection problems, means that this is only really feasible for $d \leq 4$.

CHAPTER 2

Nonparametric Regression

1. Introduction

Nonparametric regression is a regression which doesn't assume a parametric relation between a design matrix X and the response variable Y.

In the univariate fixed design setting, the design X consists of ordered real numbers $x_1 < x_2 < \cdots < x_n$, and the response variable Y we have

$$Y_i = m(x_i) + v(x_i)^{\frac{1}{2}}\epsilon_i \tag{2.1}$$

where the ϵ_i are IID, $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{V}(\epsilon_i) = 1$.

In the random design setting, we have

$$Y_i = m(X_i) + v(X_i)^{\frac{1}{2}}\epsilon_i \tag{2.2}$$

where ϵ_i are IID, $\mathbb{E}(\epsilon_i|X_i) = 0$, and $\mathbb{V}(E_i|X_i) = 1$. m_i is the regression function that is our interest to estimate. When $v(x_i) = v$ (constant), we call it homoscedastic. If it is not, we call it heteroscedastic.

2. Local polynomial estimator

Assume a fixed design. The local polynomial estimator $\hat{m}_h(x;p)$ of degree p with kernel K with a bandwidth h is constructed by fitting a polynomial of degree p using weighted least squares. The weight $K_h(x_i - x)$ is associated with the weight (x_i, Y_i).

More precisely, $\hat{m}_h(x;p) = \hat{\beta}_0$ where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ which is minimizing

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p)^2 K_h(x_i - x) \tag{2.3}$$

where $\beta \in \mathbb{R}^{p+1}$

The theory of weighted least squares gives

$$(X^T K X)\hat{\beta} = X^T K y \tag{2.4}$$

For $p = 0$, then a simple expression (Nadaraya-Watson, local constant) exists:

$$\hat{m}_h(x;0) = \frac{\sum_{i=1}^n K_h(x_i - x)Y_i}{\sum_{i=1}^n k_h(x_i - x)} \tag{2.5}$$

For $p = 1$, we call this a local linear estimator, and we have the explicit result

$$\hat{m}_h(x;1) = \frac{1}{n}\sum_{i=1}^n \frac{S_{2,h}(x) - S_{1,h}(x)(x_i - x)}{S_{2,h}(x)S_{0,h}(x) - S_{1,h}(x)^2} K_h(x_i - x)Y_i \tag{2.6}$$

with

$$S_{r,h}(x) = \frac{1}{n}\sum_{i=1}^n (x_i - x)^r K_h(x_i - x) \tag{2.7}$$

All local polynomial estimators of the form

$$\sum_{i=1}^n W(x_i, x)Y_i \tag{2.8}$$

This type of estimator is called a linear estimator. This set of weights $\{W(x_i, x)\}$ is called the **effective kernel**.

3. MSE approximations

For convenience, let $x_i = \frac{i}{n}$. We consider the following conditions:

(i) m is twice continuously differentiable on $[0, 1]$ and is bounded, v is continuous.
(ii) $h = h_n, h_n \to 0$, $nh \to \infty$.
(iii) K is a nonnegative probability density, symmetric, has zeros outside of $[-1, 1]$. $R(K) = \int K^2(x)dx < \infty$, and $\mu_2(K) = \int xK^2(x) < \infty$.

Theorem 2.1. *Under the conditions previously, for $x \in (0, 1)$, we have*

$$MSE(\hat{m}_h(x;1)) = \frac{1}{nh}R(K)v(x) + \frac{1}{4}h^4(m''(x))^2\mu_2(K) + o(\frac{1}{nh} + h^4) \tag{2.9}$$

Proof (Sketch of proof). As usual, we use a bias^2 + variance calculation.

$$
\begin{aligned}
\text{BIAS} &= \mathbb{E}(\hat{m}_h, x; 1) - m(x) && (2.10)\\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{S_{0,h}(x) - S_{1,h}(x)(x_i - x)}{DEN}K_h(x_i - x)Y_i\right) - m(x) && (2.11)\\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{S_{2,h}(x) - S_{1,h}(x)(x_i - x)}{DEN}K_h(x_i - x)\underbrace{m(x_i)}_{m(x)+(x_i-x)m'(x)+\frac{1}{2}(x_i-x)^2m''(x)} - m(x) && (2.12)\\
&= \frac{m(x)}{DEN}\{\frac{S_{2,h}(x)S_{0,h}(x) - S_{1,h}^2(x)}{\dots}\} && (2.13)\\
&\quad + \frac{m'(x)}{DEN}\{\frac{S_{2,h}(x)S_{1,h}(x) - S_{1,h}(x)S_{2,h}(x)}{DEN}\} && (2.14)\\
&\quad + \frac{1}{2}m''(x)\{\frac{S_{2,h}^2(x) - S_{1,h}(x)S_{3,h}(x)}{S_{2,h}(x)S_{0,h}(x) - S_{1,h}^2(x)}\} - m(x) && (2.15)\\
&= m(x) + 0 + \frac{1}{2}m''(x)\{\frac{(h^2\mu_2(K) + o(h^2))^2 - o(h)o(h^3)}{h^2\mu_2(K)(1 + o(1)) - o(h^2)}\} - m(x) && (2.16)\\
&= m(x) + \frac{1}{2}m''(x)\frac{h^4\mu_2^2(K) + o(h^4)}{h^2\mu_2(K) + o(h^2)} - m(x) && (2.17)\\
&= m(x) + \frac{1}{2}m''(x)h^2\mu_2(K) + o(h^2) + REM - m(x) && (2.18)\\
&= \frac{1}{2}m''(x)h^2\mu_2(K) + o(h^2) && (2.19)
\end{aligned}
$$

since $|REM| = o(h^2)$. Note that we have

$$
\begin{aligned}
S_{r,h}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^r K_h(x_i - x) && (2.20)\\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^r \frac{1}{h}K(\frac{x_i - x}{h}) && (2.21)\\
&= \frac{1}{nh}h^r\sum_{i=1}^{n}(\frac{x_i - x}{h})^r K(\frac{x_i - x}{h}) && (2.22)\\
&= h^r\{\int_{-1}^{1}u^r K(u)du + o(1)\} && (2.23)\\
&= h^r\mu_r(K) + o(h^r) && (2.24)
\end{aligned}
$$

from bounded support of K, with $\frac{|x_i - x|}{h} \leq 1$.

For the variance, we need the preliminary calculations that

$$t_{r,h}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^r K_h^2(x_i - x) \tag{2.25}$$

$$= h^{r-1}\mu_r(K^2) + o(h^{r-1}) \tag{2.26}$$

$$\mathbb{V}(\hat{m}_h(x;1)) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{S_{2,h}(x) - S_{1,h}(x)(x_i - x)}{DEN}\right)^2 K_h^2(x_i - x)v(x_i) \tag{2.27}$$

$$= \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\frac{S_{2,h}^2(x) - 2(x_i - x)S_{1,h}(x)S_{2,h}(x) + (x_i - x)^2 S_{1,h}^2(x)}{DEN^2}K_h^2(x_i - x)v(x) + REM_2 \tag{2.28}$$

$$= \frac{1}{n}\frac{S_{2,h}^2(x)t_{0,h}(x) - 2S_{1,h}(x)S_{2,h}(x)t_{1,h}(x) + S_{1,h}^2(x)t_{2,h}(x)}{DEN}v(x) + REM_2 \tag{2.29}$$

$$= \frac{v(x)}{n}\frac{(h^2\mu_2(K) + o(h^2))^2(h^{-1}\mu_0(K^2) + o(h^{-1})) - 2o(h)(h^2\mu_2(K) + o(h^2))(\mu_1(K^2) + o(1)) + o(h^2)(h\mu_2(K}{(h^2\mu_2(K)(1 + o(1)) + o(h^2))^2} \tag{2.30}$$

$$= \frac{v(x)}{n}\frac{h^3\mu_2^2(K)\mu_0(K^2) + o(h^3)}{h^4\mu_2^2(K) + o(h^4)} + REM_2 \tag{2.31}$$

$$= \frac{v(x)}{n}\frac{1}{h}R(K) + o\left(\frac{1}{nh}\right) \tag{2.32}$$

where $|REM_2| = o(\frac{1}{nh})$ With some further work, we can integrate term by term the asymptotic expansion to obtain $MISE(\hat{m}(\cdot;1))$. $\square$

For p even, the bias is more complicated. Moreover, for p even, the bias at boundary point $x = \alpha h$, $\alpha \in [0, 1)$ has larger order than the bias at the interior point.[1]

4. Splines

4.1. Motivation. Let $n \geq 3$, and consider for a fixed homoscedastic design

$$Y_i = m(x_i) + \sigma\epsilon_i \tag{2.33}$$

where ϵ_i are IID with $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{V}(\epsilon_i) = 1$.

Another natural idea to estimate the regression curve m is to balance the fidelity of the fit to the data and the roughness of the resulting curve. This can be done by minimizing

$$\sum_{i=1}^{n}(Y_i - \tilde{g}(x_i))^2 + \lambda\int \tilde{g}''(x)^2 dx \tag{2.34}$$

[1]In the demonstration, asymmetry of BETA$(2, 4)$ distributions combined with the negative slope of the true regression function, we see that local constant estimators has an upward bias. In contrast, local linear estimators adapts to this

over $\tilde{g} \in S_2[a,b]$, the set of twice continuously differentiable functions on $[a,b]$. λ is a regularization parameter. As $\lambda \to \infty$, the curve is very close to the linear regression line. As $\lambda \to 0$, the resulting curve closely fits the observations.

4.2. Cubic Spline.

DEFINITION 2.2. A cubic spline is a function $g : [a,b] \to \mathbb{R}$ satisfies

(i) g is a cubic polynomial on $[(a,x_1),(x_1,x_2),\dots,(x_n,b)]$.
(ii) g is twice continuously differentiable on $[a,b]$.

PROPOSITION 2.3. *For a given* $\mathbf{g} = (g_1,\dots,g_n^T)$*, there exists a unique natural cubic spline* g *with knots* $x_1,\dots,x_n$ *- so* $g(x_i) = g_i$ *for* $i = 1,\dots,n$*. Moreover, there exists a nonnegative definite matrix* K *such that*

$$\int_a^b g''(x)^2 dx = g^T K g \tag{2.35}$$

We call g *the* ***natural cubic spline*** *interpolant to* g *at* $x_1,\dots,x_n$*.*

THEOREM 2.4. *For any* $\tilde{g} \in S_2[a,b]$ *satisfying* $\tilde{g}(x_i) = g_i, i = 1,\dots,n$*, the cubic spline interpolant to* g *at* $\mathbf{g} = g_1,\dots,g_n$ *uniquely minimizes*

$$\int_a^b \tilde{g}''(x)^2 dx \tag{2.36}$$

over $\tilde{g} \in S_2[a,b]$*.*

PROOF. Let $\tilde{g} \in S_2[a,b]$ satisfy $\tilde{g}(x_i) = g_i$, $i = 1,\dots,n$. Let $h = \tilde{g} - g$ such that $h(x_i) = 0$. Then

$$R(\tilde{g}'') = \int_a^b (h'' + g'')^2 dx = R(h'') + R(g'') + 2\int_a^b h''(x)g''(x)dx \tag{2.37}$$

Then

$$\begin{aligned}
\int_a^b h''(x)g''(x) &= -\int_a^b g'''(x)h'(x)dx + g''h'(x)|_a^b && (2.38)\\
&= -\int_{x_1}^{x_n} g'''(x)h'(x)dx && (2.39)\\
&= -\sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}} g'''(x)h'(x)dx && (2.40)\\
&= -\sum_{i=1}^{n-1} g'''(x_i)\int_{x_i}^{x_{i+1}} h'(x)dx && (2.41)\\
&= -\sum_{i=1}^{n} g'''(x_{i+1})(h(x_{i+1}) - h(x_i)) && (2.42)\\
&= 0 && (2.43)
\end{aligned}$$

since $g''(x) = 0$ at a and b.

Thus,

$$R(\tilde{g}'') = R(g'') + R(h'') \geq R(g'') \tag{2.44}$$

with equality when $R(h) = 0 \iff h$ is linear on (x_i, x_{i+1}), with $h(x_{i+1}) = h(x_i) = 0$. Thus, $h \equiv 0$. $\square$

4.3. Natural Cubic Smoothing Spline. Recall that $Y_i = m(x_i) + \sigma\epsilon_i$, $m \in S_2[a,b]$, $0 < x_1 < \cdots < x_n < b$. We seek to minimize

$$\mathcal{G}_\lambda(\tilde{g}) = \sum_{i=1}^{n}(Y_i - \tilde{g}(x_i))^2 + \lambda \int_a^b \tilde{g}''(x)^2 dx \tag{2.45}$$

over $\tilde{g} \in S_2[a,b]$.

THEOREM 2.5. *For each $\lambda > 0$, there is a unique solution $\hat{g}$ minimizing $\mathcal{G}(\tilde{g})$over $\tilde{g} \in S_2[a,b]$. This is the natural cubic spline*

$$\hat{g} = (I + \lambda K)^{-1} Y \tag{2.46}$$

PROOF. Suppose $\tilde{g}$ is not a natural cubic spline. Then, there exists a unique natural cubic spline interpolant g to $\tilde{g}(x_1, \ldots, \tilde{g}(x_n))$. Then, by the previous theorem, we know

$$\int_a^b g''(x)^2 dx < \int_a^b \tilde{g}''(x)^2 dx \Rightarrow \mathcal{G}(g) > \mathcal{G}_\lambda(g) \tag{2.47}$$

We may therefore suppose g as a natural cubic spline.

Let $\mathbf{g} = (g(x_1), \ldots, g(x_n))$. Then

$$\begin{aligned}
\mathcal{G}_\lambda(g) &= (Y - \mathbf{g})^T (Y - \mathbf{g}) + \lambda g^T K g && (2.48)\\
&= Y^T Y - 2\mathbf{g}^T Y + \mathbf{g}^T \mathbf{g} + \lambda \mathbf{g}^T K \mathbf{g} && (2.49)\\
&= \mathbf{g}^T (I + \lambda K)\mathbf{g} + Y^T Y - 2\mathbf{g}^T Y && (2.50)\\
&= (\mathbf{g} - (I + \lambda K)^{-1} Y)^T (I + \lambda K)(g - (I + \lambda K)^{-1} Y) && (2.51)\\
&\quad + Y^T Y - Y^T (1 + \lambda K)^{-1} Y && (2.52)
\end{aligned}$$

We know K is nonnegative definite and $\lambda > 0$, so $I + \lambda K$ is positive definite.

Thus $\mathcal{G}_\lambda(g)$ is uniquely minimized by $\hat{g} = (I + \lambda K)^{-1} Y$.

$\square$

We call $\hat{g}$ that **natural cubic smoothing spline with data** (x_i, Y_i).

4.4. Choice of λ. Cross validation method validates the estimated curve without the i-th observation by comparing the i-th value

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i,\lambda}(x_i))^2 \tag{2.53}$$

where $\hat{g}_{-i,\lambda}$ is chosen by minimizing $\mathcal{G}_\lambda$ over all data points except the i-th,

$$\sum_{j\neq i}^{n}(Y_j - \tilde{g}(x_j))^2 + \lambda\int_a^b \tilde{g}''(x)^2dx \tag{$\star$}$$

Theorem 2.6.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - \hat{g}_\lambda(x_i)}{1 - A_{ii}})^2 \tag{2.54}$$

where $A = (I + \lambda K)^{-1}$ *and*

$$\int_{-\infty}^{\infty}\hat{g}_\lambda''(x)^2dx = \hat{g}_\lambda(\mathbf{x})^T K \hat{g}_\lambda(\mathbf{x}) \tag{2.55}$$

Proof. Note that $\hat{g}_{-i,\lambda}$ also minimizes

$$\hat{g}_{-i,\lambda}(x_i) - \tilde{g}(x_i)^2 + (\star) \tag{$\star\star$}$$

over $\tilde{g} \in S_2[a,b]$.

Then

$$\begin{aligned}
(\star\star) &\geq (\star) && (2.56)\\
&\geq \sum_{j\neq i}^{n}(Y_j - \hat{g}_{-i,\lambda}(x_j))^2 + \int_a^b \hat{g}_{-i,\lambda}(x)^2dx && (2.57)\\
&= (\hat{g}_{-i,\lambda}(x_i) - \hat{\mathbf{g}}_{-\mathbf{i},\lambda})^2 + \sum_{j\neq i}^{n}(Y_i - \hat{\mathbf{g}}_{-\mathbf{i},\lambda}(x_j)) + \int_a^b \mathbf{g}_{\mathbf{i},\lambda}(x)^2dx && (2.58)
\end{aligned}$$

Note that $(\star\star) = \sum_{j=1}^{n}(Y_j^{[i]} - \tilde{g}(x_i))^2 + \lambda\tilde{g}''(x)^2dx$ where

$$Y_j^{[i]} = \begin{cases} Y_j & i \neq j \\ \hat{g}_{-i,\lambda}(x_i) & i = j \end{cases} \tag{2.59}$$

Then, we can see that $(\star\star)$ has the same form as the original problem, so

$$\hat{g}_{-i,\lambda} = (I + \lambda K)^{-1}Y^{[i]} = AY^{[i]} \tag{2.60}$$

$$\hat{g}_{-i,\lambda}(x_i) = \sum_{j=1}^{n}A_{ij}Y_j^{[i]} = A_{ii}\hat{g}_{-i,\lambda}(x_i) + \sum_{j\neq i}A_{ij}Y_j. \tag{2.61}$$

and so

$$\hat{g}_{-i,\lambda}(x_i) = \frac{\sum_{j\neq i} A_{ij} Y_j}{1 - A_{ii}}. \tag{2.62}$$

□

Therefore

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \frac{\sum_{j\neq i} A_{ij} Y_j}{1 - A_{ii}})^2 \tag{2.63}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - \sum_{j=1}^{n} A_{ij} Y_j}{1 - A_{ii}}) \tag{2.64}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - \hat{g}_\lambda(x_i)}{1 - A_{ii}})^2. \tag{2.65}$$

By replacing A_{ii} with the average of diagonal elements of A, we have a generalized cross-validation

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{g}_\lambda(x_i)}{1 - \frac{1}{n}\operatorname{Tr} A}\right)^2 \tag{2.66}$$

A_{ii} is analogous to the leverage of the i-th observation in the linear regression. Modified (GCV) CV down-weights observations with high leverage.

Consider the model $Y_i = m(x_i) + \sigma\epsilon_i$, with fixed design. m is twice continuously differentiable on $[a, b]$, so

$$\sum_{i=1}^{n}(Y_i - \tilde{g}(x_i))^2 + \lambda\int \tilde{g}''(x)^2 dx \tag{2.67}$$

with $\tilde{g} \in S_2[a, b]$.

Cubic spline can be expanded with truncated power series basis functions: $1, x, x^2, x^3, (x - x_1)_+^3, \ldots (x - x_n)_+^3$, ($n$ number of basis functions can be obtained — see example sheet).

4.5. Regression Spline and Penalized Spline. One possible issue with cubic spline is that we need to estimate parameters of dimension n. One possible solution is to use a smaller number of knots — say N — and locate them at $\xi_1, \ldots, \xi_N$. Then, we fit the curve using standard least squares, and so minimize

$$\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{p}\beta_j x_i^j - \sum_{j=1}^{N}\beta_{pj}(x_i - \xi_j)_+^p)^2 \tag{2.68}$$

over $\beta = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{p1}, \ldots, \beta_{pN})^T \in \mathbb{R}^{p+1+N}$

Using a matrix form, this is equivalent to $\|Y - X\beta\|_2^2$, where

$$X = \begin{Bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p & (x_1 - \xi_1)_+^p & \ldots & (x_1 - \xi_N)_+^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p & (x_2 - \xi_1)_+^p & \ldots & (x_2 - \xi_N)_+^p \end{Bmatrix} \tag{2.69}$$

The solution $\hat{\beta} = (X^T X)^{-1} X^T Y$ gives the estimated curve at the observations $\mathbf{x} = (x_1, \ldots, x_n)$. The points

$$((x_1, (X\hat{\beta})_1), \ldots, (x_n, (X\hat{\beta})_n)) \tag{2.70}$$

give the fitted curve. The curve corresponding to $\hat{\beta}$ is called the **regression spline** of order p with knots at $(\xi_1, \ldots, \xi_N)$.

It is recommended to use $N = \min(\frac{n}{4}, 35)$ and locate the k-th knot at $(\frac{k}{N+1})$-th sample quantile of design points.

Computationally, it is better to use the equivalent β-splines (de Boor, 1978).

Note that N is playing the role of a smoothing parameter that controls the bias-variance tradeoff. Higher N reduces the bias but increases the variance.

An alternative to choosing N is to use large N but penalize large estimated coefficients. That is, we add a penalty term $\lambda B^T DB$ where D is a $(p+1+N \times p+1+N)$ matrix with all elements zero except the bottom-right $N \times N$ block, which is the I_N, the N-dimensional identity matrix.

We have that this then has the solution $\hat{\beta}_\lambda = (X^T Y + \lambda D)^{-1} X^T Y$.

The fitted curve corresponding to $\hat{\beta}_\lambda$ is called the **penalized spline** of order p with knots $(\xi_1, \ldots, \xi_N)$.

4.6. Equivalent Kernel. From the solution $\hat{g}_\lambda(\mathbf{x}) = (I + \lambda K)^{-1} Y$, we have

$$\hat{g}_\lambda(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \tag{2.71}$$

where the $W_{ni}(x)$ does not depend on Y_i.

Connections between smoothing splines and kernel regression estimators is established by Silverman (1984). He proved that under some regularity conditions, and random design,

$$W_{ni}(x) \approx \frac{1}{nf(x_i)} \mathcal{K}_{h(x_i)}(X_i - x) \tag{2.72}$$

where f is a density of distribution of X, $h(X_i) = (\frac{n}{f(X_i)})^{\frac{1}{4}}$, and

$$\mathcal{K}(t) = \frac{1}{2} \exp(-\frac{|t|}{\sqrt{2}}) \sin(\frac{|t|}{\sqrt{2}} + \frac{\pi}{4}) \tag{2.73}$$

This provides intuition to help understand how smoothing splints assign weights to x near the observations.

We have $\hat{m}_h(x; 1) = \sum_{i=1}^{n} W(x_i, x) Y_i$ where $W(x_i, x) = \frac{1}{nf(X_i)} K_h(x_i - x)$.

5. Multivariate Regression and Additive Models

A d-dimensional nonparametric regression suffers the same curse of dimensionality as we saw in kernel density estimation.

However, if m is smooth around $x_0 \in \mathbb{R}^d$, so $m(x) \approx m(x_0) + \sum_{i=1}^{d} (x_j - x_{0j}) \frac{\partial}{\partial x_j} m(x_0)$.

This motivates us to use

$$Y_i = \alpha + \sum_{j=1}^{d} g_j x_{ij} + \epsilon_i, i = 1, \ldots, n \tag{2.74}$$

and we minimize

$$\sum_{i=1}^{n}(Y_i - \alpha - \sum_{j=1}^{d} g_j(x_{ij}))^2 + \sum_{j=1}^{d} \lambda_j \int g_j''(x)^2 dx \tag{2.75}$$

Note that $g_j(x_{ij}) = Y_i - \alpha - \sum_{k \neq j} g_k(x_{ik})$.

We have then a back-fitting algorithm that solves the minimization problem

(i) $\hat{\alpha} = 0, \hat{g}_j = 0, j = 1, \ldots, d.$

(ii) For $j = 1, \ldots, d$,

$$\hat{g}_j = \textsc{Smooth}((x_i, Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{g}_k(x_{ik}))\forall i \tag{2.76}$$

and $\hat{g}_j = \hat{g}_j - \frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(x_{ij})$

(iii) Repeat until convergence.

CHAPTER 3

Nearest Neighbor Classification

We have $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \{0, 1\}$. The regression function $\mathbb{E}(Y|X = x)$ is denoted by $\nu(x)$, and we let μ be the distribution of X - so $\mathbb{P}(X \in A) = \mu(A)$.

A function $g : \mathbb{R}^d \to \{0, 1\}$ is called a classifier. If the distribution of (X, Y) are known, we can minimize the risk $\mathbb{P}(g(X) \neq Y) = L(g)$ over $g : \mathbb{R}^d \to \{0, 1\}$. The minimizer $g^\star$ is called a Bayes classifier, and $L(g^d)$ is called the Bayes risk.

LEMMA 3.1. *For a classifier $\tilde{g}$ which has the form*

$$\tilde{g}(x) = \begin{cases} 1 & \hat{\nu}(x) > \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases} \tag{3.1}$$

we have

$$\mathbb{P}(\tilde{g}(X) \neq Y) - L^\star \leq 2\mathbb{E}(\|\hat{\nu}(X) - \nu(X)) \tag{3.2}$$

When we have data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we want to construct a sequence of classifiers $\{g_n\}$ such that the risk using g_n is close to the Bayes risk with high probability.

DEFINITION 3.2 (k-nearest neighbor classification). A k-NN classifier g_n is defined by

$$g_n(x) = \begin{cases} 1 & \sum_{i=1}^n W_{ni}(X)\mathbb{I}(Y_i = 1) > \sum_{i=1}^n W_{ni}(X)\mathbb{I}(Y_i = 0) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

which is equivalent to

$$\sum_{i=1}^n W_{ni}(X)\mathbb{I}(Y_i = 1) > \frac{1}{2} \iff \sum_{i=1}^n W_{ni}(X)Y_i > \frac{1}{2} \tag{3.4}$$

where

$$W_{ni}(X) = \frac{1}{k} \tag{3.5}$$

if X_i is a k-nearest neighbor of X, and zero otherwise.

DEFINITION 3.3. For a certain distribution of (X, Y), we say g_n is consistent if $\mathbb{P}(g_n(X) \neq Y) - L^\star \to 0$.

We say g_n is strongly consistent if

$$\mathbb{P}\Big(\lim_{n\to\infty} L(g_n) = L(g^\star)\Big) = 1 \tag{3.6}$$

THEOREM 3.4. *If $k \to \infty$, $\frac{k}{n} \to 0$, then for all distributions of (X, Y), the k-NN estimates g_n are consistent.*

PROOF. Preliminaries:

(i) By a corollary of Lemma 1,

$$\mathbb{P}(g_n(X) \neq Y|D_n) - L^\star \leq 2\sqrt{\int_{\mathbb{R}^d} (\eta_n(x) - \eta(x))^2 d\mu(x)} \tag{3.7}$$

(ii) If $k \to \infty$, $\frac{k}{n} \to 0$, then

$$\|X_{(k)}(X) - X\| \overset{as}{\to} 0 \tag{3.8}$$

(examples class)

(iii) Stones Lemma - for any integrable function f, any n,

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}(|f(X_i(X))|) \leq \gamma_d \mathbb{E}(|f(X)|) \tag{3.9}$$

where γ_d is a constant only depending on d.

We can now complete the proof. By the first result, it suffices to prove

$$\mathbb{E}\big((\eta_n(X) - \eta(X))^2\big) \to 0 \tag{3.10}$$

with $\eta_n(X) = \sum_{i=1}^n W_{ni}(X)Y_i$.

Recall that $\eta_n(X) = \sum_{i=1}^n W_{ni}(X)Y_i$ and $W_{ni}(X)$ is $\frac{1}{k}$ if and only if X_i is among the k-nearest neighbors of X. In order to use the bias-variance decomposition, let $\mathbb{E}(\eta_n(X)|X, X_1, \dots, X_n) = \sum_{i=1}^n W_{ni}(X)\eta(X_i) := \tilde{\eta}(X_i)$. Then

$$\mathbb{E}\big((\eta_n(X) - \eta(X))^2\big) \leq 2\mathbb{E}\big((\eta_n(X) - \tilde{\eta}(X))^2\big) + 2\mathbb{E}\big((\tilde{\eta}(X) - \eta(X))^2\big) \tag{3.11}$$

or 2 time variance + 2 times Bias squared.

As $\sum_{i=1}^n W_{ni}(X) = 1$, and Cauchy-Swartz, we have

$$\text{BIAS}^2 = \mathbb{E}\left((\sum_{i=1}^n W_{ni}(X)(\eta(X_i) - \eta(X)))^2 \right) \tag{3.12}$$

$$\leq \mathbb{E}\left((\sum_{i=1}^n W_{ni}(X)(\eta(X_i) - \eta(X))^2) \right) \tag{3.13}$$

Now, consider a continuous function $0 \leq \eta^\star \leq 1$ which approximates η such that (there exists $\eta^\star$ since a continuous function is dense in $L^2(\mu)$), $\mathbb{E}\big((\eta^\star(X) - \eta(X))^2\big) \leq \epsilon$.

Also, we require $\eta^\star$ satisfies (using uniform continuity of $\eta^\star$) that, for a given $\epsilon > 0$, there exists $\delta > 0$ such that $(\eta^\star(x) - \eta^\star(y))^2 leq\epsilon$ when $\|x - y\| \leq \delta$. Then, by using the previous result, uniform continuity of $\eta^\star$, and the approximating property of $\eta^\star$ for each three splitted terms,

$$\text{BIAS}^2 \leq \mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X)(\eta(X_i) - \eta(X))^2\right) \tag{3.14}$$

$$\leq 3\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}((\eta(X_i) - \eta^\star(X_i))^2 + (\eta^\star(X_i) - \eta^\star(X))^2 + (\eta^\star(X) - \eta(X))^2)\right) \tag{3.15}$$

$$\leq 3((\gamma_d \mathbb{E}\big((\eta(X) - \eta^\star(X))^2\big) + \sum_{i=1}^{n} W_{ni}(X)(\epsilon + \mathbb{I}(\|X_i - X\| > \delta))) + \epsilon) \tag{3.16}$$

$$\leq 3(\gamma_d\epsilon + 2\epsilon + \sum_{i=1}^{n} W_{ni}(X)\mathbb{I}(\|X_i - X\| > \delta)) \tag{3.17}$$

$$\to 0. \tag{3.18}$$

For the variance term, we use the fact that for $i \neq j$, $\mathbb{E}((Y_i - \eta(X_i))(Y_j - \eta(X_j))|X, X_1, \ldots, X_n) = 0$. Then

$$\text{VARIANCE} = \mathbb{E}\big((\eta_n(X) - \tilde{\eta}(X))^2\big) \tag{3.19}$$

$$= \mathbb{E}\left((\sum_{i=1}^{n} W_{ni}(X)(Y_i - \eta(X_i)))^2\right) \tag{3.20}$$

$$= \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(W_{ni}(X)W_{nj}(X)(Y_i - \eta(X_i)(Y_j - \eta(X_j))))|X, X_1, \ldots, X_n\right)\right) \tag{3.21}$$

$$= \mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X)^2(Y_i - \eta(X_i))^2\right) \tag{3.22}$$

$$\leq \mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X)^2\right) \tag{3.23}$$

$$\leq \mathbb{E}\left(\max_i W_{ni}(\sum_{i=1}^{n} W_{ni}(X))\right) \tag{3.24}$$

$$= \mathbb{E}\Big(\max_i W_{ni}\Big) \tag{3.25}$$

$$= \frac{1}{k} \to 0. \tag{3.26}$$

where the second last line follows as $|Y_i - \eta(X_i)| \leq 1$. $\square$

CHAPTER 4

Minimax Lower Bounds

As a first attempt to understand a nonparametric estimation problem, we consider a minimax risk,

$$R(\Theta) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\tilde{j}, \theta). \tag{4.1}$$

If we can find our $\hat{\theta}^\star$, which minimizes $\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\tilde{\theta}, \theta)$ we call $\hat{\theta}^\star$ our minimax estimator. However, it is very difficult to find $\hat{\theta}^\star$. Let $c\gamma_n \leq R(\Theta) \leq C\gamma_n$, we call γ_n is minimax rate of convergence.

For instance, for $\Theta = \{m$, m is twice continuously differentiable on $[0, 1]$, $m''(x) < \infty\}$, then

$$\sup_{m \in \Theta} \mathbb{E}\big((\hat{m}_h(x; 1) - m(x))^2\big) \leq Cn^{-\frac{4}{5}} \tag{4.2}$$

Question — can we also calculate

$$\int_{\tilde{m}} \sup_{m \in \Theta} \mathbb{E}\big((\tilde{m}(x_0) - m(x_0))^2\big) \geq cn^{-\frac{4}{5}} \tag{4.3}$$

Lemma 4.1 (Le Cam's two points lemma). *Let $\mathcal{P}$ be probability measures on $(\mathcal{X}, \mathcal{A})$, and let (Θ, d) be the pseudo-metric space, with*

$$d : \Theta \times \Theta \to [0, \infty) \tag{4.4}$$

given by

$$d(\theta_1, \theta_2) = d(\theta_2, \theta_1), d(\theta_1, \theta_2) + d(\theta_2, \theta_3) \geq Ad(\theta_1, \theta_3) \tag{4.5}$$

Let $\theta : \mathcal{P} \to \Theta$, $\theta(P)$ is the parameter of interest ($P \in \mathcal{P}$). With $\theta_0 = \theta(P_0)$, $\theta_1 = \theta(P_1)$, under two conditions,

(i) $d(\theta_0, \theta_1) \geq \delta > 0$,
(ii) $h^2(P_0, P_1) \leq C < 1$

where $h^2(P_0, P1)$ is the Hellinger distance $\int(\sqrt{dP_0} - \sqrt{dP_1})^2$, the we have for all estimators $\tilde{\theta}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P d(\tilde{\theta}, \theta(P)) \geq \frac{A\delta}{2}(1 - \sqrt{C}) \tag{4.6}$$

PROOF.

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P d(\tilde{\theta},\theta(P)) \geq \max_{P\in\{P_0,P_1\}} \mathbb{E}_P d(\tilde{\theta},\theta(P)) \tag{4.7}$$
$$\geq \frac{1}{2}(\mathbb{E}_{P_0} d(\tilde{\theta},\theta(P_0)) + \mathbb{E}_{P_1} d(\tilde{\theta},\theta(P_1))) \tag{4.8}$$
$$\tag{4.9}$$

Let $d(\tilde{\theta},\theta(P_0)) + d(\tilde{\theta},\theta(P_1)) = DEN$, and $\frac{d(\tilde{\theta},\theta(P_0))}{DEN} = f_0$, $\frac{d(\tilde{\theta},\theta(P_1))}{DEN} = f_1$.
Note that $DEN \geq Ad(\theta(P_0),\theta(P_1)) \geq A\delta$, by our assumptions.
Then our RHS is given as

$$\frac{1}{2}(\mathbb{E}_{P_0}(f_0 \cdot DEN) + \mathbb{E}_{P_1}(f_1 \cdot DEN)) \geq \frac{1}{2}A\delta(\mathbb{E}_{P_0} f_0 + \mathbb{E}_{P_1} f_1) \tag{4.10}$$

By the Neyman-Pearson lemma, we have

$$\geq \frac{1}{2}A\delta \int \min(P_0,P_1) = \frac{1}{2}A\delta(1 - TV(P_0,P_1)) \tag{4.11}$$

From the third example sheet, we can show that $(TV(P_0,P_1))^2 \leq h^2(P_0,P_1)$. By assumption, this is bounded above by C. Using this result, we have

$$\geq \frac{1}{2}A\delta(1-\sqrt{C}) \tag{4.12}$$

□

REMARK 4.2. *From the proof,*

(i) Sample size n does not seem to appear in the lemma. However, P is usually the joint distribution of n samples. Thus, the condition on the Hellinger distance gives some conditions on n.
(ii) The two conditions work also in the opposite direction.
(iii) We can extend the two points lemma to the multiple testing case.

THEOREM 4.3 (Nonparametric regression). *Let $Y_i = m(x_i) + \epsilon_i$, $\epsilon_i \sim N(0,1)$, $x_i = \frac{i}{n}$, $m \in \Theta$ with Θ the set of all twice continuously differentiable functions on $[0,1]$, $m''(x) < \infty$. Then for any estimator $\tilde{m}$ and any $x_0 \in [0,1]$,*

$$\sup_{m\in\Theta} \mathbb{E}\big((\tilde{m}(x) - m(x_0))^2\big) \geq Cn^{-\frac{4}{5}} \tag{4.13}$$

PROOF. Let $\mathcal{P}$ be the set of distributions of $Y_1,\ldots,Y_n$ with $Y_i = m(x_i) + \epsilon_i$ and $\epsilon_i \sim N(0,1)$, $m \in \Theta$. Let Θ be as given before.

Then using $(x-y)^2 + (y-z)^2 \geq \frac{1}{4}(x-z)^2$, we have

$$d(m_0,m_1) = (m_0(x_0) - m_1(x_0))^2 \tag{4.14}$$

with $A = \frac{1}{4}$.

Let $m_0 = 0$ on $x \in [0,1]$. Let m_1 be bounded away from zero at some point $x_0 > 0$. Thus $m_1(x) = h^2 K(\frac{x-x_0}{h})$, where $K(t) = a\exp(-\frac{1}{1-t^2})$ for $t \le 1$ and a a normalizing constant so $K(t)$ is a kernel, and let $h = \tilde{c}n^{-\frac{1}{5}}$.

Let P_0 be the distribution of $Y_1, \dots, Y_n$, with $Y_i = m_0(x_i) + \epsilon_i = \epsilon_i$, and P_1 be the equivalent with $Y_i = m_1(x_i) + \epsilon_i$.

Checking the first condition, we have $(d(m_0, m_1)) = (h^2 K(0))^2 = h^2 a^2 \exp(-2) = \delta$. Checking the second condition, we have

$$\begin{aligned}
h^2(P_0, P_1) &\le KL(P_0, P_1) && (4.15)\\
&= \int \dots \int \prod_{i=1}^n \phi(u_i) \log \frac{\prod_{i=1}^n \phi(i_i)}{\prod_{i=1}^n \phi(u_i - m_1(x_i))} du_1 \dots du_n && (4.16)\\
&= \int \dots \int \prod_{i=1}^n \phi(u_i) \sum_{i=1}^n \log\exp(-u_i m_1(x_i) + \frac{1}{2} m_1(x_i)^2) && (4.17)\\
&= \int \dots \int \prod_{i=1}^n \phi(u_i) \sum_{i=}^n (-u_i m_1(x_i) + \frac{1}{2} m_1(x_i)^2) du_1 \dots du_n && (4.18)\\
&= \frac{1}{2} \sum_{i=1}^n m_1(x_i)^2 && (4.19)\\
&= \frac{1}{2} \sum_{i=1}^n h^4 a^2 \exp^2(-\frac{1}{1-(\frac{x_i - x_0}{h})^2}) \mathbb{I}(|x_i - x_0| \le h) && (4.20)\\
&\le \frac{1}{2} h^4 a^2 \sum_{i=1}^n \mathbb{I}(x_0 - h \le x_i \le x_0 + h) && (4.21)\\
&\le \frac{1}{2} h^4 a^2 2nh && (4.22)\\
&= a^2 n h^5 && (4.23)
\end{aligned}$$

and as $h \sim n^{-\frac{1}{5}}$, we have our result. with the conclusion that this is bounded by $\frac{1}{8}\delta(1 - \frac{1}{\sqrt{2}})$. $\square$

CHAPTER 5

Extreme Value Theory

Let X_n be an IID sample from a distribution function F, and denote $X_{(n)} = \max\{X_1, \dots, X_n\}$ as the maximum order statistic.

Without any normalization, $X_{(n)} \to x_\star = \inf\{x : F(x) = 1\}$.

This is not overly interesting, since the limit distribution is degenerate (we call F non-degenerate if there does not exists $a \in \mathbb{R}$ such that $F(x) = \mathbb{I}(x \geq a)$)

We may ask if there exists $\{a_n\} > 0$, $\{b_n\} > 0$, and a non-degenerate G such that

$$\mathbb{P}\left(\frac{X_{(n)} - b_n}{a_n} \leq x\right) \to G(x) \tag{5.1}$$

for all continuity points x of G

Classical extreme value theory starts by asking:

(i) What kind of G appears in the limit of (5.1)?

(ii) Can we characterize F such that (5.1) holds for a specific limit distribution G?

For the first question, we have the Extremal Types theorem. For the second question, we have the "domain of attraction" problem.

1. Preliminaries

Recall that $\mathbb{P}\big(X_{(n)} \leq x\big) = F(x)^n$. We say that F is in the domain of attraction of G ($F \in D(G)$) if there exists $\{a_n\} > 0$, $\{b_n\}$ and a non-degenerate G such that

$$\mathbb{P}\left(\frac{X_{(n)} - b_n}{a_n} \leq x\right) = [F(a_n x + b_n)^n \to G(x) \text{ for all continuity points } x \text{ of } G]. \tag{5.2}$$

and write $F(a_n x + b_n)^n \hookrightarrow G(x)$.

We say that G_1 and G_2 are of same type if $G_1(ax + b) = G_2(x)$ for some $a > 0, b$.

The next lemma shows that if $F \in D(G_1)$ and $F \in D(G_2)$, then G_1 and G_2 are of the same type.

Lemma 5.1. *Suppose X_n is an IID sample from F and there exists $\{a_n\} > 0, \{b_n\}$ and non-degenerate G such that $F(a_n x + b_n)^n \hookrightarrow G(x)$. Then there exists $\{\alpha_n\} > 0, , \{\beta_n\}$ and non-degenerate $G_\star$ such that $F(\alpha_n x_{\beta_n})^n \hookrightarrow G_\star(x)$. if and only if $\frac{\alpha_n}{a_n} \to a$ for some $a > 0$, and $\frac{\beta_n - \beta}{a_n} \to b$ for some b.*

Then we can let $G_\star(x) = G(ax+b)$.

Proof. See Galambos (1978), Lemma 2.2.3 $\square$

Definition 5.2. G is **max-stable** if for every $n \in \mathbb{N}$, there exists $\{a_n\} > 0, \{b_n\}$ such that $G^n(a_n x + b_n) = G(x)$

Theorem 5.3. *$D(g)$ is non-empty if and only if G is max-stable.*

Proof. ($\Leftarrow$) If G is max-stable, $G^n(a_n x + b_n) \hookrightarrow G(x)$. Thus, by definition, $G \in D(G)$.

($\Rightarrow$) Let $F \in D(G)$. Then, there exists $\{a_n\} > 0, \{b_n\}$ such that $F^n(a_n x + b_n) \hookrightarrow G(x)$. For each $k \in \mathbb{N}$, we replace n by nk, and then

$$F^{nk}(a_{nk}x + b_{nk}) \hookrightarrow G(x) \tag{5.3}$$

Thus $F^n(a_{nk}x + b_{nk}) \hookrightarrow G^{\frac{1}{k}}(x)$. Since $G^{\frac{1}{k}}$ is also non-degenerate, $G^{\frac{1}{k}}(x) = G(a_k x + b_k)$, which implies $G(x) = G^k(a_k x + b_k)$ as they are of the same type. $\square$

Theorem 5.4. *If $F \in D(G)$, then G must belong to the following distributions (within type):*

(i) Frechet — $G_{1,\alpha}(x) = \exp(-x^{-\alpha})$, $x > 0$, $\alpha > 0$
(ii) Negative Weibull — $G_{2,\alpha} = \exp(-(-x)^\alpha)$, $x < 0$, $\alpha > 0$
(iii) Gumbel — $G_3(x) = \exp(-\exp(-x))$, $x \in \mathbb{R}$.

Conversely, these distributions can appear as such limits in (5.1).

Remark 5.5. *We have*

(i) Using $X_{(1)} = -\max\{-X_1, \dots, -x_n\}$, we have equivalent theorems in terms of normalized minima.
(ii) Sometimes, we cannot have non-degenerate G of normalized maxima — for example $X_1, \dots, X_n \sim Bern(\frac{1}{2})$, $X_{(n)}$.
(iii) We can combine these three types into Generalized Extreme Value Distribution (GEV) —

$$G(x;\mu,\sigma,\gamma) = \exp(-(1+\gamma(\frac{x-\mu}{\sigma}))^{-\frac{1}{\gamma}}) \tag{5.4}$$

with $1+\gamma(\frac{x-\mu}{\sigma}) > 0$, $\mu \in \mathbb{R}, \gamma \in \mathbb{R}, \sigma > 0$.

We have Frechet corresponds to $\gamma > 0$, $\alpha = \frac{1}{\gamma}$, NW is $\gamma < 0$, $\alpha = -\frac{1}{\gamma}$, and Gumbel corresponds to the case where $\gamma \to 0$.

Proof (non-examinable). We show $Y_n = \frac{X_{(n)} - b_n}{a_n} \xrightarrow{d} Y$, with $G_\gamma(x) = \exp(-(1+rx)^{-\frac{1}{r}})$

Then, using Helly's theorem, we have $\mathbb{E}(z(Y_n)) \to \mathbb{E}(z(Y))$ for all continuous bounded z. Then the LHS is given by

$$\int z\frac{x-b_n}{a_n} dF_{X_{(n)}}(x) = n\int z(\frac{x-b_n}{a_n})F(x)^{n-1}dF(x) \tag{5.5}$$

and changing variables so $F(x) = 1 - \frac{v}{n}, x = \dots$ $\square$

2. Necessary and Sufficient Conditions for Convergence

We say a function $l : [C, \infty] \to (0, \infty)$ is "slowly varying" if $\lim_{x\to\infty} \frac{l(tx)}{l(x)} = 1$ for all $t > 0$. For example, $l(x) = \log x, \log \log x, (\log x)^\alpha$.

We say a function $r_\alpha : [C, \infty) \to (0, \infty)$ is "regularly varying" with an index $a \in \mathbb{R}$ if $r_\alpha(x) = x^{-\alpha} l(x)$ where l is slowly varying - so $r_2(x) = x^{-2} \log x$.

We define an **expected residual lifetime** as

$$R(x) = \mathbb{E}(X - x | X > x) = \frac{1}{1 - F(x)} \int_x^{x_\star} (1 - F(y))dy \tag{5.6}$$

where $x_\star = \inf\{x : F(x) = 1\}$, and $\overline{F}(x) = 1 - F(x)$

Theorem 5.6. *$F \in D(G_{1,\alpha})$ if and only if $x_\star = \infty$, $\overline{F}(x) = x^{-\alpha} l(x)$ where l is slowly varying. We can choose $b_n = 0$, $a_n = F^{-1}(1 - \frac{1}{n})$ for which $F^n(a_n x + b_n) \hookrightarrow G_{1,\alpha}(x)$ is satisfied.*

$F \in D(G_{2,\alpha})$ if and only if $x_\star < \infty$, $\overline{F}(x_\star - \frac{1}{x}) = x^{-\alpha} l(x)$, with l slowly varying for $x > 0$. We can choose $b_n = x_\star$, $a_n = x_\star - F^{-1}(1 - \frac{1}{n})$ for convergence.

$F \in D(G_3)$ if and only if

$$\frac{\overline{F}(x + tR(x))}{\overline{F}(x)} \to e^{-t} \tag{5.7}$$

We can choose $b_n = F^{-1}(1 - \frac{1}{n})$, $a_n = R(b_n)$.

Example 5.7. *(i) Let $F(x) = 1 - \frac{\log_2(x+1)}{x^2}$ where x geq1. Then $F \in G_{1,2}$.*
(ii) Let $F(x) = 1 - (x_\star - x)^3$ where $x_\star - 1 \leq x \leq x_\star$ for some $x_\star \in \mathbb{R}$. Then $F \in G_{2,3}$.
(iii) Let $F(x) = 1 - \frac{1}{1+e^x}$. Then $F \in G_3$.

Lemma 5.8. *Suppose there exists $a_n > 0$, b_n such that $n(1 - F(a_n x + b_n)) \to u(x)$. Then*

$$F^n(a_n x + b_n) \hookrightarrow \exp(-u(x)) \tag{5.8}$$

Proof. Taking the log of the left hand side, we have

$$n \log F(a_n x + b_n) = n \log(1 - (1 - F(a_n x + b_n))) \tag{5.9}$$

$$= n(-(1 - F(a_n x + b_n)) - \frac{1}{2}(1 - F(a_n x + b_n))^2 + \dots) \tag{5.10}$$

$$= -u(x) \tag{5.11}$$

Thus the left hand side converges to $\exp(-u(x))$. □

Proof (Proof of sufficient part of first part of theorem). Proof of (1) - the sufficient part. Suppose $x_\star = \infty$, $\overline{F}(x) = x^{-\alpha} l(x)$. Use a_n and b_n as in the theorem. Then we want to prove $F^n(a_n x + b_n) \hookrightarrow G_{1,\alpha}(x) = \exp(-x^{-\alpha} \mathbb{I}(x > 0))$.

Using the lemma, we instead prove

$$n(1 - F(a_n x)) \to x^{-\alpha}\mathbb{I}(x > 0) + \infty\mathbb{I}(x < 0)\,. \tag{5.12}$$

Let $x < 0$. Note that $a_n = F^{-1}(1 - \frac{1}{n}) \to x_\star = \infty$. Thus $a_n x \to -\infty$, and $n(1 - F(a_n x)) \to \infty$.

Let $x > 0$. Note that $F(a_n) = F(F^{-1}(1 - \frac{1}{n})) \geq 1 - \frac{1}{n}$, and $F(a_n - \delta) \leq 1 - \frac{1}{n}$. Rearranging, this gives $n \geq \frac{1}{1 - F(a_n - \delta)}$

Note also we have

$$n\frac{(1 - F(a_n x))}{(1 - F(a_n x))}(1 - F(a_n)) \tag{5.13}$$

which converges to $x^{-\alpha}$, as $\overline{F} = x^{-\alpha}l(x)$.

Thus, it suffices to show that $n(1 - F(a_n)) \to 1$. Note that

$$\begin{aligned}
1 &\geq n(1 - F(a_n)) && (5.14)\\
&\geq \frac{1 - F(a_n)}{1 - F(a_n - \delta)} && (5.15)\\
&\geq \frac{1 - F(a_n)}{1 - F(a_n(1 - \epsilon))} && (5.16)\\
&= \frac{a_n^{-\alpha} l(a_n)}{a_n^{-\alpha}(1 - \epsilon)^{-\alpha} l(a_n(1 - \epsilon))} && (5.17)\\
&= (1 - \epsilon)^{\alpha} && (5.18)
\end{aligned}$$

and as ϵ can be made arbitrarily close to zero, we obtain our result. □

Proof (Proof of sufficient part of third part of theorem). Suppose

$$\frac{\overline{F}(x + tR(x))}{\overline{F}(x)} \to e^{-t} \tag{5.19}$$

and we use a_n, b_n as in the theorem. As in the lemma, we seek to prove

$$n(1 - F(a_n x + b_n)) = n(1 - F(R(b_n)x_n + b_n)) \to e^{-x}. \tag{5.20}$$

To use the condition, note that the left hand side is given as

$$\frac{n(1 - F(b_n + xR(b_n)))}{1 - F(b_n)}(1 - F(b_n)) \tag{5.21}$$

and the inner term converges to e^{-x} by assumption.

Thus, it suffices to prove $n(1 - F(b_n)) \to 1$.

$$
\begin{aligned}
1 &\geq n(1 - F(b_n)) && (5.22)\\
&\geq \frac{1 - F(b_n)}{1 - F(b_n - \delta)} && (5.23)\\
&\geq \frac{1 - F(b_n)}{1 - F(b_n - \epsilon R(b_n))} && (5.24)\\
&\to \frac{1}{e^{-(-\epsilon)}} = e^{-\epsilon} \to 1 && (5.25)
\end{aligned}
$$

Choose ϵ such that $1 - F(b_n - \delta) \leq 1 - F(b_n - \epsilon R(b_n))$. $\square$

Bibliography