Nonparametric and Semiparametric Econometrics
Lecture Notes for Econ 227

Yixiao Sun

Department of Economics,

University of California, San Diego

Winter 2015

Contents

[1] This chapter is based on Chen (2007) Handbook of Econometrics Chapter on sieve estimation. I have borrowed some sections directly from Chen (2007), as we planned to work a book project at one point. All errors are my own.

Preface

The primary goal of this course is to introduce modern nonparametric and semiparametric techniques in Econometrics. The course contains two parts. In the first part, we provide a rigorous introduction to linear smoothers. We give a thorough treatment of the so-called kernel estimators and briefly discuss an alternative class of nonparametric estimators, the class of so-called series estimators. We consider both the estimation of probability densities and the estimation of their functionals, such as conditional density and conditional moments. In the second part, we examine the general method of sieves. A prototypical example is the conditional moment restriction model that contains both the finite dimensional parameter and infinite dimensional parameter. We provide a unified framework to analyze the asymptotic properties of the semiparametric estimator of the finite dimensional parameter as well as the infinite dimensional parameter. We will consider semiparametric sieve methods and semiparametric models with endogenous covariates.

Chapter 1

Kernel Smoothing: Density Estimation

1.1 Introduction

The estimation of probability density functions and cumulative distribution function are cornerstones of applied data analysis in social sciences. Testing for the equality of two distributions is probably the most basic test in all of applied data analysis. Though PDF and CDF are often the objects of direct interest, their estimation also serves an important building blocks for other estimation and inference problem. For example, in quantile regression, a density enters the formula for the asymptotic variance.

In this chapter, we consider the so-called kernel density estimator. This is a class of nonparametric estimators of densities. As such, they impose no parametric restrictions on the functional form of the density. We first derive its theoretical properties for the univariate case, showing consistency and asymptotic normality, and derive the optimal convergence rate. The implementation of the kernel estimator is discussed; in particular, a number of selection rules for the so-called bandwidth are presented. We also introduce the concept of higher-order kernels which are able to reduce the bias of the kernel estimator. We then generalize the results to the multivariate case, where we also present a kernel estimator of the conditional density.

The framework is the following: Given an i.i.d. sample $X_i \sim X, i = 1, ..., n$, we are interested in estimating the marginal density of X. We assume that $X \sim f$ for some density f, i.e. $P(X \in A) = \int_A f(x)dx$ for any Borel set $A \subseteq \mathbb{R}$. We are then interested in estimating f.

In a parametric framework, we would specify a class of densities parameterized by some finite dimensional vector $\theta \in \Theta \subseteq \mathbb{R}^k$, $\{f(\cdot;\theta)|\theta \in \Theta\}$, and then assume that the true density

$f = f(\cdot;\theta_0)$ for some $\theta_0 \in \Theta$. A standard example is the normal case where

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

and $\theta = \left(\mu, \sigma^2\right)$. An obvious estimator of the density would then be $\hat{f} = f(\cdot;\hat{\theta})$, where

$$\hat{\theta} = \arg\min_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta).$$

Under some regularity conditions (e.g. Lecture note for 220C)

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N(0, H_0^{-1}),$$

where H_0 is the negative Hessian matrix

$$H_0 = -E\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}|_{\theta=\theta_0}.$$

It then easily follows that

$$\sqrt{n}\left(\hat{f}(x) - f(x)\right) \to_d N(0, B_0(x) H_0^{-1} B_0(x))$$

where $B_0(x) = \partial f(x,\theta_0)/\partial\theta$. So if one is willing to assume f belongs to the specified parametric class, we have a well-behaved estimator at hand.

However, there is a risk of misspecification. Assume that $f \notin \{f(\cdot;\theta)|\theta \in \Theta\}$. Then, the parametric estimator will be biased and we get:

$$\sqrt{n}\left(\hat{f}(x) - f(x,\bar{\theta})\right) \to_d N(0, \bar{B}(x) \bar{H}^{-1}\bar{\Omega}\bar{H}^{-1}\bar{B}(x))$$

where $\bar{\theta}$ is the pseudo-value

$$\bar{\theta} = \arg\min_{\theta\in\Theta} KL\left(f(x;\theta), f(x)\right),$$

and

$$\begin{aligned}
\bar{\Omega} &= E\frac{\partial \log f(X;\bar{\theta})}{\partial\theta}\frac{\partial \log f(X;\bar{\theta})}{\partial\theta} \\
\bar{H} &= -E\frac{\partial^2 \log f(X;\bar{\theta})}{\partial\theta^2}, \bar{B}(x) = \frac{\partial f(x,\bar{\theta})}{\partial\theta}.
\end{aligned}$$

Here $KL\left(f(x;\theta), f(x)\right) = E_f \log f - E_f \log f(x;\theta)$ where E_f is the expectation under f so that

$$KL\left(f(x;\theta), f(x)\right) = \int\left[-\log\frac{f(x;\theta)}{f(x)}\right] f(x)dx \geq -\log\int\frac{f(x;\theta)}{f(x)} f(x)dx = 0.$$

Exercise 1 *Suppose* $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ *and* $f(x;\theta) = \frac{1}{\sqrt{8\pi}} \exp\left[-(x-\theta)^2/8\right]$ *and* $\Theta = \mathbb{R}$. *What is* $\bar{\theta}$?

One solution to this problem is to choose a very large, flexible parametric class but no matter how large this class is chosen, one can never completely safeguard oneself against the risk of $f \notin \{f(\cdot;\theta)|\theta \in \Theta\}$. Furthermore, the numerical problems associated with actually obtaining $\hat{\theta}$ grow as the parametric class does.

An alternative solution is to construct a nonparametric estimator that works for any (sufficiently well-behaved) density, without imposing any parametric form on it. This should remove any risk of misspecification. One example of a nonparametric density estimator is the so-called kernel density estimator. Of course, we can not achieve the robustness without incurring any cost.

1.2 Kernel Estimator

If f is smooth in a small neighborhood $[x - h/2, x + h/2]$ of x, we can justify the following approximation,

$$hf(x) \approx \int_{x-h/2}^{x+h/2} f(u)du = P\left(X \in [x - h/2, x + h/2]\right)$$

by the mean value theorem. The right hand side can be approximated by counting the percentage of observations that fall in the interval $[x - h/2, x + h/2]$. A natural estimate of $f(x)$ is then given by

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{X_i \in [x - h/2, x + h/2]\right\}, \tag{1.1}$$

The above estimator is simply the classic histogram estimator of a density, where h is the so-called binwidth. Figure 1.1 presents a typical histogram.

An equivalent expression is

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} 1\left\{\frac{X_i - x}{h} \in \left[-\frac{1}{2}, \frac{1}{2}\right]\right\} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1.2}$$

with

$$K(u) = 1\left\{-\frac{1}{2} \leq u \leq \frac{1}{2}\right\} \tag{1.3}$$

is the density of the uniform distribution over the interval $[-1/2, 1/2]$. This density estimator is discontinuous however, which is a less appealing feature. This owes to the fact that our choice of K in (1.3) is a discontinuous function. In addition, the uniform kernel weights each

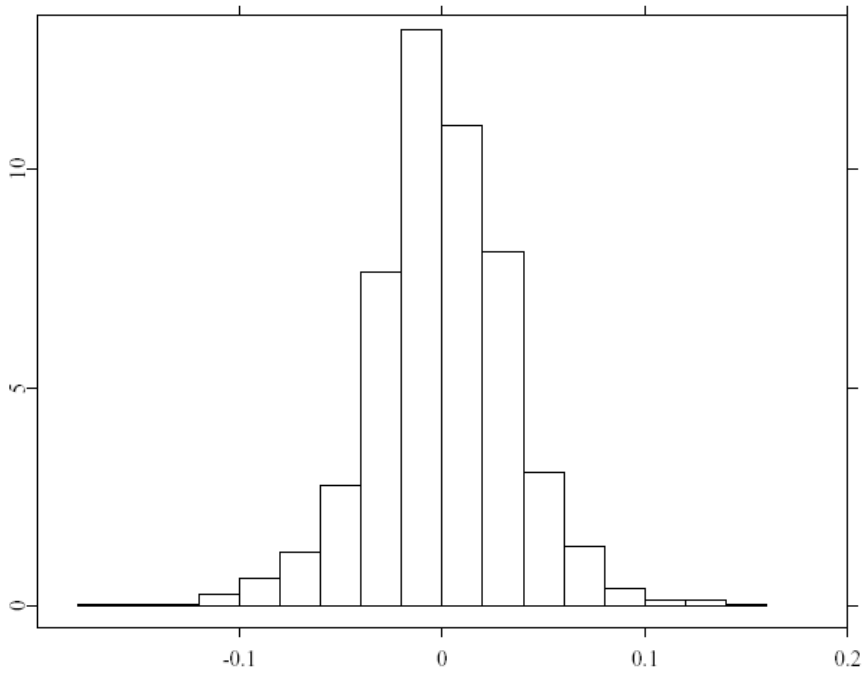


Figure 1.1: Histogram of Stock Returns

observation inside the window equally, even though observations closer to x should possess better information than more distant ones. So instead of using the uniform density in (1.3), we will choose K as some smooth density. This leads to a general class of density estimators of the form (1.2), which is known as kernel density estimators or *Parzen-Rosenblatt* estimators.

For any density estimator of the form (1.2) for some function $K(\cdot)$ and some positive number h, we will refer to K as the kernel, and h as the bandwidth. The specific choice of K given in (1.3) is normally called the uniform kernel. The general kernel estimator where $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is chosen to be some continuous (and differentiable) density can be seen as a smoothed version of the histogram estimator. In fact, the kernel density estimator in (1.2) can be motivated in a different way. Let $\varepsilon_i \sim iid$ with CDF F_ε and density $K(\cdot)$, then

$$P(X_i + h\varepsilon_i \leq x) = P(\varepsilon_i \leq \frac{x - X_i}{h}) = E_X F_\varepsilon \left(\frac{x - X}{h} \right)$$

So the density of $X_i + h\varepsilon_i$ is $(1/h) E_X K[(x - X)/h]$ which can be estimated by

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left(\frac{x - X_i}{h} \right).$$

But this is exactly the kernel density estimator $\widehat{f}(x)$. Therefore, $\widehat{f}(x)$ is a smoothed version of the histogram estimator.

Exercise 2 *Let* $1\{x \in I\}$ *be the indicator function of an interval* I*. Define*

$$g(y) = \int_{-\infty}^{\infty} 1\{x \in I\} \frac{1}{\sigma}\phi\left(\frac{y-x}{\sigma}\right) dy$$

which is the convolution between $1\{x \in I\}$ *and the standard normal pdf. Show that*

$$g(x) = E1\{x + \sigma Z \in I\}$$

for $Z \sim N(0,1)$ *and that* $g(x)$ *is smooth.*

When we derive the asymptotic properties of $\hat{f}$, we need to impose regularity conditions on $K(x)$. We usually maintain the following assumptions:

$$\begin{aligned} \int K(u)du &= 1, \\ \int u^2 K(u)du &= \mu_2 < \infty, \\ K(u) &= K(-u). \end{aligned} \tag{1.4}$$

The first condition ensures that $\widehat{f}(x)$ integrates to one approximately:

$$\begin{aligned} \int_{x_L}^{x_U} \widehat{f}(x)dx &= \int_{x_L}^{x_U} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{x_L}^{x_U} K\left(\frac{x-X_i}{h}\right) dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{(x_L-X_i)/h}^{(x_U-X_i)/h} K(u)\, du = 1 + o(1). \end{aligned} \tag{1.5}$$

If the support of X is $\mathbb{R}$, then $\int_{-\infty}^{\infty} \widehat{f}(x)dx = 1$. Some commonly used kernels are given in the following table:

Kernel	$K(u)$
Uniform	$\frac{1}{2}1\{\lvert u\rvert \le 1\}$
Triangle	$(1-\lvert u\rvert)\,1\{\lvert u\rvert \le 1\}$
Epanechnikov	$\frac{3}{4}\left(1-u^2\right)1\{\lvert u\rvert \le 1\}$
Quartic	$\frac{15}{16}\left(1-u^2\right)^2 1\{\lvert u\rvert \le 1\}$
Triweight	$\frac{35}{32}\left(1-u^2\right)^3 1\{\lvert u\rvert \le 1\}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$
Cosinus	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)1\{\lvert u\rvert \le 1\}$

Figure 1.2 illustrates the kernel density estimation as the sum of $\left\{n^{-1}K_h\left(x - X_i\right)\right\}$, a set of functions indexed by the observation points.

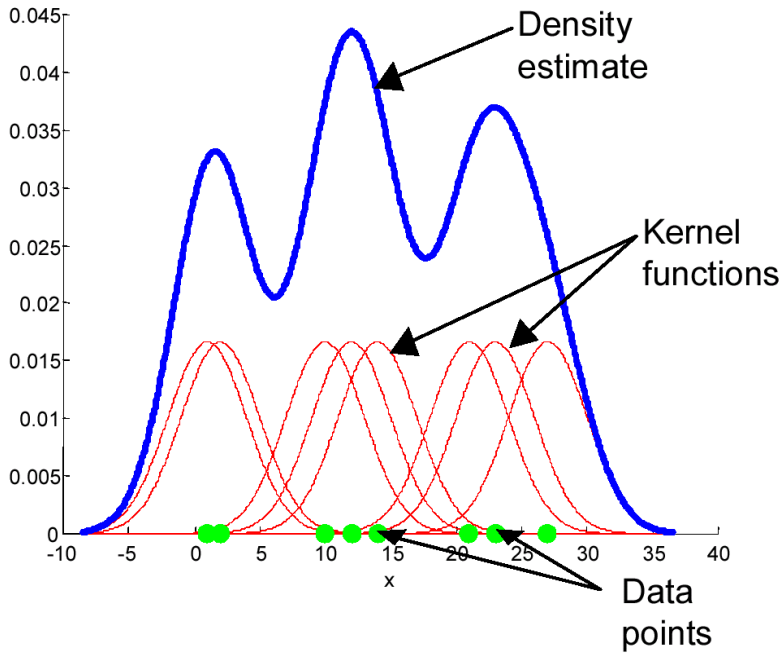


Figure 1.2: Graphical Illustration of Kernel Density Estimation

The value of h basically decides how many observations should be included in the estimation of f at the point x. Assuming that K has a compact support, $[-1, 1]$, then only observations satisfying $X_i \in [x - h, x + h]$ are included in the estimation of $f(x)$. So a small choice of bandwidth means that only observations very close to x are used in the estimation, while a large bandwidth includes most of the observations in the sample. Since the observations close to x are more likely to carry information about the density's behavior at that point, we would expect precision of the density estimator to increase, and thereby the bias to decrease, as we decrease h. On the other hand, as we decrease h, fewer observations are used to estimate $f(x)$, so we would expect the variance of our estimator to increase as we decrease h.

Figures 1.3 and 1.4 illustrate the effect of bandwidth choice on the kernel density estimation. Figure 1.3 is based on a real data set while Figure 1.4 is based on a simulated data set. Clearly, there is a trade-off between choosing a small vs. a large bandwidth. We shall see that in order to show consistency, we have to balance the bias and variance effect of h. We need

$h \to 0$ to get rid of the bias, but not too fast because then the variance will explode.

1.3 Bias and Variance

To evaluate the performance of the KDE, we often consider the pointwise mean squared error (MSE) and integrated mean squared error. The pointwise MSE is defined to be

$$E\left(\widehat{f}(x) - f(x)\right)^2$$

for a given point x in the support of the density function $f(x)$. This criterion is used to evaluate the performance of $\widehat{f}(x)$ at a given point. The Integrated MSE is a weighted average of the pointwise MSE:

$$IMSE\left(\hat{f}(x)\right) = \int E(\hat{f}(x) - f(x))^2 \pi(x) dx$$

for some weighting function $\pi(x)$. Since IMSE can also be rewritten as

$$E \int (\hat{f}(x) - f(x))^2 \pi(x) dx,$$

it is sometimes called Mean Integrated Squared Error (MISE). We will use IMSE and MISE interchangingly. The MISE criterion is used to evaluate the performance of $\widehat{f}(x)$ as a function in a functional space. For simplicity, we consider $\pi(x) = 1$, in which case

$$MISE = E\left(\left\|\hat{f} - f\right\|_2^2\right).$$

In the standard terminology, $\left\|\hat{f} - f\right\|_2^2$ is called the L_2 loss, which is random and $E\left(\left\|\hat{f} - f\right\|_2^2\right)$ is called the risk, which by definition is random.

Mean squared errors are not the only criterion for evaluating the performance of $\widehat{f}$. In fact, a more natural choice is the L_1 type criterion:

$$E\left(\left\|\hat{f} - f\right\|_1\right) = E\left(\int \left|\hat{f}(x) - f(x)\right| dx\right).$$

Here $\int \left|\hat{f}(x) - f(x)\right| dx$ is the total variation distance between the two probability measures associated with $\hat{f}(x)$ and $f(x)$ respectively. We can also use other distance measures, such as the Hellinger distance defined by

$$E\left(\left\|\hat{f} - f\right\|_H^2\right) = \frac{1}{2} E \int \left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right)^2 dx.$$

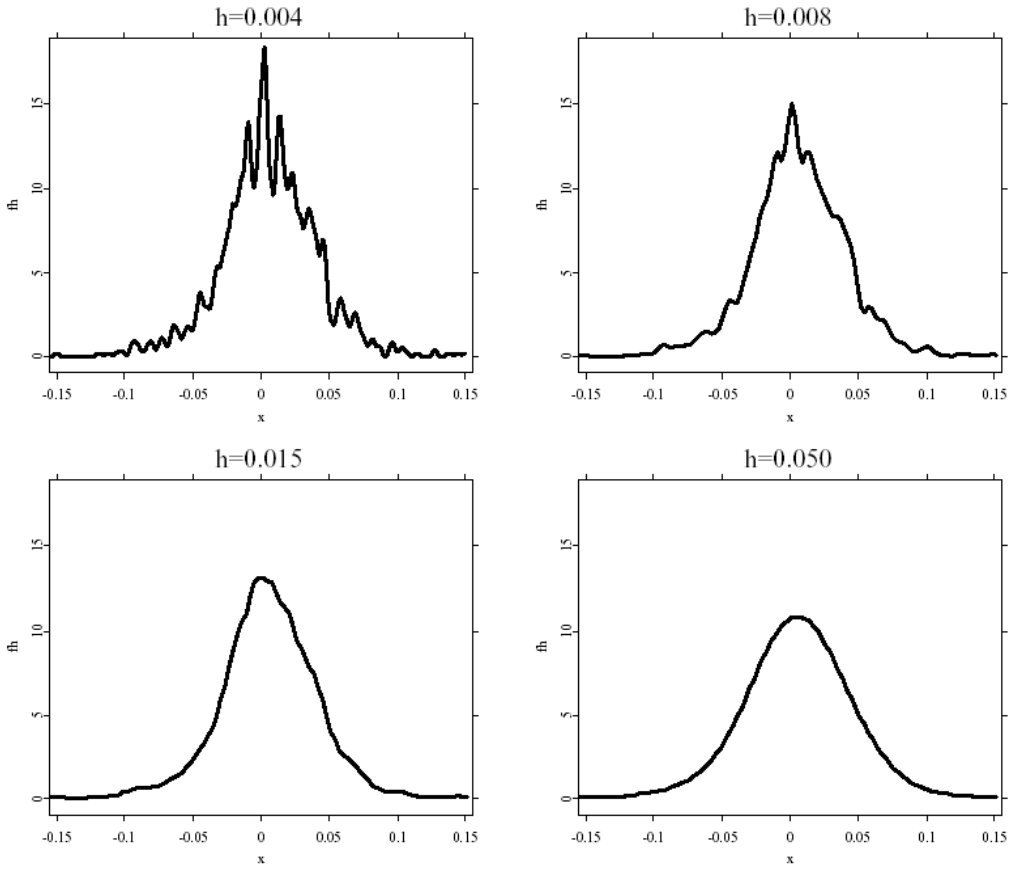


Figure 1.3: Four kernel density estimates for the stock returns data with different bandwidths and Quartic kernel

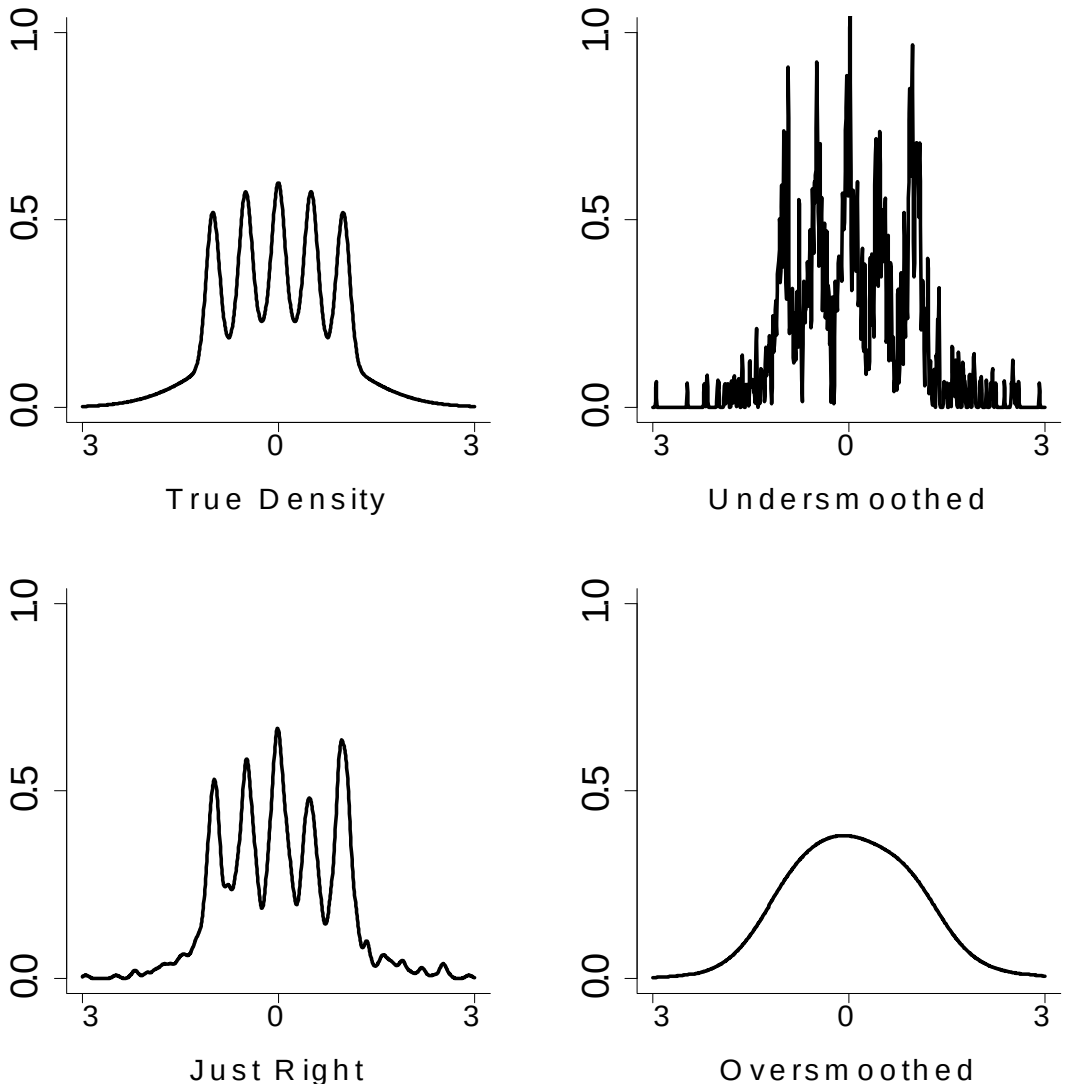


Figure 1.4: Effect of Bandwidth Choice on the Kernel Density Estimation

For discussions along this line, see Devroye and Lugosi (2001).

We use MSE because of its mathematical tractability. In the rest of this chapter, we focus on the MSE but you are encouraged to think about "what if we use the expected L_1 distance instead."

Theorem 1.3.1 *Assume that $f(x) \in C^2(\mathcal{X})$ where $\mathcal{X} = supp(X)$ and $K(\cdot)$ is a bounded function that satisfies condition (1.4) and for any $c_1 > 0$ and $c_2 > 0$, $\int_{-c_1/h}^{c_2/h} K(u)\,du = 1+o(h^2)$ and $\int_{-c_1/h}^{c_2/h} uK(u)\,du = o(h^2)$ as $h \to 0$.*

(i) For an interior point $x \in \mathcal{X}$, we have, as $h \to 0$ such that $nh \to \infty$,

$$\begin{aligned} E\widehat{f}(x) - f(x) &= \frac{1}{2}f''(x)\left(\int K(u)u^2du\right)h^2 + o(h^2) \\ var\left[\widehat{f}(x)\right] &= f(x)\left(\int K^2(u)\,du\right)\frac{1}{nh} + o\left(\frac{1}{nh}\right). \end{aligned}$$

(ii) For an interior point $x \in \mathcal{X}$, we have, as $h \to 0$ such that $nh \to \infty$,

$$E\left(\widehat{f}(x) - f(x)\right)^2 = \frac{1}{4}\left[f''(x)\right]^2\left(\int K(u)u^2du\right)^2 h^4 + f(x)\left(\int K^2(u)\,du\right)\frac{1}{nh} + o(h^4) + o(\frac{1}{nh})$$

(iii) Suppose $\mathcal{X} = \mathbb{R}$. As $h \to 0$ such that $nh \to \infty$,

$$MISE = \left(\frac{1}{4}\int\left[f''(x)\right]^2 dx\right)\left(\int K(u)u^2du\right)^2 h^4 + \left(\int K^2(u)\,du\right)\frac{1}{nh} + o(h^4) + o(\frac{1}{nh}) \quad (1.6)$$

Proof: We prove only (i) as (ii) and (iii) follow immediately from (i). First, let $\mathcal{X} = [x_L, x_U]$, then for any interior point x

$$\frac{1}{h}\int_{x_L}^{x_U} K\left(\frac{v-x}{h}\right)dv = \int_{(x_L-x)/h}^{(x_U-x)/h} K(u)\,du = 1 + o(h^2).$$

Using this, we have

$$\begin{aligned} & E\widehat{f}(x) - f(x) = \frac{1}{nh}\sum_{i=1}^{n} EK\left(\frac{x-X_i}{h}\right) - f(x) \\ = & \frac{1}{h}EK\left(\frac{x-X_i}{h}\right) - f(x) = \frac{1}{h}\int_{x_L}^{x_U} K\left(\frac{v-x}{h}\right)[f(v)-f(x)]\,dv \\ = & \int_{(x_L-x)/h}^{(x_U-x)/h} K(u)\,[f(x+uh)-f(x)]\,du + o(h^2) \\ = & \int_{(x_L-x)/h}^{(x_U-x)/h} K(u)\left[f'(x)uh + \frac{1}{2}f''(x)u^2h^2 + \frac{1}{2}\left[f''(\tilde{x})-f''(x)\right]u^2h^2\right]du + o(h^2) \quad (1.7) \\ = & hf'(x)\int_{(x_L-x)/h}^{(x_U-x)/h} K(u)udu + \frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2) \\ = & \frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2). \quad (1.8) \end{aligned}$$

Second,

$$\begin{aligned} var\left[\widehat{f}(x)\right] = & var\left\{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right\} = \frac{1}{nh^2}var\left[K\left(\frac{x-X_i}{h}\right)\right] \\ = & \frac{1}{nh^2}\left\{\int_{x_L}^{x_U} K^2\left(\frac{v-x}{h}\right)f(v)dv - \left[\int_{x_L}^{x_U} K\left(\frac{v-x}{h}\right)f(v)dv\right]^2\right\} \\ = & \frac{1}{nh^2}\left\{h\int_{(x_L-x)/h}^{(x_U-x)/h} K^2(u)\,f(x+uh)du - \left[h\int_{(x_L-x)/h}^{(x_U-x)/h} K(u)\,f(x+uh)du\right]^2\right\} \\ = & \frac{1}{nh}\int_{(x_L-x)/h}^{(x_U-x)/h} K^2(u)\,f(x+uh)du + o\left(\frac{1}{nh}\right) \\ = & \frac{f(x)}{nh}\int K^2(u)\,du + o\left(\frac{1}{nh}\right).\blacksquare \quad (1.9) \end{aligned}$$

Remark 1.3.1 *The theorem implies that*

$$\widehat{f}(x) = f(x) + O_p\left(\frac{1}{\sqrt{nh}} + h^2\right)$$

So when $h \to 0$ and $nh \to \infty$, $\hat{f}(x)$ is consistent for $f(x)$. In fact, the consistency of $\hat{f}(x)$ can be established under much weaker conditions, see Theorem 9.2 in Devroye and Lugosi (2001).

Remark 1.3.2 *Figure 1.5 illustrates the bias effect. The bias is positive when $f''(x) > 0$ and is negative when $f''(x) < 0$.*

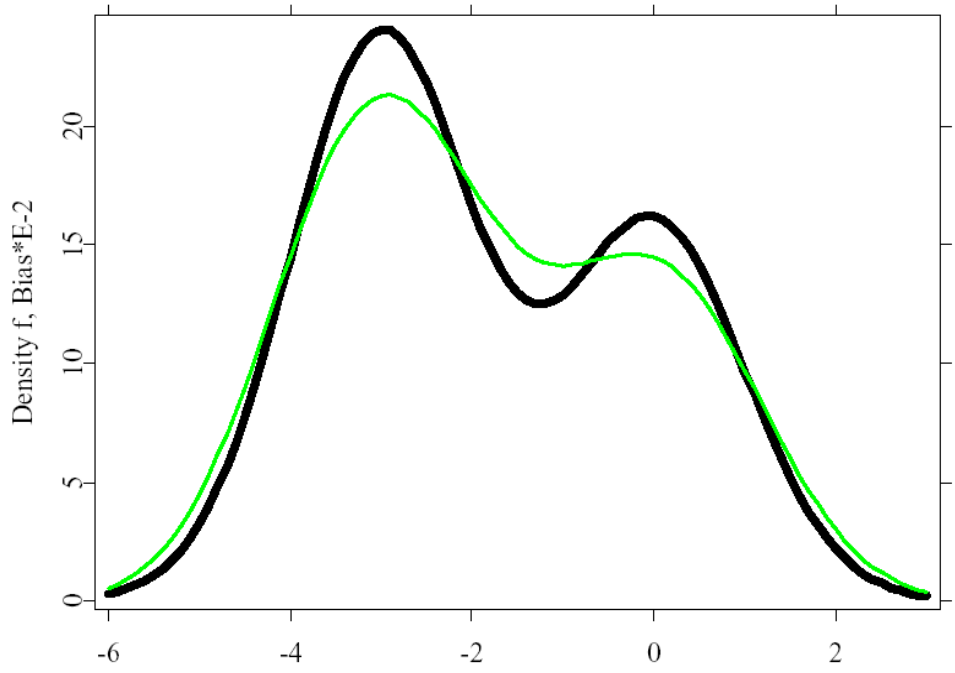


Figure 1.5: $f(x)$ (thick black line) and approximation for $E\hat{f}(x)$ (thin green line)

Remark 1.3.3 *The consistency result requires that x is an interior point of the support of X. For x at the boundary, the bias may not go to zero. As an example, consider $X \in [0,1]$ and $x = 0$. In this case*

$$\hat{f}(0) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i}{h}\right)$$

and

$$\begin{aligned} E\hat{f}(0) &= \frac{1}{h}\int_0^1 K(\frac{v}{h})f(v)dv = \int_0^{1/h} K(u)f(uh)du \\ &= f(0)\int_0^\infty K(u)du\,(1+O(h)) = \frac{f(0)}{2}(1+O(h)). \end{aligned}$$

Therefore, $\hat{f}(0)$ is downward biased. In the literature, various boundary kernels are proposed to overcome this boundary bias problem. See page 31 of Li and Racine (2007, equation 1.43).

Remark 1.3.4 *If the support of X is R, then there is no boundary problem.*

1.4 Optimal Bandwidth Choice: Plug-in Approach

We have derived the bias and variance of the kernel density estimator under appropriate conditions on the bandwidth and the kernel. In practice, it might not be very clear how to choose either of these. While we have been able to give conditions on the rate at which h should go to zero as $n \to \infty$, we need some guidelines for how to choose h in finite samples. Similarly, we would like to know which kernel will be appropriate.

Let us first consider bandwidth choice. The asymptotic mean squared error is, up to smaller order terms:

$$MSE\left(\hat{f}(x)\right) = \frac{1}{4}\left[f''(x)\right]^2 \left(\int K(u)u^2 du\right)^2 h^4 + f(x)\left(\int K^2(u)\,du\right)\frac{1}{nh}$$

The MSE-optimal h is then

$$h = \left(\frac{f(x)}{[f''(x)]^2}\frac{\left(\int K^2(u)\,du\right)}{\left(\int K(u)u^2 du\right)^2}\right)^{1/5} n^{-1/5}. \tag{1.10}$$

The above bandwidth is optimal for a given point x. For different points, the optimal bandwidth should be different. To emphasize the dependence, we may write $h = h(x)$.

Now suppose we are interested in choosing the bandwidth not for a given point x but for all points in the support of $f(x)$. In this case, we can choose h to minimize the integrated MSE (IMSE). Using the theorem in the previous section, we have, up to smaller order terms

$$IMSE\left(\hat{f}(x)\right) = \frac{1}{4}\int \left[f''(x)\right]^2 dx \left(\int K(u)u^2 du\right)^2 h^4 + \frac{1}{nh}\int K^2(u)\,du$$

The optimal bandwidth is now

$$h_{opt} = \left[\frac{1}{\int [f''(x)]^2 dx}\frac{\int K^2(u)\,du}{\left(\int K(u)u^2 du\right)^2}\right]^{1/5} n^{-1/5}. \tag{1.11}$$

The IMSE-optimal bandwidth choice rule is very simple. Unfortunately, h_{opt} depends on the unknown density f through its second order derivative f'' so one cannot calculate it directly. A number of methods have been proposed to circumvent this problem. The first is the so-called Silverman's Rule of Thumb: One simply chooses a known density f^*, e.g. the Gaussian, as a point of reference for which one calculates $f^*(x)$ and $\partial^2 f^*(x)/\partial x^2$ and then plugs these into the RHS of (1.11). More specifically, let

$$f^*(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

then some simple calculation gives

$$\left(\int \frac{\partial^2 f^*(x)}{\partial x^2} dx \right)^2 = \sigma^{-5} \frac{3}{8\sqrt{\pi}}.$$

Assume that the Gaussian kernel is used, then

$$\begin{aligned} \int K^2(u)\, du &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-u^2) du = \frac{1}{2\sqrt{\pi}}, \\ \left(\int K(u) u^2 du \right)^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 \exp(-\frac{u^2}{2}) du = 1. \end{aligned}$$

Combining these calculations, we obtain

$$h_{opt} = \left(\frac{1}{2\sqrt{\pi}\sigma^{-5}\frac{3}{8\sqrt{\pi}}} \right)^{1/5} n^{-1/5} = 1.059 \sigma n^{-1/5}.$$

In practice, one often uses the sample standard deviation $\hat{\sigma}$ to estimate σ, leading to the most popular bandwidth choice rule in applied work:

$$h_{opt} = 1.06 \hat{\sigma} n^{-1/5} \tag{1.12}$$

Estimates of other dispersion measures may be used in place of $\hat{\sigma}$. Silverman (1986, p.47) suggests replacing $\hat{\sigma}$ by

$$\min \left(\hat{\sigma}, \frac{\textit{empirical interquartile range}}{1.349} \right).$$

where 1.349 is the *interquartile range* of the standard normal distribution.

The bandwidth choice in (1.12) is optimal with respect to the chosen family of distributions. However if f^* is very different from the true density f, the resulting bandwidth choice will be very different from the optimal one. To improve on this rule of thumb, one can use an iterative procedure: One starts out with an initial bandwidth choice, this can for example be obtained from Silverman's rule of thumb, and then use this to get a first kernel estimate of the density. This kernel estimate will then in turn deliver an estimate of $f(x)$ and its 2nd derivative. Use these two estimates to calculate a new "optimal" bandwidth, and so on. Hopefully, the procedure will converge at some point.

A more objective procedure for choosing the bandwidth is the so-called cross-validation method which we discuss in the next section.

1.5 Optimal Bandwidth Choice: Cross Validation

Cross-validation is statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. The basic idea of cross-validation can be used to derive an automatic and data-driven method for choosing the bandwidth.

The main idea is still to choose h such that the associated integrated square error (ISE) is minimized. The ISE is defined as:

$$ISE(h) = \int \left[\hat{f}(x) - f(x)\right]^2 dx.$$

Taking expectation gives the mean integrated squared error:

$$E\left[ISE(h)\right] = E_X \int \left[\hat{f}(x) - f(x)\right]^2 dx = \int E\left[\hat{f}(x) - f(x)\right]^2 dx = IMSE(h).$$

ISE is often preferred as a criterion rather than its expected value, since ISE determines how closely $\hat{f}$ approximates f for a given data set, whereas MISE is concerned with the average over all possible data sets. However, in certain situations, MISE may actually be a better performance criterion than ISE. In large samples, the difference is expected to be small as it can be shown that

$$\frac{ISE(h)}{IMSE(h)} \to_p 1.$$

See Hall (1982).

The ISE can be written as

$$ISE(h) = \int \hat{f}^2(x)dx - 2\int \hat{f}\left(x\right) f\left(x\right) dx + \int f^2\left(x\right) dx.$$

Observe that the last integral does not depend on h, so we can ignore this, and consider the minimization of the first two integrals. This is equivalent to solving the following problem

$$\max \int \hat{f}\left(x\right) f\left(x\right) dx \text{ s.t. } \int \hat{f}^2(x)dx \leq C$$

for some constant C such that the Lagrangian multiplier for the constraint is $1/2$. Note that $\int \hat{f}\left(x\right) f\left(x\right) dx = \left\langle \hat{f}, f \right\rangle$ which measures the angle between $\hat{f}$ and f. So we want to select an $\hat{f}$ to be in the same direction as f subject to a normalization constraint.

We now go back to the two integrals in $ISE(h)$. The first integral can easily be computed for any given $h > 0$:

$$\begin{aligned}\int \hat{f}^2(x)dx &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx \\ &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(s)\,K\left(\frac{X_j-X_i}{h}-s\right)ds \\ &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}K*K\left(\frac{X_j-X_i}{h}\right) \\ &:= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{K}\left(\frac{X_i-X_j}{h}\right)\end{aligned}$$

where $K * K$ denotes the convolution of K with itself:

$$K * K(u) = \int_{\mathbb{R}} K(u-v)K(v)dv.$$

If K is the N(0,1)-density, the $K * K$ is the $N(0, 2)$ density.

The second one cannot be easily computed since f is unknown. To overcome this difficulty, we find an estimate of the second integral, which is

$$\begin{aligned}-2\int \hat{f}(x)\,f(x)\,dx &= -2\int \hat{f}(x;X_1,X_2,...,X_n)\,f(x)\,dx \\ &= -2E\left\{\hat{f}(\mathbb{X};X_1,X_2,...,X_n)\,|X_1,....,X_n\right\}\end{aligned}$$

where $\mathbb{X}$ is a random variable that has the same distribution as X_i and is independent of X_1^n. In the above expression, we have written $\hat{f}(x) = \hat{f}(x;X_1,X_2,...,X_n)$ to emphasize its dependence on the sample $(X_1,X_2,...,X_n)$.

In the ideal situation when we have a 'ghost sample' $\tilde{X}_1,...,\tilde{X}_n$, which is independent of $X_1,...,X_n$, we can estimate $-2\int \hat{f}(x)\,f(x)\,dx$ by

$$-\frac{2}{n}\sum_{i=1}^{n}\hat{f}(\tilde{X}_i;X_1,X_2,...,X_n).$$

Of course, such a 'ghost sample' is not available in practice. Note that the basic requirement for the above infeasible estimator to be a 'good' estimator is that $\tilde{X}_i$ is independent of $\hat{f}(x;X_1,X_2,....,X_n)$. This point is at the core of any CV procedure and justifies the splitting of the sample into a training set—used to compute the estimator— and a test set—used to assess the quality of the latter estimator.

If we replace $\tilde{X}_i$ by X_i, we have to leave X_i out from the construction of $\hat{f}$ so that the resulting estimator, say $\hat{f}_{-i}(x|X_1,...,X_{i-1},X_{i+1},...,X_n)$, is independent of X_i. This intuitive argument suggests that the second integral can be estimated by

$$-\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i;X_1,...,X_{i-1},X_{i+1},...,X_n)$$

where

$$\hat{f}_{-i}(x;X_1,...,X_{i-1},X_{i+1},...,X_n)=\frac{1}{(n-1)h}\sum_{j=1,\neq i}^{n}K\left(\frac{x-X_j}{h}\right)$$

is the so-called *leave-one-out* estimator, that is, we leave out the i-th observation in the estimation of $f(x)$. We can view $\{X_1,...,X_{i-1},X_{i+1},...,X_n\}$ as the training sample and $\{X_i\}$ as the test sample.

We then define the cross-validated bandwidth as

$$h_{cv}=\arg\min_{h>0}CV(h)$$

where

$$CV(h)=\frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{K}\left(\frac{X_i-X_j}{h}\right)-\frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}K\left(\frac{X_i-X_j}{h}\right).$$

An alternative estimator of the second integral would be $1/n\sum_{i=1}^{n}\hat{f}(X_i)$. That is, we don't leave out any observations. In this case,

$$CV_{naive}(h)=\frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\bar{K}\left(\frac{X_i-X_j}{h}\right)-2K\left(\frac{X_i-X_j}{h}\right)\right]. \tag{1.13}$$

When h is small enough, the dominating term is obtained by setting $i=j$ in the above expression, leading to

$$\frac{1}{nh}\left[\bar{K}(0)-2K(0)\right]$$

Assume that $K(u)\leq K(0)$ for all u, then $\bar{K}(0)=\int K^2(u)du\leq\int K(0)K(u)du=K(0)<2K(0)$. Therefore, minimizing (1.13) w.r.t. h will cause $h\to 0$ *for any fixed* $n\geq 1$. It is a good exercise to rigorously prove this statement. See exercise 1.6 in Li and Racine (2007). The leave-one-out estimator avoids the convergence of $\hat{h}\to 0$ for any fixed n.

Note that

$$CV(h)\doteq CV_{naive}(h)+\frac{2K(0)}{nh}$$

The second term $2K(0)/(nh)$ can be regarded as the penalty term. The smaller h is, the more complex the model is and the higher the penalty is.

It can be shown (Stone, 1984) that h_{cv} converges towards the optimal bandwidth h_{opt} given in (1.11) as $n \to \infty$. Intuitively, the leading term of $CV(h)$ is $E[CV(h)]$, which, up to an irrelevant constant, is an alternative expression for MISE. However, the rate of convergence is very slow, $n^{-1/10}$, so one should not uncritically use the cross-validated bandwidth. We do not prove these theoretical results here but will do so in the case of conditional moment estimation.

The cross-validation function $CV(h)$ can have more than one local minimum (Hall and Marron, 1991). Thus, in practice, it is prudent to plot $CV(h)$ and not just rely on the result of a minimization routine. Jones, Marron and Sheather (1996) recommended that the largest local minimizer of $CV(h)$ be used as h_{cv}, since this value produces better empirical performance than the global minimizer.

Before we leave this section, we note that practical issues concerning the actual implementation/calculation of the kernel estimator are discussed in Härdle (1990), Härdle and Linton (1994) and Ichimura and Todd (2007). See Jones, Marron, and Sheather (1996) for a short survey in the statistics literature.

1.6 Optimal Kernel

When the MISE-optimal bandwidth is used, the IMSE is:

$$IMSE = \frac{5}{4}\left\{\int \left[f''(x)\right]^2 dx\right\}^{1/5} \{C(K)\}^{2/5} n^{-4/5}(1+o(1))$$

where

$$C(K) = \left(\int K^2(u)\,du\right)^2 \left(\int K(u)u^2 du\right).$$

To minimize IMSE, we can use the kernel with the smallest $C(K)$ value. We solve the following optimization problem:

$$\min_K \left(\int K^2(u)\,du\right)^2 \left(\int K(u)u^2 du\right)$$
$$s.t.\ K(u) \geq 0,\ K(u) = K(-u),\ \text{and}\ \int K(u)du = 1.$$

So an obvious choice of K would be the one that minimizes $C(K)$. A solution to this minimization problem is the Epanechnikov kernel:

$$K_e(u) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right), & |u| \leq \sqrt{5} \\ 0, & otherwise \end{cases}$$

Proof: If $K(u)$ satisfies the constraints, then $K_b(u) := 1/bK(u/b)$ also satisfies the constraints for any $b > 0$. But

$$\begin{aligned}\left(\int_{-\infty}^{\infty} K_b^2(u)\,du\right)^2\left(\int_{-\infty}^{\infty} K_b(u)u^2du\right) &= \left(\int_{-\infty}^{\infty}\frac{1}{b^2}K^2\left(\frac{u}{b}\right)du\right)^2\left(\int_{-\infty}^{\infty}\frac{1}{b}K\left(\frac{u}{b}\right)u^2du\right)\\ &= \left(\int_{-\infty}^{\infty}K^2\left(\frac{u}{b}\right)d\frac{u}{b}\right)^2\left(\int_{-\infty}^{\infty}K\left(\frac{u}{b}\right)\left(\frac{u}{b}\right)^2d\frac{u}{b}\right)\\ &= \left(\int_{-\infty}^{\infty}K^2(v)\,dv\right)^2\left(\int_{-\infty}^{\infty}K(v)\,v^2dv\right).\end{aligned}$$

So if $K(u)$ is an optimal kernel, then $K_b(u)$ is also optimal. However $\left(\int K_b(u)u^2du\right) = 1/b^2\left(\int K(u)u^2du\right)$. That is, the original optimization problem does not have a unique solution. Different solutions have different values of $\left(\int K(u)u^2du\right)$. To find an optimal kernel, we can set $\int K(u)u^2du$ to be any positive value. In the following we set $\int K(u)u^2du = \int u^2K_e(u)du$.

Let $K(u)$ be a non-negative kernel function satisfying:

$$\begin{aligned}\int K(u)du &= \int K_e(u)du = 1 \text{ and}\\ \int u^2K(u)du &= \int u^2K_e(u)du,\end{aligned}$$

it suffices to show that

$$\int K^2(u)du \geq \int K_e^2(u)\,du. \tag{1.14}$$

Write

$$K(u) = K_e(u) + \varepsilon(u),$$

then

$$\int \varepsilon(u)\,du = \int u^2\varepsilon(u)\,du = 0. \tag{1.15}$$

Since $K(u) \geq 0$, we must also have

$$\varepsilon(u) \geq 0 \text{ for } |u| \geq \sqrt{5}.$$

Now

$$\begin{aligned}\int K^2(u)du &= \int K_e^2(u)\,du + 2\int \varepsilon(u)K_e(u)du + \int \varepsilon^2(u)\,du\\ &:= \int K_e^2(u)\,du + 2U_1 + U_2.\end{aligned}$$

and

$$\begin{aligned} U_1 &= \frac{3}{4\sqrt{5}} \int_{-\sqrt{5}}^{\sqrt{5}} \varepsilon(u) \left(1 - \frac{u^2}{5}\right) du \\ &= \frac{3}{4\sqrt{5}} \left[\int_{\mathbb{R}} \varepsilon(u) \left(1 - \frac{u^2}{5}\right) du - \int_{|u| \geq \sqrt{5}} \varepsilon(u) \left(1 - \frac{u^2}{5}\right) du \right]. \end{aligned}$$

But the first term above vanishes by (1.15) and therefore $U_1 \geq 0$. Also, it is obvious that $U_2 \geq 0$. So (1.14) holds with equality if and only if $\varepsilon(u) = 0$.

The above proof is based on the arguments in Priestley (1981, page 570). See Pagan and Ullah (1999, p. 27) for an alternative argument.

In practice however, the choice of kernel appears to have very little effect on the performance of the kernel estimator, and in most cases the Gaussian kernel is used for simplicity. However, in the econometric literature on spectral density estimation and robust standard error estimation, the quadratic spectral kernel is very popular.

It is important to point out that the Epanechnikov kernel is optimal within the class of positive kernels. If we remove the positiveness requirement, then there are other kernels that dominate the Epanechnikov kernel. For example, some high order kernels in the next section may be better than the Epanechnikov kernel.

1.7 Bias Reduction: High Order Kernels

1.7.1 High-order Kernels

To derive the bias formula in Theorem 1.3.1, we take a second order Taylor expansion of the density function $f(x)$ and use the the symmetry of the kernel function to zero out the term of order $O(h)$. In general, assuming $f(x) \in C^q$ for $q \geq 2$, we can employ high order kernel to achieve bias reduction.

Definition 1.7.1 *A kernel is said to have order q if*

$$\int u^j K(u) du = 0 \text{ for } j = 1, 2, ..., q-1 \text{ and } 0 \neq \int u^q K(u) du < \infty. \tag{1.16}$$

Consider the case that $\mathcal{X} = \mathbb{R}$, $f(x) \in C^q$, and q-th order kernel is used. Then

$$\begin{aligned} E\widehat{f}(x) - f(x) &= \int K(u) \left[f(x + uh) - f(x) \right] du \\ &= h^q \frac{f^{(q)}(x)}{q!} \int u^q K(u) du \left(1 + o((1))\right). \end{aligned} \tag{1.17}$$

So by imposing additional assumptions on the smoothness on $f(\cdot)$, and choosing $K(\cdot)$ to be a high order kernel, we are able to reduce the bias to be of order $O(h^q)$.

In order for the conditions in (1.16) to be satisfied, higher order kernels will in most cases be negative on parts of its support. As a result, in finite samples the kernel density estimator will be negative in some intervals. This property is sometimes emphasized as a drawback of estimators with higher order kernels, since the density $f(x)$ itself is nonnegative. However, this remark is of minor importance because we can always use the positive part estimator

$$\tilde{f}(x) = \max(0, \widehat{f}(x))$$

whose risk is smaller than or equal to that of $\widehat{f}(x)$:

$$E\left[\tilde{f}(x) - f(x)\right]^2 \leq E\left[\widehat{f}(x) - f(x)\right]^2 \text{ for any } x \in R \tag{1.18}$$

[Try to prove this as an exercise]

Higher order kernels prove very useful in theoretical econometrics. In particular, in semiparametric models where the bias of a nonparametric component needs to be reduced in order to achieve faster rate of convergence.

The construction of higher order kernels satisfying (1.16) is fairly straightforward. The standard procedure is to choose an initial symmetric kernel, say K^*, define

$$K(u) = K^*(u)\left(a_0 + a_1 u^2 + ... + a_r u^{2r}\right),$$

for constants $a_0, a_1, ..., a_r \in \mathbb{R}$ and $q = 2r$ and then choose $a_0, ..., a_r$ to maintain the conditions in (1.16). For example, letting

$$K^*(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

be a second order Gaussian kernel, we could begin with the construction

$$K(u) = \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)\right]\left(a_0 + a_1 u^2\right).$$

For $K(u)$ to be a fourth order kernel, we need

$$\int K(u)du = 1, \int K(u)udu = 0, \int K(u)u^2du = 0, \int K(u)u^3du = 0.$$

By construction, $\int K(u)udu = 0$, and $\int K(u)u^3du = 0$. From the remaining two restrictions $\int K(u)du = 1$ and $\int K(u)u^2 = 0$, we can easily solve for the two unknowns $a_0 = 3/2$ and $a_1 = -1/2$. So the fourth order Gaussian kernel is given by

$$K(u) = \left(\frac{3}{2} - \frac{1}{2}u^2\right)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

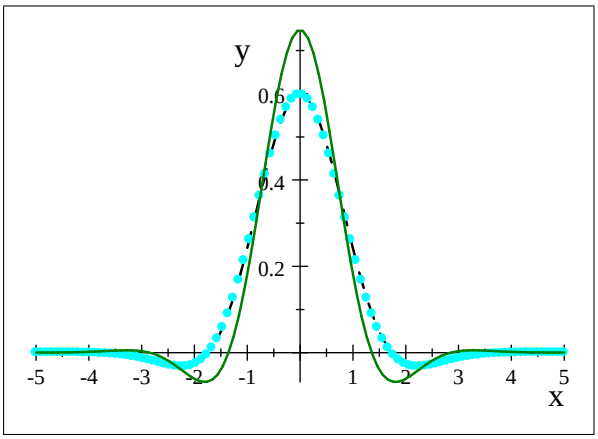


Figure 1.6: Solid Line: Sixth Order Gaussian Kernel. Dot Line: Fourth Order Gaussian Kernel

Figure 1.6 presents some examples of high order kernels.

In fact, when $f(x) \in C^q$, the $O(h^q)$ bias term can be eliminated by using a kernel of order $q+1$. In order to control the remaining bias, we need introduce to the Holder class of functions:

Definition 1.7.2 *Let T be an interval in R and let s and L be two positive numbers. The Hölder class $H(s, L)$ on T is defined as the set of $\ell = \lfloor s \rfloor$ times differentiable functions $f : T \to R$ whose ℓ-th derivative $f^{(\ell)}(\cdot)$ satisfies*

$$\left| f^{(\ell)}(x) - f^{(\ell)}(x') \right| \leq L \left| x - x' \right|^{s-\ell}, \text{ for all } x \text{ and } x' \in T$$

For $f \in H(s, L)$, we can use a kernel of order $\lfloor s \rfloor + 1$ and obtain

$$E\widehat{f}(x) - f(x) = O\left(h^s\right)$$

A proof of the above result is not difficult and is omitted here. Using the same argument before, we can show that

$$var(\widehat{f}(x)) = O\left(\frac{1}{nh}\right)$$

As a result, for $f \in H(s, L)$, we have

$$MSE(\widehat{f}(x)) = O(h^{2s}) + O\left(\frac{1}{nh}\right) = O\left(n^{-\frac{2s}{2s+1}}\right)$$

and thus

$$\widehat{f}(x) - f(x) = O_p\left(n^{-\frac{s}{2s+1}}\right).$$

This turns out to be the best rate that can be achieved in a uniform sense. More precisely, the optimal rate of convergence is often defined in terms of a minimax risk. A detailed discussion along this line is beyond the scope of this course.

1.7.2 Fourier Analysis of KDE

So far, we have studied the MISE of kernel density estimators under classical but restrictive assumptions. Indeed, the results were valid only for densities whose derivatives of given order satisfy certain conditions. In this section, we will show that more general and elegant results can be obtained using Fourier analysis.

For any function $g \in L_1(R)$, define its Fourier transform $\mathcal{F}[g]$ as

$$\mathcal{F}[g](\omega) = \int_{-\infty}^{\infty} g(x) e^{i\omega x} dx, \omega \in \mathbb{R}$$

Parseval's theorem states that

$$\int_{-\infty}^{\infty} g^2(t)\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathcal{F}[g](\omega)|^2\, d\omega$$

for any $g \in L_1(R) \cap L_2(R)$. More generally, the Fourier transform is defined in a standard way for any $g \in L_2(R)$ using the fact that $L_1(R) \cap L_2(R)$ is dense in $L_2(R)$. With this extension, Parseval's theorem holds for any $g \in L_2(R)$.

For any $g \in L_2(R)$, we have

$$\begin{aligned} \mathcal{F}[g_h](\omega) &:= \int_{-\infty}^{\infty} \frac{1}{h} g(\frac{x}{h}) e^{i\omega x} dx \\ &= \int_{-\infty}^{\infty} g(u) e^{i(\omega h)u} du = \mathcal{F}[g](h\omega), \text{ for all } h > 0 \end{aligned} \tag{1.19}$$

and

$$\begin{aligned} \mathcal{F}[g(t-\cdot)](\omega) &= \int_{-\infty}^{\infty} g(t-x) e^{i\omega x} dx = \int_{-\infty}^{\infty} g(y) e^{i\omega(t-y)} dy \\ &= e^{i\omega t} \int_{-\infty}^{\infty} g(y) e^{-i\omega y} dy = e^{i\omega t} \mathcal{F}[g](-\omega), \text{ for all } t \in R \end{aligned} \tag{1.20}$$

Define the characteristic function associated with the density $f(x)$ by

$$\phi(\omega) = E e^{i\omega X} = \mathcal{F}[f](\omega)$$

and consider the empirical characteristic function:

$$\phi_n(\omega) = \frac{1}{n} \sum_{j=1}^{n} e^{i\omega X_j}.$$

Using (1.19) and (1.20), we may write the Fourier transform of the estimator $\hat{f}(x)$ with symmetric kernel $K \in L_2(R)$, in the form

$$\begin{aligned}\mathcal{F}[\hat{f}](\omega) &= \mathcal{F}[\frac{1}{n}\sum_{j=1}^{n} K_h(x - X_i](\omega) = \frac{1}{n}\sum_{j=1}^{n}\mathcal{F}[K_h(X_j - x)](\omega) \\ &= \frac{1}{n}\sum_{j=1}^{n} e^{i\omega X_j}\mathcal{F}[K_h](-\omega) = \phi_n(\omega)\mathcal{F}[K_h](-\omega) = \phi_n(\omega)\mathcal{F}[K](h\omega)\end{aligned}$$

Typically $\mathcal{F}[K](h\omega) \in [0,1]$ almost everywhere. The multiplicative factor$\mathcal{F}[K](h\omega)$ plays the role of shrinking $\phi_n(\omega)$ towards zero.

We can now write the MISE of $\hat{f}$ as

$$\begin{aligned}& E\int\left(\hat{f}(x) - f(x)\right)^2 dx \\ =\ & \frac{1}{2\pi}E\int\left|\mathcal{F}[\hat{f}](\omega) - \mathcal{F}[f](\omega)\right|^2 d\omega \\ =\ & \frac{1}{2\pi}E\int|\phi_n(\omega)\mathcal{F}[K](h\omega) - \phi(\omega)|^2 d\omega \\ =\ & \frac{1}{2\pi}E\int|\phi_n(\omega)\mathcal{F}[K](h\omega) - \phi(\omega)\mathcal{F}[K](h\omega) + \phi(\omega)\mathcal{F}[K](h\omega) - \phi(\omega)|^2 d\omega \\ =\ & \frac{1}{2\pi}E\int|\mathcal{F}[K](h\omega) - 1|^2|\phi(\omega)|^2 d\omega + \frac{1}{2\pi}E\int|\phi_n(\omega) - \phi(\omega)|^2|\mathcal{F}[K](h\omega)|^2 d\omega \\ =\ & Bias^2 + variance\end{aligned}$$

The variance term can be further simplified as

$$\begin{aligned}E|\phi_n(\omega) - \phi(\omega)|^2 &= E|\phi_n(\omega)|^2 + E|\phi(\omega)|^2 - 2E\phi_n(\omega)\bar{\phi}(\omega) \\ &= E|\phi_n(\omega)|^2 - |\phi(\omega)|^2 \\ &= E\left|\frac{1}{n}\sum_{j=1}^{n}\exp(i\omega X_j)\right|^2 - |\phi(\omega)|^2 \\ &= \frac{n-1}{n}|\phi(\omega)|^2 + \frac{1}{n} - |\phi(\omega)|^2 \\ &= \frac{1}{n}\left(1 - |\phi(\omega)|^2\right)\end{aligned}$$

Hence:

$$\begin{aligned} E\int\left(\hat{f}(x)-f(x)\right)^2 dx \;=\; & \frac{1}{2\pi}E\int\left|\mathcal{F}[K]\left(h\omega\right)-1\right|^2\left|\phi\left(\omega\right)\right|^2 d\omega \text{ (bias}^2\text{)} \\ & +\frac{1}{2\pi n}\int\left|\mathcal{F}[K](h\omega)\right|^2 d\omega \text{ (variance I)} \\ & -\frac{1}{2\pi n}\int\left|\mathcal{F}[K](h\omega)\right|^2\left|\phi\left(\omega\right)\right|^2 d\omega \text{ (variance II)} \end{aligned} \tag{1.21}$$

Remark 1.7.1 *variance I constitutes the main term of the variance. It is similar to the expression obtained before where we did not use Fourier analysis. In fact, by Parseval's theorem, we have*

$$\frac{1}{2\pi n}\int\left|\mathcal{F}[K](h\omega)\right|^2 d\omega=\frac{1}{n}\int K_h^2(x)dx=\frac{1}{nh}\int K^2(u)du=O\left(\frac{1}{nh}\right)$$

which coincides with the variance term given in 1.6.

Remark 1.7.2 *variance II is typically of smaller order than variance I. In fact, if $|\mathcal{F}[K]|$ is bounded by a constant C, we have*

$$\frac{1}{2\pi n}\int\left|\mathcal{F}[K](h\omega)\right|^2\left|\phi\left(\omega\right)\right|^2 d\omega\leq\frac{C}{2\pi n}\int\left|\phi\left(\omega\right)\right|^2 d\omega=\frac{C}{n}\int f^2(x)dx=O\left(\frac{1}{n}\right)$$

Remark 1.7.3 *The bias term in (1.21) is different. It does not necessarily reduce to an expression involving a derivative of f.*

To access the order of MISE, what we really need is the rate that $\left|\phi\left(\omega\right)\right|^2$ converges to zero as $\omega\to\infty$. Based on this observation, we introduce another class of functions:

Definition 1.7.3 *Sobolev class of densities:*

$$\mathcal{S}(s,L)=\left\{f\;|\;f\geq 0,\int f\left(x\right)dx=1,\int\left|\omega\right|^{2s}\left|\phi\left(\omega\right)\right|^2\leq 2\pi L^2\right\}$$

For an integer s, $\mathcal{F}[f^{(s)}]\left(\omega\right)=\left(i\omega\right)^s\phi\left(\omega\right)$. As a result

$$\int\left|\omega\right|^{2s}\left|\phi\left(\omega\right)\right|^2=\int\left|\mathcal{F}[f^{(s)}]\left(\omega\right)\right|^2 ds=2\pi\int\left|f^{(s)}(x)\right|^2 dx$$

So the bound condition is equivalent to

$$\int\left|f^{(s)}(x)\right|^2 dx\leq L^2$$

Using characteristic functions in the definition adds flexibility as it allows for any $s>0$.

Theorem 1.7.1 *Let $K \in L_2(R)$ be symmetric. Assume that for some $s > 0$, there exists a constant A such that*

$$|1 - \mathcal{F}[K](\omega)| \leq A|\omega|^s$$

then

$$\sup_{f \in \mathcal{S}(s,L)} E \int \left(\hat{f}(x) - f(x)\right)^2 dx \leq Cn^{-\frac{2s}{2s+1}}$$

Given the above derivations, the proof is very simple and omitted here.

1.7.3 Flat-top Kernels

The Fourier method enables us to analyze the MISE of kernel estimators with infinite order kernel. An example of infinite order kernel is

$$K_\infty(u) = \frac{1 + \cos u - 2\cos^2 u}{\pi u^2}$$

Define

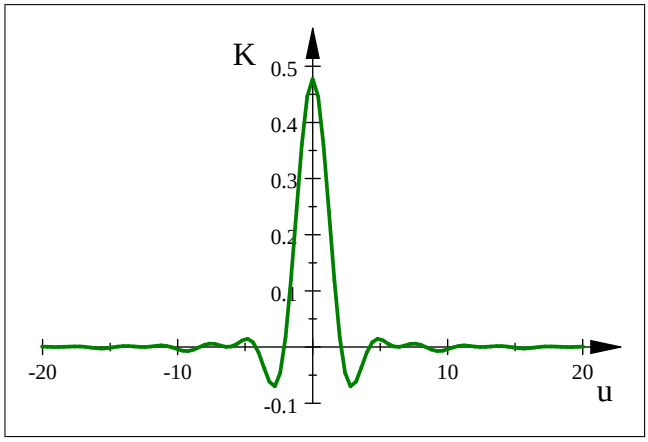


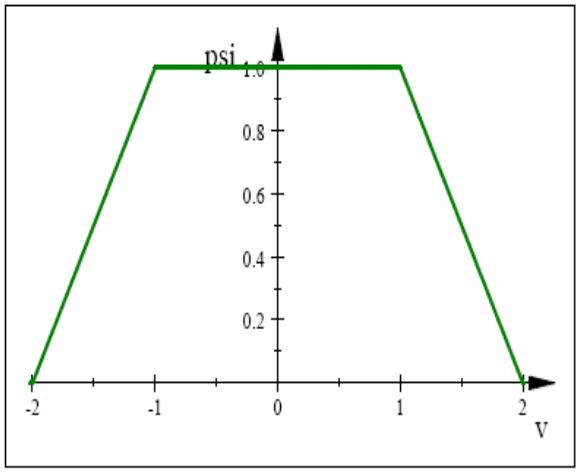
Figure 1.7: Graph of $K_\infty(u)$

$$g(v) = \begin{cases} v & v \geq 0 \\ 0 & v < 0 \end{cases}$$

then

$$\psi_\infty(v) = 2g\left(1 - \frac{|v|}{2}\right) - g(1 - |v|)$$

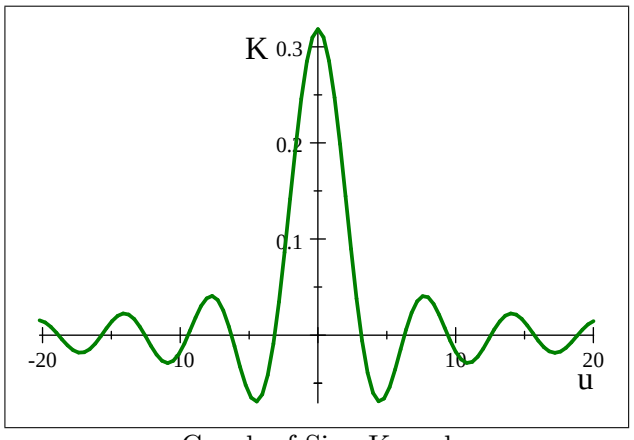
is the characteristic function of $K_\infty(u)$.



Graph of $\psi_\infty(\cdot)$

Another example of infinite order kernel is the so called sinc kernel

$$K(u) = \begin{cases} \frac{\sin u}{\pi u}, & \text{if } u \neq 0 \\ \frac{1}{\pi}, & \text{if } u = 0 \end{cases}$$



Graph of Sinc Kernel

Its Fourier transform (actually it is a version of the Fourier transform) has the form $F[K](\omega) = I(|\omega| \leq 1)$.

In a sequence of papers, Politis (e.g. 1999) and his co-authors have advocated the use of the above infinite order kernels. Politis called them *flat-top* kernels. Figures 1.7 and 1.7.3 graph the functions $K_\infty(u)$ and $\psi_\infty(v)$.

What is the role of the flatness of $\psi_\infty(v)$ play? Using the facts that $\mathcal{F}[K](\omega) = 1$ for $\omega \in [-C, C]$ for some C and that $|\mathcal{F}[K](\omega) - 1|^2 \leq 1$, we have

$$\frac{1}{2\pi} E \int |\mathcal{F}[K](h\omega) - 1|^2 |\phi(\omega)|^2 d\omega \leq \frac{1}{2\pi} \int_{\|\omega\| \geq Ch^{-1}} |\phi(\omega)|^2 d\omega.$$

If we assume that

$$|\phi(\omega)|^2 \leq C_1 e^{-C_2 t} \text{ for some constants } C_1 > 0 \text{ and } C_2 > 0,$$

then

$$\begin{aligned} & \frac{1}{2\pi} E \int |\mathcal{F}[K](h\omega) - 1|^2 |\phi(\omega)|^2 d\omega \\ \leq & \frac{1}{2\pi} \int_{\|\omega\| \geq Ch^{-1}} C_1 e^{-C_2 \omega} d\omega \leq O\left[\exp\left(-C_2 h^{-1}\right)\right] \\ = & o\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{1.22}$$

by letting $h = A/\log n$ for $A > 0$. For the asymptotic variance, we still have

$$var(\hat{f}(x)) = O\left(\frac{1}{nh}\right) = O\left(\frac{\log n}{n}\right). \tag{1.23}$$

Combining (1.22) and (1.23), we get

$$\hat{f}(x) - f(x) = O_p\left(\sqrt{\frac{\log n}{n}}\right),$$

which is only a logarithm factor away from the parametric rate $1/\sqrt{n}$.

Note that in the above derivation, we only used the flatness and boundedness of $\mathcal{F}[K](\omega)$. We did not use other features of $\mathcal{F}[K](\omega)$. This implies that, given any characteristic function $\psi(\cdot)$ with a flat top, we can take the inverse Fourier transform to get an infinite order kernel. So there are many infinite order kernels. Figure 1.8 illustrates a flat-top kernel in the two-dimensional case. This figure comes from Politis and Romano (1999).

1.8 Asymptotic Normality

To prove asymptotic normality, we need a central limit theorem for triangular arrays known as the Lindeberg-Feller theorem. Let $\{U_{n,i} | i = 1, ..., n, n \geq 1\}$ be a triangular sequence of

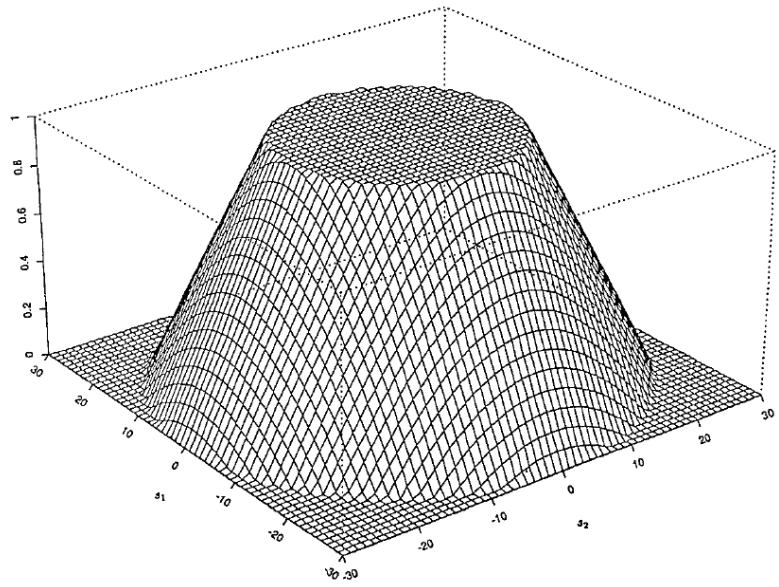


Figure 1.8: A flat-top kernel in R^2

random variables, where $U_{n,i}, i = 1, ..., n$, are independent for any $n \geq 1$. Let

$$\mu_{ni} = E[U_{n,i}], \sigma_{ni}^2 = var(U_{n,i}),\ \sigma_n^2 = \sum_{i=1}^{n} \sigma_{ni}^2$$

Theorem 1.8.1 *Davidson (Stochastic Limit Theory, 1994, Theorem 23.6). If*

$$Z_{ni} = \frac{U_{ni} - \mu_{ni}}{\sigma_n}$$

satisfies

$$\forall \varepsilon > 0 : \lim_{n \to \infty} \sum_{i=1}^{n} E\left[Z_{ni}^2 \{|Z_{ni}| \geq \varepsilon\}\right] = 0, \tag{1.24}$$

then

$$\sum_{i=1}^{n} Z_{ni} \to_d N(0,1).$$

A sufficient condition for the Lindeberg condition (1.24) to hold is obtained from Lyapunov's Theorem (see Davidson (1994, Theorem 23.11):

Lemma 1.8.1 *A sufficient condition for (1.24) to hold is*

$$\lim_{n\to\infty}\sum_{i=1}^{n} E\left|Z_{ni}\right|^{2+\delta}=0 \text{ for some } \delta>0.$$

This enables us to prove the following result:

Theorem 1.8.2 *Assume that*
(i) x is an interior point of the support of X,
(i) f is twice continuously differentiable at x with $\|f''(x)\|<C$ for some constant C,
(iii) $K(u)$ is a second order kernel with $\int|K(u)|^{2+\delta}du<\infty$,
(iv) $nh\to\infty$ and $nh^5\to C\in[0,\infty)$ as $n\to\infty$,
Let

$$B\left(x\right)=\frac{1}{2}f''(x)\left(\int K(u)u^2du\right) \text{ and } V(x)=f(x)\left(\int K^2\left(u\right)du\right),$$

then

$$\sqrt{nh}\left(\hat{f}\left(x\right)-f\left(x\right)-h^2B(x)\right)\to N(0,V(x)).$$

Proof. The proof is based on Pagan and Ullah (1999, page 40). We write

$$\begin{aligned}\sqrt{nh}\left(\hat{f}\left(x\right)-f\left(x\right)-h^2B(x)\right) &= \sqrt{nh}\left[\hat{f}(x)-E\hat{f}(x)\right]+\sqrt{nh}\left(E\hat{f}(x)-f(x)-h^2B(x)\right)\\ &= \sqrt{nh}\left[\hat{f}(x)-E\hat{f}(x)\right]+\sqrt{nh}\times o(h^2)\\ &= \sqrt{nh}\left[\hat{f}(x)-E\hat{f}(x)\right]+o\left(1\right)\end{aligned}$$

by the assumption that $nh^5\to C$ as $n\to\infty$.

In order to apply Theorem 1.8.1 to the first term, we define

$$U_{n,i}=\frac{1}{\sqrt{nh}}K\left(\frac{X_i-x}{h}\right),\ Z_{ni}=\frac{U_{ni}-\mu_{ni}}{\sigma_n},$$

$$\sigma_n^2=n\cdot var\left[\frac{1}{\sqrt{nh}}K\left(\frac{X_i-x}{h}\right)\right]=V(x)(1+o(1),$$

such that

$$\begin{aligned}\frac{\sqrt{nh}\left[\hat{f}(x)-E\hat{f}(x)\right]}{\sqrt{V(x)}} &= \sum_{i=1}^{n}\frac{U_{ni}-\mu_{ni}}{\sigma_n}(1+o(1))\\ &:= \sum_{i=1}^{n}Z_{ni}(1+o(1)).\end{aligned}$$

Now

$$\begin{aligned}\sum_{i=1}^{n} E\left|Z_{ni}\right|^{2+\delta} &= \left(\sigma_n^2\right)^{-(1+\delta/2)} \sum_{i=1}^{n} E\left|U_{ni}-\mu_{ni}\right|^{2+\delta} \\ &= \left(\sigma_n^2\right)^{-(1+\delta/2)} nE\left|U_{n1}-\mu_{n1}\right|^{2+\delta}.\end{aligned}$$

Using the c_r inequality (Davidson (p. 140), $E\left(|X+Y|^r\right) \leq c_r\left\{E\left(|X|^r\right)+E\left(|Y|^r\right)\right\}$) and Jensen's inequality, we have

$$\begin{aligned}E\left|U_{n1}-\mu_{n1}\right|^{2+\delta} &\leq 2^{1+\delta}\left[E\left(\left|U_{n1}\right|^{2+\delta}\right)+\left(E\left|U_{n1}\right|\right)^{2+\delta}\right] \\ &\leq 2^{2+\delta}E\left(\left|U_{n1}\right|^{2+\delta}\right).\end{aligned}$$

Alternatively

$$\begin{aligned}E\left|U_{n1}-\mu_{n1}\right|^{2+\delta} &\leq E\left|2\max\left(\left|U_{n1}\right|,\left|\mu_{n1}\right|\right)\right|^{2+\delta} \\ &\leq 2^{2+\delta}E\max\left(\left|U_{n1}\right|^{2+\delta},\left|\mu_{n1}\right|^{2+\delta}\right) \\ &\leq 2^{2+\delta}\left(E\left|U_{n1}\right|^{2+\delta}+\left|\mu_{n1}\right|^{2+\delta}\right) \leq 2^{3+\delta}E\left(\left|U_{n1}\right|^{2+\delta}\right).\end{aligned}$$

This bound is not as good as the bound from the c_r inequality but it suffices for the proof here. Therefore

$$\begin{aligned}\sum_{i=1}^{n} E\left|Z_{ni}\right|^{2+\delta} &\leq 2^{2+\delta}\left(\sigma_n^2\right)^{-(1+\delta/2)} nE\left(\left|U_{n1}\right|^{2+\delta}\right) \\ &= O\left[nE\left(\left|U_{n1}\right|^{2+\delta}\right)\right] \\ &= O\left[n\left(nh\right)^{-(1+\delta/2)}\int\left|K(\frac{v-x}{h})\right|^{2+\delta} f(v)dv\right] \\ &= O\left[nh\left(nh\right)^{-(1+\delta/2)}\int\left|K(u)\right|^{2+\delta} f(x+uh)du\right] \\ &= O\left(\left(nh\right)^{-\delta/2}\right).\end{aligned}$$

Combining the above proof yields the stated result. ■

Remark 1.8.1 *If $nh^5 \to 0$, then it follows from the above theorem that pointwise confidence interval can be constructed as follows:*

$$\left[\hat{f}\left(x\right)-z_{\alpha/2}\sqrt{\frac{\hat{V}(x)}{nh}},\hat{f}\left(x\right)+z_{\alpha/2}\sqrt{\frac{\hat{V}(x)}{nh}}\right]$$

where $\hat{V}(x)$ *is a consistent estimator of* $V(x)$:

$$\hat{V}(x) = \hat{f}(x) \int K^2(u)\, du.$$

Remark 1.8.2 *It can be shown that* $cov(\hat{f}(x), \hat{f}(y)) \to 0$ *for any two distinct points* x *and* y. *Using the Cramer-Wold device, we can show that for any finite collection of points* $x_1, ..., x_N$:

$$\begin{pmatrix} \sqrt{nh}\left[\hat{f}(x_1) - f(x_1) - h^2 B(x_1)\right] \\ ... \\ \sqrt{nh}\left[\hat{f}(x_N) - f(x_N) - h^2 B(x_N)\right] \end{pmatrix} \to N(0, diag(V(x_1), ..., V(x_N))).$$

This is very useful when constructing pointwise confidence intervals since one does not have to take into account any dependence between the different points. This independence also implies that functional central limit theorem does not hold for $\hat{f}(x)$ *when* $\hat{f}(x)$ *is regarded as a random function in a certain function space.*

1.9 Uniform Consistency

The pointwise consistency result in section 1.3.1 can be strengthened to uniform consistency, that is, we show convergence in the sup-norm $\sup_{x \in S} \left\| \hat{f}(x) - f(x) \right\|$where S is a compact set excluding the boundary range of the support of X.

The uniform consistency result will be needed to prove the uniform consistency of conditional moment estimators in the next chapter. The uniform consistency result is also an ingredient for establishing asymptotic properties of semiparametric estimators in later chapters.

Theorem 1.9.1 *Assume*

(i) $f(x) \in C^2(\mathcal{X})$ *where* $\mathcal{X} = supp(X)$

(ii) $K(\cdot)$ *is a bounded function that satisfies condition* (1.4) *and for any* $c_1 > 0$ *and* $c_2 > 0$, $\int_{-c_1/h}^{c_2/h} K(u)\, du = 1 + o(h^2)$ *and* $\int_{-c_1/h}^{c_2/h} uK(u)\, du = o(h^2)$ *as* $h \to 0$.

(iii) $K(\cdot)$ *is Lipschtz continunous. Then*

$$\sup_{x \in S} \left\| \hat{f}(x) - f(x) \right\| = O_p\left(\sqrt{\frac{\log n}{nh}} + h^2 \right), \tag{1.25}$$

as $n \to \infty$, $h \to 0$ *such that* $nh / \log n \to \infty$.

To prove the theorem, we write

$$\hat{f}(x) - f(x) = \hat{f}(x) - E\hat{f}(x) + E\hat{f}(x) - f(x)$$

and show that

$$\begin{aligned} \sup_{x\in S}\left|E\hat{f}(x) - f(x)\right| &= O_p(h^2), \\ \sup_{x\in S}\left|\hat{f}(x) - E\hat{f}(x)\right| &= O_p\left(\sqrt{\frac{\log n}{nh}}\right). \end{aligned} \tag{1.26}$$

The first result in (1.26) holds because for $\tilde{x} \in [x, x+uh]$, we have

$$\begin{aligned} \left|E\hat{f}(x) - f(x)\right| &= \left|\int [f(x+uh) - f(x)]\, K(u)du\right| + o(h^2) \\ &= \left|\frac{h^2}{2}\int u^2 f''(\tilde{x})K(u)du\right| + o(h^2) \\ &\leq Ch^2 \int u^2 K(u)du = O(h^2) \end{aligned}$$

uniformly over $x \in S$.

We now prove the second result in (1.26). Since S is compact, it can be covered by finite number of intervals, i.e.,

$$S \subseteq \cup_{\ell=1}^{L(n)} I_\ell,\ I_\ell := [x_{n\ell} - \delta_n, x_{n\ell} + \delta_n]$$

where $L(n)$ is a finite constant but depends on n and S. We will specify δ_n below. Given this, we can write

$$\begin{aligned} \sup_{x\in S}\left|\hat{f}(x) - E\hat{f}(x)\right| &= \max_{1\leq\ell\leq L(n)} \sup_{x\in S\cap I_\ell}\left|\hat{f}(x) - E\hat{f}(x)\right| \\ &\leq \max_{1\leq\ell\leq L(n)} \sup_{x\in S\cap I_\ell}\left|\hat{f}(x) - \hat{f}(x_{n,\ell})\right| \\ &\quad + \max_{1\leq\ell\leq L(n)}\left|\hat{f}(x_{n,\ell}) - E\hat{f}(x_{n,\ell})\right| \\ &\quad + \max_{1\leq\ell\leq L(n)} \sup_{x\in S\cap I_\ell}\left|E\hat{f}(x_{n,\ell}) - E\hat{f}(x)\right| \\ &:= Q_1 + Q_2 + Q_3. \end{aligned}$$

(Step 1. The term Q_2) Write

$$W_n(x) = \sum_{i=1}^{n} W_{n,i}(x),$$

where

$$W_{ni}(x) = \frac{1}{n}\left[K_h(x - X_i) - EK_h(x - X_i)\right].$$

Given that $K(\cdot)$ is a bounded function, we have

$$|W_{n,i}(x)| \leq C_0\left(\frac{1}{nh}\right)$$

for some constant $C_0 > 0$ and

$$\begin{aligned} EW_{n,i}^2(x) &\leq \frac{1}{n^2h^2}\int K^2\left(\frac{x-v}{h}\right)f(v)dv \\ &= \frac{1}{n^2h}\int K^2(u)\,f(x+uh)du \leq \frac{C_1}{n^2h} \end{aligned}$$

for some constant C_1 that does not depend on x.

Let

$$\lambda_n = \sqrt{nh\log n},$$

then

$$|\lambda_n W_{n,i}(x)| \leq C_0\left(\frac{\sqrt{nh\log n}}{nh}\right) = C_0\left(\sqrt{\frac{\log n}{nh}}\right) \leq 1/2$$

for n large enough. Using the inequalities that

$$\begin{aligned} 1 + x &\leq \exp(x) \text{ for } x \geq 0 \text{ and} \\ \exp(x) &\leq 1 + x + x^2 \text{ for } |x| \leq 1/2, \end{aligned}$$

we have

$$\begin{aligned}
P\left(W_n(x) > \eta\right) &= P\left(\sum_{i=1}^{n} W_{n,i}(x) > \eta\right) \\
&\leq \frac{E\exp\left[\lambda_n \sum_{i=1}^{n} W_{n,i}(x)\right]}{\exp\left(\lambda_n \eta\right)} \leq \frac{\prod_{i=1}^{n} E\exp\left[\lambda_n W_{n,i}(x)\right]}{\exp\left(\lambda_n \eta\right)} \\
&= \exp\left(-\lambda_n \eta\right) \prod_{i=1}^{n} \left(1 + E\lambda_n W_{n,i}(x) + E\left[\lambda_n^2 W_{n,i}^2(x)\right]\right) \\
&= \exp\left(-\lambda_n \eta\right) \prod_{i=1}^{n} \left(1 + E\left[\lambda_n^2 W_{n,i}^2(x)\right]\right) \\
&\leq \exp\left(-\lambda_n \eta\right) \prod_{i=1}^{n} \exp\left\{E\left[\lambda_n^2 W_{n,i}^2(x)\right]\right\} \\
&\leq \exp\left(-\lambda_n \eta\right) \exp\left[n\lambda_n^2 E W_{n,i}^2(x)\right] \\
&\leq \exp\left(-\lambda_n \eta + C_1 \frac{\lambda_n^2}{nh}\right)
\end{aligned}$$

We want to choose η such that the exponent to be of order $-\log n$. This can be done by setting

$$\eta = \eta_n = C_2 \frac{\lambda_n}{nh} = C_2 \frac{\sqrt{nh \log n}}{nh} = C_2 \sqrt{\frac{\log n}{nh}}$$

for C_2 sufficiently large. We have thus shown that

$$P\left(W_n(x) > \eta_n\right) = O(n^{-\alpha}), \text{ for } \alpha = C_2 - C_1$$

Similarly, we can show that

$$P\left(W_n(x) < -\eta_n\right) = O(n^{-\alpha}), \text{ for } \alpha = C_2 - C_1$$

Therefore

$$\begin{aligned}
P\left(Q_2 > \eta_n\right) &\leq P\left[\max_{1 \leq \ell \leq L(n)} \left|W_n\left(x_{n,\ell}\right)\right| > \eta_n\right] \\
&\leq L(n) \sup_{x \in S} P\left(\left|W_{n,i}(x)\right| > \eta_n\right) = O\left(L(n) n^{-\alpha}\right).
\end{aligned}$$

(Step 2. The terms $\mathbf{Q}_1$ and $\mathbf{Q}_3$) By the Lipschiz condition on $K(\cdot)$, we know that

$$\sup_{x \in S \cap I_\ell} \left|K\left(\frac{X_i - x}{h}\right) - K\left(\frac{X_i - x_{n,\ell}}{h}\right)\right| \leq C_3 h^{-1} \delta_n.$$

Upon choosing

$$\delta_n = h^{3/2}\sqrt{\frac{\log n}{n}},$$

we get

$$|Q_1| \leq C_3 h^{-2}\left(h^{3/2}\sqrt{\frac{\log n}{n}}\right) = O_p\left(\sqrt{\frac{\log n}{nh}}\right).$$

By the same argument, we can show that

$$|Q_3| = O_p\left(\sqrt{\frac{\log n}{nh}}\right).$$

For this choice of δ_n, $L_n = O(\delta_n^{-1})$ and we have $P(Q_2 > \eta_n) \to 0$ as $n \to \infty$.

1.10 Multivariate Density Estimation

We now wish to extend the above results to the general case where $X \in \mathbb{R}^d$. To do this, we need some additional notation. For any square matrix A, let $|A|$ denote the associated determinant. Define

$$\hat{f}(x) = \frac{1}{n|H|}\sum_{i=1}^{n}\mathbf{K}(H^{-1}(x - X_i))$$

where $\mathbf{K} : \mathbb{R}^d \to \mathbb{R}$ is a multivariate kernel with $\int \mathbf{K}(u)du = 1$ and $\int \mathbf{K}(u)udu = 0$, and $H \in \mathbb{R}^{d\times d}$ is positive definite. This is the multivariate kernel density estimator, where H now is a matrix of bandwidths. Often, one chooses $H = hI_d$, $h > 0$, so the same bandwidth h is used for all variables, in which case

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n}\mathbf{K}\left(\frac{x - X_i}{h}\right).$$

Similarly, the kernel $\mathbf{K}$ is often chosen as a product kernel,

$$\mathbf{K}(x_1, x_2, ..., x_d) = \prod_{i=1}^{d} K(x_i)$$

for some univariate kernel $K : \mathbb{R} \to \mathbb{R}$.

Under some regular conditions and using the same arguments for the univariate case, we can show that

$$\begin{aligned} Bias\left(\hat{f}(x)\right) &= h^2 B(x)(1 + o(1)) \\ var(\hat{f}(x)) &= \frac{1}{nh^d}V(x)(1 + o(1)) \end{aligned}$$

where

$$B(x) = \frac{1}{2}\int_{\mathbb{R}^d} \nabla^2 f(x,u)\mathbf{K}(u)du, V(x) = f(x)\int_{\mathbb{R}^d} \mathbf{K}^2(u)\,du.$$

$$\nabla^2 f(x,u) = \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_i \partial x_j} u_i u_j.$$

Following the arguments used in the univariate case, we can obtain that the optimal choice in the IMSE-sense is

$$h^* = Cn^{-1/(d+4)}.$$

With this choice of bandwidth, the MSE is of order $n^{-4/(d+4)}$ and the estimator $\hat{f}(x)$ converges to $f(x)$ at the rate of $n^{-2/(d+4)}$. The rate of convergence depends on the dimension of X. As d increases, the rate of convergence decreases. This is normally referred to as the curse of dimensionality: A large number of variables implies that the associated kernel estimator has a slow convergence rate. As a consequence, a greater number of observations is needed in the multivariate case for the kernel estimator to be reasonably precise. This is in contrast to the parametric case, where the rate of convergence is affected neither by the dimension of the data variables, nor by the dimension of the parameter space.

We can also obtain the CLT for $\hat{f}(x)$:

$$\sqrt{nh^d}\left(\hat{f}(x) - f(x) - h^2 B(x)\right) \to N(0, V(x)),$$

as $nh^d \to \infty$ and $nh^{d+4} \to C \in [0,\infty)$. As expected, the convergence in distribution is also slowed down compared to the univariate case.

1.11 Conditional Density Estimation

We briefly introduce conditional density estimation. See Ch 5 of Li and Racine (2007) for a detailed treatment.

Let $f_{Y|X}(y|x)$ be the conditional density of Y conditional on $X = x \in \mathbb{R}^d$. By definition,

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)}$$

where $f_{Y,X}(y,x)$ is the joint density of (Y,X) and $f_X(x)$ is the marginal density of X. We can estimate both these densities by kernel methods:

$$\begin{aligned}
\hat{f}_{Y,X}(y,x) &= \frac{1}{nh^{d+1}}\sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right)\mathbf{K}\left(\frac{X_i - x}{h}\right), \\
\hat{f}_X(x) &= \frac{1}{nh^d}\sum_{i=1}^{n} \mathbf{K}\left(\frac{X_i - x}{h}\right),
\end{aligned}$$

where $K : \mathbb{R} \to \mathbb{R}$ and $\mathbf{K} : R^d \to \mathbb{R}$, and thereby obtain an estimator of the conditional density:

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{Y,X}(y,x)}{\hat{f}_X(x)}. \tag{1.27}$$

For simplicity, we here use the same bandwidth for all variables; one could choose to use different bandwidths for each variable. We can now use the previous results to establish the asymptotic properties of $\hat{f}_{Y|X}(y|x)$. The important thing to note here is that $\hat{f}_{Y,X}(y,x)$ converges at a slower rate than $\hat{f}_X(x)$ since it has a higher dimension, so $\hat{f}_{Y,X}(y,x)$ will supply all leading terms in the asymptotic results.

To show consistency, we observe that under $h \to 0$, and $nh^{d+1} \to \infty$, both the numerator and denominator of (1.27) are pointwise consistent. Thus,

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{Y,X}(y,x)}{\hat{f}_X(x)} \to \frac{f_{Y,X}(y,x)}{f_X(x)} = f_{Y|X}(y|x).$$

Asymptotic normality of $\hat{f}_{Y|X}(y|x)$ is shown by the so-called delta-method. First, we expand the function $h(a,b) = a/b$ around a_0/b_0 :

$$\frac{a}{b} - \frac{a_0}{b_0} = \frac{1}{b_0}(a - a_0) - \frac{a_0}{b_0^2}(b - b_0) + O((a-a_0)^2) + O((b-b_0)^2).$$

Therefore

$$\begin{aligned} & \hat{f}_{Y|X}(y|x) - f_{Y|X}(y|x) \\ = \; & \frac{1}{f_X(x)}\left(\hat{f}_{Y,X}(y,x) - f_{Y,X}(y,x)\right) + \frac{f_{Y,X}(y,x)}{f_X^2(x)}\left(\hat{f}_X(x) - f_X(x)\right) \\ & + O\left[\left(\hat{f}_{Y,X}(y,x) - f_{Y,X}(y,x)\right)^2\right] + O\left[\left(\hat{f}_X(x) - f_X(x)\right)^2\right]. \end{aligned}$$

The four terms are $O_p\left(1/\sqrt{nh^{d+1}}\right)$, $O_p\left(1/\sqrt{nh^d}\right)$, $O_p\left(1/(nh^{d+1}\right)$ and $O_p\left(1/(nh^d)\right)$ respectively. So the first term will drive the asymptotic distribution. Since

$$\sqrt{nh^{d+1}}\left[\hat{f}_{Y,X}(y,x) - f_{Y,X}(y,x)\right] \to_d N\left[0, f_{Y,X}(y,x)\int_{\mathbb{R}} K^2(u)du \int_{\mathbb{R}^d} \mathbf{K}^2(v)\,dv\right],$$

when $nh^{d+1} \to \infty$ and $nh^{d+5} \to 0$, we have

$$\sqrt{nh^{d+1}}\left[\hat{f}_{Y|X}(y|x) - f_{Y|X}(y|x)\right] \to N\left[0, \frac{f_{Y,X}(y,x)}{f_X^2(x)}\int_{\mathbb{R}} K^2(u)du \int_{\mathbb{R}^d} \mathbf{K}^2(v)\,dv\right]$$

when $nh^{d+1} \to \infty$ and $nh^{d+5} \to 0$.

1.12 Time Series KDE

For strictly stationary time series data, the asymptotic bias of $\hat{f}(x)$ is the same as the iid case. The difference lies in variance calculation. We assume that X_t is an α-mixing process with the mixing coefficient

$$\alpha(\ell) = \sup_{A \in \mathcal{F}_t, B \in \mathcal{F}_{t+\ell}} |P(A)P(B) - P(AB)|$$

satisfying

$$\|\alpha(\ell)\| \leq C\ell^{-\beta} \text{ for some } C \text{ and } \beta > 2.$$

Let $Z_t = \sqrt{h}K_h(X_t - x) - E\sqrt{h}K_h(X_t - x)$, then

$$\sqrt{nh}\left[\hat{f}(x) - E\hat{f}(x)\right] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t$$

and

$$var\left\{\sqrt{nh}[\hat{f}(x) - E\hat{f}(x)]\right\} = var(Z_1) + 2\sum_{\ell=1}^{T-1}\left(1 - \frac{\ell}{T}\right)cov(Z_1, Z_{\ell+1}).$$

As before

$$\begin{aligned} var(Z_1) &= \frac{1}{h}E\left[K\left(\frac{X_t - x}{h}\right)\right]^2 (1 + o(1)) \\ &= \left[\int K^2(u)du\right] f(x)(1 + o(1)). \end{aligned}$$

It remains to examine the covariance terms. We will show that

$$\sum_{\ell=1}^{T-1} |cov(Z_1, Z_{\ell+1})| = o(1).$$

First, let $g_\ell(u, v)$ be the pdf of $(X_1, X_{\ell+1})$, then

$$\begin{aligned} |cov(Z_1, Z_{\ell+1})| &\leq \left|h\int\int K_h(u - x)K_h(v - x)g_\ell(u, v)dudv\right| \\ &\quad + \left|h\left(\int K_h(u - x)f_X(u)du\right)\left(\int K_h(v - x)f_X(v)dv\right)\right| \\ &= h\|g_\ell\|_\infty + hC = O(h). \end{aligned}$$

Second, by Lemma 3 in Sec 1.2.2 of Doukhan (1994, page 10), we have

$$\begin{aligned} cov(Z_1, Z_{\ell+1}) &\leq 4\alpha(\ell)\max(|Z_1|)\max(|Z_{\ell+1}|) \\ &= 4\alpha(\ell)\|K_\infty\|^2/h. \end{aligned}$$

Using the above two results, we obtain, for some d_T

$$\begin{aligned}\sum_{\ell=1}^{T-1}|cov(Z_1,Z_{\ell+1})| &= \sum_{\ell=1}^{d_T}|cov(Z_1,Z_{\ell+1})|+\sum_{\ell=d_T+1}^{T-1}|cov(Z_1,Z_{\ell+1})| \\ &\leq O\left(d_T h\right)+O\left(\frac{1}{h}\sum_{\ell=d_T+1}^{\infty}\ell^{-\beta}\right) \\ &= O\left(d_T h\right)+O\left(\frac{1}{h}\frac{1}{d_T^{\beta-1}}\right).\end{aligned}$$

Taking $d_T = h^{-2/\beta}$ yields

$$\begin{aligned}\sum_{\ell=1}^{T-1}|cov(Z_1,Z_{\ell+1})| &= O\left(h^{1-2/\beta}\right)+O\left(h^{\frac{2(\beta-1)}{\beta}-1}\right) \\ &= O\left(h^{1-2/\beta}\right)+O\left(h^{1-2/\beta}\right)=O\left(h^{1-2/\beta}\right),\end{aligned}$$

which implies that

$$\sum_{\ell=1}^{T-1}|cov(Z_1,Z_{\ell+1})| = o\left(1\right).$$

Consequently,

$$var\left\{\sqrt{nh}[\hat{f}\left(x\right)-E\hat{f}\left(x\right)]\right\}=\left[\int K^2(u)du\right]f\left(x\right)\left(1+o\left(1\right)\right).$$

Under some additional conditions, we can show that $\sqrt{nh}\left[\hat{f}\left(x\right)-f\left(x\right)\right]$ is asymptotically normal with variance $\left[\int K^2(u)du\right]f\left(x\right)$.

From the asymptotics given above, the time series dependence does not affect the bias and variance of the KDE. It may be regarded as a benefit of the KDE, as we can use the same procedure for the iid data in the time series setting. However, in finite samples, the time series dependence does affect the sampling distribution of the KDE. So the asymptotics that does not capture time series dependence should be used with caution. Following Chen, Liao and Sun (2014), we can design an inference procedure that is easy to use and at the same time captures the time series dependence. This is reported in the working paper Kim, Yang and Sun (2015).

1.13 Problems

1. (Theoretical question) Prove equation (1.18).

2. (Theoretical question) Let $X_1, ..., X_n \sim iidf(x)$. Suppose we want to estimate, not the pdf $f(x)$, but its derivative $f'(x)$ and its integral $F(x)$ (*i.e.* CDF). Reasonable estimators of $f'(x)$ and $F(x)$ are

$$\begin{aligned}\hat{f}'(x) &= \frac{1}{nh}\sum_{i=1}^{n}\frac{d}{dx}K\left(\frac{X_i - x}{h}\right), \\ \hat{F}(x) &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x}K\left(\frac{X_i - v}{h}\right)dv.\end{aligned}$$

Derive the AMSE's of $\hat{f}'(x)$ and $\hat{F}(x)$ and obtain the optimal bandwidths. Compare them to the optimal bandwidth for the estimation of $f(x)$ itself. Make appropriate smoothness assumptions if needed.

Hint: (i) Note that

$$\begin{aligned}E\hat{f}'(x) &= -\frac{1}{h^2}EK'\left(\frac{X_i - x}{h}\right) = -\frac{1}{h^2}\int K'\left(\frac{z - x}{h}\right)f_X(z)\,dz = -\frac{1}{h}\int K'(u)\,f_X(x + uh)\,du \\ &= -\frac{1}{h}\int K'(u)\left[f_X(x) + f'_X(x)\,uh + \frac{1}{2}f''_X(x)\,u^2h^2 + \frac{1}{6}f'''_X(x)\,u^3h^3\right]du + s.o. \\ &= f'_X(x)\left(-\int K'(u)\,udu\right) - \left(\int K'(u)\,u^3du\right)h^2\left[\frac{1}{6}f'''_X(x)\right] + s.o.\end{aligned}$$

Assume that $\int K'(u)\,udu = -1$, then the asymptotic bias of $\hat{f}'(x)$ is

$$-\left(\int K'(u)\,u^3du\right)h^2\left[\frac{1}{6}f'''_X(x)\right]$$

On the other hand, the asymptotic variance of $\hat{f}'(x)$ is of order $1/\left(nh^3\right)$. So the optimal bandwidth is of order $n^{-1/7}$.

(ii) See Li and Racine (2007, 22).

3. (Theoretical question) Exercise 1.6 in Li and Racine (2007)

4. (Theoretical question) Let $f_a(x)$, where $a > 0$, be the class of all probability densities on R such that the support of the characteristic function $\phi(t)$ included in a given interval $[-a, a]$. Show that for any $n \geq 1$ the kernel density estimator $\hat{f}(x)$ with the sinc kernel and appropriately chosen bandwidth h satisfies

$$E\left[\hat{f}(x) - f_a(x)\right]^2 \leq \frac{a}{\pi n}$$

This example, due to Ibragimov and Has'minskii (1983, page 48), shows that it is possible to estimate the density with the $\sqrt{n}$ rate on sufficiently small nonparametric classes of functions.

Hint: Use Fourier transform to characterize the pointwise bias

5. (Computer exercise) Consider a mixture normal density

$$f(x;p,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2) = p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)+(1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

Generate a single random sample of $n = 1000$ i.i.d. observations, $X_1,..,X_{1000}$, $X_i \sim f(x;1,2,0.8,\mu_2,\sigma_2^2)$, *i.e.* $X_i \sim iidN(2,0.8)$.

(a) Given this sample, calculate the pointwise kernel estimator

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$$

using the Gaussian kernel:

$$K(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$$

over a grid of points $x \in \{x_k, k = 1,...,M\}$. We here choose

$$x_k = \frac{k}{M}z_1(X) + (1-\frac{k}{M})z_{99}(X)$$

where $M = 50$, and where $z_\alpha(X)$ is the α-th percentile/quantile of $\{X\}$. (You can simulate the two quantiles first). Do this for four different bandwidths: $h = 0.5, 0.15, 0.05$, and 0.01. Plot the four kernel estimates in the same diagram together with the true density. Comment on the results. Which bandwidth does the best job?

(b) Let's compare the nonparametric estimator with the parametric version. Calculate the MLE for the simulated data set, and plot the associated parametric density estimate in the same diagram as the kernel estimate (using the bandwidth that gave the best fit). Which appears to do the best job?

(c) Assume that $f(\cdot)$ is known. Calculate the optimal bandwidth. How does this perform compared to the ones chosen in the previous question?

(d) Assume that $f(\cdot)$ is not known. Compute Silverman's rule of thumb bandwidth.

(e) Now let us find the optimal bandwidth via cross validation. Plot CV(h) over a suitable grid of bandwidths and find the cross-validated bandwidth.

(f) We now have three different bandwidths from (c), (d), (e). Calculate and plot the kernel estimate for each bandwidth in the same diagram together with the true density. How do the three resulting kernel estimate perform? Do the results make sense?

(g) We wish to estimate the pointwise 95%-confidence band. Do this using $h = 0.15$. The confidence interval can be estimated by

$$\left[\hat{f}(x) - 1.96\sqrt{\frac{\hat{V}(x)}{nh}}, \hat{f}(x) + 1.96\sqrt{\frac{\hat{V}(x)}{nh}}\right]$$

where $\hat{V}(x)$ is a consistent estimator of $V(x)$:

$$\hat{V}(x) = \hat{f}(x) \int K^2(u)\, du.$$

We can also construct the confidence band of the parametric density estimator using the asymptotic normality. Plot the kernel estimate (with Gaussian kernel) and the parametric estimate with their respective confidence bands. From these, which appears to be the most precise?

1.14 References

1. Chen, X., Liao, Z. and Y. Sun (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics* 178(3), 639–658

2. Davidson, J. (1994). *Stochastic Limit Theory.* Oxford University Press.

3. Devroye, L. and G. Lugosi (2001). *Combinatorial Methods in Density Estimation.* Springer.

4. Doukhan, P. (1994). *Mixing: Properties and Examples* (Lecture Notes in Statistics). Springer.

5. Hall, P. (1982). Limit theorems for stochastic measures of the accuracy of density estimators. *Stochastic Processes and Applications* 13(1), l1–25.

6. Hall, P. and Marron, J. S. (1991). Local minima in crossvalidation functions. J. Roy. Statist. Soc. B(53), 245–252.

7. Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

8. Härdle, W. and O. Linton (1994). Applied Nonparametric Methods. In *Handbook of Econometrics*, Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2295–2339. Elsevier.

9. Ibragimov and Has'minskii (1983). Estimation of distribution density, *Journal of Soviet Mathematics*, 21(1), 40–57.

10. Ichimura, H. and P. Todd (2007). Implementing nonparametric and semiparametric estimators, *Handbook of Econometrics* vol. 6 (eds. J.J. Heckman & E.E. Leamer). Elsevier.

11. Jones, M. C., Marron, J. S. and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation, *J. Am. Statist. Assoc.* 91, 401–407.

12. Li, Q. and J.S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press

13. Politis, D. and J. Romano (1999). Multivariate Density Estimation with General Flat-Top Kernels of Infinite Order, *Journal of Multivariate Analysis*, 68(1), 1–25.

14. Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press.

15. Priestley, M. B. (1981). *Spectral Analysis and Time Series*, Academic Press.

16. Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

17. Sheather, S.J. (2004). Density Estimation, *Statistical Science*, 19(4), 588–597.

18. Stone, C.J. (1984). An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *Annals of Statistics* 12, 1285–1297.

19. Hjort, N. L. and M. C. Jones (1996). Locally parametric nonparametric density estimation, *Annals of Statistics* 24(4), 1619–1647.

Chapter 2

Kernel Smoothing: Regression Estimation

2.1 Introduction

Suppose that we observe an iid sample $\{Y_i, X_i\}_{i=1}^n$ generated from

$$Y_i = m(X_i) + \varepsilon_i$$

where $X_i \in \mathbb{R}$, $Y_i \in \mathbb{R}$, and $\{\varepsilon_i\}$ are random errors with

$$E(\varepsilon_i | X_i = x) = 0 \text{ and } var(\varepsilon_i | X_i = x) = \sigma^2(x).$$

Then $m(\cdot)$ is the regression function of Y on X. Given that $m(x) = E(Y|X = x)$, we can also refer to $m(\cdot)$ as the conditional moment function. So the problem of estimating $m(\cdot)$ can be called either nonparametric regression estimation as in Li and Racine (2007) or nonparametric conditional moment estimation as in Pagan and Ullah (1999).

We focus on the case with one regressor. The extension to cases with multiple regressors is straightforward. We discuss a number of estimators $\hat{m}(x)$ of $m(x)$. We often refer to $\hat{m}(x)$ as a smoother. Most the estimators considered here are linear smoothers.

Definition 2.1.1 *An estimator $\hat{m}(x)$ of $m(x)$ is a linear smoother if, for each x, there exists a vector $w_n(x) = (w_{n1}(x), ..., w_{nn}(x))$ such that*

$$\hat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i,$$

where the weighting sequence depends only on $X_1, X_2, ..., X_n$, i.e. $w_{ni}(x) = w_{ni}(x; X_1, ..., X_n)$.

Different linear smoothers arise from different motivations and have different statistical properties. The methods we consider are appropriate for both random design, where X_i are iid and fixed design where X_i are fixed in repeated samples. Fixed design appears more often in statistical literature than in the econometric literature. In econometrics, we always treat X_i as random variables and so we have random designs. In the random design case, X is an ancillary statistics, and standard statistical practice, see Cox and Hinkley (1974), is to conduct inference conditional on the sample $\{X_i\}_{i=1}^n$. However, many papers in the literature prove theoretical properties unconditionally, and we will, for ease of presentation, present results in this form.

2.2 Kernel Estimators: Local Smoothing

2.2.1 Nadaraya-Watson Estimator

To motivate the Nadaraya-Watson estimator, we assume that both Y and X have unbounded supports. In this case, we do not have to worry about the boundary bias problem. The unbounded support assumption is not necessary, and we use it only to simplify the presentation.

Recall that

$$m(x) = E(Y|X=x) = \int yf(y|X=x)dy = \int \frac{yf(x,y)}{f_X(x)}dy = \frac{\int yf(x,y)dy}{\int f(x,y)dy} \tag{2.1}$$

where $f(x,y)$ is the joint density of (X,Y). A natural way to estimate $m(\cdot)$ is first to compute the estimate $f(x,y)$ and then to integrate it according to this formula. A kernel density estimator of $f(x,y)$ is

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

where $K(\cdot)$ is a kernel satisfying $\int K(u)du = 1, K(u) = K(-u)$, and $K_h(\cdot) = 1/hK(\cdot/h)$. We have

$$\int \hat{f}_h(x,y)\,dy = \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\int K_h(Y_i - y)\,dy = \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)$$

and

$$\begin{aligned}\int y\hat{f}_h(x,y)\,dy &= \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\frac{1}{h}\int yK\left(\frac{Y_i - y}{h}\right)dy \\ &= \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\int (Y_i + hu)K(u)\,du \\ &= \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)Y_i\end{aligned}$$

using the properties that $\int K(u)du = 1$ and $\int uK(u)du = 0$. Plugging these into the numerator and denominator of (2.1), we obtain the Nadaraya-Watson kernel estimator:

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

The bandwidth h determines the degree of 'roughness' of $\hat{m}_h(x)$. This can be immediately seen by considering the limits for h tending to zero or to infinity, respectively. Indeed, at an observation X_i, $hK_h(X_j - X_i) \to K(0)1\{X_j = X_i\}$ as $h \to 0$, and so $\hat{m}_h(X_i) \to Y_i$ as $h \to 0$. On the other hand, at an arbitrary point x, $K_h(X_i - x) \approx K_h(X_j - x)$ as $h \to \infty$, and so $\hat{m}_h(x) \to \bar{Y}$ as $h \to \infty$. These two limit considerations make it clear that the smoothing parameter h should not converge to zero too rapidly or too slowly.

2.2.2 k-Nearest Neighbor Estimators

The kernel estimator is defined as a weighted average of the dependent variable in a fixed neighborhood of x. The k-nearest neighbor (kNN) estimator is defined as a weighted average of the dependent variable in a varying neighborhood. This neighborhood is defined as the k-nearest neighbors of a point x.

Let $\mathcal{N}(x) = \{i : X_i \text{ is one of the kNN to } x\}$ be the set of indices of the k-nearest neighbors of x. The kNN estimator is the average of Y's with index in $\mathcal{N}(x)$:

$$\hat{m}_k = \frac{1}{k} \sum_{i \in \mathcal{N}(x)} Y_i.$$

The kNN estimator can be regarded as a kernel estimator with uniform kernel $K(u) = 1/2\{|u| \leq 1\}$ and variable bandwidth $h = R(k)$, the distance between x and its furthest kNN:

$$\hat{m}_k(x) = \frac{\sum_{i=1}^n K_R(X_i - x) Y_i}{\sum_{i=1}^n K_R(X_i - x)}. \tag{2.2}$$

Note that in the above equation, the (normalized) denominator is

$$\frac{1}{n} \sum_{i=1}^n K_R(X_i - x) = \frac{k}{2nR} \tag{2.3}$$

which is the kNN density estimate of $f_X(x)$. Intuitively, the CDF of X at the point x can be estimated by

$$\hat{F}_X(x) = \frac{1}{n} \{\# \text{ of } X_i\text{'s} \leq x\}.$$

In view of

$$f_X(x) = \lim_{h \to 0} \frac{F_X(x+h) - F_X(x-h)}{2h},$$

we can estimate $f_X(x)$ by

$$f_X(x) = \frac{1}{2nh}\{\#\text{ of } X_i\text{'s falling in the interval } [x-h, x+h]\}$$

Letting $h = R(k)$ gives equation (2.3).

For certain econometrics models (e.g. quantile regression), we need to estimate $1/f_X(x)$ in order to make inferences. A natural estimate is then

$$\widehat{\frac{1}{f_X(x)}} = \frac{2nR}{k}.$$

This is the so-called spacing estimator. See Kaplan (2015) on how to select k in order to design an inference procedure with improved finite sample performances.

Formula (2.2) provides sensible estimators for arbitrary kernels. The bias and variance of the more general kNN estimator is given in a theorem by Mack (1981). In contrast to kernel smoothing, the variance of the kNN regressor smoother does not depend on f. This makes sense, since the kNN estimator always averages over exactly k observations independently of the distribution of the X-variables. The bias constant $B_n(x)$ is also different from the one for kernel estimators. An approximate identity between kNN and kernel estimators can be obtained by setting

$$k = 2nhf_X(x)$$

or equivalently $h = k/(2nf_X(x))$. For this choice of h or k respectively, the asymptotic mean squared error formulas for the kNN estimator and the corresponding NW estimator are the same.

2.2.3 Local Polynomial Estimators

The NW estimator can be regarded as the solution to the minimization problem:

$$\hat{m}_h(x) = \arg\min_{\theta} \sum_{i=1}^{n} K_h(X_i - x)\{Y_i - \theta\}^2. \tag{2.4}$$

Thus $\hat{m}_h(x)$ is obtained by a local constant least squares approximation of the outputs Y_i. The locality is determined by a kernel K that downweighs all the X_i that are not close to x whereas θ plays the role of a local constant to be fitted. More generally, we may define a local polynomial least squares approximation, replacing in (2.4) the constant θ by a polynomial of given degree. Let $m(x)$ be sufficiently smooth such that for z sufficiently close to x we may write

$$\begin{aligned} m(z) &= m(x) + m'(x)(z-x) + ... + \frac{m^{(p)}(x)}{p!}(z-x)^p \\ &= \theta_0 + \theta_1\left(\frac{z-x}{h}\right) + ... + \theta_p\frac{1}{p!}\left(\frac{z-x}{h}\right)^p \end{aligned}$$

where

$$(\theta_0, ..., \theta_p)' = \left(m(x), m'(x) h, ..., m^{(p)}(x) h^p \right)'$$

This motivates the local polynomial class of estimators. Let $\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_p$ minimize

$$\sum_{i=1}^{n} K_h (X_i - x) \left\{ Y_i - \theta_0 - \theta_1 \left(\frac{X_i - x}{h} \right) - ... - \theta_p \frac{1}{p!} \left(\frac{X_i - x}{h} \right)^p \right\}^2. \tag{2.5}$$

Then $\hat{\theta}_0$ serves as an estimator of $m(x)$, while $\hat{\theta}_j h^{-j}$ serves as an estimator of the j-th derivative of $m(x)$

A variation on these estimators called LOESS or LOWESS (LOcal (W)eighted Regression + Explained Sum of Squares) was first considered in Cleveland (1979) who employed a nearest neighbor window and local polynomial regression. For some informal discussions on LOESS, you can search for "local regression" on wiki.

The local linear estimator is unbiased when $m(\cdot)$ is linear, while the NW estimator may be biased depending on the marginal density of the design. High order polynomial can achieve bias reduction, see Fan and Gijbels (1996).

The principle underlying the local polynomial estimator can be generalized in a number of ways. Tibshirani (1994) introduced the local likelihood procedure in which an arbitrary parametric regression function $g(x, \theta)$ substitutes the polynomial in (2.5). Loader (1999) has written a monograph on this topic. The idea of local polynomial approximations can be used in different contexts. For example, Andrews and Sun (2004) use the idea to improve long memory estimation. An adaptive procedure on selecting the degree of the polynomial is also given there.

2.2.4 Robust Smoothing

As in parametric methods, it may be desirable to down-weight the effects of large error terms. This is achieved by using a loss function different from the quadratic loss function. Let $\rho(\cdot)$ be a general loss function, we can estimate $m(x)$ by $\hat{\theta}(x)$, the minimizer of the following criterion function:

$$\hat{\theta}(x) = \arg\min \sum_{i=1}^{n} K \left(\frac{x - X_i}{h} \right) \rho(Y_i - \theta)$$

To achieve robustness, we typically assume that $\rho(\cdot)$ is minimized at zero with $\rho(t) \leq t^2$ for large $|t|$. For example, if $\rho(t) = |t|$, then we have the least absolute deviations criterion and the resulting estimator can be interpreted as a local median regression function. More generally, if $\rho(t)$ is the check function:

$$\rho(t) := \rho_\tau(t) = \begin{cases} t(\tau - 1) & t < 0 \\ t\tau & t \geq 0 \end{cases}$$

for some $\tau \in (0,1)$, then the resulting estimator can be interpreted as a local quantile regression function at quantile τ. When $\tau = 0.8$, the check function is given in Figure 2.1

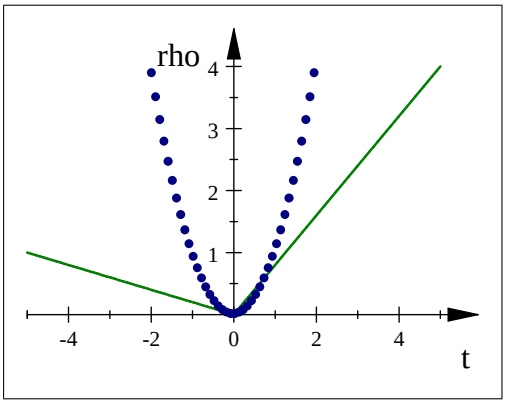


Figure 2.1: Check function $\rho_\tau(t)$ with $\tau = 0.8$ vs quadratic function t^2

Computation of robust smoothing estimator can be carried out by linear programming techniques. Chaudhuri (1991) provides asymptotic theory for this estimator in a general multidimensional context. See Koneker (2005) for an excellent introduction to quantile regression.

2.3 Asymptotic Properties: Local Constant Estimator

2.3.1 Consistency

Recall

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{i=1}^n K_h(x - X_i)},$$

so

$$\begin{aligned}
\hat{m}(x) - m(x) &= \frac{\sum_{i=1}^n K_h(x - X_i)[Y_i - m(x)]}{\sum_{i=1}^n K_h(x - X_i)} \\
&= \frac{\sum_{i=1}^n K_h(x - X_i)[m(X_i) - m(x)]}{\sum_{i=1}^n K_h(x - X_i)} + \frac{\sum_{i=1}^n K_h(x - X_i)\varepsilon_i}{\sum_{i=1}^n K_h(x - X_i)} \\
&:= \frac{\hat{e}_1(x)}{\hat{f}_X(x)} + \frac{\hat{e}_2(x)}{\hat{f}_X(x)}
\end{aligned}$$

where $\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_i)$ and

$$\hat{e}_1(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_i)[m(X_i) - m(x)], \ \hat{e}_2(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\varepsilon_i.$$

By definition,

$$\begin{aligned} E\hat{e}_1(x) &= \frac{1}{h}\int K\left(\frac{v-x}{h}\right)[m(v)-m(x)]\,f_X(v)\,dv \\ &= \int_{(x_L-x)/h}^{(x_U-x)/h} K(u)\,[m(x+uh)-m(x)]\,f_X(x+uh)\,du \end{aligned}$$

where x_L and x_U are the lower and upper limits of the support of X. If x is an interior point such that $(x_U - x)/h > c$ and $(x - x_L)/h > c$ for some finite c and the kernel K is of finite support, we can replace the limits of the integration by the support of the kernel. In the sequel, we shall assume this is the case. Therefore

$$\begin{aligned} E\hat{e}_1(x) &= \int K(u)\,[m(x+uh)-m(x)]\,f_X(x+uh)\,du \\ &= \int K(u)\left[m'(x)uh + \frac{1}{2}m''(x)u^2h^2 + O(h^3)\right]\left[f_X(x) + f_X'(x)uh + O(h^2)\right]du \\ &= \frac{\mu_2 h^2}{2}\left[m''(x)f_X(x) + 2m'(x)f_X'(x)\right] + O(h^3) \end{aligned} \tag{2.6}$$

where

$$\mu_2 = \int u^2 K(u)\,du.$$

Also

$$\begin{aligned} var\left[\hat{e}_1(x)\right] &= var\left\{\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\,[m(X_i)-m(x)]\right\} \\ &= \frac{1}{n}var\left\{K_h(x-X_i)\,[m(X_i)-m(x)]\right\} \\ &\le \frac{1}{n}E\left\{K_h^2(x-X_i)\,[m(X_i)-m(x)]^2\right\} \\ &= \frac{1}{nh^2}\int K^2(\frac{v-x}{h})\,[m(v)-m(x)]^2\,f_X(v)\,dv \\ &= \frac{1}{nh}\int K^2(u)\,[m(x+uh)-m(x)]^2\,f_X(x+uh)\,du \\ &= O\left(\frac{1}{nh}h^2\right) = O\left(\frac{h}{n}\right). \end{aligned} \tag{2.7}$$

Combining (2.6) and (2.7) gives

$$\hat{e}_1(x) = \frac{\mu_2 h^2}{2}\left[m''(x)f_X(x) + 2m'(x)f_X'(x)\right] + O_p(h^3) + O_p\left(\sqrt{\frac{h}{n}}\right). \tag{2.8}$$

Next, we observe that $E\hat{e}_2(x) = 0$ and that

$$\begin{aligned} E\hat{e}_2^2(x) &= \frac{1}{n^2}\sum_{i=1}^{n} EK_h^2\left(x - X_i\right)\sigma^2\left(X_i\right) \\ &= \frac{1}{nh^2}\int K^2\left(\frac{x-v}{h}\right)\sigma^2(v)f_X(v)dv \\ &= \frac{1}{nh}\sigma^2\left(x\right)f_X\left(x\right)\int K^2(u)du\left(1+o\left(1\right)\right). \end{aligned}$$

Hence

$$\hat{e}_2(x) = O_p\left(\frac{1}{\sqrt{nh}}\right).$$

We have thus shown that

$$\hat{e}_1\left(x\right) + \hat{e}_2\left(x\right) = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right).$$

Combining this with

$$\hat{f}_X\left(x\right) = f_X\left(x\right) + o_p\left(1\right),$$

we obtain

$$\hat{m}\left(x\right) - m(x) = O_p\left(\frac{\left\|\hat{e}_1\left(x\right)\right\| + \left\|\hat{e}_2\left(x\right)\right\|}{f_X\left(x\right) + o_p\left(1\right)}\right) = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right), \tag{2.9}$$

provides that $f_X\left(x\right) > \Delta$ for some $\Delta > 0$.

It now follows from (2.9) that when $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $\hat{m}(x)$ is consistent for $m(x)$. To get some intuition, we consider $K\left(u\right) = \frac{1}{2}\left\{|u| < 1\right\}$. In this case

$$\begin{aligned} \hat{m}\left(x\right) &= \frac{\sum_{i=1}^{n} K_h\left(x - X_i\right)\left[m(X_i) + \varepsilon_i\right]}{\sum_{i=1}^{n} K_h\left(x - X_i\right)} = \frac{\sum_{i:|X_i - x| \le h}\left[m(X_i) + \varepsilon_i\right]}{\sum_{i:|X_i - x| \le h}} \\ &= \underbrace{\textit{average of } m(X_i)} + \underbrace{\textit{average of } \varepsilon_i} \text{ for } i \text{ such that } |X_i - x| \le h. \end{aligned}$$

Now when $|X_i - x| \le h$, we have $m(X_i) - m(x) = m'(x)\left(X_i - x\right) + \frac{1}{2}m''\left(\tilde{x}_i\right)\left(X_i - x\right)^2$. Because $K\left(u\right)$ is symmetric, $m'(x)\left(X_i - x\right)$ in the approximation of $m(X_i) - m(x)$ will be "averaged out", leaving only the term of order $O(h^2)$. This gives rise to the bias of $\hat{m}\left(x\right)$. Next, the second term above involves the average of ε_i. Its order depends on the cardinality of the set $\left\{i : |X_i - x| \le h\right\}$ which is approximately $n \times 2hf_X\left(x\right) = O(nh)$. the variance of the second term is then $O(1/(nh))$. We often call nh the effective sample size or the local sample size.

2.3.2 Asymptotic Normality

The asymptotic normality of $\hat{m}(x)$ can also be established at this stage using the Lindeberg-Feller CLT.

Theorem 2.3.1 *Assume that (i) x is an interior point of the support of X,*

(ii) $m(x)$ and $f_X(x)$ have continuous and bounded second order derivatives and $f_X(x) > b > 0$ for some b.

(iii) $K(u)$ is a second order kernel with $\int [K(u)]^{2+\delta} du < \infty$ for some $\delta > 0$,

(iv) $E(|\varepsilon_i|^{2+\delta} | X_i) < C$ almost surely for some constant C,

(v) $h \to 0$ and $nh^5 \to M \in [0, \infty)$ as $n \to \infty$.

Then

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2 B(x)\right) \to_d N(0, V(x))$$

where

$$B(x) = \frac{\mu_2}{2}\left[m''(x) + 2m'(x)\frac{f_X'(x)}{f_X(x)}\right],$$
$$V(x) = \frac{\sigma^2(x)}{f_X(x)}\int K^2(u)du.$$

Proof. Note that

$$\sqrt{nh}\left[\hat{m}(x) - m(x) - \frac{\hat{e}_1(x)}{\hat{f}_X(x)}\right] = \frac{1}{\sqrt{nh}}\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\varepsilon_i}{\hat{f}_X(x)}$$
$$= \frac{1}{\sqrt{nh}}\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\varepsilon_i}{f_X(x)}(1 + o(1)) := \frac{1}{f_X(x)}\sum_{i=1}^n \omega_{ni}\varepsilon_i$$

where

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

In order to apply the Lindeberg-Feller CLT given in Chapter 1, we compute

$$\sigma_n^2 = var\left(\sum_{i=1}^n \omega_{ni}\varepsilon_i\right) = E\sum_{i=1}^n \frac{1}{nh}K^2\left(\frac{x - X_i}{h}\right)\sigma^2(X_i)$$
$$= \frac{1}{h}\int K^2\left(\frac{x-v}{h}\right)\sigma^2(v) f_X(v)dv = \sigma^2(x) f_X(x)\int K^2(u)du + o(1).$$

Let

$$Z_{ni} = \frac{\omega_{ni}\varepsilon_i}{\sigma_n},$$

then it remains to verify the condition that

$$\lim_{n\to\infty}\sum_{i=1}^{n} E\left|Z_{ni}\right|^{2+\delta}=0 \text{ for some } \delta>0,$$

see Lemma 1.8.1 in Chapter 1. But

$$\begin{aligned}\sum_{i=1}^{n} E\left|Z_{ni}\right|^{2+\delta} &= \sum_{i=1}^{n} E\left|\omega_{ni}\right|^{2+\delta}\left|\varepsilon_i\right|^{2+\delta}\\ &\leq CnE\left|\omega_{ni}\right|^{2+\delta}=Cn\left(nh\right)^{-1-\delta/2}\int K^{2+\delta}\left(\frac{x-v}{h}\right)f_X(v)dv\\ &= Cnh\left(nh\right)^{-1-\delta/2}\int K^{2+\delta}\left(u\right)f_X(x+uh)du\\ &= O\left[\left(nh\right)^{-\delta/2}\right]\to 0 \text{ as } n\to\infty.\end{aligned}$$

Therefore we have shown that

$$\sum_{i=1}^{n}\omega_{ni}\varepsilon_i \to N\left[0,\sigma^2\left(x\right)f_X(x)\int K^2(u)du\right]$$

and

$$\sqrt{nh}\left[\hat{m}\left(x\right)-m(x)-\frac{\hat{e}_1(x)}{\hat{f}_X(x)}\right]\to_d N\left[0,\frac{\sigma^2\left(x\right)}{f_X(x)}\int K^2(u)du\right].$$

Combining this with (2.8) yields the desired result. ■

Remark 2.3.1 *What is especially notable is the presence of the term* $2m'(x)\frac{f'_X(x)}{f_X(x)}$ *in the bias. We call it the design bias since it depends on the design, that is, the distribution of the* X_i*'s. This means that the bias is sensitive to the position of the* X_i*'s. Furthermore, it can be shown that kernel estimators also have high bias near the boundaries. This is known as boundary bias. We will see that we can reduce these biases by using a refined local polynomial regression.*

Remark 2.3.2 *The theorem shows that the asymptotic bias of* $\hat{m}(x)$ *is*

$$asymbias(\hat{m})=h^2B(x)$$

and the asymptotic variance is

$$asymvar(\hat{m}\left(x\right))=\frac{1}{nh}V\left(x\right).$$

So the asymptotic mean squared error for an interior point x is

$$AMSE = h^4 B^2(x) + \frac{1}{nh} V(x).$$

The best rate is obtained by taking $h \sim n^{-1/5}$. In this case, the asymptotic MSE is of order $n^{-4/5}$ and $\hat{m}(x)$ converges to $m(x)$ at the rate of $n^{-2/5}$. This rate is slower than the parametric rate $n^{-1/2}$.

Remark 2.3.3 *In the above discussion, "asymptotic bias" refers to the difference of the mean of the asymptotic distribution of the estimator and the true value $m(x)$. For nonlinear models, this notion of asymptotic bias is typical. This is because it is difficult, if not impossible, to calculate $E\hat{m}(x) - m(x)$ in nonlinear models. When $E\hat{m}(x) - m(x)$ exists and can be computed without much difficulty, "asymptotic bias" often refers to the dominating term of $E\hat{m}(x) - m(x)$. In the present case, we can actually show that*

$$E\hat{m}(x) - m(x) = h^2 B(x)(1 + o(1)).$$

Getting this result requires additional steps. See Pagan and Ullah (1999, p.101).

Remark 2.3.4 *A key feature in the bias calculation is that the kernel is of second order so that it has zero first moment. In turns out that if $K(\cdot)$ is a higher order kernel and $m(\cdot)$ and $f_X(\cdot)$ are sufficiently smooth, then the bias can be reduced even further. Specifically, suppose $K(\cdot)$ is a q-th order kernel, then*

$$\begin{aligned}
E\hat{e}_1(x) &= \int K(u)\left[m(x+uh) - m(x)\right] f_X(x+uh)\, du \\
&= \int K(u)\left[m(x+uh) f_X(x+uh) - m(x) f_X(x+uh)\right] du \\
&= \int K(u)\left[\sum_{j=1}^{q-1} \frac{(m f_X)^{(j-1)}}{(j-1)!}(uh)^{j-1} + \frac{(m f_X)^{(q)}}{q!}(uh)^q\right] du \\
&\quad - m(x)\int K(u)\left[\sum_{j=1}^{q-1} \frac{f_X^{(j-1)}}{(j-1)!}(uh)^{j-1} + \frac{f_X^{(q)}}{q!}(uh)^q\right] du + O\left(h^{q+1}\right) \\
&= \frac{h^q}{q!}\left[(m f_X)^{(q)}(x) - \left(m f_X^{(q)}\right)(x)\right]\int u^q K(q) du.
\end{aligned}$$

In this case, the optimal bandwidth is $h \sim n^{-1/(2q+1)}$ and the rate of convergence for $\hat{m}(x)$ is $n^{-q/(2q+1)}$. This rate is arbitrarily close to the parametric rate when q is large enough.

Remark 2.3.5 ***Curse of dimensionality.*** *When there are many $X's$ and the bandwidth matrix is of the form $H = hI_d$, the variance of the NW estimator is of order $1/(nh^d)$, where d is the dimension of X. Because of the high variance, the optimal bandwidth is $h \sim n^{-1/(2q+d)}$ and the resulting rate of convergence for $\hat{m}(x)$ is only $n^{-q/(2q+d)}$. The reason for this high variance is that in high dimensions observations are more spread out. Consider an example. Suppose we have n data points uniformly distributed on the interval $[0,1]$. How many data points will we find in the interval $[0,0.1]$? The answer is: about $n/10$ points. Now suppose we have n data points on the 10-dimensional unit cube $[0,1]^{10}$. How many data points will we find in the cube $[0,0.1]^{10}$? The answer is about*

$$n \times \frac{1}{10^{10}}$$

Thus, n has to be huge to ensure that small neighborhoods have any data in them.

Remark 2.3.6 *The asymptotic distribution can be used to calculate pointwise confidence interval for the NW estimator. In practice, it is usual to ignore the bias term, since this is rather complicated, depending on higher derivatives of the regression function and perhaps the derivatives of the density of X. This approach can be justified when a bandwidth is chosen so that the bias is relatively small. That is, we suppose that h^2 is small relative to $1/\sqrt{nh}$, i.e. $h = o(n^{-1/5})$. In this case, the interval*

$$\left[\hat{m}(x) - z_{\alpha/2}\sqrt{\widehat{var}(\hat{m}(x))}, \hat{m}(x) + z_{\alpha/2}\sqrt{\widehat{var}(\hat{m}(x))}\right]$$

where $\widehat{var}(\hat{m}(x))$ is a consistent estimate of the asymptotic variance of $\hat{m}(x)$, is a valid $(1-\alpha)$ confidence set. However, the rate condition $h = o(n^{-1/5})$ does not provide any practical guidance.

Remark 2.3.7 *To get consistent estimates of $var(\sqrt{nh}\hat{m}(x))$, we exploit the linearity of the estimator*

$$\hat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i,$$

where $w_{ni}(x)$ depends only on the design. Note that $var(\sqrt{nh}\hat{m}(x)) = Evar(\sqrt{nh}\hat{m}(x)|X_1^n]$ where

$$var(\sqrt{nh}\hat{m}(x)|X_1^n] = nh\sum_{i=1}^{n} w_{ni}^2(x)\,\sigma^2(X_i).$$

Therefore, consider

$$\hat{V}_1(x) = nh\sum_{i=1}^{n} w_{ni}^2(x)\,\hat{\varepsilon}_i^2,$$

$$\hat{V}_2(x) = \frac{\hat{\sigma}^2(x)}{\hat{f}(x)}\int K^2(u)du,$$

where

$$\hat{\varepsilon}_i = Y_i - \hat{m}(X_i), \ \hat{\sigma}^2(x) = \sum_{i=1}^{n} w_{ni}^2(x)\hat{\varepsilon}_i^2.$$

Both estimators are consistent, and so are the confidence intervals based on them.

2.4 Bandwidth Selection

2.4.1 Discrepancy Measures

In this section, we describe two methods of bandwidth selection for nonparametric regression estimation. We first define some performance criteria for an estimator $\hat{m}(x)$ of the function $m(\cdot)$. In the sequel, $\pi(\cdot)$ denote some weighting function defined on the support of X.

1. Pointwise MSE (MSE or PMSE):

$$d_{MSE}(\hat{m}(x), m(x)) = E[(\hat{m}(x) - m(x))^2].$$

The MSE measures the squared deviation of the estimator $\hat{m}$ from m at a single point x.

2. Integrated SE (ISE):

$$d_{ISE}(\hat{m}, m) = \int [(\hat{m}(x) - m(x))^2]\pi(x)dx$$

is a global discrepancy measure. But it is still a random variable as different samples will produce different $d_{ISE}(\hat{m}, m)$. The weight function $\pi(x)$ may be used to assign less weight to observations in regions of sparse data (to reduce the variance in this region) or at the tail of the distribution of (to trim away boundary effects).

3. Mean Integrated SE (MISE) or Integrated Mean SE (IMSE):

$$d_{MISE}(\hat{m}, m) = Ed_{ISE}(\hat{m}, m) = \int E[(\hat{m}(x) - m(x))^2]\pi(x)dx.$$

MISE is not a random variable. It is the expected value of the random variable d_{ISE} with the expectation being taken with respect to all possible samples of X and Y.

4. Average SE (ASE):

$$d_{ASE}((\hat{m}, m)) = \frac{1}{n}\sum_{i=1}^{n} [\hat{m}(X_i) - m(X_i)]^2 \omega(X_i)$$

is a discrete approximation to ISE, and just like the ISE, it is both a random variable and a global measure of discrepancy.

Which discrepancy measure should be used to derive a rule for choosing h? A natural choice would be MISE or its asymptotic version AMISE since we have some experience of its optimization from the density case. It turns out that when $\pi(x) = \omega(x) f_X(x)$ all three measures $d_{ISE}, d_{MISE}, d_{ASE}$ lead asymptotically to the same level of smoothing (Marron and Hardle (1986)). Hence, we can use different measures to motivate different bandwidth selection rules.

2.4.2 MSE-optimal Bandwidth and Plug-in Implementation

When the pointwise MSE criterion is used, the optimal bandwidth has a closed form solution. The PMSE is

$$h^4 B^2(x) + \frac{1}{nh} V(x),$$

and the MSE-optimal bandwidth is

$$h_{MSE} = \left(\frac{V(x)}{4B^2(x)} \right)^{1/5} n^{-1/5}.$$

Frequently, people work with an integrated MSE (IMSE or MISE) criterion, in which case, the optimal bandwidth is

$$h_{IMSE} = \left(\frac{\int V(x)\pi(x)\,dx}{4\int B^2(x)\pi(x)\,dx} \right)^{1/5} n^{-1/5}.$$

which depends on the unknown quantities:

$$\sigma^2(x), f_X(x), f_X'(x), f_X''(x), m'(x), m''(x).$$

The plug-in approach is the same as in the case of density estimation. It involves nonparametrically estimating the unknown quantities $B(x)$ and $V(x)$ by $\hat{B}(x)$ and $\hat{V}(x)$, say, and then let

$$\hat{h}_{MSE} = \left(\frac{\hat{V}(x)}{4\hat{B}^2(x)} \right)^{1/5} n^{-1/5}$$

and

$$\hat{h}_{IMSE} = \left(\frac{\int \hat{V}(x)\pi(x)\,dx}{4\int \hat{B}^2(x)\pi(x)\,dx} \right)^{1/5} n^{-1/5}.$$

If $\hat{B}(x) \to B(x)$ and $\hat{V}(x) \to V(x)$, then

$$\frac{\hat{h}_{MSE}(x) - h_{MSE}(x)}{h_{MSE}(x)} \to_p 0.$$

On the other hand, if

$$\sup_{x:\pi(x)>0}\left(\left|\hat{B}\left(x\right)-B\left(x\right)\right|+\left|\hat{V}\left(x\right)-V(x)\right|\right)\rightarrow_{p}0,$$

then

$$\frac{\left|\hat{h}_{IMSE}-h_{IMSE}\right|}{h_{IMSE}}\rightarrow_{p}0.$$

The disadvantage of this method is that one must estimate the derivatives of m and f, which are typically poorly behaved estimates [the variance of a kernel estimate of $m''\left(x\right)$ is of order $1/(nh^{5})$]. In addition, to estimate the derivatives, we need to select some initial or "pilot" bandwidth. However, if the pilot bandwidth is not close to its optimal value, the second step plug-in bandwidth can be far away from its optimal value. For more details, see Loader (1999).

Silverman (1986) suggests a compromise method he called rule of thumb. This involves specifying an auxiliary parametric model for the data distribution and using this to infer a simple formula for the optimal bandwidth. Analogous approach has been used in the time series literature on HAC estimation (e.g. Andrews (1991)).

2.4.3 Cross Validation

General Theory

Another performance criterion for bandwidth choice is the average sum of squared residuals (ASSR):

$$d_{ASSR}=\frac{1}{n}\sum_{i=1}^{n}\left[Y_{i}-\hat{m}(X_{i})\right]^{2}\omega\left(X_{i}\right).$$

We choose h to minimize d_{ASSR}. Unfortunately, this method will always lead us to select $h=0$, in which case $\hat{m}(X_{i})=Y_{i}$ for all i and $d_{ASSR}=0$. The basic problem is that (X_{i},Y_{i}) is included in the sample when we try to predict Y_{i} and the dimension of the model is potentially very large. When $h=0$, we can think that the model is of dimension n and we have overfitting. To overcome this problem, we predict the value of Y_{i} without including the observation point (X_{i},Y_{i}) in the sample. The criterion function becomes:

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}\left[Y_{i}-\hat{m}_{-i}(X_{i})\right]^{2}\omega\left(X_{i}\right)$$

where

$$\hat{m}_{-i}\left(X_{i}\right)=\frac{\sum_{\ell\neq i}K_{h}\left(X_{i}-X_{\ell}\right)Y_{\ell}}{\sum_{\ell\neq i}K_{h}\left(X_{i}-X_{\ell}\right)}$$

is the leave-one-out (LOO) kernel estimator of $m(X_i)$.

The proof of the next theorem shows that, up to a constant, $CV(h)$ is a consistent estimator of the dominating terms in $d_{IMSE}(\hat{m}, m)$ when $\pi(x) = \omega(x) f_X(x)$.

Theorem 2.4.1 *Let*

$$H_n = \left\{ cn^{-1/5} : c \in [c_L, c_U] \text{ for some } 0 < c_L \leq c_U < \infty \right\}.$$

Assume

(i) the support of X is $\mathbb{S}$, a compact set,
(ii) $m(x) \in C^2(\mathbb{S})$ and $f_X(x) \in C^2(\mathbb{S})$, $\sigma^2(x) \in C^2(\mathbb{S})$;
(iii) $f_X(x) > \Delta > 0$ for $x \in \mathbb{S}$ and $\omega(x) \in C(\mathbb{S})$
(iv) ε_i is independent of X_i,
(v) $K \in C^2(\mathbb{R})$ with $\int u^2 K(u) < \infty$ and $\int K^2(u) < \infty$.
Then we have, for $\pi(x) = \omega(x) f_X(x)$,

$$\begin{aligned} CV(h) &= \frac{1}{n}\sum_{i=1}^{n} [\hat{m}(X_i) - m(X_i)]^2 \omega(X_i) + constant + o_p\left(n^{-4/5}\right) \\ &= n^{-4/5}\left[c^4 \int B(x)\pi(x)\,dx + \frac{1}{c}\int V(x)\pi(x)\,dx\right] + cons + o_p\left(n^{-4/5}\right) \end{aligned}$$

uniformly over $h \in H_n$.

Proof: By definition,

$$\begin{aligned} CV(h) &= \frac{1}{n}\sum_{i=1}^{n} [\hat{m}_{-i}(X_i) - m(X_i) - \varepsilon_i]^2 \omega(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} (\hat{m}_{-i}(X_i) - m(X_i))^2 \omega(X_i) - \frac{2}{n}\sum_{i=1}^{n} [\hat{m}_{-i}(X_i) - m(X_i)]\omega(X_i)\varepsilon_i + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \omega(X_i) \end{aligned}$$

The last term does not depend on h and can thus be ignored. For the first term, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n} (\hat{m}_{-i}(X_i) - m(X_i))^2 \omega(X_i) = \frac{1}{n}\sum_{i=1}^{n} (\hat{m}(X_i) - m(X_i))^2 \omega(X_i) + o_p\left(n^{-4/5}\right).$$

For the second term, we use Lemma 1 of Härdle and Marron (1985), which says that

$$\sup_{x\in\mathbb{S}} \sup_{h\in H_n} \left| \frac{1}{n}\sum_{\ell=1}^{n} K_h(X_\ell - x) - f_X(x) \right| = o_p(1).$$

So

$$
\begin{aligned}
&-\frac{2}{n}\sum_{i=1}^{n}\left[\hat{m}_{-i}(X_i)-m(X_i)\right]\omega\left(X_i\right)\varepsilon_i \\
&=-\frac{2}{n}\sum_{i=1}^{n}\left[\frac{\sum_{\ell\neq i}K_h\left(X_i-X_\ell\right)\left[Y_\ell-m(X_i)\right]}{\sum_{\ell\neq i}K_h\left(X_i-X_\ell\right)}\right]\omega\left(X_i\right)\varepsilon_i \\
&=-\frac{2}{n^2}\sum_{i=1}^{n}\sum_{\ell\neq i}K_h\left(X_\ell-X_i\right)\left[Y_\ell-m(X_i)\right]\frac{\omega\left(X_i\right)\varepsilon_i}{f_X\left(X_i\right)}\left(1+o_p(1)\right) \\
&=-\left(2T_{n1}+2T_{n2}\right)\left(1+o_p(1)\right).
\end{aligned}
$$

uniformly over $h\in H_n$ where

$$
\begin{aligned}
T_{n1}&=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{e}_1\left(X_i\right)\omega\left(X_i\right)}{f_X\left(X_i\right)}\varepsilon_i \\
T_{n2}&=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{\ell\neq i}\frac{K_h\left(X_\ell-X_i\right)\omega\left(X_i\right)}{f_X\left(X_i\right)}\varepsilon_\ell\varepsilon_i.
\end{aligned}
$$

Note that $T_{n2}\neq\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{e}_2(X_i)}{f_X(X_i)}\omega\left(X_i\right)\varepsilon_i$ because one observation is left out in the second sum in T_{n2}.

The term T_{n1} has mean zero and variance:

$$
\begin{aligned}
&E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_1\left(X_i;X_{-i}\right)\omega\left(X_i\right)\frac{\varepsilon_i}{f_X\left(X_i\right)}\right)^2 \\
&=\frac{1}{n^2}E\sum_{i=1}^{n}\left\{\hat{e}_1\left(X_i;X_{-i}\right)\right\}^2\frac{\omega^2\left(X_i\right)\sigma^2\left(X_i\right)}{f_X^2\left(X_i\right)}\ \{\text{by conditioning on } X_1,...,X_n\} \\
&=\frac{1}{n}E\left\{\frac{\hat{e}_1^2(X_i;X_{-i})\omega^2\left(X_i\right)\sigma^2\left(X_i\right)}{f_X^2\left(X_i\right)}\right\}=\frac{1}{n}\int\frac{\left[E\hat{e}_1^2(x;X_{-i})\right]\omega^2\left(x\right)\sigma^2\left(x\right)}{f_X\left(x\right)}dx \\
&=O\left[\frac{1}{n}\left(\frac{h}{n}+h^4\right)\right]=O\left(n^{-9/5}\right),
\end{aligned}
$$

uniformly over $c\in[c_L,c_U]$. Here the first $O\left(\cdot\right)$ term follows from the mean and variance calculation.

The term T_{n2} is asymptotically equivalent to a statistic of the form:

$$
\frac{1}{n(n-1)}\sum_{i\neq\ell}\tilde{U}_{i\ell}\text{ for }\tilde{U}_{i\ell}=K_h\left(X_\ell-X_i\right)\omega\left(X_i\right)\frac{\varepsilon_\ell\varepsilon_i}{f_X\left(X_i\right)},
$$

which can be written as a U-statistic of the form:

$$\frac{1}{2n(n-1)}\sum_{i\neq\ell}U_{i\ell} \text{ for } U_{i\ell}=K_h\left(X_\ell-X_i\right)\varepsilon_\ell\varepsilon_i\left(\frac{\omega\left(X_i\right)}{f_X\left(X_i\right)}+\frac{\omega\left(X_\ell\right)}{f_X\left(X_\ell\right)}\right).$$

One may use the theory for U-statistics to compute its variance[1]. A brute force approach is also an option:

$$\begin{aligned}
&var\left(\frac{1}{2n(n-1)}\sum_{i\neq\ell}U_{i\ell}\right)\\
&=\frac{1}{4n^2(n-1)^2}E\sum_{i\neq\ell}\sum\sum_{j\neq k}\sum K_h\left(X_\ell-X_i\right)K_h\left(X_j-X_k\right)\varepsilon_\ell\varepsilon_i\varepsilon_j\varepsilon_k\\
&\times\left(\frac{\omega\left(X_i\right)}{f_X\left(X_i\right)}+\frac{\omega\left(X_\ell\right)}{f_X\left(X_\ell\right)}\right)\left(\frac{\omega\left(X_j\right)}{f_X\left(X_j\right)}+\frac{\omega\left(X_k\right)}{f_X\left(X_k\right)}\right)\\
&=\frac{1}{2n^2(n-1)^2}E\sum_{i\neq\ell}\sum K_h^2\left(X_\ell-X_i\right)\sigma^2\left(X_\ell\right)\sigma^2\left(X_i\right)\left(\frac{\omega\left(X_i\right)}{f_X\left(X_i\right)}+\frac{\omega\left(X_\ell\right)}{f_X\left(X_\ell\right)}\right)^2\\
&=\frac{1}{2n(n-1)}EK_h^2\left(X_\ell-X_i\right)\sigma^2\left(X_\ell\right)\sigma^2\left(X_i\right)\left(\frac{\omega\left(X_i\right)}{f_X\left(X_i\right)}+\frac{\omega\left(X_\ell\right)}{f_X\left(X_\ell\right)}\right)^2\\
&=\frac{1}{2n(n-1)h}E\left[\int\frac{1}{h}K^2\left(\frac{v-X_i}{h}\right)\sigma^2\left(v\right)\sigma^2\left(X_i\right)\left(\frac{\omega\left(X_i\right)}{f_X\left(X_i\right)}+\frac{\omega\left(v\right)}{f_X\left(v\right)}\right)^2 f_X\left(v\right)dv\right]\\
&=O\left(\frac{1}{n^2h}\right)E\left[\int K^2\left(u\right)\sigma^2\left(X_i+uh\right)\sigma^2\left(X_i\right)\left(\frac{\omega\left(X_i\right)}{f_X\left(X_i\right)}+\frac{\omega(X_i+uh)}{f_X\left(X_i+uh\right)}\right)^2 f_X\left(X_i+uh\right)du\right]\\
&=O\left(\frac{1}{n^2h}\right)\left(\int K^2\left(u\right)du\right)E\left[\frac{\sigma^4\left(X_i\right)\omega^2\left(X_i\right)}{f_X^2\left(X_i\right)}\right]\\
&=O\left(\frac{1}{n^2h}\right)=O\left(n^{-9/5}\right)
\end{aligned}$$

uniformly over $c\in[c_L,c_U]$.

Note: If we do not use the LOO estimator, then T_{n2} will contain a term of the form:

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{h}K\left(0\right)\frac{\varepsilon_i^2}{f_X\left(X_i\right)}=O_p\left(\frac{1}{nh}\right).$$

[1]Here is a link to some concise note on U statistic : http://www.math.ucla.edu/~tom/Stat200C/Ustat.pdf. You can also refere to Meng, White and Sun (2014) for an example how the theory of U statistics and processes is used for nonparametric specification testing.

This term will be of the same order of magnitude as the MISE. As a consequence, d_{ASSR}, as an estimator of the out-of-sample mean squared predicted error, is downward biased. The out-of-sample mean squared predicted error is defined to be $E(Y_i^* - \hat{m}(X_i^*))^2$, where (X_i^*, Y_i^*) is a pair of observations not in the estimation sample.

Therefore, $T_{n1} = O_p\left(n^{-9/10}\right)$ and $T_{n2} = O_p\left(n^{-9/10}\right)$. As a result,

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}(X_i) - m(X_i)\right)^2 + O_p\left(n^{-9/10}\right) + \text{cons.}$$

It is not hard to show that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{-i}(X_i) - m(X_i)\right)^2\omega\left(X_i\right) \\
=\ &\frac{1}{n}\sum_{i=1}^{n}\left[h^4B^2(X_i)\omega\left(X_i\right) + \frac{1}{nh}V(X_i)\omega\left(X_i\right)\right] + o_p\left(n^{-4/5}\right) \\
=\ &h^4\int B^2(x)\pi\left(x\right)dx + \frac{1}{nh}\int V(x)\pi\left(x\right)dx + o_p\left(n^{-4/5}\right) \\
=\ &n^{-4/5}\left[c^4\int B(x)\pi\left(x\right)dx + \frac{1}{c}\int V(x)\pi\left(x\right)dx\right] + o_p\left(n^{-4/5}\right).
\end{aligned}$$

So

$$CV(h) = n^{-4/5}\left[c^4\int B(x)\pi\left(x\right)dx + \frac{1}{c}\int V(x)\pi\left(x\right)dx\right] + cons + o_p\left(n^{-4/5}\right)$$

as desired. ■

Remark 2.4.1 *The theorem shows that* $\hat{c}_{cv} = \arg\min CV(h) \to c^*$ *where*

$$c^* = \left(\frac{\int V(x)\pi\left(x\right)dx}{4\int B^2(x)\pi\left(x\right)dx}\right)^{1/5}$$

so $\hat{h}_{cv} = \hat{c}_{cv}n^{-1/5} \to h^* = c^*n^{-1/5}$ *with* h^* *being the optimal bandwidth that minimizes*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{-i}(X_i) - m(X_i)\right)^2\omega\left(X_i\right).$$

Remark 2.4.2 *It follows from the theorem that*

$$\begin{aligned}
&\frac{CV(\hat{h}_{cv}) - n^{-1}\sum_{i=1}^{n}\varepsilon_i^2\omega\left(X_i\right)}{\min_h \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{-i}(X_i) - m(X_i)\right)^2\omega\left(X_i\right)} \\
&= \frac{\min_h \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{-i}(X_i) - m(X_i)\right)^2\omega\left(X_i\right) + O\left(n^{-9/10}\right)}{\min_h \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{-i}(X_i) - m(X_i)\right)^2\omega\left(X_i\right)} \\
&= 1 + O_p\left(n^{-9/10}n^{-4/5}\right) = 1 + O_p\left(n^{-1/10}\right).
\end{aligned}$$

Therefore, $CV(\hat{h}_{cv})-n^{-1}\sum_{i=1}^{n}\varepsilon_i^2\omega(X_i)$ *converges to* $\min_h \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{-i}(X_i)-m(X_i)\right)^2\omega(X_i)$. *Nevertheless, the rate of convergence is quite low.*

Remark 2.4.3 *The conditions in the theorem are not the weakest possible. See Härdle and Marron (1985) for somewhat relaxed assumptions.*

2.4.4 Computation and GCV

For simplicity, we let $\omega(X_i)=1$ in the rest of this chapter.

When calculating cross validation, some care is needed to avoid high computational cost. If one directly computes

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

then n versions of $\hat{m}_{-i}(X_i)$ are required, which leads to an order n^2 algorithm. Fortunately, there now exists a fast order-n algorithm for computation of $CV(h)$ for most common smoothing techniques.

Recall that

$$\hat{m}(X_i)=\frac{\sum_{j=1}^{n}K_h(X_i-X_j)Y_j}{\sum_{j=1}^{n}K_h(X_i-X_j)}:=\sum_{j=1}^{n}w_{nj}(X_i)Y_j$$

with

$$w_{nj}(X_i)=\frac{K_h(X_i-X_j)}{\sum_{\ell=1}^{n}K_h(X_i-X_\ell)}\text{ and }\sum_{j=1}^{n}w_{nj}(X_i)=1.$$

Now

$$\begin{aligned}\hat{m}_{-i}(X_i)&=\frac{\sum_{j\neq i}K_h(X_i-X_j)Y_j}{\sum_{j\neq i}K_h(X_i-X_j)}=\frac{\sum_{j\neq i}w_{nj}(X_i)Y_j}{\sum_{j\neq i}w_{nj}(X_i)}\\&=\frac{\sum_{j\neq i}w_{nj}(X_i)Y_j}{1-w_{ni}(X_i)}=\frac{\hat{m}(X_i)-w_{ni}(X_i)Y_i}{1-w_{ni}(X_i)}.\end{aligned}$$

The above proof is based on a brute force argument and works for the NW estimator only. We now provide a more general proof that is valid for any linear smoother. For each i, consider the new data set

$$D_i:=\{(X_j,Y_j):j\neq i\}\cup(X_i,\hat{m}_{-i}(X_i)):=\left\{X_j,\tilde{Y}_j:j=1,2,...,n\right\}$$

The new data set is the same as the original data set $D=\{(X_j,Y_j):j=1,2,...,n\}$ except that Y_i is replaced by the LOO estimator $\hat{m}_{-i}(X_i)$. Let $\hat{m}_{D_i}(X_j),j=1,2,...,n$ be the fitted value of the linear smoother using the newly constructed data D_i. Suppose that

$$\hat{m}_{D_i}(X_i)=\hat{m}_{-i}(X_i), \tag{2.10}$$

then

$$\begin{aligned}\hat{m}_{-i}(X_i) = \hat{m}_{D_i}(X_i) &= \sum_{j=1}^{n} w_{nj}(X_i)\tilde{Y}_j = \sum_{j\neq i} w_{nj}(X_i)\tilde{Y}_j + w_{ni}(X_i)\tilde{Y}_i \\ &= \sum_{j\neq i} w_{nj}(X_i) Y_j + w_{ni}(X_i)\hat{m}_{-i}(X_i) \\ &= \sum_{j=1}^{n} w_{nj}(X_i) Y_j + w_{ni}(X_i)\hat{m}_{-i}(X_i) - w_{ni}(X_i) Y_i \\ &= \hat{m}(X_i) - w_{ni}(X_i) Y_i + w_{ni}(X_i)\hat{m}_{-i}(X_i)\end{aligned}$$

Therefore

$$\hat{m}_{-i}(X_i) = \frac{\hat{m}(X_i) - w_{ni}(X_i) Y_i}{1 - w_{ni}(X_i)}$$

which gives the exactly the same result as before.

The general proof replies only on the linearity assumption and condition (2.10). The condition holds for most of the linear smoothers. For the NW estimator, we have

$$\begin{aligned}\hat{m}_{-i}(X_i) &= \arg\min_{\theta} \sum_{j\neq i} K_h(X_j - X_i)(Y_j - \theta)^2, \\ \hat{m}_{D_i}(X_i) &= \arg\min_{\theta} \sum_{j\neq i} K_h(X_j - X_i)\left(\tilde{Y}_j - \theta\right)^2 \\ &= \arg\min_{\theta} \sum_{j\neq i} K_h(X_j - X_i)(Y_j - \theta)^2 + K_h(0)(\hat{m}_{-i}(X_i) - \theta)^2.\end{aligned}$$

It is obvious that $\hat{m}_{D_i}(X_i) = \hat{m}_{-i}(X_i)$.

Given the explicit formula relating $\hat{m}_{-i}(X_i)$ to $\hat{m}(X_i)$, we have

$$\begin{aligned}CV(h) &= \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}_{-i}(X_i)]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \frac{\hat{m}(X_i) - w_{ni}(X_i) Y_i}{1 - w_{ni}(X_i)}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{m}(X_i)}{1 - w_{ni}(X_i)}\right]^2.\end{aligned} \tag{2.11}$$

The beauty of the above expression is that $CV(h)$ can be computed using only ordinary residuals.

Remark 2.4.4 *Craven and Wahba (1978) suggest that instead one should minimize*

$$GCV = \frac{n^{-1}\sum_{i=1}^{n}[Y_i - \hat{m}(X_i)]^2}{\left(n^{-1}\sum_{i=1}^{n}[1 - w_{ni}(X_i)]\right)^2}.$$

This is a variant of the cross-validation approach in which the quantities in the denominator of (2.11) are replaced by their sample average. They call it generalized cross-validation, or GCV.

Remark 2.4.5 *In the linear regression case with k regressors, we have* $\hat{Y} = \hat{m}(X) = HY$ *for* $H = X\left(X'X\right)^{-1}X'$. *H is often referred to as the Hat matrix. If the error term is* $iid(0, \sigma^2)$, *then*

$$var(\hat{Y}) = \sigma^2 H$$

and

$$var(\hat{Y}_i) = \sigma^2 H_{ii},\ H_{ii} \text{ is the } (i,i)th \text{ element of } H.$$

More specifically,

$$H_{ii} = x_i\left(X'X\right)^{-1}x_i'$$

where x_i *is the i-th row of X.* H_{ii} *indicates the effect of the given observation on the predictive value* $\hat{Y}$. *For the residual* $e_i = Y_i - \hat{Y}_i$, *we have*

$$var(e_i) = \sigma^2\left(1 - H_{ii}\right).$$

To estimate σ^2, *we can use*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)^2}{(1 - H_{ii})},$$

which is analogous to (2.11). If we follow the idea of Craven and Wahba (1978), we would use

$$\begin{aligned}\hat{\sigma}^2 &= \left(1 - \frac{\sum H_{ii}}{n}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \left(1 - \frac{rank(H)}{n}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \left(1 - \frac{k}{n}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,\end{aligned}$$

which is exactly the recommended formula for small samples.

2.5 Model Selection Perspective

2.5.1 Mallows C_p Criterion

In the previous section, we claim that

$$d_{ASSR} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(X_i)\right]^2$$

is not an appropriate criterion for bandwidth selection. Some adjustment such as that given in (2.11) is needed. Here we derive another adjustment and obtain a criterion called Mallows' C_p criterion.

Let

$$L = L_h = [w_{nj}(X_i)]' = \begin{pmatrix} w_{n1}(X_1) & w_{n2}(X_1) & ... & w_{nn}(X_1) \\ w_{n1}(X_2) & w_{n2}(X_2) & ... & w_{nn}(X_2) \\ ... & ... & ... & ... \\ w_{n1}(X_n) & w_{n2}(X_n) & ... & w_{nn}(X_n) \end{pmatrix},$$

then the fitted value of $\hat{\mathbf{m}} = (\hat{m}(X_1), ..., \hat{m}(X_n))$ is simply

$$\hat{\mathbf{m}} = L_h Y \text{ for } Y = (Y_1, ..., Y_n).$$

This is analogous to the linear regression case where $L = X(X'X)^{-1}X'$. However, L_h is not symmetric and is not a projection matrix while $L = X(X'X)^{-1}X'$ is symmetric and is a projection matrix.

Let

$$d_{ASE} = \frac{1}{n}\|\mathbf{m} - \hat{\mathbf{m}}\|^2.$$

Then

$$\begin{aligned} & d_{ASSR} - d_{ASE} \\ = & \frac{1}{n}\|\mathbf{m} + \boldsymbol{\varepsilon} - \hat{\mathbf{m}}\|^2 - \frac{1}{n}\|\mathbf{m} - \hat{\mathbf{m}}\|^2 \\ = & \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \mathbf{m} - \hat{\mathbf{m}} \rangle \\ = & \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 + \frac{2}{n}\langle \boldsymbol{\varepsilon}, \mathbf{m} - L_h(\mathbf{m} + \boldsymbol{\varepsilon}) \rangle \\ = & \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 + \frac{2}{n}\langle \boldsymbol{\varepsilon}, (I - L_h)\mathbf{m} \rangle - \frac{2}{n}\langle \boldsymbol{\varepsilon}, L_h\boldsymbol{\varepsilon} \rangle. \end{aligned}$$

So

$$d_{ASE} = d_{ASSR} + \frac{2}{n}\langle \boldsymbol{\varepsilon}, L_h\boldsymbol{\varepsilon} \rangle - \frac{2}{n}\langle \boldsymbol{\varepsilon}, (I - L_h)\mathbf{m} \rangle - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2.$$

Assume that $E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma^2 I$, then

$$d_{AMSE}(\hat{m}, m) := E(d_{ASE}) = E(d_{ASSR}) + \underbrace{\frac{2\sigma^2}{n} Etr(L_h)}_{Penalty} - \sigma^2.$$

Note that σ^2 is a constant that does not depend on h, we can define

$$\begin{aligned} C_p &= d_{ASSR} + \frac{2\sigma^2}{n} tr\left(L_h\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(X_i)\right]^2 + \frac{2\sigma^2}{n} tr\left(L_h\right) \end{aligned}$$

which is Mallows' C_p criterion. According to Mallows original paper, he chooses "C" to honor his friend Cuthbert Daniel, an American statistician. The criterion is originally designed for selecting the number of variables in a linear regression where "p" signifies the number of variables in the regression.

Remark 2.5.1 *Obviously*

$$d_{ASE} = C_p + \underbrace{\left[\frac{2}{n}\left\langle \boldsymbol{\varepsilon}, L_h\boldsymbol{\varepsilon}\right\rangle - \frac{2\sigma^2}{n} Etr\left(L_h\right)\right] - \frac{2}{n}\left\langle \boldsymbol{\varepsilon}, \left(I - L_h\right)\mathbf{m}\right\rangle}_{} - \frac{1}{n}\left\|\boldsymbol{\varepsilon}\right\|^2$$

where $-n^{-1}\left\|\boldsymbol{\varepsilon}\right\|^2$ *does not depend on* h*, and under some conditions the middle term can be shown to be equal to* $o_p\left(C_p\right)$*.*

Remark 2.5.2 *In the linear regression case,* $tr\left(L_h\right) = k$*, where* k *is the number of regressors. So*

$$C_p = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(X_i)\right]^2 + \frac{1}{n}\left(2\sigma^2 k\right).$$

For nonparametric regression, we call $tr\left(L_h\right)$ the effective (or equivalent) number of parameters *or* the degrees of freedom of the nonparametric fit.

Remark 2.5.3 *To implement Mallows* C_p*, we need to estimate* σ^2*. It is often recommended to use the largest model to estimate* σ^2*. For kernel smoothing, we choose a very small bandwidth* h^* *to construct* L_{h^*} *and obtain the following estimator of* σ^2 *:*

$$\hat{\sigma}^2 = \frac{nd^*_{ASSR}}{tr\left[\left(I - L_{h^*}\right)'\left(I - L_{h^*}\right)\right]} = \frac{nd^*_{ASSR}}{n - 2tr(L_{h^*}) + tr(L'_{h^*}L_{h^*})}$$

where d^*_{ASSR} *is the ASSR associated with the small bandwidth* h^**. To motivate this formula, we note that*

$$\begin{aligned} E\left(d_{ASSR}\right) &= \frac{1}{n}\left\|\left(I - L_h\right)\mathbf{m}\right\|^2 + \frac{1}{n} tr\left[\left(I - L_h\right)'\left(I - L_h\right) E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \\ &= \frac{1}{n} E\left\|\left(I - L_h\right)\mathbf{m}\right\|^2 + \frac{\sigma^2}{n} Etr\left[\left(I - L_h\right)'\left(I - L_h\right)\right] \end{aligned}$$

When h is small, we can expect that

$$E\left(d_{ASSR}\right)=\frac{\sigma^2}{n}Etr\left[\left(I-L_h\right)'\left(I-L_h\right)\right]+s.o. \tag{2.12}$$

as the bias term $\frac{1}{n}E\left\|\left(I-L_h\right)\mathbf{m}\right\|^2$ is of smaller order.

Remark 2.5.4 *Note that $n-tr\left(L_h\right)$ is not* the degrees of freedom for the residuals. *From the remark above, we can use $n-2tr(L_h)+tr(L_h'L_h)$ as* the degrees of freedom for the residuals. *For linear regressions, $L^2=L$ and so $n-tr\left(L\right)=n-2tr(L)+tr(LL)$. For nonparametric regression, these two quantities are different in general.*

Remark 2.5.5 *When the loss function is based on the log-likelihood, we can derive the Akaike Information Criterion (AIC, Akaike (1974)). This criterion is based on the following approximation:*

$$-2E\log P_{\hat{\theta}}(Y,X)\approx-\frac{2}{n}\sum_{i=1}^{n}\log P_{\hat{\theta}}(Y_i,X_i)+\frac{2}{n}k$$

where k is the number of parameters. In the Gaussian case, we have

$$\begin{aligned}
&-\frac{2}{n}\sum_{i=1}^{n}\log P_{\hat{\theta}}(Y_i,X_i)+\frac{2}{n}k\\
=\ &const-\frac{2}{n}\times\left[-\frac{1}{2}\frac{\sum\left(Y_i-\hat{m}(X_i)\right)^2}{\sigma^2}\right]+\frac{2}{n}k\\
=\ &\frac{1}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{m}(X_i)\right)^2+\frac{2\sigma^2}{n}k\right)+const,
\end{aligned}$$

and so the AIC is asymptotically equivalent to Mallows' C_p criterion.

Remark 2.5.6 *The similarity between GCV and AIC can be seen from the following approximation:*

$$\begin{aligned}
\log\left(GCV\right)&=\log\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\hat{m}(X_i)\right]^2-2\log(1-\frac{1}{n}tr(L_h))\\
&\approx\log\left\{\left[\frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{m}(X_i)\right)^2-\sigma^2}{\sigma^2}+1\right]\sigma^2\right\}+\frac{2}{n}tr(L_h)\\
&\approx\log\left(\sigma^2\right)+\frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{m}(X_i)\right)^2-\sigma^2}{\sigma^2}+\frac{2}{n}tr(L_h)\\
&=\frac{1}{\sigma^2}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{m}(X_i)\right)^2+\frac{2}{n}tr(L_h)\sigma^2\right]+const
\end{aligned}$$

2.5.2 Zero Trace Smoother

The second term in the C_p criterion is $\frac{2\sigma^2}{n} tr(L_h)$. Can we design an estimator with $tr(L_h) = 0$ so that no adjustment is needed? We have two estimators at our disposal: the linear smoother $L_h Y$ and and the naive estimator Y. It is suggested that we may consider a linear combination of these two estimators, yielding

$$L_h^* Y = \alpha L_h Y + (1-\alpha) I_n Y = [\alpha L_h + (1-\alpha) I_n] Y$$

Setting $\text{tr}(L_h^*) = 0$ gives

$$\alpha = \frac{n}{n - tr(L_h)}$$

and

$$L_h^* Y = \frac{n}{n - tr(L_h)} L_h Y - \frac{tr(L_h)}{n - tr(L_h)} Y.$$

L_h^* is a zero trace linear smoother. Now

$$\begin{aligned} ASSR(L_h^*) &= \frac{1}{n} \|Y - L_h^* Y\|^2 = \frac{1}{n} \left\| Y - \frac{n}{n - tr(L_h)} L_h Y + \frac{tr(L_h)}{n - tr(L_h)} Y \right\|^2 \\ &= \frac{1}{n} \left\| \frac{n(Y - L_h Y)}{n - tr(L_h)} \right\|^2 = \frac{n^{-1} \|(Y - L_h Y)\|^2}{\left[1 - \frac{1}{n} tr(L_h)\right]^2} \end{aligned}$$

which is exactly the same as the GCV criterion.

Recall that the LOO estimator is

$$\hat{m}_{-i}(X_i) = \frac{\hat{m}(X_i)}{1 - w_{ni}(X_i)} - \frac{w_{ni}(X_i) Y_i}{1 - w_{ni}(X_i)}.$$

Representing $\hat{m}_{-i}$ in a vector form, we have

$$\hat{\mathbf{m}}_- = L_h^{loo} Y$$

for some matrix L_h^{loo}. Note that

$$\frac{\partial \hat{m}_{-i}(X_i)}{\partial Y_i} = \frac{w_{ni}(X_i)}{1 - w_{ni}(X_i)} - \frac{w_{ni}(X_i)}{1 - w_{ni}(X_i)} = 0$$

and $\text{tr}\left(L_h^{loo}\right) = \sum_{i=1}^n \frac{\partial \hat{m}_{-i}(X_i)}{\partial Y_i} = 0$. So the LOO estimator is also a zero trace linear smoother.

2.5.3 A Theoretical Development of AIC

The AIC can be motivated from the Kullback-Leibler (KL) information. For two probability densities f and g, the KL information $I(g;f)$ is defined as the following

$$I(g;f) = E_Y \log \frac{g(Y)}{f(Y)} = \int g(y) \log \frac{g(y)}{f(y)} dy.$$

In the above expression, Y is a random variable with probability density $g(y)$. Note that $I(g;f) \neq I(f;g)$. Two important properties of the KL information (also called the KL distance) are (i) $I(g;f) \geq 0$ (ii) $I(g,f) = 0$ iff $g(y) = f(y)$ almost everywhere. These two properties can be proved using Jensen's inequality:

$$I(g;f) = E_Y \log \frac{g(Y)}{f(Y)} = -E_Y \log \frac{f(Y)}{g(Y)} \geq -\log E_Y \frac{g(Y)}{f(Y)} = -\log \int f(y) dy = 0$$

using the convexity of $-\log(x)$.

Let $g(y)$ be the true probability density. Given a sequence of candidate models f_j, $j = 1, 2, ..., J$, the best model f_{j^*} in terms of the KL distance satisfies

$$I(g; f_j^*) \leq I(g; f_j) \text{ for all } j = 1, 2, ..., J. \tag{2.13}$$

Note that

$$\begin{aligned} I(g;f) &= \int g(y) \log g(y) dy - \int g(y) \log f(y) dy \\ &= \int g(y) \log g(y) dy - E_Y \log f(Y) \end{aligned}$$

where the first term is a constant. Therefore, the model selection criterion in (2.13) is equivalent to selecting the model with the maximum expected log-likelihood.

Given an iid sample $Y_1, ..., Y_n$ with density $g(\cdot)$, the expected log-likelihood for model $f_j(\cdot|\theta_j)$ with parameter θ_j can be estimated by

$$E_Y \log f_j(Y|\theta_j) \approx \frac{1}{n} \sum_{i=1}^{n} \log f_j(Y_i|\theta_j) := \ell(\theta_j)$$

Therefore, the use of the log-likelihood is motivated. When θ_j is known, we have

$$E_Y \log f_j(Y|\theta_j) = E\ell(\theta_j) = \frac{1}{n} \sum_{i=1}^{n} E \log f_j(Y_i|\theta_j).$$

So $\ell(\theta_j)$ is an unbiased estimator of $E_Y \log f_j(Y|\theta_j)$. However, in practice, θ_j is not known and is often estimated by MLE $\hat{\theta}_j$. The problem is that the above equality does not hold when we plug in $\hat{\theta}_j$, that is

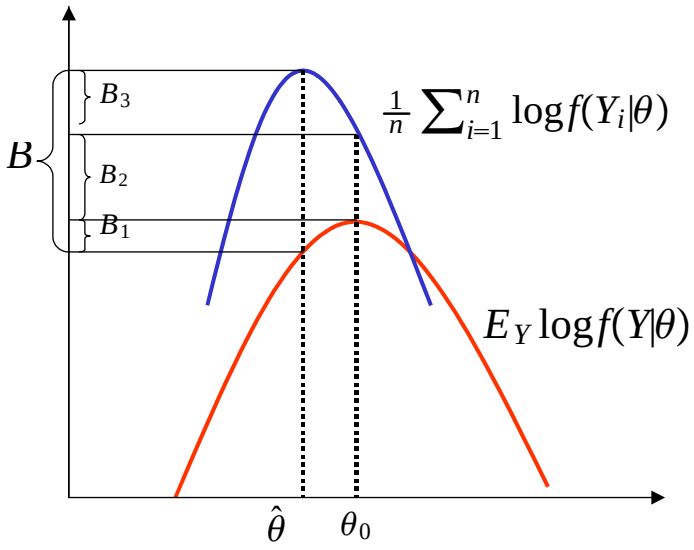
$$E_{Y_1^n} E_Y \log f_j(Y|\hat{\theta}_j) \neq E_{Y_1^n} \frac{1}{n} \sum_{i=1}^{n} \log f_j\left(Y_i|\hat{\theta}_j\right)$$

where $E_{Y_1^n}$ is the expectation operator with respect to the randomness of $(Y_1, ..., Y_n)$ and E_Y is the expectation operator with respect to the randomness of Y conditioning on $(Y_1, ..., Y_n)$. Conditioning is irrelevant here as Y is a random variable not in the sample so that Y and $(Y_1, ..., Y_n)$ are independent. While Y is independent of $\hat{\theta}_j$ on the lhs, Y_i is not independent of $\hat{\theta}_j$ on the rhs as $\{Y_i\}$ is used to estimate θ_j. Because of this, the maximized log-likelihood is a biased estimator of the average expected log-likelihood.

We want to show that the bias is approximately equal to the number of parameters estimated in the model. Since our proof is applicable to all candidate models $f_j(\cdot|\theta_j)$, we drop the subscript 'j' for notational simplicity. We also use E_n to stand for $E_{Y_1^n}$. Let B denote the bias so that

$$\begin{aligned}
B &= E_n\left[E_Y \log f(Y|\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \log f\left(Y_i|\hat{\theta}\right)\right] \\
&= E_n\left[E_Y \log f(Y|\hat{\theta}) - E_Y \log f(Y|\theta_0)\right] \\
&+ E_n\left[E_Y \log f(Y|\theta_0) - \frac{1}{n}\sum_{i=1}^{n} \log f\left(Y_i|\theta_0\right)\right] \\
&+ E_n\left[\frac{1}{n}\sum_{i=1}^{n} \log f\left(Y_i|\theta_0\right) - \frac{1}{n}\sum_{i=1}^{n} \log f\left(Y_i|\hat{\theta}\right)\right] \\
&:= B_1 + B_2 + B_3
\end{aligned}$$

where θ_0 is the pseudo value of $\theta : \theta_0 = \arg\min_\theta E_Y \log f(Y|\theta)$. It is easy to see that $B_2 = 0$. It remains to evaluate B_1 and B_3.



Evaluation of B_1

Using the familiar Taylor expansion, we have

$$E_Y \log f(Y|\hat{\theta}) - E_Y \log f(Y|\theta_0) \approx -\frac{1}{2}\left(\hat{\theta}-\theta_0\right)' H \left(\hat{\theta}-\theta_0\right)$$

where

$$H = -E\left(\frac{\partial^2 \log f(Y|\theta)}{\partial\theta\partial\theta'}\right)|_{\theta=\theta_0}.$$

For the general theory on extreme estimators, we know that

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \to N(0, H^{-1}SH^{-1}),$$

where

$$S = E\frac{\partial \log f(Y|\theta)}{\partial\theta}\left[\frac{\partial \log f(Y|\theta)}{\partial\theta}\right]'|_{\theta=\theta_0}.$$

Therefore

$$B_1 \approx -\frac{1}{2n}Etr\left(HH^{-1}SH^{-1}\right) = -\frac{1}{2n}Etr\left(SH^{-1}\right).$$

Evaluation of B_3

Consider a Taylor expansion of $n^{-1}\sum_{i=1}^{n}\log f\left(Y_i|\theta_0\right)$ around $n^{-1}\sum_{i=1}^{n}\log f\left(Y_i|\hat{\theta}\right)$:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\log f\left(Y_i|\theta_0\right)-\frac{1}{n}\sum_{i=1}^{n}\log f\left(Y_i|\hat{\theta}\right)\\&=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log f\left(Y_i|\hat{\theta}\right)}{\partial\theta}\left(\theta_0-\hat{\theta}\right)+\frac{1}{2}\left(\theta_0-\hat{\theta}\right)'\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f\left(Y_i|\tilde{\theta}\right)}{\partial\theta\partial\theta'}\left(\theta_0-\hat{\theta}\right).\end{aligned}$$

The first term is zero by the definition of $\hat{\theta}$. A ULLN can be used to show that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f\left(Y_i|\tilde{\theta}\right)}{\partial\theta\partial\theta'}\to -H,$$

so the second term can be approximated by

$$\frac{1}{2n}tr(HH^{-1}SH^{-1})=-\frac{1}{2n}tr(SH^{-1}).$$

That is

$$B_3\approx -\frac{1}{2n}tr(SH^{-1}).$$

Combining the above results, we get

$$B\approx -\frac{1}{n}tr\left(SH^{-1}\right).$$

In the special case when the true model belongs to the family of candidate models, it is known that $S=H$ and $tr(SH^{-1})=tr\left(I\right)=k$, the number of estimated parameters. A model selection can be biased on maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\log f_j\left(Y_i|\hat{\theta}_j\right)-\frac{k}{n}=\frac{1}{n}\ell(\hat{\theta})-\frac{k}{n}$$

which directly yields

$$\begin{aligned}AIC&=-2\ell(\hat{\theta})+2k\\&=-2\left(\text{maximized log-likelihood}\right)+2\left(\text{number of estimated parameters}\right)\end{aligned}$$

2.6 A Shrinkage Interpretation of Kernel Smoothing

The smoothing matrix L_h for kernel smoothing is not symmetric but it can be written as the product of two symmetric matrices:

$$L_h = D^{-1}A, \text{ where } A_{i,j} = \frac{1}{nh}K\left(\frac{X_i - X_j}{h}\right),$$
$$D = diag\left(\sum_{j=1}^{n}\frac{1}{nh}K\left(\frac{X_i - X_j}{h}\right)\right).$$

With this representation, we have

$$D^{1/2}\mathbf{m} = D^{1/2}\left(L_h Y\right) = \left(D^{-1/2}AD^{-1/2}\right)\left(D^{1/2}Y\right).$$

If the kernel function is symmetric and positive semidefinite, then A is symmetric and positive semidefinite. By definition, a kernel function is positive semidefinite if

$$\sum_{i=1}^{n}\sum_{j=1}^{n}K(X_i - X_j)X_iX_j \geq 0 \text{ for all } X_i \text{ and } X_j.$$

Under the assumption of positive semidefiniteness, we have

$$D^{-1/2}AD^{-1/2} = \sum_{i=1}^{n}\lambda_i e_i e_i'$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are the ordered eigenvalues of $D^{-1/2}AD^{-1/2}$ with corresponding eigenvectors $e_1, \ldots, e_n$. The eigenvectors $\{e_1, \ldots, e_n\}$ form an orthonormal basis for $\mathbb{R}^n$ with respect to the usual inner product, so for a general dependent vector Y one has the representation:

$$D^{1/2}Y = \sum_{i=1}^{n}\theta_i e_i, \ \theta_i = Y'D^{1/2}e_i$$

or

$$Y = \sum_{i=1}^{n}\theta_i \tilde{e}_i \text{ for } \tilde{e}_i = D^{-1/2}e_i.$$

Here $\{\tilde{e}_1, ..., \tilde{e}_n\}$ form an orthonormal basis for $\mathbb{R}^n$ with respect to the following inner product:

$$(x, y)_D = x'Dy$$

and $(Y, \tilde{e}_i)_D = Y'D\tilde{e}_i = Y'D^{1/2}e_i = \theta_i$.

The fitted values can now be represented as

$$\hat{Y} = D^{-1/2}\left[D^{-1/2}AD^{-1/2}\right]D^{1/2}Y = D^{-1/2}\sum_{i=1}^{n}\lambda_i e_i e_i' \sum_{i=1}^{n}\theta_i e_i$$
$$= D^{-1/2}\sum_{i=1}^{n}\lambda_i\theta_i e_i = \sum_{i=1}^{n}\lambda_i\theta_i\tilde{e}_i.$$

Now that the eigenvalues of $D^{-1/2}AD^{-1/2}$ are the same as those of $D^{-1}A$. But the eigenvalues of $D^{-1}A$ are less or equal to 1. This is because the row sum of $D^{-1}A$ for each row is 1. Therefore $\lambda_i \in [0,1]$. Comparing

$$Y = \sum_{i=1}^{n}\theta_i\tilde{e}_i \text{ with } \hat{Y} = \sum_{i=1}^{n}\lambda_i\theta_i\tilde{e}_i,$$

we can see that $\hat{Y}$ is a shrinkage estimator — the coefficient θ_i is shrunk by a factor λ_i to $\lambda_i\theta$.

2.7 Uniform Consistency

Let

$$\|m\|_p = \left[\int_C |m(x)|^p \, d\mu(x)\right]^{1/p}$$
$$\|m\|_\infty = \sup_{x\in C}|m(x)|.$$

where C is a closed interval, $C \subseteq S$, the support of X. We investigate the conditions under which

$$\|\hat{m} - m\|_p = O_p(\delta_n)$$

for some sequence $\delta_n \to 0$. We shall concentrate on the L_∞ distance, which is usually the most difficult to establish. The uniform consistency results are especially important for the analysis of semiparametric estimators, which involves averages of nonparametric estimates evaluated at a large number of points.

Note that the NW estimation can be written as

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

for weights $\{w_{ni}(x)\}$ that depend only on X_1^n. By the triangle inequality,

$$|\hat{m}(x) - m(x)| \le \left|\sum_{i=1}^{n} w_{n,i}(x)\varepsilon_i\right| + \left|\sum_{i=1}^{n} w_{n,i}(x)[m(X_i) - m(x)]\right|$$

The second term is purely deterministic conditional on X_1^n. Using a Taylor series expansion, this term can be shown to be $O\left(h^2\right)$ uniformly. We can establish

$$\sup_{x\in C}\left|\sum_{i=1}^{n} w_{n,i}\left(x\right)\varepsilon_i\right| = O_p\left(\sqrt{\frac{\log n}{nh}}\right) := O_p\left(\delta_n\right)$$

using the same proof for the uniform consistency of the kernel density estimator. See Chapter 1 or Li and Racine (p. 78).

2.8 Uniform Confidence Intervals (Optional)

In the previous section, we establish that, under suitable conditions,

$$\left\|\hat{m}-m\right\|_\infty = O_p\left(\sqrt{\frac{\log n}{nh}}\right) + O_p\left(h^2\right).$$

We now seek to refine this result into a limiting distribution. Let

$$T_n = \sup_{x\in C} T_n\left(x\right); T_n\left(x\right) = \frac{\hat{m}\left(x\right)-m\left(x\right)}{\sqrt{var\left[\hat{m}\left(x\right)\right]}}$$

where C is some compact set contained in the support of X and $\text{var}(\hat{m}(x))$ is the asymptotic variance or conditional variance. We know that with undersmoothing $T_n\left(x\right)$ is asymptotically standard normal for each x, but that $T_n = \sup_{x\in C} T_n\left(x\right) = O_p\left(\sqrt{\log n}\right)$. It can be shown that there exist increasing sequences a_n and b_n such that

$$P\left(a_n\left(T_n - b_n\right) \leq t\right) \to \exp\left[-\left(2\exp(-t)\right)\right] \tag{2.14}$$

i.e., T_n is asymptotically extreme value.

The main use of (2.14) is simultaneous confidence intervals. The confidence intervals we have provided

$$\left[\hat{m}\left(x\right) - z_{\alpha/2}\sqrt{var\left[\hat{m}\left(x\right)\right]}, \hat{m}\left(x\right) + z_{\alpha/2}\sqrt{var\left[\hat{m}\left(x\right)\right]}\right]$$

is valid for a single point. However, we are usually interest in the function m at a (possibly infinite) number of different points, in which case simply plotting out the above interval for each x will not give the right level.

There are two main approaches to providing correct confidence intervals. One is to use Bonferroni type inequality to correct the level For this method, the simultaneous confidence intervals at points $x_1, ..., x_N$ is given by

$$I_i = \left[\hat{m}\left(x_i\right) - z_{\alpha/(2N)}\sqrt{var\left[\hat{m}\left(x_i\right)\right]}, \hat{m}\left(x_i\right) + z_{\alpha/(2N)}\sqrt{var\left[\hat{m}\left(x_i\right)\right]}\right]$$

By the Bonferroni inequality,

$$\begin{aligned} P(\cap_{i=1}^{N} m(x_i) \in I_i) &= 1 - P(\cup_{i=1}^{N} m(x_i) \notin I_i) \\ &\geq 1 - \sum_{i=1}^{N} P(m(x_i) \notin I_i) \\ &= 1 - N\left(\frac{\alpha}{N}\right) = 1 - \alpha. \end{aligned}$$

The confidence band based on this approach is often too wide. In fact, the band grows with N.

The second is to treat function $\hat{m}$ as a random element and use stochastic process limit theory. In other words, we find a set of functions $C(\hat{m})$ with the property that

$$P[m \in C(\hat{m})] \to 1 - \alpha$$

for large n. This is provided by the limit theory in (2.14) by letting

$$C(\hat{m}) = \{m : a_n(T_n - b_n) \leq c_\alpha\}$$

where c_α solves $\exp[-(2\exp(-t))] = 1 - \alpha$. This leads to the bands of the form

$$I(x) = \left[\hat{m}(x) - \left(b_n + \frac{c_\alpha}{a_n}\right)\sqrt{\widehat{var}(\hat{m}(x))}, \hat{m}(x) + \left(b_n + \frac{c_\alpha}{a_n}\right)\sqrt{\widehat{var}(\hat{m}(x))}\right]$$

where

$$a_n = \sqrt{-2\log h}, b_n = \sqrt{-2\log h} + \frac{\log\left(\lambda/\left[4\pi^2\right]\right)}{\sqrt{-2\log h}},$$

$$\lambda = \left\|K'\right\|^2 / \|K\|^2,$$

and $\widehat{var}(\hat{m}(x))$ is some estimate of $var(\hat{m}(x))$. These intervals has the correct coverage in large samples, i.e.

$$\lim_{n\to\infty} P[m(x) \in I(x), \forall x] = 1 - \alpha.$$

In their Handbook of Econometrics Chapter, Härdle and Linton (1994) discussed the above confidence intervals. In practice, however, these intervals do not work terribly well for the reasons discussed in Hall (1993). A better approach is based on bootstrap, see Härdle and Marron (1991).

2.9 Time Series Case

Consider time series data (X_t, Y_t) that are generated from a strictly stationary process. As in the iid case, the conditional mean estimator is:

$$\hat{m}(x) = \sum_{t=1}^{n} w_{nt}(x) Y_t,$$

where

$$w_{nt}(x) = K_h(x - X_t) / \sum_{t=1}^{T} K_h(x - X_t).$$

The asymptotic bias of $\hat{m}(x)$ is the same as the iid case. Here we focus on the asymptotic variance of $\sqrt{nh}[\hat{m}(x) - m(x)]$. That is, we compute the asymptotic variance of

$$\sqrt{Th}\sum_{t=1}^{n} w_{nt}(x)\varepsilon_t = \left[\frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t)\right]^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{h}K_h(x - X_t)\varepsilon_t\right).$$

We have shown that $T^{-1}\sum_{t=1}^{T} K_h(x - X_t)$ converges to $f_X(x)$. So it remains to compute the asymptotic variance of

$$Q_T = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \text{ for } Z_t = \sqrt{h}K_h(x - X_t)\varepsilon_t.$$

By strict stationarity, we have

$$var(Q_T) = var(Z_1) + 2\sum_{\ell=1}^{T-1}\left(1 - \frac{\ell}{T}\right)cov(Z_1, Z_{\ell+1}).$$

Here $var(Z_1) \asymp 1$. As in the kernel density estimation, we let d_T be a constant such that $d_T \to \infty$ and $d_T h \to 0$. Define

$$J_1 = \sum_{\ell=1}^{d_T-1} |cov(Z_1, Z_{\ell+1})|, \; J_2 = \sum_{\ell=d_T}^{T-1} |cov(Z_1, Z_{\ell+1})|.$$

Now, assuming that

(i) $X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}$ has a continuous density $f_{X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}}(w_1, w_2, w_3, w_4)$ wrt to the Lesbegue measure,

(ii) $\int |K(v)| dv < \infty$ and $\int\int |w_3 w_4| f_{X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}}(x, x, w_3, w_4) dw_3 dw_4 < \infty$,

we have

$$\begin{aligned}
& |cov(Z_1, Z_{\ell+1})| \\
= \; & h \left| E K_h \left(x - X_1 \right) K_h \left(x - X_{\ell+1} \right) \varepsilon_1 \varepsilon_{\ell+1} \right| \\
\leq \; & h \int\int\int\int \frac{1}{h} K \left(\frac{x - w_1}{h} \right) \frac{1}{h} K \left(\frac{x - w_2}{h} \right) |w_3 w_4| \\
& \times f_{X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}}(w_1, w_2, w_3, w_4) dw_1 dw_2 dw_3 dw_4 \\
= \; & h \int\int\int\int K(v_1) K(v_2) |w_3 w_4| f_{X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}}(x - h v_1, x - h v_2, w_3, w_4) dv_1 dv_2 dw_3 dw_4 \\
= \; & O\left(h \int\int\int\int K(v_1) K(v_2) |w_3 w_4| f_{X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}}(x, x, w_3, w_4) dv_1 dv_2 dw_3 dw_4 \right) \\
= \; & O\left(h \int\int |w_3 w_4| f_{X_1, X_{\ell+1}, \varepsilon_1, \varepsilon_{\ell+1}}(x, x, w_3, w_4) dw_3 dw_4 \right) = O(h).
\end{aligned}$$

So

$$J_1 = O\left(d_T h \right) = o\left(1 \right).$$

If $\{Z_t\}$ is a ρ mixing process satisfying $\sum_{\ell=1}^{\infty} |\rho(\ell)| < \infty$ uniformly over h, where

$$\rho(\ell) = \sup_{A \in L_2(\mathcal{F}_t), B \in L_2(\mathcal{F}_{t+\ell})} corr(A, B)$$

then

$$J_2 \leq var(Z_1) \sum_{\ell=d_T}^{\infty} |\rho(\ell)| = o\left[var(Z_1) \right].$$

Combining the above two results, we obtain

$$2 \sum_{\ell=1}^{T-1} \left(1 - \frac{\ell}{T} \right) cov(Z_1, Z_{\ell+1}) = o\left(var\left(Z_1 \right) \right).$$

We have therefore shown that

$$var(Q_T) = var\left(Z_1 \right) \left(1 + o\left(1 \right) \right).$$

That is, the time series dependence can be ignored in the asymptotic variance calculation. However, the same comments on KDE in the previous chapter apply here.

In the above arguments, we have assumed that $\{Z_t\}$ is a ρ mixing process, which is a high level assumption. We can also impose some mixing conditions on $\{X_t, \varepsilon_t\}$, say, α-mixing as in the previous chapter and show that $var(Q_T) = var\left(Z_1 \right) \left(1 + o\left(1 \right) \right)$.

2.10 Asymptotic Properties of Local Linear Estimator

The local polynomial estimator of $m(x)$ is defined as $\hat{\theta}_0(x)$, where

$$\left(\hat{\theta}_0, ..., \hat{\theta}_p\right) = \arg\min \sum_{i=1}^{n} K_h(X_i - x) \left\{ Y_i - \theta_0 - \theta_1 \left(\frac{X_i - x}{h}\right) - ... - \theta_p \frac{1}{p!} \left(\frac{X_i - x}{h}\right)^p \right\}^2. \tag{2.15}$$

Let

$$Z_{ix} = \begin{pmatrix} 1 \\ \frac{X_i - x}{h} \\ ... \\ \frac{1}{p!}\left(\frac{X_i - x}{h}\right)^p \end{pmatrix}, Z_x = \begin{pmatrix} Z_{1x}' \\ ... \\ Z_{nx}' \end{pmatrix}$$

and W_x be the $n \times n$ diagonal matrix whose (i,i)-th component is $K_h(X_i - x)$. We can write (2.15) as

$$\hat{\theta} = \arg\min_{\theta} (Y - Z_x\theta)' W_x (Y - Z_x\theta)$$

with $\hat{\theta} = \left(\hat{\theta}_0, ..., \hat{\theta}_p\right)'$. Minimizing the above objective function gives the weighted least squares estimator:

$$\hat{\theta} = \left(Z_x' W_x Z_x\right)^{-1} \left(Z_x' W_x Y\right).$$

In particular, $\hat{\theta}_0$ is the inner product of the first row of $\left(Z_x' W_x Z_x\right)^{-1} Z_x' W_x$ with Y. More specifically,

$$\hat{\theta}_0 = \underbrace{e_1'}_{1\times(p+1)} \underbrace{\left(Z_x' W_x Z_x\right)^{-1}}_{(p+1)\times(p+1)} \underbrace{Z_x' W_x}_{(p+1)\times n} Y := \sum_{i=1}^{n} w_{ni}(x) Y_i$$

where $e_1 = (1, 0, ..., 0)'$ and $w_{ni}(x)$ is the i-th element of the row vector

$$e_1' \left(Z_x' W_x Z_x\right)^{-1} \left(Z_x' W_x\right)$$

A more explicit expression for $w_{ni}(x)$ is

$$w_{ni}(x) = e_1' \left(Z_x' W_x Z_x\right)^{-1} \left[Z_x' W_x\right]_i = \left[e_1' \left(Z_x' W_x Z_x\right)^{-1} Z_{ix}\right] K_h(X_i - x),$$

where $\left[Z_x' W_x\right]_i$ is the i-th column of $Z_x' W_x$. That is $\left[Z_x' W_x\right]_i = Z_{ix} K_h(X_i - x)$. Therefore, the local polynomial estimator $\hat{\theta}_0$ is a linear estimator. The weight $w_{ni}(x)$ combines the weighting kernel K and the least squares operations, and is sometimes referred to as *the equivalent kernel.*

Note that $e_1'\left(Z_x'W_xZ_x\right)^{-1}Z_{ix}$ is a scalar that depends on x and the design points $\{X_i\}_{i=1}^n$, we can write

$$\begin{aligned} w_{ni}(x) &= C(X_1,\ldots,X_n,x)K_h(X_i-x) \\ &= \sum_{\ell=0}^{p} c_\ell^* \frac{1}{\ell!}\left(\frac{X_i-x}{h}\right)^\ell K_h(X_i-x) \end{aligned}$$

for some c_ℓ^* depending on the design points $\{X_i\}_{i=1}^n$ and x. More specifically, $\left(c_0^*, c_1^*, ..., c_p^*\right) = e_1'\left(Z_x'W_xZ_x\right)^{-1}$, the first row of $\left(Z_x'W_xZ_x\right)^{-1}$.

In addition,

$$\sum_{i=1}^n w_{ni}(x) = 1 \text{ and } \sum_{i=1}^n w_{ni}(x)\left(\frac{X_i-x}{h}\right)^j = 0 \text{ for } j = 1,2,...,p \tag{2.16}$$

This is because $e_1'\left(Z_x'W_xZ_x\right)^{-1}Z_x'W_xZ_x = (1,0,....0)$. So the equivalent kernel depends on the design points and x. This explains why local polynomial fit can adapt automatically to various designs and to the boundary problem.

Local polynomial regression can be regarded as a pseudo-regression based on

$$Y_i = \theta_0 + \theta_1\left(\frac{X_i-x}{h}\right) + \ldots + \theta_p\frac{1}{p!}\left(\frac{X_i-x}{h}\right)^p + \tilde{\varepsilon}_i.$$

where $\tilde{\varepsilon}_i = error_i + \varepsilon_i$ and $error_i$ contains the error in approximating $m(X_i)$ by the polynomial $\theta_0 + \theta_1(X_i-x)/h + \ldots + \theta_p\left((X_i-x)/h\right)^p/p!$. We note that

$$\begin{aligned} \sum_{i=1}^n w_{ni}(x)Y_i &= \left(\sum_{i=1}^n w_{ni}(x)\right)\theta_0 + \left[\sum_{i=1}^n w_{ni}(x)\left(\frac{X_i-x}{h}\right)\right]\theta_1 \\ &\quad + \ldots + \left[\sum_{i=1}^n w_{ni}(x)\frac{1}{p!}\left(\frac{X_i-x}{h}\right)^p\right]\theta_p + \sum_{i=1}^n w_{ni}(x)\tilde{\varepsilon}_i. \end{aligned}$$

For $\sum_{i=1}^n w_{ni}(x)Y_i$ to be a good estimator of θ_0, we require that the weights $\{w_{ni}(x)\}$ satisfy (2.16), in which case

$$\theta_0 = \sum_{i=1}^n w_{ni}(x)Y_i - \sum_{i=1}^n w_{ni}(x)\tilde{\varepsilon}_i.$$

Local polynomial regression can be regarded as a way to construct the weights that satisfy (2.16).

In the next two subsections, we investigate the asymptotic bias and variance of the local polynomial estimator. For notational simplicity, we focus on the local linear case. The results can be easily extended to the general local polynomial case.

2.10.1 Asymptotic Variance of Local Linear Estimators

Suppose x is an interior point. It is easy to see that

$$asymvar\left(\sqrt{nh}\hat{\theta}_0\right) = \text{plime}_1'\left(\frac{Z_x'W_xZ_x}{n}\right)^{-1}\left(\frac{Z_x'W_x\Sigma W_xZ_x}{nh^{-1}}\right)\left(\frac{Z_x'W_xZ_x}{n}\right)^{-1}e_1,$$

where

$$\begin{aligned}\frac{Z_x'W_xZ_x}{n} &= \begin{pmatrix} \frac{1}{n}\sum\left[K_h\left(X_i-x\right)\right] & \frac{1}{n}\sum_{i=1}^n\left[K_h\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right) \\ \frac{1}{n}\sum_{i=1}^n\left[K_h\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right) & \frac{1}{n}\sum_{i=1}^n\left[K_h\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)^2 \end{pmatrix} \\ &:= \begin{pmatrix} S_{n,0}\left(x\right) & S_{n,1}\left(x\right) \\ S_{n,1}\left(x\right) & S_{n,2}\left(x\right) \end{pmatrix} = S_n^{2\times 2}\left(x\right)\end{aligned}$$

and

$$\begin{aligned}&\frac{Z_x'W_x\Sigma W_xZ_x}{nh^{-1}} \\ &= \begin{pmatrix} \frac{h}{n}\sum_{i=1}^n\left[K_h^2\left(X_i-x\right)\right]\sigma^2\left(X_i\right) & \frac{h}{n}\sum_{i=1}^n\left[K_h^2\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)\sigma^2\left(X_i\right) \\ \frac{h}{n}\sum_{i=1}^n\left[K_h^2\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)\sigma^2\left(X_i\right) & \frac{h}{n}\sum_{i=1}^n\left[K_h^2\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)^2\sigma^2\left(X_i\right) \end{pmatrix} \\ &:= \begin{pmatrix} \Gamma_{n,0}\left(x\right) & \Gamma_{n,1}\left(x\right) \\ \Gamma_{n,1}\left(x\right) & \Gamma_{n,2}\left(x\right) \end{pmatrix} := \Gamma_n^{2\times 2}\left(x\right)\end{aligned}$$

Note that[2]

$$\begin{aligned}S_{n,j}\left(x\right) &= \frac{1}{n}\sum_{i=1}^n\left[K_h\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)^j \\ &\to^p E\left[K_h\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)^j = \int\frac{1}{h}K\left(\frac{v-x}{h}\right)\left(\frac{v-x}{h}\right)^j f(v)dv \\ &= \int K(u)u^j f_X(x+uh)du = \mu_j f_X(x)+o\left(1\right),\end{aligned}$$

[2] $S_{n,j}\left(x\right)$ can be regarded as a density estimator with $K(u)u^j$ as the kernel function.

where $\mu_j = \int K(u)u^j du$, and[3]

$$\begin{aligned}
\Gamma_{n,j}(x) &= \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{X_i - x}{h}\right)\right]^2\left(\frac{X_i - x}{h}\right)^j \sigma^2(X_i) \\
&\to^p hE\left[K_h^2(X_i - x)\right]\left(\frac{X_i - x}{h}\right)^j \sigma^2(X_i) \\
&= \frac{1}{h}\int K^2\left(\frac{v - x}{h}\right)\left(\frac{v - x}{h}\right)^j \sigma^2(v) f(v) dv \\
&= \int K^2(u) u^j \sigma^2(x + uh) f(x + uh) du \\
&= \sigma^2(x) f(x) \nu_j + o(1),
\end{aligned}$$

where $\nu_j = \int K^2(u)u^j du$. Combining the above results yields:

$$\begin{aligned}
asymvar\left(\sqrt{nh}\hat{\theta}_0\right) &= e_1'\begin{pmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{pmatrix}^{-1}\begin{pmatrix} \Gamma_{n,0}(x) & \Gamma_{n,1}(x) \\ \Gamma_{n,1}(x) & \Gamma_{n,2}(x) \end{pmatrix}\begin{pmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{pmatrix}^{-1} e_1 \\
&\to e_1'\begin{pmatrix} f_X(x) & 0 \\ 0 & \mu_2 f_X(x) \end{pmatrix}^{-1}\begin{pmatrix} \sigma^2(x) f_X(x)\nu_0 & \sigma^2(x) f_X(x)\nu_1 \\ \sigma^2(x) f_X(x)\nu_1 & \sigma^2(x) f_X(x)\nu_2 \end{pmatrix} \\
&\times \begin{pmatrix} f_X(x) & 0 \\ 0 & \mu_2 f_X(x) \end{pmatrix}^{-1} e_1 = \frac{\sigma^2(x)}{f(x)}\int K^2(u).
\end{aligned}$$

So the asymptotic variance of $\hat{\theta}_0$ is the same as the NW estimator.

2.10.2 Asymptotic Bias of Local Linear Estimators

Let

$$\mathbf{m} = \begin{pmatrix} m(X_1) \\ ... \\ m(X_n) \end{pmatrix}.$$

In view of

$$m(X_i) = m(x) + m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2 + \frac{1}{6}m'''(\xi_i)(X_i - x)^3$$

[3] $\Gamma_{n,j}(x)$ can be regarded as a smoothing of $\sigma^2(x)$ with kernel $K^2(u)u^j$

where ξ_i is between x and X_i, we have

$$\mathbf{m} = Z_x \begin{pmatrix} m(x) \\ hm'(x) \end{pmatrix} + \mathcal{R}$$

where $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2$ and

$$\mathcal{R}_1 = \begin{pmatrix} \frac{1}{2}m''(x)(X_1 - x)^2 \\ ... \\ \frac{1}{2}m''(x)(X_n - x)^2 \end{pmatrix}, \ \mathcal{R}_2 = \begin{pmatrix} \frac{1}{6}m'''(\xi_1)(X_1 - x)^3 \\ ... \\ \frac{1}{6}m'''(\xi_n)(X_n - x)^3 \end{pmatrix}.$$

Now

$$E\left(Z_x' W_x Y\right) = Z_x' W_x Z_x \begin{pmatrix} m(x) \\ hm'(x) \end{pmatrix} + Z_x' W_x \mathcal{R}.$$

The dominating asymptotic bias of $\hat{\theta}_0$ is $h^2 B$ where

$$\begin{aligned} B =& \operatorname{p\,lim} \frac{1}{h^2} \left\{ e_1' \left(Z_x' W_x Z_x\right)^{-1} \left[Z_x' W_x Z_x \begin{pmatrix} m(x) \\ hm'(x) \end{pmatrix} + Z_x' W_x \mathcal{R} \right] - m(x) \right\} \\ =& \operatorname{p\,lim} \frac{1}{h^2} e_1' \left(\frac{Z_x' W_x Z_x}{n} \right)^{-1} \frac{Z_x' W_x \mathcal{R}_1}{n} + \operatorname{p\,lim} \frac{1}{h^2} e_1' \left(\frac{Z_x' W_x Z_x}{n} \right)^{-1} \frac{Z_x' W_x \mathcal{R}_2}{h} \\ =& \left(f_X^{-1}(x),\ 0 \right) \operatorname{p\,lim} \frac{1}{nh^2} Z_x' W_x \mathcal{R}_1 + \left(f_X^{-1}(x),\ 0 \right) \operatorname{p\,lim} \frac{1}{nh^2} Z_x' W_x \mathcal{R}_2. \end{aligned}$$

Under some smoothness condition, we can show that the second term is of smaller order than the first term. We now focus on the first term, starting with

$$\frac{1}{nh^2} Z_x' W_x \mathcal{R}_1 = \begin{pmatrix} \frac{m''(x)}{2nh^2} \sum_{i=1}^n \left[K_h(X_i - x)\right](X_i - x)^2 \\ \frac{m''(x)}{2nh^2} \sum_{i=1}^n \left[K_h(X_i - x)\right](X_i - x)^3 \end{pmatrix}.$$

But

$$\begin{aligned} & \frac{m''(x)}{2nh^2} \sum_{i=1}^n \left[K_h(X_i - x)\right](X_i - x)^2 \\ & \to^p \frac{m''(x)}{2h^2} E\left[K_h(X_i - x)\right](X_i - x)^2 = \frac{m''(x)}{2h^2} \int K_h(v - x)(v - x)^2 f(v) dv \\ & = \frac{m''(x)}{2} \int K\left(\frac{v - x}{h}\right) \left(\frac{v - x}{h}\right)^2 f(v) d\frac{v}{h} = \frac{m''(x) f_X(x)}{2} \int K(u) u^2 du + o(1). \end{aligned}$$

So

$$\begin{aligned} asymbias\left(\hat{\theta}_0\right) &= h^2 \frac{1}{f_X(x)} \frac{m''(x) f_X(x)}{2} \int K(u) u^2 du \left(1 + o(1)\right) \\ &= h^2 \frac{m''(x)}{2} \int K(u) u^2 du \left(1 + o(1)\right). \end{aligned}$$

Alternatively, it follows from $\theta_0 = \sum_{i=1}^n w_{ni}(x) Y_i - \sum_{i=1}^n w_{ni}(x) \tilde{\varepsilon}_i$ and $\hat{\theta}_0 = \sum_{i=1}^n w_{ni}(x) Y_i$ that

$$\hat{\theta}_0 - \theta_0 = \sum_{i=1}^n w_{ni}(x) \frac{1}{2} m''(x) (X_i - x)^2 + \sum_{i=1}^n w_{ni}(x) \varepsilon_i + s.o.$$

The dominating asymptotic bias of $\hat{\theta}_0 - \theta_0$ is

$$\frac{1}{2} m''(x) h^2 \mathrm{p}\lim \sum_{i=1}^n w_{ni}(x) \left(\frac{X_i - x}{h}\right)^2,$$

which can be shown to be equal to $h^2 \frac{m''(x)}{2} \int K(u) u^2 du$.

Recall that the asymptotic bias of the NW estimator is

$$\frac{\mu_2}{2} \left[\frac{m''(x) f_X(x) + 2m'(x) f_X'(x)}{f_X(x)}\right] h^2.$$

Thus, the local linear estimator is free from design bias. The above bias derivation holds regardless of whether x is on the boundary or not. At the boundary points, the NW kernel estimator has asymptotic bias of order $O(h)$ while the local linear estimator has bias of order $O\left(h^2\right)$. In this sense, the local linear estimation eliminates the boundary bias.

2.10.3 An odd World

The asymptotic variance of the local linear estimator is the same as that of the local constant estimator. Similarly, under the assumption that $K(\cdot)$ is symmetric, the asymptotic variance of the local cubic estimator is the same as that of the local quadratic estimator. To see this,

we note that for the local quadratic estimator, the 'S' matrix satisfies:

$$\begin{aligned} e_{1,3}' \left[S_n^{3\times 3}(x)\right]^{-1} &= e_{1,3}' \begin{pmatrix} S_{n,0}(x) & S_{n,1}(x) & S_{n,2}(x) \\ S_{n,1}(x) & S_{n,2}(x) & S_{n,3}(x) \\ S_{n,2}(x) & S_{n,3}(x) & S_{n,4}(x) \end{pmatrix}^{-1} \\ &\to {}^{p} e_{1,3}' \begin{pmatrix} S_0(x) & 0 & S_2(x) \\ 0 & S_2(x) & 0 \\ S_2(x) & 0 & S_4(x) \end{pmatrix}^{-1} \\ &= \left[\frac{S_4(x)}{S_0(x)S_4(x) - S_2^2(x)}, 0, -\frac{S_2^2(x)}{S_0(x)S_4(x) - S_2^2(x)}\right] := \ell_3(x)' \end{aligned}$$

where the last line follows from simple calculations. The asymptotic variance is

$$\ell_3(x)' \Gamma^{3\times 3} \ell_3(x).$$

For the local cubic estimator, the 'S' matrix satisfies:

$$\begin{aligned} e_{1,4}' S_{n,4}(x) &\to e_{1,4}' \begin{pmatrix} S_{n,0}(x) & S_{n,1}(x) & S_{n,2}(x) & S_{n,3}(x) \\ S_{n,1}(x) & S_{n,2}(x) & S_{n,3}(x) & S_{n,4}(x) \\ S_{n,2}(x) & S_{n,3}(x) & S_{n,4}(x) & S_{n,5}(x) \\ S_{n,3}(x) & S_{n,4}(x) & S_{n,5}(x) & S_{n,6}(x) \end{pmatrix}^{-1} \\ &\to e_{1,4}' \begin{pmatrix} S_0(x) & S_1(x) & S_2(x) & 0 \\ S_1(x) & S_2(x) & S_3(x) & S_4(x) \\ S_2(x) & S_3(x) & S_4(x) & 0 \\ 0 & S_4(x) & 0 & S_6(x) \end{pmatrix}^{-1} \\ &= e_{1,3}' \begin{pmatrix} \left[S_{1\cdot 2}^{33}\right]^{-1}, & -\left[S_{1\cdot 2}^{33}\right]^{-1} S_6^{-1}(x) \begin{pmatrix} 0 \\ S_4(x) \\ 0 \end{pmatrix} \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} S_{1\cdot 2}^{33} &= \left[S^{33}(x) - S_6^{-1}(x) \begin{pmatrix} 0 \\ S_4(x) \\ 0 \end{pmatrix} \begin{pmatrix} 0 & S_4(x) & 0 \end{pmatrix} \right] \\ &= \begin{pmatrix} S_0(x) & 0 & S_2(x) \\ 0 & S_2(x) - \frac{S_4^2(x)}{S_6(x)} & 0 \\ S_2(x) & 0 & S_4(x) \end{pmatrix} \end{aligned}$$

and so

$$\left(S_{1\cdot 2}^{33}\right)^{-1} = \begin{pmatrix} \frac{S_0(x)S_4(x)}{S_0(x)S_4(x)-S_2(x)S_2(x)} & 0 & -\frac{S_2^2(x)}{S_0(x)S_4(x)-S_2^2(x)} \\ \# & \# & \# \\ \# & \# & \# \end{pmatrix} = \begin{pmatrix} \ell_3(x)' \\ \# \end{pmatrix}$$

where "# " stands for numbers that we do not care about. Therefore

$$\begin{aligned} e_{1,4}' S_{n,4}(x) &\to \left[\ell_3'(x), -\ell_3'(x) S_6^{-1}(x) \begin{pmatrix} 0 \\ S_4(x) \\ 0 \end{pmatrix} \right] \\ &= \left[\ell_3'(x), 0 \right]. \end{aligned}$$

The asymptotic variance is

$$\begin{aligned} &\left[\ell_3'(x), 0 \right] \begin{pmatrix} \Gamma^{3\times 3} & c \\ c' & d \end{pmatrix} \left[\ell_3'(x), 0 \right]' \\ &= \ell_3'(x) \Gamma^{3\times 3} \ell_3(x), \end{aligned}$$

which is exactly the same as the asymptotic variance of the local quadratic estimator. However, the local cubic estimator has the advantage of removing one more bias term than the local quadratic estimator and hence is preferred in large samples.

Our asymptotic result shows that for estimating the function $m(x)$, a polynomial of odd degree $(2k+1)$ is preferred over the polynomial of even degree $(2k)$. Asymptotically, it is an odd world. However, I do not know whether this is a wise advice in finite samples. It is worthwhile examining this by Monte Carlo simulations. In empirical applications, it seems safe to use local linear regression instead of local constant regression.

2.11 Application: Regression Discontinuity

2.11.1 Overview

A typical application of local linear regressions is the regression discontinuity (RD) or regression discontinuity design (RDD). The idea can be best illustrated using the "sharp" RD design. In this case, an individual receives the treatment if and only if the running variable (also called the forcing variable) is larger than a threshold. Under some assumptions, the individuals lying closely on the two sides of the threshold can be regarded as statistically identical. As a result, we can identify and estimate the average treatment effect even though we do no have a controlled randomized experiment. The RDD has become increasingly popular in economics, political science, epidemiology, and related disciplines in recent years.

The RDD can be traced back to Thistlewaite and Campbell (1960) which aims at measuring the causal effect on career aspirations of a student receiving a scholarship. There was a precise test score cutoff, above which all students received a particular scholarship and below which none did. Call this score cutoff c, say $c = 100$. The idea is that we don't think that students whose test score was $X = 99$ are (significantly) different from students who scored $X = 101$. That is, we believe the two point difference is mostly random noise—one student got less sleep and scored 99, another ate a good breakfast and got 101, etc., not reflecting any underlying, persistent differences among the students. Consequently, we have a quasi-experimental design, where we can compare the outcomes of the treatment group ($X = 101$) to the outcomes of the control group ($X = 99$).

We will formalize the model following Hahn, Todd, and Van Der Klaauw (2001, HTV hereafter). See Imbens and Lemieux (2008) and Lee and Lemieux (2010) for recent surveys and practical guidance. In the "sharp" design, where all units above the cutoff are treated and all below are untreated (as in the scholarship example). In the "fuzzy" design, the probability of treatment for individuals above the cutoff is higher than the probability of treatment for individuals below the cutoff. Sometimes a governmental/institutional policy seemingly leads to a sharp design with a clear cutoff, but in practice the enforcement is fuzzy.

Let $D_i \in \{0, 1\}$ be a binary treatment variable, and let Y_i be the outcome of interest. In the "potential outcomes" treatment effect framework, let $Y_i(1)$ denote the outcome individual i would have if treated (e.g., what would i's current wage be if he had received a scholarship before college?), and let $Y_i(0)$ denote the outcome that the same individual i would have if not treated (e.g., what would i's current wage be if they had not received a scholarship before college?). In the real world, individual i either got a scholarship or did not get a scholarship—these are mutually exclusive states, so we only observe $Y_i(0)$ or $Y_i(1)$, but never both. The *observed* Y_i is

$$\begin{aligned} Y_i &= (1 - D_i) Y_i(0) + D_i Y_i(1) \\ &= Y_i(0) + X_i [Y_i(1) - Y_i(0)] = \alpha_i + D_i \beta_i \end{aligned}$$

where

$$\alpha_i \equiv Y_i(0) \text{ and } \beta_i \equiv Y_i(1) - Y_i(0).$$

Let X_i be a continuous, observable "running variable" or "forcing variable" such as a student's test score. In the sharp design, we have

$$D_i = 1\{X_i \geq c\}$$

where c is the cutoff point. Note that the potential outcomes $Y_i(0)$ and $Y_i(1)$ may depend on X_i. Conceptually, we can define the following two conditional means:

$E(Y(0)|X = x)$: the conditional mean of $Y(0)$ given $X = x$ had i been treated;
$E(Y(1)|X = x)$: the conditional mean of $Y(1)$ given $X = x$ had i been untreated.

We can estimate $E(Y(0)|X = x)$ only when $X < c$, as we do not observe $Y(0)$ when $X \geq c$. Similarly, we can estimate $E(Y(1)|X = x)$ only when $X \geq c$, as we do not observe $Y(1)$ when $X < c$.

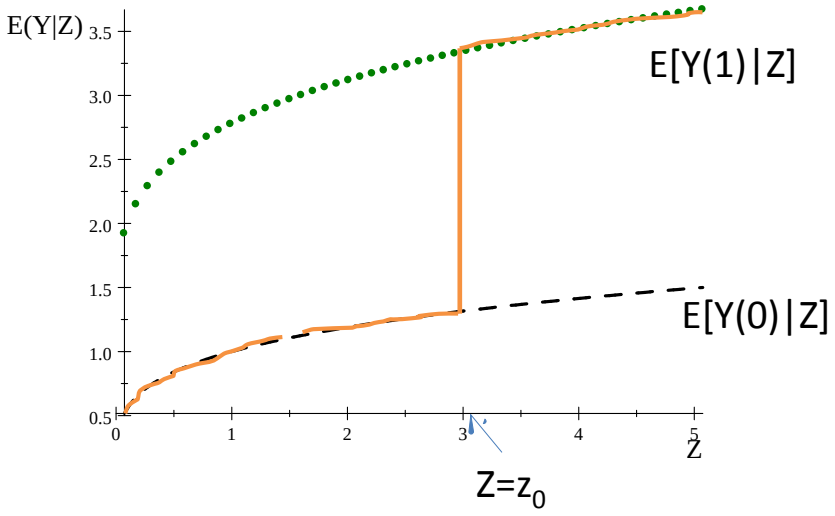


Figure 2.2: Three conditional means. The figure needs to be re-drawn to replace the notational change $(Z, z_0) \to (X, c)$

The two conditional means $E(Y(0)|X = x)$ and $E(Y(1)|Z = z)$ are illustrated in Figure 2.2 (dotted and dashed lines). The solid line signifies the third condition mean $E(Y|X)$, which

is equal to

$$E(Y|X=x)=\begin{cases} E(Y(0)|X=x), & \text{if } X<c \\ E(Y(1)|X=x), & \text{if } X\geq c \end{cases}$$

If both $E(Y(0)|X=x)$ and $E(Y(1)|X=x)$ are continuous but there is a jump in $E(Y|X=x)$ at $x=c$, then we may attribute the jump to the treatment because nothing else besides the treatment status has changed for individuals lying closely on the two sides of the threshold.

In the fuzzy design, the probability of treatment assignment is discontinuous at c. That is

$$\lim_{x\to c+} P(D=1|X=x) \neq \lim_{x\to c-} P(D=1|X=x)$$

without requiring the jump to equal 1. Such a situation can arise if incentives to participate in a program change discontinuously at a threshold, without these incentives being powerful enough to move all units from nonparticipation to participation. Again, any jump in $E(Y|X=x)$ at $x=c$ indicates a nonzero average treatment effect. However,$\lim_{x\to c+} E(Y|X=x) - \lim_{x\to c-} E(Y|X=x)$ is not the treatment effect in this case. The reason is that there are some individuals whose treatment status does not change regardless of whether their X is larger than c or not.

Given that the sharp RD is a special case of the fuzzy RD, we focus the latter and make the following assumptions.

Assumption 2.11.1 *For binary treatment indicator D_i and continuous running variable X_i having positive density in a neighborhood around c, the following limits exist and are not equal to each other ($d^+ \neq d^-$):*

$$d^+ \equiv \lim_{x\to c+} E[D_i \mid X_i=x], \quad d^- \equiv \lim_{x\to c^-} E[D_i \mid X_i=x].$$

Assumption 2.11.2 *In the potential outcomes model, $E[Y_i(0) \mid X_i=x]$ and $E[Y_i(1) \mid X_i=x]$ are continuous in x at $x=c$.*

Assumption 2.11.3 *In the potential outcomes model, $\beta_i=\beta$ for all i.*

Let $ITT_i=1\{X_i\geq c\}$ be the intention to treat. Define

$$D_i=1\left\{ITT_i-\frac{1}{2}>\varepsilon_i\right\}$$

for a continuous random variable $\varepsilon_i \sim (0,\sigma_\varepsilon^2)$ that is independent of X_i. Let $F_{\sigma_\varepsilon^2}(e)=P(\varepsilon_i\leq e)$. Then

$$d^+ \equiv \lim_{x\to c+} E[D_i \mid X_i=x] = F(\frac{1}{2}) \text{ and } d^- \equiv \lim_{x\to c-} E[D_i \mid X_i=x] = F(-\frac{1}{2})$$

It is clean that $d^+ \neq d^-$ and so Assumption 2.11.1 holds. As $\sigma_\varepsilon^2 \to 0$, the treatment assignment rule becomes $D_i = ITT_i$, the sharp design case.

Assumptions 2.11.2 and 2.11.3 are sufficient for identification of a common (i.e., same for all i) treatment effect. By "identification," we mean that the common treatment effect β can be expressed in terms of the joint distribution of observable variables (X_i, Y_i, Z_i): if we know the joint distribution of observables (which asymptotically we will), then we can compute the (unique, true) treatment effect β.

Theorem 2.11.1 *Let Assumptions 2.11.1, 2.11.2, and 2.11.3 hold, then*

$$\beta = \frac{y^+ - y^-}{d^+ - d^-}, \tag{2.17}$$

where

$$y^+ \equiv \lim_{x \to c+} E[Y_i \mid X_i = x], \quad y^- \equiv \lim_{x \to c^-} E[Y_i \mid X_i = x].$$

This is stated as part of Theorem 1 in HTV (2001). We have to be sure that the observed discontinuity in outcomes is driven only by the discontinuity in treatment probability. If for some reason the effect of X_i on the (potential) outcome were also discontinuous at $x = c$, we couldn't disentangle the two effects.

Proof. Consider an arbitrarily small $e > 0$. Then

$$\begin{aligned} & E[Y_i | X_i = c + e] \\ = \; & E[(1 - D_i) Y_i(0) + D_i Y_i(1) | X_i = c + e] \\ = \; & E\{Y_i(0) + D_i [Y_i(1) - Y_i(0)] | X_i = c + e\} \\ = \; & E[Y_i(0) | X_i = c + e] + E[D_i | X_i = c + e] \beta. \end{aligned}$$

Similarly,

$$\begin{aligned} & E[Y_i | X_i = c - e] \\ = \; & E[Y_i(0) | X_i = c - e] + E[D_i | X_i = c - e] \beta. \end{aligned}$$

Hence

$$\begin{aligned} & \lim_{e \to 0} E[Y_i | X_i = c + e] - E[Y_i | X_i = c - e] \\ = \; & \lim_{e \to 0} \{E[D_i | X_i = c + e] - E[D_i | X_i = c - e]\} \beta, \end{aligned}$$

where we have used Assumption 2.11.2. That is,

$$y^+ - y^- = \left(d^+ - d^-\right) \beta.$$

Dividing to isolate β yields (2.17). The denominator in (2.17) is nonzero by Assumption 2.11.1. ■

We may not believe Assumption 2.11.3. We can relax this assumption by allowing for heterogeneous treatment effect but we need to impose Assumption 2.11.4 below. This new assumption is similar to a "selection on observables" or "conditional independence" or "unconfoundedness" assumption often made in the treatment effects literature, but here it is weaker because it need only hold near c. So we hope the discontinuity gives us a sort of locally-randomized experiment. Of course, with our relaxation of assumptions, we cannot identify everybody's treatment effect, but rather the "local average treatment effect" (LATE) for individuals with $X_i = c$: instead of β, our object of interest is $E[\beta_i \mid X_i = c]$.

Assumption 2.11.4 (Local conditional independence) *In a neighborhood of the cutoff c, D_i is independent of β_i conditional on X_i.*

Theorem 2.11.2 *Let Assumptions 2.11.1, 2.11.2, and 2.11.4 hold, then*

$$E[\beta_i \mid X_i = c] = \frac{y^+ - y^-}{d^+ - d^-}.$$

Proof. From the same starting point as the proof of Theorem 2.11.1,

$$\begin{aligned} & E[Y_i|X_i = c + e] = E\left\{Y_i(0) + D_i\beta_i|X_i = c + e\right\} \\ = \; & E\left[Y_i(0)|X_i = c + e\right] + E\left[D_i|X_i = c + e\right] E\left[\beta_i|X_i = c + e\right], \end{aligned}$$

by Assumption 2.11.4. Similarly,

$$\begin{aligned} & E[Y_i|X_i = c - e] \\ = \; & E\left[Y_i(0)|X_i = c - e\right] + E\left[D_i|X_i = c - e\right] E\left[\beta_i|X_i = c - e\right]. \end{aligned}$$

From this and Assumptions 2.11.2, we have

$$\begin{aligned} & \lim_{x \to c+} E[Y_i \mid X_i = x] - \lim_{x \to c^-} E[Y_i \mid X_i = x] \\ & \quad = E[\beta_i \mid X_i = c] \left\{ \lim_{x \to c+} E[D_i \mid X_i = x] - \lim_{x \to c^-} E[D_i \mid X_i = x] \right\}. \end{aligned}$$

The theorem then follows from Assumption 2.11.1. ■

Assumption 2.11.4 may be still strong in the absence of a randomized experiment. The formula $E[\beta_i \mid X_i = x] = (y^+ - y^-)/(d^+ - d^-)$ reminds us of the Wald estimator in an IV regression. We want to see whether $(y^+ - y^-)/(d^+ - d^-)$ still has a causal interpretation without Assumption 2.11.4.

We first introduce some notation. Define the binary random variable $D_i(x)$ as

$$D_i(x) = \begin{cases} 1, & \text{if an individual would participate had his running variable been } x \\ 0, & \text{if an individual would not participate had his running variable been } x \end{cases}$$

$D_i(x)$ indicates the potential treatment status or potential participation. Clearly, we cannot observe the entire set of potential participation indicators $\{D_i(x) \text{ for all } x \in \text{supp}(X)\}$, but we can think about them in the same way as we think about $Y_i(0)$ and $Y_i(1)$ even though they are not observed.

We can divide individuals into four groups according to their decisions under different values of X near c :

	$D_i(c-e)=0$	$D_i(c-e)=1$
$D_i(c+e)=0$	never taker ($\mathcal{N}$)	defier ($\mathcal{D}$)
$D_i(c+e)=1$	complier ($\mathcal{C}$)	always taker ($\mathcal{A}$)

Alternatively, we can define the event

$$\mathcal{N} = \{(D_i(c-e), D_i(c+e)) = (0,0)\}$$

and similarly for $\mathcal{C}$, $\mathcal{D}$, $\mathcal{A}$. For example, if individual i makes her decision according to $D_i(x) = 1\{x > c\}$, then she is a complier.

Assumption 2.11.5 *(i) There exists an $\varepsilon > 0$ such that $D_i(c+e) \geq D_i(c-e)$ for all $e \leq \varepsilon$. (ii) $P(\mathcal{I}|X_i = x) = P(\mathcal{I})$ for $\mathcal{I} = \mathcal{N}$, $\mathcal{C}$, $\mathcal{A}$ and for x in a neighborhood of c.*

Theorem 2.11.3 *Let Assumptions 2.11.1, 2.11.2 and 2.11.5 hold, then*

$$E[Y_i(1) - Y_i(0) | X_i = c, \mathcal{C}] = \frac{y^+ - y^-}{d^+ - d^-}.$$

Proof. Note that

$$\begin{aligned} E(Y_i|X_i = c+e) &= E\{Y_i(0) + [Y_i(1) - Y_i(0)] D_i | X_i = c+e\} \\ E(Y_i|X_i = c-e) &= E\{Y_i(0) + [Y_i(1) - Y_i(0)] D_i | X_i = c-e\}, \end{aligned}$$

we have

$$\begin{aligned} & E(Y_i|X_i = c+e) - E(Y_i|X_i = c-e) \\ = \; & E[Y_i(0)|X_i = c+e] - E[Y_i(0)|X_i = c-e] \\ & + E\{[Y_i(1) - Y_i(0)] D_i | X_i = c+e\} - E\{[Y_i(1) - Y_i(0)] D_i | X_i = c-e\}. \end{aligned}$$

Under Assumption 2.11.1, the first term converges to zero as $e \to 0$. For the second term, we first observe that $P(\mathcal{D}) = 0$ under the monotonicity condition in Assumption 2.11.5(i). So

$$\begin{aligned}
& E\left\{\left[Y_i(1) - Y_i(0)\right] D_i | X_i = c + e\right\} \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] D_i(c+e) | X_i = c + e, \mathcal{C}\right\} P\left\{\mathcal{C}\right\} \\
& + E\left\{\left[Y_i(1) - Y_i(0)\right] D_i(c+e) | X_i = c + e, \mathcal{N}\right\} P\left\{\mathcal{N}\right\} \\
& + E\left\{\left[Y_i(1) - Y_i(0)\right] D_i(c+e) | X_i = c + e, \mathcal{A}\right\} P\left\{\mathcal{A}\right\} \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] | X_i = c + e, \mathcal{C}\right\} P\left\{\mathcal{C}\right\} \\
& + E\left\{\left[Y_i(1) - Y_i(0)\right] | X_i = c + e, \mathcal{A}\right\} P\left\{\mathcal{A}\right\}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& E\left\{\left[Y_i(1) - Y_i(0)\right] D_i | X_i = c - e\right\} \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] D_i(c-e) | X_i = c - e, \mathcal{C}\right\} P\left\{\mathcal{C}\right\} \\
& + E\left\{\left[Y_i(1) - Y_i(0)\right] D_i(c-e) | X_i = c - e, \mathcal{N}\right\} P\left\{\mathcal{N}\right\} \\
& + E\left\{\left[Y_i(1) - Y_i(0)\right] D_i(c-e) | X_i = c - e, \mathcal{A}\right\} P\left\{\mathcal{A}\right\} \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] | X_i = c - e, \mathcal{A}\right\} P\left\{\mathcal{A}\right\}.
\end{aligned}$$

Hence, using Assumption 2.11.5(ii), we have

$$\begin{aligned}
& E\left\{\left[Y_i(1) - Y_i(0)\right] D_i | X_i = c + e\right\} - E\left\{\left[Y_i(1) - Y_i(0)\right] D_i | X_i = c - e\right\} \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] | X_i = c + e, \mathcal{C}\right\} P\left\{\mathcal{C}\right\}.
\end{aligned}$$

We have therefore shown that

$$\begin{aligned}
& E\left(Y_i | X_i = c + e\right) - E\left(Y_i | X_i = c - e\right) \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] | X_i = c + e, \mathcal{C}\right\} P\left\{D_i(c+e) - D_i(c-e) = 1\right\} \\
= \; & E\left\{\left[Y_i(1) - Y_i(0)\right] | X_i = c + e, \mathcal{C}\right\} E\left[D_i(c+e) - D_i(c-e)\right]
\end{aligned}$$

where the last line follows because $D_i(c+e) - D_i(c-e)$ is a binary random variable under the monotonicity condition given in Assumption 2.11.5(i).

Letting $e \to 0$ yields

$$y^+ - y^- = E\left[Y_i(1) - Y_i(0) | X_i = c, C\right]\left(d^+ - d^-\right).$$

That is

$$E\left[Y_i(1) - Y_i(0) | X_i = c, C\right] = \frac{y^+ - y^-}{d^+ - d^-}$$

as desired. ■

The two tables below present the basic intuition.

The subset of individuals whose $X_i = c + e$

Individual	treated or not: D_i	$Y_i(1)$	$Y_i(0)$	Y_i
$\mathcal{C}_1$	1	$Y_{\mathcal{C}_1}(1)$	?	$Y_{\mathcal{C}_1}(1)$
$\mathcal{C}_2$	1	$Y_{\mathcal{C}_2}(1)$	?	$Y_{\mathcal{C}_2}(1)$
$\mathcal{A}_1$	1	$Y_{\mathcal{A}_1}(1)$	?	$Y_{\mathcal{A}_1}(1)$
$\mathcal{A}_2$	1	$Y_{\mathcal{A}_2}(1)$	?	$Y_{\mathcal{A}_2}(1)$
$\mathcal{N}_1$	0	?	$Y_{\mathcal{N}_1}(0)$	$Y_{\mathcal{N}_1}(0)$
$\mathcal{N}_2$	0	?	$Y_{\mathcal{N}_2}(0)$	$Y_{\mathcal{N}_2}(0)$

$$E(Y_i|X_i = c + e) = \bar{Y}_{\mathcal{C}}(1) P(\mathcal{C}) + \bar{Y}_{\mathcal{A}}(1) P(\mathcal{A}) + \bar{Y}_{\mathcal{N}}(0) P(\mathcal{N})$$

The subset of individuals whose $X_i = c - e$

Individual	treated or not: D_i	$Y_i(1)$	$Y_i(0)$	Y_i
$\mathcal{C}_1$	0	?	$Y_{\mathcal{C}_1}(0)$	$Y_{\mathcal{C}_1}(0)$
$\mathcal{C}_2$	0	?	$Y_{\mathcal{C}_2}(0)$	$Y_{\mathcal{C}_2}(0)$
$\mathcal{A}_1$	1	$Y_{\mathcal{A}_1}(1)$	?	$Y_{\mathcal{A}_1}(1)$
$\mathcal{A}_2$	1	$Y_{\mathcal{A}_2}(1)$	?	$Y_{\mathcal{A}_2}(1)$
$\mathcal{N}_1$	0	?	$Y_{\mathcal{N}_1}(0)$	$Y_{\mathcal{N}_1}(0)$
$\mathcal{N}_2$	0	?	$Y_{\mathcal{N}_2}(0)$	$Y_{\mathcal{N}_2}(0)$

$$E(Y_i|X_i = c - e) = \bar{Y}_{\mathcal{C}}(0) P(\mathcal{C}) + \bar{Y}_{\mathcal{A}}(1) P(\mathcal{A}) + \bar{Y}_{\mathcal{N}}(0) P(\mathcal{N})$$

So

$$E(Y_i|X_i = c + e) - E(Y_i|X_i = c - e) = \left[\bar{Y}_{\mathcal{C}}(1) - \bar{Y}_{\mathcal{C}}(0)\right] P(\mathcal{C}).$$

The theorem is proved under the assumptions weaker than HTV (2001). HTV assumes that $(Y_i(1) - Y_i(0), D_i(x))$ is independent of X_i for X_i in a neighborhood of $x = c$.Here we allows for the (local) dependence between $Y_i(1) - Y_i(0)$ and X_i.

2.11.2 Estimation under Sharp RDD and Fuzzy RDD

For estimation, we could obtain a consistent estimator of the LATE by plugging in any consistent estimators of y^+, y^-, d^+, and d^-. (Only the former two are needed for the sharp RD

case, since the denominator is one.) Recall that the local linear estimator (or any local polynomial) performs better at the boundary than the local constant; consequently, the former is preferred in the RD setting, since we are looking at observations with $X_i < c$ but estimating the value at c, and then looking at observations with $X \geq c$ but again estimating the value at the boundary c. So the issue of estimation essentially boils down to nonparametric estimation at a boundary.

Sharp RDD

Let

$$\begin{aligned}\left(\hat{\theta}_0^+, \hat{\theta}_1^+\right) &= \arg\min_{(\theta_0,\theta_1)} \sum_{i=1}^{n} \left[Y_i - \theta_0 - \theta_1 (X_i - c)\right]^2 K_h(X_i - c) 1\{X_i \geq c\} \\ &= \left(Z_c' W_{c,r} Z_c\right)^{-1} Z_c' W_{c,r} Y\end{aligned}$$

where

$$W_{c,r} = diag\left\{K_h(X_i - c) 1\left\{\frac{X_i - c}{h} \geq 0\right\}\right\} = diag\left\{K_{r,h}(X_i - c)\right\}$$

with

$$K_r(u) = K(u) 1\{u \geq 0\}.$$

Given that c is a constant, we write $Z = Z_c$ and $W_r = W_{c,r}$ hereafter to suppress their dependence on c.

Similarly,

$$\begin{aligned}\left(\hat{\theta}_0^-, \hat{\theta}_1^-\right) &= \arg\min_{(\theta_0,\theta_1)} \sum_{i=1}^{n} \left[Y_i - \theta_0 - \theta_1 (X_i - c)\right]^2 K_h(X_i - c) 1\{X_i < c\} \\ &= \left(Z' W_l Z\right)^{-1} Z' W_l Y\end{aligned}$$

where

$$W_l = diag\left\{K_{l,h}(X_i - c)\right\} \text{ with } K_l(u) = K(u) 1\{u < 0\}.$$

Let $\hat{y}^+ = \hat{\theta}_0^+$ and $\hat{y}^- = \hat{\theta}_0^-$. Then we can estimate $y_\Delta = y^+ - y^-$ by

$$\hat{y}_\Delta = \hat{y}_0^+ - \hat{y}_0^-$$

which is our proposed estimator of the ATE in the sharp RDD.

Let $\Sigma_y = Var(Y|X) = diag(var(Y_i|X_i))$. Following the same proof for the local linear estimator at an interior point, we can show that the asymptotic variance of $\hat{y}^+$ is

$$asymvar\left(\sqrt{nh}\hat{y}^+\right) = \text{plime}_1' \left(\frac{Z' W_r Z}{n}\right)^{-1} \left(\frac{Z' W_r \Sigma_y W_r Z}{nh^{-1}}\right) \left(\frac{Z' W_r Z}{n}\right)^{-1} e_1,$$

where

$$\frac{Z'W_rZ}{n} = \begin{pmatrix} \frac{1}{n}\sum\left[K_{rh}\left(X_i-c\right)\right] & \frac{1}{n}\sum_{i=1}^{n}\left[K_{rh}\left(X_i-c\right)\right]\left(\frac{X_i-c}{h}\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left[K_{rh}\left(X_i-c\right)\right]\left(\frac{X_i-c}{h}\right) & \frac{1}{n}\sum_{i=1}^{n}\left[K_{rh}\left(X_i-c\right)\right]\left(\frac{X_i-c}{h}\right)^2 \end{pmatrix}$$
$$\rightarrow^p \begin{pmatrix} \int K_r\left(u\right)du & \int K_r\left(u\right)udu \\ \int K_r\left(u\right)udu & \int K_r\left(u\right)u^2du \end{pmatrix} f_X\left(c\right) := \Gamma_r f_X\left(c\right)$$

and

$$\frac{Z'W_r\Sigma_yW_rZ}{nh^{-1}}$$
$$= \begin{pmatrix} \frac{h}{n}\sum_{i=1}^{n}\left[K_{rh}^2\left(X_i-x\right)\right]\sigma_y^2\left(X_i\right) & \frac{h}{n}\sum_{i=1}^{n}\left[K_{rh}^2\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)\sigma_y^2\left(X_i\right) \\ \frac{h}{n}\sum_{i=1}^{n}\left[K_{rh}^2\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)\sigma_y^2\left(X_i\right) & \frac{h}{n}\sum_{i=1}^{n}\left[K_{rh}^2\left(X_i-x\right)\right]\left(\frac{X_i-x}{h}\right)^2\sigma_y^2\left(X_i\right) \end{pmatrix}$$
$$\rightarrow^p \begin{pmatrix} \int K_r^2\left(u\right)du & \int K_r^2\left(u\right)udu \\ \int K_r^2\left(u\right)udu & \int K_r^2\left(u\right)u^2du \end{pmatrix} \sigma_{yr}^2\left(c\right)f_X\left(c\right) := V_r\sigma_{yr}^2\left(c\right)f_X\left(c\right),$$

with $\sigma_{yr}^2\left(c\right) = \lim_{x\rightarrow c+} Var(Y_i|X_i=x)$. Hence

$$asymvar\left(\sqrt{nh}\hat{y}^+\right) = e_1'\left[\Gamma_r f_X\left(c\right)\right]^{-1}\left[V_r\sigma_r^2\left(c\right)f_X\left(c\right)\right]\left[\Gamma_r f_X\left(c\right)\right]^{-1}e_1 = \frac{\sigma_{yr}^2\left(c\right)}{f_X\left(c\right)}e_1'\Gamma_r^{-1}V_r\Gamma_r^{-1}e_1.$$

The approximate bias of $\hat{y}^+$ is

$$abias\left(\hat{y}^+\right)$$
$$= h^2 \times abias\left[\left(\frac{Z'W_rZ}{n}\right)^{-1}\frac{Z'W_rY}{nh^2}\right] = h^2 \times e_1'\left[\Gamma_r f_X\left(c\right)\right]^{-1}\left(\operatorname{p}\lim_{n\rightarrow\infty}\frac{Z'W_r\mathcal{R}_1}{nh^2}\right),$$

where

$$\frac{Z'W_r\mathcal{R}_1}{nh^2} = \begin{pmatrix} \frac{1}{n}\sum\left[K_{rh}\left(X_i-c\right)\right]\frac{1}{2}m_{y,r}''\left(X_i\right)\left(\frac{X_i-c}{h}\right)^2 \\ \frac{1}{n}\sum_{i=1}^{n}\left[K_{rh}\left(X_i-c\right)\right]\frac{1}{2}m_{yr}''\left(X_i\right)\left(\frac{X_i-c}{h}\right)^3 \end{pmatrix}$$
$$\rightarrow^p \frac{1}{2}m_{yr}''\left(c\right)f_X\left(c\right)\begin{pmatrix} \int K_r\left(u\right)u^2du \\ \int K_r\left(u\right)u^3du \end{pmatrix}$$

and $m''_{yr}(c)$ is the second order right derive of $E(Y_i|X_i = x)$ at $x = c$. So

$$\begin{aligned} & abias\left(\hat{y}^{+}\right) \\ = & \frac{1}{2}m''_{yr}(c)\,h^2 \times e_1'\Gamma_r^{-1}\begin{pmatrix}\int K_r(u)\,u^2du \\ \int K_r(u)\,u^3du\end{pmatrix} := \frac{1}{2}m''_{yr}(c)\,h^2 \times e_1'\Gamma_r^{-1}\mu_r. \end{aligned}$$

We can carry out the same calculations for $\hat{y}^{-}$. We have

$$\begin{aligned} asymvar\left(\sqrt{nh}\hat{y}^{-}\right) &= \frac{\sigma_{yl}^2(c)}{f_X(c)}e_1'\Gamma_l^{-1}V_l\Gamma_l^{-1}e_1 = \frac{\sigma_{yl}^2(c)}{f_X(c)}e_1'\Gamma_r^{-1}V_r\Gamma_r^{-1}e \\ abias\left(\hat{y}^{-}\right) &= \frac{1}{2}m''_{yl}(c)\,h^2 \times e_1'\Gamma_l^{-1}\mu_l = \frac{1}{2}m''_{yl}(c)\,h^2 \times e_1'\Gamma_r^{-1}\mu_r \end{aligned}$$

where the second equality in each of the above two lines follows from simple calculations using the symmetry of $K(\cdot)$.

Using a triangular array CLT, we have

$$\begin{aligned} & \begin{pmatrix} \sqrt{nh}\left(\hat{y}^{+} - y^{+} - \frac{1}{2}m''_{yr}(c)\,h^2 \times e_1'\Gamma_r^{-1}\mu_r\right) \\ \sqrt{nh}\left(\hat{y}^{-} - y^{-} - \frac{1}{2}m''_{yl}(c)\,h^2 \times e_1'\Gamma_r^{-1}\mu_r\right)\end{pmatrix} \\ \rightarrow\ & {}^{d}N\left[0, \begin{pmatrix} \frac{\sigma_{yr}^2(c)}{f_X(c)}e_1'\Gamma_r^{-1}V_r\Gamma_r^{-1}e_1 & 0 \\ 0 & \frac{\sigma_{yl}^2(c)}{f_X(c)}e_1'\Gamma_r^{-1}V_r\Gamma_r^{-1}e_1\end{pmatrix}\right] \end{aligned}$$

as $n \to \infty, h \to 0$ such that $\sqrt{nh}h^2 < \infty$. Let

$$B_{y_\Delta} = \frac{1}{2}\left(m''_{yr}(c) - m''_{yl}(c)\right)h^2 \times e_1'\Gamma_r^{-1}\mu_r,$$

then

$$\sqrt{nh}\left(\hat{y}_\Delta - y_\Delta - h^2B_{y_\Delta}\right) \rightarrow^{d} N\left[0, \frac{\sigma_{yr}^2(c) + \sigma_{yl}^2(c)}{f_X(c)}e_1'\Gamma_r^{-1}V_r\Gamma_r^{-1}e_1\right]$$

as $n \to \infty, h \to 0$ such that $\sqrt{nh}h^2 < \infty$.

The bias and variance formulae can be used to find the AMSE optimal bandwidth. See Sun (2005) for details.

Fuzzy RDD

For the Fuzzy RDD, we need to estimate $d_\Delta = d^{+} - d^{-}$. We can follow the same procedure for estimating $y^{+} - y^{+}$ and obtain

$$\sqrt{nh}\left(\hat{d}_\Delta - d_\Delta - h^2B_{d_\Delta}\right) \rightarrow^{d} N\left[0, \frac{\sigma_{dr}^2(c) + \sigma_{dl}^2(c)}{f_X(c)}e_1'\Gamma_r^{-1}V_r\Gamma_r^{-1}e_1\right]$$

where

$$B_{d_\Delta} = \frac{1}{2}\left(m''_{dr}(c) - m''_{dl}(c)\right) h^2 \times e_1'\Gamma_r^{-1}\mu_r$$

with $m''_{dr}(c), m''_{dl}(c), \sigma^2_{dr}(c), \sigma^2_{dl}(c)$ defined in the same manner as $m''_{yr}(c), m''_{yl}(c), \sigma^2_{yr}(c), \sigma^2_{yl}(c)$ is respectively defined.

The LATE estimator is then $\hat{y}_\Delta/\hat{d}_\Delta$. Under the rate condition that $h = Cn^{-\frac{1}{5}}$ for some finite constant $C > 0$, we have

$$\sqrt{nh}\left(\frac{\hat{y}_\Delta}{\hat{d}_\Delta} - \frac{y_\Delta}{d_\Delta}\right) = \frac{1}{d_\Delta}\sqrt{nh}\left(\hat{y}_\Delta - y_\Delta\right) - \frac{y_\Delta}{d_\Delta^2}\sqrt{nh}\left(\hat{d}_\Delta - d_\Delta\right) + s.o.$$

So

$$\sqrt{nh}\left(\frac{\hat{y}_\Delta}{\hat{d}_\Delta} - \frac{y_\Delta}{d_\Delta} - h^2 B\right) \to N(0, V)$$

where

$$B = \frac{B_{d_{\Delta y}}}{d_\Delta} - \frac{y_\Delta}{d_\Delta^2} B_{d_{\Delta d}}$$

and

$$\begin{aligned} V \;=\; & \frac{1}{d_\Delta^2}\frac{\sigma^2_{yr}(c) + \sigma^2_{yl}(c)}{f_X(c)} e_1'\Gamma_r^{-1} V_r \Gamma_r^{-1} e_1 \\ & + \frac{y_\Delta^2}{d_\Delta^4}\frac{\sigma^2_{dr}(c) + \sigma^2_{dl}(c)}{f_X(c)} e_1'\Gamma_r^{-1} V_r \Gamma_r^{-1} e_1 \\ & - \frac{2y_\Delta}{d_\Delta^3}\frac{\sigma^2_{dyr}(c) + \sigma^2_{dyl}(c)}{f_X(c)} e_1'\Gamma_r^{-1} V_r \Gamma_r^{-1} e_1. \end{aligned}$$

Here $\sigma^2_{dyr}(c) = \lim_{x\to c+} cov(Y_i, D_i | X_i = x)$ and $\sigma^2_{dyl}(c) = \lim_{x\to c-} cov(Y_i, D_i | X_i = x)$.

Using the above bias and variance formulae, Imbens and Kalyanarama (2012) obtain the AMSE optimal bandwidth. They also investigate its plug-in implementation which regularizes the denominator estimator.

2.12 Problems

1. The local polynomial estimator of order p has a remarkable property: It reproduces polynomials of degree $\leq p$ in the following sense. Let $Q(x)$ be a polynomial of degree less than or equal to p. Show that

$$\sum_{i=1}^{n} w_{ni}(x) Q(X_i) = Q(x).$$

In particular,

$$\sum_{i=1}^{n} w_{ni}(x) = 1$$

$$\sum_{i=1}^{n} w_{ni}(x)(X_i - x)^k = 0 \text{ for } k = 1, 2, ..., p.$$

Hint: To prove this property, we note

$$Q(X_i) = Q(x) + Q'(x)(X_i - x) + ... + Q^{(p)}(x)\frac{(X_i - x)^p}{p!}$$

and

$$Q := \begin{pmatrix} Q(X_1) \\ Q(X_2) \\ ... \\ Q(X_n) \end{pmatrix} = \begin{pmatrix} 1 & \frac{X_1 - x}{h} & ... & \frac{1}{p!}\left(\frac{X_1 - x}{h}\right)^p \\ 1 & \frac{X_2 - x}{h} & ... & \frac{1}{p!}\left(\frac{X_2 - x}{h}\right)^p \\ ... & ... & \ddots & ... \\ 1 & \frac{X_n - x}{h} & .. & \frac{1}{p!}\left(\frac{X_n - x}{h}\right)^p \end{pmatrix} \begin{pmatrix} Q(x) \\ hQ'(x) \\ ... \\ h^p Q^{(p)}(x) \end{pmatrix} = Z_x Q_x$$

So

$$e_1'\left(Z_x' W_x Z_x\right)^{-1} Z_x' W_x Q = e_1' Q_x = Q(x)$$

That is

$$\sum_{i=1}^{n} w_{ni}(x) Q(X_i) = Q(x)$$

2. Consider the local polynomial regression with p-th order polynomial. Assume that (i) there exists a real number $\lambda_0 > 0$ and a positive integer n_0 such that the smallest eigenvalue $\lambda_{\min}(Z_x' W_x Z_x / n) > \lambda_0$ almost surely for all $n \geq n_0$; (ii) the kernel function has a compact support $[-1, 1]$ and $|K(u)| \leq K_{\max}$ for all $u \in R$; (iii) $m(x) \in H(s, L)$ the Hölder class with bound parameter L and smoothness parameter s satisfying $\lfloor s \rfloor = p$; (iv) $f_X(x) \in (0, \infty)$.

(a) Prove $\sum_{i=1}^{n} |w_{ni}(x)| \leq C$ almost surely for $n \geq n_0$, where C is a constant depending only on $\lambda_0, K_{\max}$ and $f_X(x)$.

(b) Prove $w_{ni}(x) = 0$ if $|x - X_i| > h$.

(c) Using (a) and (b), prove that the asymptotic bias is of order h^s. That is, show

$$E\hat{e}^*(x) = E\sum_{i=1}^{n} w_{ni}(x)[m(X_i) - m(x)] = O(h^s)$$

Hint: if $m(x) \in H(s, L)$, then

$$m(z) - m(x) = \sum_{i=1}^{p} m^{(k)}(x) \frac{(z-x)^k}{k!} + \left[m^{(p)}(x + \tau(z-x)) - m^{(p)}(x) \right] \frac{(z-x)^p}{p!}$$

for some $\tau \in [0, 1]$.

3. Prove that the weights $w_{ni}(x)$ for the local polynomial smoother satisfy

$$w_{ni}(x) = K\left(\frac{X_i - x}{h}\right) Q\left(\frac{X_i - x}{h}\right)$$

for some polynomial

$$Q(z) = \alpha_0 + \alpha_1 z + ... + \alpha_p z^p$$

where $\alpha_0, ..., \alpha_p$ may depend on the kernel function and the design points $X_1, ..., X_n$. This result shows that the local polynomial smoother can be regarded as the NW estimator with high order kernels that is design adaptive.

4. The local polynomial estimator is a linear smoother. That is, $\hat{Y} = LY$ for some matrix L. Investigate the relationship among the (expected) trace of the smoother, the degree of the local polynomial, and the bandwidth. Assuming that X_i is iid uniform on [0,1], graph the expected trace against the bandwidth for each local polynomial considered (local constant, local linear, local quadratic and local cubic polynomials). Discuss what you find from the perspective of the equivalent degree of freedom.

5. Consider the local polynomial regression with $n = 100$, $X_i = i/n$, $K(x) = \left(\sqrt{2\pi}\right)^{-1} \exp\left(-x^2/2\right)$ and $h = 0.2$. For each $x \in \{0,\ 0.25,\ 0.50,\ 0.75,\ 1\}$ and $p = 0, 1, 2, 3$, plot $w_{ni}(x)$ against X_i. Describe the graphs. For example, discuss whether the shape of the curve changes with the target point x.

6. Consider a nonparametric regression

$$Y_i = m(X_i) + \sigma\varepsilon_i$$

where $X_i = i/n$, $\varepsilon_i \sim iidN(0,\ 1)$ and

$$m(x) = \sqrt{x(1-x)} \sin\left(\frac{2.1\pi}{x + 0.05}\right), x \in [0, 1]$$

The function is plotted in Figure (2.3).

(a) Make a data set with sample size $n = 1000$ and $\sigma = 0.1$. Plot the data.

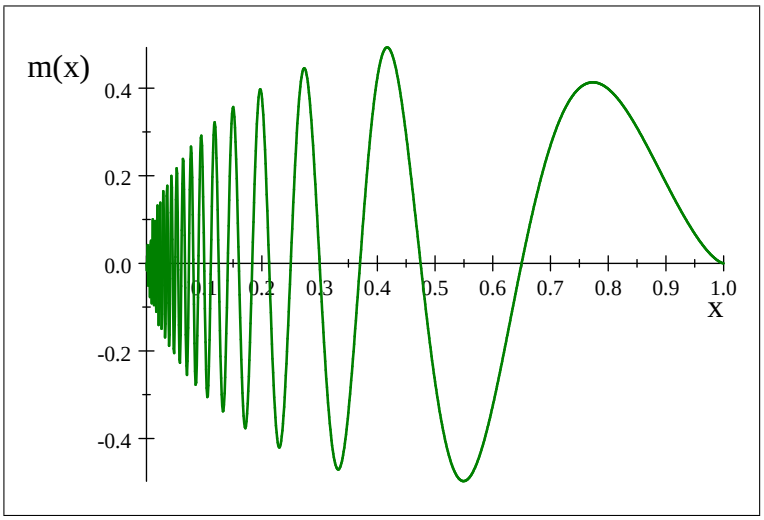


Figure 2.3: Doppler function

(b) Estimate the function using local linear regression with a Gaussian kernel $K(x) = \left(\sqrt{2\pi}\right)^{-1} \exp\left(-x^2/2\right)$. Plot the cross-validation criterion, generalized cross-validation criterion, Mallows' C_p criterion versus the bandwidth. For Mallows' C_p criterion, use half of the cross-validation bandwidth as the pilot bandwidth to estimate σ^2. Report each of the bandwidth choice.

(c) Plot the fitted function for each bandwidth choice. Find and plot a 95 percent confidence band for each fitted function.

(d) Repeat (a)-(c) for $\sigma = 2$. Compare the results for this case with those for $\sigma = 0.1$.

2.13 References

1. Akaike, H. (1974). "A new look at the statistical identification model." *IEEE Transactions in Automatic Control*, 19, 716–25.

2. Andrews, D. W. K. (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* Vol. 59, 817–858.

3. Andrews, D. W. K. and Y. Sun (2004): "Adaptive Local Polynomial Whittle Estimation of Long-range Dependence." *Econometrica* 72(2) 569–614

4. Cox, D. R., and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman & Hall, London.

5. Chaudhuri, P. (1991). "Nonparametric estimates of regression quantiles and their local Bahadur representation." *The Annals of Statistics*, 19, 760–777.

6. Cleveland, W. S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74 (368): 829–836.

7. Craven, P. and G. Wahba (1979). "Smoothing noisy data with spline functions." *Numerische Mathematik*, 31, 377–403.

8. Hahn, J., Todd, P. and W. Van Der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design", *Econometrica*, 69, 201–209.

9. Hall, P. (1993). "On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation." *The Annals of Statistics*, 55, 291–304.

10. Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

11. Härdle, W. and O. Linton (1994). Applied Nonparametric Methods. In *Handbook of Econometrics*, Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2295–2339. Elsevier.

12. Härdle, W. and J.S. Marron (1985). Optimal Bandwidth Selection in Nonparametric Regression Function Estimation, *The Annals of Statistics,* 13(4), 1465–1481 .

13. Härdle, W. and J.S. Marron (1991). "Bootstrap simultaneous error bars for nonparametric regression." *The Annals of Statistics,* 19(2), pp. 778–796.

14. Imbens, G. and T. Lemieux. (2008), "Regression Discontinuity Designs", Journal of Econometrics, 142, 615–635.

15. Imbens, G. and K. Kalyanarama (2008): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies*, 79, 933–959

16. Kaplan, D. M. (2015). "Improved quantile inference via fixed-smoothing asymptotics and Edgeworth expansion." *Journal of Econometrics*, 185(1): 20–32.

17. Koneker, R. (2005). *Quantile Regression*, Cambridge University Press.

18. Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall.

19. Loader C. (1999). "Bandwidth selection: classical or plug-in?" *The Annals of Statistics,* 27(2), 415–438.

20. Loader C. (1999). *Local Regression and Likelihood*, Springer: New York.

21. Lee, D. S., and T. Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281-355.

22. Li, Q. and J.S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

23. Mack, Y.P. (1981). "Local properties of k-NN regression estimates", *SIAM Journal on Algebraic and Discrete Methods* 2, 311–323.

24. Mallows, C.L. (1973). "Some Comments on Cp." *Technometrics*, 15, 661–675.

25. Marron, J.S. and W. Hardle (1986): "Random approximations to some measures of accuracy in nonparametric curve estimation." *Journal of Multivariate Analysis.* pp. 91–113.

26. Meng, H., Hal White and Y. Sun (2014): "A Flexible Nonparametric Test for Conditional Independence." Working paper, Department of Economics, UC San Diego

27. Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press.

28. Sun, Y. (2005). "Adaptive Estimation of the Regression Discontinuity Model." Unpublished paper. Department of Economics, UC San Diego.

29. Thistlewaite, D. and D. Campbell, D. (1960). "Regression-Discontinuity Analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology* 51 (6): 309–317.

Chapter 3

All of Series Estimation

As an alternative to kernel estimation, another popular device to perform nonparametric estimation is the so-called method of sieves. Sieve estimators include spline estimators and series estimators as special cases. Sometimes, sieve estimators are identified with series estimators in the econometrics literature but the class of sieve estimators is in fact much larger than the class of series estimators. Sieve estimators can be used to estimate both conditional moment and conditional density.

3.1 Examples and Motivations

To motivate the method of sieves, consider the nonparametric regression:

$$Y = m(X) + \varepsilon, E[\varepsilon|X] = 0.$$

A naive approach to estimating $m(\cdot)$ is to solve the following minimization problem:

$$\hat{m}(x) = \arg \min_{m(\cdot)\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - m(X_i))^2,$$

where $\mathcal{M}$ is a function space that the true $m(\cdot)$ lies in. If $\mathcal{M}$ is rich enough, the estimator $\hat{m}(x)$ will interpolate the data, assuming no ties in X_i's. In general, the so-chosen function will not converge to the true regression function.

The difficulties encountered in moving from finite to infinite dimension space can also be illustrated by the failure of MLE in nonparametric density estimation. Let $X_1, ..., X_n$ be an iid sample from an absolutely continuous distribution with unknown density function $f(x)$. The naive nonparametric MLE of $f(x)$ is

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \ell(X_1, .., X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i) \tag{3.1}$$

where $\mathcal{F}$ is the space of 'allowable' densities. If $\mathcal{F}$ is too rich, the MLE will fail to produce a meaningful estimator. The basic problem is that any estimator of f that is continuous can be improved upon by shrinking it towards a discrete distribution with jumps at the sample points.

Grenander (1981) suggests that we attempt the optimization problem (minimization of the SSR or the negative log-likelihood) within a subspace of the parameter space and then allow this subspace to grow with sample size. This sequence of subspaces from which the estimator is drawn is called a "sieve", and the resulting procedure is called "the method of sieves". The approximating subspaces are called "sieve spaces."

For example, we can define

$$\mathcal{M}_n = \left\{ m(\cdot) : m(x) = \sum_{j=1}^{J_n} \beta_j \phi_j(x) \right\}$$

where $\{\phi_j\}$ are some basis functions. The sieve LS estimator of $m(x)$ is given by

$$\hat{m}(x) = \arg \min_{m(\cdot) \in \mathcal{M}_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m(X_i))^2.$$

It can be shown that $\hat{m}(x)$ is consistent if (i) $\mathcal{M}_n$ grows with n so that $\mathcal{M}_n$ becomes dense in $\mathcal{M}$; (ii) $\mathcal{M}_n$ does not grow too fast. In this example, the growth of $\mathcal{M}_n$ is controlled by the parameter J_n.

In the case of MLE, we can let

$$\mathcal{F}_n = \left\{ f(x) : f(x) = \left[\sum_{j=0}^{J_n} \beta_j H_j(x) \exp\left(-\frac{x^2}{2}\right) \right]^2 \right\}$$

where $\{H_j(x)\}$ are Hermit polynomials. We define

$$\hat{f} = \arg \min_{f \in \mathcal{F}_n} \ell(X_1, ..., X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log f(X_i).$$

Again, it can be shown that the above $\hat{f}$ is consistent when $J_n \to \infty$ at an appropriate rate with n.

An appealing feature of the sieve estimator is the huge range of applications; basically any nonparametric estimation problem can be handled with the method of sieves. Furthermore, its close resemblance to the parametric estimator makes it easy to compute and fairly intuitive. Finally, the extension to additive structures and semiparametric models is very convenient, see Andrews and Whang (1990) and Andrews (1991).

The downturn is that even though the optimization problem is finite-dimensional, it can still be a formidable task to actually solve it. In particular, as the dimension of the domain of $\mathcal{F}_n$ or $\mathcal{M}_n$ increases, "the size" of sieve space can become very large. So here the well-known curse of dimensionality is not just a theoretical problem, but also a practical issue. In addition, there is relatively little theory about how to select the sieve spaces and how to optimally determine their growth with the sample size. Also it is often technically more difficult to establish the asymptotic properties of sieve estimators than the kernel estimators, though the difficulty has been largely overcome in the recent literature.

3.2 Sieve Spaces

In this section, we present some commonly used sieves whose approximation properties are known in the mathematical literature. As before, we focus on the univariate case. We do so to enhance clarity at the cost of generality.

3.2.1 Hölder Class and Finite Dimensional Linear Sieves

The most popular class of functions considered in nonparametric literature is the Hölder smoothness class. Let $[p]$ be the largest integer strictly less than p so that $p = [p] + \alpha$ for $\alpha \in (0, 1]$. A real-valued function h on $\mathcal{X}$ is said to be p-smooth if $h(\cdot)$ is $[p]$ times continuously differentiable and $h^{([p])}$ satisfies a Hölder condition

$$\left|h^{([p])}(x) - h^{([p])}(y)\right| \leq c\,|x - y|^{\alpha}$$

for some constant c. Essentially, the Hölder class is the class of functions that admits a Taylor expansion with a 'well behaved' remainder term. Define a Hölder ball with smoothness $p = [p] + \alpha$ as

$$\Lambda_c^p = \left\{ h \in C^{[p]}(\mathcal{X}) : \sup_{j \leq [p]} \sup_{x \in \mathcal{X}} \left|h^{(j)}(x)\right| \leq c,\ \sup_{x,y \in \mathcal{X}} \frac{\left|h^{([p])}(x) - h^{([p])}(y)\right|}{|x - y|^{\alpha}} \leq c \right\}$$

The Hölder class can be approximated well by various linear sieves. A sieve is called a "finite dimensional linear sieve" if it is a linear span of finitely many known basis functions. Some examples of finite dimensional linear sieves on [0,1] are given as follows.

Polynomials:

$$Pol(J_n) = \left\{ \sum_{k=0}^{J_n} a_k x^k, x \in [0,1] : a_k \in \mathbb{R} \right\}$$

A justification for the polynomial basis is the following famous theorem:

Theorem 3.2.1 *(Weierstrass) Let $m(\cdot)$ be a continuous bounded function with compact domain $\mathcal{X} \subset \mathbb{R}$. Then for any $\varepsilon > 0$, there exists a polynomial function $P(x) = \sum_{k=0}^{\infty} a_k x^k$ such that*

$$\sup_{x \in \mathcal{X}} |m(x) - P(x)| < \varepsilon.$$

The Weierstrass Approximation Theorem shows that the continuous real-valued functions on a compact interval can be uniformly approximated by polynomials. In other words, the polynomials are uniformly dense in $C(\mathcal{X}; R)$ with respect to the sup-norm.

Trigonometric polynomials. Let $\text{TriPol}(J_n)$ denote the space of trigonometric polynomials on $[0, 1]$ of degree J_n or less. That is,

$$\text{TriPol}(J_n) = \left\{ a_0 + \sum_{k=1}^{J_n} [a_k \cos(2k\pi x) + b_k \sin(2k\pi x)],\ x \in [0,1] :\ a_k, b_k \in \mathcal{R} \right\}.$$

Theorem 3.2.2 *The Fourier series of a continuous, piecewise smooth function f (on $[0, 1]$) converges uniformly to f.*

Corollary 3.2.1 *The set of all trigonometric polynomials are uniformly dense in $C_{per}([0; 2\pi]; R)$, the set of continuous 2π periodic function.*

The Weierstrass approximation theorem and the above theorem (and corollary) are special cases of the general Stone–Weierstrass theorem:

Theorem 3.2.3 *Suppose X is a compact Hausdorff space and A is a subalgebra of $C(X, R)$ which contains a non-zero constant function. Then A is dense in $C(X, R)$ if and only if it separates points.*

This theorem and its proof can be found in standard textbooks on real analysis.

3.2.2 L_p and Finite Dimensional Linear Sieves

In nonparametric and semiparametric econometrics, sometimes the parameters of interest are functions with unbounded supports. Here we present a finite-dimensional linear sieve that can approximate functions with unbounded supports well. In the following we let $L_p(\mathcal{X}, \omega)$, $1 \leq p < \infty$ denote the space of real-valued functions h such that $\int_{\mathcal{X}} |h(x)|^p \omega(x) dx < \infty$ for a smooth weight function ω: $\mathcal{X} \to (0, \infty)$.

Hermite polynomials Hermite polynomial series $\{H_k : k = 1, 2, ...\}$ is an orthonormal basis of $L_2(\mathcal{X}, \omega)$ with $\omega(x) = \exp\{-x^2\}$. It can be obtained by applying the Gram-Schmidt procedure to the polynomial series $\{x^{k-1} : k = 1, 2, ...\}$ under the inner product $\langle f, g\rangle_\omega = \int_{\mathcal{R}} f(x)g(x)\exp\{-x^2\}dx$. That is, $H_1(x) = 1/\sqrt{\int_{\mathcal{R}} \exp\{-x^2\}dx} = \pi^{-1/4}$, and for all $k \geq 2$,

$$H_k(x) = \frac{x^{k-1} - \sum_{j=1}^{k-1} \langle x^{k-1}, H_j\rangle_\omega H_j(x)}{\sqrt{\int_{\mathcal{R}}[x^{k-1} - \sum_{j=1}^{k-1} \langle x^{k-1}, H_j\rangle_\omega H_j(x)]^2 \exp\{-x^2\}dx}}.$$

Let $\text{HPol}(J_n)$ denote the space of Hermite polynomials on $\mathcal{R}$ of degree J_n or less:

$$\text{HPol}(J_n) = \left\{ \sum_{j=0}^{J_n} a_j H_j(x) \exp\left(-\frac{x^2}{2}\right), a_j \in \mathbb{R} \right\}$$

where

$$H_j(x) = (-1)^j \exp\left(\frac{x^2}{2}\right) \frac{d^j}{dx^j}\left[\exp\left(-\frac{x^2}{2}\right)\right].$$

Any function in $L_2(\mathcal{X}) = \left\{h(x) : \int h^2(x)\,dx < \infty\right\}$ can be approximated well by the $Hpol(J_n)$ sieve as $J_n \to \infty$. Intuitively, when $\int h^2(x)\,dx < \infty$, we have $\int \left[h(x)/\sqrt{\omega(x)}\right]^2 \omega(x)\,dx < \infty$ so $h(x)/\sqrt{\omega(x)} \in L_2(\mathcal{X}, \omega)$. Any function in $L_2(\mathcal{X}, \omega)$ can be represented by the series expansion in $\{H_j(x)\}$. As a result,

$$h(x) = \sum_{j=0}^{\infty} a_j H_j(x) \exp\left(-\frac{x^2}{2}\right) \text{ a.e.}$$

When the $\text{HPol}(J_n)$ sieve is used to approximate an unknown $\sqrt{f(x)}$ over $\mathbb{R}$, the corresponding sieve MLE is also called the semi-nonparametric (SNP) estimator in econometrics; See, e.g. Gallant and Nychka (1987) and Gallant and Tauchen (1989). More precisely,

$$f(x) \doteq \left[\sum_{j=0}^{J_n} a_j H_j(x) \exp\left(-\frac{x^2}{2}\right)\right]^2$$

3.2.3 Other Smoothness Classes and Finite Dimensional Nonlinear Sieves.

Nonlinear sieves can also be used for sieve estimation. A popular class of nonlinear sieves in econometrics is single hidden layer feedforward Artificial Neural Networks (ANN). Here we present a typical ANN.

Sigmoid ANN (Artificial Neural Networks):

$$sANN(J_n) = \left\{ b_0 + \sum_{j=1}^{J_n} b_j \psi\left(a_{j0} + a_{j1} x\right),\ a_{j,0}, a_{j,1}, b_j \in \mathbb{R} \right\},$$

where ψ is a sigmoid function — a bounded function with $\psi(u) \to 1$ as $u \to \infty$ and $\psi(u) \to 0$ as $u \to -\infty$. For example, ψ is the logit CDF. See Chen (2006) for more examples of ANNs.

$$\begin{array}{ccccccc} & \nearrow & a_{10} + a_{11}x & \to\to & \psi\left(a_{10} + a_{11}x\right) & \searrow & \\ x & \to & \ldots & \to\to & \ldots & \to & b_0 + \sum_{j=1}^{J_n} b_j \psi\left(a_{j0} + a_{j1}x\right) \\ & \searrow & a_{J0} + a_{J1}x & \to\to & \psi\left(a_{J0} + a_{J1}x\right) & \nearrow & \end{array}$$

Neural Networks can provide flexible approximations to nonlinear functions in certain class of functions. Let $\mathcal{X}$ be a compact set in $\mathcal{R}^d$, and $C(\mathcal{X})$ be the space of continuous functions mapping from $\mathcal{X}$ to $\mathcal{R}$. Gallant and White (1988a) first established that the sANN sieve with the cosine squasher activation function is dense in $C(\mathcal{X})$ under the sup-norm. Cybenko (1989) and Hornik et al. (1989) show that the sANN(J_n), with any sigmoid activation function, is dense in $C(\mathcal{X})$ under the sup-norm. For more details, see White (2006) in the Handbook of Economic Forecasting.

3.2.4 Infinite Dimensional Sieves

Most commonly used sieve spaces are finite-dimensional truncated series such as those listed above. However, the general theory on sieve estimation can also allow for infinite-dimensional sieve. For example, consider the smoothness class $\Theta = \Lambda_p(\mathcal{X})$ with $\mathcal{X} = [0, 1]$, $p > 1/2$. Then any function $\theta \in \Theta$ can be expressed as an infinite Fourier series

$$\theta(x) = \sum_{k=0}^{\infty} a_k \cos 2k\pi x + \sum_{k=0}^{\infty} b_k \sin 2k\pi x,$$

Previously, we suggest using truncation to obtain

$$\theta_{J_n}(x) = \sum_{k=0}^{J_n} a_k \cos 2k\pi x + \sum_{k=0}^{J_n} b_k \sin 2k\pi x,$$

and search estimators of the above form, and then let $J_n \to \infty$. We can also allow for the infinite expansion but restrict the Fourier coefficients a_k and b_k using certain metric:

$$\Theta_n = \{\theta \in \Theta : pen\left(\theta\right) \leq c_n\}$$

with $c_n \to \infty$ as $n \to \infty$ arbitrarily slowly; see e.g. Shen (1997). Here $pen(\theta)$ is a metric or a penalty. Typically $pen(\theta)$ is large if the a_k and b_k are large for large k.

When the objective function is concave and $pen(\theta)$ is convex, the sieve estimator $\arg\min_{\theta\in\Theta_n} \hat{Q}_n(\theta)$ is equivalent to the penalized extremum estimator:

$$\max_{\theta\in\Theta}\left[\hat{Q}_n(\theta) - \lambda_n pen(\theta)\right]$$

where the Lagrangian multiplier λ_n is chosen such that the solution satisfies $\text{pen}(\hat{\theta}) = c_n$. See Eggermont and LaRiccia (2001, Sec 1.6). Adding a penalty term to the criterion we are optimizing is sometimes called **regularization**. λ_n is often called as the regularization parameter.

3.2.5 Tensor product spaces.

Let $\mathcal{U}_\ell$, $1 \le \ell \le d$, be compact sets in Euclidean spaces and $\mathcal{U} = \mathcal{U}_1 \times ... \times \mathcal{U}_d$ be their Cartesian product. Let $\mathbb{G}_\ell$ be a linear space of functions on $\mathcal{U}_\ell$ for $1 \le \ell \le d$, each of which can be any of the sieve spaces described above, among others. The tensor product, $\mathbb{G}$, of $\mathbb{G}_1, \ldots, \mathbb{G}_d$ is defined as the space of functions on $\mathcal{U}$ spanned by the functions $\prod_{\ell=1}^d g_\ell(x_\ell)$, where $g_\ell \in \mathbb{G}_\ell$ for $1 \le \ell \le d$. We note that $\dim(\mathbb{G}) = \prod_{\ell=1}^d \dim(\mathbb{G}_\ell)$. Tensor-product construction is a standard way to generate linear sieves of multivariate functions from linear sieves of univariate functions.

3.3 Conditional Moment Estimation: Splines

Consider a regression model of the form

$$Y = m(X) + \varepsilon, E[\varepsilon|X] = 0.$$

We restrict our attention to the univariate case.

3.3.1 Penalized OLS and Cubic Spline

Following the previous section, we consider the penalized sum of squares:

$$R_\lambda(m) = \frac{1}{n}\sum_{i=1}^n [Y_i - m(X_i)]^2 + \lambda J(m) \tag{3.2}$$

where $J(m)$ is some roughness penalty. With the loss of generality, we assume the data are ordered such that $X_1 \le X_2 \le \ldots \le X_n$. The rougher the function is, the higher the penalty is. The parameter λ controls the severity of the penalty. Let $\hat{m}_\lambda$ be the resulting function that minimizes $R_\lambda(m)$

$$\hat{m} = \arg \min_{m(\cdot)\in W_2^2[a,b]} R_\lambda(m), \tag{3.3}$$

where

$$W_k^p(\mathcal{X}) = \{f \in L^p(\mathcal{X}) : D^\alpha f \in L^p(\mathcal{X}), \text{ for } \|\alpha\|_1 \leq k\}.$$

is the Sobolev space with $D^\alpha f$ being the α-th weak derivative of f.

We will focus on the special case that $J(\hat{m}) = \int \hat{m}''(u)^2 du$. In macroeconomics, the following $J(\hat{m})$ is often used:

$$J(\hat{m}) = \sum_{i=2}^{n} [\hat{m}(X_{i-1}) - 2\hat{m}(X_i) + \hat{m}(X_{i+1})]^2,$$

which can be regarded as a discrete version of $J(\hat{m}) = \int \hat{m}''(u)^2 du$. The resulting estimator $\hat{m}(x)$ is the Hodrick-Prescott filter estimator. There are more general penalty functions in the statistics and econometrics. Gu (2002) embeds $m(x)$ in a Reproducing Kernel Hilbert Space (RKHS) and uses the norm in that RKHS as the penalty. The theory of RKHS is beautiful and we will briefly discuss it later.

When λ is larger, the roughness penalty is higher and the resulting $\hat{m}(x)$ is less rough (or more smooth). On the other hand, when λ is smaller, the roughness penalty is lower and the resulting $\hat{m}(x)$ is more rough. In the extreme case, as $\lambda \to 0$, $\hat{m}_\lambda$ interpolates the observations. When $J(\hat{m}) = \int \hat{m}''(u)^2 du$, $\hat{m}(X_i)$ becomes the linear least square fit as $\lambda \to \infty$.

In general, what does $\hat{m}_\lambda$ look like? To answer this question, we need to define splines. Spline is a special piecewise polynomial, the most commonly used splines are piecewise cubic splines:

Definition 3.3.1 *Let $\xi_1 < \xi_2 < ... < \xi_k$ be a set of ordered points – called knots– contained in some interval $[a, b]$. A cubic spline is a continuous function $m(\cdot)$ such that (i) $m(\cdot)$ is a cubic polynomial between two successive ξ values, (ii) m, m'_λ, and m''_λ are continuous at the knots. A spline that is linear beyond the boundary knots is called a natural cubic spline (NCS).*[1]

Theorem 3.3.1 *The function $\hat{m}(x)$ that minimizes $R_\lambda(\hat{m})$ in (3.2) with $J(\hat{m}) = \int \hat{m}''(u)^2 du$ is a natural cubic spline with knots at the data points. The estimator $\hat{m}_\lambda$ is called a smoothing spline.*

Note: let $\hat{Y}_i = \hat{m}(X_i)$, then by definition NCS $\hat{m}(x)$ interpolates the points $\left(X_i, \hat{Y}_i\right)$. Among all the functions that interpolate these points, the NCS has the smallest norm, i.e., $J(\cdot)$.

Proof. The proof is based on calculus of variations. We will give a different proof using the theory of reproducing kernel Hilbert space in the next section. Consider a small perturbation

[1] The term spline is adopted from the name of a flexible strip of metal commonly used by drafters to assist in drawing curved lines.

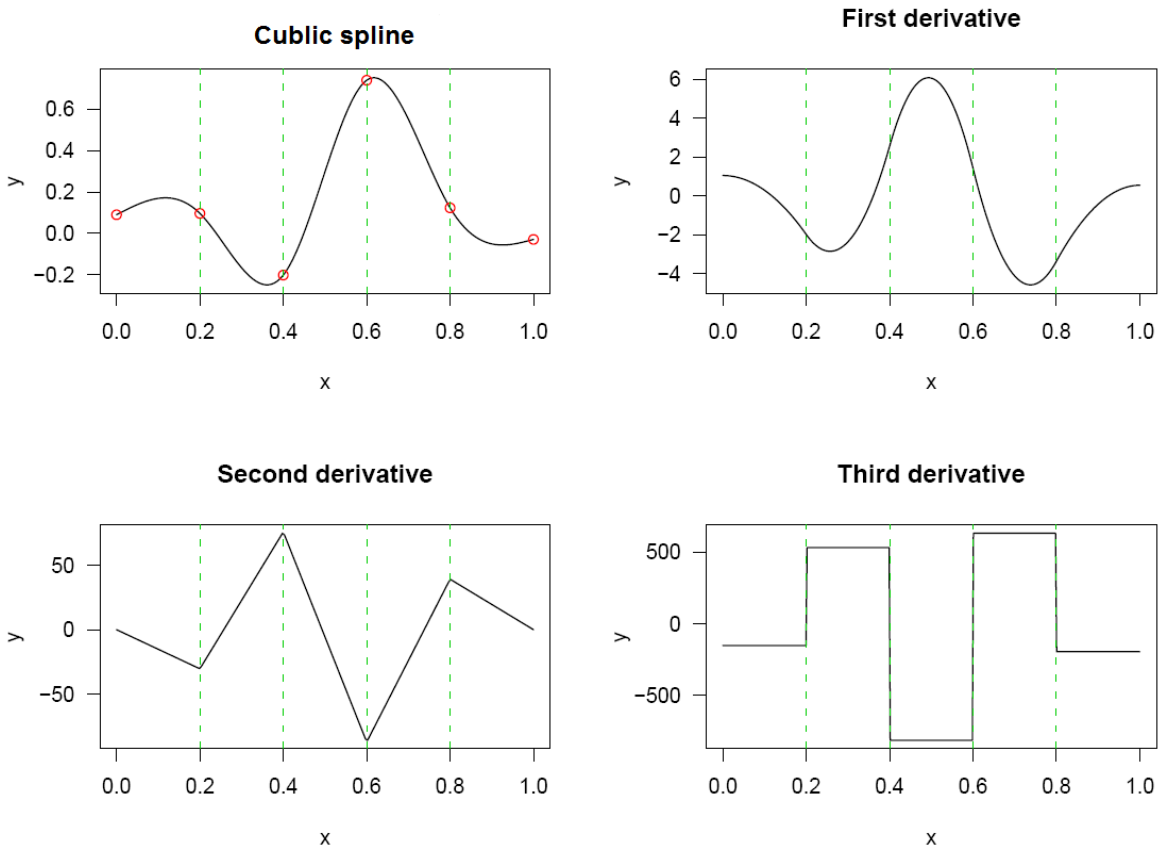


Figure 3.1: Natural Cublic Spline

$\tilde{m}(x) = \hat{m}(x) + \delta\eta(x)$ for some 'nicely behaved' function $\eta(x)$. Then

$$R_\lambda(\tilde{m}) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(X_i) - \delta\eta(X_i)\right]^2 + \lambda\int\left[\hat{m}''(u) + \delta\eta''(u)\right]^2 du,$$

and

$$\frac{\partial R_\lambda(\tilde{m})}{\partial\delta} = -\frac{2}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(X_i) - \delta\eta(X_i)\right]\eta(X_i) + 2\lambda\int\left[\hat{m}''(u) + \delta\eta''(u)\right]\eta''(u)du.$$

So

$$\frac{\partial R_\lambda(\tilde{m})}{\partial\delta}|_{\delta=0} = -\frac{2}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}(X_i)\right]\eta(X_i) + 2\lambda\int\hat{m}''(u)\,\eta''(u)du = 0.$$

If $\hat{m}(\cdot)$ is the optimal function, then it is necessary that

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\hat{m}(X_i)\right]\eta(X_i)=\lambda\int_a^b\hat{m}''(u)\,\eta''(u)du \text{ for all } \eta(\cdot).$$

Let $\eta(u)=\eta'(u)=0$ when $u=a$ or b, then integration by part gives

$$\begin{aligned}\int_a^b\hat{m}''(u)\,d\eta'(u) &= \hat{m}''(u)\,\eta'(u)\,|_a^b-\int_a^b\eta'(u)\,\hat{m}'''(u)du\\ &= -\int_a^b\hat{m}'''(u)d\eta(u)=\hat{m}'''(u)\eta(u)\,|_a^b+\int_a^b\eta(u)\,d\hat{m}'''(u)\\ &= \int_a^b\hat{m}^{(4)}(u)\eta(u)\,du=\sum_{i=0}^{n}\int_{X_i}^{X_{i+1}}\hat{m}^{(4)}(u)\eta(u)\,du,\end{aligned}$$

where we define $X_{n+1}=b$ and $X_0=a$. Combining the above two equations, we get

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\hat{m}(X_i)\right]\eta(X_i)=\lambda\sum_{i=0}^{n}\int_{X_i}^{X_{i+1}}\hat{m}^{(4)}(u)\eta(u)\,du$$

for all $\eta(u)$ such that $\eta(u)=\eta'(u)=0$ when $u=a$ or b. Let $\eta(u)$ be such a function that concentrates on $[X_i,X_{i+1})$, that is, $\eta(u)=0$ for all $u\notin[X_i,X_{i+1})$. Then

$$\frac{1}{n}\left[Y_i-\hat{m}(X_i)\right]\eta(X_i)=\lambda\int_{X_i}^{X_{i+1}}\hat{m}^{(4)}(u)\eta(u)\,du$$

where $\hat{m}^{(4)}$ is the 4-th order weak derivative of $\hat{m}$. As a result

$$\hat{m}^{(4)}(u)=\frac{1}{n\lambda}\left[Y_i-\hat{m}(X_i)\right]\delta_{X_i}(u) \text{ for } u\in[X_i,X_{i+1})$$

where $\delta_{X_i}(\cdot)$ is the Dirac delta function. Thus a solution takes the form

$$\begin{aligned}\hat{m}(x) &= a_i+b_i(x-X_i)+c_i(x-X_i)^2+d_i(x-X_i)^3\\ \text{for } x &\in [X_i,X_{i+1}),\ i=0,1,2,...,n-1.\end{aligned}\tag{3.4}$$

■

The following counting exercise may serve to clarify the essentially finite dimensional nature of the estimator $\hat{m}(\cdot)$. We have $(n-1)$ intervals with 4 parameters on each interval. So there are $4(n-1)$ parameters for the interior regions. If we wish to extend $\hat{m}(\cdot)$ beyond X_1 and X_n, we have another 4 parameters, say, a_0,b_0, a_n,b_n. Note that the c's and d's in these outer regions are zero; were they not, the roughness penalty could be reduced without

disturbing the value of the first and the last terms in the objective function. In total, there are $4n$ parameters. Now at each of the design points, X_i, we have the following continuity restrictions: $m^{(k)}(X_i+) = m^{(k)}(X_i-)$ for $k = 0, 1, 2$. So there are $3n$ such constraints in total and we are left with a problem in n parameters.

We can reparametrize (3.4) as

$$\begin{aligned}\hat{m}(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\ &\quad + \sum_{j=1}^{n} \theta_{0j} 1\{x \geq X_j\} + \sum_{j=1}^{n} \theta_{1j} (x - X_j)_+ + \sum_{j=1}^{n} \theta_{2j} (x - X_j)_+^2 + \sum_{j=1}^{n} \theta_j (x - X_j)_+^3\end{aligned}$$

where $a_+ = a1\{a > 0\} = \max(a, 0)$. The coefficients $\theta_{0j}, \theta_{1j}, \theta_{2j}$ and θ_{3j} record the jumps in the different derivatives at the knots. Since $m^{(k)}(X_j+) = m^{(k)}(X_j-)$ for $k = 0, 1, 2$, we have $\theta_{0j} = \theta_{1j} = \theta_{2j} = 0$ for all $j = 1, 2, ..., n$. As a result, we can write $\hat{m}(x)$ as

$$\hat{m}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{n} \theta_j (x - X_j)_+^3 \tag{3.5}$$

where $a_+ = a1\{a > 0\} = \max(a, 0)$. This is a general representation of a cubic spline.

For a natural cubic spline, four constraints are imposed on the coefficients $\{\beta_i\}$ and $\{\theta_j\}$. More specifically, when $x \leq X_1$, we have

$$\hat{m}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

By the linearity of $\hat{m}(x)$ for $x \leq X_1$, we have $\hat{m}''(x) = 0$, which implies that $\beta_2 = \beta_3 = 0$. So

$$\hat{m}(x) = \beta_0 + \beta_1 x + \sum_{j=1}^{n} \theta_j (x - X_j)_+^3. \tag{3.6}$$

When $x \geq X_n$, we have

$$\hat{m}(x) = \beta_0 + \beta_1 x + \sum_{j=1}^{n} \theta_j (x - X_j)^3 \text{ and so } \hat{m}''(x) = 6 \sum_{j=1}^{n} \theta_j (x - X_j).$$

By the linearity of $\hat{m}(x)$ for $x \geq X_n$, we have $\hat{m}''(x) = 0$ and so

$$\sum_{j=1}^{n} \theta_j = 0, \sum_{j=1}^{n} \theta_j X_j = 0.$$

To sum up, $\hat{m}$ can be represented by (3.6) subject to the above two constraints. There are $n + 2$ parameters with 2 constraints.

It can be shown that $\hat{m}(x)$ can be represented by n basis functions:

$$N_1(x) = 1,\ N_2(x) = x, N_{k+2} = d_k(x) - d_{n-1}(x),\ k = 1, 2, \ldots, n-2 \tag{3.7}$$

where

$$d_k(x) = \frac{(x - X_k)_+^3 - (x - X_n)_+^3}{X_n - X_k}.$$

That is

$$\hat{m}(x) = \sum_{k=1}^{n} \gamma_k N_k(x) \tag{3.8}$$

without any restrictions on the γ_k's. By construction, $N_k(x) = 0$ for $x \leq X_1$ and $N_k(x)$ is linear in x for $x \geq X_n$. That is to say, each of these basis functions has zero 2nd and 3rd derivative outside the boundary knots.

Given the above representation, the original $R_\lambda(m) := R_\lambda(\gamma)$ becomes

$$R_\lambda(\gamma) = \frac{1}{n}\|Y - N\gamma\|^2 + \lambda\gamma^T\Omega\gamma$$

where $N = (N_{ij}) = (N_j(X_i))$ and $\{\Omega_{ij}\} = \left\{\int N_j''(x) N_k''(x)\, dx\right\}$. The solution is easily seen to be

$$\hat{\theta} = \left(N'N + n\lambda\Omega\right)^{-1} N'Y.$$

Since $\hat{m}$ is a linear smoother, the model selection criterion C_p can be used to determine the smoothing parameter λ.

The fitted values at $\{X_i\}$ is

$$\hat{Y} = N\left(N'N + n\lambda\Omega\right)^{-1} N'Y = \left(I + n\lambda\tilde{\Omega}\right)^{-1} Y \text{ for } \tilde{\Omega} = \left(N'\right)^{-1}\Omega\left(N\right)^{-1}$$

Let $\tilde{\Omega} = \sum_{k=1}^{n} d_k u_k u_k'$ be the eigen decomposition of $\tilde{\Omega}$, then

$$\hat{Y} = \sum_{k=1}^{n} \frac{1}{1 + nd_k} u_k\left(u_k'Y\right).$$

Note that

$$Y = \sum_{k=1}^{n} u_k\left(u_k'Y\right).$$

which is the representation of Y using the complete bases $\{u_k\}$. Comparing the above two equations, we see that the smoothing spline operates by shrinking the coefficients $(u_k'Y)$ using $(1 + n\lambda d_k)^{-1} \leq 1$.

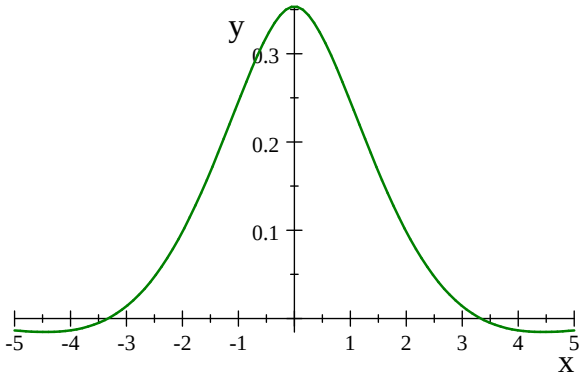


Figure 3.2: Equivalent kernel for cubic smoothing splines

Although $\hat{m}_\lambda$ is linear in the Y data, its dependency on the design is rather complicated. This has resulted in rather less treatment of the statistical properties of these estimators, except in rather simple settings, although see Wahba (1990)— in fact, the extension to multivariate design is not straightforward. However, splines are asymptotically equivalent to kernel smoothers as Silverman (1984) showed. The equivalent kernel is

$$K(u) = \frac{1}{2}\exp\left(-\frac{|u|}{\sqrt{2}}\right)\sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right)$$

which is of fourth order while the equivalent bandwidth is

$$h = h(\lambda; x) = \left(\frac{\lambda}{n f_X(x)}\right)^{1/4}.$$

One advantage of the spline estimators over kernels is that global inequality and equality constraints can be imposed more conveniently. For example, it may be desirable to restrict the function to pass through a particular point, see Jones (1985).

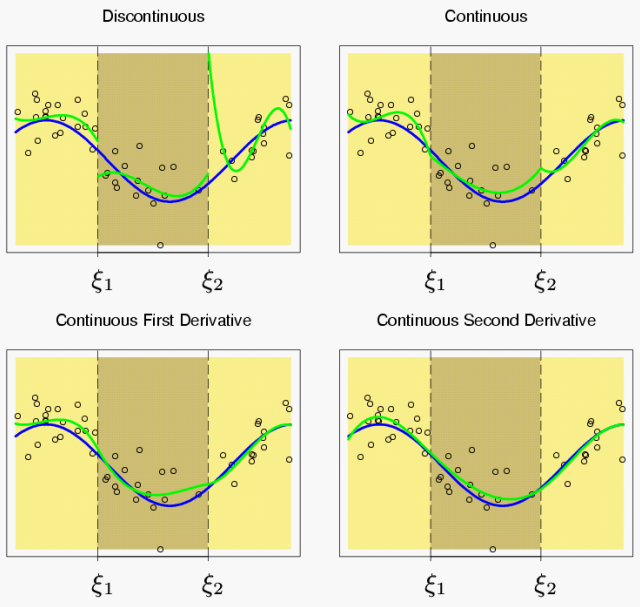
3.3.2 Regression Spline

Another popular usage of spline is the so called regression spline. Instead of placing knots at every data point, regression spline method places knots at some selected points, usually at selected quantiles (i.e. the terciles, or quartiles, or quintiles, depending on how many knots you want). A smarter strategy would place more knots in regions where $m(x)$ is changing more rapidly. Knot placement is an arcane art form, which is often cited as the first disadvantage of regression splines.

For a given set of knots $(\xi_1, ..., \xi_J)$, regression cubic spline method approximates $m(x)$ by a series expansion:

$$m(x) \doteq \sum_{j=0}^{3} \beta_j x^j + \sum_{j=1}^{J} \theta_j \left(x - \xi_j\right)_+^3$$

In Section 3.6, we consider general expansions that include the above expansion as a special case. Figure 3.3.2 illustrates the fit by regression cubic spline with $J = 2$.



3.4 Conditional Moment Estimation: RK Methods

3.4.1 Hilbert Space and Reproducing Kernel Hilbert Space

Definition 3.4.1 *A Hilbert space $\mathcal{H}$ is a vector space endowed with an inner product and associated norm and metric, such that every Cauchy sequence in $\mathcal{H}$ has a limit in $\mathcal{H}$,i.e., $\mathcal{H}$ is complete.*

Example 3.4.1 *R^n, $L^2[a, b]$ with usual inner product are Hilbert spaces*

Example 3.4.2 *Sobolev space*

$$W_q[0,1] := W_q^2[0,1] = \left\{ f : f^{(1)}, ..., f^{(q-1)} \text{ are absolutely continuous and } \int_0^1 \left(f^{(q)}(u)\right)^2 du < \infty \right\}$$

with inner product

$$\langle f, g\rangle_S = \sum_{v=0}^{q-1} f^{(v)}(0)\, g^{(v)}(0) + \int_0^1 f^{(q)}(u)\, g^{(q)}(u)\, du$$

is a Hilbert space.

Note: (i) A function f is absolutely continuous on $[0,1]$, if it has a derivative f' almost everywhere; the derivative is Lebesgue integrable; and

$$f(x) = f(0) + \int_0^x f'(u)\, du \text{ for all } x \in [0,1].$$

(ii) For any function $h \in W_q[0,1]$, we have

$$h(x) = \sum_{v=0}^{q-1} \frac{h^{(v)}(0)\, x^v}{v!} + \int_0^1 G_q(x,u)\, h^{(q)}(u)\, du,$$

where

$$G_q(x,u) = \frac{(x-u)_+^{q-1}}{(q-1)!}.$$

This is simply the Taylor expansion with the integral form of the remainder (Lagrange Remainder).

(iii) The induced norm by the inner product is

$$\|f\|_S^2 = \sum_{v=0}^{q-1} \left[f^{(v)}(0)\right]^2 + \int_0^1 \left[f^{(q)}(u)\right]^2 du$$

(iv) Let $\phi_k(x) = \sqrt{2}\sin 2\pi kx$. The L_2 norm of ϕ_k is

$$\|\phi_k\|_2 = \left[2\int_0^1 (\sin 2\pi kx)^2\, dx\right]^{1/2} = 1.$$

But for $q = 1$,

$$\|\phi_k\|_S^2 = \int_0^1 \left(\frac{d}{dx}\left(\sqrt{2}\sin 2\pi kx\right)\right)^2 dx = 4\pi^2 k^2;$$

and for $q = 2$,

$$\begin{aligned}\|\phi_k\|_S^2 &= \left(2\sqrt{2}\pi k\right)^2 + \int_0^1 \left(\frac{d}{dx^2}\left(\sqrt{2}\sin 2\pi kx\right)\right)^2 dx \\ &= 8\pi^2 k^2 + 16\pi^4 k^4.\end{aligned}$$

While the L_2 norm of $\phi_k(x)$ is a constant regardless of the value of k, the Sobolev norm of $\phi_k(x)$ increases with k. Essentially, the L_2 norm measures only the "height" and "width" of a function. The Sobolev norm measures the "height" and "width" and the frequency scale of a function, i.e., how quickly the function oscillates. As k increases, ϕ_k oscillates more often, leading to a larger Sobolev norm.

Definition 3.4.2 *A functional $L(\cdot) : \mathcal{H} \to R$ is bounded if there exists a constant M such that $|L(h)| \leq M \|h\|_{\mathcal{H}}$ for all $h \in \mathcal{H}$. Note that the constant M does not depend on h.*

Example 3.4.3 *$\mathcal{H} = L^2[a,b]$ and $L(h) = L_g(h) := \int_a^b h(u)g(u)du$ for some $g \in L^2[a,b]$. $L_g(\cdot)$ is a linear and bounded functional. However, $L(h) = L_x(h) = h(x)$ for some $x \in R$ is not a bounded functional as the evaluation functional is not well-defined on $L^2[a,b]$.*

Example 3.4.4 *$\mathcal{H} = W_q[0,1]$ and $L(h) = L_x(h) = h(x)$ for some $x \in R$. $L_x(\cdot)$ is a linear and bounded functional. The boundedness follows because*

$$\begin{aligned}
|L_x(h)| &= |h(x)| = \left| \sum_{v=0}^{q-1} \frac{h^{(v)}(0)\, x^v}{v!} + \int_0^1 G_q(x,u) h^{(q)}(u)\, du \right| \\
&\leq \left(\sum_{v=0}^{q-1} h^{(v)}(0)^2 \right)^{1/2} \left(\sum_{v=0}^{q-1} \left(\frac{x^v}{v!} \right)^2 \right)^{1/2} \\
&\quad + \left(\int \left(h^{(q)}(u) \right)^2 du \right)^{1/2} \left\{ \int_0^1 G_q^2(x,u) du \right\}^{1/2} \\
&\leq M_x \left(\sum_{v=0}^{q-1} h^{(v)}(0)^2 + \int_0^1 \left(h^{(q)}(u) \right)^2 du \right)^{1/2} \\
&\leq M_x \|h\|_S \quad \textit{for some constant } M_x \textit{ depending on } x \textit{ but not on } h.
\end{aligned}$$

Definition 3.4.3 *A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space on which all evaluation functionals are bounded (or continuous), i.e., $L_x(\cdot)$ is a bounded functional for all $x \in \mathcal{X}$, the domain of the functions in this space.*

Note that the first part of the penalized LS regression depends on $m(\cdot)$ evaluated at the data points $X_1, ..., X_n$. It seems natural to assume that $L_x(m) = m(x)$ to be a continuous functional for each x.

According to this definition, R^n and $\mathcal{H} = W_q[0,1]$ are RKHS's but $L^2[a,b]$ is not. To motivate the above definition using $W_q[0,1]$ as an example, we need the Riesz representation theorem. We will also use this theorem frequently in later chapters.

Theorem 3.4.1 (Riesz representation theorem) *For every bounded linear functional L on a Hilbert space $\mathcal{H}$, there exists a unique $g \in \mathcal{H}$ such that $L(h) = \langle g, h \rangle$ for any $h \in \mathcal{H}$.*

Proof. Define the null space of L as $N_L = \{h : Lh = 0\} = L^{-1}\{0\}$. Because L is continuous, N_L is a closed linear subspace of $\mathcal{H}$. If $N_L = \mathcal{H}$, then $L(h) = 0$ for any $h \in \mathcal{H}$ and we can take $g = 0$ be the representator. Otherwise, N_L is a proper subspace of $\mathcal{H}$ and there exists $g_0 \in \mathcal{H} \ominus N_L$ (i.e., $g_0 \in \mathcal{H}$ but $g_0 \notin N_L$ and $\langle g_0, n\rangle_{\mathcal{H}} = 0$ for any $n \in N_L$. Without the loss of generality, we can assume that $\|g_0\|_{\mathcal{H}} = 1$. Note that

$$(Lh)\, g_0 - (Lg_0)\, h \in N_L,$$

and so

$$\langle (Lh)\, g_0 - (Lg_0)\, h, g_0\rangle_{\mathcal{H}} = 0.$$

Rearranging this yields

$$(Lh) = \langle h, (Lg_0)\, g_0\rangle_{\mathcal{H}} = \langle h, g\rangle_{\mathcal{H}}$$

where we take $g = (Lg_0)\, g_0$. The uniqueess obviously holds. ■

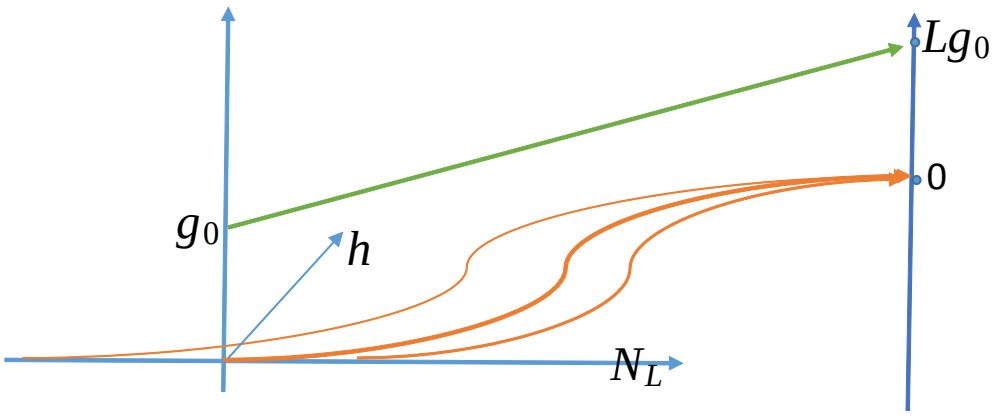


Figure 3.3: Geometry Behind the Riesz Representation Theorem

By the Riesz representation theorem, there exists $K_x(\cdot) \in W_q[0,1]$, the representer of $L_x(\cdot)$, such that $\langle K_x(\cdot),\ f(\cdot)\rangle_S = f(x)$, $\forall f \in W_q[0,1]$. That is

$$f(x) = \sum_{v=0}^{q-1} f^{(v)}(0)\, K_x^{(v)}(0) + \int_0^1 f^{(q)}(u)\, K_x^{(q)}(u)\, du$$

for a fixed $x \in [0,1]$. But

$$f(x) = \sum_{v=0}^{q-1} f^{(v)}(0)\, \frac{x^v}{v!} + \int_0^1 f^{(q)}(u)\, G_q(x,u)\, du.$$

Comparing the above two equations, we obtain

$$K_x^{(v)}(0) = \frac{x^v}{v!} \text{ and } K_x^{(q)}(u) = G_q(x,u).$$

Plugging these into the definition of $K_x(y)$ yields:

$$\begin{aligned} K_x(y) &= \sum_{v=0}^{q-1} K_x^{(v)}(0) \frac{y^v}{v!} + \int_0^1 K_x^{(q)}(y) G_q(y,u)\, du \\ &= \sum_{v=0}^{q-1} \frac{x^v}{v!}\frac{y^v}{v!} + \int_0^1 G_q(x,u) G_q(y,u)\, du. \end{aligned} \tag{3.9}$$

We write $K_x(y) = K(x,y)$. Then

$$\langle K(x,\cdot), f(\cdot)\rangle_S = f(x), \;\; \langle K(x,\cdot), K(\cdot,y)\rangle_S = K(x,y) = K(y,x).$$

The above properties are the reproducing properties. We call $K(x,y)$ the reproducing kernel.

The reproducing kernel is clearly continuous and symmetric. It is also positive-semidefinite. For any a_i and x_i

$$\begin{aligned} &\sum_{i=1}^n \sum_{j=1}^n a_i K(x_i,x_j) a_j \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i \langle K(x_i,\cdot), K(\cdot,x_j)\rangle_S a_j = \left\| \sum_{i=1}^n a_i K(x_i,\cdot) \right\|_S^2 \geq 0. \end{aligned}$$

Alternatively, for any $f(x) \in L^2[0,1]$, we have

$$\begin{aligned} \int_0^1 \int_0^1 f(x) K(x,y) f(y)\, dxdy &= \int_0^1 \int_0^1 f(x) \langle K(x,\cdot), K(\cdot,y)\rangle_S f(y)\, dxdy \\ &= \left\langle \int_0^1 f(x) K(x,\cdot)\, dx, \int K(\cdot,y) f(y)\, dy \right\rangle_S \\ &= \left\| \int_0^1 f(x) K(x,\cdot)\, dx \right\|_S^2 \geq 0. \end{aligned}$$

So we have found a continuous, symmetric, and positive semi-definite kernel for the space $W_q[0,1]$. This kernel reproduces the function in $W_q[0,1]$ in the sense that $\langle K(x,\cdot), f(\cdot)\rangle_S = f(x)$ for any x.

Next, we show that the functions in $W_q[0,1]$ are generated by the kernel function. To this end, we need Mercer's theorem.

Theorem 3.4.2 (Mercer's theorem) *Suppose that K is a continuous, symmetric, and positive semi-definite kernel on $[0,1]^2$. Then there exists orthonormal bases $\phi_j(\cdot) \in L^2[0,1]$ such that*

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\, \phi_j(y)$$

where the rhs converges absolutely and uniformly over $(x,y) \in [0,1]^2$ and $\{\phi_j\}$ and $\{\lambda_j\}$ are eigen functions and eigenvalues satisfying:

$$\lambda_j \phi_j(x) = \int_0^1 K(x,y)\phi_j(y)\, dy. \tag{3.10}$$

Note: when continuity of the kernel is not assumed, the expansion no longer converges uniformly.

Using the representation in the Mercer's theorem, we have

$$\begin{aligned} f(x) &= \left\langle \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\, \phi_j(y), f(y) \right\rangle_S = \sum_{j=1}^{\infty} \lambda_j \phi_j(x) \left\langle \phi_j(y), f(y) \right\rangle_S \\ &= \sum_{j=1}^{\infty} a_j \phi_j(x) \text{ for } a_j = \lambda_j \left\langle \phi_j(y), f(y) \right\rangle_S . \end{aligned} \tag{3.11}$$

That is, the functions in $W_q[0,1]$ are spanned by the eigen functions $\{\phi_j(x)\}$ of the reproducing kernel. In essence, Mercer's theorem provides a concrete way to construct a RKHS. It provides a coordinate basis representation of a RKHS.

The above representation is the same as the usual basis expansion in a Hilbert space. It looks a little different because $\{\phi_j(x)\}$ are not an orthonormal in $W_q[0,1]$. Using (3.10), we have

$$\left\langle \phi_\ell(x), \lambda_j \phi_j(x) \right\rangle = \int_0^1 \left\langle \phi_\ell(x), K(x,y) \right\rangle \phi_j(y)\, dy = \int_0^1 \phi_\ell(y)\, \phi_j(y)\, dy = \delta_{\ell j}$$

and so $\left\langle \phi_\ell(x), \phi_j(x) \right\rangle = \delta_{\ell j}/\lambda_j$. As a result, $\{\sqrt{\lambda_j}\phi_j(x)\}$ are orthonormal bases in $W_q[0,1]$. Now the usual basis expansion is

$$f(x) = \sum_{j=1}^{\infty} \left\langle \sqrt{\lambda_j}\phi_j, f \right\rangle_S \sqrt{\lambda_j}\phi_j(x)$$

with $\sum_{j=1}^{\infty} \left\langle \sqrt{\lambda_j}\phi_j, f \right\rangle_S^2 < \infty$, which is identical to (3.11).

For $f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x) = \sum_{j=1}^{\infty} \left[a_j/\sqrt{\lambda_j}\right] \sqrt{\lambda_j}\phi_j(x)$ and $g(x) = \sum_{j=1}^{\infty} b_j \phi_j(x) = \sum_{j=1}^{\infty} \left[b_j/\sqrt{\lambda_j}\right] \sqrt{\lambda_j}\phi_j(x)$, we have

$$\langle f(x), g(x) \rangle_S = \sum_{j=1}^{\infty} \frac{a_j b_j}{\lambda_j}.$$

This inner product is different from the usual inner product in the L^2 space. Dividing by the eigenvalues in effect amounts to imposing a smoothness condition on the space. For a function to be in RKHS, the coefficients a_j must go to zero quickly so that the norm $\sum_{j=1}^{\infty} a_j^2/\lambda_j$ is finite.

To sum up, we start with $W_q[0,1]$ on which all evaluation functionals are linear and bounded. Using the Riesz representation theorem, we obtain a reproducing kernel and show that functions in $W_q[0,1]$ are spanned by the eigen functions $\{\phi_j(x)\}$ of the reproducing kernel. Our analyses apply to any Hilbert space on which all evaluation functionals are bounded.

On the other hand, given a continuous, symmetric, and positive semi-definite kernel K on $[0,1]^2$, we can construct a Hilbert space as follows:

(i) Find the eigen decomposition: $K(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x) \phi_j(y)$ where the convergence holds uniformly and absolutely.

(ii) Define the space of functions $\mathcal{H}_K = \left\{ f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x) \text{ such that } \sum_{j=1}^{\infty} a_j^2/\lambda_j < \infty \right\}$ or equivalently $\mathcal{H}_K = cls\left\{ \sum_{i=1}^{J} c_i K(x, y_i) \text{ for all finite } J \right\}$ where "cls" denotes "closed linear span". In the second definition, $\mathcal{H}_K$ contains all linear combinations of — often infinitely many— "shifts" of the kernel K.

(iii) Endow $\mathcal{H}_K$ with the inner product $\langle f(x), g(x) \rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} a_j b_j/\lambda_j$ and the norm $\|f\|_{\mathcal{H}_K}^2 = \sum_{j=1}^{\infty} a_j^2/\lambda_j$.

This space is a Reproducing Kernel Hilbert Space (RKHS). It follows from (iii) that

$$\left\langle \phi_j(x), \phi_k(x) \right\rangle_{\mathcal{H}_K} = \frac{\delta_{jk}}{\lambda_j} \text{ and so } \left\langle K(x,y), \phi_k(x) \right\rangle_{\mathcal{H}_K} = \phi_k(y).$$

As a result

$$\langle f(x), K(x,y) \rangle_{\mathcal{H}_K} = \left\langle \sum_{j=1}^{\infty} a_j \phi_j(x), K(x,y) \right\rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \lambda_j \left\langle \phi_j(x), K(x,y) \right\rangle_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \lambda_j \phi_j(y) = f(y).$$

That is, $K(x,y)$ is indeed a reproducing kernel for $\mathcal{H}_K$.

If $K(\cdot,\cdot)$ is translation invariant, i.e, $K(x,y) = \tilde{K}(x-y)$, then we can show that

$$\mathcal{H}_K = \left\{ f \in L^2[0,1] \cap C[0,1] : \frac{\mathcal{F}f}{\sqrt{\mathcal{F}\tilde{K}}} \in L^2(\mathbb{R}) \right\}$$

A function f belongs to the native space $\mathcal{H}_K$ if the decay of its Fourier transform $\mathcal{F}f$ relative to that of the Fourier transform $\mathcal{F}\tilde{K}$ of the kernel is rapid enough. This characterization encodes the kind of smoothness information embedded in $\|f\|^2_{\mathcal{H}_K} = \sum_{j=1}^{\infty} a_j^2/\lambda_j < \infty$.

3.4.2 Reproducing Kernel Methods

We are now ready to solve the smoothing spline problem using the theory of RKHS. Embedding an unknown nonparametric function in a RKHS has been very popular in the computing science literature, within the subject of machine learning. There is a huge machine learning literature on reproducing kernel methods or kernel methods (not to be confused with the local kernel smoothing method in Chapters 1 and 2). Schölkopf and Smola (2002) gives a comprehensive overview of reproducing kernel methods in machine learning research.

We first consider the problem

$$\arg \min_{m \in W_q[0,1]} \frac{1}{n} \sum_{i=1}^{n} [Y_i - m(X_i)]^2 + \lambda J(m) \tag{3.12}$$

where $J(m) = \|m\|^2_{\mathcal{H}_K}$, the squared-norm in the RKHS. The criterion function can be motivated from the Lagrange method used to solve the constrained minimization problem:

$$\arg \min_{m \in W_q[0,1]} \frac{1}{n} \sum_{i=1}^{n} [Y_i - m(X_i)]^2, \text{ st } J(m) \leq C$$

for some constant $C > 0$. Instead of imposing a rigid parametric function form restriction on $m(\cdot)$, we effectively impose a soft constraint that $J(m) \leq C$.

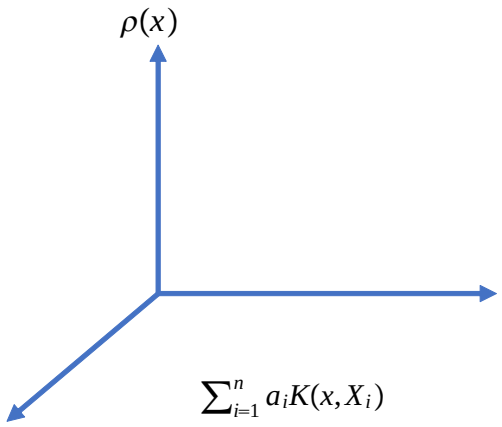
Let S_K be the closed linear subspace of $W_q[0,1]$ spanned by the functions $K(\cdot, X_i), i = 1, ..., n$. Every $m(x)$ in the Hilbert space has a unique decomposition, a component in the subspace and a component orthogonal to it:

$$m(x) = \sum_{i=1}^{n} a_i K(x, X_i) + \rho(x)$$

where $\rho(x)$ is perpendicular to the subspace S_K. That is

$$\langle \rho(\cdot), K(\cdot, X_i) \rangle_{\mathcal{H}_K} = 0 \text{ for } i = 1, 2, ..., n$$

See the figure below.



Using the reproducing property, we have

$$m(X_j) = \langle m(\cdot), K(\cdot, X_j)\rangle_{\mathcal{H}_K} = \sum_{i=1}^{n} \langle a_i K(\cdot, X_i), K(\cdot, X_j)\rangle_{\mathcal{H}_K} = \sum_{i=1}^{n} a_i K(X_i, X_j).$$

The values of $m(\cdot)$ at the data points only depend on the coefficients a_i and not on the perpendicular component $\rho(x)$.

Why is this fact important? Because the SSR is pointwise, so the first term only depends on the values of $m(x)$ at the data points. We can establish equivalence classes for the functions in $W_q[0,1]$ s.t. m and $\tilde{m}$ are equivalent if and only if $m(X_j) = \tilde{m}(X_j)$ for all the data point X_j. For the second term, we have

$$\|m\|^2_{\mathcal{H}_K} = \left\| \sum_{i=1}^{n} a_i K(x, X_i) \right\|^2_{\mathcal{H}_K} + \|\rho\|^2_{\mathcal{H}_K}.$$

So the optimal solution within each equivalent class must have $\rho = 0$. That is, the optimal solution is of the form

$$m(x) = \sum_{i=1}^{n} a_i K(x, X_i).$$

Let $\mathbf{K} = (K(X_i, X_j))$ be an $n \times n$ matrix, then the problem in (3.12) becomes

$$\min_a \frac{1}{n} (Y - \mathbf{K}a)'(Y - \mathbf{K}a) + \lambda a' \mathbf{K} a.$$

The solution for a is obtained simply as

$$a = \left(\mathbf{K}'\mathbf{K} + n\lambda \mathbf{K}\right)^{-1} \mathbf{K}' Y = (\mathbf{K} + n\lambda I)^{-1} Y \tag{3.13}$$

and the estimator $\hat{\mathbf{m}} = (\hat{m}(X_1), ..., \hat{m}(X_n))'$ is

$$\hat{\mathbf{m}} = \mathbf{K}(\mathbf{K} + n\lambda I)^{-1} Y := L(\lambda) Y,$$

which is a linear smoother.

More generally, consider the following problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \mathcal{L}[f; (Y_i, X_i)_{i=1}^n] + \Omega\left(\|f\|_{\mathcal{H}}^2\right)$$

where $\mathcal{H}$ is a RKHS with kernel $K(x, y)$. If (i) the loss function $\mathcal{L}[f; (Y_i, X_i)_{i=1}^n]$ is pointwise, i.e. $\mathcal{L}[f; (Y_i, X_i)_{i=1}^n] = \sum_{i=1}^n \ell(Y_i, f(X_i))$ for some loss function $\ell(\cdot, \cdot)$; (ii) $\Omega(\cdot)$ is a monotonically increasing function. Then the representer Theorem (Kimeldorf and Wahba (1971)) says that the optimal solution can be represented as

$$f^*(x) = \sum_{i=1}^n a_i K(x, X_i).$$

This is a powerful result. It shows that although we search for the optimal solution in an infinite-dimensional space, adding the regularization term reduces the problem to finite-dimensional. This phenomenon has been dubbed the *kernel property* or *kernel trick* in the machine learning literature.

We now go back to the original problem where $J(m) = \int_0^1 \left[m^{(q)}(x)\right]^2 dx$. Note that $J(m)$ is not squared norm of m in $W_q[0,1]$. However, we can decompose $W_q[0,1]$ into two orthogonal RKHS's: $W_q^{(1)}[0,1] \oplus W_q^{(2)}[0,1]$ where $W_q^{(1)}[0,1]$ is generated by the kernel $\sum_{v=0}^{q-1} \frac{x^v}{v!} \frac{y^v}{v!}$ and $W_q^{(2)}[0,1]$ is generated by the kernel $K_\perp(x, y) := \int_0^1 G_q(x, u) G_q(y, u) du$. Obviously, $W_q^{(1)}[0,1]$ consists of only polynomials of order at most $q-1$. $J(m)$ can be rewritten as the squared norm in $W_q[0,1]$ for the function $\mathbb{P}m$ where $\mathbb{P}$ is the projection operator onto the subspace $W_q^{(2)}[0,1]$. For any polynomial $f \in W_q^{(1)}[0,1]$, we have $\mathbb{P}f = 0$. Hence $W_q^{(1)}[0,1]$ is the null space of the roughness penalty $\|\mathbb{P}m\|_{\mathcal{H}}$.

Now we can write

$$m(x) = m^*(x) + \rho(x)$$

where $m^*(x) = \sum_{i=1}^n a_i K_\perp(x, X_i) + \sum_{\ell=0}^{q-1} b_\ell x^\ell$ and $\rho(x)$ is orthogonal to $W_q^{(1)}[0,1]$ and $\{K_\perp(x, X_i) : i = 1, ..., n\}$. That is

$$\rho^{(v)}(0) = 0, \ j = 0, 1, ..., q-1, \text{ and } \langle \rho(\cdot), K_{K_\perp}(\cdot, X_i) \rangle = 0.$$

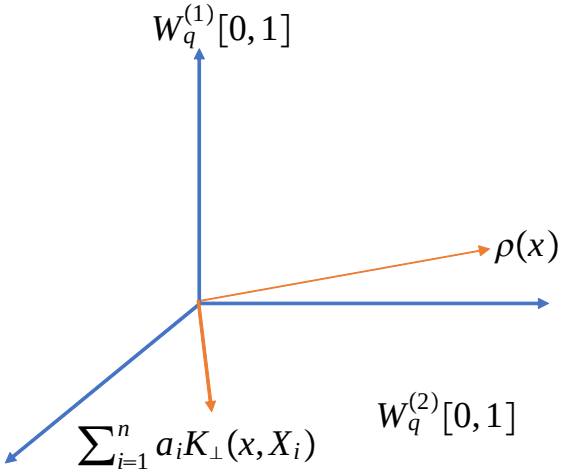
Now

$$m(X_j) = \langle m(\cdot), K(\cdot, X_j) \rangle = \sum_{i=1}^n a_i K_\perp(X_j, X_i) + \sum_{\ell=0}^{q-1} b_\ell X_j^\ell = m^*(X_j),$$

which does not depend on ρ. In addition[2],

$$\begin{aligned} J(m) &= \|m\|_{\mathcal{H}}^2 - \sum_{v=0}^{q-1} \left[m^{(v)}(0)\right]^2 \\ &= \|m^*\|_{\mathcal{H}}^2 + \|\rho\|_{\mathcal{H}}^2 - \sum_{v=0}^{q-1} \left[m^{(v)}(0)\right]^2 \\ &= \|m^*\|_{\mathcal{H}}^2 + \|\rho\|_{\mathcal{H}}^2 - \sum_{v=0}^{q-1} \left[m^{*(v)}(0)\right]^2. \end{aligned}$$

So at the optimal solution, $\rho(x) = 0$.



The decomposition of the RKHS:
$\rho(x) \in W_q^{(2)}[0,1]$ and
$\rho(x) \perp \sum_{i=1}^n a_i K_\perp(x, X_i)$ and
$\rho(x) \perp W_q^{(1)}[0,1]$

[2] Alternatively,

$$\begin{aligned} J(m) &= \|\mathbb{P}m\|_{\mathcal{H}}^2 = \|\mathbb{P}m^* + \mathbb{P}\rho\|_{\mathcal{H}}^2 \\ &= \|Pm^*\|_{\mathcal{H}}^2 + \|\rho\|_{\mathcal{H}}^2 \end{aligned}$$

because $\mathbb{P}\rho = \rho$ and $\langle \mathbb{P}m^*, \rho \rangle_{\mathcal{H}} = 0$. Hence at the optimal solution, $\rho(x) = 0$.

The optimal $m(x)$ is then

$$m^*(x) = \sum_{i=1}^{n} a_i K_\perp(x, X_i) + \sum_{j=0}^{q-1} b_j x^j.$$

When $q = 4$, we can show that the above representation is exactly the same as (3.5).

The above discussions are based on a particular RKHS $W_q[0,1]$. It is easy to see that all the results can be generalized to any RKHS with a continuous, symmetric and positive semi-definite kernel. This leads to a huge literature on kernel methods in the machine learning and statistical learning communities. See Sec 5.8 of Hastie, Tibshirani and Friedman (2009) for more examples and additional references.

3.5 Conditional Moment Estimation: Ridge and Lasso

3.5.1 Motivation

The form of a in (3.13) reminds us of the the ridge estimator for linear regression models. In ridge regression, one estimates the parameter vector β_0 of the linear model

$$Y = X\beta_0 + \varepsilon$$

by minimizing the objective function

$$\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

The first term of the object function is the residual sum of squares. The second term penalizes a large norm ("roughness") of the parameter vector. The regularization parameter (ridge parameter) λ determines the trade-off between minimizing the residual sum of squares and minimizing the norm of the estimate. For a given regularization parameter λ, the regularized estimate is

$$\hat{\beta}_\lambda = \left(X'X + n\lambda I\right)^{-1} X'Y.$$

For regularization parameter $\lambda \to 0$, the estimator $\hat{\beta}_\lambda$ reduces to the ordinary least squares estimator. For regularization parameter $\lambda \to \infty$, the estimator $\hat{\beta}_\lambda$ approaches zero. For intermediate values of the regularization parameter, the estimator $\hat{\beta}_\lambda$ is "shrunk" toward zero compared with the ordinary least squares estimator. Hence, it is a biased estimator. Even if the design matrix X is rank-deficient, so that $X'X$ is singular, the regularized matrix $X'X + n\lambda I$ is nonsingular for any nonzero value of λ.

The ridge estimator has a Bayesian interpretation. Consider the model $Y_i = X_i\beta + \varepsilon_i$ with $\varepsilon_i | X \sim iidN(0, \sigma_\varepsilon^2)$. Assume that the prior distribution of β given X is iid $N(0, \sigma_\beta^2)$, then the

joint distribution of Y and β is

$$\begin{aligned} & f(Y,\beta|X) = f(Y|X,\beta)f(\beta|X) \\ = & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(Y_i - X_i\beta)^2}{2\sigma_\varepsilon^2}\right) \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left(-\frac{\beta_j^2}{2\sigma_\beta^2}\right) \\ = & \ const. \times \exp\left(-\left[\frac{\sum_{i=1}^{n}(Y_i - X_i\beta)^2}{2\sigma_\varepsilon^2} + \frac{\sum_{j=1}^{p}\beta_j^2}{2\sigma_\beta^2}\right]\right) \\ = & \ const. \times \exp\left(-\left[\frac{\|Y - X\beta\|_2^2 + n\lambda\|\beta\|_2^2}{2\sigma_\varepsilon^2}\right]\right) \end{aligned}$$

where $\lambda = \sigma_\varepsilon^2 / \left(n\sigma_\beta^2\right)$. The posterior distribution of β is proportional to the above expression. The mode of this posterior distribution is obtained by minimizing $\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$.

If one replaces the L_2 norm by the L_1 norm, then the penalized OLS becomes

$$\arg\min_{\beta}\left(\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\sum_{i=1}^{p}|\beta_i|\right) = \arg\min_{\beta}\left(\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right).$$

The resulting estimator is the so called LASSO/Lasso (Least Absolute Shrinkage and Selection Operator) estimator, as it shrinks some coefficients and set other coefficients to zero. The estimator is proposed by Tibshirani (1996, JRSS). It is designed to retain good features of both subset selection and ridge regression. The method is described by Hastie, Tibshirani, and Friedman (2009), among others. Asymptotics for the lasso estimator has been established by Knight and Fu (2000).

The following graph compares the ridge regression with Lasso when $p = 2$.

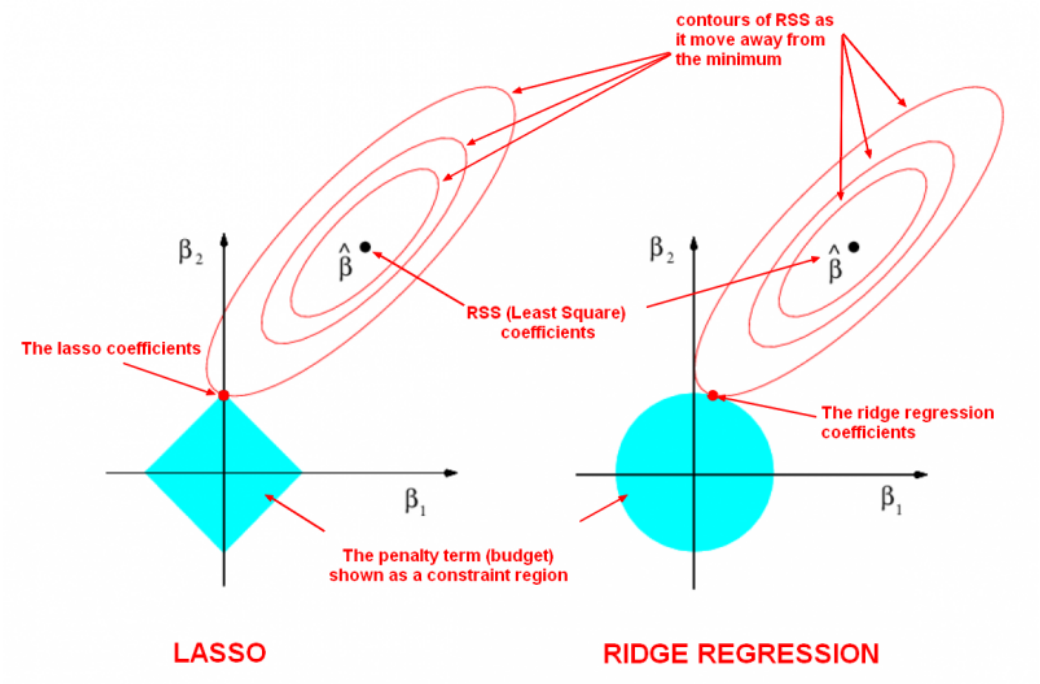
In both the ridge regression and Lasso, it is standard practice not to restrict the intercept. In the above discussion, we have implicitly assumed that the data have been demeaned, i.e., $Y_i = \tilde{Y}_i - n^{-1}\sum_{i=1}^{n}\tilde{Y}_i$ and $X_i = \tilde{X}_i - n^{-1}\sum_{i=1}^{n}\tilde{X}_i$ where $\left\{\left(\tilde{X}_i, \tilde{Y}_i\right)\right\}$ are the raw data so that β is the vector of slope coefficients. In addition, we assume that the data have been standardized, as otherwise it does not make sense to talk about the size of β when it is not scale invariant.

As a middle ground between Ridge regression and LASSO, we may consider a penalty that is a linear combination of L_1 and L_2 penalties, leading to the *elastic-net* penalty:

$$\sum_{i=1}^{p}\left[\alpha\beta_i^2 + (1-\alpha)|\beta_i|\right]$$

The minimization problem becomes

$$\arg\min_{\beta}\left(\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\sum_{i=1}^{p}\left[\alpha\beta_i^2 + (1-\alpha)|\beta_i|\right]\right).$$



The above approach is introduced by Zou and Hastie (2005). There are many other types of penalty functions in the literature.

3.5.2 Basic Idea of Lasso

Lasso has become increasingly popular in statistics. Among others, the main reasons are

(i) Its capacity to handle high-dimensional estimation problems, including the case when the number of regressors is greater than the sample size n.

(ii) Its statistical accuracy for prediction and variable selection.

(iii) Its computational feasibility.

To illustrate the shrinkage and selection mechanism behind Lasso, consider the case that $p = n$ and $X'X/n = I_p$, the identity matrix. The OLS estimator of β is then $\hat{\beta}_{OLS} = X'Y/n$. The ridge estimator is

$$\hat{\beta}_\lambda = \left(\frac{X'X}{n} + \lambda I\right)^{-1} \frac{1}{n} X'Y = \frac{1}{1+\lambda}\hat{\beta}_{OLS},$$

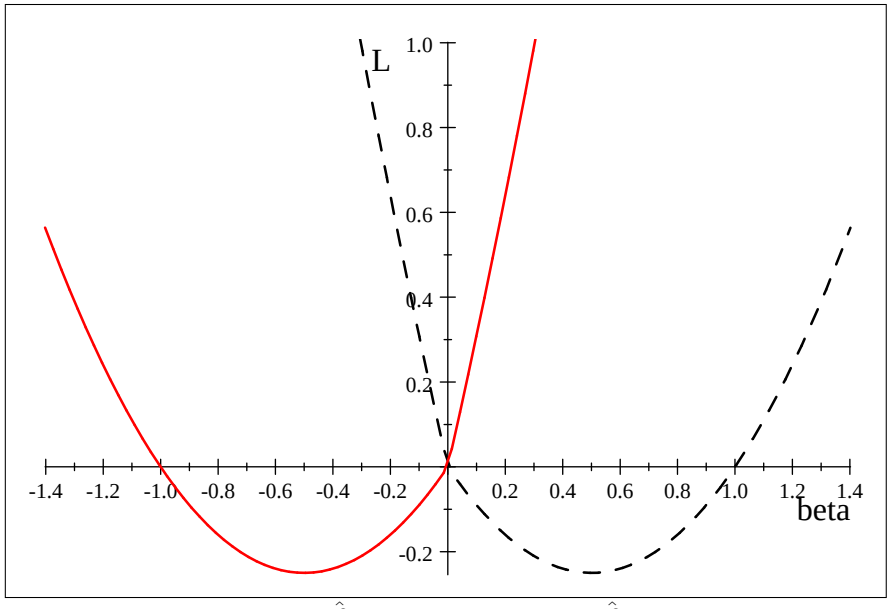
which shrinks $\hat{\beta}_{OLS}$ by the same fraction. Note that

$$\begin{aligned}\frac{1}{n}\|Y - X\beta\|^2 + \lambda\sum_{i=1}^{p}|\beta_i| &= \frac{1}{n}Y'Y + \frac{1}{n}\beta'X'X\beta - \frac{2}{n}\beta'X'Y + \lambda\sum_{i=1}^{p}|\beta_i| \\ &= \frac{1}{n}Y'Y + \beta'\beta - 2\beta'\hat{\beta}_{OLS} + \lambda\sum_{i=1}^{k}|\beta_i| \\ &= \sum_{i=1}^{n}\left(\beta_i^2 - 2\beta_i\hat{\beta}_{i,OLS} + \lambda|\beta_i|\right) + const.\end{aligned}$$

To minimize the Lasso objective function, we can minimize each individual term

$$L_i = \beta_i^2 - 2\beta_i\hat{\beta}_{i,OLS} + \lambda|\beta_i|$$

in the above sum. Figure 3.5.2 graphs L_i as a function of β_i.



dashed line: $\hat{\beta}_{OLS} > 0$; solid line: $\hat{\beta}_{OLS} > 0$

It is obvious to see that if $\hat{\beta}_{i,OLS} > 0$, then $\beta_{i,lasso} > 0$. So when $\hat{\beta}_{i,OLS} > 0$, it suffices to consider

$$\tilde{L}_i(\beta_i) = \beta_i^2 - 2\beta_i\hat{\beta}_{i,OLS} + \lambda\beta_i, \text{ s.t. } \beta_i > 0.$$

Ignoring the constraint, the first order condition is $2\hat{\beta}_i - 2\hat{\beta}_{i,OLS} + \lambda = 0$, which implies $\hat{\beta}_i = \hat{\beta}_{i,OLS} - \frac{\lambda}{2}$. But the LASSO estimator has to be positive, and so in this case the actual

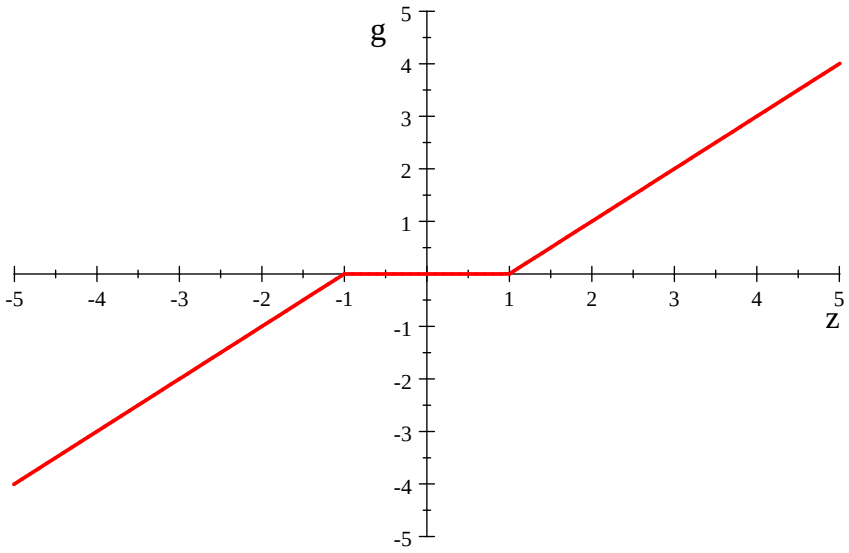


Figure 3.4: Soft Thesholding Function $g_{soft,\lambda/2}$ for $\lambda = 2$.

solution is

$$\hat{\beta}_{i,lasso} = sign\left(\hat{\beta}_{i,OLS}\right)\left(\hat{\beta}_{i,OLS} - \frac{\lambda}{2}\right)_+ = sign\left(\hat{\beta}_{i,OLS}\right)\left(\left|\hat{\beta}_{i,OLS}\right| - \frac{\lambda}{2}\right)_+$$

Similarly, when $\hat{\beta}_{i,OLS} < 0$, it is enough to minimize

$$\tilde{L}_i\left(\beta_i\right) = \beta_i^2 - 2\beta_i\hat{\beta}_{i,OLS} - \lambda\beta_i, \text{s.t. } \beta_i < 0.$$

The solution is

$$\hat{\beta}_{i,lasso} = -\left[-\left(\hat{\beta}_{i,OLS} + \frac{\lambda}{2}\right)\right]_+ = sign\left(\hat{\beta}_{i,OLS}\right)\left[\left|\hat{\beta}_{i,OLS}\right| - \frac{\lambda}{2}\right]_+.$$

Finally, when $\hat{\beta}_{i,OLS} = 0$, we have $\hat{\beta}_{i,lasso} = 0$. Combining the three cases, we have

$$\hat{\beta}_{i,lasso} = sign\left(\hat{\beta}_{i,OLS}\right)\left(\left|\hat{\beta}_{i,OLS}\right| - \frac{\lambda}{2}\right)_+ = g_{soft,\lambda/2}(\hat{\beta}_{i,OLS})$$

where $g_{soft,\lambda/2}\left(z\right) = sign(z)\max\left(|z| - \frac{\lambda}{2}, 0\right)$. The function is given in Figure 3.4.

3.5.3 Some Theory for the Lasso

Prediction Consistency of Lasso

Consider the classical linear regression

$$Y_i = X_i\beta_0 + \varepsilon_i, i = 1, ..., n.$$

where X_i's are fixed and $\varepsilon_i \sim iidN(0, \sigma^2)$. Suppose for the moment that $p = \dim(\beta) < n$. Let the OLS estimator be $\hat{\beta}_{OLS} = (X'X)^{-1} X'Y$ with the corresponding prediction error $\hat{\varepsilon}_{i,OLS} = \hat{Y} - X\beta_0 = X\left(\hat{\beta}_{OLS} - \beta_0\right) = X(X'X)^{-1} X'\varepsilon$. Then

$$\frac{\|\hat{\varepsilon}_{OLS}\|^2}{\sigma^2} := \frac{\left(\hat{\beta}_{OLS} - \beta_0\right)' X'X\left(\hat{\beta}_{OLS} - \beta_0\right)}{\sigma^2} = \frac{\varepsilon' X(X'X)^{-1} X'\varepsilon}{\sigma^2} \sim \chi^2_p.$$

It follows that

$$\frac{E\left\|X\left(\hat{\beta}_{OLS} - \beta_0\right)\right\|^2}{n} = \frac{p}{n}\sigma^2.$$

If we orthonormalize the design matrix X such that $X'X = I_p$, then $E\left\|\left(\hat{\beta}_{OLS} - \beta_0\right)\right\|_2^2 = p\sigma^2/n$. That is, each parameter is estimated with the squared accuracy σ^2/n. The overall squared accuracy is $\sigma^2/n \times p$.

Now we turn to the case $p > n$. The OLS is not possible and we attempt to use the Lasso. Let

$$S_0 = \left\{j : \beta_{0,j} \neq 0\right\} \text{ and } s_0 = |S_0|, \text{ the cardinality of } S_0,$$

where s_0 is often referred to as the sparsity index of the true β_0. We assume that s_0 is relatively small. In particular, $s_0 < n$. Had we known the active set S_0, we would neglect all covariates not in S_0 and estimate the coefficients for those in S_0 only by OLS. We would end up with the squared accuracy $\sigma^2/n \times s_0$. Now that S_0 is not known, we propose to estimate β by Lasso. Under some conditions, we will show that with a large probability

$$\frac{\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2}{n} \leq C \times \frac{\sigma^2}{n} \times s_0 \times \log(p) \tag{3.14}$$

for some constant $C > 0$.

Lemma 3.5.1 *We have*

$$\frac{1}{n}\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_L\right\|_1 \leq \frac{2}{n}\varepsilon' X\left(\hat{\beta}_L - \beta_0\right) + \lambda\|\beta_0\|_1 \tag{3.15}$$

Proof. By definition

$$\frac{1}{n}\left\|Y - X\hat{\beta}_L\right\|_2^2 + \lambda\left\|\hat{\beta}_L\right\|_1 \leq \frac{1}{n}\left\|Y - X\beta_0\right\|_2^2 + \lambda\left\|\beta_0\right\|_1.$$

That is

$$\frac{1}{n}\left\|\varepsilon - X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_L\right\|_1 \leq \frac{1}{n}\left\|\varepsilon\right\|_2^2 + \lambda\left\|\beta_0\right\|_1.$$

Rewriting the above leads to the lemma. ■

The lhs of (3.15) is the square loss plus the penalty term. We proceed to bound the rhs of (3.15). Note that

$$\begin{aligned}\left|2\varepsilon' X\left(\hat{\beta}_L - \beta_0\right)\right| &= \left|2\varepsilon'[X^{(1)}, ..., X^{(p)}]\left(\hat{\beta}_L - \beta_0\right)\right| \\ &= \left|2[\varepsilon' X^{(1)}, ..., \varepsilon' X^{(p)}]\left(\hat{\beta}_L - \beta_0\right)\right| \\ &\leq 2\max_{j=1,...,p}\left|\varepsilon' X^{(j)}\right|\left\|\hat{\beta}_L - \beta_0\right\|_1.\end{aligned}$$

Define the event

$$\mathcal{E} = \left\{\frac{\max_{j=1,...,p} 2\left|\varepsilon' X^{(j)}\right|}{n} \leq \lambda_0\right\}$$

for some λ_0 to be defined later. The lemma below shows that the event happens with a large probability.

Lemma 3.5.2 *Suppose that* $\left\|X^{(j)}\right\|^2/n = 1$ *for all* j*, then for all* $t > 0$ *and for*

$$\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log p}{n}},$$

we have

$$P(\mathcal{E}) \geq 1 - \exp\left(-\frac{t^2}{2}\right).$$

Proof. It follows from $\left\|X^{(j)}\right\|^2/n = 1$ and the normality assumption that $\varepsilon' X^{(j)}/\sqrt{n\sigma^2} \sim$

$N(0,1)$. Let χ_1^2 and χ_2^2 be the χ^2 random variables with 1 and 2 degrees of freedom. Then

$$\begin{aligned} P(\mathcal{E}^c) &= P\left\{\frac{\max_{j=1,\ldots,p} 2\left|\varepsilon' X^{(j)}\right|}{n} \geq 2\sigma\sqrt{\frac{t^2+2\log p}{n}}\right\} \\ &= P\left\{\frac{\max_{j=1,\ldots,p}\left|\varepsilon' X^{(j)}\right|}{\sqrt{n\sigma^2}} \geq \sqrt{t^2+2\log p}\right\} \\ &\leq \sum_{j=1}^{p} P\left\{\chi_1^2 \geq t^2+2\log p\right\} \leq \sum_{j=1}^{p} P\left\{\chi_2^2 \geq t^2+2\log p\right\} \\ &= p\exp\left(-\frac{t^2+2\log p}{2}\right) = \exp\left(-\frac{t^2}{2}\right) \end{aligned}$$

where the last line uses the fact that the CDF of χ_2^2 is $1-\exp(x/2)$. ■

Corollary 3.5.1 *Suppose that* $\left\|X^{(j)}\right\|^2/n = 1$ *for all* j. *Let*

$$\lambda = 4\hat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$$

where $\hat{\sigma}$ *is an estimator of* σ. *Then with probability of least* $1-\alpha$ *for*

$$\alpha = \exp\left(-\frac{t^2}{2}\right) + P\left\{\frac{1}{2} \leq \hat{\sigma}/\sigma \leq 1\right\},$$

we have

$$\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2/n \leq 3\lambda_0 \left\|\beta_0\right\|_1$$

Proof. Conditional on the event $\mathcal{E}$ and $\left\{\frac{1}{2} \leq \hat{\sigma}/\sigma \leq 1\right\}$, we have

$$\begin{aligned} \left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2/n + \lambda\left\|\hat{\beta}_L\right\|_1 &\leq 2\varepsilon' X\left(\hat{\beta}_L - \beta_0\right)/n + \lambda\left\|\beta_0\right\|_1 \\ &\leq \lambda_0\left\|\hat{\beta}_L - \beta_0\right\|_1 + 2\lambda_0\frac{\hat{\sigma}}{\sigma}\left\|\beta_0\right\|_1 \\ &\leq \lambda_0\left\|\hat{\beta}_L\right\|_1 + \lambda_0\left\|\beta_0\right\|_1 + 2\lambda_0\left\|\beta_0\right\|_1 \\ &= \lambda_0\left\|\hat{\beta}_L\right\|_1 + 3\left\|\beta_0\right\|_1 \end{aligned} \tag{3.16}$$

and so

$$\begin{aligned} \left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2/n &\leq 3\lambda_0\left\|\beta_0\right\|_1 + \lambda_0\left\|\hat{\beta}_L\right\|_1 - 2\lambda_0\frac{\hat{\sigma}}{\sigma}\left\|\hat{\beta}_L\right\|_1 \\ &\leq 3\lambda_0\left\|\beta_0\right\|_1. \end{aligned}$$

■

To use the corollary, we can pick up a large enough t and a large enough estimator $\hat{\sigma}$ such that α is small. In this case, the result in the corollary holds with high probability.

It follows from the corollary that if $\|\beta_0\|_1$ is of smaller order than $\sqrt{n/\log p}$, then $\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2/n \to 0$. This is, Lasso is prediction consistent. When $\|\beta_0\|_1 \propto s_0$, we have

$$\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2/n \leq 6\sigma\sqrt{\frac{t^2+2\log p}{n}}s_0$$

with high probability. Comparing with (3.14), this upper bound is not good enough but we obtain it easily.

Prediction Consistency and L_1 Consistency

We now turn to more refined results. Let

$$\beta_{j,S} = \beta_j 1\{j \in S\} \text{ and } \beta_{j,S^c} = \beta_j 1\{j \notin S\}$$

and $\beta_S = \left(\beta_{j,S}\right)_{p\times 1}$, $\beta_{j,S^c} = \left(\beta_{j,S}\right)_{p\times 1}$. Clearly $\beta = \beta_S + \beta_{S^c}$.

Lemma 3.5.3 *On $\mathcal{E}$, with $\lambda \geq 2\lambda_0$, we have*

$$\frac{2}{n}\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_{S_0^c}\right\|_1 \leq 3\lambda\left\|\hat{\beta}_{S_0} - \beta_{0,S_0}\right\|_1$$

Proof. It follows from Lemma 3.5.1 that on $\mathcal{E}$, we have

$$\frac{1}{n}\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_L\right\|_1 \leq \lambda_0\left\|\hat{\beta}_L - \beta_0\right\|_1 + \lambda\|\beta_0\|_1, \tag{3.17}$$

and so

$$\frac{2}{n}\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 + 2\lambda\left\|\hat{\beta}_L\right\|_1 \leq \lambda\left\|\hat{\beta}_L - \beta_0\right\|_1 + 2\lambda\|\beta_0\|_1. \tag{3.18}$$

But

$$\begin{aligned}\left\|\hat{\beta}_L\right\|_1 &= \left\|\hat{\beta}_{L,S_0}\right\|_1 + \left\|\hat{\beta}_{L,S_0^c}\right\|_1 \\ &\geq \|\beta_{0,S_0}\|_1 - \left\|\hat{\beta}_{L,S_0} - \beta_{0,S_0}\right\|_1 + \left\|\hat{\beta}_{L,S_0^c}\right\|_1\end{aligned}$$

and

$$\left\|\hat{\beta}_L - \beta_0\right\|_1 = \left\|\hat{\beta}_{L,S_0} - \beta_{0,S_0}\right\|_1 + \left\|\hat{\beta}_{L,S_0^c} - \beta_{0,S_0^c}\right\|_1,$$

we have

$$\begin{aligned}
&\frac{2}{n}\left\|X\left(\hat{\beta}_L-\beta_0\right)\right\|^2+2\lambda\left\|\beta_{0,S_0}\right\|_1-\underbrace{2\lambda\left\|\hat{\beta}_{L,S_0}-\beta_{0,S_0}\right\|_1}+\underbrace{2\lambda\left\|\hat{\beta}_{L,S_0^c}\right\|_1}\\
\leq\ &\underbrace{\lambda\left\|\hat{\beta}_{L,S_0}-\beta_{0,S_0}\right\|_1}+\underbrace{\lambda\left\|\hat{\beta}_{L,S_0^c}-\beta_{0,S_0^c}\right\|_1}+2\lambda\left\|\beta_0\right\|_1,
\end{aligned}$$

which implies

$$\frac{2}{n}\left\|X\left(\hat{\beta}_L-\beta_0\right)\right\|^2+\lambda\left\|\hat{\beta}_{L,S_0^c}\right\|_1\leq 3\lambda\left\|\hat{\beta}_{L,S_0}-\beta_{0,S_0}\right\|_1.$$

■

Remark 3.5.1 *The lemma implies that* $\left\|\hat{\beta}_{S_0^c}\right\|_1\leq 3\left\|\hat{\beta}_{S_0}-\beta_{0,S_0}\right\|_1=3\left\|\hat{\beta}_{S_0}\right\|_1$.

Assumption 3.5.1 (Compatibility) *For all* β *satisfying* $\left\|\beta_{S_0^c}\right\|_1\leq 3\left\|\beta_{S_0}\right\|_1$, *the following condition holds*

$$\left\|\beta_{S_0}\right\|_1^2\leq\beta'\left(\frac{X'X}{n}\right)\beta\frac{s_0}{\phi_0^2}$$

for some constant ϕ_0^2

Theorem 3.5.1 *Suppose that the compatibility assumption holds for* S_0. *Then on* $\mathcal{E}$, *we have, for* $\lambda\geq 2\lambda_0$

$$\frac{1}{n}\left\|X\left(\hat{\beta}_L-\beta_0\right)\right\|_2^2+\lambda\left\|\hat{\beta}_L-\beta_0\right\|_1\leq\frac{4\lambda^2 s_0}{\phi_0^2}.$$

Proof. Using the compatibility assumption and Lemma 3.5.3, we have

$$\begin{aligned}
&\frac{2}{n}\left\|X\left(\hat{\beta}_L-\beta_0\right)\right\|_2^2+\lambda\left\|\hat{\beta}_L-\beta_0\right\|_1\\
=\ &\frac{2}{n}\left\|X\left(\hat{\beta}_L-\beta_0\right)\right\|_2^2+\lambda\left\|\hat{\beta}_{L,S_0}-\beta_{0,S_0}\right\|_1+\lambda\left\|\hat{\beta}_{L,S_0^c}-\beta_{0,S_0^c}\right\|_1\\
=\ &\frac{2}{n}\left\|X\left(\hat{\beta}_L-\beta_0\right)\right\|_2^2+\lambda\left\|\hat{\beta}_{L,S_0}-\beta_{0,S_0}\right\|_1+\lambda\left\|\hat{\beta}_{L,S_0^c}\right\|_1\\
\leq\ &4\lambda\left\|\hat{\beta}_{L,S_0}-\beta_{0,S_0}\right\|_1\leq\frac{4\lambda\sqrt{s_0}}{\phi_0}\frac{\left\|X'\left(\hat{\beta}_L-\beta_0\right)\right\|_2}{\sqrt{n}}\\
\leq\ &\frac{\left\|X'\left(\hat{\beta}_L-\beta_0\right)\right\|_2^2}{n}+\frac{4\lambda^2 s_0}{\phi_0^2}
\end{aligned}$$

which implies the stated result. ■

Corollary 3.5.2 *Suppose that* $\left\|X^{(j)}\right\|^2/n = 1$ *for all* j *and the compatibility assumption holds for* S_0. *Let*

$$\lambda = 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}}$$

where $\hat{\sigma}$ *is an estimator of* σ. *Then with probability of least* $1-\alpha$ *for*

$$\alpha = \exp\left(-\frac{t^2}{2}\right) + P\left\{\frac{1}{2} \leq \hat{\sigma}/\sigma \leq 1\right\},$$

we have

$$\frac{1}{n}\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_L - \beta_0\right\|_1 \leq \frac{4\lambda^2 s_0}{\phi_0^2}.$$

It follows from the above corollary that with high probability

$$\frac{1}{n}\left\|X\left(\hat{\beta}_L - \beta_0\right)\right\|_2^2 \leq C_1 \times \frac{\sigma^2}{n} \times s_0 \times \log(p)$$

and

$$\left\|\hat{\beta}_L - \beta_0\right\|_1 \leq C_2 s_0\sqrt{\frac{\log p}{n}}$$

for some constants C_1 and C_2.

3.6 Conditional Moment Estimation: Series Estimators

3.6.1 Series Estimator and its Convergence in Weak and Strong Norms

Consider the model

$$Y_i = m_0(X_i) + u_i \text{ where } E(u_i|X_i) = 0 \text{ a.e.}$$

for $i = 1, 2, ..., n$. Suppose that $m_0 \in (\mathcal{M}, d)$, which is an infinite dimensional metric space. We assume that $\mathcal{M}$ can be well approximated by a sequence of the sieve spaces $\mathcal{M}_{J_n}$:

$$\mathcal{M}_{J_n} = \left\{ m(x) : m(x) = \sum_{j=1}^{J_n} \phi_j(x)\beta_j \text{ for some } (\beta_1, ..., \beta_J) \right\}.$$

That is, for any $m_0 \in \mathcal{M}$, there exists $m_{0J_n} \in \mathcal{M}_{J_n}$ such that $d(m_{0J_n}, m_0) \to 0$ as $J_n \to \infty$. We often write $m_{0J_n} := \Pi_{J_n} m_0$ which reminds us that m_{0J_n} is the "projection" of m_0 onto the sieve space.

The method of sieves involves using m_{0J_n} as an approximation to m_0 and regressing Y_i on $\phi_i^{J_n} =: \phi^{J_n}(X_i) := \left(\phi_1(X_i), ..., \phi_{J_n}(X_i)\right)$. Let $\left\{\hat{\beta}_j\right\}_{j=1}^{J_n}$ be the least squares estimators, then

$$\hat{m}(x) = \sum_{j=1}^{J_n} \phi_j(x)\hat{\beta}_j = \sum_{i=1}^{n} W_{ni}(x) Y_i$$

where

$$W_n(x) = (W_{n1}, ..., W_{nn})' = \left[\phi^{J_n}(x)\right]\left(\Phi'_{J_n}\Phi_{J_n}\right)^{-1}\Phi'_{J_n}$$
$$\phi^{J_n}(x) = \left(\phi_1(x), ..., \phi_{J_n}(x)\right),$$

and

$$\Phi_{J_n} = \begin{pmatrix} \phi_1^{J_n} \\ ... \\ \phi_n^{J_n} \end{pmatrix}.$$

To derive the asymptotic properties, one splits up in a variance term and a bias term:

$$\hat{m}(x) - m_0(x) = \sum_{j=1}^{J_n} \phi_j(x)\left(\hat{\beta}_j - \beta_j\right) + \sum_{j=J_n+1}^{\infty} \phi_j(x)\beta_j.$$

The second term is a deterministic bias term, which goes to to zero as $J_n \to \infty$. The first term is the variance term. To control this term, J_n cannot be allowed to go to infinity too fast. So the choice of J_n plays the same role here as the bandwidth h does in the kernel estimation.

To rigorously derive the asymptotic properties of $\hat{m}(x)$, we maintain the following assumptions, where for notational economy we use J for J_n.

Assumption 1. $\{X_i, Y_i\}$ is iid and $var(Y_i|X_i)$ is bounded on the support of X.

Assumption 2. (i) The smallest eigenvalue of $E\left[\phi^J(X_i)\right]'\phi^J(X_i)$ is bounded away from zero uniformly in $J = 1, 2, ...$

(ii) There exists a sequence of constants $\zeta(J)$ such that

$$\sup_{x \in \text{supp}(X)} \left\|\phi^J(x)\right\| \leq \zeta(J)$$

and $\zeta^2(J) J/n \to 0$ as $n \to \infty$.

(iii) There exists $\alpha > 0$ such that

$$\sup_{x \in \text{supp}(X)} |m_0(x) - \phi^J(x)\beta^J| = O(J^{-\alpha})$$

uniformly in J where $\beta^J = (\beta_1, ..., \beta_J)'$.

(iv) $1/J + J/n \to 0$ as $n \to \infty$.

Remark 3.6.1 *In OLS, we require $EX_i'X_i$ to be invertible, that is, the smallest eigenvalue of $EX_i'X_i$ to be positive. Assumption 2(i) basically imposes that $E\left[\phi^J(X_i)\right]'\phi^J(X_i)$ is invertible uniformly in J.*

Remark 3.6.2 *For assumption 2(ii), consider an example where $\phi^J(x) = \left(1, x, ..., x^{J-1}\right)$. Then $\left\|\phi^J(x)\right\|^2 = 1 + x^2 + ... + x^{2J-2}$. If $x \in [0,1]$, then $\sup\left\|\phi^J(x)\right\| = O\left(\sqrt{J}\right)$ and we can take $\zeta(J) = C\sqrt{J}$ for some constant C. So the rate condition $\zeta^2(J)J/n \to 0$ reduces to $J^2/n \to 0$, which limits the growth of J. The rate condition $\zeta^2(J)J/n \to 0$ may not be the weakest possible but we are content with it, as it simplifies our proof.*

Remark 3.6.3 *In Assumption 2(iii), α is related to the smoothness of the function $m_0(x)$ and the dimensionality of x. For example, for splines and power series, this assumption will be satisfied with $\alpha = s/d$ where s is the number of continuous derivatives of $m_0(x)$ that exist, and d is the dimension of x. The smoother $m_0(x)$ is, the easier it is to approximate $m_0(x)$.*

Remark 3.6.4 *Sometimes we replace Assumption 2(iii) by a stronger assumption in order to obtain stronger results. For any integer $d \geq 0$, letting*

$$\|m\|_d = \max_{\|\lambda\| \leq d} \sup_{x \in supp(X)} \left| \frac{\partial^{|\lambda|}}{\partial x^{|\lambda|}} m(x) \right|,$$

we require

$$\left\|m_0(\cdot) - \phi^J(\cdot)\beta^J\right\|_d = O(J^{-\alpha}). \tag{3.19}$$

When $d = 0$, the above condition reduces to Assumption 2(iii).

Define the norm

$$\|m\|_w = \left\{E\left[m(X)\right]^2\right\}^{1/2} = \left(\int_{\text{supp}(X)} [m(x)]^2 \, dF(x)\right)^{1/2}$$

where $F(x)$ is the CDF of X.

Theorem 3.6.1 *Under Assumptions 1 and 2, we have*

$$\|\hat{m} - m_0\|_w^2 = O_p\left(\frac{J}{n} + J^{-2\alpha}\right).$$

Proof. Without the loss of generality, we assume that $E\left[\phi^J(X_i)\right]'\phi^J(X_i) = I_J$, a $J \times J$ identity matrix. If $E\left[\phi^J(X_i)\right]'\phi^J(X_i) := Q_J \neq I_J$, we can simply change the basis functions $\phi^J(X_i)$ to $\phi^J(X_i)Q_J^{-1/2}$ and β^J to $Q_J^{1/2}\beta^J$. It is easy to see that all assumptions continue to hold for the transformed basis functions. In fact, we only to check the nontrivial condition that $\sup_x \left\|\phi^J(x)Q_J^{-1/2}\right\|^2 J/n \to 0$. But

$$\begin{aligned}\sup_{x\in\text{supp}(X)} \left\|\phi^J(x)Q_J^{-1/2}\right\|^2 J/n &\leq \sup_{x\in\text{supp}(X)} \left\|\phi^J(x)Q_J^{-1}\phi^J(x)'\right\|^2 J/n \\ &\leq \left[\lambda_{\min}\left(Q_J\right)\right]^{-1} \sup_{x\in\text{supp}(X)} \left\|\phi^J(x)\right\|^2 J/n\end{aligned}$$

So if $\lambda_{\min}\left(Q_J\right)$ is bounded away from zero uniformly over J, then $\sup_{x\in\text{supp}(X)} \left\|\phi^J(x)Q_J^{-1/2}\right\|^2 J/n \to 0$. This derivation shows that Assumption 2(i) can be replaced by $\left[\lambda_{\min}\left(Q_J\right)\right]^{-1} \sup_x \left\|\phi^J(x)\right\|^2 J/n \to 0$.

Now

$$\begin{aligned}&\left\|\hat{m}(x) - m_0(x)\right\|_w^2 = \int \left[\hat{m}(x) - m_0(x)\right]^2 dF(x) \\ =& \int \left[\phi^J(x)\hat{\beta}^J - \phi^J(x)\beta^J + \phi^J(x)\beta^J - m_0(x)\right]^2 dF(x) \\ \leq& \underbrace{2\int \left[\phi^J(x)\left(\hat{\beta}^J - \beta^J\right)\right]^2 dF(x)}_{\text{Estimation error}} + \underbrace{2\int \left[\phi^J(x)\beta^J - m_0(x)\right]^2 dF(x)}_{\text{Approximation error}} \\ =& 2\left[\hat{\beta}^J - \beta^J\right]'\left[\int \left[\phi^J(x)\right]'\phi^J(x)dF(x)\right]\left(\hat{\beta}^J - \beta^J\right) + O\left(J^{-2\alpha}\right) \\ =& 2\left\|\hat{\beta}^J - \beta^J\right\|^2 + O\left(J^{-2\alpha}\right),\end{aligned}$$

it suffices to show that

$$\left\|\hat{\beta}^J - \beta^J\right\|^2 = O_p\left(\frac{J}{n} + J^{-2\alpha}\right).$$

Let $\hat{Q}_J = n^{-1}\Phi_J'\Phi_J$ and $\mathbf{m} = (m_0(X_1), ..., m_0(X_n))'$, then

$$\hat{\beta}^J - \beta^J = \hat{Q}_J^{-1}n^{-1}\Phi_J'\left(Y - \Phi_J\beta^J\right) = \hat{Q}_J^{-1}n^{-1}\Phi_J'u + \hat{Q}_J^{-1}n^{-1}\Phi_J'\left(\mathbf{m} - \Phi_J\beta^J\right)$$

where $Y = (Y_1, ..., Y_n)'$ and $u = (u_1, ..., u_n)$. We proceed to show the following three results:

Step 1. Show that $E\left(\left\|\hat{Q}_J - I_J\right\|^2\right) = O\left[\zeta^2(J)\, J/n\right]$ [3].

Proof: For a $J \times J$ matrix A, define $\|A\|^2 = \sum_{j=1}^J \sum_{i=1}^J a_{ij}^2 = tr(A'A)$ and $\|A\|_2^2 = \lambda_{\max}(A'A)$. The former is called the (squared) Frobenius norm or the Hilbert-Schmidt norm and the latter is called the spectral norm. These two norms satisfy

$$\lambda\left(A'A\right) \leq \|A\|_2^2 \leq \|A\|^2$$

where $\lambda(A'A)$ is any of the eigenvalues of $A'A$. To derive an upper bound for $\lambda(A'A)$, we can work with $\|A\|^2$. Note that the (i,j)-th element of $\hat{Q}_J$ is $n^{-1}\sum_{k=1}^n \left[\phi_i(X_k)\phi_j(X_k)\right]$, so

$$\begin{aligned}
E\left(\left\|\hat{Q}_J - I_J\right\|^2\right) &= \sum_{i=1}^J \sum_{j=1}^J E\left(\frac{1}{n}\sum_{k=1}^n \left[\phi_i(X_k)\phi_j(X_k) - \delta_{ij}\right]\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^J \sum_{j=1}^J E\left[\phi_i(X)\phi_j(X) - \delta_{ij}\right]^2 \leq \frac{1}{n}\sum_{i=1}^J \sum_{j=1}^J E\phi_i^2(X)\phi_j^2(X) \\
&= \frac{1}{n} E\left(\sum_{i=1}^J \phi_i^2(X)\right)\left(\sum_{i=1}^J \phi_i^2(X)\right) \leq \frac{1}{n}\zeta^2(J)\, E\left(\sum_{i=1}^J \phi_i^2(X)\right) \\
&= \frac{1}{n}\zeta^2(J)\, tr\left\{E\left[\phi^J(X_i)\right]' \phi^J(X_i)\right\} = \frac{J}{n}\zeta^2(J).
\end{aligned}$$

As a result,

$$E\left[\lambda\left(\left(\hat{Q}_J - I_J\right)\left(\hat{Q}_J - I_J\right)'\right)\right] \leq \frac{J}{n}\zeta^2(J).$$

But

$$E\left[\lambda\left(\left(\hat{Q}_J - I_J\right)\left(\hat{Q}_J - I_J\right)'\right)\right] = E\left\{\left[\lambda\left(\hat{Q}_J - I_J\right)\right]^2\right\} = E\left\{\left[\lambda\left(\hat{Q}_J\right) - 1\right]^2\right\}.$$

This implies that, with probability approaching one, the smallest eigenvalue of $\hat{Q}_J$ converges to 1. More specifically, let $A_n = \left\{\omega : \left\|\lambda_{\min}\left(\hat{Q}_J\right) - 1\right\| > \varepsilon\right\}$, then $P(A_n) = o(1)$ as $n \to \infty$ for any $\varepsilon > 0$. In the rest of the proof, we can focus on A_n^c, the complement of A_n. We can do this because by definition $\xi_n = O_p(1)$ if for any $\varepsilon > 0$, there exists $M_\varepsilon < \infty$ such that

[3] Roughly speaking, each element of $\hat{Q}_J$ converges to that of I_J at the rate of $1/n$. Pretending that the convergence happens independently, we have $E\left(\left\|\hat{Q}_J - I_J\right\|^2\right) = O(J^2/n)$. But the independence does not hold. That is why we need a bound for $\left\|\phi^J\right\|_\infty$ to obtain the result.

$P(|\xi_n| \geq M_\varepsilon) \leq \varepsilon$ for n sufficiently large. But

$$\begin{aligned} P_n(|\xi_n| \geq M_\varepsilon) &= P_n(|\xi_n| \geq M_\varepsilon, A_n) + P_n(|\xi_n| \geq M_\varepsilon, A_n^c) \\ &\leq P_n(A_n) + P_n(|\xi_n| \geq M_\varepsilon, A_n^c) \\ &= o(1) + P_n(|\xi_n| \geq M_\varepsilon, A_n^c) \\ &= o(1) + P_n(|\xi_n| \geq M_\varepsilon | A_n^c) \\ &= o(1) + \frac{E\left(|\xi_n| 1_{A_n^c}\right)}{M_\varepsilon} \end{aligned}$$

where $1_{A_n^c} = 1\{\omega \in A_n^c\}$. So it suffices to focus on A_n^c.

Step 2. Show that $\left\|\hat{Q}_J^{-1} n^{-1} \Phi_J' u\right\|^2 = O_p\left(\frac{J}{n}\right)$.

Proof: We prove this by using: $P_n(|\xi_n| \geq M_\varepsilon) = o(1) + E\left(|\xi_n| \cdot 1_{A_n^c}\right)/M_\varepsilon$.

$$\begin{aligned} & E\left(\left\|\hat{Q}_J^{-1} n^{-1} \Phi_J' u\right\|^2 \cdot 1_{A_n^c}\right) \\ = \; & \frac{1}{n^2} E\left(u' \Phi_J \hat{Q}_J^{-1} \hat{Q}_J^{-1} \Phi_J' u \cdot 1_{A_n^c}\right) = \frac{1}{n^2} E\left(u' \Phi_J \hat{Q}_J^{-1/2} \hat{Q}_J^{-1} \hat{Q}_J^{-1/2} \Phi_J' u \cdot 1_{A_n^c}\right) \\ \leq \; & \frac{C}{n^2} E\left(u' \Phi_J \hat{Q}_J^{-1/2} \hat{Q}_J^{-1/2} \Phi_J' u \cdot 1_{A_n^c}\right) \leq \frac{C}{n^2} E\left(u' \Phi_J \hat{Q}_J^{-1/2} \hat{Q}_J^{-1/2} \Phi_J' u\right) \\ = \; & \frac{C}{n^2} Etr\left[\Phi_J \hat{Q}_J^{-1} \Phi_J' E\left(uu'|X\right)\right] = \frac{C}{n} Etr\left(\Phi_J \left(\Phi_J' \Phi_J\right)^{-1} \Phi_J'\right) = O\left(\frac{J}{n}\right). \end{aligned}$$

Step 3. Show that $\left\|\hat{Q}_J^{-1} n^{-1} \Phi_J' \left(\mathbf{m} - \Phi_J \beta^J\right)\right\|^2 = O_p\left(J^{-2\alpha}\right)$.

Proof:

$$\begin{aligned} & P\left\{\left\|\hat{Q}_J^{-1} \frac{1}{n} \Phi_J' \left(\mathbf{m} - \Phi_J \beta^J\right)\right\|^2 > M_\varepsilon J^{-2\alpha} | A_n^c\right\} \\ = \; & P\left\{\frac{1}{n^2} \left(\mathbf{m} - \Phi_J \beta^J\right)' \Phi_J \hat{Q}_J^{-1} \hat{Q}_J^{-1} \Phi_J' \left(\mathbf{m} - \Phi_J \beta^J\right) > M_\varepsilon J^{-2\alpha} | A_n^c\right\} \\ \leq \; & P\left\{\frac{C}{n^2} \left(\mathbf{m} - \Phi_J \beta^J\right)' \Phi_J \hat{Q}_J^{-1} \Phi_J' \left(\mathbf{m} - \Phi_J \beta^J\right) > M_\varepsilon J^{-2\alpha} | A_n^c\right\} \\ = \; & P\left\{\frac{C}{n} \left(\mathbf{m} - \Phi_J \beta^J\right)' \Phi_J \left(\Phi_J' \Phi_J\right)^{-1} \Phi_J' \left(\mathbf{m} - \Phi_J \beta^J\right) > M_\varepsilon J^{-2\alpha} | A_n^c\right\} \\ \leq \; & P\left\{\frac{C}{n} \left(\mathbf{m} - \Phi_J \beta^J\right)' \left(\mathbf{m} - \Phi_J \beta^J\right) > M_\varepsilon J^{-2\alpha} | A_n^c\right\} \\ = \; & P\left\{O\left(J^{-2\alpha}\right) > M_\varepsilon J^{-2\alpha} | A_n^c\right\} = o(1) \text{ for large enough } M_\varepsilon. \end{aligned}$$

Combining steps 1-3 yields the stated result. ■

Remark 3.6.5 *Under the stronger condition in* (3.19) *and that*

$$\left\| \sqrt{\phi_1^2(x) + ... + \phi_J^2(x)} \right\|_d \leq \zeta_d(J),$$

we have

$$\begin{aligned}
\|\hat{m}(x) - m_0(x)\|_d &= \left\| \phi_J \hat{\beta}^J - m_0(x) \right\|_d \leq \left\| \phi_J \left(\hat{\beta}^J - \beta^J \right) \right\|_d + \left\| m_0(x) - \phi^J(x) \beta^J \right\|_d \\
&\leq \left\| \sqrt{\phi_1^2(x) + ... + \phi_J^2(x)} \times \left\| \hat{\beta}^J - \beta^J \right\| \right\|_d + O\left(J^{-\alpha}\right) \\
&= \left\| \sqrt{\phi_1^2(x) + ... + \phi_J^2(x)} \right\|_d \left\| \hat{\beta}^J - \beta^J \right\| + O\left(J^{-\alpha}\right) \\
&\leq \zeta_d(J) \left(O_p\left(\sqrt{\frac{J}{n}} + J^{-\alpha} \right) \right) + O\left(J^{-\alpha}\right) \\
&= O_p\left[\zeta_d(J) \left(\sqrt{\frac{J}{n}} + J^{-\alpha} \right) \right].
\end{aligned}$$

Remark 3.6.6 $\|\hat{m}(x) - m_0(x)\|_w^2$ *is related to the average squared error (ASE) defined by* $ASE = 1/n \sum_{i=1}^n \left[\hat{m}(X_i) - m(X_i)\right]^2$. *We have*

$$\begin{aligned}
ASE &= n^{-1} \sum_{i=1}^n \left[\hat{m}(X_i) - \phi^J(X_i) \beta^J + \phi^J(X_i) \beta^J - m(X_i) \right]^2 \\
&\leq 2n^{-1} \sum_{i=1}^n \left[\hat{m}(X_i) - \phi^J(X_i) \beta^J \right]^2 + 2n^{-1} \sum \left[\phi^J(X_i) \beta^J - m(X_i) \right]^2 \\
&= 2n^{-1} \sum_{i=1}^n \left[\phi^J(X_i) \left(\hat{\beta}^J - \beta^J \right) \right]^2 + O\left(J^{-2\alpha}\right) \\
&= 2n^{-1} \left(\hat{\beta}^J - \beta^J \right)' \left(\Phi_J' \Phi_J \right) \left(\hat{\beta}^J - \beta^J \right) + O\left(J^{-2\alpha}\right) \\
&= 2n^{-1} \left[u + \mathbf{m} - \Phi_J \beta^J \right]' \Phi_J \left(\Phi_J' \Phi_J \right)^{-1} \Phi_J' \left[u + \mathbf{m} - \Phi_J \beta^J \right] + O\left(J^{-2\alpha}\right) \\
&= O_p\left(\frac{J}{n} + J^{-2\alpha} \right)
\end{aligned}$$

which is the same as the rate for $\|\hat{m}(x) - m_0(x)\|_w^2$. *More refined analysis shows that* $\|\hat{m}(x) - m_0(x)\|_w^2 = ASE\left(1 + o_p(1)\right)$.

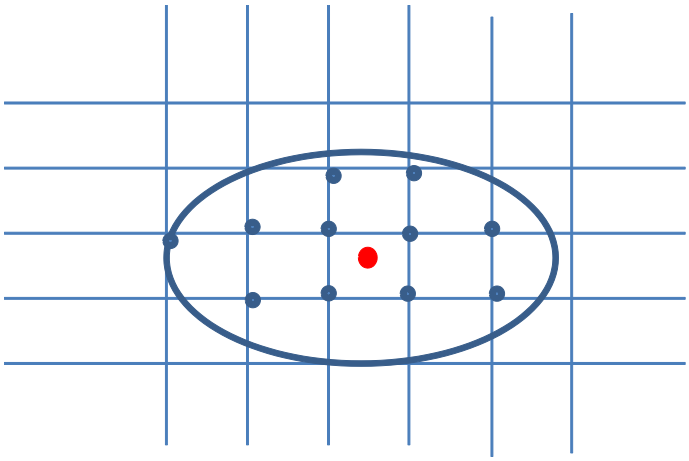


Figure 3.5: $\mathcal{B}_{J_n}$ and m_0

3.6.2 Local Geometry

Theorem 3.6.1 implies that $\hat{m} \in \mathcal{B}_{J_n} := \mathcal{B} \cap \mathcal{M}_{J_n}$ with probability approaching one, where

$$\mathcal{B} \equiv \{\|m - m_0\|_w \leq \xi_n \log(\log(n))\} \tag{3.20}$$

and $\xi_n = \left(\sqrt{J/n} + J^{-\alpha}\right)$. See Figure 3.5 for an illustration.

Consider the space $\mathcal{V} = \mathcal{B} - m_0$. Define the inner product on $\mathcal{V}$ as

$$\langle m_1 - m_0, m_2 - m_0 \rangle = E\left[m_1(X) - m_0(X)\right]\left[m_2\left(X\right) - m_0(X)\right].$$

Assume that $\mathcal{V}$ is a Hilbert space under the above inner product. We have $\|m - m_0\|_w^2 = \langle m_1 - m_0, m_2 - m_0 \rangle$.

3.6.3 Asymptotic Normality of Evaluation Functionals of $\hat{m}(\cdot)$

We proceed to establish the asymptotic normality of $\rho(\hat{m})$ for some functional $\rho(\cdot)$. For example, $\rho(m) := \rho_x(m) = m(x)$ for some given x or $\rho(m) := \rho_g(m) = \int m(x)\, g(x)\, dx$ for some function $g(x)$.

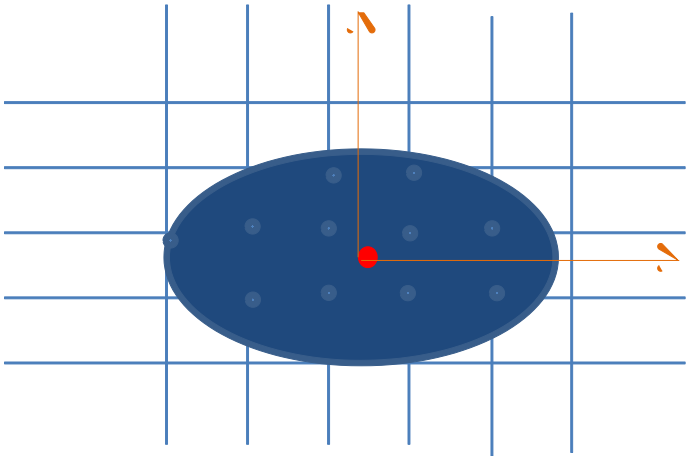


Figure 3.6: $\mathcal{V} = \mathcal{B} - m_0$.

We first consider evaluation functionals. We note

$$\begin{aligned}\sqrt{\frac{n}{J}}\left[\hat{m}(x) - m_0(x)\right] &= \sqrt{\frac{n}{J}}\phi^J(x)\left(\hat{\beta}^J - \beta^J\right) + \sqrt{\frac{n}{J}}\left[\phi^J(x)\beta^J - m_0(x)\right] \\ &= \sqrt{\frac{n}{J}}\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J' u \\ &\quad + \sqrt{\frac{n}{J}}\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m} - \Phi_J\beta^J\right) + \underbrace{o\left(1\right)}_{Bias}\end{aligned} \tag{3.21}$$

where $\mathbf{m} = (m(X_1), ..., m(X_2))'$. Here the $o(1)$ term holds provided that $\sqrt{\frac{n}{J}}J^{-\alpha} \to 0$. We now deal with the above two terms.

Let

$$V_J = \phi^J(x)Q_J^{-1}\Omega_J Q_J^{-1}\left[\phi^J(x)\right]'$$

where

$$\begin{aligned}\Omega_J &= E\left[\phi^J(X_i)'\phi^J(X_i)\sigma^2\left(X_i\right)\right] = n^{-1}E\left(\Phi_J' u u'\Phi_J\right), \\ Q_J &= E\left[\phi^J(X_i)'\phi^J(X_i)\right] = \frac{1}{n}E\left(\Phi_J'\Phi_J\right).\end{aligned}$$

For the first test term, we have

$$\begin{aligned}\sqrt{\frac{n}{J}}\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'u &= \sqrt{\frac{n}{J}}\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\sum_{i=1}^n\phi^J\left(X_i\right)u_i \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\underbrace{\frac{1}{\sqrt{J}}\phi^J(x)\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}\left[\phi^J\left(X_i\right)\right]'}u_i \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n K(x,X_i)u_i \to^d N(0,V),\end{aligned}$$

where

$$K(u,v)=\frac{1}{\sqrt{J}}\phi^J(u)\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}\left[\phi^J\left(v\right)\right]'=\frac{1}{\sqrt{J}}\sum_{j=1}^J\phi_j(u)\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}\phi_j(v) \tag{3.22}$$

is the generalized kernel function and

$$V=\lim_{n\to\infty}\frac{V_J}{J}=\lim_{n\to\infty}\frac{1}{J}\phi^J(x)Q_J^{-1}\Omega_JQ_J^{-1}\left[\phi^J(x)\right]'.$$

Here the asymptotic normality follows from Liapunov's CLT. To verify the CLT conditions, we assume $E\left[(Y_i-m_0(X_i))^4|X_i\right]$ is bounded and V is well-defined. For more discussion on the generalized kernel function, see Chen, Liao and Sun (2014).

For the second term, we have

$$\begin{aligned}&\left|\sqrt{\frac{n}{J}}\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m}-\Phi_J\beta^J\right)\right|^2 \\ =\ &\left|\sqrt{\frac{n}{J}}\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1/2}\left(\Phi_J'\Phi_J\right)^{-1/2}\Phi_J'\left(\mathbf{m}-\Phi_J\beta^J\right)\right|^2 \\ =\ &\frac{n}{J}\left\langle\left(\Phi_J'\Phi_J\right)^{-1/2}\phi^J(x),\left(\Phi_J'\Phi_J\right)^{-1/2}\Phi_J'\left(\mathbf{m}-\Phi_J\beta^J\right)\right\rangle_{\mathbb{R}^J} \\ \le\ &\frac{n}{J}\left[\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\phi^J(x)'\right]\left(\mathbf{m}-\Phi_J\beta^J\right)\Phi_J\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m}-\Phi_J\beta^J\right) \\ \le\ &\frac{n}{J}\left[\phi^J(x)\left(\Phi_J'\Phi_J\right)^{-1}\phi^J(x)'\right]\left\|\mathbf{m}-\Phi_J\beta^J\right\|^2 \\ =\ &\left[\frac{1}{J}\phi^J(x)\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}\phi^J(x)'\right]O_p\left(nJ^{-2\alpha}\right).\end{aligned}$$

To ensure the above term is $o\left(V\right)$, we need to assume $nJ^{-2\alpha}\to 0$ and

$$\frac{1}{J}\phi^J(x)\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}\phi^J(x)'=O_p\left(V\right).$$

The latter is satisfied if $var(Y_i|X_i)$ is bounded away from zero.

We gather the assumptions in Assumption 3 below:

Assumption 3 (i) $E\left[(Y_i - m_0(X_i))^4|X_i\right]$ is bounded,
(ii) $nJ^{-2\alpha} = o(1)$,
(iii) $var(Y_i|X_i)$ is bounded away from zero.

Theorem 3.6.2 *Under Assumptions 1-3, we have*

$$\sqrt{n}V_J^{-1/2}\left[\hat{m}(x) - m_0(x)\right] \to_d N(0,1).$$

3.6.4 Asymptotic Normality of General Functionals of $\hat{m}(\cdot)$

We consider a general functional $\rho(m)$ such that $\rho(m) - \rho(m_0)$ may not be linear. To derive the asymptotic distribution of $\rho(\hat{m})$, we assume that it can be approximated by a linear functional:

$$\rho(m) - \rho(m_0) \doteq \frac{\partial \rho(m_0)}{\partial m}[m - m_0]$$

in some sense, where the rhs is defined as follows

$$\frac{\partial \rho(m_0)}{\partial m}[m - m_0] = \lim_{\tau \to 0} \frac{\rho(m_0 + \tau(m - m_0)) - \rho(m_0)}{\tau}$$

and m_0 is the true function. More precisely, we require

$$\left|\rho(m) - \rho(m_0) - \frac{\partial \rho(m_0)}{\partial m}[m - m_0]\right| \leq C\|m - m_0\|_w^2 \tag{3.23}$$

for $m \in \mathcal{B}$ and some constant C not depending on m. For example, if $\rho(m) = \int m(x)\omega(x)dx$ for some weighting function $\omega(x)$. Then

$$\frac{\partial \rho(m_0)}{\partial m}[m - m_0] = \int \left[(m(x) - m_0(x))\right]\omega(x)dx.$$

More generally, the pathwise/directional derivative of $\rho(m)$ at $m = m_0$ in the direction of ϕ is defined to be

$$\frac{\partial \rho(m_0)}{\partial m}[\phi] = \lim_{\tau \to 0} \frac{\rho(m_0 + \tau\phi) - \rho(m_0)}{\tau}$$

for any $\phi \in \mathcal{M}$.

Now using (3.23), we have

$$\begin{aligned}\rho(\hat{m})-\rho(m_0) &= \frac{\partial\rho(m_0)}{\partial m}[\hat{m}-m_0]+\underbrace{\rho(\hat{m})-\rho(m_0)-\frac{\partial\rho(m_0)}{\partial m}[\hat{m}-m_0]}\\ &= \frac{\partial\rho(m_0)}{\partial m}[\hat{m}-m_0]+O_p\left(\frac{J}{n}+J^{-2\alpha}\right)\end{aligned}$$

The linear functional $\frac{\partial\rho(m_0)}{\partial m}[\cdot]$ on $\mathcal{V}$ may and may not be bounded with respect to the norm $\|m-m_0\|_w=\left\{E\left[m(X)-m_0(X)\right]^2\right\}^{1/2}$ on $\mathcal{V}$.

We write

$$\begin{aligned}&\frac{\partial\rho(m_0)}{\partial m}[\hat{m}-m_0]\\ =\ &\frac{\partial\rho(m_0)}{\partial m}\left[\phi^J(x)\beta^J-m_0\right]+\frac{\partial\rho(m_0)}{\partial m}[\hat{m}-\phi^J(x)\beta^J]:=I_1+I_2.\end{aligned}$$

Under the assumption that $\frac{\partial\rho(m_0)}{\partial m}[\cdot]$ is continuous wrt the sup-norm $\|\cdot\|_\infty$, we have

$$|I_1|=\frac{\partial\rho(m_0)}{\partial m}\left[\phi^J(x)\beta^J-m_0\right]=O\left(\left\|\phi^J(x)\beta^J-m_0\right\|_\infty\right)=O(J^{-\alpha}).$$

It remains to deal with I_2. Let $D_J=\left\{\frac{\partial\rho(m_0)}{\partial m}\left[\phi^J(x)\right]\right\}'$, then

$$\begin{aligned}I_2 &= \frac{\partial\rho(m_0)}{\partial m}[\hat{m}-\phi^J(x)\beta^J]=\frac{\partial\rho(m_0)}{\partial m}[\phi^J(x)\hat{\beta}^J-\phi^J(x)\beta^J]\\ &= \frac{\partial\rho(m_0)}{\partial m}\left[\phi^J(x)\right]\left(\hat{\beta}^J-\beta^J\right)=D_J\left(\hat{\beta}^J-\beta^J\right)\\ &= D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'u+D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m}-\Phi_J\beta^J\right).\end{aligned}$$

The two expressions are similar to those in (3.21).

To add more details, we write

$$\begin{aligned}I_2 &= \frac{1}{n}D_J'Q_J^{-1}\Phi_J'u+\frac{1}{n}D_J'\left(\hat{Q}_J^{-1}-Q_J^{-1}\right)\Phi_J'u+D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m}-\Phi_J\beta^J\right)\\ &:= I_{21}+I_{22}+I_{23}.\end{aligned}$$

Let $V_J=D_J'Q_J^{-1}\Omega_JQ_J^{-1}D_J$, which is a scalar. It is not hard to show that

$$\begin{aligned}\frac{\sqrt{n}I_{21}}{\sqrt{V_J}} &= \frac{1}{\sqrt{nV_J}}D_J'Q_J^{-1}\Phi_J'u=\frac{1}{\sqrt{nV_J}}D_J'Q_J^{-1}\sum_{i=1}^n\left[\phi^J(X_i)\right]'u_i\\ &= \frac{1}{\sqrt{nV_J}}\sum_{i=1}^nD_J'Q_J^{-1}\left[\phi^J(X_i)\right]'u_i=\frac{1}{\sqrt{n}}\sum_{i=1}^nK_D(X_i)u_i\\ &\to {}^dN(0,1),\end{aligned}$$

where $K_D(X_i) = \sum_{j=1}^{J} D_j Q_J^{-1} \phi_j(X_i) / \sqrt{V_J}$. The rate of convergence is determined by the magnitude of V_J.

To show that I_{22} and I_{23} are smaller order than I_{21}, we need to evaluate the order of V_J. Under the condition that $var(Y_i|X_i) \geq c > 0$, we have for any $a \in \mathbb{R}^J$,

$$a'\Omega_J a = E\left\{\left[\phi^J(X_i)a\right]^2 \sigma^2(X_i)\right\} \geq ca'E\left[\phi^J(X_i)'\phi^J(X_i)\right]a = ca'Q_J a$$

and so

$$V_J = \left(D_J'Q_J^{-1}\right)\Omega_J\left(Q_J^{-1}D_J\right) \geq cD_J'Q_J^{-1}D_J \geq c_*\|D_J\|^2$$

for some $c_* > 0$. Combining the same argument with $var(Y_i|X_i) \leq C < \infty$, we have

$$V_J \leq C^*\|D_J\|^2$$

for some $C^* < \infty$. We have therefore shown that $V_J \propto \|D_J\|^2$.

It remains to show that $I_{22} = o_p\left(\sqrt{V_J}/\sqrt{n}\right)$ and $I_{23} = o_p\left(\sqrt{V_J}/\sqrt{n}\right)$. That is

$$\begin{aligned}
\sqrt{n}V_J^{-1/2}\frac{1}{n}D_J'\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]\left(\Phi_J'u\right) &= o_p(1), \\
\sqrt{n}V_J^{-1/2}D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m} - \Phi_J\beta^J\right) &= o_p(1).
\end{aligned}$$

The first equation is easy to show, as

$$\begin{aligned}
& E\left(\left\|\sqrt{n}V_J^{-1/2}\frac{1}{n}D_J'\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]\left(\Phi_J'u\right)\right\|^2 \cdot 1_{A_n^c}\right) \\
=\ & EV_J^{-1}D_J'\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]\frac{\Phi_J'E(uu'|X)\Phi_J}{n}\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]D_JV_J^{-1} \cdot 1_{A_n^c} \\
\leq\ & CEV_J^{-1}D_J'\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]\hat{Q}_J\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]D_JV_J^{-1} \cdot 1_{A_n^c} \\
\leq\ & C\left\|V_J^{-1/2}D_J\right\|^2 E\left\|\left[\hat{Q}_J^{-1} - Q_J^{-1}\right]\right\|^2 \\
\leq\ & C\left\|V_J^{-1/2}D_J\right\|^2 \frac{J}{n}\zeta^2(J) = O\left(\frac{J}{n}\zeta^2(J)\right) = o(1).
\end{aligned}$$

For the second equation, we have:

$$\begin{aligned}
& \left\|\sqrt{n}V_J^{-1/2}D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m} - \Phi_J\beta^J\right)\right\|^2 \\
\leq\ & n\left\|V_J^{-1/2}D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\right\|^2\left\|\left(\mathbf{m} - \Phi_J\beta^J\right)\right\|^2 \\
=\ & V_J^{-1}D_J'\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}D_JO_p\left(nJ^{-2\alpha}\right) = O_p\left(V_J^{-1}D_J'Q_J^{-1}D_J\right)O_p\left(nJ^{-2\alpha}\right).
\end{aligned}$$

So

$$\left\|\sqrt{n}V_J^{-1/2}D_J'\left(\Phi_J'\Phi_J\right)^{-1}\Phi_J'\left(\mathbf{m} - \Phi_J\beta^J\right)\right\|^2 = O_p\left(nJ^{-2\alpha}\right) = o_p(1).$$

Theorem 3.6.3 *Let Assumptions 1-3 hold. In addition (i)*

$$\left|\rho(m)-\rho(m_0)-\frac{\partial\rho(m_0)}{\partial m}[m-m_0]\right|\leq C\|m-m_0\|_w^2$$

for $m\in\mathcal{B}_{J_n}$ and for some constant C not depending on m. (ii) $\frac{\partial\rho(m_0)}{\partial m}[\cdot]$ is a linear functional on $\mathcal{V}$ (iii) For some constant C

$$\frac{\partial\rho(m_0)}{\partial m}\left[\phi^J(x)\beta^J-m_0\right]\leq C\left\|\phi^J(x)\beta^J-m_0\right\|_\infty.$$

(iii)$\sqrt{n}\|D_J\|^{-1}J/n=o(1)$. Then

$$\sqrt{n}V_J^{-1/2}\left(\rho(\hat{m})-\rho(m)\right)\to N(0,1),$$

where $V_J=D_J'Q_J^{-1}\Omega_JQ_J^{-1}D_J$ and $D_J=\frac{\partial\rho(m_0)}{\partial m}\left[\phi^J(x)\right]$.

Remark 3.6.7 *The above condition (iii) and Assumption 3(ii) ensure that $\sqrt{n}V_J^{-1/2}(O_p\left(\frac{J}{n}+J^{-2\alpha}\right))=o_p(1)$.*

Remark 3.6.8 *When the functional $\frac{\partial\rho(m_0)}{\partial m}[\cdot]$ is unbounded on $\mathcal{V}$, that is*

$$\sup_{m\in\mathcal{B}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}[m-m_0]\right|}{\|m-m_0\|_w}=\infty,$$

it is typical that $\|D_J\|_2\to\infty$ as $J\to\infty$. To see this, we first note

$$\begin{aligned}
\sup_{m\in\mathcal{B}_{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}[m-m_0]\right|}{\|m-m_0\|_w} &= \sup_{m\in\mathcal{B}_{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}[m-\Pi_{J_n}m_0+\Pi_{J_n}m_0-m_0]\right|}{\|m-\Pi_{J_n}m_0+\Pi_{J_n}m_0-m_0\|_w}\\
&\leq \sup_{m\in\mathcal{B}_{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}[m-\Pi_{J_n}m_0]\right|+O(J_n^{-\alpha})}{\|m-\Pi_{J_n}m_0\|_w-O\left(J_n^{-\alpha}\right)}\\
&= \sup_{m\in\mathcal{B}_{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}[m-\Pi_{J_n}m_0]\right|}{\|m-\Pi_{J_n}m_0\|_w}+o(1)\\
&= \sup_{\beta^{J_n}\in\mathbb{R}^{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}\left[\phi^{J_n}\beta^{J_n}\right]\right|}{\left\|\phi^{J_n}\beta^{J_n}\right\|_w}+o(1).
\end{aligned}$$

Similarly we can show that

$$\sup_{m\in\mathcal{B}_{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}[m-m_0]\right|}{\|m-m_0\|_w}\geq\sup_{\beta^{J_n}\in\mathbb{R}^{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}\left[\phi^{J_n}\beta^{J_n}\right]\right|}{\left\|\phi^{J_n}\beta^{J_n}\right\|_w}+o(1).$$

Hence

$$\sup_{m\in\mathcal{B}_{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}\left[m-m_0\right]\right|}{\left\|m-m_0\right\|_w}=\sup_{\beta^{J_n}\in\mathbb{R}^{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}\left[\phi^{J_n}\beta^{J_n}\right]\right|}{\left\|\phi^{J_n}\beta^{J_n}\right\|_w}+o\left(1\right)$$

Since the lhs diverges to ∞ *as* $J_n\to\infty$*, we have*

$$\sup_{\beta^{J_n}\in\mathbb{R}^{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}\left[\phi^{J_n}\beta^{J_n}\right]\right|}{\left\|\phi^{J_n}\beta^{J_n}\right\|_w}\to\infty\ as\ J_n\to\infty.$$

But

$$\sup_{\beta^{J_n}\in\mathbb{R}^{J_n}}\frac{\left|\frac{\partial\rho(m_0)}{\partial m}\left[\phi^{J_n}\beta^{J_n}\right]\right|}{\left\|\phi^{J_n}\beta^{J_n}\right\|_w}=\sup_{\beta^{J_n}\in\mathbb{R}^{J_n}}\frac{D_J'\beta^{J_n}}{\left\|\beta^{J_n}\right\|_2}=\left\|D_J\right\|_2,$$

and so $\left\|D_J\right\|_2$ *as* $J_n\to\infty$*.*

Remark 3.6.9 *When the functional* $\frac{\partial\rho(m_0)}{\partial m}\left[\cdot\right]$ *is bounded on* $\mathcal{V}$*, we can use the same argument above to show that*

$$\left\|D_J\right\|_2\to D<\infty.$$

Remark 3.6.10 *The rate of convergence is given by* $\sqrt{n}/\left\|D_J\right\|_2$*. For bounded functionals,* $\left\|D_J\right\|_2$ *remains bounded as* $J\to\infty$*. So the rate of convergence is* $\sqrt{n}$*. For unbounded functionals,* $\left\|D_J\right\|_2\to\infty$ *as* $J\to\infty$*. So the rate of convergence is slower than*$\sqrt{n}$*. The bottom line: If* $\rho\left(m\right)-\rho(m_0)$ *can be approximated well enough by a linear functional and the approximating linear functional is bounded with respect to the norm* $\left\|m-m_0\right\|_w=\left\{E\left(\left[m\left(X\right)-m_0\left(X\right)\right]^2\right)\right\}^{1/2}$*, then* $\rho\left(m\right)$ *is* $\sqrt{n}$*-estimable.*

3.6.5 Asymptotic Normality of Bounded Functionals: Further Remarks

When $\frac{\partial\rho(m_0)}{\partial m}[m-m_0]$ is a linear and bounded functional with respect to the norm $\left\|m-m_0\right\|_w$, we can invoke the Riesz representation theorem to show that there exists a $\nu^*\left(x\right)\in\mathcal{V}$ such that

$$\frac{\partial\rho\left(m_0\right)}{\partial m}[m-m_0]=E\left[m\left(X\right)-m_0\left(X\right)\right]\nu^*\left(X\right).$$

In this case

$$\begin{aligned}I_2&=\frac{\partial\rho\left(m_0\right)}{\partial m}[\hat{m}-\phi^J(x)\beta^J]=\frac{\partial\rho\left(m_0\right)}{\partial m}[\hat{m}-m_0]+\frac{\partial\rho\left(m_0\right)}{\partial m}[\phi^J(x)\beta^J-m_0]\\&=E\left[\hat{m}\left(X\right)-\phi^J(X)\beta^J\right]\nu^*\left(X\right)=\left[E\phi^J(X)\nu^*\left(X\right)\right]\left(\hat{\beta}^J-\beta^J\right)\end{aligned}$$

and

$$D_J=E\left[\phi^J(X)'\nu^*\left(X\right)\right].$$

So

$$\begin{aligned} V_J &= D_J' Q_J^{-1} E\left\{\left[\phi^J(X)\right]' \phi^J(X)\sigma^2(X)\right\} Q_J^{-1} D_J \\ &= E\left\{\left[\phi^J(X) Q_J^{-1} D_J\right]' \left[\phi^J(X) Q_J^{-1} D_J\right] \sigma^2(X)\right\} := E\left\{\left[\nu_J^*(X)\right]^2 \sigma^2(X)\right\} \end{aligned}$$

where

$$\nu_J^*(X) = \phi^J(X) Q_J^{-1} D_J = \phi^J(X)\left(E\left[\phi^J(X)'\phi^J(X)\right]\right)^{-1} E\left[\phi^J(X)'\nu^*(X)\right]$$

is the projection of $\nu^*(X)$ onto the space spanned by $\phi^J(X)$. So

$$V_J \leq CE\left[\nu_J^*(X)\right]^2 \leq CE\left[\nu^*(X)\right]^2 < \infty.$$

In fact, $V_J \to E\left\{\left[\nu^*(X)\right]^2 \sigma^2(X)\right\}$. As a result, $\rho(\hat{m})$ is $\sqrt{n}$ consistent.

Remark 3.6.11 *In the case of unbounded evaluation functionals, $\rho(\hat{m}) - \rho(m_0) = \hat{m}(x) - m_0(x) = \phi^J(x)\left(\hat{\beta}^J - \beta^J\right)(1 + o_p(1))$. That is, $\rho(\hat{m}) - \rho(m_0)$ is asymptotically equivalent to a linear combination of $\left(\hat{\beta}^J - \beta^J\right)$ with weights that may not decay. In contrast, in the case of bounded linear functionals, $\rho(\hat{m}) - \rho(m_0)$ is asymptotically equivalent to $\left[E\phi^J(X)\nu^*(X)\right]\left(\hat{\beta}^J - \beta^J\right)$. When $E\left[\phi^J(X)\right]'\phi^J(X) = I_J$, $\left[E\phi^J(X)\nu^*(X)\right]$ is the vector of projection coefficients on the space spanned by $\phi^J(X)$. Given that $\left\|\left[E\phi^J(X)\nu^*(X)\right]\right\|^2 < \infty$, $\rho(\hat{m}) - \rho(m_0)$ is asymptotically equivalent to a linear combination of $\left(\hat{\beta}^J - \beta^J\right)$ with weights decaying to zero as they are square-summable.*

Remark 3.6.12 *The series estimator is based on OLS with the objective function $n^{-1}\sum\left[Y_i - m(X_i)\right]^2$. The objective function converges to*

$$\begin{aligned} E\left(Y - m(X)\right)^2 &= E\left[m_0(X) + u - m(X)\right]^2 \\ &= E\left[m(X) - m_0(X)\right]^2 + const \\ &= \|m - m_0\|_w^2 + const. \end{aligned}$$

So in the limit we learn how far away $m(X)$ is from $m_0(X)$ as measured by the metric $\|\cdot\|_w$. Any linear functional on $\mathcal{V}$ that is bounded wrt the norm $\|m - m_0\|_w$ is $\sqrt{n}$ estimable. Otherwise, it is not estimable at the $\sqrt{n}$ rate.

For more details on the results in this section, see Andrews (1991) and Newey (1997).

3.7 IV Regression with Nonparametric First Stage

To be done.

3.8 Bibliographical Remarks

Wahba (1990), Eubank (1999), and Gu (2002) give detailed treatments of smoothing and regression splines. See Gu (2002) and Eggermont and LaRiccia (2001) for more details on the penalized maximum likelihood approach to density estimation and conditional moment estimation. The RKHS approach is very popular in the machine learning literature but it has not attracted much attention in econometrics. There are many books on this approach often with "kernel methods" in the title. Recent overview of the method of sieves can be found in Chen (2006). Series estimators of regression functions have been considered in Andrews (1991) and Newey (1997). Recently, Chen, Liao and Sun (2014) develop a unified framework to establish the asymptotic properties of plug-in estimators of bounded and unbounded functionals. Both iid data and time series data are considered.

3.9 Problems

1. Prove (3.8).

Hint: First, solve for θ_{n-1} (and θ_n) from the two constraints

$$\begin{aligned}\theta_{n-1}+\theta_n &= -\sum_{j=1}^{n-2}\theta_j \\ \theta_{n-1}X_{n-1}+\theta_n X_n &= -\sum_{j=1}^{n-2}\theta_j X_j\end{aligned}$$

and obtain

$$\theta_{n-1}=-\frac{\sum_{j=1}^{n-2}\theta_j\left(X_n-X_j\right)}{\left(X_n-X_{n-1}\right)}.$$

Second, plugging θ_n and θ_{n-1} into $\beta_0+\beta_1 x+\sum_{j=1}^{n}\theta_j(x-X_j)_+^3$ yields

$$\begin{aligned}&\beta_0+\beta_1 x+\sum_{j=1}^{n}\theta_j(x-X_j)_+^3 \\ =\ &\beta_0+\beta_1 x+\sum_{j=1}^{n-2}\theta_j\left(X_n-X_j\right)\left[\frac{(x-X_j)_+^3-(x-X_n)_+^3}{(X_n-X_j)}-\frac{\left[(x-X_{n-1})_+^3-(x-X_n)_+^3\right]}{(X_n-X_{n-1})}\right].\end{aligned}$$

Finally, define γ_j in terms of θ_j,β_0,β_1. It then follows that (3.8) holds.

2. (i) Compute $K_\perp(x,y)=\int_0^1 G_2(x,u)\,G_2(y,u)\,du$ in (3.9). (ii) Show that $K_\perp(x,y)$ is cubic in x when $x\in[0,y]$ and is linear in x when $x\in[y,1]$. (iii) As a function of x,

show that $K_\perp(x, X_i)$ is continuously differentiable up to the second order. (iv) Assume $X_1 \leq X_2 \leq ... \leq X_n$. Show that $K_\perp(x, X_i)$ is linear in x when $x > X_n$

3. Consider the problem

$$\arg \min_{m \in W_2[0,1]} R_\lambda(m) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - m(X_i)]^2 + \lambda J(m) \tag{3.24}$$

where $J(\hat{m}) = \int \hat{m}''(u)^2 \, du$. It has been shown that the solution can be represented by

$$m(x) = \sum_{i=1}^{n} a_i K_\perp(x, X_i) + b_0 + b_1 x.$$

Show that $m(x)$ is a natural cubic spline when a_i and b_i are chosen to solve (3.24) (Hint: Use the results in problem 2)

4. Prove that

$$K(x, y) = (1 - |x - y|)_+$$

is a positive semidefinite kernel on $[0, 1]^2$ (This kernel is related to the Bartlett kernel in the HAC estimation). Hint: Mercer's representation for $K(x, y)$ can be obtained by a Fourier series expansion of $(1 - |z|)_+$ for $z \in [0, 1]$.

5. Consider the estimator

$$\hat{\beta}_e = \arg \min_{\beta} \left(\frac{1}{n} \|Y - X\beta\|^2 + \lambda \sum_{i=1}^{p} \left[\alpha \beta_i^2 + (1 - \alpha) |\beta_i| \right] \right)$$

where $X'X/n = I_p$. Derive the relationship between $\hat{\beta}_e$ and $\hat{\beta}_{OLS}$.

6. Consider the linear model

$$Y_i = \sum_{k=1}^{p} X_{ki} \beta_k + \varepsilon_i, i = 1, ..., n$$

where $X_i = (X_{1,i}, ..., X_{p,i})' \sim iidN(0, I_p)$, $\varepsilon_i \sim iidN(0, 1)$ and $\{X_i\}$ are independent of $\{\varepsilon_i\}$. Let $\beta_k = 1$ for all $k = 1, ..., p$

(i) Simulation the data with $n = 100$ and $p = 100$.

(ii) Let $m = n/2$. Fit a linear regression model to the first half of the data, i.e. $\{X_i, Y_i\}_{i=1}^{m}$ by means of the ridge and lasso with $\lambda_{ridge} = 0.64$ and $\lambda_{lasso} = 0.24$. (Note: these values of λ are obtained by cross validation based on the same DGP. You can obtain and employ your own cross validated λ for each procedure if you prefer)

(iii) Use the ridge and lasso estimators in (ii) and the covariates $\{X_i\}_{i=m+1}^n$ to obtain the predicted values $\hat{Y}_i$ for $i = m+1, ..., n$.

(iv) Plot Y_i against $\hat{Y}_i$ for both the ridge and lasso procedures. Which procedure does a better job in the sense of having a smaller $\frac{1}{n-m}\sum_{i=m+1}^n \left(Y_i - \hat{Y}_i\right)^2$?

7. Consider the same model as in Problem 6. Now we let $\beta_k = k$ for $k = 1, 2, ..., 5$ and $\beta_k = 0$ for $k > 5$. Repeat (i)-(iv) in Problem 6 but with $\lambda_{ridge} = 0.50$ and $\lambda_{lasso} = 0.98$. Comparing the results here and those in Problem 6, what can you conclude?

8. Suppose that $X_i \sim iid$ uniform$[0,1]$ and $\phi_j(x) = \sqrt{2}\sin\left[\left(j - \frac{1}{2}\right)\pi x\right]$. Define the generalized kernel as in (3.22):

$$K_J(u,v) = \frac{1}{\sqrt{J}}\sum_{j=1}^J \phi_j(u)\left(\frac{\Phi_J'\Phi_J}{n}\right)^{-1}\phi_j(v). \tag{3.25}$$

Let $n = 10000$. Simulate and graph $K_J(0.5, v)$ as a function of v for $v \in [0,1]$ for $J = 10, 30, 50$ and 70. Discuss your findings (i.e., the shape of $K_J(0.5, v)$ for each J and change in the shapes as J increases).

9. Let Y and $X \in R$ and define $m_0(x) = E[Y|X = x]$. Let $\hat{m}(x)$ be the series estimator for $m_0(x)$. Suppose we estimate the derivative of $m_0'(x^*)$ by $\hat{m}'(x^*)$. Obtain the rate of convergence for $\hat{m}'(x^*) - m_0'(x^*)$. (Impose any conditions if necessary).

10. Let Y and $X \in \mathbb{R}$. We want to estimate $\int_{\text{supp}(X)} E[Y|X = x]dx$. Under what conditions, it is can be estimated at the parametric $\sqrt{n}$ rate? Please provide conditions that are as weak as possible.

3.10 References

1. Andrews, D.W.K. (1991): Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models. *Econometrica* 59, 307-45.

2. Andrews, D.W.K. and Y. J. Whang (1990): Additive interactive regression models: circumvention of the curse of dimensionality, *Econometric Theory* 6: 466-479.

3. Aronszajn N. (1950): Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, Vol. 68, No. 3, pp. 337-404.

4. Chen, X. (2006): Large Sample Sieve Estimation of Semi-Nonparametric Models. *Handbook of Econometrics* Vol. 6 (eds. J.J. Heckman & E.E. Leamer). North-Holland.

5. Chen, X., Liao, Z and Y. Sun (2014): Sieve Inference on Possibly Misspecified Semi-nonparametric Time Series Models, *Journal of Econometrics*, Volume 178(3), pp. 639-658

6. Chen, X. and X. Shen (1998): Sieve Extremum Estimates for Weakly Dependent Data, *Econometrica*, Vol. 66, No. 2, pp. 289-314.

7. Eggermont, P.P.B. and V.N. LaRiccia (2001): *Maximum Penalized Likelihood Estimation*, Volume I, Springer.

8. Eggermont, P.P.B. and V.N. LaRiccia (2009): *Maximum Penalized Likelihood Estimation*, Volume II, Springer.

9. Eubank, R.L. (1999): *Nonparametric Regression and Spline Smoothing.* Marcel Dekker.

10. Fenton, V.M. & A.R. Gallant (1996): Convergence Rates of SNP Density Estimators. *Econometrica* 64, 719-727.

11. Gallant, A.R. & D.W. Nychka (1987): Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica* 55, 363-390.

12. Gallant, A.R. and G. Tauchen (1996): Which Moments to Match? *Econometric Theory*, 12, 657-681.

13. Green, P. J., and Silverman, B.W. (1994): *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, New York: Chapman & Hall.

14. Gu, C. (2002): *Smoothing Spline ANOVA Models*, Springer.

15. Hastie, T., R. Tibshirani, and J. Friedman (2009): *Elements of Statistical Learning*, Springer.

16. Kimeldorf, G. and Wahba, G. (1971): Some results on Tchebycheffian Spline Functions, *J. Mathematical Analysis and Applications*, 33(1), 82-95.

17. Knight, K. and W. Fu (2000): Asymptotics for lasso-type estimators, *The Annals of Statistics*, 28, no. 5, 1356–1378

18. Newey, W.K. (1997): Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* 79, 147-168.

19. Schölkopf, B., and Smola, A. J. (2002), *Learning with Kernels*, Cambridge, MA: MIT Press.

20. Shen, X. and W.H. Wong (1994): Convergence Rate of Sieve Estimates, *The Annals of Statistics*, Vol. 22, No. 2., pp. 580-615.

21. Silverman (1984): Spline Smoothing: The Equivalent Variable Kernel Method, *The Annals of Statistics*, Vol. 12, No. 3, pp. 898-916.

22. Tibshirani, R. (1994): Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288.

23. Wahba, G. (1990): *Spline Models for Observational Data.* Cbms-Nsf Regional Conference Series in Applied Mathematics, 59.

24. White, H. (2006): Approximate Nonlinear Forecasting Methods, in *Handbook of Economic Forecasting.* pp. 460–512.

Chapter 4

Examples of Semiparametric Models

4.1 Introduction

In terms of the trade-off between precision and misspecification, semiparametric models are intermediate cases lying between the two extremes: the fully parametric and fully nonparametric model. There are many different definitions of semiparametric models. The linear regression model without imposing a parametric distribution assumption on the error term can be regarded as a semiparametric model. Here we call a model semiparametric if it contains both a nonparametric component and a parametric component that need to be estimated. So we are interested in estimating models which are characterized by a finite-dimensional parameter, $\beta \in \beta \subseteq \mathbb{R}^k$, and infinite-dimensional ones, τ and h, which are normally functions.

Semiparametric models can be seen as a compromise between the high flexibility and robustness of the nonparametric case, and the faster convergence rate obtained in the parametric one: a fully nonparametric model will be more robust than semiparametric and parametric models since it doesn't suffer from the risk of misspecification. On the other hand, nonparametric estimators suffer from low convergence rates, which deteriorate as one consider higher order derivatives and/or higher dimension data. In contrast, the parametric model carries a risk of misspecification, but if it is correctly specified, it will normally enjoy $\sqrt{n}$-consistency with no deterioration caused by derivatives and/or multi-dimensional data. The idea of a semiparametric model is to take the best of both worlds.

We give a number of examples of semiparametric models in this chapter.

4.2 Traditional semiparametric models

In the previous chapters, we considered the general regression model,

$$Y = f(X) + u, E[u|X] = 0 \tag{4.1}$$

where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$ and the function $f : \mathbb{R}^d \to \mathbb{R}$ is unknown.

In the fully parametric case, we assumed that the regression function was known up to some finite-dimensional parameter β, $f(x) = f(x;\beta)$. We derived regularity conditions under which the NLLS estimator β was $\sqrt{n}$-consistent and asymptotically normally distributed. This in turn implies that $\hat{f}(x) = f(x;\hat{\beta})$ is $\sqrt{n}$-consistent and asymptotically normal under suitable smoothness conditions. However, the parametric family may be misspecified in which case the estimator is inconsistent. This leads us to consider the fully nonparametric case, where we derived kernel and sieve estimators of f. These estimators are very robust, but unfortunately suffer from a lower precision relative to parametric estimators: we found that the optimal convergence rate is $n^{-2/(d+4)}$ which is slower than $\sqrt{n}$. Moreover, the precision of the nonparametric estimator is influenced by the dimension of X, $d \geq 1$. As d increases, the convergence rate of the nonparametric estimator deteriorates.

In the following, we present a number of semiparametric models which still allows for high degree of flexibility of the model while improving on the convergence rate.

4.2.1 Partially Linear Model

A semiparametric partially linear model (PLM) (Engle, Granger, Rice and Weiss (1986) and Robinson (1988)) is defined by

$$Y = X\beta_0 + f_0(V) + u,\ E(u|(X,V)) = 0 \tag{4.2}$$

where $X \in \mathbb{R}^{d_1}$, $V \in \mathbb{R}^{d_2}$, and $d = d_1 + d_2$. The function form $f_0(\cdot)$ is not specified. The finite dimensional parameters β_0 constitute the parametric part of the model and the unknown function $f_0(\cdot)$ the nonparametric part. By imposing the partial linear structure, we reduce the nonparametric dimension from d to d_2.

The partially linear model is one of the most popular semiparametric models. It can be regarded as a first step in relaxing the parametric function form in $E(Y|X,V)$. To motivate the model, we consider the control function approach in causal inference. We have a linear causal model with endogeneity:

$$Y = X\beta_0 + \varepsilon$$

where X is the causal variable of interest. A control variable V is a variable that satisfies the conditional mean independence assumption:

$$E(\varepsilon|X,V) = E(\varepsilon|V).$$

Under this assumption, we have

$$E(Y|X,V) = X\beta_0 + E(\varepsilon|V).$$

Suppose that we do not want to make any parametric assumption on $E(\varepsilon|V)$. Writing $E(\varepsilon|V) = f_0(V)$ for a nonparametric function $f_0(V)$, we then have

$$Y = X\beta_0 + f_0(V) + u, \tag{4.3}$$

where $u = Y - E(Y|X,V) = \varepsilon - E(\varepsilon|V)$ satisfies $E(u|(X,V)) = 0$.

As a second example, consider the regression discontinuity design with an effect independent of the running variable V. Using the standard notation in the potential outcomes framework, we have

$$\begin{aligned} E(Y_i|V_i, D_i = 0) &= E(Y_i(0)|V_i) := f_0(V_i) \\ E(Y_i|V_i, D_i = 1) &= E(Y_i(1)|V_i) = f_0(V_i) + \beta_0 \end{aligned}$$

for some nonparametric function $f_0(\cdot)$. So

$$E(Y_i|V_i, D_i) = D_i\beta_0 + f_0(V_i)$$

and we can write

$$Y_i = D_i\beta_0 + f_0(V_i) + u$$

where $E(u|D_i, V_i) = 0$. This is a partially linear model where the linear component involves a dummy variable .

To identify the partially linear model, we have to impose some restrictions: (i) X can not consist of a constant because the constant will be absorbed by $f_0(\cdot)$. (ii) None of the component of X is a deterministic function of V because deterministic functions of V will be absorbed into $f_0(\cdot)$.

To estimate this model, we take conditional expectation of both sides of equation (4.2), leading to

$$E(Y|V) = E(X|V)\beta_0 + f_0(V).$$

Consequently

$$Y - E(Y|V) = [X - E(X|V)]\beta_0 + u.$$

So if we know $E(Y|V) := h_{10}(V)$ and $E(X|V) := h_{20}(V)$, we can do least squares on this equation, and the moment equation that generates our estimator is

$$Q(\beta) = Em(Z, h_{10}(V), h_{20}(V), \beta) = 0$$

where $Z = (X, V, Y)$ and

$$m(Z, h_{10}(V), h_{20}(V), \beta) = \{Y - h_{10}(V) - [X - h_{20}(V)]\beta\}[X - h_{20}(V)]$$

The semiparametric analogue of this is to obtain a nonparametric estimate of the two regression functions, $h_{10}(V), h_{20}(V)$, say $h_{1n}(V), h_{2n}(V)$ in a first stage, and then, after substituting these two estimates for their true values, find the value of β that minimizes a norm in

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, h_{1n}(V_i), h_{2n}(V_i), \beta).$$

Partially linear models can be extended to partial parametric models:

$$Y = g(X, \beta_0) + f_0(V) + u,\ E(u|X) = 0 \tag{4.4}$$

where $g(X, \beta_0)$ is a known parametric function and $f_0(V)$ is, as before, a nonparametric function. For more details, see Andrews (1994, sec 3.2).

4.2.2 The Single Index Model

The second semiparametric regression model we will consider is the single index model (SIM, e.g., Ichimura (1993) and Klein and Spady (1993)). It is given by

$$Y = f_0(X\beta_0) + u, E[u|X] = 0, \tag{4.5}$$

where the function $f_0 : \mathbb{R} \to \mathbb{R}$ and the parameter $\beta_0 \in \mathbb{R}^d$ are unknown. Observe that f_0 now has $\mathbb{R}$ as its domain in contrast to before where its domain was $\mathbb{R}^d$. So this removes the curse of dimensionality from the problem. The SIM in (4.5) is contained in the general model (4.1). The name single-index comes from the fact that f_0 here is a function of the scalar $X\beta_0$ instead of the vector X. In this case, our infinite-dimensional parameter is $f_0(\cdot)$.

Semiparametric single index models arises naturally in binary choice settings. A binary choice model may be represented by

$$Y = 1\{X\beta_0 + \varepsilon > 0\}$$

where ε is independent of X. Equivalently,

$$Y = 1 - F_0(-X\beta_0) + u := f_0(X\beta_0) + u,$$

$E[u|X] = 0$ and $var(u|X) = F_0(-X\beta_0)(1 - F_0(-X\beta_0))$ where F_0 is the CDF of ε. When ε is a standard normal RV, we get the probit model. When ε is a standard logistic RV, we get the logit model. When we do not want to specify the distribution of ε, we obtain a semiparametric single index model.

The SIM can not be identified without further restriction. First, β_0 is not identified if $f_0(\cdot)$ is a linear function. Second, as in the linear regression case, β_0 is not identified if X is perfectly multicollinear. Third, $f_0(\cdot)$ is not identified if X contains no continuous random

variables. Intuitively, when X contains only discrete variables, the support of $X\beta_0$ is finite. As a result, there exist many functions $f_0(\cdot)$ and many choices of β_0 that satisfy the finite set of restrictions. Fourth, X can not contain a constant. That is, β_0 can not contain a location parameter and β_0 is identified only up to scale. This follows because for any nonzero constants α_1 and α_2, and for any function $f_0(\cdot)$, we can always find another function, say $\tilde{f}_0(\cdot)$, defined by $\tilde{f}_0(\alpha_1 + \alpha_2 X\beta_0) = f_0(X\beta_0)$. To identify the model, we usually set one of the coefficients β_0 equal to one. However, we can do so only if we know that this coefficient is not zero. More generally, we can set $\|\beta_0\| = 1$.

Let

$$h_0(\beta, \gamma) = E(Y|X\beta = \gamma),$$

then

$$f_0(X\beta_0) = h_0(\beta_0, X\beta_0).$$

The tricky thing is that $f_0(\cdot)$, as a function, depends on β_0. So a more strict notation would be $h_0(\beta_0, X\beta_0) := f_{\beta_0}(X\beta_0)$. That is, we put a subscript β_0 on f. In general, we use $h_0(\beta, \gamma)$ to denote the unknown nonparametric function.

Next, we wish to set up an estimator of β_0 and h_0. If the function h_0 was known, then an obvious estimator of β would be the NLLS one,

$$\widehat{\beta}_n^{NLLS} = \arg\min \frac{1}{n} \sum_{i=1}^{n} (Y_i - h_0(\beta, X_i\beta))^2$$

The population moment condition is

$$Q(\beta) = Em(Z, h_0(\beta, X_i\beta)) := E\frac{\partial h_0(\beta, X\beta)}{\partial \beta} [Y - h_0(\beta, X\beta)] = 0.$$

Since $h_0(\beta, X_i\beta)$ is unknown, this is not a feasible estimator. The semiparametric analogue is to obtain a nonparametric estimate of $h_0(\beta, X_i\beta)$ in a first stage, say

$$h_n(\beta, X_i\beta) = \frac{\sum_{j\neq i}^{n} Y_j K(\frac{X_j\beta - X_i\beta}{b})}{\sum_{j\neq i}^{n} K(\frac{X_j\beta - X_i\beta}{b})},$$

and then, after substituting $h_n(\beta, X_i\beta)$ for $h_0(\beta, X_i\beta)$, find the value of β that minimizes a norm in

$$Q_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, h_n(\beta, X_i\beta)) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial h_n(\beta, X_i\beta)}{\partial \beta} [Y_i - h_n(\beta, X_i\beta)].$$

One can extend the single index model to the following more general class of models,

$$Y = f_0(\tau(X, \beta_0)) + u, E[u|X] = 0, \tag{4.6}$$

for some function $\tau : \mathbb{R} \times B \to \mathbb{R}$, which is known up to β_0. The estimation strategy mentioned above carries through to this more general setting.

4.2.3 Nonlinear model with nonparametric heteroskedasticity

Consider the standard nonlinear regression model:

$$Y = f_0(X, \beta_0) + u, E[u|X] = 0,$$

where the errors are heteroskedastic:

$$E\left(u^2|X\right) = \sigma^2(X).$$

The standard NLS estimator

$$\hat{\beta}_{NLS} = \arg\min \frac{1}{n} \sum_{i=1}^{n} [Y_i - f_0(X_i, \beta)]^2$$

is consistent and asymptotically normally distributed, but not asymptotically efficient. If the conditional variance function $\sigma^2(X)$ is known, then we can do weighted least squares (WLS),

$$\tilde{\beta}_{WLS} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \frac{[Y_i - f_0(X_i, \beta)]^2}{\sigma^2(X_i)} \tag{4.7}$$

which is more efficient than $\hat{\beta}_{NLS}$. The first order condition for this problem is

$$Q_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f_0(X_i, \beta)}{\partial \beta} \frac{[Y_i - f_0(X_i, \beta)]}{h(X_i)} := \frac{1}{n} \sum_{i=1}^{n} m(Z_i, h(X_i), \beta)$$

where $h(X_i) = \sigma^2(X_i)$. So the population moment condition is

$$Q(\beta) = Em(Z_i, h_0(X_i), \beta_0) = 0.$$

If we do not know $\sigma^2(X_i)$, we can estimate it in a preliminary stage, form $m(Z_i, h_n(X_i), \beta)$ and "approximate" sample analogue to the population moment condition, and then choose β to minimize a norm of this sample analogue.

Note that there are at least two ways of doing this, and they have a somewhat different structure. In the first way, we follow the steps below:

1. Obtain $\hat{\beta}_{NLS}$

2. (a) Calculate the associated residuals, $\hat{u}_i = Y_i - X_i\hat{\beta}_{NLS}$, $i = 1, 2, ..., n$.

(b) Estimate the conditional variance nonparametrically, e.g.

$$h_n\left(X_i, \hat{\beta}_{NLS}\right) := \frac{\sum_{j \neq i} K_b(X_j - X_i)\hat{u}_j^2}{\sum_{j \neq i} K_b(X_j - X_i)}$$

3. Obtain $\hat{\beta}_{WLS}$ by solving

$$\hat{\beta}_{WLS} = \arg\min \left\| \frac{1}{n} \sum_{i=1}^{n} m \left[Z_i, h_n \left(X_i, \hat{\beta}_{NLS} \right), \beta \right] \right\|$$

Alternatively, we could recompute the weighting function for different values of β, in which case we define

$$\hat{\beta}_{WLS} = \arg\min \left\| \frac{1}{n} \sum_{i=1}^{n} m(Z_i, h_n \left(X_i, \beta \right), \beta) \right\|$$

Note that now the nonparametric function is indexed by β. These types of nesting structures happen in some empirical applications.

4.2.4 Selectivity Models

Consider the following set of latent variable equations

$$\begin{aligned} Y_1^* &= f_0(X_1, \beta_0) + u_1, \\ Y_2^* &= -q_0(X_2) + u_2, \end{aligned}$$

where (Y_1^*, Y_2^*) are unobserved latent variables. Moreover, we assume that

$$u_1 = \rho u_2 + \varepsilon$$

where u_2 and ε are independent. The observed variables are

$$\begin{aligned} Y_1 &= Y_1^* \mathbf{1} \{ Y_2^* > 0 \} \\ Y_2 &= \mathbf{1} \{ Y_2^* > 0 \}. \end{aligned}$$

The selection problem arises because the unobservables in the selection equation depends on the unobservables in the "equation of interest." As a result, as long as X_1 and X_2 are correlated (and they often contain the same variables), we would expect regression estimates of the first equation to provide biased estimates of the parameters of interest.

Very similar setups to this one can be found in labor, in the literature on the evaluation of experiments, and in I.O. An example from labor is when there is a preliminary equation determining whether a person works, and then an equation determining what a person would earn (or how many hours a person will work) should that person be working. In industrial organization we have equations determining production given that a firm decides to operate during the period, and an equation determining whether a firm should operate. In both cases it is assumed that there are unobservables which affect both the equation determining whether the agent is active, and what the agent would obtain were that agent to be active.

Here is one set of assumptions that makes it easy to illustrate how this problem can be analyzed with a semiparametric estimator.

1. Let $X = (X_1, X_2)$ then

$$Pr\{u_2 \leq v | X = x\} = F_0(v), \text{ for any } (v, x),$$

i.e., u_2 distributes independently of X.

2. F_0 strictly increasing, and

3. ε distributes independently of X and u_2.

We maintain Assumption 1 for simplicity. If X contains variables which are at least partly subject to choice, the assumption that the distribution of u_2 is independent of them is not very realistic. This assumption can be relaxed, see Olley and Pakes (1996).

Under the above assumptions,

$$\begin{aligned} E(Y_1|X, Y_2 &= 1) = f_0(X_1, \beta_0) + E(u_1|u_2 > q_0(X_2), X) \\ &= f_0(X_1, \beta_0) + \rho E(u_2|u_2 > q_0(X_2)) \\ &= f_0(X_1, \beta_0) + \frac{\rho}{1 - F_0(q_0(X_2))} \int_{q_0(X_2)}^{\infty} u_2 dF_0(u_2). \end{aligned}$$

Let

$$\frac{\rho}{1 - F_0(q_0(X_2))} \int_{q_0(X_2)}^{\infty} u_2 dF_0(u_2) := G(q_0(X_2)),$$

for some function $G(\cdot)$. Then

$$E(Y_1|X, Y_2 = 1) = f_0(X_1, \beta_0) + G(q_0(X_2)).$$

Also let

$$P(Y_2 = 1|X) = P(u > q_0(X_2)|X) = 1 - F_0(q_0(X_2)) := \tilde{F}_0(q_0(X_2)) := \tau_0(X_2)$$

with $\tilde{F}$ strictly increasing. This implies that we can invert this relationship to find

$$q_0(X_2) = \tilde{F}_0^{-1}\left[\tilde{F}_0(q_0(X_2))\right].$$

Consequently,

$$\begin{aligned} E(Y_1|X, Y_2 &= 1) = f_0(X_1, \beta_0) + G\left\{\tilde{F}_0^{-1}\left[\tilde{F}_0(q_0(X_2))\right]\right\} \\ &:= f_0(X_1, \beta_0) + h_0(\tau_0(X_2)) \end{aligned}$$

where $h_0(\cdot) = G_0(\tilde{F}_0^{-1}(\cdot))$ and $\tau_0(\cdot) = \tilde{F}_0(q_0(\cdot))$. With some abuse of noation, we can rewrite $h_0(\tau_0(X_2))$ as

$$h_0(\tau_0(X_2)) = h_0(\tau_0(X_2), \beta_0)$$

where

$$h_0\left(\tau_0\left(X_2\right),\beta\right) = E(Y_1 - f_0(X_1,\beta)|\tau_0\left(X_2\right), Y_2 = 1).$$

Now if we knew $h_0\left(\cdot,\beta\right)$ and $\tau_0(\cdot)$ we could do nonlinear least squares on this equation to provide consistent estimates of β. The moment condition defining the estimator in this case can be written as

$$Q(\beta) = Em(Z, h_0\left(\tau_0\left(X_2\right),\beta\right),\beta) = 0$$

where

$$m(Z, h_0\left(\tau_0\left(X_2\right),\beta\right),\beta) = \frac{\partial\left[f(X_1,\beta) + h_0\left(\tau_0\left(X_2\right),\beta\right)\right]}{\partial\beta}\left[Y_1 - f(X_1,\beta) - h_0\left(\tau_0\left(X_2\right),\beta\right)\right].$$

We do not know either h_0 or τ_0. Hence we proceed in steps. In the first step we provide a nonparametric estimate of the selection equation, thus producing $\tau_n\left(X_i\right)$. In the next step we run a partially nonlinear model with $\tau_n\left(X_i\right)$ as the nuisance regressor. The parameter β can then be estimated by

$$\hat{\beta} = \arg\min_{\beta\in\beta}\left\|\frac{1}{n}\sum_{i:Y_{2i}=1} m(Z_i, h_n\left(\tau_n\left(X_{2i}\right),\beta\right),\beta)\right\|$$

where for example

$$h_n\left(\tau_n\left(x\right),\beta\right) = \frac{\sum_{i=1}^n\left[Y_{1i} - f(X_{1i},\beta)\right]K_h\left(\tau_n\left(X_{2i}\right) - \tau_n\left(x\right)\right)1\left\{Y_{2i}=1\right\}}{\sum_{i=1}^n K_h\left(\tau_n\left(X_{2i}\right) - \tau_n\left(x\right)\right)1\left\{Y_{2i}=1\right\}}.$$

There are several points that should be noted here:

- We need no particular assumptions on $F(\cdot)$, other than it be strictly increasing, or on $h(\cdot)$.
- The special case $X_1 = X_2$ poses no particular problems here. The identification does not rely on a variable that is included in one equation and is excluded from another. Rather it relies on the fact that one equation sets a parametric model while the nuisance variable enters in another equation.
- This methodology does not separate out the constant term in the equation of interest, from the constant term in $G(\cdot)$ and sometimes the constant term can be of interest.

Some common features of the models in sec 4.2.1 to sec 4.2.4 are

1. a preliminary estimator of the nonparametric functions are available, most often in closed forms

2. the finite dimensional parameters β can be estimated by

$$\hat{\beta} = \arg\min \|Q_n(\beta)\|$$

for

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, h_n(Z_i, \tau_n(Z_i), \beta), \beta)$$

3. the nonparametric estimators h_n and τ_n are uniformly consistent with certain convergence rate.

In the next section, we introduce some models that do not share the above features. In particular, the estimators of unknown functions may not be consistent under the strong norm (sup norm). Instead, they are consistent under certain weak norm.

4.3 Nonparametric Regression with Endogeneity

Example[1] *(Engel curve)*: Blundell et al. (2003) have shown that a system of Engel curves that satisfies Slutsky's symmetry condition and allows for demographic effects on budget shares in a given year must take the following form:

$$Y_{1\ell i} = h_{1\ell}(Y_{2i} - h_0(X_{1i})) + h_{2\ell}(X_{1i}) + \varepsilon_{\ell i}, \quad \ell = 1, ..., N,$$

where $Y_{1\ell i}$ is the *i-th* household budget share on ℓ*-th* goods, Y_{2i} is the *i-th* household log-total non-durable expenditure, X_{1i} is a vector of the i-th household demographic variables that affect the household's non-durable consumption. Note that $h_0(X_{1i})$ is common among all the goods and is called an "equivalence scale" in the consumer demand literature. Citing strong empirical evidence and many existing works, Blundell et al. (2003) have argued that popular parametric linear and quadratic forms for $h_{1\ell}(\cdot)$ are inadequate, and that consumer demand theory only suggests the purely nonparametric specification:

$$E[Y_{1\ell i} - \{h_{1\ell}(Y_{2i} - h_0(X_{1i})) + h_{2\ell}(X_{1i})\}|X_{1i}, Y_{2i}] = E[\varepsilon_{\ell i}|X_{1i}, Y_{2i}] = 0, \tag{4.8}$$

where $h_{1\ell}$, $h_{2\ell}$ and h_0 are all unknown functions. For the identification of all these unknown functions $\theta = (h_0, h_{11}, ..., h_{1N}, h_{21}, ..., h_{2N})'$ satisfying (4.8), it suffices to assume that at least one of $h_{1\ell}, \ell = 1, ..., N$ is nonlinear and that $h_{2\ell}(x_1^*) = 0$, $\ell = 1, ..., N$, for some x_1^* in the support of X_1.

Unfortunately, when X_{1i} contains too many household demographic variables (say when $\dim(X_{1i}) \geq 3$), the fully nonparametric specification (4.8) cannot lead to precise estimates of

[1] This section borrows liberally from Chen's handbook chapter.

the unknown functions $h_0, h_{21}, ..., h_{2N}$ due to the so-called "curse of dimensionality". Therefore, applied researchers must impose more structure on the model. Using the British family expenditure survey (FES) data, Blundell et al. (1998) found the following semi-nonparametric specification to be reasonable:

$$E[Y_{1\ell i} - \{h_{1\ell}(Y_{2i} - g(X_{1i}'\beta_1)) + X_{1i}'\beta_{2\ell}\}|X_{1i}, Y_{2i}] = 0, \tag{4.9}$$

where $h_{1\ell}, \ell = 1, ..., N$ are still unknown functions, but now $h_0(X_{1i}) = g(X_{1i}'\beta_1)$ and $h_{2\ell}(X_{1i}) = X_{1i}'\beta_{2\ell}$ are known up to unknown finite-dimensional parameters β_1 and $\beta_{2\ell}$. Here the parameters of interest are $\theta = (\beta_1', \beta_{21}', ..., \beta_{2N}', h_{11}, ..., h_{1N})'$. This semi-nonparametric specification has been estimated by Blundell et al. (1998) using the kernel method and Blundell et al. (2007) using the sieve method.

Both the specifications (4.8) and (4.9) assume that the total non-durable expenditure Y_{2i} is exogenous. However, this assumption has been rejected empirically. Noting the endogeneity of total non-durable expenditure, Blundell et al. (2007) considered the following semi-nonparametric instrumental variables (IV) regression:

$$E[Y_{1\ell i} - \{h_{1\ell}(Y_{2i} - g(X_{1i}'\beta_1)) + X_{1i}'\beta_{2\ell}\}|X_{1i}, X_{2i}] = 0, \tag{4.10}$$

where the parameters of interest are still $\theta = (\beta_1', \beta_{21}', ..., \beta_{2N}', h_{11}, ..., h_{1N})'$, and X_{2i} is the gross earnings of the head of the *i-th* household which is used as an instrument for the total non-durable expenditure Y_{2i}. They estimated this model via the sieve method and their empirical findings demonstrate the importance of accounting for the endogenous total expenditure semi-nonparametrically.

We note that the above example and many other economic models imply semi-nonparametric conditional moment restrictions of the form

$$E[\rho(Z_i; \theta_o)|X_i] = 0, \quad \theta_o \equiv (\beta_o', h_o')', \tag{4.11}$$

where $\rho(\cdot;\cdot)$ is a column vector of residual functions whose functional forms are known up to unknown parameters, $\theta \equiv (\beta', h')'$, and $\{Z_i' = (Y_i', X_i')\}_{i=1}^n$ is the data where Y_i is a vector of endogenous variables and X_i is a vector of conditioning variables. Here $E[\rho(Z_i, \theta)|X_i]$ denotes the conditional expectation of $\rho(Z_i, \theta)$ given X_i, and the true conditional distribution of Y_i given X_i is unspecified (and is treated as a nuisance function). The parameters of interest $\theta_o \equiv (\beta_o', h_o')'$ contain a vector of finite dimensional unknown parameters β_o and a vector of infinite dimensional unknown functions $h_o(\cdot) = (h_{o1}(\cdot), ..., h_{oq}(\cdot))'$, where the arguments of $h_{oj}(\cdot)$ could depend on Y, X, known index function $\delta_j(Z, \beta_o)$ up to unknown β_o, other unknown function $h_{ok}(\cdot)$ for $k \neq j$, or could also depend on unobserved random variables. Motivated by the asset pricing and rational expectations models, Hansen (1982) studied the conditional moment restriction $E[\rho(Z_t; \beta_o)|X_t] = 0$ (i.e., without unknown h_o) for stationary ergodic time series data (where typically $Z_t' = (Y_t', X_t')$ and X_t includes lagged Y_t and other

pre-determined variables known at time t). Newey and Powell (2003), Ai and Chen (2003) and Chen and Pouzo (2009) studied the general case $E[\rho(Z_t;\beta_o,h_o)|X_t]=0$ for i.i.d. data.

The semi-nonparametric conditional moment models given by (4.11) can be classified into two broad subclasses. The first subclass consists of *models without endogeneity* in the sense that $\rho(Z_i,\theta)-\rho(Z_i,\theta_o)$ does not depend on any endogenous variables (Y_i). In this case,

$$E\left(\rho(Z_i,\theta)-\rho(Z_i,\theta_o)|X_i\right)=\rho(Z_i,\theta)-\rho(Z_i,\theta_o)$$

hence the true parameter θ_o can be identified as the unique maximizer of

$$Q(\theta)=-E[\rho(Z_i,\theta)'\{\Sigma(X_i)\}^{-1}\rho(Z_i,\theta)],$$

where $\Sigma(X_i)$ is a positive definite weighting matrix. The second subclass consists of *models with endogeneity* in the sense that $\rho(Z_i,\theta)-\rho(Z_i,\theta_o)$ does depend on endogenous variables (Y_i). Here the true parameter θ_o can be identified as the unique maximizer of

$$Q(\theta)=-E[m(X_i,\theta)'\{\Sigma(X_i)\}^{-1}m(X_i,\theta)] \quad \text{with} \quad m(X_i,\theta)\equiv E[\rho(Z_i,\theta)|X_i].$$

Although the second subclass includes the first subclass as a special case, when θ contains unknown functions, it is much easier to derive asymptotic properties for various nonparametric estimators of θ identified by the conditional moment models belonging to the first subclass. The first subclass includes, as special cases, many semi-nonparametric regression models that have been well studied in econometrics. For example, it includes the specifications (4.8) and (4.9) of the Engel curve example , the partially linear regression $E[Y_i-X_{1i}'\beta_o-h_o(X_{2i})|X_{1i},X_{2i}]=0$, the index regression $E[Y_i-h_o(X_i'\beta_o)|X_i]=0$, the varying coefficient model $E[Y_i-\sum_{j=1}^q h_{oj}(D_{ji})X_{ji}\mid (D_{ki},X_{ki}),k=1,...,q]=0$ and the additive model with a known link (F) function $E[Y_i-F(\sum_{j=1}^q h_{oj}(X_{ji}))|X_{1i},...,X_{qi}]=0$.

The second subclass includes, as a special case, the specification (4.10) in the Engle curve example. A leading, yet difficult example of this subclass, is the purely nonparametric instrumental variables (IV) regression $E[Y_{1i}-h_o(Y_{2i})|X_i]=0$ studied by Newey and Powell (2003), Hall and Horowitz (2005) and Carrasco et al. (2006), among others. A more difficult example is the nonparametric IV quantile regression $E[1\{Y_{1i}\le h_o(Y_{2i})\}-\gamma|X_i]=0$ for some known $\gamma\in(0,1)$ considered by Chernozhukov and Hansen (2006), and Chen and Pouzo (2009). Kaplan and Sun (2014) consider a smoother version of the moment conditions underlying the quantile regression. They show the practical and theoretical advantages of the SEE (smoothed estimating equation) approach. See Blundell and Powell (2003), Florens (2003), Newey and Powell (2003), Carrasco et al. (2006) and Chen and Pouzo (2009) for additional examples.

4.4 Bibiographical Remarks

For models in this chapter, most of the results derived in the literature use kernel estimators for the nonparametric part. However, one can in most cases substitute this for other nonparametric estimators such as the sieve ones.

Bickel et al (1993) treat all the models in sections 4.2.1 to 4.2.4, and many more in detail, but the book is fairly technical and therefore not a very good starting point for further reading on semiparametric models.

Robinson (1988) considers the estimation of the partially linear model. Andrews (1994) give results for the extended version. We will study partially linear models in the detail in the next chapter. The index model is treated in detail in Horowitz (1998); see also Ichimura (1993) and Klein and Spady (1993).

Ai and Chen (2003) and Blundell et al (2006) deal with nonparametric models with endogeneity. The estimators they consider belong to the class of sieve minimum distance estimators. Their approach is applicable to the models in sections 4.2.1 to 4.2.4 but involves quite different asymptotic arguments. We will study these two papers in a later chapter.

4.5 References

1. Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions." *Econometrica*, 71, 1795-1843.

2. Andrews, D.W.K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity." *Econometrica* 62, 43-72.

3. Bickel (1982): "On Adaptive Estimation." *Annals of Statistics*, Vol. 10, No. 3, 647-671.

4. Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. The John Hopkins University Press.

5. Blundell, R. and J. Powell (2003) "Endogeneity in Nonparametric and Semiparametric Regression Models", in M. Dewatripont, L.P. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications*, 2, 312-357, Cambridge: Cambridge University Press.

6. Blundell, R., M. Browning and I. Crawford (2003): "Non-parametric Engel Curves and Revealed Preference." *Econometrica*, 71, 205-240.

7. Blundell, R., X. Chen and D. Kristensen (2007): "Semiparametric Engel Curves with Endogenous Expenditure." *Econometrica*, Vol. 75, No. 6 (November, 2007), 1613–1669

8. Blundell, R., A. Duncan and K. Pendakur (1998): "Semiparametric Estimation and Consumer Demand." *Journal of Applied Econometrics*, 13, 435-461.

9. Carrasco, M., J.-P. Florens and E. Renault (2006): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization", in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6. North-Holland, Amsterdam.

10. Chen, X. and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals." *Journal of Econometrics*. Volume 152, Issue 1, Pages 46-60.

11. Chernozhukov, V. and C. Hansen (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics* 132 (2006) 491–525

12. Engle, Granger, Rice and Weiss (1986): "Semi-parametric estimates of the relation between weather and electricity demand," *Journal of American Statistical Association* 81 (1986): 310-320.

13. Florens, J.P. (2003) "Inverse Problems and Structural Econometrics: the Example of Instrumental Variables", in M. Dewatripont, L.P. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications*, 2, 284-311, Cambridge: Cambridge University Press.

14. Hansen, L.P., (1982): "Large Sample Properties of Generalized Methods of Moments Estimators." *Econometrica*, Vol. 50, page 1029-1054.

15. Hall, P. and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables", *Annals of Statistics*, 33, 2904-2929.

16. Horowitz (1998): *Semiparametric Methods in Econometrics*. Springer-Verlag.

17. Ichimura, H. (1993): "Semiparametic Least Squares (SLS) and Weighted SLS Estimationof Single-Index Models." *Journal of Econometrics* 58, 71-120.

18. Klein, R. and Spady, R. (1993): "An Efficient Semiparametric Estimator of Binary Response Models." *Econometrica* 61, 387-421.

19. Newey, W.K. and J.L Powell (2003) "Instrumental Variable Estimation of Nonparametric Models", *Econometrica,* 71, 1565-1578. Working paper version, 1989.

20. Olley, G.S. and Ariel Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry". *Econometrica*, Vol. 64, No. 6, 1263-1297.

21. Robinson, P.M. (1988): "Root-N-Consistent Semiparametric Regression." *Econometrica* 56, 931-954.

22. Kaplan, D. and Sun, Y. (2014): "Smoothed Estimating Equations for Instrumental Variables Quantile Regression." Working paper. Department of Economics, UC San Diego.

Chapter 5

Case Study: Partially Linear Model

5.1 Introduction

A semiparametric partially linear model (PLM) is defined by

$$Y_i = X_i\beta_0 + g(V_i) + u_i \tag{5.1}$$

where X_i is $1 \times p$ vector, β_0 is $p \times 1$ vector and $V_i \in \mathbb{R}^q$. The function form $g(\cdot)$ is not specified. The finite dimensional parameters β_0 constitute the parametric part of the model and the unknown function $g(\cdot)$ the nonparametric part. The data is assumed to be iid with $E(u_i|X_i, V_i) = 0$ and $E(u_i^2|X_i, V_i) = \sigma^2$. Conditional heteroskedasticity can be allowed but we consider the homoscedastic case for simplicity and transparency.

Partially linear models have many applications. Engle, Granger, Rice and Weiss (1986) are among the first to consider the partially linear model. They use data based on the monthly electricity sales Y_i for four cities, the monthly price of electricity X_1, income X_2, and average daily temperature t. They model the electricity demand as the sum of a smooth function of monthly temperature t, a linear function of X_1 and X_2, and 11 monthly dummy variables. That is, their model is:

$$Y = X_1\beta_1 + X_2\beta_2 + \sum_{k=1}^{11} \delta_k D_k + g(t) + u.$$

For more applications of PLMs, see the recent monograph of Hardle, Liang and Gao (2000), which also provides a thorough treatment of PLMs.

Following the work of Engle, Granger, Rice and Weiss (1986), much attention has been directed to estimating the partially linear model. For example, Engle, Granger, Rice and Weiss (1986) use the spline smoothing technique and defined the penalized estimators of β

and g as the solution of

$$\arg\min_{\beta,g} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - X_i\beta - g(V_i)\right]^2 + \lambda \int \left\{g''(u)\right\}^2 du$$

where λ is a penalty parameter.

We can also use the method of sieves to estimate the parameter β. We approximate $g(V)$ by a series expansion:

$$g(V) \approx \sum_{j=1}^{J_n} \phi_j(V)\,\theta_j$$

and regress Y_i on X_i and the basis functions $\left\{\phi_j(V_i)\right\}_{j=1}^{J_n}$. See 15.3 of Li and Racine (2007), which discusses this approach in some detail.

Robinson (1988) constructs a feasible least squares estimator of β based on estimating the nonparametric component by a Nadaraya-Waston kernel estimator. We will study Robinson's estimator in the next section. The proofs are based on Andrews (1994), who provides a general framework for proving the $\sqrt{n}$ consistency and asymptotic normality for a wide class of semiparametric estimators. Andrews names the estimators MINPIN because they are estimators that MINimize a criterion function that may depend on Preliminary Infinite-dimensional Nuisance parameter estimators.

The following presentation is inspired from Section 5.2 in Pagan and Ullah (1999) and Section 7.3 in Li and Racine (2007).

5.2 A Semiparametric Estimator

Taking the expectation of (5.1) conditional on V_i, we obtain

$$E(Y_i|V_i) = E(X_i|V_i) + g(V_i). \tag{5.2}$$

Subtracting (5.2) from (5.1) yields

$$Y_i - E(Y_i|V_i) = \left[X_i - E(X_i|V_i)\right]\beta_0 + u_i. \tag{5.3}$$

The conditional mean assumption that $E(u_i|X_i, V_i) = 0$ implies:

$$E\left[\left(X_i - q_i^X\right)\left(Y_i - q_i^Y - \left(X_i - q_i^X\right)\beta_0\right)\right] = 0$$

where $q_i^X = E(X_i|V_i)$ and $q_i^Y = E(Y_i|V_i)$. Let $q_i = (q_i^X, q_i^Y)'$ and

$$m_i(\beta, q) = \left(X_i - q_i^X\right)\left\{Y_i - q_i^Y - \left(X_i - q_i^X\right)\beta\right\}$$

then the moment condition can be compactly written as

$$Em_i(\beta_0, q_{0,i}) = 0. \tag{5.4}$$

For clarity, we sometimes add a subscript '0' to q_i to emphasize that it is the true conditional expectation. The sample analogue of (5.4) is

$$\bar{m}(\beta, q) = \frac{1}{n}\sum_{i=1}^{n} m_i(\beta, q_i) = 0.$$

We can define a semiparametric estimator $\tilde{\beta}$ of β as the parameter that solves the equation $\bar{m}(\beta, q) = 0$. Of course, this estimator is not feasible as we do not know q_i and has to replace it by some nonparametric estimator. Let

$$\begin{aligned}\hat{q}_i^Y &= \frac{1}{n}\sum_{j=1}^{n} Y_j K_h(V_i, V_j)/\hat{f}(V_i) \\ \hat{q}_i^X &= \frac{1}{n}\sum_{j=1}^{n} X_j K_h(V_i, V_j)/\hat{f}(V_i)\end{aligned}$$

where

$$\hat{f}(V_i) = \frac{1}{n}\sum_{j=1}^{n} K_h(V_i, V_j)$$

and

$$K_h(V_i, V_j) = \prod_{s=1}^{q} h_s^{-1} k\left(\frac{V_{is} - V_{js}}{h_s}\right).$$

Then the feasible semiparametric estimator $\hat{\beta}$ of β satisfies

$$\bar{m}(\hat{\beta}, \hat{q}) = \frac{1}{n}\sum_{i=1}^{n} m_i(\hat{\beta}, \hat{q}_i) = 0,$$

where $\hat{q} = (\hat{q}_1, ..., \hat{q}_n)$. A closed form solution of $\hat{\beta}$ is :

$$\hat{\beta} = \left[\sum_{i=1}^{n}(X_i - \hat{q}_i^X)'(X_i - \hat{q}_i^X)\right]^{-1} \sum_{i=1}^{n}(X_i - \hat{q}_i^X)'(Y_i - \hat{q}_i^Y).$$

This estimator can be regarded as the OLS estimator based on the equation

$$Y_i - \hat{q}_i^Y = \left[X_i - \hat{q}_i^X\right]\beta + error.$$

5.3 Asymptotics via Stochastic Equicontinuity

To examine the asymptotic properties of $\hat{\beta}$, we may expand $m_i(\hat{\beta}, \hat{q}_i)$ around $(\beta_0, \hat{q}_i)$ as in the parametric case, leading to

$$0 = \frac{1}{n}\sum_{i=1}^{n} m_i(\hat{\beta}, \hat{q}_i) = \frac{1}{n}\sum_{i=1}^{n} m_i(\beta_0, \hat{q}_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial m_i(\tilde{\beta}, \hat{q}_i)}{\partial \beta}\left(\hat{\beta} - \beta_0\right)$$

where $\tilde{\beta}$ is between $\hat{\beta}$ and β_0. So

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = -\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m_i(\tilde{\beta}, \hat{q}_i)}{\partial \beta}\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\beta_0, \hat{q}_i).$$

One may proceed to expand $m_i(\beta_0, \hat{q}_i)$ around $q_{0,i}$ as in the parametric case. However, $\hat{q}_i$ is now infinite-dimensional. A precise treatment is needed.

Lemma 5.3.1 *Assume that*

(i) $\hat{q} \to_p q_0 = (q_{0,1}, ..., q_{0,n})$ *with respect to some metric* $\rho(\cdot, \cdot)$, *for example*

$$\rho(q, q_0) = \lim_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n} E\left\|m_i(\beta_0, q_i) - m_i(\beta_0, q_{0,i})\right\|^s\right]^{1/s}, \ 1 \le s \le \infty$$

(ii) $n^{-1/2}\sum_{i=1}^{n} Em_i(\beta_0, \hat{q}_i) \to_p 0$ *where* $Em_i(\beta_0, \hat{q}_i) = Em_i(\beta_0, q_i)|_{q_i=\hat{q}_i}$.

(iii) $n^{-1/2}\sum_{i=1}^{n}[m_i(\beta_0, q_{0,i}) - Em_i(\beta_0, q_{0,i})] \to_d N(0, \Omega)$.

(iv) $v_n(q) = n^{-1/2}\sum_{i=1}^{n}[m_i(\beta_0, q_i) - Em_i(\beta_0, q_i)]$ *is stochastically equicontinuous with respect to the metric* $\rho(\cdot, \cdot)$.

Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m_i(\beta_0, \hat{q}_i) - Em_i(\beta_0, \hat{q}_i)] \to_d N(0, \Omega).$$

Note: $\{v_n(q)\}$ is stochastically equicontinuous on Q if $\forall \varepsilon > 0$, $\exists \delta > 0$ such that

$$\overline{\lim}_{n\to\infty} P(\sup_{q\in Q}\sup_{q'\in B(q,\delta)} |v_n(q') - v_n(q)| > \varepsilon) < \varepsilon.$$

where

$$B(q, \delta) = \left\{q' : \rho\left(q', q\right) \le \delta\right\}.$$

To understand the origin of these conditions, we write

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\beta_0, \hat{q}_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\beta_0, q_{0,i}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[m_i(\beta_0, \hat{q}_i) - m_i(\beta_0, q_{0,i})\right]$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\underline{[m_i(\beta_0, \hat{q}_i) - Em_i(\beta_0, \hat{q}_i)]} - \underline{[m_i(\beta_0, q_{0,i}) - Em_i(\beta_0, q_{0,i})]}\right\}$$

where $Em_i(\beta_0, q_{0,i}) = 0$. The last two terms need to be $o_p(1)$ for $\sqrt{n}(\hat{\beta}-\beta_0)$ to have a limiting normal distribution that does not depend on $\hat{q}$. The last term is $o_p(1)$ by the stochastic equicontinuity assumption in (iv). The second term is $o_p(1)$ if condition (ii) holds. Because $Em_i(\beta_0, \hat{q}_i)$ involves $\hat{q}_i$, in general cases, the second term may not be $o_p(1)$. Formally,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[m_i(\beta_0, \hat{q}_i) - m_i(\beta_0, q_{0,i})\right] \\
\approx\ &\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial q}Em_i(\beta_0, q_{0,i})\right]\sqrt{n}\left(\hat{q}_i - q_{0,i}\right) + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial q^2}Em_i(\beta_0, q_{0,i})\right]\sqrt{n}\left(\hat{q}_i - q_{0,i}\right)^2.
\end{aligned}$$

Thus condition (ii), sometimes refer to as the "asymptotic independence" condition, effectively combines in a single format the requirement that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial q}Em_i(\beta_0, q_{0,i})\right]\sqrt{n}\left(\hat{q}_i - q_{0,i}\right) \text{ and } \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial q^2}Em_i(\beta_0, q_{0,i})\right]\sqrt{n}\left(\hat{q}_i - q_{0,i}\right)^2$$

be $o_p(1)$ if the distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ is to be independent of $\hat{q}$. For this reason,

$$\sqrt{n}\left(\hat{q}_i - q_{0,i}\right)^2 = o_p(1) \text{ or } \hat{q}_i - q_{0,i} = o_p\left(n^{-1/4}\right)$$

uniformly over $i = 1, 2, ..., n$, arises frequently in later discussions. Note that when

$$n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial q^2}Em_i(\beta_0, q_{0,i}) = o_p(1),$$

the rate condition that $\hat{q}_i - q_{0,i} = o_p\left(n^{-1/4}\right)$ can be relaxed.

Observing that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial q}Em_i(\beta_0, q_{0,i})\right]\sqrt{n}\left(\hat{q}_i - q_{0,i}\right) \\
=\ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial q_i^X}\left(X_i - q_i^X\right)\left\{Y_i - q_i^Y - \left(X_i - q_i^X\right)\beta_0\right\}\left(\hat{q}_i^X - q_i^X\right) \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial q_i^Y}\left(X_i - q_i^X\right)\left\{Y_i - q_i^Y - \left(X_i - q_i^X\right)\beta_0\right\}\left(\hat{q}_i^Y - q_i^Y\right) \\
=\ &-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\hat{q}_i^X - q_i^X\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\left(\hat{q}_i^X - q_i^X\right)\beta_0 - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\left(\hat{q}_i^Y - q_i^Y\right)
\end{aligned}$$

where $v_i = \left(X_i - q_i^X\right)$. So if each of the above terms is $o_p(1)$, then

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial q}Em_i(\beta_0, q_{0,i})\right]\sqrt{n}\left(\hat{q}_i - q_{0,i}\right) = o_p(1).$$

In the above lemma, $Em_i(\beta_0, \hat{q}_i)$ should be read as follows. First, letting p_X and p_Y be arbitrary functions of V, we compute

$$\begin{aligned} Em_i(\beta_0, p) &= E\left(X_i - p_X(V_i)\right)\left\{Y_i - p_Y(V_i) - \left(X_i - p_X(V_i)\right)\beta_0\right\} \\ &= E\left(X_i - p_X(V_i)\right)\left\{Y_i - X_i\beta_0 - p_Y(V_i) + p_X(V_i)\beta_0\right\} \end{aligned}$$

with the expectation taken with respect to the density of X_i, Y_i and V_i. Note that

$$Y_i - X_i\beta_0 = E\left(Y_i|V_i\right) - E\left(X_i|V_i\right)\beta_0 + u_i$$

so

$$\begin{aligned} Em_i(\beta_0, p) &= E\left(X_i - p_X(V_i)\right)\left\{E\left(Y_i|V_i\right) - E\left(X_i|V_i\right)\beta_0 + u_i - p_Y(V_i) + p_X(V_i)\beta_0\right\} \\ &= E\left[E(X_i|V_i) - p_X(V_i)\right]\left\{\underbrace{\left[E\left(Y_i|V_i\right) - p_Y(V_i)\right]} - \underbrace{\left[E\left(X_i|V_i\right) - p_X(V_i)\right]\beta_0}\right\} \\ &= E\left(q_i^X - p_X(V_i)\right)\left[q_i^Y - p_Y(V_i) - \left(q_i^X - p_X(V_i)\right)\beta_0\right] \\ &= \int \left(q^X(v) - p_X(v)\right)\left[q^Y(v) - p_Y(v) - \left(q^X(v) - p_X(v)\right)\beta_0\right] f_V(v)\, dv. \end{aligned}$$

Second, we substitute $\hat{q}_i$, the specific nonparametric estimates of $E(X_i|V_i)$ and $E(Y_i|V_i)$, for $p_X(V_i)$ and $p_Y(V_i)$, giving

$$Em_i(\beta_0, \hat{q}_i) = \int \left(q^X(v) - \hat{q}^X(v)\right)\left[q^Y(v) - \hat{q}^Y(v) - \left(q^X(v) - \hat{q}^X(v)\right)\beta_0\right] f_V(v)\, dv$$

Theorem 5.3.1 *Let $\tilde{X}_i = X_i - E(X_i|V_i)$. Assume*

(i) (X_i, Y_i, V_i) is iid with $E(\tilde{X}_i u_i) = 0$, $E(u_i^2|\tilde{X}_i) = \sigma^2$,

(ii) $Eu_i^4 < \infty, E\left\|\tilde{X}_i\right\|^4 < \infty$,

(iii) CLT and LLN apply to $n^{-1/2}\sum \tilde{X}_i u_i$ and $n^{-1}\sum \tilde{X}_i\tilde{X}_i'$, respectively,

(iv) $E\left[\left(q_i^X - \hat{q}_i^X\right)\right]^4$, $E\left[\left(q_i^Y - \hat{q}_i^Y\right)\right]^4$, $E\left[n^{1/4}\left(q_i^X - \hat{q}_i^X\right)\right]^2$ and $E\left[n^{1/4}\left(q_i^Y - \hat{q}_i^Y\right)\right]^2$ are $o_p(1)$.

(v) Condition (iv) in Lemma 5.3.1 holds.

Then

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d N(0, \sigma^2 V^{-1})$$

where $V = E\left(\tilde{X}_i\tilde{X}_i'\right)$.

In the above conditions $E\left[\left(q_i^X - \hat{q}_i^X\right)\right]^4$ is defined to be

$$E\left[\left(q_i^X - \hat{q}_i^X\right)\right]^4 = \int \left[q\left(x\right) - \hat{q}\left(x\right)\right]^4 f_X\left(x\right) dx$$

other moments are similarly defined.

Proof (for scalar X_i and V_i) : Note that

$$\begin{aligned}
\hat{\beta} &= \left[\sum_{i=1}^n (X_i - \hat{q}_i^X)'(X_i - \hat{q}_i^X)\right]^{-1} \sum_{i=1}^n (X_i - \hat{q}_i^X)'(Y_i - \hat{q}_i^Y). \\
&= \left[\sum_{i=1}^n (X_i - \hat{q}_i^X)'(X_i - \hat{q}_i^X)\right]^{-1} \sum_{i=1}^n (X_i - \hat{q}_i^X)'\left[q_i^Y - \hat{q}_i^Y + \left(X_i - q_i^X\right)\beta_0 + u_i\right]
\end{aligned}$$

So

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta} - \beta_0\right) = \left[\frac{1}{n}\sum_{i=1}^n \hat{X}_i\hat{X}_i'\right]^{-1} \Bigg[& \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{X}_i(q_i^Y - \hat{q}_i^Y)}_{\text{Due to the estimation of } q_i^Y} \\
& \underbrace{-\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{X}_i\left(q_i^X - \hat{q}_i^X\right)\beta_0}_{\text{Due to the estimation of } q_i^X} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{X}_i u_i\Bigg]
\end{aligned}$$

We first consider the denominator, starting with

$$\begin{aligned}
n^{-1}\sum_{i=1}^n \hat{X}_i\hat{X}_i' &= \frac{1}{n}\sum_{i=1}^n \left(X_i - q_i^X + q_i^X - \hat{q}_i^X\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(\tilde{X}_i\right)^2 + \frac{2}{n}\sum_{i=1}^n \tilde{X}_i\left(q_i^X - \hat{q}_i^X\right) + \frac{1}{n}\sum_{i=1}^n \left(q_i^X - \hat{q}_i^X\right)^2.
\end{aligned}$$

Now the last term tends to zero by the results in previous chapters[1]. More specifically, we can show that the expectation of the dominant term in $n^{-1}\sum_{i=1}^n \left(q_i^X - \hat{q}_i^X\right)^2$ converges to zero. For more details, see the section that establishes the validity of cross-validation bandwidth choice. The absolute value of the second term is bounded by

$$\begin{aligned}
&\left|\frac{2}{n}\sum \tilde{X}_i\left(q_i^X - \hat{q}_i^X\right)\right| \\
&\leq 2\left(\frac{1}{n}\sum \tilde{X}_i^2\right)^{1/2}\left(\frac{1}{n}\sum \left(q_i^X - \hat{q}_i^X\right)^2\right)^{1/2} = o_p\left(1\right)
\end{aligned}$$

[1] When we estimate q_i^X, we could leave the observation (X_i, Y_i, Z_i) out to simplify the asymptotic analysis. In the sequel, we assume that leave-one-out estimators are used.

by the Cauchy inequality and the same argument for the last term. Finally, by assumption, $n^{-1}\sum_{i=1}^{n}\left(\tilde{X}_i\right)^2$ satisfies a LLN. We have therefore shown that

$$n^{-1}\sum_{i=1}^{n}\hat{X}_i\hat{X}_i' \to_p V = E\left(\tilde{X}_i\tilde{X}_i'\right).$$

Next, we consider the numerator. Let

$$m_i(\beta,\tilde{q}_i) = \left(X_i - \tilde{q}_i^X\right)\left(q_i^Y - \tilde{q}_i^Y\right) + \left(X_i - \tilde{q}_i^X\right)\left(\tilde{q}_i^X - q_i^X\right)\beta + \left(X_i - \tilde{q}_i^X\right)u_i$$

so that $m_i(\beta_0, q_{0,i}) = \tilde{X}_i u_i$. We apply Lemma 5.3.1 by verifying the conditions given above.

[Condition (i). For this condition to hold, we need to show that $\hat{q}_i^Y = q_i^Y + o_p(1)$ and $\hat{q}_i^X = q_i^X + o_p(1)$ in certain metric. Taking the metric in Lemma 5.3.1(i) with $s = 2$, we obtain

$$\begin{aligned}
\rho^2(\hat{q}, q_0) &= \lim n^{-1}\sum_{i=1}^{n} E\left[m_i(\beta_0,\hat{q}_i) - m_i(\beta_0, q_{0,i})\right]^2 \\
&= \lim n^{-1}\sum_{i=1}^{n} E\left[\hat{X}_i(q_i^Y - \hat{q}_i^Y) + \hat{X}_i\left(\hat{q}_i^X - q_i^X\right)\beta_0 + \hat{X}_i u_i - \tilde{X}_i u_i\right]^2 \\
&= E\left[\tilde{X}_i(q_i^Y - \hat{q}_i^Y) + \left(q_i^X - \hat{q}_i^X\right)\left(q_i^Y - \hat{q}_i^Y\right) + \tilde{X}_i\left(\hat{q}_i^X - q_i^X\right)\beta_0\right. \\
&\quad \left. - \left(q_i^X - \hat{q}_i^X\right)^2\beta_0 + \left(q_i^X - \hat{q}_i^X\right)u_i\right]^2
\end{aligned}$$

Using the Cauchy inequality twice, we have, for some constant C,

$$\begin{aligned}
&\rho^2(\hat{q}, q_0) \\
\leq\ & C\left\{E\left[\tilde{X}_i(q_i^Y - \hat{q}_i^Y)\right]^2 + E\left[\tilde{X}_i\left(\hat{q}_i^X - q_i^X\right)\beta_0\right]^2 + E\left[\left(q_i^X - \hat{q}_i^X\right)u_i\right]^2\right\} \\
&+C\left\{E\left[\left(q_i^X - \hat{q}_i^X\right)\left(q_i^Y - \hat{q}_i^Y\right)\right]^2 + E\left[\left(q_i^X - \hat{q}_i^X\right)^2\beta_0\right]^2\right\} \\
\leq\ & C\left[E\left(\tilde{X}_i\right)^4\right]^{1/2} E\left[(q_i^Y - \hat{q}_i^Y)^4\right]^{1/2} + C\left[E\left(\tilde{X}_i\right)^4\right]^{1/2}\left[E(q_i^X - \hat{q}_i^X)^4\right]^{1/2} \\
&+C\left[E\left(q_i^X - \hat{q}_i^X\right)^4\right]^{1/2}\left(Eu_i^4\right)^{1/2} + C\left[E\left(q_i^X - \hat{q}_i^X\right)^4\right]^{1/2}\left[E\left(q_i^Y - \hat{q}_i^Y\right)^4\right]^{1/2} \\
&+CE\left[\left(q_i^X - \hat{q}_i^X\right)^4\right]
\end{aligned}$$

But

$$E\left(\hat{X}_i\right)^4 = E\left(X_i - q_i^X + q_i^X - \hat{q}_i^X\right)^4 \leq C\left[E\left(\tilde{X}_i\right)^4 + E\left(q_i^X - \hat{q}_i^X\right)^4\right]$$

for some constant $C > 0$. Therefore, if $E(q_i^Y - \hat{q}_i^Y)^4 = o_p(1)$ and $E(q_i^X - \hat{q}_i^X)^4 = o_p(1)$, then $\rho^2(\hat{q}, q_0) \to 0$.

[**Condition (ii).** We have computed $Em_i(\beta_0, \hat{q}_i)$ before. It remains to show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left(q_i^X - \hat{q}_i^X\right)\left[q_i^Y - \hat{q}_i^Y - \left(q_i^X - \hat{q}_i^X\right)\beta_0\right] \to_p 0.$$

But the absolute value of the lhs is bounded by

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[E\left(q_i^X - \hat{q}_i^X\right)^2\right]^{1/2} \left[E\left(q_i^Y - \hat{q}_i^Y\right)^2\right]^{1/2} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left(q_i^X - \hat{q}_i^X\right)^2 |\beta_0| \\
= \;& \sqrt{n}\left[E\left(q_i^X - \hat{q}_i^X\right)^2\right]^{1/2} \left[E\left(q_i^Y - \hat{q}_i^Y\right)^2\right]^{1/2} + \sqrt{n} E\left(q_i^X - \hat{q}_i^X\right)^2 |\beta_0| \\
= \;& \left\{E\left[n^{1/4}\left(q_i^X - \hat{q}_i^X\right)\right]^2\right\}^{1/2} \left\{E\left[n^{1/4}\left(q_i^Y - \hat{q}_i^Y\right)\right]^2\right\}^{1/2} \\
& + \left\{E\left[n^{1/4}\left(q_i^X - \hat{q}_i^X\right)\right]^2\right\} |\beta_0|.
\end{aligned}$$

It is clear that $n^{1/4}$ consistency is indeed for the estimators $\hat{q}_i^X$ and $\hat{q}_i^Y$.

Condition (iii) is assumed

Condition (iv) is a technical condition that can be checked using the results in Andrews (1994). We will come back to stochastic equicontinuity later.

Using Lemma 5.3.1, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m_i(\beta_0, \hat{q}_i) - Em_i(\beta_0, \hat{q}_i)\right] \\
= \;& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{X}_i u_i + o_p(1) \to_d N(0, \sigma^2 V)
\end{aligned}$$

and

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d N(0, \sigma^2 V^{-1}),$$

which completes the proof. ■

Remark 5.3.1 *The Theorem shows that the nonparametric estimator has to converge at a fast enough rate to ensure the $\sqrt{n}$-consistency of $\hat{\beta}$. In the previous chapters, we have shown that, when a second order kernel is used, $\sqrt{nh^d}\left(\hat{q}_i^X - q_i^X\right)$ is stochastically bounded under the rate conditions that $h \to \infty$ and $nh^{4+d} \to C$ for some constant C. Hence $n^{1/4}\left(\left(\hat{q}_i^X - q_i^X\right)\right)$ converges to zero if $n^{1/4}\left(nh^d\right)^{-1/2} = n^{-1/4}h^{d/2}$ converges to zero, i.e. $nh^{2d} \to \infty$. This*

may contradict with $nh^{4+d} \to C$. *To resolve this problem, we can use higher order kernels in order to achieve bias reduction and a faster rate of convergence. When a q-th order kernel is used, the variance of* $\left(\hat{q}_i^X - q_i^X\right)$ *is of order* (nh^d) *and the bias is of order* h^{2q}. *At the optimal bandwidth* $h \sim n^{-1/(2q+d)}$, *the rate of convergence of* $\hat{q}_i^X$ *is* $n^{-q/(2q+d)}$. *To ensure* $n^{1/4}\left(\left(\hat{q}_i^X - q_i^X\right)\right) = o_p(1)$, *we need*

$$\frac{1}{4} - \frac{q}{2q+d} < 0.$$

When we use a second order kernel, i.e. $q = 2$, *we have*

$$\frac{1}{4} - \frac{2}{4+d} < 0.$$

which requires $d \leq 3$. *This is the condition invoked by Robinson (1988). When* $d \geq 4$, *a kernel with order greater than 2 may have to be used. See Li (1996) or Li and Racine's book (page 227) for weaker requirement on the order of the kernel employed.*

Remark 5.3.2 *To ensure the stochastic equicontinuity, we typically have to trim out some observations. This is to say, we define*

$$\hat{\beta} = \left[\sum_{i=1}^{n}(X_i - \hat{q}_i^X)'(X_i - \hat{q}_i^X)\right]^{-1} \sum_{i=1}^{n}(X_i - \hat{q}_i^X)'(Y_i - \hat{q}_i^Y)1_i$$

where $1_i = 1\{\hat{f}(V_i) \geq b_n\}$ *is a trimming function and* $b_n \to 0$ *at certain rate. There are two reasons for trimming. First, trimming can eliminate observation for the computation of* $\hat{\beta}$ *for which the nuisance parameter estimator* $\hat{q}_i$ *is estimated with relatively large error in comparison to the non-trimmed observations. Second, trimming makes it easy to establish the stochastic equicontinuity. One can obtain the stochastic equicontinuity over a bound set but not over unbounded set in general. We have not explicitly dealt with trimming here. See Robinson (1988), Andrews(1994), Li and Racine's textbook for detailed theoretical development. Some authors argue that trimming can be ignored in practical implementations. Others disagree. For practical guidances, see Ichimura and Todd's (2007) Handbook of Econometrics chapter.*

Remark 5.3.3 *For discussions on the estimation of the asymptotic variance and the non-parametric component, see Li and Racine (2007, page 228).*

5.4 References

1. Andrews, D.W.K. (1994) Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity. *Econometrica* 62, 43-72.

2. Härdle, Liang and Gao (2000): *Partially linear models.* Springer.

3. Engle, Granger, Rice and Weiss (1986): "Semi-parametric estimates of the relation between weather and electricity demand," *Journal of American Statistical Association* 81 (1986): 310-320.

4. Ichimura, H. and P.E. Todd (2007) Implementing Nonparametric and Semiparametric Estimators. Forthcoming in *Handbook of Econometrics* vol. 6. North-Holland.

5. Li, Q. and J.S. Racine (2007): *Nonparametric Econometrics: Theory and Practice.* Princeton University Press

6. Pagan, A. and A. Ullah (1999): *Nonparametric Econometrics.* Cambridge University Press.

7. Robinson, P.M. (1988) Root-N-Consistent Semiparametric Regression. *Econometrica* 56, 931-954.

Chapter 6

Semiparametric Methods: Two-step Estimation

In a previous chapter, we presented a number of examples of semiparametric models, and designed two-step estimators of the finite dimensional parameters of interest. In this chapter, we set up a framework which allows us to derive the asymptotic properties of these two-step estimators.

6.1 The Framework

Consider an econometric model that specifies a set of conditions on a vector of population moments

$$Q(\beta_0) = 0$$

where

$$Q(\beta) = \int m(z, h_0(v_1, \tau_0(v_2), \beta), \beta) dP(z)$$

and v_1 and v_2 are subvectors of z. For notational simplicity, we often write $h_0(v_1, \tau_0(v_2), \beta) = h_0(v, \tau_0(v), \beta)$ for $v = (v_1', v_2')'$. The form of moment conditions is motivated from an IO application, see Pakes and Olley (1995).

Generally, both $h(\cdot)$ and $\tau(\cdot)$ are unknown functions. Estimators of β are obtained by drawing a random sample of size n from the distribution $P(\cdot)$, forming nonparametric estimates of h_0 and τ_0, say h_n and τ_n, and finding that value of β that makes a norm of the sample moment

$$Q_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, h_n(V_{1i}, \tau_n(V_{2i}), \beta), \beta)$$

as close to zero as possible.

All the estimators have the common feature that a preliminary estimator of the nonparametric component is used to estimate the parametric one. So the set-up has much in common with the two-step estimators for parametric models, but now the preliminary estimator is infinite-dimensional. The strategy of proof can however be adapted to this more general setting.

As in the parametric case we will require uniform convergence, stochastic equicontinuity, and asymptotic normality conditions. Most of these conditions can be found in an advance econometrics textbook, e.g., Wooldridge (2002).

We begin by assuming that all functions are sufficiently smooth in all of their arguments, and that the data are an i.i.d. sample from some population. A key assumption is that the nonparametruc functions can be estimated with faster enough rates of convergence in the sup norm. The assumption can be relaxed but it allows us to prove limit theorems under the conditions that are transparent to those who are not familiar with function spaces. For this case, the results are based on Pakes and Olley (1995). We then go back to show how to amend the proofs for cases where the objective function may be discontinuous.

6.2 Limit Theorem

6.2.1 Preliminaries

The population moment conditions are

$$Q(\beta_0) = 0$$

where

$$Q(\beta) = Em(Z, h_0\left(V_1, \tau_0(V_2), \beta\right), \beta) \in \mathbb{R}^{d_m}$$

and

- $V_1 \in \mathbb{R}^{d_{V_1}}, V_2 \in \mathbb{R}^{d_{V_2}}, V = (V_1', V_2')' \in \mathbb{R}^{d_v}$ is subvector of $Z \in \mathbb{R}^d$. $d_V = d_{V_1} + d_{V_2} \leq d$.
- $\beta \in \mathcal{B}$, a compact subset of $\mathbb{R}^k$.
- $\tau_0 \in \Gamma$ and $h_0\left(\cdot\right) \in \mathcal{H}$ where Γ and $\mathcal{H}$ are pseudo metric space of functions:

$$\begin{aligned} \Gamma &= \left\{\tau\left(\cdot\right) : \mathcal{V}_2 \to \mathbb{R}^{d_\tau}\right\} \\ \mathcal{H} &= \left\{h\left(\cdot\right) : \mathcal{V}_1 \times \Gamma \times \mathcal{B} \to \mathbb{R}^{d_h}\right\} \end{aligned}$$

Thus

$$m\left(\cdot\right) : \mathcal{Z} \times \mathcal{H} \times \mathcal{B} \to \mathbb{R}^{d_m}.$$

Our problem is to estimate β, and we would use a method of moments estimator if h_0 and τ_0 were known. Because these two functions are not known, we plug preliminary estimators of them, say $\tau_n(\cdot)$, and $h_n(\cdot, \tau_n(\cdot), \beta)$ into $Q(\cdot)$ and consider minimizing a norm of

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, h_n(V_{1i}, \tau_n(V_{2i}), \beta), \beta).$$

Starting with the Euclidean norm, β_n, our estimator of β, will be defined as

$$\beta_n = \arg\min_{\beta \in \mathcal{B}} \|Q_n(\beta)\| + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{6.1}$$

In order to find β_n, we will have to compute $h_n(\cdot)$ for each different value of β.

The goal is to provide a limit distribution for β_n which satisfies (6.1). Assuming that $m(\cdot)$ and $h(\cdot)$ are sufficiently smooth, we can follow the steps below to establish the asymptotic properties of β_n.

- Take a second order expansion to the moment conditions about the true values of τ_0 and h_0 at each data point. These expansions are done pointwise; i.e., separately at every value $\tau(V_{2i})$ and $h(V_{1i}, \tau(V_{2i}), \beta)$. This is possible only if (i) $h(V_i, \cdot)$ depends on τ only through $\tau(V_{2i})$ and (*ii*) $m(Z_i, \cdot)$ depends on h only through $h(V_{1i}, \tau(V_{2i}), \beta)$. This may be restrictive but it generally holds in economic applications.

- Produce a new objective function from the expansion by dropping the second and higher order terms in it.

- Impose conditions which ensure that τ_n and h_n are very "close" to τ_0 and h_0 (uniformly over the arguments of these functions) so that the approximation errors do not matter asymptotically.

- Show that the argmin of this "approximate" objective function has the same asymptotic properties of the argmin of $\|Q_n(\beta)\|$.

The rest of the argument is entirely analogous to the fully parametric case. We derive the limit distribution for the estimator which minimizes the new "simpler" objective function. However, we cannot actually compute this estimator and its asymptotic covariance matrix because the estimator and its covariance matrix depend on objects that we do not know. We can, of course, calculate the estimator given in (6.1), but we cannot provide a closed form expression for its value, and hence find it difficult to analyze its limit distribution directly. Since the two estimators are very "close" to each other, the limit distribution of the estimator given in (6.1) is the same as that of the new estimator: the estimator which we cannot calculate, but whose limit distribution is easy to find.

Our proof stategy here is the same as what we use in the fully parametric case. If we view $\delta_{\tau i} = \tau_0(V_{2i})$ and $\delta_{hi}(\beta) = h_0(V_{1i}, \tau_0(V_{2i}), \beta)$ as $2n$ parameters to be estimated, then we can proceed as if we are in the parametric two-step world. The only difference is that the number of the first step estimators (i.e. $2n$) grows with the sample size. We have to control the overall estimation errors in these $2n$ parameters. That is why we assume that the functions τ_0 and h_0 can be estimated uniformly well over their respective domains. The uniform conditions are stronger than needed but they make our proof close to the parametric case.

We begin by assuming that: $m(\cdot)$ is twice continuously differentiable in $h = h(\cdot)$ $(\forall h \in \mathcal{H})$ and once continuously differentiable in β; and that $h(\cdot)$ is twice continuously differentiable in $\tau = \tau(\cdot)$ $(\forall \tau \in \Gamma)$. Further these derivatives are assumed to be continuously differentiable in β in a region of β_0 and bounded by functions (envelopes) which are square integrable. Also for expositional convenience, we consider the case where both $\tau(\cdot)$ and $h(\cdot)$ map into a subset of $\mathbb{R}$.

6.2.2 Smoothness Assumptions

Assumption R: Regularity Conditions.

(i) For each $(\beta, h(\cdot), \tau(\cdot)) \in (\mathcal{B} \times \mathcal{H} \times \Gamma)$, let

$$m(z, h, \tau, \beta) := m[z, h(v_1, \tau(v_2), \beta), \beta]$$

and assume that

$$\begin{aligned} \dot{m}(z, h, \tau, \beta) &= \left.\frac{\partial m[z, h, \beta]}{\partial h}\right|_{h=h(v_1,\tau(v_2),\beta)} \\ \ddot{m}(z, h, \tau, \beta) &= \left.\frac{\partial^2 m[z, h, \beta]}{\partial h^2}\right|_{h=h(v_1,\tau(v_2),\beta)} \end{aligned}$$

exist. Also assume that there is an envelop function $M(z)$ such that

$$|\dot{m}(z, h, \tau, \beta)| \leq M(z), \;\; |\ddot{m}(z, h, \tau, \beta)| \leq M(z)$$

and

$$\int M(z)^2 dP(z) \leq \kappa \text{ for some } \kappa < \infty.$$

(ii) Let $h(v, \tau, \beta) = h(v_1, \tau(v_2), \beta)$ and assume that

$$\begin{aligned} \dot{h}(v, \tau, \beta) &= \left.\frac{\partial h(v_1, \tau, \beta)}{\partial \tau}\right|_{\tau=\tau(v_2)} \\ \ddot{h}(v, \tau, \beta) &= \left.\frac{\partial^2 h(v_1, \tau, \beta)}{\partial \tau^2}\right|_{\tau=\tau(v_2)} \end{aligned}$$

exist. Also assume that there is an envelope function $H(z)$ such that

$$\left|\dot{h}(v,\tau,\beta)\right| \leq H(z),\ \left|\ddot{h}(v,\tau,\beta)\right| \leq H(z)$$

and

$$\int M(z)H(z)^2 dP(z) \leq \kappa,\ \int H(z)^2 dP(z) \leq \kappa \text{ for some } \kappa < \infty.$$

(iii) $m(z,h_0,\tau_0,\beta), \dot{m}(z,h_0,\tau_0,\beta), h(v,\tau_0,\beta)$ and $\dot{h}(v,\tau_0,\beta)$ are all continuously differentiable in β (a.e. P). Moreover, for all β in some neighborhood of β_0, $m(z,h_0,\tau_0,\beta)$, $\partial\dot{m}(z,h_0,\tau_0,\beta)/\partial\beta$ and $\partial\dot{h}(v,\tau_0,\beta)/\partial\beta$ exists and are bounded by the square integrable functions $M(z)$ and $H(z)$ given above:

$$\begin{aligned} |m(z,h_0,\tau_0,\beta)| &\leq M(z) \\ \left|\frac{\partial\dot{m}(z,h_0,\tau_0,\beta)}{\partial\beta}\right| &\leq M(z) \\ \left|\frac{\partial\dot{h}(v,\tau_0,\beta)}{\partial\beta}\right| &\leq H(z). \end{aligned}$$

Furthermore

$$\left\|\frac{\partial\dot{h}(v,\tau_0,\beta)}{\partial\beta} - \frac{\partial\dot{h}(v,\tau_0,\beta_0)}{\partial\beta}\right\| \leq \|H(z)\|\,\|\beta-\beta_0\|.$$

(iv) As $n\to\infty$, $h_n(v_1,\tau_n(v_2),\beta)$ will be contained in $\mathcal{H}$ with probability approaching one, and β_0 is in the interior of $\mathcal{B}$.

The last condition in (iii) is needed for a brute force proof of the stochastic equicontinuity condition. We assume stochastic equicontinuity directly so that the last condition in (iii) is not explicitly used in our proof.

In the above notation, we use a dot to denote a derivative with respect to the nonparametric component.

Example 6.2.1 *Partially linear model:*

$$Y = X\beta_0 + f_0(V) + u \tag{6.2}$$

where $E(u|X,V) = 0$ *and* $var(u|X,V) = \sigma^2$. *Let* $Z = (X,V,Y)$ *and with some abuse of notation*

$$h(Z) := h(V) = (h_1(V), h_2(V))$$

and

$$h_0(V) = (h_{01}(V), h_{02}(V)) = [E(X|V), E(Y|V)].$$

The moment condition is:

$$Em(Z, \beta_0, h_0) = 0$$

where

$$m(Z, \beta, h) = (X - h_1(V))[Y - h_2(V) - (X - h_1(V))\beta].$$

There is no τ function in m but h is a bivariate nonparametric function. We have

$$\begin{aligned}
\dot{m}_1(Z, h, \tau, \beta) &= \left.\frac{\partial m[Z, h, \beta]}{\partial h_1}\right|_{h=h(V)} = -[Y - h_2(V) - (X - h_1(V))\beta] + (X - h_1(V))\beta \\
\dot{m}_2(Z, h, \tau, \beta) &= \left.\frac{\partial m[Z, h, \beta]}{\partial h_2}\right|_{h=h(V)} = -[X - h_1(V)] \\
\ddot{m}_{12}(Z, h, \tau, \beta) &= \left.\frac{\partial m[Z, h, \beta]}{\partial h_1}\right|_{h=h(V)} = 1 \text{ and } \ddot{m}_{11}(Z, h, \tau, \beta) = \ddot{m}_{22}(Z, h, \tau, \beta) = 0.
\end{aligned}$$

6.2.3 Stochastic Approximations

Under Assumption R, we can use the mean value theorem to approximate $m[Z_i, h_n(V_i, \tau_{ni}, \beta), \beta] - m[Z_i, h_0(v_i, \tau_{0i}, \beta), \beta]$ for each fixed β. To simplify the notation, we suppress the dependence of h_n, h_0 and m on β for now and we write

$$\begin{aligned}
m_i[h_{ni}(\tau_{ni})] &= m[Z_i, h_n(V_{1i}, \tau_n(V_{2i}), \beta), \beta] \\
m_i[h_{0i}(\tau_{0i})] &= m[Z_i, h_0(V_{1i}, \tau_0(V_{2i}), \beta), \beta]
\end{aligned}$$

where $\tau_{ni} = \tau_n(V_{2i})$, $\tau_{0i} = \tau_0(V_{2i})$,

$$h_{ni}(\tau_{ni}) = h_n(V_{1i}, \tau_n(V_{2i}), \beta) \text{ and } h_{0i}(\tau_{0i}) = h_0(V_{1i}, \tau_0(V_{2i}), \beta).$$

To simplify the notation further, we suppress the subscript i for now. We have

$$\begin{aligned}
& m[h_n(\tau_n)] - m[h_0(\tau_0)] \\
=\ & \dot{m}(h_0(\tau_0))[h_n(\tau_n) - h_0(\tau_0)] + \frac{1}{2}\ddot{m}(\tilde{h}_n)[h_n(\tau_n) - h_0(\tau_0)]^2 \\
=\ & \dot{m}(h_0(\tau_0))[h_n(\tau_0) - h_0(\tau_0)] + \dot{m}[h_0(\tau_0)][h_n(\tau_n) - h_n(\tau_0)] \\
& + \frac{1}{2}\ddot{m}(\tilde{h}_n)[h_n(\tau_n) - h_0(\tau_0)]^2
\end{aligned} \tag{6.3}$$

where $\tilde{h}_n$ is between $h_0(\tau_0)$ and $h_n(\tau_n)$. Moreover

$$\begin{aligned}
& h_n(\tau_n) - h_n(\tau_0) \\
=\ & \dot{h}_n(\tau_0)[\tau_n - \tau_0] + \frac{1}{2}\ddot{h}_n(\tilde{\tau}_n)[\tau_n - \tau_0]^2 \\
=\ & \dot{h}_0(\tau_0)[\tau_n - \tau_0] + \left[\dot{h}_n(\tau_0) - \dot{h}_0(\tau_0)\right][\tau_n - \tau_0] + \frac{1}{2}\ddot{h}_n(\tilde{\tau}_n)[\tau_n - \tau_0]^2
\end{aligned} \tag{6.4}$$

where $\tilde{\tau}_{ni}$ is between $\tau_n(v_i)$ and $\tau_0(v_i)$.

Now plugging (6.4) into (6.3) and putting everything except the first two terms in these expansions into the remainder, we have

$$\begin{aligned} m[h_n(\tau_n)] &= m[h_0(\tau_0)] + \dot{m}\left(h_0(\tau_0)\right)\left[h_n(\tau_0) - h_0(\tau_0)\right] \\ &\quad + \dot{m}\left(h_0(\tau_0)\right)\dot{h}_0(\tau_0)\left[\tau_n - \tau_0\right] + e_n(\beta) \end{aligned}$$

where

$$\begin{aligned} e_n\left(\beta\right) &= \dot{m}\left(h_0(\tau_0)\right)\left[\dot{h}_n(\tau_0) - \dot{h}_0(\tau_0)\right]\left(\tau_n - \tau_0\right) \\ &\quad + \frac{1}{2}\dot{m}\left(h_0(\tau_0)\right)\ddot{h}_n(\tilde{\tau}_n)\left(\tau_n - \tau_0\right)^2 + \frac{1}{2}\ddot{m}\left(\tilde{h}_n\right)\left[h_n(\tau_n) - h_0(\tau_0)\right]^2. \end{aligned}$$

Ignoring $e_n\left(\beta\right)$, we define

$$\begin{aligned} & m_i^*(\beta) = m[Z_i, h_{0i}(V_i, \tau_{0i}, \beta), \beta] \\ & + \dot{m}[Z_i, h_0(V_i, \tau_{0i}, \beta), \beta]\left[h_n(V_i, \tau_{0i}, \beta) - h_0(V_i, \tau_{0i}, \beta)\right] \\ & + \dot{m}[Z_i, h_0(V_i, \tau_{0i}, \beta), \beta]\left[\dot{h}_0(V_i, \tau_{0i}, \beta)\tau_{ni} - \tau_{0i})\right] \end{aligned}$$

and consider the problem

$$\beta_n^* = \arg\min_{\beta \in \mathcal{B}} \left\| Q_n^*\left(\beta\right)\right\| := \arg\min_{\beta \in \mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} m^*(\beta)\right\|. \tag{6.5}$$

or equivalently from an asymptotic point of view:

$$\left\| Q_n^*\left(\beta_n^*\right)\right\| = \min_{\beta \in \mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} m^*(\beta)\right\| + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{6.6}$$

Although $m_i^*(\beta)$ has a very simple form, we cannot actually calculate it, and as a result we cannot actually calculate $Q_n^*\left(\beta\right)$. However, we can derive the limit distribution of this estimator. We will show that the limit distribution of $\arg\min \left\| Q_n^*\left(\beta\right)\right\|$ is the same as that of $\arg\min \left\| Q_n\left(\beta\right)\right\|$.

Below we provide conditions which ensure that

$$\sup_{\beta \in \mathcal{B}} \left\| Q_n\left(\beta\right) - Q_n^*\left(\beta\right)\right\| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

This will imply that our estimator, which is defined on the objective function $Q_n\left(\beta\right)$, will also minimize $Q_n^*\left(\beta\right)$ up to small order term.

We now introduce a set of three assumptions on the rates of convergence of the estimators of h_0 and τ_0 (i to iii), which together ensure that $\sup_{\beta\in\mathcal{B}} \|Q_n(\beta) - Q_n^*(\beta)\| = o_p(1/\sqrt{n})$. Primitive conditions for the validity of these thee assumptions have been considered in the first part of the course. See also the sections on uniform rates of convergence in Li and Racine (2007, sec. 1.10 and 2.3).

Assumption r: rate of convergence.
(i) (rate for h)

$$n^{\alpha_1} \sup_{v\in\mathcal{V},\beta\in\mathcal{B}} \|h_n(v,\tau_0(v),\beta) - h_0(v,\tau_0(v),\beta)\| = O_p(1),$$

(ii) (rate for τ)

$$n^{\alpha_2} \sup_{v\in\mathcal{V}} \|\tau_n(v) - \tau_0(v)\| = O_p(1),$$

(iii) (rate for $\dot{h}$)

$$n^{\alpha_3} \sup_{v\in\mathcal{V},\beta\in\mathcal{B}} \left\|\dot{h}_n(v,\tau_0(v),\beta) - \dot{h}_0(v,\tau_0(v),\beta)\right\| = O_p(1),$$

with

$$\alpha_1 > 1/4,\ \alpha_2 > 1/4,\ \text{and } \alpha_2 + \alpha_3 > 1/2.$$

(v) (condition on $\partial\dot{h}_n(v,\tau_0(v),\beta)/\partial\beta$)

$$\sup_{v\in\mathcal{V}} \left\| \frac{\partial\dot{h}_n(v,\tau_0(v),\beta_0)}{\partial\beta} - \frac{\partial\dot{h}_0(v,\tau_0(v),\beta_0)}{\partial\beta} \right\| = o_p(1).$$

Remarks: $\alpha_1 > 1/4$ will be needed to ensure that the second order term in the expansion about $h_0(\cdot)$ is $o_p(1)$. $\alpha_2 > 1/4$ will be needed to ensure that the second order term in the expansion about $\tau_0(\cdot)$ is $o_p(1)$. The condition $\alpha_2 + \alpha_3 > 1/2$ ensures that the second order interaction between $h_0(\cdot)$ and $\tau_0(\cdot)$ is $o_p(1)$. We require that all the $o_p(1)$ terms hold uniformly over their arguments.

Assumption (iv) is introduced here for convenience, as it follows from (iii) in the most applications and helps simplify the asymptotic normality proof below. In most applications, we can write h_0 as an unknown function of a known (possibly vector valued) function of v, τ, and β, say $x(v,\tau,\beta)$. That is, $h = h(x(v,\tau,\beta))$. Then (iii) will imply (iv) if $\partial x(v,\tau_0,\beta_0)/\partial\beta \neq 0$ whenever $\partial x(v,\tau_0,\beta_0)/\partial\tau \neq 0$. Assumption (iv) is one of the primitive sufficient conditions for (6.10).

Only Assumptions (i) to (iii) are needed for the following lemma. The lemma allows us to analyze the simpler problem $\arg\min \|Q_n^*(\beta)\|$.

Lemma 6.2.1 *Let Assumptions r(i)-(iii) hold, then*

$$\sup_{\beta\in\mathcal{B}}\|Q_n(\beta)-Q_n^*(\beta)\|=o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof: In view of $Q_n(\beta)-Q_n^*(\beta)=n^{-1}\sum_{i=1}^n e_{ni}(\beta)$, we have

$$\begin{aligned}
&\|Q_n(\beta)-Q_n^*(\beta)\| \\
\leq\ &\left\|n^{-1}\sum_{i=1}^n \dot{m}[h_{0i}(\tau_{i0}))]\left[\dot{h}_{ni}(\tau_{0i})-\dot{h}_{0i}(\tau_{0i})\right](\tau_{ni}-\tau_{0i})\right\| \\
&+\left\|(2n)^{-1}\sum_{i=1}^n \frac{1}{2}\dot{m}\left(h_{0i}(\tau_{0i})\right)\ddot{h}_{ni}(\tilde{\tau}_{ni})\left(\tau_{ni}-\tau_{0i}\right)^2\right\| \\
&+\left\|(2n)^{-1}\sum_{i=1}^n \frac{1}{2}\ddot{m}\left(\tilde{h}_{ni}\right)\left[h_{ni}(\tau_{ni})-h_{0i}(\tau_{0i})\right]^2\right\|.
\end{aligned}$$

We consider each of the three terms in turn. For the first term, we use assumptions (ii) and (iii) to obtain:

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left\|n^{-1}\sum_{i=1}^n \dot{m}[h_{0i}(\tau_{i0}))]\left[\dot{h}_{ni}(\tau_{0i})-\dot{h}_{0i}(\tau_{0i})\right](\tau_{ni}-\tau_{0i})\right\| \\
\leq\ &\sup_{\beta\in\mathcal{B}} n^{-1}\sum_{i=1}^n |M(Z_i)|\left[\dot{h}_{ni}(\tau_{0i})-\dot{h}_{0i}(\tau_{0i})\right](\tau_{ni}-\tau_{0i}) \\
\leq\ &n^{-1}\sum_{i=1}^n |M(Z_i)|\,O_p\left(n^{-(\alpha_3+\alpha_2)}\right)=o_p\left(1/\sqrt{n}\right)
\end{aligned}$$

where we have used the law of large numbers for i.i.d. random variables. Similarly, the second term is bounded by

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left\|(2n)^{-1}\sum_{i=1}^n \frac{1}{2}\dot{m}\left(h_{0i}(\tau_{0i})\right)\ddot{h}_{ni}(\tilde{\tau}_{ni})\left(\tau_{ni}-\tau_{0i}\right)^2\right\| \\
\leq\ &(2n)^{-1}\sum_{i=1}^n |M(Z_i)H(Z_i)|\left(\tau_{ni}-\tau_{0i}\right)^2 \\
=\ &O_p(1)\,O_p\left(n^{-2\alpha_2}\right)=o_p\left(1/\sqrt{n}\right)
\end{aligned}$$

where the last equality follows because $\alpha_2>1/4$. Finally, we handle the third term, starting

with

$$\begin{aligned} & [h_{ni}(\tau_{ni}) - h_{0i}(\tau_{0i})]^2 \\ = \; & [h_{ni}(\tau_{ni}) - h_{ni}(\tau_{0i}) + h_{ni}(\tau_{0i}) - h_{0i}(\tau_{0i})]^2 \\ \leq \; & 2\left[h_{ni}(\tau_{ni}) - h_{ni}(\tau_{0i})\right]^2 + 2\left[h_{ni}(\tau_{0i}) - h_{0i}(\tau_{0i})\right]^2. \end{aligned}$$

But

$$\begin{aligned} & h_{ni}(\tau_{ni}) - h_{ni}(\tau_{0i}) \\ = \; & \dot{h}_n(\tau_{0i})(\tau_{ni} - \tau_{0i}) + \frac{1}{2}\ddot{h}_n(\tilde{\tau}_{ni})(\tau_{ni} - \tau_{0i})^2 \\ = \; & \left[\dot{h}_n(\tau_{0i}) - \dot{h}_0(\tau_{0i})\right](\tau_{ni} - \tau_{0i}) + \dot{h}_0(\tau_{0i})(\tau_{ni} - \tau_{0i}) + \frac{1}{2}\ddot{h}_n(\tilde{\tau}_{ni})(\tau_{ni} - \tau_{0i})^2 \end{aligned} \tag{6.7}$$

so

$$\sup_{\beta \in \mathcal{B}} [h_{ni}(\tau_{ni}) - h_{ni}(\tau_{0i})]^2 \leq H^2(Z_i) o_p\left(n^{-2\alpha_2}\right).$$

Combining this with assumption (i), we get

$$\begin{aligned} & \left\| (2n)^{-1} \sum_{i=1}^{n} \frac{1}{2}\ddot{m}\left(\tilde{h}_{ni}\right)[h_{ni}(\tau_{ni}) - h_{0i}(\tau_{0i})]^2 \right\|. \\ \leq \; & (2n)^{-1} \sum_{i=1}^{n} |M(Z_i)| H^2(Z_i) O_p\left(n^{-2\alpha_2}\right) + (2n)^{-1} \sum_{i=1}^{n} |M(Z_i)| O_p\left(n^{-2\alpha_1}\right) = o_p\left(1/\sqrt{n}\right) \end{aligned}$$

where we have used: $\alpha_1 > 1/4$ and $\alpha_2 > 1/4$. ■

Recall that β_n minimizes the objective function $\|Q_n(\beta)\|$, and β_n^* minimizes $\|Q_n^*(\beta)\|$. Then from above

$$\begin{aligned} \|Q_n^*(\beta_n)\| \; &= \; \|Q_n(\beta_n)\| + o_p\left(\frac{1}{\sqrt{n}}\right) \text{ (by the Lemma)} \\ &\leq \; \|Q_n(\beta_n^*)\| + o_p\left(\frac{1}{\sqrt{n}}\right) \text{ (by the definition of } \beta_n) \\ &\leq \; \|Q_n^*(\beta_n^*)\| + o_p\left(\frac{1}{\sqrt{n}}\right) \text{ (by the Lemma)} \\ &= \; \min_{\beta \in \mathcal{B}} \|Q_n^*(\beta)\| + o_p\left(\frac{1}{\sqrt{n}}\right) \text{ (by the definition of } \beta_n^*). \end{aligned}$$

But we also have $\|Q_n^*(\beta_n)\| \geq \min_{\beta \in \mathcal{B}} \|Q_n^*(\beta)\|$. Hence

$$\|Q_n^*(\beta_n)\| = \min_{\beta \in \mathcal{B}} \|Q_n^*(\beta)\| + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{6.8}$$

We now use (6.8) to prove consistency, $\sqrt{n}$-consistency, and asymptotic normality of β_n. For the consistency proof, Assumption 1 could be replaced by weaker conditions that the approximation errors are of order $o_p(1)$.

6.2.4 Consistency

Recall that β_n is defined according to

$$\beta_n = \arg\min_{\beta\in\mathcal{B}} \|Q_n(\beta)\| + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{6.9}$$

By Lemma 6.2.1, β_n also satisfies

$$\beta_n = \arg\min_{\beta\in\mathcal{B}} \|Q_n^*(\beta)\| + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Assumption I: identification. For $\forall\delta > 0$, $\exists\varepsilon(\delta) > 0$ such that

$$\inf_{\|\beta-\beta_0\|>\delta} \|Q(\beta)\| \geq \varepsilon(\delta) > 0$$

The identification condition is the same as in the parametric case.

Theorem 6.2.1 *Let Assumptions R, r and I hold. Then* $\beta_n - \beta_0 = o_p(1)$.

Proof: In view of (6.8), we only have to show that

$$\sup_{\beta\in\mathcal{B}} \|Q_n^*(\beta) - Q(\beta)\| = o_p(1).$$

Note that

$$\begin{aligned}
& \sup_{\beta\in\mathcal{B}} \|Q_n^*(\beta) - Q(\beta)\| \\
\leq & \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n m[Z_i, h_0(V_i,\tau_0,\beta),\beta] - Em[Z_i, h_0(V_i,\tau_0,\beta),\beta] \right\| \\
& + \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n \dot{m}[Z_i, h_0(V_i,\tau_0,\beta),\beta]\left[h_n(V_i,\tau_0,\beta) - h_0(V_i,\tau_0,\beta)\right] \right\| \\
& + \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n \dot{m}[Z_i, h_0(V_i,\tau_0,\beta),\beta]\left[\dot{h}(V_i,\tau_0,\beta)\left(\tau_n(V_i) - \tau_0(V_i)\right)\right] \right\| \\
\leq & \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n m[Z_i, h_0(V_i,\tau_0,\beta),\beta] - Em[Z_i, h_0(V_i,\tau_0,\beta),\beta] \right\| \\
& + \frac{1}{n}\sum_{i=1}^n |M(Z_i)| O_p\left(n^{-\alpha_1}\right) + \frac{1}{n}\sum_{i=1}^n |M(Z_i) H(Z_i)| O_p\left(n^{-\alpha_2}\right) \\
= & \sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n m[Z_i, h_0(V_i,\tau_0,\beta),\beta] - Em[Z_i, h_0(V_i,\tau_0,\beta),\beta] \right\| + o_p(1).
\end{aligned}$$

Recall that $m[Z_i, h_0(V_i,\tau_0,\beta),\beta]$ does not depend on any infinite dimensional parameters (the unknown functions are set at their true values). So we are back to working with uniform convergence conditions that are uniform over only a finite set of parameters, the same conditions we need for the parametric case. In particular,

$$\sup_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^n m[Z_i, h_0(V_i,\tau_0,\beta),\beta] - Em[Z_i, h_0(V_i,\tau_0,\beta),\beta] \right\| = o_p(1)$$

follows from the differentiability of $m[Z_i, h_0(V_i,\tau_0,\beta),\beta]$ with respect to β, the compactness of $\mathcal{B}$ and the square integrability of the envelope function for $m[Z_i, h_0(V_i,\tau_0,\beta),\beta]$. ■

6.2.5 Root-n Consistency

Assume that $Q(\beta)$ is continuously differentiable. Denote

$$D(\beta) = \partial Q(\beta)/\partial\beta'.$$

Theorem 6.2.2 *Let Assumptions R, r and I hold. In addition, assume*
(i) For any sequence β_n such that $\beta_n - \beta_0 = o_p(1)$

$$\left\|\sqrt{n}\left[Q_n^*(\beta_n) - Q(\beta_n)\right] - \sqrt{n}\left[Q_n^*(\beta_0) - Q(\beta_0)\right]\right\| = o_p(1)\left[1 + \left\|\sqrt{n}(\beta_n - \beta_0)\right\|\right], \quad (6.10)$$

(ii) $Q_n^*(\beta_0) = O_p(1/\sqrt{n})$,
(iii) $D(\beta)$ *is continuous at* $\beta = \beta_0$ *and* $D(\beta_0)$ *is of full rank* k.
Then

$$\beta_n - \beta_0 = O_p\left(1/\sqrt{n}\right).$$

Remark 6.2.1 *Assumption (i) is like the "equicontinuity" condition, i.e. it states that, after we multiply by* $\sqrt{n}$*, the distribution of the objective functions at* $\beta = \beta_0$ *will be asymptotically the same as the distribution of the objective function at* $\beta = \beta_n$ *if* $\beta_n - \beta_0 = o_p(1)$. *As stated, Assumption (i) is slightly weaker than the equicontinuity assumption, which says*

$$\left\|\sqrt{n}\left[Q_n^*(\beta_n) - Q(\beta_n)\right] - \sqrt{n}\left[Q_n^*(\beta_0) - Q(\beta_0)\right]\right\| = o_p(1)$$

for any β_n *such that* $\beta_n - \beta_0 = o_p(1)$. *It is however all we need for smooth problems, and has the advantage that, given our smoothness conditions, it can be proved by "brute force"; i.e. without appealing to more general theorems on the stochastic equicontinuity of empirical processes defined on metric spaces satisfying specific assumptions. For more details, see the proof in Pakes and Olley (1995).*

Remark 6.2.2 *Assumption (ii) is more complicated to verify than the analogous assumption for the finite dimensional case, since now* $\sqrt{n}Q_n^*(\beta_0)$ *involves objects like*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}[Z_i, h_0(V_i,\tau_0,\beta),\beta]\left[h_n(V_i,\tau_0,\beta) - h_0(V_i,\tau_0,\beta)\right]$$

which are not transparently $\sqrt{n}$ *times the mean of i.i.d. random variables. As a result we will have to come back to this assumption and provide primitive sufficient conditions.*

Remark 6.2.3 *The derivative matrix* $D(\cdot)$ *is the derivative of the limit function evaluated at* $h = h_0$ *and* $\tau = \tau_0$. *The estimated functions do not enter it at all. Note that if* $h_0(\cdot)$ *is a function of* β*, then* β *enters the definition of* $Q(\cdot)$ *in two places. As a result,* $D(\cdot)$ *will consist of the sum of two derivatives, one of which is obtained from the chain rule.*

Proof: The differentiability of $Q(\beta)$ at $\beta = \beta_0$ and Assumption (iii) imply that

$$\begin{aligned}\sqrt{n}\left\|Q(\beta_n) - Q(\beta_0)\right\| &= \left\|D\left(\tilde{\beta}_n\right)\sqrt{n}(\beta_n - \beta)\right\| \\ &= \left\|\left[D(\beta_0) + o_p(1)\right]\sqrt{n}(\beta_n - \beta)\right\|\end{aligned}$$

where $\tilde{\beta}_n$ is between β_n and β_0. So it suffices to show that $\sqrt{n}\left\|Q\left(\beta_n\right)-Q(\beta_0)\right\|=O_p\left(1\right)$. Note that

$$\begin{aligned}
& \sqrt{n}\left\|Q\left(\beta_n\right)-Q(\beta_0)\right\| \\
\leq\ & \sqrt{n}\left\|Q\left(\beta_n\right)-Q(\beta_0)-\left[Q_n^*\left(\beta_n\right)-Q_n^*\left(\beta_0\right)\right]\right\| \\
& +\sqrt{n}\left\|Q_n^*\left(\beta_n\right)-Q_n^*\left(\beta_0\right)\right\| \text{ [triangle Inequality]} \\
\leq\ & o_p\left(1\right)\left[1+\sqrt{n}\left\|\beta_n-\beta_0\right\|\right]+\sqrt{n}\left\|Q_n^*\left(\beta_n\right)\right\| \\
& +\sqrt{n}\left\|Q_n^*\left(\beta_0\right)\right\| \text{ [triangle Inequality]} \\
\leq\ & o_p\left(1\right)\left[1+\sqrt{n}\left\|\beta_n-\beta_0\right\|\right]+2\sqrt{n}\left\|Q_n^*\left(\beta_0\right)\right\|+o_p(1) \\
=\ & o_p\left(1\right)\left[1+\sqrt{n}\left\|\beta_n-\beta_0\right\|\right]+O_p(1)
\end{aligned}$$

where the last inequality follows from the fact that β_n minimizes the objective function $Q_n^*\left(\beta\right)$ (up to a term of order $o_p\left(1/\sqrt{n}\right)$). As a result,

$$\left\|\left[D\left(\beta_0\right)+o_p\left(1\right)\right]\sqrt{n}(\beta_n-\beta)\right\|\leq o_p\left(1\right)\left[1+\left\|\sqrt{n}\left(\beta_n-\beta_0\right)\right\|\right]+O_p(1),$$

from which we deduce that

$$\left\|\sqrt{n}(\beta_n-\beta)\right\|=O_p\left(1\right)$$

as required. ■

6.2.6 Asymptotic Normality

The argument for asymptotic normality is now essentially the same as in the parametric case. The typical argument starts with an approximation of the first order conditions or the moment conditions:

$$Q_n^*(\beta_n)=Q_n^*(\beta_0)+\underbrace{\frac{\partial}{\partial\beta}Q_n^*(\tilde{\beta}_n)}_{d_m\times k}(\beta_n-\beta)$$

where $\tilde{\beta}_n$ is between β_0 and β_n. We then show that a ULLN holds for $\frac{\partial}{\partial\beta}Q_n^*(\tilde{\beta}_n)$ so that

$$\frac{\partial}{\partial\beta}Q_n^*(\tilde{\beta}_n)=D(\beta_0)+o_p\left(1\right):=D+o_p\left(1\right)$$

provided that $\beta_n=\beta_0+o_p\left(1\right)$. Therefore, the argmin of $\|Q_n^*(\beta)\|$ is asymptotically equivalent to the argmin of $\|D(\beta-\beta_0)+Q_n^*(\beta_0)\|$ when β is in the $\sqrt{n}$ neighborhood of β_0. To solve the latter minimization problem, we recognize that it is similar to an OLS problem. The solution is

$$\sqrt{n}\left(\beta_n^{**}-\beta_0\right)=-\left(D'D\right)^{-1}D'\sqrt{n}Q_n^*(\beta_0).$$

We complete the argument by proving that $\sqrt{n}Q_n^*(\beta_0)$ is asymptotically normal.

The above traditional argument relies on the differentiability of $Q_n^*(\beta)$. We now present an argument without assuming differentiability. Instead, we maintain a stochastic equicontinuity condition. Let $B_n = \{\beta : \|\beta - \beta_0\| \leq (\log n)/\sqrt{n}\}$. Since β_n is $\sqrt{n}$ consistent, we have $\beta_n \in B_n$ with probability approaching one.

Note that

$$\begin{aligned} Q_n^*(\beta) &= \{Q_n^*(\beta) - Q(\beta) - [Q_n^*(\beta_0) - Q(\beta_0)]\} \\ &\quad + \{Q_n^*(\beta_0) + Q(\beta)\}. \end{aligned}$$

Under the assumption of stochastic equicontinuity, the first term

$$\{Q_n^*(\beta) - Q(\beta) - [Q_n^*(\beta_0) - Q(\beta_0)]\} = o_p\left(1/\sqrt{n}\right)$$

if $\beta \in B_n$. So

$$\begin{aligned} \|Q_n^*(\beta)\| &= \|Q_n^*(\beta_0) + Q(\beta)\| + o_p\left(1/\sqrt{n}\right) \\ &= \|Q_n^*(\beta_0) + D(\beta - \beta_0)\| + o_p\left(1/\sqrt{n}\right) \end{aligned}$$

where the second equality uses only the differentiability of $Q(\beta) := Em(Z, h_0(V, \tau_0(V), \beta), \beta)$ and continuity of $\partial Q(\beta)/\partial\beta$ at $\beta = \beta_0$. While $m(Z, h_0(V, \tau_0(V), \beta), \beta)$ may not be differentiable, its expectation is likely to be differentiable as taking an expectation is a smoothing operator. For more discussions on this idea, see Andrews' Handbook of Econometrics Chapter (Ch 37, 1994).

Theorem 6.2.3 *Let Assumptions R, r, and I hold. In addition, assume*
(i) $Q_n^*(\beta) - Q(\beta)$ *is stochastically equicontinuous in that*

$$\sup_{\beta \in B_n} \left\| \sqrt{n}\left[Q_n^*(\beta) - Q(\beta)\right] - \sqrt{n}\left[Q_n^*(\beta_0) - Q(\beta_0)\right] \right\| = o_p(1).$$

(ii) $\sqrt{n}Q_n^*(\beta_0) \to_d N(0, \Omega)$ *with* $\|\Omega\| < \infty$.
(iii) $D(\beta)$ *is continuous at* $\beta = \beta_0$ *and* $D(\beta_0)$ *is of full rank* k.
(iv) β_0 *is in the interior of* $\mathcal{B}$.
Then

$$\sqrt{n}(\beta_n - \beta_0) \to_d N[0, (D'D)^{-1}D'\Omega D(D'D)^{-1}].$$

Note: with the result of $\sqrt{n}$-consistency, Assumption (i) here is equivalent to assumption (i) in the previous theorem. The only assumption we have added to the assumptions for the $\sqrt{n}$-consistency, is the strengthening of (ii) from$\sqrt{n}Q_n^*(\beta_0) = O_p(1)$ to the asymptotic normality assumption here, and the assumption that β_0 is an interior point. Also, as always, the form of the derivative matrix D depends only on the limit function $Q(\beta) = Em[Z, h_0(V, \tau_0(V), \beta), \beta]$.

Proof: By the stochastically equicontinuity assumption (i), we have

$$\|Q_n^*(\beta)\| = \|Q_n^*(\beta_0) + Q(\beta)\| + o_p\left(1/\sqrt{n}\right)$$

uniformly over

$$\beta \in \mathcal{B}_n = \left\{\beta : \|\beta - \beta_0\| \leq (\log n)/\sqrt{n}\right\}.$$

Under the condition that $D(\beta)$ is continuous at $\beta = \beta_0$, we have

$$\begin{aligned} Q(\beta) &= Q(\beta) - Q(\beta_0) = \frac{\partial Q\left(\tilde{\beta}_n\right)}{\partial \beta}(\beta - \beta_0) \\ &= D(\beta - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

uniformly over $\beta \in \mathcal{B}_n$. As a result

$$\|Q_n^*(\beta)\| = \|Q_n^*(\beta_0) + D(\beta - \beta_0)\| + o_p\left(\frac{1}{\sqrt{n}}\right)$$

uniformly over $\beta \in \mathcal{B}_n$ and

$$\arg\min_{\beta \in \mathcal{B}_n} \|Q_n^*(\beta)\| = \arg\min_{\beta \in \mathcal{B}_n} \|Q_n^*(\beta_0) + D(\beta - \beta_0)\| + o_p\left(\frac{1}{\sqrt{n}}\right).$$

That is,

$$\begin{aligned} \sqrt{n}(\beta_n - \beta_0) &= -\left(D'D\right)^{-1} D'\sqrt{n}Q_n^*(\beta_0) + o_p(1) \\ &\to {}^d N[0, (D'D)^{-1}D'\Omega D(D'D)^{-1}]. \end{aligned}$$

6.3 Some Technical Details

6.3.1 Asymptotic Normality of Objective Function

The asymptotic normality is a requirement that a weighted average of the difference between the nonparametric estimates and the true value of the estimated function (evaluated at β_0) distributes normally. Clearly sufficient conditions for this will be

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{m}(Z_i, h_0, \tau_0, \beta_0)\left(h_n(V_i, \tau_0, \beta_0) - h_0(V_i, \tau_0, \beta_0)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_1(Z_i) + o_p(1)$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{m}(Z_i, h_0, \tau_0, \beta_0)\,\dot{h}(V_i, \tau_0, \beta_0)\left(\tau_n(V_{2i}) - \tau_0(V_{2i})\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_2(Z_i) + o_p(1)$$

with $Ef_i(Z_i) = 0$ and $Ef_i^2(Z_i) < \infty$ for $i = 1, 2$. If the above are satisfied, then

$$\sqrt{n}Q_n^*(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[m(Z_i, h_0(\tau_0), \beta) + f_1(Z_i) + f_2(Z_i)\right] + o_p(1) \to_d N(0, \Omega)$$

for

$$\Omega = E\left[m(Z_i, h_0(\tau_0), \beta) + f_1(Z) + f_2(Z)\right]\left[m(Z_i, h_0(\tau_0), \beta) + f_1(Z) + f_2(Z)\right]'.$$

To illustrate the steps for establishing the asymptotic normality, we consider an example of the first case:

$$I_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}(Z_i, h_0, \tau_0, \beta_0)\left(h_n(V_i, \tau_0, \beta_0) - h_0(V_i, \tau_0, \beta_0)\right)$$

with

$$h_0(V_i, \tau_0, \beta_0) = E(Y_{1i}|V_i)$$

for some Y_{1i} that is possibly constructed from Z_i, *i.e.*, $Y_{1i} := Y_1(Z_i, \tau_0, \beta_0)$ and

$$h_n(v, \tau_0, \beta_0) = \left[\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{V_j - v}{b}\right)Y_{1j}\right]\left[\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{V_j - v}{b}\right)\right]^{-1}.$$

We assume a stochastic equicontinuity condition of the form: as $\delta_n \to 0$,

$$\sup_{\|h-h_0\|\le\delta_n}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}(Z_i, h_0, \tau_0, \beta_0)\left(h(V_i, \tau_0, \beta_0) - h_0(V_i, \tau_0, \beta_0)\right)\right.$$
$$\left.-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[\dot{m}(Z_i, h_0, \tau_0, \beta_0)\left(h(V_i, \tau_0, \beta_0) - h_0(V_i, \tau_0, \beta_0)\right)\right]\right\| = o_p(1).$$

That is,

$$\sup_{\|h-h_0\|\le\delta_n}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}(Z_i, h_0, \tau_0, \beta_0)\left(h(V_i, \tau_0, \beta_0) - h_0(V_i, \tau_0, \beta_0)\right)\right.$$
$$\left.-\sqrt{n}E\left[\dot{m}(Z_i, h_0, \tau_0, \beta_0)\left(h(V_i, \tau_0, \beta_0) - h_0(V_i, \tau_0, \beta_0)\right)\right]\right\| = o_p(1).$$

There is a quite large literature that provides sufficient conditions for stochastic equicontinuity when the index set is a function space. See Andrews (1994) for an overview.

Let

$$\begin{aligned}\rho(h) &= E\left[\dot{m}(Z, h_0, \tau_0, \beta_0)\, h(V, \tau_0, \beta_0)\right] = E\left\{E\left[\dot{m}(Z, h_0, \tau_0, \beta_0)\,|V\right] h(V, \tau_0, \beta_0)\right\} \\ &= \int\left[\dot{m}_E(v, h_0, \tau_0, \beta_0)\, h(v, \tau_0, \beta_0)\right] f_0(v)\, dv\end{aligned}$$

where $\dot{m}_E(v, h_0, \tau_0, \beta_0) = E[\dot{m}(Z, h_0, \tau_0, \beta_0) | V = v]$ and $f_0(v)$ is the pdf of V. Then

$$\begin{aligned}
I_1 &= \sqrt{n}[\rho(h_n) - \rho(h_0)] \\
&= \sqrt{n}\int \dot{m}_E(v, h_0, \tau_0, \beta_0)[h_n(v, \tau_0, \beta_0) - h_0(v, \tau_0, \beta_0)] f_0(v)\, dv \\
&= \sqrt{n}\int \dot{m}_E(v, h_0, \tau_0, \beta_0)\left[\frac{q_n(v, \tau_0, \beta_0)}{f_n(v)} - h_0(v, \tau_0, \beta_0)\right] f_0(v)\, dv \\
&= \sqrt{n}\int \frac{\dot{m}_E(v, h_0, \tau_0, \beta_0)}{f_n(v)}[q_n(v, \tau_0, \beta_0) - f_n(v) h_0(v, \tau_0, \beta_0)] f_0(v)\, dv \\
&= \sqrt{n}\int \dot{m}_E(v, h_0, \tau_0, \beta_0)[q_n(v, \tau_0, \beta_0) - f_n(v) h_0(v, \tau_0, \beta_0)]\, dv + o_p(1)
\end{aligned}$$

where

$$q_n(v, \tau_0, \beta_0) = \frac{1}{nb}\sum_{j=1}^{n} K\left(\frac{V_j - v}{b}\right) Y_{1j} \text{ and } f_n(v) = \frac{1}{nb}\sum_{j=1}^{n} K\left(\frac{V_j - v}{b}\right).$$

Now

$$\begin{aligned}
I_1 &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[\int \frac{1}{b} K\left(\frac{V_j - v}{b}\right)\dot{m}_E(v, h_0, \tau_0, \beta_0)[Y_{1j} - h_0(v, \tau_0, \beta_0)]\, dv\right] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}_E(V_i, h_0, \tau_0, \beta_0)[Y_{1i} - h_0(V_i, \tau_0, \beta_0)] + o_p(1).
\end{aligned}$$

The intuition here is very clear. The contribution of the first stage to the variance of the estimator depends on:

- The variance about the regression function (as this determines the variance of the estimator of the regression function about its true value) and
- The derivative of the moment function m with respect to the unknown function h evaluated at the true function.

By the same argument, if $\tau(V_{2i}) = E(Y_{2i}|V_{2i})$ for some $Y_{2i} := Y_2(Z_i)$, then

$$I_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[\dot{m}(Z_i, h_0, \tau_0, \beta_0)\dot{h}(V_i, \tau_0, \beta_0) | V_{2i}\right](Y_{2i} - \tau_0(V_{2i})) + o_p(1)$$

6.3.2 Orthogonality Conditions

As in the two step estimation in the parametric case, the asymptotic variance simplifies when "orthogonality" conditions hold. We now consider two special cases:

(a) $E\left[\dot{m}\left(Z_i, h_0, \tau_0, \beta_0\right)|V_i\right] = 0$ (a.e. P),

In this case, we have

$$\Omega = Em\left(Z_i, h_0, \tau_0, \beta_0\right)\left[m\left(Z_i, h_0, \tau_0, \beta_0\right)\right]'.$$

(b) $E\left\{\dot{m}\left(Z_i, h_0, \tau_0, \beta_0\right)\dot{h}_0\left(V_i, \tau_0, \beta_0\right)|V_{2i}\right\} = 0$ (a.e. P),

In this case, we have

$$\Omega = E\left[m\left(Z_i, h_0, \tau_0, \beta_0\right) + f_1\left(Z_i\right)\right]\left[m\left(Z_i, h_0, \tau_0, \beta_0\right) + f_1\left(Z_i\right)\right]'$$

In case (a) we need not correct the variance-covariance of our estimate of β for the fact that we are using estimated $h_n(\cdot)$ and $\tau_n(\cdot)$ rather than the actual unknown functions $h(\cdot)$ and $\tau(\cdot)$, while in the second case we must make a correction for the fact that $h(\cdot)$ is estimated but no correction is needed for the fact that $\tau(\cdot)$ is estimated.

These are the analogues of the orthogonality conditions for the two step estimators in the finite dimensional case. There the difference in the first step estimator and its true value has a first order impact that does not differ across observations. Consequently, an average derivative has to be zero for the orthogonality condition to hold. Here we require that the derivative be zero for every value of V, since the difference between the estimated and actual values will be different for different values of V.

Going back to the semiparametric examples, note that if

$$m(Z, \beta) = q(Z, \beta)h(X, \tau(X), \beta)$$

and we use a first stage estimate of $(h(\cdot), \tau(\cdot), \beta)$ and minimize a norm in the mean of

$$m(Z, \beta; h_n, \tau_n, \beta_n) = q(Z, \beta)h_n(X, \tau_n(X), \beta_n),$$

then the derivative of the limit function with respect to any of the first stage parameters will be some function of X times $q(Z, \beta)$. So if $E[q(Z, \beta)|X] = 0$, the orthogonality condition holds and we do not have to adjust for the estimation uncertainty in h_n and τ_n.

It is a good exercise to how this is true in the example where we use a nonparametric estimate of the variance of the regression function to perform WLLS. This is not true however in the selection problem or in the more complicated problems referred to in the notes. We should keep in mind that even in the cases where the orthogonality conditions hold, there is a real issue of how well we do in finite samples when we use those approximations, as we are relying on asymptotics in a rather integral way for all of the arguments.

6.4 Case Studies

I assume that all the conditions can be verified for the models we consider. Many papers in the literature provide primitive conditions for our assumptions. The objective here is to derive the form of the asymptotic distribution for each popular semiparametric model.

6.4.1 Partial Linear Model

The model:

$$Y = X\beta_0 + f_0(V) + u \tag{6.11}$$

where $E(u|X,V) = 0$ and $var(u|X,V) = \sigma^2$. Let

$$h(V) = (h_1(V), h_2(V))$$

and

$$h_0(V) = (h_{01}(V), h_{02}(V)) = [E(X|V), E(Y|V)].$$

The moment condition is:

$$Em(\beta_0, h_0) = 0$$

where

$$m(\beta, h) = (X - h_1(V))[Y - h_2(V) - (X - h_1(V))\beta].$$

Compute $D(\beta)$

Note that

$$\begin{aligned} Q(\beta) &= Em(\beta, h_0) = E[X - h_{01}(V)][Y - h_{02}(V) - (X - h_{01}(V))\beta] \\ &= E[[X - h_{01}(V)]E\{[Y - h_{02}(V) - (X - h_{01}(V))\beta]|X, V\}] \\ &= E\{[X - h_{01}(V)][h_{02}(V) + (X - h_{01}(V))\beta_0 - h_{02}(V) - (X - h_{01}(V))\beta]\} \\ &= E\{[X - h_{01}(V)][(X - h_{01}(V))](\beta_0 - \beta)\} \\ &= E[X - h_{01}(V)]^2(\beta_0 - \beta). \end{aligned}$$

So

$$D(\beta) = \frac{\partial Q(\beta)}{\partial \beta} = -E[X - h_{01}(V)]^2.$$

Limiting Distribution

$$m^*(\beta_0, h) = m(\beta_0, h_0) + \dot{m}(\beta_0, h_0)(h - h_0)$$

where

$$\begin{aligned} m(\beta_0, h_0) &= (X - h_{01}(V))[Y - h_{02}(V) - (X - h_{01}(V))\beta_0] \\ &= (X - h_{01}(V))u \end{aligned}$$

and

$$\begin{aligned} \dot{m}_1(\beta_0, h_0) &= -[Y - h_2(V) - (X - h_1(V))\beta_0] + (X - h_1(V))\beta_0|_{h=h_0} \\ &= -u + (X - h_{01}(V))\beta_0 \end{aligned}$$

$$\dot{m}_2(\beta_0, h_0) = -(X - h_1(V))|_{h=h_0} = -(X - h_{01}(V))$$

so

$$\begin{aligned} \sqrt{n}Q_n^*(\beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n [X_i - h_{01}(V_i)]u_i + \beta_0 \frac{1}{\sqrt{n}}\sum_{i=1}^n [X_i - h_{01}(V_i)](h_{n1} - h_{01}) \\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - h_{01}(V_i))(h_{n2} - h_{02}) - \frac{1}{\sqrt{n}}\sum u_i (h_{n1} - h_{01}) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n [X_i - h_{01}(V_i)]u_i + o(1) \end{aligned}$$

where we have used the stochastic equicontinuity of $\nu_n(h)$:

$$\nu_n(h) = \frac{1}{\sqrt{n}}\sum_{i=1}^n [X_i - h_{01}(V_i)]h(V_i) \text{ or } \nu_n(h) = \frac{1}{\sqrt{n}}\sum u_i (h_{n1} - h_{01})$$

and

$$E(X_i - h_{01}(V_i)|V_i) = 0, \; E(u_i|V_i) = 0.$$

So for the partial linear model, the orthogonal condition holds. As a result, the estimation uncertainty in h_{n1} and h_{n2} does not factor into the asymptotic variance of $Q_n^*(\beta_0)$. Therefore

$$\sqrt{n}Q_n^*(\beta_0) \to^d N(0, \sigma^2 E[X - E(X|V)]^2)$$

and

$$\begin{aligned} \sqrt{n}(\beta_n - \beta_0) &= -\left(D'D\right)^{-1} D'\sqrt{n}Q_n^*(\beta_0) + o_p(1) = -D^{-1}\sqrt{n}Q_n^*(\beta_0) + o_p(1) \\ &\to {}^d N(0, \sigma^2 \{E[X - E(X|V)]\}^{-2}). \end{aligned}$$

6.4.2 Single Index Model

The model:

$$Y = f(X\beta_0) + u \tag{6.12}$$

where $E(u|X) = 0$ and so $E(Y|X\beta_0) = f(X\beta_0)$. Let

$$h_0(\beta, \gamma) = E(Y|X\beta = \gamma)$$

then

$$f(X\beta_0) = h_0(\beta_0, X\beta_0).$$

To understand $h_0(\beta, \gamma)$, we consider an example where $X = (X_1, X_2)' \sim N(0, I_2)$, $u \sim N(0,1)$, $X \perp u$ and $f(w) = w$. Then

$$h_0(\beta, \gamma) = E(Z_{\beta_0} + u|Z_\beta = \gamma)$$

where $Z_\beta := X\beta$. Noting that we can write

$$Z_{\beta_0} = \frac{cov(Z_\beta, Z_{\beta_0})}{var(Z_\beta)} Z_\beta + V_\beta = \frac{\beta'\beta_0}{\beta'\beta} Z_\beta + V_\beta$$

for $V_\beta \perp Z_\beta$, we have

$$\begin{aligned} h_0(\beta, \gamma) &= \frac{\beta'\beta_0}{\beta'\beta}\gamma + E\left(V_\beta^2|Z_\beta = \gamma\right) = \frac{\beta'\beta_0}{\beta'\beta}\gamma + E\left(V_\beta^2\right) \\ &= \frac{\beta'\beta_0}{\beta'\beta}\gamma + \beta_0'\beta_0 - \frac{\left(\beta'\beta_0\right)^2}{\left(\beta'\beta\right)^2}\left(\beta'\beta\right) \end{aligned}$$

That is

$$h_0(\beta, X\beta) = \frac{\beta'\beta_0}{\beta'\beta} X\beta + \beta_0'\beta_0 - \frac{\left(\beta'\beta_0\right)^2}{\left(\beta'\beta\right)^2}\left(\beta'\beta\right).$$

If we impose the normalization that $\beta'\beta = 1$, then

$$h_0(\beta, X\beta) = \left(\beta'\beta_0\right) X\beta + \beta_0'\beta_0 - \left(\beta'\beta_0\right)^2.$$

For notational simplicity, we drop the subscript on $h_0(\beta, \gamma)$ and write it as $h(\beta, \gamma)$. This should not cause any confusion.

Compute $D(\beta)$

The moment condition can be written as

$$Em(Z, h(\beta_0, X\beta_0)) = 0$$

where

$$\begin{aligned} m(Z, h(\beta, X\beta)) &= \frac{dh(\beta, X\beta)}{d\beta}\left[Y - h(\beta, X\beta)\right] \\ &= \left\{h_\beta(\beta, X\beta) + h_\gamma(\beta, X\beta) X\right\}\left[Y - h(\beta, X\beta)\right]. \end{aligned}$$

The limiting function is

$$\begin{aligned} Q(\beta) &= E\left\{h_\beta(\beta, X\beta) + h_\gamma(\beta, X\beta) X\right\}\left[Y - h(\beta, X\beta)\right] \\ &= E\left\{h_\beta(\beta, X\beta) + h_\gamma(\beta, X\beta) X\right\}\left[h(\beta_0, X\beta_0) + u - h(\beta, X\beta)\right] \\ &= E\left\{h_\beta(\beta, X\beta) + h_\gamma(\beta, X\beta) X\right\}\left[h(\beta_0, X\beta_0) - h(\beta, X\beta)\right]. \end{aligned}$$

Note the the function $h(\cdot, \cdot)$ in the above expression is the true conditional mean function $E(Y|X\beta = \gamma)$. So

$$D(\beta_0) = \frac{\partial Q(\beta)}{\partial \beta}|_{\beta=\beta_0} = -E\left[h_\beta(\beta_0, X\beta_0) + h_\gamma(\beta_0, X\beta_0) X\right]'\left[h_\beta(\beta_0, X\beta_0) + h_\gamma(\beta_0, X\beta_0) X\right]$$

where

$$h_\beta(\beta_0, X\beta_0) = \frac{\partial E(Y|X\beta = \gamma)}{\partial \beta}|_{\beta=\beta_0, \gamma=X\beta_0} = \frac{\partial E\left[f(X\beta_0)|X\beta = \gamma\right]}{\partial \beta}|_{\beta=\beta_0, \gamma=X\beta_0}$$

and

$$h_\gamma(\beta_0, X\beta_0) = \frac{\partial E(Y|X\beta = \gamma)}{\partial \gamma}|_{\beta=\beta_0, \gamma=X\beta_0} = \frac{\partial E\left[f(X\beta_0)|X\beta_0 = \gamma\right]}{\partial \gamma}|_{\gamma=X\beta_0} = f'(X\beta_0).$$

To obtain an analytical expression for $h_\beta(\beta_0, X\beta_0)$, we note that as $\|\varepsilon\| \to 0$

$$\begin{aligned} & E\left[f(X\beta_0)|X(\beta_0 + \varepsilon) = \gamma\right] - E\left[f(X\beta_0)|X\beta_0 = \gamma\right] \\ & E\left[f(X\beta_0)|X\beta_0 = \gamma - X\varepsilon\right] - E\left[f(X\beta_0)|X\beta_0 = \gamma\right] \\ = \; & Ef(\gamma - X\varepsilon|X\beta_0 = \gamma - X\varepsilon) - f(\gamma) \\ = \; & -f'(\gamma)E(X\varepsilon|X\beta_0 = \gamma - X\varepsilon) + o(\|\varepsilon\|) \\ = \; & -f'(\gamma)E(X|X\beta_0 = \gamma)\varepsilon + o(\|\varepsilon\|) \end{aligned}$$

which implies that

$$h_\beta(\beta_0, X\beta_0) = -f'(X\beta_0)E(X|X\beta_0).$$

Hence

$$D(\beta_0) = E\left\{\left[f'(X\beta_0)\right]^2\left[X - E(X|X\beta_0)\right]'\left[X - E(X|X\beta_0)\right]\right\}$$

Limiting Distribution

The objective function is the norm of:

$$\begin{aligned} Q_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{dh_n(\beta, X_i\beta)}{d\beta}\left[Y_i - h_n(\beta, X_i\beta)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{dh_0(\beta, X_i\beta)}{d\beta}\left[Y_i - h_n(\beta, X_i\beta)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{dh_n(\beta, X_i\beta)}{d\beta} - \frac{dh_0(\beta, X_i\beta)}{d\beta}\right]\left[Y_i - h_n(\beta, X_i\beta)\right]. \end{aligned}$$

Note that we can approximate $Q_n(\beta)$ by

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\frac{dh_0(\beta, X_i\beta)}{d\beta}\left[Y_i - h_0(\beta, X_i\beta)\right] \\ &+\frac{1}{n}\sum_{i=1}^{n}\frac{dh_0(\beta, X_i\beta)}{d\beta}\left[h_0(\beta, X_i\beta) - h_n(\beta, X_i\beta)\right]. \\ &+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{dh_n(\beta, X_i\beta)}{d\beta} - \frac{dh_0(\beta, X_i\beta)}{d\beta}\right]\left[Y_i - h_0(\beta, X_i\beta)\right] \\ &+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{dh_n(\beta, X_i\beta)}{d\beta} - \frac{dh_0(\beta, X_i\beta)}{d\beta}\right]\left[h_0(\beta, X_i\beta) - h_n(\beta, X_i\beta)\right] \\ &+\text{high order terms.} \end{aligned}$$

The first order approximation is

$$\begin{aligned} Q_n^*(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{dh_0(\beta, X_i\beta)}{d\beta}\left[Y_i - h_0(\beta, X_i\beta)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{dh_0(\beta, X_i\beta)}{d\beta}\left[h_0(\beta, X_i\beta) - h_n(\beta, X_i\beta)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{dh_n(\beta, X_i\beta)}{d\beta} - \frac{dh_0(\beta, X_i\beta)}{d\beta}\right]\left[Y_i - h_0(\beta, X_i\beta)\right] \end{aligned}$$

Plugging in $\beta = \beta_0$ yields

$$\begin{aligned}\sqrt{n}Q_n^*(\beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\left[X_i - E(X_i|X_i\beta_0)\right]u_i \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{dh_n(\beta, X_i\beta)}{d\beta} - \frac{dh_0(\beta, X_i\beta)}{d\beta}\right]_{\beta=\beta_0} u_i \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\left[X_i - E(X_i|X_i\beta_0)\right]\left[h_0(\beta_0, X_i\beta_0) - h_n(\beta_0, X_i\beta_0)\right]. \\ &\rightarrow {}_dN(0, \Omega_\beta).\end{aligned}$$

where

$$\Omega_\beta = E\left\{\left[f'(X_i\beta_0)\right]^2\left[X_i - E(X_i|X_i\beta_0)\right]'\left[X_i - E(X_i|X_i\beta_0)\right]Var(u_i|X_i)\right\}.$$

Here we have assumed that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{dh_n(\beta, X_i\beta)}{d\beta} - \frac{dh_0(\beta, X_i\beta)}{d\beta}\right]_{\beta=\beta_0} u_i = o_p(1)$$

which is reasonable, and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\left[X_i - E(X_i|X_i\beta_0)\right]\left[h_0(\beta_0, X_i\beta_0) - h_n(\beta_0, X_i\beta_0)\right] = o_p(1)$$

which holds if the process

$$v_n(h) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\left[X_i - E(X_i|X_i\beta_0)\right]\left[h_0(\beta_0, X_i\beta_0) - h(\beta_0, X_i\beta_0)\right]$$

is stochastically equicontinuous such that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\left[X_i - E(X_i|X_i\beta_0)\right]\left[h_0(\beta_0, X_i\beta_0) - h(\beta_0, X_i\beta_0)\right] \\ &= E\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\left[X_i - E(X_i|X_i\beta_0)\right]\left[h_0(\beta_0, X_i\beta_0) - h(\beta_0, X_i\beta_0)\right] + o_p(1) \\ &= E\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f'(X_i\beta_0)\, E\left\{\left[X_i - E(X_i|X_i\beta_0)\right]|X_i\beta_0\right\}\left[h_0(\beta_0, X_i\beta_0) - h(\beta_0, X_i\beta_0)\right] + o_p(1) \\ &= o_p(1).\end{aligned}$$

So

$$\sqrt{n}\left(\hat{\beta}-\beta\right)=-D^{-1}\sqrt{n}Q_n^*\left(\beta_0\right)+o_p\left(1\right)\rightarrow_d N(0,\Omega_\beta)$$

where

$$\Omega_\beta=D^{-1}\left(\beta_0\right)\Omega\left[D^{-1}\left(\beta_0\right)\right]'.$$

In the case of homoscedasticity, we have

$$\Omega_\beta=\sigma_u^2\left\{E\left[f'\left(X\beta_0\right)\right]^2\left[X-E\left(X|X\beta_0\right)\right]'\left[X-E\left(X|X\beta_0\right)\right]\right\}^{-1}$$

Had we know the function form of $f\left(\cdot\right)$, the asymptotic variance of the NLLS would be

$$\Omega_{NLLS}=\sigma_u^2\left\{E\left[f'\left(X\beta_0\right)\right]^2X'X\right\}^{-1}.$$

Clearly $\Omega_\beta\geq\Omega_{NLLS}$. This is the cost of the semiparametric estimation.

6.4.3 Sample Selection Model

The model:

$$\begin{aligned}Y_1^* &= f_0(X_1,\beta_0)+u_1,\\ Y_2^* &= -q_0(X_2)+u_2,\end{aligned}$$

where (Y_1^*,Y_2^*) are unobserved latent variables. Moreover, we assume that

$$u_1=\rho_0u_2+\varepsilon$$

where u_2 and ε are independent. The observed variables are

$$\begin{aligned}Y_1 &= Y_1^*\mathbf{1}\left\{Y_2^*>0\right\}\\ Y_2 &= \mathbf{1}\left\{Y_2^*>0\right\}.\end{aligned}$$

Compute $D(\beta)$

The moment condition is

$$E\left[m(Z,h_0\left(\tau_0\left(X_2\right),\beta_0\right),\beta_0)\right]=0$$

where

$$\begin{aligned}&m(Z,h\left(\tau\left(X_2\right),\beta\right),\beta)\\ =\ &\frac{\partial\left[f(X_1,\beta)+h\left(\tau\left(X_2\right),\beta\right)\right]}{\partial\beta}\left[Y_1-f(X_1,\beta)-h\left(\tau\left(X_2\right),\beta\right)\right]\\ =\ &\left\{\frac{\partial f(X_1,\beta)}{\partial\beta}-E\left[\frac{\partial f(X_1,\beta)}{\partial\beta}|\tau\left(X_2\right)\right]\right\}\left[Y_1-f(X_1,\beta)-h\left(\tau\left(X_2\right),\beta\right)\right]\end{aligned}$$

and the expectation here and in the rest of this subsection is conditional on $Y_2 = 1$. Here we have used the definition that

$$h\left(\tau\left(X_2\right), \beta\right) = E(Y_1 - f(X_1, \beta) | \tau\left(X_2\right), Y_2 = 1)$$

to compute $\partial h\left(\tau\left(X_2\right), \beta\right) / \partial \beta$.

The limiting function is

$$\begin{aligned}
& Q(\beta) \\
= \; & E \frac{\partial\left[f_0(X_1, \beta) + h_0\left(\tau\left(X_2\right), \beta\right)\right]}{\partial \beta}\left[Y_1 - f_0(X_1, \beta) - h_0\left(\tau_0\left(X_2\right), \beta\right)\right] \\
= \; & E \frac{\partial\left[f_0(X_1, \beta) + h_0\left(\tau\left(X_2\right), \beta\right)\right]}{\partial \beta}\left[Y_1 - f_0(X_1, \beta) - h_0\left(\tau_0\left(X_2\right), \beta\right)\right] \\
= \; & E \frac{\partial\left[f_0(X_1, \beta) + h_0\left(\tau\left(X_2\right), \beta\right)\right]}{\partial \beta}\left[f_0(X_1, \beta_0) - f(X_1, \beta) + h_0\left(\tau_0\left(X_2\right), \beta_0\right) - h_0\left(\tau_0\left(X_2\right), \beta\right)\right].
\end{aligned}$$

So

$$\begin{aligned}
& D = D(\beta_0) \\
= \; & E\, \frac{\partial\left[f_0(X_1, \beta) + h_0\left(\tau\left(X_2\right), \beta\right)\right]}{\partial \beta} \frac{\partial\left[f(X_1, \beta) - h_0\left(\tau_0\left(X_2\right), \beta\right)\right]}{\partial \beta'}\Bigg|_{\beta=\beta_0} \\
= \; & E\left\{\frac{\partial f_0(X_1, \beta)}{\partial \beta} - E\left[\frac{\partial f_0(X_1, \beta)}{\partial \beta} | \tau_0\left(X_2\right)\right]\right\}\left\{\frac{\partial f_0(X_1, \beta)}{\partial \beta} - E\left[\frac{\partial f_0(X_1, \beta)}{\partial \beta} | \tau_0\left(X_2\right)\right]\right\}'\Bigg|_{\beta=\beta_0}
\end{aligned}$$

To obtain the asymptotic distribution of β, we compute $\sqrt{n} Q_n^*(\beta_0)$ as follows:

$$\begin{aligned}
& m(Z, h\left(\tau\left(X_2\right), \beta\right), \beta) \\
= \; & \frac{\partial\left[f(X_1, \beta) + h\left(\tau\left(X_2\right), \beta\right)\right]}{\partial \beta}\left[Y_1 - f(X_1, \beta) - h\left(\tau\left(X_2\right), \beta\right)\right] \\
= \; & \left\{\frac{\partial f(X_1, \beta)}{\partial \beta} - E\left[\frac{\partial f(X_1, \beta)}{\partial \beta} | \tau\left(X_2\right)\right]\right\}\left[Y_1 - f(X_1, \beta) - h\left(\tau\left(X_2\right), \beta\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\sqrt{n}Q_n^*(\beta_0) \\
= &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}-E\left[\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}|\tau_0\left(X_{2i}\right)\right]\right\}e_i \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial h_n\left(\tau_n\left(X_2\right),\beta_0\right)}{\partial\beta}-\frac{\partial h_0\left(\tau_0\left(X_2\right),\beta_0\right)}{\partial\beta}\right]e_i \\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}-E\left[\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}|\tau_0\left(X_{2i}\right)\right]\right\}\left[h_n\left(\tau_0\left(X_{2i}\right),\beta_0\right)-h_0\left(\tau_0\left(X_{2i}\right),\beta_0\right)\right] \\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}-E\left[\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}|\tau_0\left(X_{2i}\right)\right]\right\}\dot{h}_0\left(\tau_0\left(X_{2i}\right),\beta_0\right)\left[\tau_n\left(X_{2i}\right)-\tau_0\left(X_{2i}\right)\right]
\end{aligned}
$$

where

$$e_i=\left[Y_{1i}-f(X_{1i},\beta_0)-h_0\left(\tau_0\left(X_{2i}\right),\beta_0\right)\right].$$

Because the orthogonality conditions hold, the last two terms can be shown to be $o_p\left(1\right)$. Let

$$g_n\left(X_2,\beta_0\right)=\frac{\partial h_n\left(\tau_n\left(X_2\right),\beta_0\right)}{\partial\beta}\text{ and }g_0\left(X_2,\beta_0\right)=\frac{\partial h_0\left(\tau_0\left(X_2\right),\beta_0\right)}{\partial\beta}$$

By a stochastic equicontinuity argument,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g_n\left(X_2,\beta_0\right)-g_0\left(X_2,\beta_0\right)\right]e_i=\sqrt{n}E\left[g_n\left(X_2,\beta_0\right)-g_0\left(X_2,\beta_0\right)\right]e_i+o_p\left(1\right)=o_p\left(1\right)$$

Hence

$$\sqrt{n}Q_n^*(\beta_0)\to N(0,\Omega)$$

where

$$
\begin{aligned}
\Omega = &\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}-E\left[\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}|\tau_0\left(X_{2i}\right)\right]\right\} \\
&\times\left\{\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}-E\left[\frac{\partial f_0(X_{1i},\beta_0)}{\partial\beta}|\tau_0\left(X_{2i}\right)\right]\right\}'\sigma_e^2\left(X_i\right)
\end{aligned}
$$

and $\sigma_e^2\left(X_i\right)=var(e_i|X_i)$.

Therefore, we have $\sqrt{n}\left(\beta_n-\beta_0\right)\to N(0,D^{-1}\Omega D^{-1\prime})$.

6.5 Non-differentiable Objective Function

This section is based on Chen, Linton and Keilegom (2003) and Chen (2006, Sec. 4.1) which assume away τ. Their approach can be generalized easily to the case with $\tau\left(\cdot\right)$ function.

Following Chen, Linton and Keilegom (2003), we consider the moment condition:

$$Q\left(h_0, \beta_0\right) = 0 \text{ for } Q\left(h, \beta\right) = Em(Z_i, h(Z_i, \beta), \beta).$$

The sample analogue is

$$Q_n(h, \beta) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, h(Z_i, \beta), \beta).$$

We will assume that $Q\left(h, \beta\right)$ is a continuous functions of its arguments, although $m(\cdot, \cdot, \cdot)$ need not be. Our notation is slightly different from the previous section as we emphasize the dependence of Q and Q_n on the unknown function h.

As an example of non-differentiable objective function, consider the model:

$$Y_i = X_{1i}\beta_0 + h_*\left(X_{2i}\right) + \varepsilon_i$$

where

$$P(\varepsilon_i < 0 | X_{1i}, X_{2i}) = \alpha \in (0, 1)$$

That is, conditioning on $X_i = (X_{1i}, X_{2i})$, the 100α-th quantile is $X_{1i}\beta_0 + h_*\left(X_{2i}\right)$. We could write $X_{1i}\beta_0 + h_*\left(X_{2i}\right)$ as $X_{1i}\beta_0^{\alpha} + h_*^{\alpha}\left(X_{2i}\right)$ if we want to emphasize that it is the 100α-th conditional quantile. Let $h_0(X_2, \beta) = 100\alpha$-th quantile of $Y - X_1\beta$ given X_2. Obviously, $h_0(X_2, \beta_0) = h_*\left(X_{2i}\right)$. To estimate β_0, we can solve $\arg\min_{\beta} \left\|Q_n(h_n, \beta)\right\|$ where

$$Q_n\left(h_n, \beta\right) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, h_n(X_{2i}, \beta), \beta)$$

$$m(Z_i, h(X_{2i}, \beta), \beta) = \left[1\left\{Y_i < X_{1i}\beta + h\left(X_{2i}, \beta\right)\right\} - \alpha\right] X_{1i}$$

and $h_n(X_{2i}, \beta)$ is a first stage nonparametric estimator of $h_0\left(X_{2i}, \beta\right)$.

6.5.1 Consistency

Theorem 6.5.1 *Assume $\beta_0 \in \mathcal{B}$ and $Q(h_0, \beta_0) = 0$ and*

(i) (Definition of β_n)

$$\left\|Q_n(h_n\left(\cdot, \beta_n\right), \beta_n)\right\| \leq \inf \left\|Q_n(h_n\left(\cdot, \beta\right), \beta)\right\| + o_p\left(1\right).$$

(ii) (identification) $\forall \delta > 0, \exists \varepsilon\left(\delta\right) > 0$ such that

$$\inf_{\|\beta - \beta_0\| > \delta} \left\|Q(h_0\left(\cdot, \beta\right), \beta)\right\| \geq \varepsilon\left(\delta\right) > 0.$$

(iii) (Continuity of Q) Uniformly over $\beta \in \mathcal{B}, Q(h, \beta)$ is continuous in $h\left(\cdot\right)$ at $h = h_0$.

(iv) (Consistency of h_n)

$$\|h_n - h_0\|_{\mathcal{H}} := \sup_{\beta \in \mathcal{B}} \|h_n(Z, \beta) - h_0(Z, \beta)\| = o_p(1).$$

(v) For all δ_n such that $\delta_n = o_p(1)$

$$\sup_{\beta \in \mathcal{B}, \|h - h_0\|_{\mathcal{H}} \leq \delta_n} \|Q_n(h(\cdot, \beta), \beta) - Q(h(\cdot, \beta), \beta)\| = o_p(1).$$

Then

$$\beta_n - \beta_0 = o_p(1).$$

The continuity in (iii) is w.r.t. a metric $\|\cdot\|_{\mathcal{H}}$ in the function space, which we will usually take the following forms:

- sup of sup-norm
- sup of L_2 norm with respect to the Lesbegue measure
- sup of L_2 norm with respect to the probability measure of the underlying random variable.

To emphasize that it is a norm for a function space, we sometimes write it as $\|\cdot\|_{\mathcal{H}}$ as we do in (iv). The assumption says that we can ensure

$$\|Q(h, \beta) - Q(h_0, \beta)\| < \epsilon$$

by choosing $\|h - h_0\|_{\mathcal{H}} \leq \delta(\epsilon)$ for some $\delta(\epsilon)$ that does not depend on β. Note that we are only requiring the continuity of the limiting function $Q(h, \beta)$.

Remark 6.5.1 *We have not required $m(\cdot, \cdot, \cdot)$ (or, as a result $Q_n(h, \beta)$) to be continuous in β or in h.*

Remark 6.5.2 *If $h_n(\cdot) = h_0(\cdot)$ everywhere, assumptions (iii) and (iv) are unnecessary, and we are back to exactly the same assumptions used in a parametric problem.*

Remark 6.5.3 *Intuitively, we expect $\beta_n = \arg\min_\beta \|Q_n(h_n(\cdot, \beta), \beta)\| \to \arg\min_\beta \|Q(h_n(\cdot, \beta), \beta)\| \to \arg\min_\beta \|Q(h_0(\cdot, \beta), \beta)\| = \beta_0$. For the first convergence to hold, we need conditions (iv) and (v). For the second convergence to hold, we need condition (iii).*

Proof: Let $\delta > 0$. By assumption (ii), there exists $\varepsilon(\delta) > 0$ such that whenever $\beta \in \mathcal{B} \backslash B(\beta_0, \delta)$ we have

$$\|Q(h_0(\cdot, \beta), \beta) - Q(h_0(\cdot, \beta_0), \beta_0)\| \geq \varepsilon(\delta) > 0 \tag{6.13}$$

Thus

$$\begin{aligned} P(||\beta_n - \beta_0|| \ > \ \delta) &= P(\beta_n \in \mathcal{B} \backslash B(\beta_0, \delta)) \\ &\leq P\left[\|Q(h_0(\cdot, \beta_n), \beta_n) - Q(h_0(\cdot, \beta_0), \beta_0)\| \geq \varepsilon(\delta)\right]. \end{aligned}$$

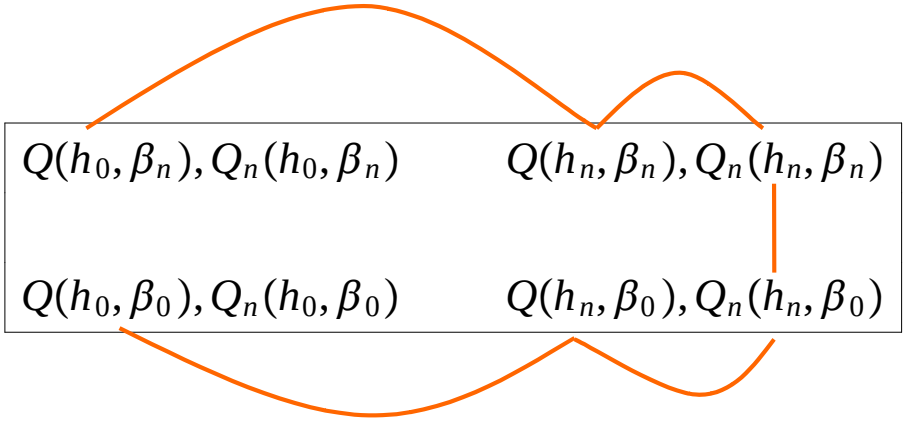
It suffices to show that

$$\|Q(h_0(\cdot, \beta_n), \beta_n) - Q(h_0(\cdot, \beta_0), \beta_0)\| = o_p(1).$$

From the triangle inequality

$$\begin{aligned} & \|Q(h_0(\cdot, \beta_n), \beta_n) - Q(h_0(\cdot, \beta_0), \beta_0)\| \\ \leq \ & \|Q(h_0(\cdot, \beta_n), \beta_n) - Q(h_n(\cdot, \beta_n), \beta_n)\| + \|Q(h_n(\cdot, \beta_n), \beta_n) - Q_n(h_n(\cdot, \beta_n), \beta_n)\| \\ & + \|Q_n(h_n(\cdot, \beta_n), \beta_n) - Q(h_0(\cdot, \beta_0), \beta_0)\| \\ = \ & o_p(1) + o_p(1) + \|Q_n(h_n, \beta_n)\| \end{aligned}$$

where the first $o_p(1)$ comes from the continuity of Q in h and the consistency of h_n (Assumptions (iii) and (iv)) and the second $o_p(1)$ from the consistency of h_n (this gets us in the δ_n neighborhood) and the ULLN (Assumptions (iv) and (v)). To help track the terms in above derivation, we may use a table of the form:



For each Q_n or Q, the second argument of the h function is the same as the second argument of the Q_n or Q function

We are left with proving that $\|Q_n(h_n(\cdot,\beta_n),\beta_n)\| = o_p(1)$. Now by the definition of β_n,

$$\begin{aligned}
\|Q_n(h_n(\cdot,\beta_n),\beta_n)\| &\leq \inf_{\beta\in\mathcal{B}} \|Q_n(h_n(\cdot,\beta),\beta)\| + o_p(1) \\
&\leq \|Q_n(h_n(\cdot,\beta_0),\beta_0)\| + o_p(1) \\
&\leq \|Q_n(h_n(\cdot,\beta_0),\beta_0) - Q(h_n(\cdot,\beta_0),\beta_0)\| \\
&\quad + \|Q(h_n(\cdot,\beta_0),\beta_0) - Q(h_0(\cdot,\beta_0),\beta_0)\| + o_p(1) \\
&\leq \sup_{\beta\in\mathcal{B},\|h-h_0\|_{\mathcal{H}}\leq\delta_n} \|Q_n(h(\cdot,\beta),\beta) - Q(h(\cdot,\beta),\beta)\| \\
&\quad + \sup_{\|h-h_0\|_{\mathcal{H}}\leq\delta_n} \|Q(h(\cdot,\beta_0),\beta_0) - Q(h_0(\cdot,\beta_0),\beta_0)\| \\
&= o_p(1)
\end{aligned}$$

using Assumptions (iii)–(v). ■

6.5.2 Asymptotic Normality

For notational simplicity, we suppress the argument for h hereafter so that

$$(h,\beta) = (h(\cdot,\beta),\beta), \quad (h_0,\beta) = (h_0(\cdot,\beta),\beta), (h_0,\beta_0) = (h_0(\cdot,\beta_0),\beta_0).$$

To conserve space, we combine the $\sqrt{n}$ consistency and asymptotic normality into one step, although the proof has the usual two parts.

Our assumption involves generalizing the concept of derivatives from the standard finite-dimensional case to the infinite-dimensional one. Let $Q(h): \mathcal{H} \to \mathbb{R}^{d_h}$ be a functional taking any given $h \in \mathcal{H}$ into a Euclidean vector. For example, $Q(h) = \int h(x)dx$.

Definition 6.5.1 *We say that $Q(\cdot)$ is pathwise differentiable at $h \in \mathcal{H}$ if there exists a linear and continuous functional $\dot{Q}(\cdot)$: $\mathcal{H} \to \mathbb{R}^{d_h}$ such that*

$$\dot{Q}_h(g) = \lim_{t\to 0} \frac{Q(h+tg) - Q(h)}{t}$$

for all $g \in \mathcal{H}$.

One normally refers to $\dot{Q}_h(\cdot)$ as the pathwise derivative of $Q(h)$ at h. In the finite-dimensional case, if Q is differentiable with derivative $\partial Q(h)/\partial h$, then $\dot{Q}_h(g)$ is the differential of Q:

$$\dot{Q}_h(g) = \frac{\partial Q(h)}{\partial h} g$$

We can normally carry over results from the finite-dimensional case when deriving the pathwise derivative. In particular, the chain-rule is still valid.

Example 1. $\mathcal{H} = \left\{h : \int |h(x)|\, dx < \infty\right\}$ and $Q(h) = \int h(x)dx$. In this case,

$$Q(h + tg) - Q(h) = \int (h(x) + tg(x))\, dx - \int h(x)dx = t \int g(x)$$

so

$$\dot{Q}_h(g) = \int g(x).$$

Example 2. $\mathcal{H} = \left\{h : \frac{\partial h(x)}{\partial x}|_{x=x_0} \text{ exists}\right\}$ and $Q(h) = \frac{\partial h(x)}{\partial x}|_{x=x_0}$. In this case

$$\begin{aligned} Q(h + tg) - Q(h) &= \frac{\partial [h(x) + tg(x)]}{\partial x}|_{x=x_0} - \frac{\partial [h(x)]}{\partial x}|_{x=x_0} \\ &= t\frac{\partial g(x)}{\partial x}|_{x=x_0}. \end{aligned}$$

so

$$\dot{Q}_h(g) = \frac{\partial g(x)}{\partial x}|_{x=x_0}.$$

The notion of derivative we used here is the Gateaux derivative. For more discussion of functional derivatives including other definitions of functional derivatives, see sec 20.2 in van de Vaart (1998). See also Gill (1989) on the use of functional derivatives in nonparametric and semiparametric ML estimators.

Theorem 6.5.2 *Assume* $Q(h_0, \beta_0) = 0, \beta_0 \in int(\mathcal{B})$ *and*

(i) (Definition of β_n*)*

$$Q_n(h_n, \beta_n) \leq \inf_{\beta \in \mathcal{B}} \|Q_n(h_n, \beta)\| + o_p\left(1/\sqrt{n}\right)$$

(ii) $D(h_0, \beta) = \partial Q(h_0(\cdot, \beta), \beta)/\partial\beta$ *exists in a neighborhood of* β_0*, is of full column rank, and is continuous at* $\beta = \beta_0$*.*

(iii) The pathwise derivative (functional) $\Delta_{h_0,\beta}(\cdot)$ *exists in all directions (i.e. for all* $h - h_0$*), and for all* (h, β) *satisfying* $\|h - h_0\|_{\mathcal{H}} = o(1)$ *and* $\|\beta - \beta_0\| = o(1)$*, it is true that*

(a)

$$\|Q(h, \beta) - Q(h_0, \beta) - \Delta_{h_0,\beta}[h(\cdot, \beta) - h_0(\cdot, \beta)]\| \leq C\|h - h_0\|^2_{\mathcal{H}}$$

(b)

$$\left\|\Delta_{h_0,\beta}[h(\cdot, \beta) - h_0(\cdot, \beta)] - \Delta_{h_0,\beta_0}[h(\cdot, \beta_0) - h_0(\cdot, \beta_0)]\right\| = o(\|\beta - \beta_0\|)$$

(iv) With probability approaching 1, $h_n \in \mathcal{H}$ *and* $\|h_n - h_0\|_{\mathcal{H}} = o_p\left(n^{-1/4}\right)$ *uniformly over* β *with* $\|\beta - \beta_0\| = o(1)$*.*

(v) For all δ_n such that $\delta_n = o_p(1)$

$$\sup_{\|\beta-\beta_0\|\leq\delta_n, \|h-h_0\|_{\mathcal{H}}\leq\delta_n} \|Q_n(h,\beta) - Q(h,\beta) - [Q_n(h_0,\beta_0) - Q(h_0,\beta_0)]\| = o_p\left(1/\sqrt{n}\right).$$

(vi) for some finite matrix V

$$\sqrt{n}\left[Q_n\left(h_0,\beta_0\right) + \Delta_{h_0,\beta_0}\left(h_n(\cdot,\beta_0) - h_0\left(\cdot,\beta_0\right)\right)\right] \to_d N(0,V).$$

Then

$$\sqrt{n}\left(\beta_n - \beta_0\right) \to_d N(0,\Omega)$$

where

$$\Omega = \left(D'D\right)^{-1}\left(D'VD\right)\left(D'D\right)^{-1}$$

and

$$D = D\left(h_0,\beta_0\right) = \partial Q(h_0,\beta)/\partial\beta|_{\beta=\beta_0}.$$

Assumption (iii) on pathwise derivatives is a "high level" assumption that we have to verify in different cases. Its usually not hard to do. For example, take the case where

$$m(z,h,\beta) = \{v_1\beta + e \leq h\left(v_2\right)\}$$

which is discontinuous in both β and in h. Letting $F(\cdot)$ be the CDF of e, we have

$$Q(h,\beta) = E\left\{F\left[h\left(v_2\right) - v_1\beta\right]\right\}$$

So

$$Q(h,\beta) - Q(h_0,\beta) = E\left\{F\left[h\left(v_2\right) - v_1\beta\right] - F\left[h_0\left(v_2\right) - v_1\beta\right]\right\}$$

Assuming that $F(\cdot)$ is smooth with $F'(\cdot) = f(\cdot)$, we get

$$\begin{aligned} Q(h,\beta) - Q(h_0,\beta) \;=\; & Ef\left(h_0\left(v_2\right) - v_1\beta\right)\left[h\left(v_2\right) - h_0\left(v_2\right)\right] \\ & + \frac{1}{2}Ef'(\cdot)\left[h\left(v_2\right) - h_0\left(v_2\right)\right]^2 \end{aligned}$$

where $\cdot$ represents a point between $h\left(v_2\right) - v_1\beta$ and $h_0\left(v_2\right) - v_1\beta$. Now let

$$\Delta_{h,\beta}\left(h - h_0\right) = Ef\left(h_0\left(v_2\right) - v_1\beta\right)\left[h\left(v_2\right) - h_0\left(v_2\right)\right].$$

That is, $\Delta_{h,\beta}$ is a "functional" of $h - h_0$. Assume that $f(\cdot)$ has a derivative that is bounded in absolute value by $2C$. Then

$$\begin{aligned} \|Q(h,\beta) - Q(h_0,\beta) - \Delta_{h,\beta}\left(h - h_0\right)\| \;&\leq\; CE\left[h\left(v_2\right) - h_0\left(v_2\right)\right]^2 \\ &=\; C\left\|h - h_0\right\|_{\mathcal{H}}^2 \end{aligned}$$

as required.

Proof of the Theorem.

We begin with the$\sqrt{n}$-consistency step. As in the parametric case the fact that the limit function is differentiable in β at $\beta = \beta_0$ and $D = D(h_0, \beta_0)$ is of full column rank implies that

$$Q(h_0, \beta_n) - Q(h_0, \beta_0) \geq C \|\beta_n - \beta_0\|,$$

so it suffices to show that $\|Q(h_0, \beta_n) - Q(h_0, \beta_0)\| = O_p(1/\sqrt{n})$. But as in the proof of consistency, we have

$$\begin{aligned} \|Q(h_0, \beta_n) - Q(h_0, \beta_0)\| &\leq \|Q(h_0, \beta_n) - Q(h_n, \beta_n)\| \\ &\quad + \|Q(h_n, \beta_n) - Q_n(h_n, \beta_n)\| \\ &\quad + \|Q_n(h_n, \beta_n) - Q(h_0, \beta_0)\| \\ &:= I_1 + I_2 + I_3 \end{aligned}$$

First term: $I_1 = o(\|\beta_n - \beta_0\|) + O_p(1/\sqrt{n})$

Using the triangle inequality, we have

$$\begin{aligned} & \|Q(h_0, \beta_n) - Q(h_n, \beta_n)\| \\ \leq\ & \left\|Q(h_0, \beta_n) - Q(h_n, \beta_n) - \Delta_{h_0, \beta_n}[h_0(\cdot, \beta_n) - h_n(\cdot, \beta_n)]\right\| \\ & + \left\|\Delta_{h_0, \beta_n}[h_0(\cdot, \beta_n) - h_n(\cdot, \beta_n)] - \Delta_{h_0, \beta_0}[h_0(\cdot, \beta_0) - h_n(\cdot, \beta_0)]\right\| \\ & + \left\|\Delta_{h_0, \beta_0}[h_0(\cdot, \beta_0) - h_n(\cdot, \beta_0)]\right\| \\ \leq\ & C\|(h_0 - h_n)\|_{\mathcal{H}}^2 + o(\|\beta_n - \beta_0\|) + O_p\left(1/\sqrt{n}\right) \\ \leq\ & o(\|\beta_n - \beta_0\|) + O_p\left(1/\sqrt{n}\right) \end{aligned} \tag{6.14}$$

from Assumptions (iii)(a), (iii)(b), (vi) and (v). Here we have used the fact that since both β_n and h_n are consistent we can choose a sequence $\{\delta_n\}$ such that

$$P(\|\beta_n - \beta_0\| \geq \delta_n \text{ or } \|h_n - h_0\|_{\mathcal{H}} \geq \delta_n) \to 0$$

as $n \to \infty$. That is, we can focus the δ_n neighborhood and ignore the other possibilities.

Second term: $I_2 = O_p(1/\sqrt{n})$

Using the triangle inequality on the second term, we have

$$\begin{aligned} & \|Q(h_n, \beta_n) - Q_n(h_n, \beta_n)\| \\ \leq\ & \|Q(h_n, \beta_n) - Q_n(h_n, \beta_n) - [Q(h_0, \beta_0) - Q_n(h_0, \beta_0)]\| \\ & + \|[Q(h_0, \beta_0) - Q_n(h_0, \beta_0)]\| \\ =\ & o_p\left(1/\sqrt{n}\right) + \|Q_n(h_0, \beta_0)\| \\ =\ & O_p\left(1/\sqrt{n}\right) \end{aligned} \tag{6.15}$$

from Assumptions (v) and (vi).

Third term: $I_3 = O_p\left(1/\sqrt{n}\right)$

Note that

$$\begin{aligned}\|Q_n(h_n,\beta_n)\| &= \inf_{\beta\in\mathcal{B}}\|Q_n(h_n,\beta)\| + o_p\left(1/\sqrt{n}\right)\\ &\leq \|Q_n(h_n(\cdot,\beta_0),\beta_0)\| + o_p\left(1/\sqrt{n}\right)\end{aligned}$$

and

$$\begin{aligned}\|Q_n(h_n(\cdot,\beta_0),\beta_0)\| &\leq \|Q_n(h_n,\beta_0) - Q(h_n,\beta_0)\|\\ &\quad + \|Q(h_n,\beta_0) - Q(h_0,\beta_0)\|\\ &\leq \|Q_n(h_0,\beta_0) - Q(h_0,\beta_0)\| + o_p\left(1/\sqrt{n}\right)\\ &\quad + \left\|Q(h_n,\beta_0) - Q(h_0,\beta_0) - \Delta_{h_0,\beta_0}(h_0(\cdot,\beta_0) - h_n(\cdot,\beta_0))\right\|\\ &\quad + \left\|\Delta_{h_0,\beta_0}(h_0(\cdot,\beta_0) - h_n(\cdot,\beta_0))\right\| + o_p\left(1/\sqrt{n}\right)\\ &= O_p\left(1/\sqrt{n}\right)\end{aligned}$$

from assumptions (v), (ii), and (vi).

Combining the above analyses yields:

$$C\|\beta_n - \beta_0\| \leq O_p\left(1/\sqrt{n}\right) + o\left(\|\beta_n - \beta_0\|\right),$$

which implies that

$$\|\beta_n - \beta_0\| = O_p\left(1/\sqrt{n}\right).$$

Define

$$L_n(h_n,\beta) = D(\beta - \beta_0) + Q_n^*(h_0,\beta_0)$$

where

$$Q_n^*(h_0,\beta_0) = \left[Q_n(h_0,\beta_0) + \Delta_{h_0,\beta_0}(h_n - h_0)\right].$$

We can now proceed as in the case with a smooth objective function. Define

$$B_n = \left\{\beta : \|\beta - \beta_0\| \leq \frac{M}{\sqrt{n}}\right\}$$

for some large enough M. Given the $\sqrt{n}$ consistency of $\hat{\beta}_n$, we have $\hat{\beta}_n \in B_n$ with probability

approaching one. So we can focus on $\beta \in B_n$. Now uniformly over any $\beta \in B_n$, we have

$$\begin{aligned} Q_n(h_n,\beta) &= Q_n(h_n,\beta) - Q(h_n,\beta) - (Q_n(h_0,\beta_0) - Q(h_0,\beta_0)) \\ &\quad + Q(h_n,\beta) + (Q_n(h_0,\beta_0) - Q(h_0,\beta_0)) \\ &= Q(h_n,\beta) + Q_n(h_0,\beta_0) + o_p\left(1/\sqrt{n}\right) \quad \text{by (v)} \\ &= Q(h_0,\beta) + \Delta_{h_0,\beta}(h_n - h_0) + Q_n(h_0,\beta_0) + o_p\left(1/\sqrt{n}\right) \text{ by (iv)} \\ &= Q(h_0,\beta) + \Delta_{h_0,\beta_0}(h_n - h_0) \\ &\quad + Q_n(h_0,\beta_0) + o_p\left(1/\sqrt{n}\right) \qquad \text{by } \sqrt{n}\text{-consistency} \\ &= D(\beta_0)(\beta - \beta_0) + \Delta_{h_0,\beta_0}(h_n - h_0) + Q_n(h_0,\beta_0) + o_p\left(1/\sqrt{n}\right) \end{aligned}$$

That is, uniformly over any $\beta \in B_n$,

$$\|Q_n(h_n,\beta) - L_n(h_n,\beta)\| = o_p\left(1/\sqrt{n}\right).$$

This implies

$$\beta_n = \beta_n^* + o_p\left(1/\sqrt{n}\right)$$

where

$$\beta_n^* = \arg\min \|L_n(h_n,\beta)\| + o_p\left(1/\sqrt{n}\right).$$

As a result

$$\begin{aligned} \sqrt{n}(\beta_n - \beta_0) &= \sqrt{n}(\beta_n^* - \beta_0) + o_p(1) \\ &= \left(D'D\right)^{-1} D'Q_n^* + o_p(1) \to^d N(0,\Omega). \end{aligned}$$

6.5.3 Example: partial linear median regression

6.6 Bibliographical Remarks

Andrews (1994), Chen, Linton and van Keilegom (2004), Newey and McFadden (1994), Newey (1994), Pakes and Olley (1995) all give similar results for consistency and asymptotic normality for two-step semiparametric estimators; see also Ai (1997). For further reading on functional derivatives, see e.g. Flett (2008).

6.7 Problems

1. Consider the standard nonlinear regression model:

$$Y = f(X,\beta_0) + \varepsilon, E[\varepsilon|X] = 0,$$

where the errors are heteroskedastic:

$$E\left(\varepsilon^2|X\right) = h_0(X)$$

The standard NLS estimator

$$\hat{\beta}_{NLS} = \arg\min \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - f(X_i,\beta)\right]^2$$

is consistent and asymptotically normally distributed, but not asymptotically efficient. If the conditional variance function $h_0(X)$ is known, then we can do weighted least squares (WLS),

$$\tilde{\beta}_{WLS} = \arg\min \frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i - f(X_i,\beta)\right]^2}{h_0(X_i)} \tag{6.16}$$

which is more efficient than $\hat{\beta}_{NLS}$. The first order condition for this problem is

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial f(X_i,\beta)}{\partial\beta}\frac{\left[Y_i - f(X_i,\beta)\right]}{h_0(X_i)} := \frac{1}{n}\sum_{i=1}^{n} m(Z_i, h_0(X_i), \beta)$$

So the population moment condition is

$$Q(\beta) = Em(Z_i, h_0(X_i), \beta_0) = 0.$$

If we do not know $h_0(X)$, we can estimate it in a preliminary stage, form $m(Z_i, h_n(X_i), \beta)$ and "approximate" sample analogue to the population moment condition, and then choose β to minimize a norm of this sample analogue. Note that there are at least two ways of doing this, and they have a somewhat different structure. In the first way, we follow the steps below:

(**a**) Obtain $\hat{\beta}_{NLS}$

(**b**) Calculate the associated residuals, $\hat{\varepsilon}_i = Y_i - X_i\hat{\beta}_{NLS}$, $i = 1, 2, ..., n$.

(**c**) Estimate the conditional variance nonparametrically, e.g.

$$h_n\left(X_i, \hat{\beta}_{NLS}\right) := \frac{\sum_{j=1}^{n} K_b(X_j - X_i)\hat{\varepsilon}_j^2}{\sum_{j=1}^{n} K_b(X_j - X_i)}$$

where b is the bandwidth parameter.

(**d**) Obtain $\hat{\beta}_{WLS}$ by solving

$$\hat{\beta}_{WLS} = \arg\min \left\| \frac{1}{n} \sum_{i=1}^{n} m \left[Z_i, h_n \left(X_i, \hat{\beta}_{NLS} \right), \beta \right] \right\|$$

Alternatively, we could recompute the weighting function for different values of β, in which case we define

$$\check{\beta}_{WLS} = \arg\min \left\| \frac{1}{n} \sum_{i=1}^{n} m(Z_i, h_n \left(X_i, \beta \right), \beta) \right\|$$

Assuming all the regularity assumptions hold, find and compare the asymptotic distributions of $\hat{\beta}_{WLS}$ and $\check{\beta}_{WLS}$.

2. Consider the partial median regression with endogenous X_1

$$Y = X_1 \beta_0 + h_* \left(X_2 \right) + \varepsilon$$

where

$$P(\varepsilon < 0 | X_2, X_3) = 0.5$$

and X_3 is an instrument for X_1. Assuming that all conditions in theorem 6.5.2 hold (or make appropriate assumptions and verify all conditions there), derive the asymptotic distribution of the two-step estimator β_n :

$$\beta_n = \arg\min \left\| \frac{1}{n} \sum m \left(Z_i, h_n \left(X_{2i}, \beta \right), \beta \right) \right\|$$

where $Z_i = \left(X_{1i}, X_{2i}, X_{3i}, Y_i \right)$, $h_n \left(X_2, \beta \right)$ is a consistent estimator of $h_0 \left(X_2, \beta \right) := median \left(Y - X_1 \beta | X_2 \right)$ and

$$m \left(Z_i, h_n \left(X_{2i}, \beta \right), \beta \right) = X_{3i} \left[1 \left\{ Y_i - X_{1i} \beta \leq h_n \left(X_{2i}, \beta \right) \right\} - 0.5 \right].$$

3. Consider the partial linear model

$$\begin{aligned} Y_i &= \alpha + \beta X_i + h(Z_i) + \varepsilon_i \\ h(Z_i) &= 0.5 \delta_1 Z_i^2 \\ X_i &= \frac{1}{\ln(\delta_2 + 1)} \exp \left(Z_i \ln \left(\delta_2 + 1 \right) \right) + \eta_i \end{aligned}$$

where (ε_i, η_i) are iid Gaussian with unit covariance matrix and $Z_i = i/n$. Let $(\delta_1, \delta_2) = (-20, 4)$. $(\alpha, \beta) = (1, 1)$. Generate 1000 samples of size 100 from the model.

Part A

(i) For each sample and each smoothing parameter (i.e. the bandwidth parameter), estimate (α, β) using Robinson's approach with unknown conditional means estimated by local linear regression with quadratic kernel. Construct the 95% confidence interval for β :

$$CI_{\hat{\beta}} = \left[\hat{\beta} - 1.96\hat{\sigma}_{\beta}, \hat{\beta} + 1.96\hat{\sigma}_{\beta}\right]$$

where $\hat{\sigma}_{\beta}^2$ is an estimator of the se of $\hat{\beta}$.

(ii) Graph the mse of $\hat{\beta}$ against the bandwidth used in the local linear regressions. Discuss whether there is an opportunity for optimal bandwidth that minimizes the mse, that is, inspect whether the mse curve is U-shaped.

(iii) Graph the absolute coverage error of $CI_{\hat{\beta}}$ against the bandwidth used used in the local linear regressions. Discuss whether there is an opportunity for optimal bandwidth that minimizes the absolute coverage error (ace).

(iv) Since we do not know the true DGP in practice, the optimal bandwidth in (ii) and (iii) if exists is not feasible. To come up with a feasible version, we can fit an approximating model to the sample data:

$$Y_i = \alpha + \beta X_i + \sum_{j=1}^{J_{pilot}} \gamma_j \phi_j (Z_i) + \varepsilon_i, \ J_{pilot} = 5 \tag{6.17}$$

where $\phi_j (Z_i)$ are Hermit polynomials. We then regard the approximating model as if it the true model. More specifically, using the approximating model, we can generate $\hat{Y}_i$ according to

$$\hat{Y}_i = \alpha^* + \beta^* X_i + \sum_{j=1}^{J_{pilot}} \gamma_j^* \phi_j (Z_i) + \varepsilon_i^*$$

where $\alpha^*, \beta^*, \gamma^*$ are estimates from (6.17) and $\{\varepsilon_i^*\}$ are iid draws from the estimated residuals $\{\hat{\varepsilon}_i\}$. Repeat (i), (ii) (iii) on 1000 samples of the form $\left\{X_i, Z_i, \hat{Y}_i, i = 1, 2, ..., 100\right\}$. Find the optimal bandwidth based on both the mse and ace criteria and compare it with the corresponding infeasible bandwidth if it exists.

Part B.

Carry out all steps in part A but now use the method of sieves to estimate (α, β). That is, use a Hermit polynomials to approximate $h(Z)$ and regress Y on constant, X and the polynomial bases. Graph the mse of $\hat{\beta}$ and the absolute coverage error of the 95% confidence interval against J, the number of terms in the sieve approximation. Find the optimal J if it exists. Finally, use the procedure in B(iv) to design a feasible rule for the choice of J and discuss the performance of this rule for the smoothing parameter choice.

6.8 References

1. Ai, C. (1997), "A Semiparametric Maximum Likelihood Estimator." *Econometrica* 65, 933-963.

2. Andrews, D.W.K. (1994): "Empirical Process Methods in Econometrics." In *Handbook of Econometrics*, Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2246-2294. Elsevier.

3. Chen, X. (2007): Semiparametric and Nonparametric Estimation via the Method of Sieves. *Handbook of Econometrics*, Vol. 6B (eds. J. Heckman and E. Leamer) 5549-5632. Elsevier.

4. Chen, X., O. Linton & I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function is not Smooth." *Econometrica* 71, 1591-1608.

5. Flett, T.M. (2008), Differential Analysis: Differentiation, Differential Equations and Differential Inequalities. Cambridge University Press.

6. Gill, R. (1989): "Non- and Semi-Parametric Maximum Likelihood Estimators and the Von Mises Method (Part 1)." *Scandinavian Journal of Statistics*, Vol. 16(2), pp. 97-128.

7. Li, Q. and Racine J. (2007), *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

8. Newey, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators." *Econometrica* 62, 1349–1362.

9. Newey, W. K., and D. L. McFadden (1999): "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*. Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2113-2245. Elsevier.

10. Pakes and Olley (1995): "A limit theorem for a smooth class of semiparametric estimators." *Journal of Econometrics*, 65(1), 295-332.

11. van de Vaart, A.W. (1998), *Asymptotic Statistics*. Cambridge University Press.

12. Wooldridge J. (2002), *Econometrics of Cross Section and Panel Data*. The MIT press.

Chapter 7

The General Sieve Extremum Estimation

7.1 Introduction [1]

In this chapter, we consider general sieve extreme estimators or M estimator that include the estimators in Chapter 3 as special cases. We focus on sieve estimators with finite dimensional sieves. The main departure is that the sieve extreme estimator may not have a closed form expression. This is in contrast to the smoothing spline and series estimators under the quadratic loss function, in which case the estimators can be represented in closed forms.

7.1.1 Basic Setting

Suppose we are interested in estimating some unknown function $\theta_0 : \mathbb{R}^d \to \mathbb{R}$. This could for example be a regression function or a density as considered in Chapter 2. Very often, this function can be characterized as the unique minimizer of a population criterion function $Q(\theta)$: $Q(\theta) > Q(\theta_0)$ for $\theta \neq \theta_0$ in some function space Θ (equipped with a suitable norm, say $d(\theta_1, \theta_2)$). The choice of $Q(\cdot)$ and the existence of θ_0 are suggested by the identification of an econometric model. The (pseudo-) true parameter $\theta_0 \in \Theta$ is unknown but is related to a joint probability measure $P(z_1, \cdots, z_n)$, from which a sample of size n observations $\{Z_i\}_{i=1}^n$, $Z_i \in \mathbb{R}^{d_z}$, $1 \leq d_z < \infty$, is available. Let $Q_n : \Theta \to \mathbb{R}$ be an empirical criterion, which is a measurable function of the data $\{Z_i\}_{i=1}^n$ for all $\theta \in \Theta$, and converges to Q in some sense as the sample size $n \to \infty$. One general way to estimate θ_0 is by minimizing Q_n over Θ;

[1] This chapter is based on Chen (2007) Handbook of Econometrics Chapter on sieve estimation. I have borrowed some sections directly from Chen (2007), as we planned to work a book project at one point. All errors are my own.

the minimizer[2], $\arg\min_{\theta\in\Theta} Q_n(\theta)$, assuming it exists, is then called the *extremum* estimator. Note that we use a notation different from the chapter on semiparametric two step estimation. There the population objective function is $\|Q(\theta)\|_E^2$ and here the objective function is $Q(\theta)$ itself.

7.1.2 Sieve Extremum Estimator

As we discussed in Chapter 3, when the space Θ is overly complex and rich, $\arg\min_{\theta\in\Theta} Q_n(\theta)$ may not provide a consistent estimator of θ_0. The method of sieves provides a general approach to resolve the difficulty by minimizing Q_n over a sequence of approximating spaces Θ_{k_n}, called *sieves* by Grenander (1981), which are less complex but are dense in Θ. Popular sieves are typically compact, non-decreasing ($\Theta_{k_n} \subseteq \Theta_{k_{n+1}} \subseteq \cdots \subseteq \Theta$). They are dense in Θ in that for any $\theta \in \Theta$ there exists an element $\pi_{k_n}\theta$ in Θ_{k_n} satisfying $d(\theta, \pi_{k_n}\theta) \to 0$ as $n \to \infty$, where $d(\cdot,\cdot)$ is a (pseudo) metric and the notation π_{k_n} can be regarded as a projection mapping from Θ to Θ_{k_n}. Very often Θ_{k_n} is a finite dimensional linear sieve space so that

$$\Theta_{k_n} = span\{\varphi_0, ..., \varphi_{k_n}\}$$

and $\theta_0 = \sum_{k=0}^{\infty} \gamma_k \varphi_k$.

An *approximate sieve extremum estimator*, denoted by $\hat{\theta}_n$, is defined as an approximate minimizer of $Q_n(\theta)$ over the sieve space Θ_{k_n}, i.e.,

$$Q_n(\hat{\theta}_n) \leq \inf_{\theta\in\Theta_{k_n}} Q_n(\theta) + O_p(\eta_{k_n}), \quad \text{with } \eta_{k_n} \to 0 \text{ as } n \to \infty. \tag{7.1}$$

When $\eta_{n_k} = 0$, we call $\hat{\theta}_n$ in (7.1) the *exact* sieve extremum estimator. The sieve extremum estimation method clearly includes the standard extremum estimation method by setting $\Theta_{k_n} = \Theta$ for all n. Sometimes, the method of sieves is not needed if the function space is not too rich, even though we consider semi-nonparametric estimation.

For a semi-nonparametric econometric model, $\theta_0 \in \Theta$ can be decomposed into two parts $\theta_0 = (\beta_0', h_0')' \in B \times \mathcal{H}$, where B denotes a finite dimensional compact parameter space, and $\mathcal{H}$ an infinite dimensional parameter space. In this case, a natural sieve space will be $\Theta_{k_n} = B \times \mathcal{H}_{k_n}$ with $\mathcal{H}_{k_n}$ being a sieve for $\mathcal{H}$, and the resulting estimate $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ in (7.1) will sometimes be called a simultaneous (or joint) sieve extremum estimator.

7.1.3 Sieve M-estimator

When $Q_n(\theta)$ can be expressed as a sample average of the form

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i), \tag{7.2}$$

[2] If the space Θ is not compact, we should write $\arg\inf_{\theta\in\Theta}$ instead of $\arg\min_{\theta\Theta}$. Here I ignore this subtle difference.

with $l : \Theta \times \mathbb{R}^{d_z} \to \mathbb{R}$ being the criterion based on a single observation, we also call the $\hat{\theta}_n$ solving (7.1) as an *approximate sieve M-estimator*. This includes sieve maximum likelihood estimation (MLE), sieve least squares (LS), sieve generalized least squares (GLS) and sieve quantile regression as special cases.

We reserve the term "sieve extremum estimator" for sieve estimators that minimize a general criterion function, i.e., a criterion function that can not be written as the sample average of certain functions as in (7.2). The consistency result in the next section applies to general sieve extremum estimators while the rate of convergence result below applies only to sieve M-estimators.

7.2 Consistency of Sieve Extremum Estimators

In this section, we shall establish that, under mild regularity assumptions, the sieve extremum estimation will consistently estimate both finite-dimensional and infinite-dimensional unknown parameters.

7.2.1 Assumptions and Consistency Theorem

Let $d(\cdot, \cdot)$ be a (pseudo) metric on Θ. In particular, when $\Theta = B \times \mathcal{H}$ where B is a subset of some Euclidean space and $\mathcal{H}$ is a subset of some normed function space, we can use $d(\theta, \widetilde{\theta}) = |\beta - \widetilde{\beta}|_e + \|h - \widetilde{h}\|_{\mathcal{H}}$, where $|\cdot|_e$ denotes the Euclidean norm, and $\|\cdot\|_{\mathcal{H}}$ is a norm imposed on the function space $\mathcal{H}$. For example, if $\mathcal{H} = C^m(\mathcal{X})$ with a bounded $\mathcal{X}$, we could take $\|h\|_{\mathcal{H}}$ to be $\|h\|_\infty$ or $\|h\|_2$. The following assumptions are needed to ensure that $d(\hat{\theta}_n, \theta_0) \to 0$ as $n \to \infty$.

Assumption 1. (Definition). $\hat{\theta}_n \in \Theta_{k_n}$ and $Q_n(\hat{\theta}_n) \leq \inf_{\theta \in \Theta_{k_n}} Q_n(\theta) + O_p(\eta_{k_n})$ for some $\eta_{k_n} \to 0$ as $n \to \infty$.

Assumption 2. (Identification) (i) $Q(\theta_0) < \infty$ and if $Q(\theta_0) = -\infty$, $Q(\theta) > -\infty$ for all $\theta \in \Theta_{k_n} \setminus \{\theta_0\}$ and all $k_n \geq 1$. (ii) There exist a non-increasing positive function $\delta(\cdot)$ and positive function $g(\cdot)$ such that $\forall \varepsilon > 0$, $\inf_{\theta \in \Theta_{k_n} \setminus B(\theta_0, \varepsilon)} Q(\theta) \geq Q(\theta_0) + \delta(k_n) g(\varepsilon)$, where by $B(\theta_0, \varepsilon)$ we denote the open ball of radius ε centered at θ_0, i.e. $B(\theta_0, \varepsilon) := \{\theta : d(\theta, \theta_0) < \varepsilon\}$.

Assumption 3. (Sieve Space) $\Theta_{k_n} \subseteq \Theta_{k_n+1} \subseteq \Theta$ for all $k_n \geq 1$; and there exists a sequence $\pi_{k_n}\theta_0 \in \Theta_{k_n}$ such that $d(\pi_{k_n}\theta_0, \theta_0) \to 0$ as $k_n \to \infty$.

Assumption 4. (Continuity) (i) For each $k_n \geq 1$, $Q(\theta)$ is lower semi-continuous[4] on Θ_{k_n} under the metric $d(\cdot, \cdot)$; (ii) $\|Q(\theta_0) - Q(\pi_{k_n}\theta_0)\| = o(\delta(k_n))$.

Assumption 5. (Compact Sieve Space) The sieve spaces Θ_{k_n} are compact under the metric $d(\cdot, \cdot)$.

[4] We say that $Q(\theta)$ is lower semi-continuous at θ_o if for every $\varepsilon > 0$, there exists a neighborhood N of θ_o such that $Q(\theta) \geq Q(\theta_o) - \varepsilon$ for all θ in N.

Assumption 6. (Uniform Convergence over Sieves). Let

$$c\left(k_n\right) = \sup_{\theta \in \Theta_{k_n}} \left|Q_n\left(\theta\right) - Q\left(\theta\right)\right|,$$

then as $n \to \infty$ and $k_n \to \infty$ (i) $c(k_n) = o_p\left(1\right)$ (ii) $c(k_n) = o_p\left(\delta\left(k_n\right)\right)$ (iii) $\eta_{k_n} = o\left(\delta\left(k_n\right)\right)$.

Theorem 7.2.1 *Let assumptions 1-6 hold, then* $d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0$ *as* $k_n \to \infty$ *and* $n \to \infty$.

Proof: By assumptions 4.i and 5, $\inf_{\theta \in \Theta_{k_n} \backslash B(\theta_0, \varepsilon)} Q(\theta)$ exists. Let $\varepsilon > 0$. By assumption 2, whenever $\theta \in \Theta_{k_n} \backslash B(\theta_0, \varepsilon)$, we have

$$Q(\theta) - Q(\theta_0) \geq \delta(k_n) g(\varepsilon). \tag{7.3}$$

Thus

$$\begin{aligned}
P(d(\hat{\theta}_n, \theta_0) \;>\; & \varepsilon) = P(\hat{\theta}_n \in \Theta_{k_n} \backslash B(\theta_0, \varepsilon)) \\
\leq\; & P\left(Q(\hat{\theta}_n) - Q(\theta_0) \geq \delta(k_n) g(\varepsilon)\right) \\
\leq\; & P(Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\hat{\theta}_n) - Q(\theta_0) \geq \delta(k_n) g(\varepsilon)) \\
\leq\; & P(Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + \underbrace{\inf_{\theta \in \Theta_{k_n}} Q_n(\theta) + O_p(\eta_{k_n})}_{Definition} - Q(\theta_0) \geq \delta(k_n) g(\varepsilon)) \\
\leq\; & P(Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\pi_{k_n}\theta_0) + O_p(\eta_{k_n}) - Q(\theta_0) \geq \delta(k_n) g(\varepsilon)) \\
\leq\; & P(Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\pi_{k_n}\theta_0) - Q\left(\pi_{k_n}\theta_0\right) \\
& + O_p(\eta_{k_n}) + Q\left(\pi_{k_n}\theta_0\right) - Q(\theta_0) \geq \delta(k_n) g(\varepsilon)) \\
\leq\; & P\left(2 \sup_{\theta \in \Theta_{k_n}} \left\|Q_n(\theta) - Q\left(\theta\right)\right\| + O_p(\eta_{k_n}) + o\left(\delta\left(k_n\right)\right) \geq \delta(k_n) g(\varepsilon))\right) \\
\leq\; & P\left(o_p\left(\delta\left(k_n\right)\right) + O_p(\eta_n) + o\left(\delta\left(k_n\right)\right) \geq \delta(k_n) g(\varepsilon))\right) \\
\to\; & 0.
\end{aligned}$$

The first inequality follows by (7.3), the second inequality by the definition of the extremum estimator, and the last equality by uniform convergence. □

Remarks

1. Theorem 7.2.1 is an extension of the same theorem in the fully parametric case. The identification assumption is more refined than that in the parametric case which typically entails the following: $\forall \varepsilon > 0, \exists \delta\left(\varepsilon\right) > 0$ such that

$$\inf_{\left\|\theta - \theta_0\right\| > \varepsilon} Q(\theta) \geq Q(\theta_0) + \delta\left(\varepsilon\right).$$

For the sieve estimation, we allow $\delta(\varepsilon)$ to depend on k_n, the "size" of the sieve space. This is necessary because for some problems $\inf_{\theta\in\Theta_{k_n}\setminus B(\theta_0,\varepsilon)} Q(\theta)$ may become closer and closer to $Q(\theta_0)$ as the sieve space increases.

2. As an example, consider the nonparametric IV regression

$$Y_1 = h_0(Y_2) + u$$

where Y_2 is endogenous and we have an IV variable X such that $E(u|X) = 0$. In this case, $E[Y_1 - h_0(Y_2)|X] = 0$ for almost all X and we can define

$$Q(h) = E\left(E[Y_1 - h(Y_2)|X]\right)^2 = E\left(E[h(Y_2) - h_0(Y_2)|X]\right)^2.$$

Suppose $d(h, h_0) = E\left\{[h(Y_2) - h_0(Y_2)]^2\right\}$. Then it is possible to find that $h_n(Y_2) \notin B(h_0, \varepsilon)$ for some $\varepsilon > 0$,i.e., $E\left\{[h_n(Y_2) - h_0(Y_2)]^2\right\} \geq \varepsilon^2$, such that $E\left([h_n(Y_2) - h_0(Y_2)]\,|X\right) \to 0$. That is to say, there exists $h_n(Y_2) \notin B(h_0, \varepsilon)$ but $Q(h_n) \to Q(h_0)$ as n increases. The identification assumption says that the rate of $Q(h_n)$ approaching $Q(h_0)$ has to be bounded below by $\delta(k_n)$.

3. If $\liminf_{k_n} \delta(k_n) > 0$, then assumption 6(iii) is automatically satisfied with $\eta_{k_n} = o(1)$, Assumption 6(ii) is implied by Assumption 6(i), and Assumption 4(ii) is implied by Assumption 3 and **Assumption 4(ii)'**: $Q(\theta)$ is continuous at θ_0 in Θ.

4. Note that when $\Theta_{k_n} = \Theta$ is compact, the assumptions for Theorem 7.2.1 become the standard assumptions imposed for consistency of parametric extremum estimation in Newey and McFadden (1994). For semi-nonparametric models, the entire parameter space Θ contains infinite-dimensional unknown functions and is generally non-compact. Nevertheless, one can easily construct compact approximating parameter spaces (sieves) Θ_{k_n}. Moreover, it is relatively easy to verify the uniform convergence over compact sieve spaces, while "$\text{plim}_{n\to\infty} \sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = 0$" may fail when the space Θ is too "large" or too "complex".

5. Assumption 6 can be replaced by (i) $Q(\theta) - Q_n(\theta) \leq c(k)$ uniformly over Θ_{k_n} (ii) $Q(\pi_{k_n}\theta_0) - Q(\theta_0) \leq c(k)$.

7.2.2 Uniform Convergence and Entropy Assumption

Among the required assumptions for consistency, the uniform convergence assumption is the most difficult to verify even if we focus on the sieve space Θ_{k_n}. Numerous results are available in the probability and statistical learning literature concerning sufficient assumptions for the uniform convergence. Most of these results rely on some sort of entropy assumption that restricts the size of the set Θ_{k_n}. As a motivating example, consider a uniform convergence

result with a bounded $l : 0 \leq l \leq M$ for some M. Then by Hoeffding's inequality, for each fixed θ, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) - El(\theta, Z)\right| > \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{M^2}\right).$$

which together with the union bound implies

$$P\left\{\sup_{\theta\in\Theta_{k_n}}\left|\frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) - El(\theta, Z)\right| > \varepsilon\right\} \leq 2\#\left(\Theta_{k_n}\right)\exp\left(-\frac{2n\varepsilon^2}{M^2}\right).$$

where $\#\left(\Theta_{k_n}\right)$ denotes the number of elements in Θ_{k_n}. Thus, if Θ_{k_n} is finite, then ULLN trivially holds. However, in econometric applications, Θ_{k_n} is always an infinite set. But sometimes it is possible to choose a finite set or countable set Θ_ε such that

$$\begin{aligned}&\left\{\sup_{\theta\in\Theta_{k_n}}\left|\frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) - El(\theta, Z)\right| > \varepsilon\right\}\\ \subset\ &\left\{\sup_{\theta\in\Theta_{k_n,\varepsilon}}\left|\frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) - El(\theta, Z)\right| > \varepsilon'\right\}\end{aligned}$$

for some ε' depending on ε but not on n. In general $\Theta_{k_n,\varepsilon}$ is a cover of Θ_{k_n} and $\#\left(\Theta_{k_n}\right)$ characterizes the complexity of the set Θ_{k_n}. If $l\left(\cdot,\theta\right)$ is well behaved, for example, Lipschitz continuous, then the complexity of Θ_{k_n} is directly related to the complexity of the function class $\mathcal{F}_n = \{l(\cdot,\theta) : \theta \in \Theta_{k_n}\}$.

Let

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i}$$

be the empirical measure and P be probability distribution of Z_i. Given a measurable function f, we write $P_n f$ as the expectation of f under the empirical measure and Pf as the expectation of f under P. Thus

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \text{ and } Pf = \int f dP$$

So

$$\frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) - El(\theta, Z) = \left(P_n - P\right) l(\theta, \cdot).$$

If the ULLN holds over $\mathcal{F}_n$, then we say $\mathcal{F}_n = \{l(\theta,\cdot) : \theta \in \Theta_{k_n}\}$ is Gelivenko-Cantelli (GC). If the empirical process $\sqrt{n}\left(P_n - P\right) g$ for $g \in \mathcal{F}_n$ converges to a tight limiting process in a certain functional space, then we say $\mathcal{F}_n$ is Donsker.

Whether a class of functions is GC or Donsker depends on the size of the function class or the complexity of the class. We now review some notions of complexity formally. Let $L_r(P)$, $r \in [1, \infty)$ denote the space of real-valued random variables with finite r-th moments and $\|\cdot\|_r$ denote the $L_r(P)$-norm. More specifically

$$\|g\|_r = \left(E\left|g(Z)\right|^r\right)^{1/r}.$$

One notion of complexity of the class $\mathcal{F}_n$ is the $L_r(P)$-*covering numbers without bracketing*, which is the minimal number of w-balls $\{\{f : ||f - g_j||_r \leq w\}, ||g_j||_r < \infty, j = 1, \ldots, N\}$ that cover $\mathcal{F}_n$, denoted as $N(w, \mathcal{F}_n, ||\cdot||_r)$. Likewise, we can define $N(w, \mathcal{F}_n, ||\cdot||_{n,r})$ as the $L_r(P_n)$-(random) covering numbers without bracketing, where $\|\cdot\|_{n,r}$ denote the $L_r(P_n)$-norm and P_n denote the empirical measure of a random sample $\{Z_i\}_{i=1}^n$. That is,

$$\|g\|_{n,r} = \left(\frac{1}{n}\sum_{i=1}^{n}\left|g\left(Z_i\right)\right|^r\right)^{1/r}.$$

Sometimes the covering numbers of $\mathcal{F}_n$ can grow to infinity very fast as n grows; it is then more convenient to measure the complexity of $\mathcal{F}_n$ using the notion of $L_r(P)$-*metric entropy without bracketing*,

$$H(w, \mathcal{F}_n, ||\cdot||_r) \equiv \log(N(w, \mathcal{F}_n, ||\cdot||_r)),$$

and the $L_r(P_n)$-(random) metric entropy without bracketing,

$$H(w, \mathcal{F}_n, ||\cdot||_{n,r}) \equiv \log(N(w, \mathcal{F}_n, ||\cdot||_{n,r})).$$

Another notion of complexity is the $L_r(P)$-*covering numbers with bracketing*, which is the smallest number $N = N_{[\,]}(w, \mathcal{F}_n, \|\cdot\|_r)$ such that there exists a collection of brackets $\left\{f_1^\ell, f_1^u\right\}, ..., \left\{f_N^\ell, f_N^u\right\}$ such that

$$\sup_{1\leq i\leq N}\left\|f_i^l - f_i^u\right\|_r \leq w$$

and for any $f \in \mathcal{F}_n$, there exists an i satisfying

$$f_i^\ell \leq f \leq f_i^u, \text{ a.e.}$$

The corresponding *metric entropy with bracketing* is defined to be

$$H_{[]}(w, \mathcal{F}_n, ||\cdot||_r) \equiv \log(N_{[]}(w, \mathcal{F}_n, \|\cdot\|_r)).$$

Detailed discussions on metric entropy can be found in Pollard (1984), Andrews (1994), van der Vaart and Wellner (1996) and van de Geer (2000).

Lemma 7.2.1 *Let* $Q_n(\theta) = n^{-1}\sum_{i=1}^n l(\theta, Z_i)$ *and* $\{Z_i\}_{i=1}^n$ *is i.i.d. Suppose that*

$$H_{[\,]}(w, \mathcal{F}_n, ||\cdot||_1) < \infty.$$

for all $w > 0$*, then Assumption 6(i) holds.*

For a proof, see Lemma 3.1 of van de Geer (2002) or van de Vaart (1998, Theorem 19.4).

Lemma 7.2.2 *Let* $Q_n(\theta) = n^{-1}\sum_{i=1}^n l(\theta, Z_i)$ *and* $\{Z_i\}_{i=1}^n$ *is i.i.d. Suppose that*

(i) $E\{\sup_{\theta\in\Theta_{k_n}} |l(\theta, Z_i)|\} < \infty$,

(ii) $H(w, \mathcal{F}_n, ||\cdot||_{n,1}) = o_p(n)$ *for all* $w > 0$.

Then Assumption 6(i) holds.

For a proof, see van de Geer (2000, Lemma 3.1).

Lemma 7.2.3 *Let* $Q_n(\theta) = n^{-1}\sum_{i=1}^n l(\theta, Z_i)$ *and* $\{Z_i\}_{i=1}^n$ *is i.i.d. Suppose that*

(i) $E\{\sup_{\theta\in\Theta_{k_n}} |l(\theta, Z_i)|\}^2 < \infty$,

(ii) $H(w, \mathcal{F}_n, ||\cdot||_{n,2}) = o_p(n)$ *for all* $w > 0$.

Then Assumption 6(i) holds.

This lemma follows from Lemma 3.6 and the proof of Theorem 3.7 in van de Geer (2000).

When the function class Θ is too complex in terms of its metric entropy being too large, then the uniform convergence over the entire parameter space Θ may fail, but the uniform convergence over a sieve space Θ_{k_n} may still hold. For example, when the space Θ is infinite-dimensional and not totally bounded, i.e., it can not be covered by finitely many subsets of a fixed "size", $H(w, \{l(\theta,\cdot) : \theta \in \Theta\}, ||\cdot||_{n,1}) = O_p(n)$ may occur. Hence $\sup_{\theta\in\Theta} | |Q_n(\theta) - Q(\theta)| \neq o_p(1)$. In such a case, the extremum estimator obtained by minimizing over the entire parameter space Θ, $\arg\min_{\theta\in\Theta} Q_n(\theta)$, may fail to exist or be inconsistent.

7.2.3 Example: Consistency of Sieve M-estimators

Let

$$\hat{\theta}_n = \arg \inf_{\theta\in\Theta_{k_n}} n^{-1}\sum_{i=1}^n l(\theta, Z_i) + o_p(1).$$

Suppose that Assumptions 3 and 5 hold, that Assumption 2 is satisfied with $Q(\theta) = E\{l(\theta, Z_i)\}$ and $\liminf_{k_n} \delta(k_n) > 0$, and that $E\{l(\theta, Z_i)\}$ is continuous at $\theta = \theta_0 \in \Theta$. Then $d(\hat{\theta}_n, \theta_0) = o_p(1)$ under the following Assumption 6M:

Assumption 6M:
(i) $\{Z_i\}_{i=1}^n$ is i.i.d., $E\{\sup_{\theta\in\Theta_{k_n}} |l(\theta, Z_i)|\}$ is bounded;
(ii) there are a finite $s > 0$ and a random variable $U(Z_i)$ with $E\{U(Z_i)\} < \infty$ such that

$$\sup_{\theta,\theta'\in\Theta_{k_n}:d(\theta,\theta')\leq\delta} |l(\theta, Z_i) - l(\theta', Z_i)| \leq \delta^s U(Z_i);$$

(iii) $\log N(\delta^{1/s}, \Theta_{k_n}, d) = o(n)$ for all $\delta > 0$.

The consistency result in this example is a direct consequence of Theorem 7.2.1 and Pollard's (1984) Theorem II.24. This is because Assumption 6M (i) and (ii) imply $H(w, \{l(\theta, \cdot) : \theta \in \Theta_{k_n}\}, ||\cdot||_{n,1}) \leq \log N(\delta^{1/s}, \Theta_{k_n}, d)$, hence Assumption 6M implies Assumption 6(i).

7.2.4 Example: Consistency of Sieve MD-estimators

Let

$$\hat{\theta}_n = \arg \inf_{\theta\in\Theta_{k_n}} \frac{1}{n}\sum_{i=1}^n \widehat{m}(X_i, \theta)'\widehat{m}(X_i, \theta) + o_p(1).$$

Suppose that Assumptions 3 and 5 hold, and that

$$m(X_i, \theta) \equiv E\{\rho(Z_i, \theta)|X_i\} = 0$$

only when $\theta = \theta_0 \in \Theta$, that for almost all X_i, $m(X_i, \theta)$ is continuous in θ_0 under the metric $d(\cdot, \cdot)$, and that $\liminf_{k_n} \delta(k_n) > 0$. Then $d(\hat{\theta}_n, \theta_0) = o_p(1)$ under the following Assumption 6MD:

Assumption 6MD:
(i) $\{Z_i\}_{i=1}^n$ is i.i.d., $E\{\sup_{\theta\in\Theta_{k_n}} |m(X_i, \theta)'m(X_i, \theta)|\}$ is bounded;
(ii) there are a finite $s > 0$ and a $U(X_i)$ with $E\{[U(X_i)]^2\} < \infty$ such that

$$\sup_{\theta,\theta'\in\Theta_{k_n}:d(\theta,\theta')\leq\delta} \left|m(X_i, \theta) - m(X_i, \theta')\right| \leq \delta^s U(X_i);$$

(iii) $\log N(\delta^{1/s}, \Theta_{k_n}, d) = o(n)$ for all $\delta > 0$;
(iv) $n^{-1}\sum_{i=1}^n |\widehat{m}(X_i, \theta) - m(X_i, \theta)|^2 = o_p(1)$ uniformly over $\theta \in \Theta_{k_n}$.
See Chen and Pouzo (2009) for a proof of the above claim.

7.3 Convergence Rates of Sieve M-estimators

There are many results on convergence rates of sieve M-estimators of unknown functions. For i.i.d. data, van de Geer (1995) obtained the rate for sieve LS regression. Shen and Wong

(1994), and Birgé and Massart (1998) derived the rates for general sieve M-estimation. van de Geer (1993) and Wong and Shen (1995) obtained the rates for sieve MLE. The general theory on convergence rates is technically involved and relies on the theory of empirical processes. In this section we present a simple version of the rate results for sieve M-estimation whose assumptions are easy to verify. For the most general theory on convergence rates of sieve M-estimates, see the papers by Shen and Wong (1994), Wong and Shen (1995) and Birgé and Massart (1998).

7.3.1 Rate of Convergence Theorem

Recall $\theta_0 \in \Theta$ and that the approximate sieve M-estimator $\hat{\theta}_n$ solves:

$$n^{-1}\sum_{i=1}^{n} l(\hat{\theta}_n, Z_i) \leq \inf_{\theta \in \Theta_{k_n}} n^{-1}\sum_{i=1}^{n} l(\theta, Z_i) + O_p(\varepsilon_n^2) \quad \text{with} \quad \varepsilon_n \to 0. \tag{7.4}$$

Let

$$K(\theta, \theta_0) = El(\theta, Z) - El\,(\theta_0, Z) \geq 0.$$

We assume that the distance $d(\theta, \theta_0) := \|\theta - \theta_0\|$ is equivalent to $\sqrt{K(\theta, \theta_0)}$, i.e., there exists constants c_1 and c_2 such that

$$c_1\sqrt{K(\theta, \theta_0)} \leq d(\theta, \theta_0) \leq c_2\sqrt{K(\theta, \theta_0)}$$

for all θ in a small neighborhood of θ_0. We use such a norm for both the rate of convergence and asymptotic normality results.

To establish the rate of convergence, we maintain the following assumptions:

Assumption 7.3.1 *$\{Z_i\}_{i=1}^n$ is an i.i.d. sequence.*

Assumption 7.3.2 *There is a $C_1 > 0$ such that for all small $\varepsilon > 0$,*

$$\sup_{\{\theta \in \Theta_{k_n} : ||\theta_0 - \theta|| \leq \varepsilon\}} var\,(l(\theta, Z_i) - l(\theta_0, Z_i)) \leq C_1\varepsilon^2.$$

Assumption 7.3.3 *For any small $\delta > 0$, there exists a constant $s \in (0, 2)$ such that*

$$\sup_{\{\theta \in \Theta_{k_n} : ||\theta_0 - \theta|| \leq \delta\}} |l(\theta, Z_i) - l(\theta_0, Z_i)| \leq \delta^s U(Z_i),$$

with $E([U(Z_i)]^\gamma) \leq C_2$ for some $\gamma \geq 2$.

Assumptions 7.3.1 and 7.3.2 imply that, within a neighborhood of θ_0,

$$var\left(n^{-1/2}\sum_{i=1}^{n}(l(\theta,Z_i)-l(\theta_0,Z_i))\right) \text{ behaves like } ||\theta-\theta_0||^2.$$

Assumption 7.3.3 implies that, when restricting to a local neighborhood of θ_0, $l(\theta,Z_i)$ is "continuous" at θ_0 with respect to a metric $||\theta-\theta_0||$, which is locally equivalent to $K^{1/2}$. Assumptions 7.3.2 and 7.3.3 are usually easily verifiable by exploiting the specific form of the criterion function.

In view of the consistency of $\hat{\theta}_n$, we can focus on a neighborhood of $\theta_0 : ||\theta_0-\theta|| \leq \delta$ for some $\delta > 0$. Define

$$\mathcal{F}_n(\delta) = \{l(\theta,Z_i)-l(\theta_0,Z_i) : ||\theta_0-\theta|| \leq \delta, \theta \in \Theta_{k_n}\},$$

and

$$J_{[]}\left(\delta,\mathcal{F}_n(\delta),\|\cdot\|_2\right) = \int_0^{\delta}\sqrt{H_{[]}(w,\mathcal{F}_n(\delta),||\cdot||_2)}dw.$$

It is clear that $J_{[]}\left(\delta,\mathcal{F}_n(\delta),\|\cdot\|_2\right)$ increases as δ or n increases.

For a given constant C that does not depend on n, define

$$\delta_n = \inf\{\delta\in(0,1) : \frac{J_{[]}\left(\delta,\mathcal{F}_n(\delta),\|\cdot\|_2\right)}{\sqrt{n}\delta^2} \leq C\}.$$

Typically, as a function of δ, $J_{[]}\left(\delta,\mathcal{F}_n(\delta),\|\cdot\|_2\right)/\left(\sqrt{n}\delta^2\right)$ is nonincreasing in δ. See Figure 7.1 for an illustration of δ_n. To calculate δ_n, an upper bound on $H_{[]}(w,\mathcal{F}_n(\delta),||\cdot||_2)$ is often enough. Many bound results are available in the literature. For instance, according to Lemma 2.1 of Ossiander (1987) we have that, $H_{[]}(w,\mathcal{F}_n(\delta),||\cdot||_2) \leq H(w,\mathcal{F}_n(\delta),||\cdot||_\infty)$.

Before presenting the result on the rate convergence, we present a very useful lemma, which characterizes the behavior of $\sup_{f\in\mathcal{F}(\delta)}|\mathbb{G}_n f|$ for

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[f(Z_i)-Ef(Z)\right].$$

We use the notation "$a \lesssim b$" for "a is smaller than b, up to a generic constant."

Lemma 7.3.1 *Assume that $\{Z_i\}$ are iid. For any class $\mathcal{F}(\delta)$ of measurable functions such that*

$$Ef^2 \leq \delta^2 \text{ for all } f\in\mathcal{F}(\delta) \text{ and } \sup_{f\in\mathcal{F}(\delta)}\|f\|_\infty < B$$

we have

$$E\sup_{f\in\mathcal{F}(\delta)}|\mathbb{G}_n f| \lesssim J_{[]}\left(\delta,\mathcal{F}(\delta),\|\cdot\|_2\right)\left(1+\frac{J_{[]}\left(\delta,\mathcal{F}(\delta),\|\cdot\|_2\right)}{\sqrt{n}\delta^2}B\right).$$

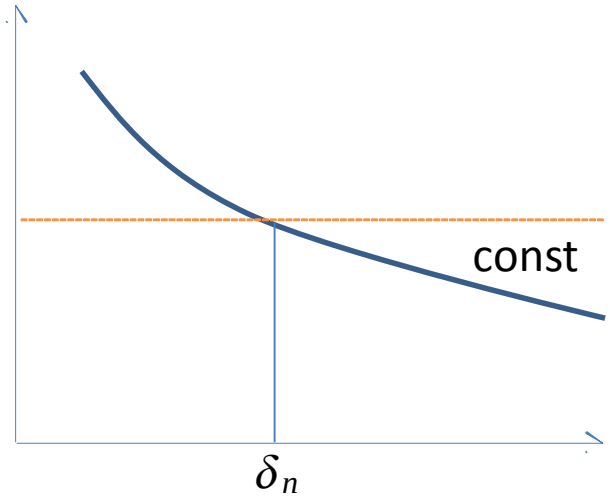


Figure 7.1: $J_{[]}\left(\delta, \mathcal{F}_n\left(\delta\right), \|\cdot\|_2\right)/\sqrt{n}\delta^2$ as a function of δ

i.e.,

$$E \sup_{f \in \mathcal{F}(\delta)} |\mathbb{G}_n f| \leq c J_{[]}\left(\delta, \mathcal{F}\left(\delta\right), \|\cdot\|_2\right)\left(1 + \frac{J_{[]}\left(\delta, \mathcal{F}\left(\delta\right), \|\cdot\|_2\right)}{\sqrt{n}\delta^2} B\right)$$

for some universal constant $c < \infty$.

Remark 7.3.1 *Roughly speaking, the lemma says that* $\sup_{f \in \mathcal{F}(\delta)} |\mathbb{G}_n f|$ *is locally equivalent to* $J_{[]}\left(\delta, \mathcal{F}\left(\delta\right), \|\cdot\|_2\right)$. *The lemma is Lemma 19.36 in van der Vaart (1998) and Lemma 3.4.2 in van der Vaart and Weller (1996). In the latter,* $J_{[]}\left(\delta, \mathcal{F}\left(\delta\right), \|\cdot\|_2\right)$ *is defined to be* $\int_0^\delta \sqrt{1 + H_{[]}(w, \mathcal{F}\left(\delta\right), \|\cdot\|_2)}dw$. *This does not seem to make any difference for our asymptotic development here. An alternative definition that leads to a tighter upper bound is*

$$J_{[]}\left(\delta, \mathcal{F}, \|\cdot\|_2\right) = \int_{b\delta^2}^{\delta} \sqrt{1 + H_{[]}(w, \mathcal{F}\left(\delta\right), \|\cdot\|_2)}dw$$

for some constant $b > 0$. *For some problems which are rare, the lower limit may dominate.*

Corollary 7.3.1 *Assume that* $\{Z_i\}$ *are iid. Let*

$$\mathcal{F}_n\left(\delta\right) = \{l(\theta, Z_i) - l(\theta_0, Z_i) : ||\theta_0 - \theta|| \leq \delta, \theta \in \Theta_{k_n}\},$$

be the class of functions such that

$$\sup_{\mathcal{F}_n} E\left[l(\theta, Z_i) - l(\theta_0, Z_i)\right]^2 \leq C_1\delta^2$$

$$\sup_{\mathcal{F}_n} \|l(\theta, Z_i) - l(\theta_0, Z_i)\|_\infty \leq \delta^s B$$

then

$$E \sup_{\mathcal{F}_n(\delta)} |\mathbb{G}_n [l(\theta, Z_i) - l(\theta_0, Z_i)]| \leq C J_{[]} (\delta, \mathcal{F}_n (\delta), \|\cdot\|_2) \left(1 + \frac{J_{[]} (\delta, \mathcal{F}_n (\delta), \|\cdot\|_2)}{\sqrt{n}\delta^2} \delta^s B\right).$$

For an empirical process $\mathbb{G}_n [l(\theta, \cdot)]$, we define

$$m_{\theta_0} (\delta) := \sup_{\|\theta - \theta_0\| \leq \delta} \mathbb{G}_n [l(\theta, \cdot) - l(\theta_0, \cdot)]$$

which is called the modulus of continuity of the process at θ_0. The "modulus of continuity" of the empirical process gives an upper bound on the rate of convergence.

Theorem 7.3.1 *Let assumptions 7.3.1-7.3.3 hold and $\eta_n = o(\varepsilon_n^2)$. If $J_{[]} (\delta, \mathcal{F}_n (\delta), \|\cdot\|_2) / \delta^\alpha$ is a nonincreasing function of δ for some constant $\alpha < 2$ and not depending on n. Then $\hat{\theta}_n = \theta_0 + O_p (\varepsilon_n)$ where $\varepsilon_n = \max \{\delta_n, \|\theta_0 - \pi_{k_n} \theta_0\|\}$.*

Proof. By the definition of $\hat{\theta}_n$ and the fact that $\pi_{k_n} \theta_0 \in \Theta_{k_n}$, we have

$$Q_n(\hat{\theta}_n) \leq Q_n(\pi_{k_n} \theta_0) + O_p (\eta_n) = Q_n(\pi_{k_n} \theta_0) + o_p(\varepsilon_n^2)$$

using $\eta_n = o(\varepsilon_n^2)$. Now:

$$\begin{aligned} d^2(\hat{\theta}_n, \theta_0) &\lesssim K(\hat{\theta}_n, \theta_0) = Q(\hat{\theta}_n) - Q(\theta_0) \\ &= Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\theta_0) - Q(\theta_0) + Q_n(\hat{\theta}_n) - Q_n(\theta_0) \\ &\leq Q_n(\pi_{k_n} \theta_0) - Q_n(\theta_0) + \frac{1}{\sqrt{n}} \left|\mathbb{G}_n \left[l(\hat{\theta}_n, Z) - l(\theta_0, Z)\right]\right| + o_p \left(\varepsilon_n^2\right) \end{aligned}$$

Next, since

$$\begin{aligned} & E [Q_n (\pi_{k_n} \theta_0) - Q_n (\theta_0)] \\ = \; & Q (\pi_{k_n} \theta_0) - Q (\theta_0) = K (\pi_{k_n} \theta_0, \theta_0) \\ \leq \; & C d^2 (\pi_{k_n} \theta_0, \theta_0) = O \left(\varepsilon_n^2\right) \text{ (by the definition of } \varepsilon_n) \end{aligned}$$

and

$$var ([Q_n (\pi_{k_n} \theta_0) - Q_n (\theta_0)]) \leq \frac{1}{n} var [l (\pi_{k_n} \theta_0, Z) - l (\theta_0, Z)] \leq \frac{1}{n} \varepsilon_n^2 = O(\varepsilon_n^4),$$

we have

$$E [Q_n (\pi_{k_n} \theta_0) - Q_n (\theta_0)]^2 = O(\varepsilon_n^4).$$

As a result,

$$P \left(Q_n(\pi_{k_n} \theta_0) - Q_n(\theta_0) \geq \kappa \varepsilon_n^2\right) \to 0$$

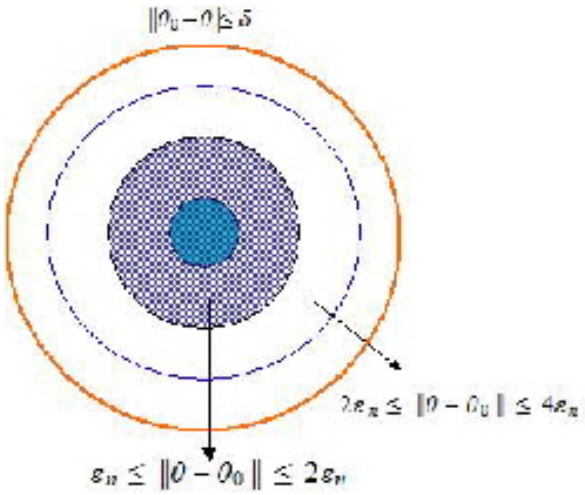


Figure 7.2: peels

as $\kappa \to \infty$. So for any M, we have

$$\begin{aligned} & P\left(d(\hat{\theta}_n, \theta_0) \geq 2^M \varepsilon_n\right) \\ = \; & P\left(d^2(\hat{\theta}_n, \theta_0) \geq \left(2^M \varepsilon_n\right)^2, \, d^2(\hat{\theta}_n, \theta_0) \leq \frac{1}{\sqrt{n}} \left|\mathbb{G}_n\left[l(\hat{\theta}_n, Z) - l(\theta_0, Z)\right]\right| + \kappa \varepsilon_n^2\right) + o(1). \end{aligned}$$

Note that (i) how the equivalence of $d^2(\theta, \theta_0)$ to $K(\theta, \theta_0)$ is used in establishing this result. If there is no sieve approximation, which occurs when the functional space is not too large, we only need $d^2(\hat{\theta}_n, \theta_0) \lesssim K(\hat{\theta}_n, \theta_0)$. (2) how the rate of convergence $d(\hat{\theta}_n, \theta_0)$ depends on the rate of change of the empirical process $\mathbb{G}_n[l(\theta, Z)]$ with respect to a small change in θ near θ_0.

To bound the above probability, we partition $\{\theta : \|\theta - \theta_0\| \leq \delta\}$ into "peels"

$$A_{n,j} = \left\{\theta \in \Theta_{k_n} : 2^{j-1}\varepsilon_n \leq \|\theta - \theta_0\| \leq 2^j \varepsilon_n\right\}, \; j = 1, ..., \mathcal{J}$$

where

$$\mathcal{J} = \min\left\{j : 2^j \varepsilon_n \geq \delta\right\}.$$

See figure 7.2.

To simplify the presentation, we assume that $\|U(\cdot)\|_\infty \leq B$ for some constant $B > 0$. We

do so in order to avoid using truncation arguments under the weak assumption 7.3.3. Now

$$\begin{aligned}
& P\left(d^2(\hat{\theta}_n,\theta_0) \geq \left(2^M\varepsilon_n\right)^2,\, d^2(\hat{\theta}_n,\theta_0) \leq \frac{1}{\sqrt{n}}\left|\mathbb{G}_n\left[l(\hat{\theta}_n,Z)-l(\theta_0,Z)\right]\right| + \kappa\varepsilon_n^2\right) \\
= & \sum_{j=M+1}^{\mathcal{J}} P\left(\hat{\theta}_n \in A_{nj} \text{ and } d^2(\hat{\theta}_n,\theta_0) \leq \left|\mathbb{G}_n\left[l(\hat{\theta}_n,Z)-l(\theta_0,Z)\right]\right| + \kappa\varepsilon_n^2\right) \\
\leq & \sum_{j=M+1}^{\mathcal{J}} P\left(\frac{1}{\sqrt{n}}\sup_{A_{nj}}\left|\mathbb{G}_n\left[l(\theta_n,Z)-l(\theta_0,Z)\right]\right| \geq \left(2^{j-1}\varepsilon_n\right)^2 - \kappa\varepsilon_n^2\right) \\
\leq & \sum_{j=M+1}^{\mathcal{J}} P\left(\frac{1}{\sqrt{n}}\sup_{\|\theta-\theta_0\|\leq 2^j\varepsilon_n}\left|\mathbb{G}_n\left[l(\theta,Z)-l(\theta_0,Z)\right]\right| \geq \left(2^{j-1}\varepsilon_n\right)^2 - \kappa\varepsilon_n^2,\right.
\end{aligned}$$

Note that how the peeling device is used. For each $\hat{\theta}_n \in A_{nj}$, we have $\left(2^{j-1}\varepsilon_n\right)^2 \leq d^2(\hat{\theta}_n,\theta_0) \leq \left(2^j\varepsilon_n\right)^2$. So $d^2(\hat{\theta}_n,\theta_0) \leq \left|\mathbb{G}_n\left[l(\hat{\theta}_n,Z)-l(\theta_0,Z)\right]\right| + \kappa\varepsilon_n^2$ implies that

$$\left(2^{j-1}\varepsilon_n\right)^2 \leq \left|\mathbb{G}_n\left[l(\hat{\theta}_n,Z)-l(\theta_0,Z)\right]\right| + \kappa\varepsilon_n^2$$

This is where we used the lower bound of peel j. On the other hand, for each $\hat{\theta}_n \in A_{nj}$, we have

$$\left|\mathbb{G}_n\left[l(\hat{\theta}_n,Z)-l(\theta_0,Z)\right]\right| \leq \sup_{\|\theta-\theta_0\|\leq 2^j\varepsilon_n}\left|\mathbb{G}_n\left[l(\theta,Z)-l(\theta_0,Z)\right]\right|$$

This is where we used the upper bound the peel j.

Using Corollary 7.3.1, we deduce that,

$$\begin{aligned}
& P\left(d(\hat{\theta}_n,\theta_0) \geq 2^M\varepsilon_n\right) \\
\leq & \sum_{j=M+1}^{\mathcal{J}} P\left(\frac{1}{\sqrt{n}}\sup_{\|\theta-\theta_0\|\leq 2^j\varepsilon_n}\left|\mathbb{G}_n\left[l(\theta,Z)-l(\theta_0,Z)\right]\right| \geq \left(2^{j-1}\varepsilon_n\right)^2 - \kappa\varepsilon_n^2,\right. \\
& \left.\sup_{\|\theta-\theta_0\|\leq 2^j\varepsilon_n}\left\|l(\theta,Z)-l(\theta_0,Z)\right\|_2 \lesssim \left(2^j\varepsilon_n\right)^2, \left\|l(\theta,Z)-l(\theta_0,Z)\right\|_\infty \leq \left(2^j\varepsilon_n\right)^s B\right) + o(1) \\
\lesssim & \sum_{j=M+1}^{\mathcal{J}}\left(\frac{J_{[]}\left(2^j\varepsilon_n,\mathcal{F}_n\left(2^j\varepsilon_n\right),\|\cdot\|_2\right)}{\sqrt{n}\left[\left(2^{j-1}\varepsilon_n\right)^2-\kappa\varepsilon_n^2\right]}\left(1+\frac{J_{[]}\left(2^j\varepsilon_n,\mathcal{F}_n\left(2^j\varepsilon_n\right),\|\cdot\|_2\right)}{\sqrt{n}\left(2^j\varepsilon_n\right)^2}\left(2^j\varepsilon_n\right)^s B\right)\right) + o(1)
\end{aligned}$$

by the Markov inequality

By the definition of δ_n,

$$\frac{J_{[]}\left(2^j\varepsilon_n,\mathcal{F}_n\left(2^j\varepsilon_n\right),\|\cdot\|_2\right)}{\sqrt{n}\left(2^j\varepsilon_n\right)^2} \leq const,$$

as $2^j\varepsilon_n > \delta_n$. In addition, $\left(2^j\varepsilon_n\right)^s \leq \delta^s$, and

$$\frac{J_{[]}\left(2^j\varepsilon_n, \mathcal{F}_n\left(2^j\varepsilon_n\right), \|\cdot\|_2\right)}{\left(2^j\varepsilon_n\right)^\alpha} \leq \frac{J_{[]}\left(\varepsilon_n, \mathcal{F}_n\left(\varepsilon_n\right), \|\cdot\|_2\right)}{\left(\varepsilon_n\right)^\alpha}$$

which implies that

$$J_{[]}\left(2^j\varepsilon_n, \mathcal{F}_n\left(2^j\varepsilon_n\right), \|\cdot\|_2\right) \lesssim 2^{j\alpha} J_{[]}\left(\varepsilon_n, \mathcal{F}_n\left(\varepsilon_n\right), \|\cdot\|_2\right) \lesssim 2^{j\alpha}\sqrt{n}\left(\varepsilon_n\right)^2.$$

Inserting these bounds into the above equation, we obtain, for some constant $\mathcal{C}$

$$\begin{aligned}
& P\left(d(\hat{\theta}_n, \theta_0) \geq 2^M\varepsilon_n\right) \\
\lesssim & \sum_{j=M+1}^{\mathcal{J}} \left(\frac{2^{j\alpha}\varepsilon_n^2}{\left(2^{j-1}\varepsilon_n\right)^2 - \kappa\varepsilon_n^2}\right)\left(1 + \mathcal{C}\left(2^j\varepsilon_n\right)^s B\right) + o(1) \\
\leq & \sum_{j=M+1}^{\mathcal{J}} \left(\frac{2^{j\alpha}}{\left(2^{j-1}\right)^2 - \kappa}\right)(1 + \mathcal{C}\delta^s B) + o(1) \\
= & \sum_{j=M+1}^{\mathcal{J}} \left(\frac{2^{j(\alpha-2)+2}}{1 - 4\kappa/2^{2j}}\right)(1 + \mathcal{C}\delta^s B) + o(1) \\
\to & \ 0 \text{ as } M \to \infty \text{ at any rate.}
\end{aligned}$$

That is, we can make $P\left(d(\hat{\theta}_n, \theta_0) \geq 2^M\varepsilon_n\right)$ arbitrarily small by choosing M large enough. ■

We note that δ_n increases with the complexity of the sieve Θ_{k_n} and can be interpreted as a measure of the standard deviation term, while the deterministic approximation error $||\theta_0 - \pi_{k_n}\theta_0||$ decreases with the complexity of the sieve Θ_{k_n} and is a measure of the bias. The best convergence rate can be obtained by choosing the complexity of the sieve Θ_{k_n} such that $\delta_n \asymp ||\theta_0 - \pi_{k_n}\theta_0||$.

The following theorem is similar to Theorem 7.3.1 but for the case when there is no need to use the method of sieves.

Theorem 7.3.2 *Assume that*

(i) $c_1\sqrt{K(\theta, \theta_0)} \leq d(\theta, \theta_0) \leq c_2\sqrt{K(\theta, \theta_0)}$

(ii) $E\sup_{d(\theta,\theta_0)\leq\delta} |Q_n(\theta) - Q(\theta) - [Q_n(\theta_0) - Q(\theta_0)]| \leq \phi_n(\delta)/\sqrt{n}$ *where* $\phi_n(\delta)/\delta^\alpha$ *is a nonincreasing function of* δ *for* $\alpha < 2$.

(iii) $d\left(\hat{\theta}_n, \theta_0\right) = o_p(1)$, *then*

$d\left(\hat{\theta}_n, \theta_0\right) = O_p(\delta_n)$ *for* δ_n *satisfying*

$$\frac{\phi_n(\delta_n)}{\sqrt{n}\delta_n^2} \leq c_3$$

for some constant c_3 *not depending on* n *and for every* n.

β	s	name / situation
1	1/2	classical smoothness
1/2	1/3	bounded monotone on $\mathbb{R}$
3/4	2/5	convex on $\mathbb{R}$
3/4	2/5	bounded second derivative on $[0,1]$
$1-d/4$	$2/(d+4)$	convex in $\mathbb{R}^d$

The above theorem can be proved using the same idea for proving Theorem 7.3.1. As an example, consider estimating a monotone decreasing density on $[0,\infty]$ by MLE. The nonparametric MLE is

$$\hat{p}(x) = \arg\min_{p\in\mathcal{P}} -\sum_{i=1}^{n} \log p(X_i)$$

where $\mathcal{P}$ is the set of all nonincreasing density functions on $[0,\infty]$. In this case, the function space is small enough that we do not need to use the method of sieves. The estimator is the left derivative of lease concave majorant of the empirical CDF.

Here are some rates when $\phi_n(\delta) = \delta^\beta$ and $\delta_n = n^{-1/[2(2-\beta)]} = n^{-s}$:

7.3.2 Example: Additive Mean Regression

Suppose that the i.i.d. data $\{Y_i, X_i' = (X_{1i}, ..., X_{qi})\}_{i=1}^n$ is generated according to

$$Y_i = h_{01}(X_{1i}) + ... + h_{0q}(X_{qi}) + u_i, \quad E[u_i|X_i] = 0.$$

Let $\theta_0 = (h_{01}, ..., h_{0q})' \in \Theta = \mathcal{H}$ be the parameters of interest with $\mathcal{H} = \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$ to be specified later. For simplicity, we assume that $\dim(X_j) = 1$ for $j = 1, ..., q$, $\dim(X) = q$ and $\dim(Y) = 1$. We estimate the regression function $\theta_0(X) = (h_{01}(X_1), ..., h_{0q}(X_q))'$ by minimizing over a sieve $\Theta_{k_n} = \mathcal{H}_n$ the criterion

$$Q_n(\theta) = n^{-1}\sum_{i=1}^{n} l(\theta, Z_i),$$

where

$$l(\theta, Z_i) = \frac{1}{2}[Y_i - \sum_{j=1}^{q} h_j(X_{ji})]^2$$

and $Z_i = (Y_i, X_i')'$. Let

$$\|\theta - \theta_0\|^2 = \mathrm{E}\left(\|\theta(X) - \theta_0(X)\|_E^2\right) = E\{\sum_{j=1}^{q}[h_j(X_j) - h_{oj}(X_j)]\}^2,$$

where $\|\cdot\|_E$ is the Euclidian norm. We will show that this squared norm is equivalent to $K(\theta, \theta_0)$.

Before we impose further assumptions, let us recall the definition of a Hölder ball:

$$\Lambda_c^p(\mathcal{X}) = \left\{h \in C^{[p]}(\mathcal{X}) : \sup_{j\leq[p]}\sup_{x\in\mathcal{X}}\left|h^{(j)}(x)\right| \leq c,\ \sup_{x,y\in\mathcal{X}}\frac{\left|h^{([p])}(x) - h^{([p])}(y)\right|}{|x-y|^{\alpha}} \leq c\right\}.$$

Assumption 7.3.4 *(i) for $j = 1, ..., q$, $h_{oj} \in \mathcal{H}^j = \Lambda_{c_j}^{p_j}([b_{1j}, b_{2j}])$ with $p_j > 1/2$; (ii) $h_{oj}(x_j^*) = 0, j = 2, ..., q$ for some known $x_j^* \in (b_{1j}, b_{2j})$.*

Assumption 7.3.5 *$\sigma^2(X) \equiv E[u^2|X]$ is bounded.*

Assumption 7.3.4(ii) is sufficient for identification, and Assumption 7.3.5 is a simple regularity assumption that has been imposed in many papers; see e.g. Newey (1997).

The sieve will be chosen to have the form $\mathcal{H}_n = \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$ where

$$\mathcal{H}_n^1 = \{h_1 \in \Theta_{1n} : ||h_1||_\infty \leq c_1\}$$

and

$$\mathcal{H}_n^j = \{h_j \in \Theta_{jn} : h_j(x_j^*) = 0, ||h_j||_\infty \leq c_j\}$$

for $j = 2, ..., q$. Here Θ_{jn} can be any of the finite-dimensional linear sieve examples such as $\Theta_{jn} = \mathrm{Pol}(k_{jn})$ or $\mathrm{TriPol}(k_{jn})$.

Proposition 7.3.1 *Let $\hat{\theta}_n$ be the sieve M-estimator. Suppose that Assumptions 7.3.4 and 7.3.4 hold. Let $k_{jn} = O(n^{1/(2p_j+1)})$ for $j = 1, ..., q$. Then $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-p/(2p+1)})$ with $p = \min\{p_1, \cdots, p_q\}$.*

Proof. Theorem 7.3.1 is readily applicable to prove this result. Let $\ell_q = (1, 1, ..., 1)' \in \mathbb{R}^q$. Then

$$\begin{aligned}
l(\theta, Z_i) - l(\theta_0, Z_i) &= \frac{1}{2}\left([Y_i - \theta'(X_i)\,\ell_q)]^2 - [Y_i - \theta_0'(X_i)\,\ell_q]^2\right) \\
&= \frac{1}{2}\left([Y_i - \theta_0'\ell_q - (\theta - \theta_0)'\ell_q]^2 - [Y_i - \theta_0'\ell_q]^2\right) \\
&= \frac{1}{2}\left([u_i - (\theta - \theta_0)'\ell_q]^2 - u_i^2\right) \\
&= (\theta - \theta_0)'\ell_q u_i + \frac{1}{2}(\theta - \theta_0)'\ell_q\ell_q'(\theta - \theta_0). \qquad (7.5)
\end{aligned}$$

As a result,

$$\begin{aligned} K(\theta,\theta_0) &= E\left[l(\theta,Z_i)-l(\theta_0,Z_i)\right] = E\frac{1}{2}(\theta-\theta_0)'\ell_q\ell_q'(\theta-\theta_0) \\ &= \frac{1}{2}E\left[\sum_{i=1}^{q}(\theta_i-\theta_{i0})\right]^2 \leq \frac{q}{2}E\left(\|\theta-\theta_0\|_E^2\right) = \frac{q}{2}\|\theta-\theta_0\|^2. \end{aligned}$$

So $\|\theta-\theta_0\|^2 \to 0$ implies $K(\theta,\theta_0)\to 0$. On the other hand, if $K(\theta,\theta_0)\to 0$, then $\ell_q'(\theta-\theta_0)\to 0$ almost surely. Given the identification assumption, we deduce that $\|\theta-\theta_0\|_E^2 \to 0$ almost surely, which implies that $\|\theta-\theta_0\|_2^2\to 0$. So $K(\theta,\theta_0)$ and $\|\theta-\theta_0\|^2$ are equivalent.

Assumption 7.3.1 is assumed. Now we check Assumptions 7.3.2 and 7.3.3. In view of (7.5), we have

$$\begin{aligned} E[l(\theta,Z_i)-l(\theta_0,Z_i)]^2 &\leq const.E(\sigma^2(X_i)[\theta_0(X_i)'\ell_q-\theta(X_i)'\ell_q]^2)+E([\theta_0(X_i)'\ell_q-\theta(X_i)'\ell_q]^4) \\ &\leq const.\|\theta-\theta_0\|^2+E([\theta_0(X_i)'\ell_q-\theta(X_i)'\ell_q]^4). \end{aligned}$$

By Theorem 1 of Gabushin (1967) when p is an integer and Lemma 2 in Chen and Shen (1998) for any $p>0$, we have

$$\|\theta-\theta_0\|_\infty \leq c\|\theta-\theta_0\|^{2p/(2p+1)}.$$

Hence

$$\begin{aligned} E([\theta_0(X_i)'\ell_q-\theta(X_i)'\ell_q]^4) &\leq \sup_x[\sum|\theta_j(x)-\theta_{0j}(x)|]^2E([\theta_0(X_i)'\ell_q-\theta(X_i)'\ell_q]^2) \\ &\leq C\|\theta-\theta_0\|^{2(1+[2p/(2p+1)])}. \end{aligned}$$

So Assumption 7.3.2 is satisfied for all $\varepsilon\leq 1$. On the other hand,

$$|l(\theta,Z_i)-l(\theta_0,Z_i)| \leq \|\theta-\theta_0\|_\infty[|u_i|+(\|\theta_0\|_\infty+\|\theta\|_\infty)/2] \quad a.s.$$

Using Lemma 2 in Chen and Shen (1998), we have

$$|l(\theta,Z_i)-l(\theta_0,Z_i)| \leq c\|\theta-\theta_0\|^{2p/(2p+1)}[|u_i|+(\|\theta_0\|_\infty+\|\theta\|_\infty)/2],$$

so Assumption 7.3.3 is satisfied with $s=2p/(2p+1)$, $U(Z_i)=|u_i|+const$ and $\gamma=2$.

To apply Theorem 7.3.1, it remains to compute the deterministic approximation error rate $||\theta_0-\pi_{k_n}\theta_0||$ and the metric entropy with bracketing $H_{[]}(w,\mathcal{F}_n,||\cdot||_2)$ of the class

$$\mathcal{F}_n=\{l(\theta,Z_i)-l(\theta_0,Z_i):\|\theta-\theta_0\|\leq\delta,\theta\in\Theta_{k_n}\}.$$

By definition,

$$\|\theta_0-\pi_{k_n}\theta_0\| \leq const.\max\{\|h_{oj}-\pi_{k_n}h_{oj}\|_\infty : j=1,...,q\}.$$

Therefore, there exists a constant C with $0 < w/C \leq \delta < 1$ such that

$$H_{[]}(w, \mathcal{F}_n, ||\cdot||_2) \leq \sum_{j=1}^{q} \log N(\frac{w}{C}, \mathcal{H}_n^j, ||\cdot||_\infty).$$

This holds because we can use $N(w/C, \mathcal{H}_n^j, ||\cdot||_\infty)$ balls to cover $\mathcal{H}_n^j$ wrt the norm $||\cdot||_\infty$. That is, $\mathcal{F}_n$ can be covered with $\Pi_{j=1}^q N(w/C, \mathcal{H}_n^j, ||\cdot||_\infty)$ balls, each with radius w/C. In other words, for any $f \in \mathcal{F}_n$, we can find f^* such that $\|f - f^*\|_\infty \leq qw/C$. We have therefore found $\Pi_{j=1}^q N(w/C, \mathcal{H}_n^j, ||\cdot||_\infty)$ pairs of brackets, each of which has the form $[f^* - qw/C, f^* + qw/C]$ with L_2-distance between the brackets less than w if we choose C appropriately.

The final bit of calculation now depends on the choice of sieves. First, for $j = 1, ..., q$, $\mathcal{H}^j = \Lambda_{c_j}^{p_j}$,

$$\|h_{oj} - \pi_{k_n} h_{oj}\|_\infty = O((k_{jn})^{-p_j})$$

by Lorentz (1966). Second, for all $j = 1, 2, ..., q$,

$$\log N(\frac{w}{C}, \mathcal{H}_n^j, ||\cdot||_\infty) \leq const \times k_{jn} \times \log(1 + \frac{4c_j}{w})$$

by Lemma 2.5 in van de Geer (2000). Hence δ_n solves

$$\begin{aligned} \frac{1}{\sqrt{n}\delta_n^2} \int_0^{\delta_n} \sqrt{H_{[]}(w, \mathcal{F}_n, ||\cdot||_2)} dw &\leq \frac{1}{\sqrt{n}\delta_n^2} \max_{j=1,...,q} \int_0^{\delta_n} \sqrt{k_{jn} \times \log(1 + \frac{4c_j}{w})} dw \\ &\leq \frac{1}{\sqrt{n}\delta_n^2} \max_{j=1,...,q} \sqrt{k_{jn}} \times \delta_n \leq const \end{aligned}$$

and the solution is $\delta_n \asymp \max_{j=1,...,q} \sqrt{\frac{k_{jn}}{n}}$. By Theorem 7.3.1,

$$\|\hat{\theta}_n - \theta_0\| = O_p\left(\max_{j=1,...,q} \{(k_{jn})^{-p_j}, \delta_n\}\right).$$

With the choice of $k_{jn} = O(n^{1/(2p_j+1)})$ for $j = 1, ..., q$, we obtain

$$\|\hat{\theta}_n - \theta_0\| = O_p(n^{-p/(2p+1)})$$

with $p = \min\{p_1, \cdots, p_q\} > 0.5$. This immediately implies

$$\|\hat{h}_{nj} - h_{0j}\|_2 = O_p(n^{-p/(2p+1)})$$

for $j = 1, ..., q$. ■

Remark: Since the parameter space $\mathcal{H} = \mathcal{H}^1 \times \cdots \times \mathcal{H}^q$ is compact with respect to the norm $||\cdot||$, we can take the original parameter space $\mathcal{H}$ as the sieve space $\mathcal{H}_n$. Applying Theorem 7.3.1

again, note that the approximation error $\|\pi_{k_n}\theta_0 - \theta_0\| = 0$, we have $\|\hat{\theta}_n - \theta_0\| = O_p(\delta_n)$, where δ_n solves:

$$\begin{aligned}
&\frac{1}{\sqrt{n}\delta_n^2}\int_0^{\delta_n}\sqrt{\sum_{j=1}^{q}\log N(w,\mathcal{H}^j,||\cdot||_\infty)}dw \\
\leq\ &\frac{1}{\sqrt{n}\delta_n^2}\int_0^{\delta_n}\sqrt{\sum_{j=1}^{q}\left(\frac{c_j}{w}\right)^{1/p_j}}dw \quad \text{by Birman and Solomjak (1967)} \\
\leq\ &\frac{1}{\sqrt{n}\delta_n^2}\max_{j=1,\ldots,q} const.\,(\delta_n)^{1-\frac{1}{2p_j}} \leq const.
\end{aligned}$$

which is satisfied if $\delta_n = O(n^{-p/(2p+1)})$ with $p = \min\{p_1, \cdots, p_q\} > 0.5$. However, it is unclear how one can implement such an optimization over the entire parameter space $\mathcal{H}$ given a finite data set.

7.3.3 Example: Multivariate Quantile Regression (Optional)

Suppose that the i.i.d. data $\{Y_i, X_i\}_{i=1}^n$ is generated according to

$$Y_i = \theta_0(X_i) + u_i, \quad P[u_i \leq 0|X_i] = \alpha \in (0,1),$$

where $X_i \in \mathcal{X} = \mathcal{R}^d$, $d \geq 1$. We estimate the conditional quantile function $\theta_0(\cdot)$ by minimizing over Θ_{k_n} the criterion

$$\widehat{Q}_n(\theta) = n^{-1}\sum_{i=1}^{n} l(\theta, Y_i, X_i),$$

where

$$l(\theta, Y_i, X_i) = \{1(Y_i < \theta(X_i)) - \alpha\}[Y_i - \theta(X_i)].$$

Let

$$\|\theta - \theta_0\|^2 = \mathrm{E}(\theta(X_i) - \theta_0(X_i))^2$$

and $W_1^m(\mathcal{X})$ be the Sobolev space where functions as well as all their partial derivatives (up to m-th order) are $L_1(\mathcal{X}, leb)$-integrable.

Assumption 7.3.6 $\theta_0 \in \Theta = W_1^1(\mathcal{X})$.

Assumption 7.3.7 *Let $f_{u|X}$ be the conditional density of u_i given X_i satisfying*

$$0 < \inf_{x\in\mathcal{X}} f_{u|X=x}(0) \leq \sup_{x\in\mathcal{X}} f_{u|X=x}(0) < \infty$$

and

$$\sup_{x\in\mathcal{X}} |f_{u|X=x}(z) - f_{u|X=x}(0)| \to 0 \ \textit{as}\ |z| \to 0.$$

It is known that the tensor product of finite-dimensional linear sieves will not be able to approximate functions in $W_1^m(\mathcal{X})$, $m \geq 1$ well, hence the sieve convergence rates based on those linear sieves will be slower than those based on nonlinear sieves; see e.g. Chen and Shen (1998, proposition 1 case 1.3(ii)) for such an example. Chen et al. (2001) have shown that neural network sieves lead to faster convergence rates for functions in $W_1^m(\mathcal{X})$. Thus we consider the following Gaussian radial basis ANN sieve Θ_{k_n} for the unknown $\theta_0 \in W_1^1(\mathcal{X})$:

$$\Theta_{k_n} = \left\{ \alpha_0 + \sum_{j=1}^{k_n} \alpha_j G\left(\frac{\{(x-\gamma_j)'(x-\gamma_j)\}^{1/2}}{\sigma_j} \right), \sum_{j=0}^{k_n} |\alpha_j| \leq c_0,\ |\gamma_j| \leq c_1, 0 < \sigma_j \leq c_2 \right\},$$

where G is the standard Gaussian density function.

Proposition 7.3.2 *Let $\hat{\theta}_n$ be the sieve M-estimate. Suppose that Assumptions* 7.3.6 *and* 7.3.7 *hold. Let $k_n^{2(1+1/(d+1))} \log(k_n) = O(n)$. Then $\|\hat{\theta}_n - \theta_0\| = O_p([n/\log n]^{-(1+2/(d+1))/[4(1+1/(d+1))]})$.*

Proof. Theorem 7.3.1 is readily applicable to prove this result. Assumption 7.3.1 is directly assumed. By the above assumptions on conditional density $f_{u|X}$, it is easy to check that $K(\theta_0, \theta) \asymp \mathrm{E}(\theta(X_i) - \theta_0(X_i))^2$; see Chen and White (1999, page 686-687) for details. Now let us check Assumptions 7.3.2 and 7.3.3. Note that

$$|l(\theta, Y_i, X_i) - l(\theta_0, Y_i, X_i)| \leq \max(\alpha, 1-\alpha)|\theta(X_i) - \theta_0(X_i)|,$$

we have

$$var\left(l(\theta, Y_i, X_i) - l(\theta_0, Y_i, X_i)\right) \leq E[l(\theta, Y_i, X_i) - l(\theta_0, Y_i, X_i)]^2 \leq E[\theta(X_i) - \theta_0(X_i)]^2,$$

and thus Assumption 7.3.2 is satisfied. Moreover, we have

$$\sup_{\{\theta \in \Theta_{k_n} : \|\theta - \theta_0\| \leq \delta\}} |l(\theta, Y_i, X_i) - l(\theta_0, Y_i, X_i)| \leq \sup_{\{\theta \in \Theta_{k_n} : \|\theta - \theta_0\| \leq \delta\}} |\theta(X_i) - \theta_0(X_i)|,$$

and

$$\|\theta - \theta_0\|_\infty \leq c\|\theta - \theta_0\|^{2/3}$$

by Theorem 1 of Gabushin (1967). Hence, Assumption 7.3.3 is satisfied with $s = 2/3$, $U(X_i) \equiv c$.

Now by results in Chen et al. (2001),

$$\|\theta_0 - \pi_{k_n}\theta_0\| \leq const.(k_n)^{-1/2-1/(d+1)}$$

and

$$\log N(w, \Theta_{k_n}, \|\cdot\|_\infty) \leq const.k_n \log(\frac{k_n}{w}).$$

With $k_n^{2(1+1/(d+1))} \log(k_n) = O(n)$, it is easy to see that

$$\|\hat{\theta}_n - \theta_0\| = O_p([n/\log n]^{-(1+2/(d+1))/[4(1+1/(d+1))]})$$

by applying Theorem 7.3.1. ■

7.4 Smooth Functional of Sieve M-Estimator

7.4.1 Asymptotic Normality of Smooth Functionals

In this subsection we present a simple $\sqrt{n}$-asymptotic normality theorem for the plug-in estimate of a smooth functional of θ_0. Let

$$\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n) = \arg \min_{(\beta,h)\in B\times\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n} l(\beta, h, Z_i)$$

be the sieve M-estimator of $\theta_0 = (\beta_0, h_0)$.

Suppose that $\Theta = B \times \mathcal{H}$ is convex in θ_0 so that $\theta_0 + \tau[\theta - \theta_0] \in \Theta$ for all small $\tau \in [0, 1]$ and for all fixed $\theta \in \Theta$. Suppose that the directional derivative

$$\frac{\partial l(\theta_0, z)}{\partial \theta}[\theta - \theta_0] \equiv \lim_{\tau\to 0} \frac{l(\theta_0 + \tau[\theta - \theta_0], z) - l(\theta_0, z)}{\tau}$$

is well-defined for almost all z in the support of Z.

Let $\Theta = B\times\mathcal{H}$ be equipped with a norm $\|\cdot\|$. Suppose the functional of interest, $f : \Theta \to \mathcal{R}$, is smooth in the sense that

$$\frac{\partial f(\theta_0)}{\partial \theta}[\theta - \theta_0] \equiv \lim_{\tau\to 0} \frac{f(\theta_0 + \tau[\theta - \theta_0]) - f(\theta_0)}{\tau}$$

is well-defined and

$$\|\frac{\partial f(\theta_0)}{\partial \theta}\| \equiv \sup_{\{\theta\in\Theta:\|\theta-\theta_0\|>0\}} \frac{|\frac{\partial f(\theta_0)}{\partial \theta}[\theta - \theta_0]|}{\|\theta - \theta_0\|} < \infty$$

Next, suppose that $\|\cdot\|$ induces an inner product $\langle\cdot,\cdot\rangle$ on the completion of the space spanned by $\Theta - \theta_0$, denoted as $\bar{V}$. That is

$$\langle\theta_1, \theta_2\rangle = \left(\|\theta_1\|^2 + \|\theta_2\|^2 - \|\theta_1 - \theta_2\|\right)/2 \text{ for any } \theta_1, \theta_2 \in \bar{V}.$$

By the Riesz representation theorem, there exists $v^* \in \bar{V}$ such that, for any $\theta \in \Theta$,

$$\frac{\partial f(\theta_0)}{\partial \theta}[\theta - \theta_0] = \langle\theta - \theta_0, v^*\rangle \quad \text{iff} \quad \|\frac{\partial f(\theta_0)}{\partial \theta}\| < \infty.$$

As a special case, consider a parametric MLE and $f(\theta) = \lambda'\theta$ for some λ. For MLE, we know that

$$\begin{aligned} K(\theta, \theta_0) &= E[\ell(\theta, Z) - \ell(\theta_0, Z)] \doteq \frac{1}{2}(\theta - \theta_0)' E\frac{\partial^2 \ell(\theta_0, Z)}{\partial\theta\partial\theta'}(\theta - \theta_0) \\ &= \frac{1}{2}(\theta - \theta_0)' H(\theta - \theta_0). \end{aligned}$$

So we can employ $\|\theta-\theta_0\|^2=(\theta-\theta_0)'H(\theta-\theta_0)$ and define

$$\langle\theta_1-\theta_0,\theta_2-\theta_0\rangle=(\theta_1-\theta_0)'H(\theta_2-\theta_0).$$

Now

$$\begin{aligned}\|\frac{\partial f(\theta_0)}{\partial\theta}\|^2 &\equiv \sup_{\{\theta\in\Theta:\|\theta-\theta_0\|>0\}}\frac{\left|\lambda'(\theta-\theta_0)\right|^2}{\|\theta-\theta_0\|^2}=\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|>0\}}\frac{(\theta-\theta_0)'\lambda\lambda'(\theta-\theta_0)}{(\theta-\theta_0)'H(\theta-\theta_0)}\\ &=\sup_{s\neq0}\frac{s'\lambda\lambda's}{s'Hs}\end{aligned}$$

It is not hard to see that the sup is achieved at

$$s^*=H^{-1}\lambda$$

and thus $v^*=H^{-1}\lambda$. To check

$$\frac{\partial f(\theta_0)}{\partial\theta}[\theta-\theta_0]=\langle\theta-\theta_0,v^*\rangle$$

we note that $\frac{\partial f(\theta_0)}{\partial\theta}[\theta-\theta_0]=\lambda'(\theta-\theta_0)$ and $\langle\theta-\theta_0,v^*\rangle=(\theta-\theta_0)'Hv^*=(\theta-\theta_0)'\lambda$. So $v^*=H^{-1}\lambda$ is indeed the representator of $\lambda'(\theta-\theta_0)$.

Comment: In general, to find the representor, we solve the maximization problem

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|>0\}}\frac{|\frac{\partial f(\theta_0)}{\partial\theta}[\theta-\theta_0]|}{\|\theta-\theta_0\|}=\sup_{\{s\neq0\}}\frac{|\frac{\partial f(\theta_0)}{\partial\theta}[s]|^2}{\|s\|^2}.$$

The maximizer s^* is the Riesz representor.

Let e_n denote any sequence satisfying $e_n=o(n^{-1/2})$, and

$$\mu_n(g)=\frac{1}{\sqrt{n}}\mathbb{G}_n(g)=\frac{1}{n}\sum_{i=1}^n\{g(Z_i)-E(g(Z_i))\}.$$

Recall that

$$K(\theta,\theta_0)\equiv E[l(\theta,Z_i)-l(\theta_0,Z_i)]=Q(\theta)-Q(\theta_0).$$

Assumption 7.4.1 *(i) there is an $\omega>0$ such that $|f(\theta)-f(\theta_0)-\frac{\partial f(\theta_0)}{\partial\theta}[\theta-\theta_0]|=O(\|\theta-\theta_0\|^\omega)$ uniformly in $\theta\in\Theta_{k_n}$ with $\|\theta-\theta_0\|=o(1)$; (ii) $\|\frac{\partial f(\theta_0)}{\partial\theta}\|<\infty$; (iii) there is a $\pi_{k_n}v^*\in\Theta_{k_n}$ such that $||\pi_{k_n}v^*-v^*||\times\|\hat\theta_n-\theta_0\|=o_p(n^{-1/2})$.*

Assumption 7.4.2 $\sup_{\{\theta\in\Theta_{k_n}:\|\theta-\theta_0\|\leq\varepsilon_n\log n\}}\mu_n(l(\theta\pm e_n\pi_{k_n}v^*,Z)-l(\theta,Z)-\frac{\partial l(\theta_0,Z)}{\partial\theta}[\pm e_n\pi_{k_n}v^*])=O_p(e_n^2)$

Assumption 7.4.3 $K(\hat{\theta}_n \pm e_n \pi_{k_n} v^*, \theta_0) - K(\hat{\theta}_n, \theta_0) = \pm e_n \langle \hat{\theta}_n - \theta_0, \pi_{k_n} v^* \rangle + o(n^{-1})$.

Assumption 7.4.4 *(i)* $\mu_n(\frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^* - v^*]) = o_p(n^{-1/2})$; *(ii)* $E\{\frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^*]\} = o(n^{-1/2})$.

Assumption 7.4.5 $n^{1/2}\mu_n(\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^*]) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*})$, *with* $\sigma^2_{v^*} > 0$.

We note that for the classical nonlinear M-estimation such as those reviewed in Newey and McFadden (1994), Assumptions 7.4.1(i)(ii), 7.4.2, 7.4.3 and 7.4.5 are still required (albeit in slightly different expressions), while Assumptions 7.4.1(iii) and 7.4.4 are automatically satisfied since $\pi_{k_n} v^* = v^*$ for the standard nonlinear M-estimation. Note that for i.i.d. data Assumption 7.4.5 is satisfied whenever $\sigma^2_{v^*} = Var\left(\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^*]\right) > 0$. If $l(\theta, Z)$ is also pathwise differentiable in $\theta \in \Theta_{k_n}$ with $\|\theta - \theta_0\| = o(1)$, then Assumptions 7.4.2 and 7.4.3 are implied by Assumptions 7.4.2**'** and 7.4.3' respectively, where

Assumption 7.4.2' $\sup_{\{\overline{\theta} \in \Theta_{k_n} : \|\overline{\theta} - \theta_0\| \leq \delta_n\}} \mu_n \left(\frac{\partial l(\overline{\theta}, Z)}{\partial \theta}[\pi_{k_n} v^*] - \frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^*] \right) = o_p(n^{-1/2})$.

Assumption 7.4.3' $E\{\frac{\partial l(\hat{\theta}_n, Z)}{\partial \theta}[\pi_{k_n} v^*]\} = \langle \hat{\theta}_n - \theta_0, \pi_{k_n} v^* \rangle + o(n^{-1/2})$.

Theorem 7.4.1 *Suppose Assumptions* 7.4.1–7.4.5 *hold and* $\|\hat{\theta}_n - \theta_0\|^\omega = o_p(n^{-1/2})$. *Then, for the sieve M-estimate* $\hat{\theta}_n$, $n^{1/2}(f(\hat{\theta}_n) - f(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*})$.

Proof. Consider a local alternative value

$$\tilde{\theta}_n = \hat{\theta}_n + e_n u_n^* \text{ with } u_n^* = \pm \pi_{k_n} v^*.$$

By definition,

$$Q_n\left(\tilde{\theta}_n\right) \geq Q_n\left(\hat{\theta}_n\right) + o_p\left(\varepsilon_n^2\right)$$

So

$$\begin{aligned}
o_p\left(\varepsilon_n^2\right) &\leq Q_n\left(\tilde{\theta}_n\right) - Q_n\left(\hat{\theta}_n\right) = \frac{1}{n}\sum_{i=1}^n \left[l\left(\tilde{\theta}_n, Z_i\right) - l\left(\hat{\theta}_n, Z_i\right) \right] \\
&= \mu_n\left[l(\hat{\theta}_n \pm e_n \pi_{k_n} v^*, Z) - l(\hat{\theta}_n, Z) \right] + K\left(\hat{\theta}_n \pm e_n \pi_{k_n} v^*, \theta_0\right) - K\left(\hat{\theta}_n, \theta_0\right) \\
&\leq \sup_{\{\theta \in \Theta_{k_n} : \|\theta - \theta_0\| \leq \varepsilon_n \log n\}} \mu_n\left[l(\theta \pm e_n \pi_{k_n} v^*, Z) - l(\theta, Z) \right] \\
&\quad + K\left(\hat{\theta}_n \pm e_n \pi_{k_n} v^*, \theta_0\right) - K\left(\hat{\theta}_n, \theta_0\right) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\partial l(\theta_0, Z)}{\partial \theta}[\pm e_n \pi_{k_n} v^*]) + O_p(e_n^2) \pm e_n \langle \hat{\theta}_n - \theta_0, \pi_{k_n} v^* \rangle + o(n^{-1}).
\end{aligned}$$

where the last equality follows from Assumptions 7.4.2 and 7.4.3. As a result,

$$\begin{aligned}
& \sqrt{n}\langle\hat{\theta}_n - \theta_0, \pi_{k_n} v^*\rangle \\
= & -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^*] + o_p(1) \\
= & -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^*] - E\frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^*] \right\} + o_p(1) \text{ by Assumption 7.4.4(ii)} \\
= & -\sqrt{n}\mu_n \left(\frac{\partial l(\theta_0, Z)}{\partial \theta}[\pi_{k_n} v^*] \right) + o_p(1) \\
= & -\sqrt{n}\mu_n \left(\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^*] \right) + o_p(1) \text{ by Assumption 7.4.4(i)} \\
\to & \ {}^d N(0, \sigma^2_{v^*}) \text{ by Assumption 7.4.5}
\end{aligned}$$

Next, we connect $\sqrt{n}(f(\hat{\theta}_n) - f(\theta_0))$ with $\sqrt{n}\langle\hat{\theta}_n - \theta_0, \pi_{k_n} v^*\rangle$:

$$\begin{aligned}
& \sqrt{n}(f(\hat{\theta}_n) - f(\theta_0)) \\
= & \sqrt{n}\frac{\partial f}{\partial \theta_0}\left[\hat{\theta}_n - \theta_0\right] + o_p(1) \text{ [by the assumption } \|\hat{\theta}_n - \theta_0\|^\omega = o_p(n^{-1/2})] \\
= & \sqrt{n}\langle\hat{\theta}_n - \theta_0, v^*\rangle + o_p(1) \text{ [by the RRT]} \\
= & \sqrt{n}\langle\hat{\theta}_n - \theta_0, \pi_{k_n} v^*\rangle + \sqrt{n}\langle\hat{\theta}_n - \theta_0, v^* - \pi_{k_n} v^*\rangle \\
= & \sqrt{n}\langle\hat{\theta}_n - \theta_0, \pi_{k_n} v^*\rangle + o_p(1) \text{ by Assumption 7.4.1(iii)} \\
\to & \ {}^d N(0, \sigma^2_{v^*}).
\end{aligned}$$

■

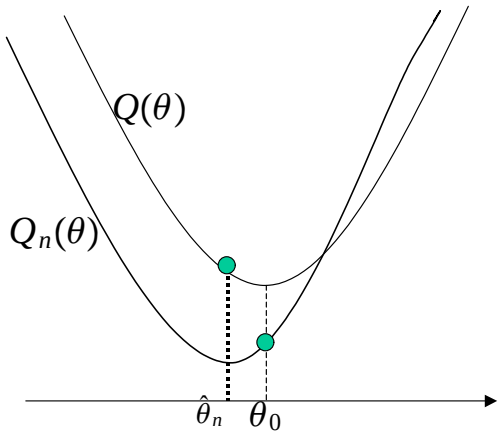
Remark: To prove the theorem, we can also mimic the argument for parametric models. Assuming that θ_n is the exact minimizer of $Q_n(\theta)$ and $Q_n(\theta)$ is twice continuously differentiable. Then $Q_n\left(\hat{\theta}_n + t\pi_{k_n} v^*\right)$ is minimized at $t = 0$. The first order condition is

$$\frac{\partial Q_n\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] = 0$$

assuming an interior solution (When this does not hold, we have to go back to our 'local perturbation' argument).

Under Assumption 7.4.2, we have

$$\left\{ \frac{\partial Q_n\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] - \frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] \right\} - \left\{ \frac{\partial Q_n(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] - \frac{\partial Q(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] \right\} = o_p\left(\frac{1}{\sqrt{n}}\right).$$



Combining the above two equations and using the assumption that $\frac{\partial Q(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] = o_p(1/\sqrt{n})$ yields:

$$\frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] = -\frac{\partial Q_n(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{7.6}$$

But the left hand side is approximately

$$K(\hat{\theta}_n + \pi_{k_n} v^*, \theta_0) - K(\hat{\theta}_n, \theta_0) = \langle \hat{\theta}_n - \theta_0, \pi_{k_n} v^* \rangle + o\left(\frac{1}{\sqrt{n}}\right),$$

so

$$\sqrt{n}\langle \hat{\theta}_n - \theta_0, \pi_{k_n} v^* \rangle = -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Remark: To connect the inner product with the population objective function, we employ the norm[5] $\|\theta - \theta_0\|^2 = 0.5[Q(\theta) - Q(\theta_0)]$ and notice that

$$\begin{aligned}
\langle \hat{\theta}_n - \theta_0, tv^* \rangle &= \left(\left\| \hat{\theta}_n + tv^* - \theta_0 \right\|^2 - \|tv^*\|^2 - \left\| \hat{\theta}_n - \theta_0 \right\|^2 \right)/2 \\
&= \left(Q\left(\hat{\theta}_n + tv^*\right) - Q(\theta_0) - [Q(tv^* + \theta_0) - Q(\theta_0)] - \left[Q(\hat{\theta}_n) - Q(\theta_0) \right] \right) \\
&= Q\left(\hat{\theta}_n + tv^*\right) - Q(\hat{\theta}_n) - [Q(tv^* + \theta_0) - Q(\theta_0)],
\end{aligned}$$

so dividing both sides by t and letting $t \to 0$ yields

$$\langle \hat{\theta}_n - \theta_0, v^* \rangle = \frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[v^*] - \frac{\partial Q(\theta_0)}{\partial \theta}[v^*] = \frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[v^*].$$

[5] This holds only when $Q(\theta) - Q(\theta_0)$ is a norm in the space $B(\theta_0, \delta) - \theta_0$ for some small δ.

Therefore, under assumption 7.4.1(iii) and

$$\frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[v^* - \pi_{k_n} v^*] = o_p\left(\frac{1}{\sqrt{n}}\right),$$

we have

$$\langle \hat{\theta}_n - \theta_0, v^* \rangle = \frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Combining this with equation (7.6), we obtain

$$\begin{aligned}
\langle \hat{\theta}_n - \theta_0, v^* \rangle &= -\frac{\partial Q_n(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&= -\frac{\partial [Q_n(\theta_0) - Q(\theta_0)]}{\partial \theta}[\pi_{k_n} v^*] - \frac{\partial Q(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&= -\mu_n\left(\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^*]\right) + \mu_n\left(\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^* - \pi_{k_n} v^*]\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&= -\mu_n\left(\frac{\partial l(\theta_0, Z)}{\partial \theta}[v^*]\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \qquad (7.7)
\end{aligned}$$

where the last two equalities follow from assumption 7.4.4. The asymptotic normality of $\langle \hat{\theta}_n - \theta_0, v^* \rangle$ now follows from assumption 7.4.5.

Remark: The essential steps in the proof are

$$\begin{aligned}
(i)\ f(\hat{\theta}_n) - f(\theta_0) &= \langle \hat{\theta}_n - \theta_0, v^* \rangle + o_p\left(\frac{1}{\sqrt{n}}\right) \ \text{(linear approximation and RRT)} \\
(ii)\ \langle \hat{\theta}_n - \theta_0, v^* \rangle &= \frac{1}{2}\frac{\partial\left(\|\theta - \theta_0\|^2\right)}{\partial \theta}\Big|_{\theta=\hat{\theta}_n}[v^*] \\
&= \frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] + o\left(\frac{1}{\sqrt{n}}\right) \ \text{(see above)} \\
(iii)\ \frac{\partial Q\left(\hat{\theta}_n\right)}{\partial \theta}[\pi_{k_n} v^*] &= -\frac{\partial Q_n(\theta_0)}{\partial \theta}[\pi_{k_n} v^*] + o_p\left(\frac{1}{\sqrt{n}}\right) \ \text{(SE)}
\end{aligned}$$

Remark: In applications, one needs to specify a Hilbert norm $\|\theta - \theta_0\|$ in order to compute the representer v^*. Wong and Severini (1991) and Shen (1997) have used the Fisher norm, $\|\theta - \theta_0\|^2 = E\{\frac{\partial l(\theta_0, Z_i)}{\partial \theta}[\theta - \theta_0]\}^2$, for the sieve MLE procedure. Ai and Chen (1999, 2003) have introduced a Fisher-like norm for their sieve MD and sieve GLS procedures. In the next subsection we specialize Theorem 7.4.1 to derive root-n asymptotic normality of parametric parts in a partially additive mean regression model.

7.4.2 Example: Partially Additive Mean Regression

Suppose that the i.i.d. data $\{Y_i, X_i' = (X_{0i}', X_{1i}, ..., X_{qi})\}_{i=1}^n$ is generated according to

$$Y = X_0'\beta_0 + h_{01}(X_1) + ... + h_{0q}(X_q) + u, \quad E[u|X] = 0.$$

Let $\theta_0 = (\beta_0, h_{01}, ..., h_{0q})' \in \Theta = B \times \mathcal{H}$ be the parameters of interests, where B is a compact subset of $\mathcal{R}^{d_\beta}$ and $\mathcal{H}$ is the same as that in Subsection 7.3.2. Since $h_{01}(\cdot)$ can have a constant, we assume that X_0 does not contain the constant regressor, $\dim(X_0) = d_\beta$, $\dim(X_j) = 1$ for $j = 1, ..., q$, $\dim(X) = d_\beta + q$, and $\dim(Y) = 1$. We estimate the regression function

$$\theta_0(X) = X_0'\beta_0 + \sum_{j=1}^q h_{0j}(X_j)$$

by minimizing over $\Theta_{k_n} = B \times \mathcal{H}_n$ the criterion

$$\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n l(\theta, Y_i, X_i), \; l(\theta, Y_i, X_i) = \frac{1}{2}\rho^2(\theta, Z_i)$$

where

$$\rho(\theta, Z) = Y - X_0'\beta - \sum_{j=1}^q h_j(X_j).$$

Let

$$\begin{aligned} \|\theta - \theta_0\|^2 &= 2E[l(\theta, Y_i, X_i)]^2 - 2E[l(\theta_0, Y_i, X_i)]^2 \\ &= E\{X_0'(\beta - \beta_0) + \sum_{j=1}^q [h_j(X_j) - h_{oj}(X_j)]\}^2. \end{aligned}$$

This norm can also be motivated from a linearization of $l(\theta, Y, X)$:

$$\begin{aligned} l(\theta, Y, X) - l(\theta_0, Y, X) &\approx \frac{\partial l(\theta_0, Y, X)}{\partial \theta}[\theta - \theta_0] \\ &= u\left\{X_0'(\beta - \beta_0) + [h_j(X_j) - h_{oj}(X_j)]\right\} \end{aligned}$$

Thus

$$\|\theta - \theta_0\|^2 = E\left(\frac{1}{\sigma_X}\frac{\partial l(\theta_0, Y, X)}{\partial \theta}[\theta - \theta_0]\right)^2$$

where $\sigma_X^2 = var(u|X)$.

Let $\bar{V}$ denote the closure of the linear span of $\Theta - \theta_0$ under the metric $\|\cdot\|$. Then $\left(\bar{V}, \|\cdot\|\right)$ is a Hilbert space with the inner product

$$\langle \theta_1, \theta_2 \rangle = E\left[X_0'\beta_1 + \sum_{j=1}^{q} h_j^1(X_j)\right]\left[X_0'\beta_2 + \sum_{j=1}^{q} h_j^2(X_j)\right].$$

Since both $\theta_1 \in \bar{V}$ and $\theta_2 \in \bar{V}$, the above inner product should be understood to be (with some abuse of notation)

$$\begin{aligned}\langle \theta_1, \theta_2 \rangle &:= \langle \theta_1 - \theta_0, \theta_2 - \theta_0 \rangle \\ &= E\left[X_0'(\beta_1 - \beta_0) + \sum_{j=1}^{q}\left[h_j^1(X_j) - h_{0j}(X_j)\right]\right]\left[X_0'(\beta_2 - \beta_0) + \sum_{j=1}^{q}\left[h_j^2(X_j) - h_{0j}(X_j)\right]\right].\end{aligned}$$

We are interested in obtaining the asymptotic distribution of $\sqrt{n}\lambda'(\hat{\beta}_n - \beta)$ for any given vector λ. That is, $f(\theta) = \lambda'\beta$ is the functional of interest. To find the representer v^*, we compute the norm of the functional $f(\theta)$:

$$\begin{aligned}\|f\|^2 &= \sup_{\{\theta \in \Theta - \theta_0, \|\theta\| > 0\}} \frac{\lambda'\beta\beta'\lambda}{\|\theta\|^2} \\ &= \sup_{\{\theta \in \Theta - \theta_0, \|\theta\| > 0\}} \frac{\lambda'\beta\beta'\lambda}{E\{X_0'\beta + \sum_{j=1}^{q}[h_j(X_j)]\}^2}.\end{aligned}$$

To find the sup, we let

$$h_j(X_j) = -w^j(X_j)'\beta$$

for $d_\beta \times 1$ vector function $w^j(X_j) = \left[w_1^j(X_j), ..., w_{d_\beta}^j(X_j)\right]'$ and write

$$X_0'\beta + \sum_{j=1}^{q}[h_j(X_j)] = \left[X_0 - \sum_{j=1}^{q} w^j(X_j)\right]'\beta := A_\theta'\beta.$$

So

$$\begin{aligned}\|f\| &= \sup_{\{\theta \in \Theta - \theta_0, \|\theta\| > 0\}} \frac{\beta'\lambda\lambda'\beta}{\beta' E\left(A_\theta A_\theta'\right)\beta} = \sup_{\{\theta \in \Theta - \theta_0, \|\theta\| > 0\}} \sup_{\mu : \mu = [E(A_\theta A_\theta')]\beta} \frac{\left[\lambda'\left[E\left(A_\theta A_\theta'\right)\right]^{-1}\mu\right]^2}{\mu'\left[E\left(A_\theta A_\theta'\right)\right]^{-1}\mu} \\ &= \sup_{\{\theta \in \Theta - \theta_0, \|\theta\| > 0\}} \lambda'\left[E\left(A_\theta A_\theta'\right)\right]^{-1}\lambda \text{ (taking } \mu = \lambda).\end{aligned}$$

Hence the optimal β is $\beta^* = [E(A_\theta A_\theta')]^{-1}\lambda$ and the problem now reduces to $\inf_{\{\theta\in\Theta-\theta_0,\|\theta\|>0\}} E(A_\theta A_\theta')$. Let

$$D_{w^*}(X) = X_0 - \sum_{j=1}^{q} w^{*j}(X_j),$$

$$\begin{aligned} W^* &= \begin{pmatrix} w_1^{*1}(X_1) & ... & w_1^{*q}(X_q) \\ ... & ... & ... \\ w_{d_\beta}^{*1}(X_1) & ... & w_{d_\beta}^{*q}(X_q) \end{pmatrix}_{d_\beta\times q} \\ &:= \begin{pmatrix} w^{*1}(X_1), & ... & w^{*q}(X_q) \end{pmatrix} \end{aligned}$$

where $w^{*j}(X_j) \in \mathcal{R}^{d_\beta}$, $j = 1, ..., q$ solves

$$\inf_{w^j, j=1,...,q} E\left(A_\theta A_\theta'\right) = E[(X_0 - \sum_{j=1}^{q} w^j(X_j))(X_0 - \sum_{j=1}^{q} w^j(X_j))']. \tag{7.8}$$

This is equivalent to minimizing $E\left[(X_0 - \sum_{j=1}^q w^j(X_j))'(X_0 - \sum_{j=1}^q w^j(X_j))\right]$, which in turn is equivalent to projecting each element of $X_0 \in \mathcal{R}^{d_\beta}$ onto the closed space spanned by $\sum_{j=1}^q w^j(X_j) \in \mathcal{R}^{d_\beta}$. This is true because we want to minimize $\lambda'[E(A_\theta A_\theta')]^{-1}\lambda$ for *any* λ. To illustrate the idea, consider the case with a scalar X_0. Then $\sum_{j=1}^q w^{*j}(X_j)$ is the conditional mean of X_0 given $(X_1, ..., X_q)$ subject to the constraint that the conditional mean is additively separable.

Taking $\beta^* = \{E[D_{w^*}(X)D_{w^*}'(X)]\}^{-1}\lambda$, $h_j^*(X_j) = -\left[w^{*j}(X_j)\right]'\beta^*$, and $A_\theta^* = D_{w^*}(X)$ yields

$$\|f\| = \lambda'\left\{E[D_{w^*}(X)D_{w^*}'(X)]\right\}^{-1}\lambda.$$

In addition,

$$E[D_{w^*}(X)]h_j(X_j) = 0$$

This holds because W^* is the optimal solution to (7.8).

If $E[D_{w^*}(X)D_{w^*}'(X)]$ is finite and positive definite, then $\|f\|$ is bounded. By Riesz representation theorem:

$$f(\theta) = \langle\theta, v^*\rangle$$

for some $v^* \in \bar{V}$. Let $v_\beta^* = \beta^* = (E\{D_{w^*}(X)D_{w^*}'(X)\})^{-1}\lambda$, a $d_\beta \times 1$ vector;

$$v_h^* = \left(h_1^*, ..., h_q^*\right)' = \left(-\left[w^{*1}(X_1)\right]'\beta^*, ..., -\left[w^{*q}(X_q)\right]'\beta^*\right)' = -(W^*)'v_\beta^*,$$

a $q \times 1$ vector of functions. Then we can let $v^* = \left((v_\beta^*)', (v_h^*)\right)'$. In fact, for $\ell_q = (1, ..., 1)'$, we have

$$\begin{aligned}
& \langle \theta, \left(v_\beta^*, v_h^*\right) \rangle \\
= \; & E[X_0'\beta + \sum_{j=1}^q h_j(X_j)] \left[X_0 v_\beta^* - \ell_q' \left(W^*\right)' v_\beta^*\right] \\
= \; & E[X_0'\beta + \sum_{j=1}^q h_j(X_j)] \left[D_{w^*}'(X) v_\beta^*\right] = E\left[X_0'\beta\right] \left[D_{w^*}(X)' v_\beta^*\right] \\
= \; & E\left\{ D_{w^*}(X) + \sum_{j=1}^q w^{*j}(X_j) \right\}' \beta \left[D_{w^*}'(X) v_\beta^*\right] \\
= \; & E\left\{ \left[D_{w^*}'(X)\right] \beta \left[D_{w^*}'(X) v_\beta^*\right] \right\} = \beta' E\left[D_{w^*}(X) D_{w^*}'(X)\right] v_\beta^* = \lambda'\beta
\end{aligned}$$

and

$$\|v^*\| = E\left\{ \left[D_{w^*}'(X) v_\beta^*\right] \right\}^2 = \lambda' \left\{ E[D_{w^*}(X) D_{w^*}'(X)] \right\}^{-1} \lambda = \|f\|$$

where the first equality follows from the definition of $v^* = \left((v_\beta^*)', (v_h^*)\right)'$. Hence we have verified that v^* is indeed the representer of the linear functional f.

Proposition 7.4.1 *Suppose that Assumption* 7.3.4 *and the followings hold:*
(i) $\beta_0 \in int(B)$*;*
(ii) $\Sigma_0(X) = \sigma^2(X)$ *is positive and bounded;*
(iii) $E[X_0 X_0']$ *is positive definite;* $E[D_{w^*}(X) D_{w^*}'(X)]$ *is positive definite;*
(iv) each element of w^{*j} *belongs to the Hölder space* Λ^{m_j} *with* $m_j > 1/2$ *for* $j = 1, ..., q$*.*
Let $k_{jn} = O(n^{1/(2p_j+1)})$ *for* $j = 1, ..., q$*. Then*

$$n^{1/2}(\widehat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}(0, V_1^{-1} V_2 V_1^{-1})$$

where

$$V_1 = E[D_{w^*}(X) D_{w^*}'(X)],$$

and

$$V_2 = E[D_{w^*}(X) \Sigma_0(X) D_{w^*}'(X)].$$

Proof. Let $\Theta_{k_n} = B \times \mathcal{H}_n$ and $\mathcal{H}_n = \mathcal{H}_n^1 \times \cdots \times \mathcal{H}_n^q$, where $\mathcal{H}_n^j$, $j = 1, 2, ..., q$, are the same as those in Subsection 7.3.2. We now verify each of the assumptions of Theorem 7.4.1.

- **Assumption 7.4.1**

Clearly, Assumption 7.4.1(i) is satisfied with $\frac{\partial f(\theta_0)}{\partial \theta}[\theta - \theta_0] = \lambda'(\beta - \beta_0)$ and $\omega = \infty$. In addition, under Assumption (iii), we have $\|v^*\| < \infty$. By the same proof as that for Proposition 7.3.1, we have $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-p/(2p+1)})$ provided that $p = \min\{p_1, \cdots, p_q\} > 0.5$. This and assumption (iv) imply that

$$\begin{aligned} ||\pi_{k_n} v^* - v^*|| \times \|\hat{\theta}_n - \theta_0\| &= O_p\left[\min\left(k_{jn}^{-p_j}\right) n^{-p/(2p+1)}\right] \\ &= O_p\left\{\min\left[n^{-p_j/(2p_j+1)}\right] n^{-p/(2p+1)}\right\} = O_p\left(n^{-\frac{2p}{2p+1}}\right) = o_p\left(n^{-1/2}\right) \end{aligned}$$

So Assumption 7.4.1(iii) holds.

- **Assumption 7.4.2'**

In the present context, Assumption 7.4.2' states that

$$\mu_n\left(\left\{X_0'[v_\beta^*] + \sum_{j=1}^q [\pi_{k_n} v_{h_j}^*(X_j)]\right\}\left\{X_0'[\beta - \beta_0] + \sum_{j=1}^q [h_j(X_j) - h_{oj}(X_j)]\right\}\right) = o_P(n^{-1/2}),$$

uniformly over $\theta \in \Theta_{k_n}$ with $\|\theta - \theta_0\| \leq O(n^{-p/(2p+1)} \log n)$. Applying theorem 3 in Chen et al. (2003), assumptions (i)-(iv) and Assumption 7.3.4 ($h_j \in \mathcal{H}^j = \Lambda_c^{m_j}$ with $m_j > 1/2$ for all $j = 1, ..., q$) imply Assumption 7.4.2'; also see van der Vaart and Wellner (1996).

- **Assumption 7.4.3'**

Assumption 7.4.3' is trivially satisfied given the definition of the metric $||\cdot||$ and the corresponding inner product. More specifically

$$\begin{aligned} E\frac{\partial l\left(\hat{\theta}_n, Z\right)}{\partial \theta}[\pi_{k_n} v^*] &= -E\left[\left(Y - X_0'\hat{\beta}_n - \sum_{j=1}^q \hat{h}_{nj}(X_j)\right)\left(X_0' v_\beta^* - \ell_q' \pi_{k_n} v^*\right)\right] \\ &= E\left[X_0'\left(\hat{\beta}_n - \beta_0\right) - \sum_{j=1}^q \left(\hat{h}_{nj}(X_j) - h_{0j}(X_j)\right)\right]\left(X_0' v_\beta^* - \ell_q' \pi_{k_n} v^*\right) \\ &= \langle \hat{\theta}_n - \theta_0, \pi_{k_n} v^* \rangle \text{ [by the definition of the inner product]} \end{aligned}$$

- **Assumption 7.4.4'**

Note that

$$\frac{\partial l(\theta_0, Z)}{\partial \theta}[\theta] = \left(X_0'\beta + \sum_{j=1}^q h_j(X_j)\right) u,$$

we have

$$E\left\{\frac{\partial l(\theta_0,Z)}{\partial\theta}[\pi_n v^*]\right\} = E\left\{\left(X_0'[v_\beta^*] + \sum_{j=1}^{q}[\pi_{k_n} v_{h_j}^*(X_j)]\right)u\right\} = 0,$$

hence Condition 7.4.4(ii) is automatically satisfied. Since

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l(\theta_0,Z_i)}{\partial\theta}[\pi_n v^* - v^*] = \frac{1}{n}\sum_{t=1}^{n}\left[\left(X_0'\left(v_\beta^* - \pi_{k_n} v_\beta^*\right) + \ell_q'[\pi_n v_h^* - v_h^*]\right)\right]u,$$

by Chebyshev inequality and Assumptions (ii) and Assumption 7.4.1(iii), we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l(\theta_0,Z_i)}{\partial\theta}[\pi_n v^* - v^*] = o_P(n^{-1/2}),$$

hence Condition 7.4.4(i) is satisfied.

- **Assumption 7.4.5'**

  It is easy to show that

$$\sigma_{v^*}^2 = \lambda' V_1^{-1} V_2 V_1^{-1}\lambda > 0,$$

Condition 7.4.5 is satisfied.

With the verification of the assumptions, we use Theorem 7.4.1 to obtain:

$$\begin{aligned}
n^{1/2}\lambda'(\widehat{\beta}_n - \beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial l(\theta_0,Z)}{\partial\theta}[v^*]\right) + o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i\left[D_{w^*}'(X_i)\right]v_\beta^* + o_p(1)\\
&= \lambda'\left(E\{D_{w^*}(X)D_{w^*}'(X)\}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[D_{w^*}(X_i)\right]u_i + o_p(1)\\
&\to {}^d N(0,\lambda' V_1^{-1}V_2V_1^{-1}\lambda).
\end{aligned}$$

■

Remark 7.4.1 *Notice that for the well-known partially linear regression model* $Y_i = X_{0i}'\beta_0 + h_{01}(X_{1i}) + u_i$, $E[u_i|X_i] = 0$, *we can explicitly solve for* $D_{w^*}(X)' \equiv X_0 - w^{*1}(X_1)$ *with* $w^{*1}(X_1) = E\{X_0|X_1\}$. *Hence assumption (iv) will be satisfied if* $E\{X_0|X_1\}$ *is smooth enough.*

Remark 7.4.2 *The proposition can be extended to general cases with conditional moment restriction*

$$E[\rho(Z, \beta_0, h_0(\cdot))|X] = 0,$$

where the difference $\rho(Z, \beta, h(\cdot)) - \rho(Z, \beta_0, h_0(\cdot))$ does depend on the endogenous variables Y. See Chen (2007?) for the sieve GLS estimator when ρ is a vector, which includes our example as a special case.

7.5 Sieve MD Estimation with Endogeneity

In this section, we deal with semiparametric models with endogeneity. See Blundell, R. and J. Powell (2003) for a survey on this problem. Here we follow Ai and Chen (2003) closely (see also Newey and Powell (2003), Hall and Horowitz (2005)). They consider an equation system and design a semiparametric efficient estimator of the parametric part. Here we ignore the efficiency issue, consider only a single equation and assume all variables are scalars. The basic idea can be well illustrated using the simplified setting.

7.5.1 Nonparametric IV

As an example, consider the partial linear model

$$Y_1 = X_1\beta + h(Y_2) + u$$

where Y_2 is endogenous in that $E(u|Y_2) \neq 0$. We assume that there is an instrument X_2 such that

$$E(u|X) = 0 \text{ for } X = (X_1, X_2).$$

In the special case when $\beta_1 = 0$, we have

$$Y_1 = h(Y_2) + u$$

and

$$E\left(Y_1|X\right) = E\left[h\left(Y_2\right)|X\right].$$

This is typically called nonparametric IV regression in the literature.

We assume without the loss of generality that Y_2 and X are in [0,1]. Let $f_{Y_2|X}\left(y|x\right)$ be the conditional pdf of Y_2 given X. Then the identification condition can be written more explicitly as

$$\int_0^1 h\left(y\right) f_{Y_2|X}\left(y|x\right) dy = E\left(Y_1|X = x\right),$$

or equivalently

$$\int_0^1 h\left(y\right) f_{X,Y_2}\left(x, y\right) dy = E\left(Y_1|X = x\right) f_X\left(x\right).$$

Now multiplying both sides by $f_{X,Y_2}(x,w)$ and integrating with respect to x yields[6]:

$$\int_0^1 \left[\int_0^1 h(y) f_{X,Y_2}(x,y)\,dy\right] f_{X,Y_2}(x,w)\,dx = \int_0^1 E(Y_1|X=x) f_X(x) f_{X,Y_2}(x,w)\,dx$$

i.e.

$$\int_0^1 h(y) \left[\int_0^1 f_{X,Y_2}(x,y) f_{X,Y_2}(x,w)\,dx\right] dy = \int_0^1 E(Y_1|X=x) f_X(x) f_{X,Y_2}(x,w)\,dx.$$

Let

$$K(y,w) = \left[\int_0^1 f_{X,Y_2}(x,y) f_{X,Y_2}(x,w)\,dx\right] \text{ and } r(w) = \int_0^1 E(Y_1|X=x) f_X(x) f_{X,Y_2}(x,w)\,dx,$$

Then

$$\int_0^1 h(y) K(y,w)\,dy = r(w),$$

which is a Fredholm equation of the first kind. Under some assumptions, the above equation is equivalent to $E(Y_1|X) = E[h(Y_2)|X]$. So it suffices to solve the above integral equation. Essentially, there is an operator Γ that maps to h into r and we want to recover h or some aspect of it. That is, we want to solve a statistical inverse problem.

To shed some light on this problem, we assume that the conditional pdf $f_{Y_2|X}(y|x)$ and marginal density of $f_X(x)$ are known. In this case, $K(y,w)$ is known for now . Under some conditions, $K(y,w)$ is a reproducing kernel. By Mercer's theorem,

$$K(y,w) = \sum_{j=1}^{\infty} \lambda_j \phi_j(y) \phi_j(w)$$

for some eigenvalue λ_j and associated eigen functions $\phi_j(\cdot)$. Using this representation, we have

$$\int_0^1 h(y) K(y,w)\,dy = \sum_{j=1}^{\infty} \lambda_j \left(\int_0^1 h(y)\phi_j(y)\,dy\right) \phi_j(w).$$

[6]If we do not follow this step, we may assume that $f_{X,Y_2}(x,y) = \sum_{j,k} \lambda_{jk}\phi_j(x)\psi_k(y)$ in $L^2[0,1]^2$. With this, we have

$$\int_0^1 h(y) f_{X,Y_2}(x,y)\,dy = \sum_{j,k} \lambda_{jk}\phi_j(x)\langle h,\psi_k\rangle = \sum_{j=1}^{\infty}\left[\sum_{k=1}^{\infty}\lambda_{jk}\langle h,\psi_k\rangle\right]\phi_j(x),$$

So

$$\sum_{k=1}^{\infty}\lambda_{jk}\langle h,\psi_k\rangle = \langle E(Y_1|X=x) f_X(x), \phi_j(x)\rangle$$

But this is a little harder to deal with.

But

$$r(w) = \sum_{j=1}^{\infty} \mu_j \phi_j(w)$$

for

$$\begin{aligned} \mu_j &= \int_0^1 r(w)\,\phi_j(w)\,dw = \int_0^1 E(Y_1|X=x)\,f_X(x)\,f_{X,Y_2}(x,w)\,dx\phi_j(w)\,dw \\ &= E\left(Y_1 \int_0^1 f_{X,Y_2}(X,w)\,\phi_j(w)\,dw\right). \end{aligned}$$

So $\lambda_j \left(\int_0^1 h(y)\,\phi_j(y)\,dy\right) = \mu_j$, which can be consistently estimated by

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{1i} \int_0^1 f_{X,Y_2}(X_i,w)\,\phi_j(w)\,dw.$$

What we 'observed' is not a noisy measure of $\int_0^1 h(y)\,\phi_j(y)\,dy$ but rather a noisy measure of $\lambda_j \int_0^1 h(y)\,\phi_j(y)\,dy$. Since $\lambda_j \to 0$, the noise will be amplified when we try to recover $\int_0^1 h(y)\,\phi_j(y)\,dy$.

More precisely, to estimate for h, we can solve

$$\arg\min_h \left\| \int_0^1 h(y)\,K(y,w)\,dy - \hat{r}(w) \right\|^2 = \arg\min_h \sum_{j=1}^{\infty} \left\| \lambda_j \left(\int_0^1 h(y)\,\phi_j(y)\,dy\right) - \hat{\mu}_j \right\|^2.$$

The solution is

$$\hat{h}(y) = \sum_{j=1}^{\infty} \frac{\hat{\mu}_j}{\lambda_j} \phi_j(y).$$

Consequently, the estimation error is

$$E\left(\left\|\hat{h}(y) - h(y)\right\|^2\right) = \sum_{j=1}^{\infty} \frac{E\left(\hat{\mu}_j - \mu_j\right)^2}{\lambda_j^2} := \frac{1}{n}\sum_{j=1}^{\infty} \frac{\sigma_j^2}{\lambda_j^2}.$$

Because $\lambda_j \to 0$ as $j \to \infty$, σ_j^2/λ_j^2 does not converge to 0 as $j \to \infty$, except in special cases, and may diverge to ∞. Therefore, except in special cases, the infinite series in the above equation diverges for every n. As a consequence, $E\left(\left\|\hat{h}(y) - h(y)\right\|^2\right) = \infty$ for each n and $\hat{h}(y)$ is not a consistent estimator of $h(y)$.

There are two approaches to overcome the inconsistency problem. The first is based on the Tikhonov regularization. We solve, for some $\{a_j\}$

$$\arg \min_{\langle h,\phi_j\rangle} \sum_{j=1}^{\infty} \left\|\lambda_j \left\langle h,\phi_j\right\rangle - \hat{\mu}_j\right\|^2 + \sum_{j=1}^{\infty} a_j \lambda_j \left\langle h,\phi_j\right\rangle^2$$

where $\sum_{j=1}^{\infty} a_j \left\langle h,\phi_j\right\rangle^2$ is the penalty term. This is equivalent to solving

$$\arg \min_{\langle h,\phi_j\rangle} \left[\lambda_j \left\langle h,\phi_j\right\rangle - \hat{\mu}_j\right]^2 + a_j \left\langle h,\phi_j\right\rangle^2$$

leading to

$$\left\langle h^*,\phi_j\right\rangle = \frac{\hat{\mu}_j}{\lambda_j + a_j}.$$

The regularized estimator is then

$$\hat{h}(y) = \sum_{j=1}^{\infty} \frac{\hat{\mu}_j}{\lambda_j + a_j} \phi_j(y).$$

We choose a_j such that $a_j \to 0$ but not too slowly.

The second is based on the method of sieves. We approximate $h(\cdot)$ by $\sum_{j=1}^{J_n} \left\langle h,\phi_j\right\rangle \phi_j(\cdot)$ for some $J_n \to \infty$ but not too fast. We choose $\left\langle h,\phi_j\right\rangle$ to minimize

$$\arg \min_{\langle h,\phi_j\rangle} \sum_{j=1}^{J_n} \left\|\lambda_j \left\langle h,\phi_j\right\rangle - \hat{\mu}_j\right\|^2,$$

leading to

$$\hat{h}(y) = \sum_{j=1}^{J_n} \frac{\hat{\mu}_j}{\lambda_j} \phi_j(y).$$

Assuming that $E\left(Y_{1i}^2 | X_i\right) \leq \sigma_{\max}^2$ and $f_X(x) \leq c_{\max}$ for constants $\sigma_{\max}^2$ and $c_{\max}$, we have

$$\begin{aligned}
var(\hat{\mu}_j) &= E\left(\hat{\mu}_j - \mu_j\right)^2 \leq \frac{1}{n} E\left\{Y_{1i} \int_0^1 f_{X,Y_2}(X_i, w)\, \phi_j(w)\, dw\right\}^2 \\
&= \frac{1}{n} E\left\{Y_{1i}^2 \left[\int_0^1 f_{X,Y_2}(X_i, w)\, \phi_j(w)\, dw\right]^2\right\} \\
&\leq \frac{\sigma_{\max}^2}{n} \left[\int_0^1 f_{X,Y_2}(x, w)\, \phi_j(w)\, dw\right]^2 f_X(x)\, dx \\
&= \frac{\sigma_{\max}^2 c_{\max}}{n} \int_0^1 \left(\int_0^1 f_{X,Y_2}(x, w_1)\, f_{X,Y_2}(x, w_2)\, dx\right) \phi_j(w_1)\, \phi_j(w_2)\, dw_1 dw_2 \\
&= \frac{\sigma_{\max}^2 c_{\max}}{n} \lambda_j.
\end{aligned}$$

So it is reasonable to assume that $E\left(\hat{\mu}_j - \mu_j\right)^2 = \lambda_j \sigma^2/n$ for all $j \geq 1$ for some $\sigma^2 > 0$. Under this assumption, the variance of the sieve estimator is

$$E\left(\left\|\hat{h}\left(y\right) - E\hat{h}\left(y\right)\right\|_2^2\right) = \sum_{j=1}^{J_n} \frac{E\left(\hat{\mu}_j - \mu_j\right)^2}{\lambda_j^2} = \frac{\sigma^2}{n}\sum_{j=1}^{J_n}\frac{1}{\lambda_j}$$

and the squared bias of the estimator is

$$\left\|h(y) - E\hat{h}\left(y\right)\right\|_2^2 = \sum_{j=J_n+1}^{\infty}\left\langle h, \phi_j\right\rangle^2.$$

If $\lambda_j \geq C_\lambda j^{-\alpha}$ and $\left|\left\langle h, \phi_j\right\rangle\right| \leq C_h j^{-\beta}$. Then

$$\begin{aligned} E\left(\left\|\hat{h}\left(y\right) - h\left(y\right)\right\|_2^2\right) &\leq C\left(\frac{\sigma^2}{n}\sum_{j=1}^{J_n} j^{\alpha} + \sum_{j=J_n+1}^{\infty} j^{-2\beta}\right) \\ &\leq C\left(\frac{\sigma^2}{n}J_n^{\alpha+1} + J_n^{-2\beta+1}\right). \end{aligned}$$

Let

$$J_n^{-2\beta+1} \sim \frac{J_n^{\alpha+1}}{n} \text{ or } n = J_n^{\alpha+2\beta},$$

we have

$$E\left(\left\|\hat{h}\left(y\right) - h\left(y\right)\right\|_2^2\right) = O\left(n^{-\rho}\right)$$

for

$$\rho = \frac{(2\beta - 1)}{\alpha + 2\beta}.$$

The above presentation gives some intuition of the ill-posedness problem. Of course, we do not know the $f_{Y_2|X}\left(\cdot|\cdot\right)$ and $f_X\left(\cdot\right)$. The next section gives a feasible implementation of the method of sieves.

7.5.2 Sieve MD Estimator

We consider a general conditional moment restriction. Let $m(X,\theta) \equiv E[\rho(Z,\theta)|X]$ where $\rho(Z,\theta)$ is a function of $Z = (X_i, Y_i)$ and θ, which may contain infinite dimensional parameters. In general, $\rho(Z,\theta)$ can be a vector of functions but we focus on the scalar case without losing any essential point. We assume that $m(X,\theta) = 0$ for almost all X if and only if $\theta = \theta_0$. For example, in the partial linear model given above, $Z = (X_1, X_2, Y_1, Y_2)$ and

$$\rho(Z,\theta) = \rho(Z,\beta,h(\cdot)) = Y_1 - X_1\beta - h(Y_2).$$

The true parameter θ_0 is identified as the unique minimizer of

$$Q(\theta) = E[m^2(X, \theta)].$$

The sieve simultaneous MD procedure jointly estimates β_0 and h_0 by minimizing a sample quadratic form

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} [\hat{m}(X_i, \theta)]^2$$

over the sieve parameter space $\Theta_{k_n} = B \times \mathcal{H}_{k_n}$, where $\hat{m}(X_i, \theta)$ is any nonparametric estimator of the conditional mean function $m(X, \theta) \equiv E[\rho(Z, \theta)|X]$. That is

$$Q_n\left(\hat{\theta}_n\right) \leq \inf_{\theta \in \Theta_{k_n}} Q_n(\theta) + O(\eta_n).$$

To improve the efficiency in the presence of heteroscedasticity, we may define $Q_n(\theta)$ as the weighted sum $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} [\hat{m}(X_i, \theta)]^2 / \sigma^2(X_i)$ where $\sigma^2(X) = var[\rho(Z, \theta)|X]$. For simplicity, we do not pursue this here.

The conditional expectation $m(X_i, \theta)$ can be estimated using any of the kernel methods or series methods. As in Ai and Chen (2003), we employ a series LS estimator of $m(\cdot, \theta)$. Let $\{p_{0j}(X), j = 1, 2, ..., J_n\}$ be a sequence of known basis functions that can approximate any real-valued square integrable functions of X as $J_n \to \infty$, $p^{J_n}(X) = (p_{01}(X), ..., p_{0J_n}(X))'$ and $P = (p^{J_n}(X_1), ..., p^{J_n}(X_n))'$. For any given θ, the series estimator of m is

$$\hat{m}(X, \theta) = \left[p^{J_n}(X)\right]' \left(P'P\right)^{-1} \sum_{i=1}^{n} p^{J_n}(X_i)\rho(Z_i, \theta).$$

In the partial linear model, $\hat{m}(X_i, \theta) = \hat{E}(Y_1|X) - X_1\beta - \hat{E}(h(Y_2)|X)$. The minimization problem becomes

$$\begin{aligned} Q_n\left(\hat{\theta}_n\right) &\leq \inf_{\theta \in \Theta_{k_n}} \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{E}(Y_{1i}|X_i) - X_{1i}\beta - \hat{E}(h(Y_{2i})|X_i)\right\}^2 + O(\eta_n) \\ &= \inf_{\beta, \gamma} \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{E}(Y_{1i}|X_i) - X_{1i}\beta - \sum_{j=1}^{k_n} \gamma_j \hat{E}(O_j(Y_{2i})|X_i)\right\}^2 + O(\eta_n) \end{aligned}$$

where $O_1(\cdot), ...O_{k_n}(\cdot)$ are basis functions for approximating $h(y_2)$.

Define the criterion function with known $m(X, \theta)$ as

$$Q_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} [m(X_i, \theta)]^2$$

with limit

$$Q(\theta) = E[m(X,\theta)]^2 \text{ and } Q(\theta_0) = 0.$$

Define the metric $\|\cdot\|$ as follows:

$$\|\theta_1 - \theta_2\|^2 = E\left(\frac{dm(X,\theta_0)}{d\theta}[\theta_1 - \theta_2]\right)^2.$$

For this metric, the corresponding inner product is

$$\langle \theta_1, \theta_2 \rangle = E\left\{\frac{dm(X,\theta_0)}{d\theta}[\theta_1]\frac{dm(X,\theta_0)}{d\theta}[\theta_2]\right\}.$$

For the partial linear model example,

$$m(X,\theta) = E(Y_1 - X\beta - h(Y_2)|X) = -X(\beta - \beta_0) - E[(h(Y_2) - h_0(Y_2))|X],$$

$$\frac{dm(X,\theta_0)}{d\theta}[\theta_1 - \theta_2] = -X(\beta_1 - \beta_2) - E[(h_1(Y_2) - h_2(Y_2))|X].$$

So

$$\|\theta_1 - \theta_2\|^2 = E\{X(\beta_1 - \beta_2) + E[(h_1(Y_2) - h_2(Y_2))|X]\}^2.$$

Implicitly, the norm we use to measure the nonparametric part is $E(E[(h_1(Y_2) - h_2(Y_2))|X])^2$, which satisfies

$$\|E[(h_1(Y_2) - h_2(Y_2))|X]\|^2 \leq \|h_1(Y_2) - h_2(Y_2)\|_2^2.$$

Hence, the implied norm for the nonparametric part is weaker than the (weighted) L_2 norm. This is the source of the ill-posedness problem.

We adopt the following definition of ill-posedness. We say the statical inverse problem is *well-posed* if for all sequence $\{\theta_k\}$ in Θ with $Q(\theta_k) - Q(\theta_0) \to 0$ then $d(\theta_k, \theta_0) \to 0$; is *ill-posed* (or *not well-posed*) if there exists a sequence $\{\theta_k\}$ in Θ with $Q(\theta_k) - Q(\theta_0) \to 0$ but $d(\theta_k, \theta_0) \nrightarrow 0$. For a given semi-nonparametric model, suppose the criterion $Q(\theta)$ and the space Θ are chosen such that $Q(\theta)$ is uniquely minimized at θ_0 in Θ. Then whether the problem is ill-posed or well-posed depends on the choice of the pseudo-metric d. This is because different metrics on an infinite-dimensional space Θ may not be equivalent to each other. This is in contrast to the fact that all the norms are equivalent on a finite-dimensional Euclidean space. In particular, it is likely that some standard norm (say $||\theta - \theta_0||_s$) on Θ is not continuous in $Q(\theta) - Q(\theta_0)$ and the problem is ill-posed under $||\cdot||_s$, but there is another pseudo-metric (say $||\theta - \theta_0||_w$) on Θ that is continuous in $Q(\theta) - Q(\theta_0)$, hence the problem becomes well-posed under this $||\cdot||_w$; such a pseudo-metric is typically weaker than $||\cdot||_s$ (i.e., $||\theta - \theta_0||_s \to 0$ implies $||\theta - \theta_0||_w \to 0$).

7.5.3 Consistency

Consistency of the MD estimator follows the arguments similar to those from subsection 7.2.4. Here we define the neighborhood $B(\theta_0, \varepsilon)$ explicitly based on some metric $\|\cdot\|_s$ that may be stronger than $\|\cdot\|$ given above. To help you become familiar with the literature, we use somewhat different proofs for consistency and the rate of convergence.

Assumption 1. (Definition). $\hat{\theta}_n \in \Theta_{k_n}$ and $Q_n(\hat{\theta}_n) \leq \inf_{\theta \in \Theta_{k_n}} Q_n(\theta) + O_p(\eta_{k_n})$ for some $\eta_{k_n} \to 0$ as $n \to \infty$.

Assumption 2. (Identification) $E[\rho(Z, \theta_0)|X] = 0$ and for any $\theta \in (\Theta, \|\cdot\|_s)$, $Q(\theta) = 0$ implies that $\|\theta - \theta_0\|_s = 0$.

Assumption 3. (Sieve Space) $\Theta_{k_n} \subseteq \Theta_{k_n+1} \subseteq \Theta$ for all $k_n \geq 1$; and there exists a sequence $\pi_{k_n}\theta_0 \in \Theta_{k_n}$ such that $\|\pi_{k_n}\theta_0 - \theta_0\|_s \to 0$ as $k_n \to \infty$.

Assumption 4. (Continuity) (i) For each $k_n \geq 1$, $Q(\theta)$ is lower semi-continuous on Θ_{k_n} under the metric $\|\cdot\|_s$; (ii) $\|Q(\pi_{k_n}\theta_0)\| = O\left(\eta_{k_n}\right)$.

Assumption 5. (Compact Sieve Space) The sieve spaces Θ_{k_n} are compact under the metric $\|\cdot\|_s$.

Assumption 6. (Convergence of Q_n). (i) $Q_n(\pi_{k_n}\theta_0) \leq c_0 Q(\pi_{k_n}\theta_0) + O_p(\eta_{k_n})$ for some constant $c_0 > 0$ (ii) $Q_n(\theta) \geq cQ(\theta) - O_p(\eta_{k_n})$ for some $c > 0$ uniformly over Θ_{k_n}.

Proposition 7.5.1 *Denote*

$$g(k_n, \varepsilon) = \inf_{\theta \in \Theta_{k_n} : \|\theta - \theta_0\|_s \geq \varepsilon} Q(\theta)$$

if $\eta_{k_n} = o(g(k_n, \varepsilon))$ *for all* $\varepsilon > 0$, *then* $\left\|\hat{\theta}_n - \theta_0\right\|_s = o_p(1)$.

Proof: Under the assumption that $Q(\theta)$ is lower semi-continuous on Θ_{k_n} on the compact sieve space $(\Theta, \|\cdot\|_s)$, for all $\varepsilon > 0$, $g(k_n, \varepsilon)$ exists and is strictly positive. So

$$\begin{aligned}
& P\left(\left\|\hat{\theta}_n - \theta_0\right\|_s \geq \varepsilon\right) \\
\leq\ & P\left(Q_n(\hat{\theta}_n) \leq Q_n(\pi_{k_n}\theta_0) + O_p(\eta_{k_n}), \left\|\hat{\theta}_n - \theta_0\right\|_s \geq \varepsilon\right) \\
=\ & P\left(cQ(\hat{\theta}_n) \leq c_0 Q(\pi_{k_n}\theta_0) + O_p(\eta_{k_n}), \left\|\hat{\theta}_n - \theta_0\right\|_s \geq \varepsilon\right) \\
=\ & P\left(cg(k_n, \varepsilon) \leq c_0 Q(\pi_{k_n}\theta_0) + O_p(\eta_{k_n})\right) \\
=\ & P\left(g(k_n, \varepsilon) \leq O_p(\eta_{k_n})\right) \to o(1)
\end{aligned}$$

which implies that $\left\|\hat{\theta}_n - \theta_0\right\|_s = o_p(1)$.

Remark: There are two different routes to the consistency and rate of convergence results. One starts with the inequality

$$P\left(\left\|\hat{\theta}_n - \theta_0\right\|_s \geq \varepsilon\right) \leq P\left(Q\left(\hat{\theta}_n\right) - Q(\theta_0) \geq g(k_n, \varepsilon)\right)$$

and the other starts with

$$P\left(\left\|\hat{\theta}_n - \theta_0\right\|_s \geq \varepsilon\right) \leq P\left(Q_n(\hat{\theta}_n) \leq Q_n(\pi_{k_n}\theta_0) + O_p(\eta_{k_n}), \left\|\hat{\theta}_n - \theta_0\right\|_s \geq \varepsilon\right).$$

7.5.4 Rate of Convergence under the Weak Norm

We now make stronger assumptions and prove the rate of convergence under the weak norm $\|\cdot\|$

Assumption 7.5.1 *$\theta_0 \in \Theta$ is the only $\theta \in \Theta$ such that $E[\rho(Z, \theta_0)|X] = 0$ for almost all X.*

Assumption 7.5.2 *(i) Θ_{k_n} is compact under the metric $\|\cdot\|$. (ii) There exists $\pi_{k_n}\theta_0$ such that $\pi_{k_n}\theta_0 \in \Theta_{k_n}$ and $\|\pi_{k_n}\theta_0 - \theta_0\| = o(\varepsilon_n)$ for $\varepsilon_n = n^{-1/4}$.*

Assumption 7.5.3 *The conditional expectation can be estimated with small error in that*
(i) $Q_n(\theta) - Q_n^*(\theta) = O_p\left[\varepsilon_n \log^{-2} n\right]$ uniformly over $\theta \in \Theta_{k_n}$.
(ii) $Q_n(\theta) - Q_n^*(\theta) - [Q_n(\theta_0) - Q_n^*(\theta_0)] = O_p\left[\xi_n \varepsilon_n \log^{-2} n\right]$ uniformly over $\theta \in \Theta_{k_n}$ with $\|\theta - \theta_0\| \leq o(\xi_n)$ where $\xi_n = o(n^{-\tau}), \tau \leq 1/4$.

Assumption 7.5.4 *Let $\hat{\theta}_n^*$ be the minimizer of $Q_n^*(\theta)$ such that*

$$Q_n^*\left(\hat{\theta}_n^*\right) \leq \arg\min_{\theta \in \Theta_{k_n}} Q_n^*(\theta) + O_p(\eta_n)$$

with $\eta_n = o\left(\varepsilon_n^2\right)$, then $\left\|\hat{\theta}_n^* - \theta_0\right\| = o_p(\varepsilon_n)$.

Remarks:

1. Assumption 7.5.3 is a high level assumption. For primitive assumptions, see Ai and Chen (2003). We can write

$$\rho(Z, \theta) = m(X, \theta) + \epsilon$$

with $E(\epsilon|X) = 0$. For a given θ, this reduces to the usual nonparametric regression but we want uniform consistency for all $\theta \in \Theta_{k_n}$. Sufficient conditions for Assumption

7.5.3 are (i) the support $\mathcal{X}$ of X is compact with nonempty interior. (ii) The density of X is bounded above and bounded away from zero. (iii) The unknown function $m(X,\theta) \in \Lambda_c^{\gamma}(\mathcal{X})$ with $\gamma \geq d_x/2$. (iv) $E(\rho(Z,\theta_0)|X)$ is bounded a.e. and $\rho(Z,\theta)$ is Hölder continuous in θ. (v) For any $\delta > 0$, the covering number $N(\delta, \Theta_{k_n}, \|\cdot\|_2)$ satisfies $\log[N(\delta, \Theta_{k_n}, \|\cdot\|_2)] = O(k_n \log(k_n/\delta))$. (vi) some conditions on the bases $p^{J_n}(x)$.

2. The rate of the convergence for the sieve M estimator in assumption 7.5.4 can be obtained using Theorem 7.3.1. One key assumption is that

$$c_1[Q(\theta) - Q(\theta_0)] \leq \|\theta - \theta_0\| \leq c_2[Q(\theta) - Q(\theta_0)].$$

That is, the metric $\|\cdot\|$ we use is equivalent to the metric that is *intrinsic* to our model. With the consistency result, we may only require that the equivalence holds locally. Another crucial assumption is that the sieve space is not too "complex".

Theorem 7.5.1 *Let Assumptions 7.5.1-7.5.4 hold and* $\eta_n = o(\varepsilon_n^2)$, *then* $\left\|\hat{\theta}_n - \theta_0\right\| = o_p(\varepsilon_n)$ *for* $\varepsilon_n = n^{-1/4}$.

Proof. Let $\alpha = 1/4, \xi_{0n} = 1/(n^{\alpha}\log n), \varepsilon_{0n} = \sqrt{\xi_{0n}} = 1/\left(n^{\alpha/2}\sqrt{\log n}\right) = n^{-1/8}(\log n)^{-1/2}$. For any $x \geq 1$,

$$\begin{aligned}
& P(\left\|\hat{\theta}_n - \theta_0\right\| > x\varepsilon_{0n}) \\
= \; & P\left(\min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n(\theta) \leq Q_n(\pi_{k_n}\theta_0) + O_p(\eta_n)\right) \\
\leq \; & P\left(\sup_{\theta\in\Theta_{k_n}} |Q_n(\theta) - Q_n^*(\theta)| > \xi_{0n}\right) \\
& + P\left(\sup_{\theta\in\Theta_{k_n}} |Q_n(\theta) - Q_n^*(\theta)| \leq \xi_{0n}, \min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n(\theta) \leq Q_n(\pi_{k_n}\theta_0) + O_p(\eta_n)\right) \\
: \; & = P_1 + P_2
\end{aligned}$$

Assumption 7.5.3(i) implies that $P_1 \to 0$ as $n \to \infty$. To show $P_2 \to 0$, we note that

$$\begin{aligned}
\min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n(\theta) &= \min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} |Q_n(\theta) - Q_n^*(\theta) + Q_n^*(\theta)| \\
&\geq \min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n^*(\theta) - \min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} |Q_n(\theta) - Q_n^*(\theta)| \\
&\geq \min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n^*(\theta) - \sup_{\theta\in\Theta_{k_n}} |Q_n(\theta) - Q_n^*(\theta)|
\end{aligned}$$

and so

$$\begin{aligned} P_2 \leq\ & P\left(\sup_{\theta\in\Theta_{k_n}}\left|Q_n(\theta)-Q_n^*(\theta)\right|\leq\xi_{0n},\right. \\ & \left.\min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n^*(\theta)-\sup_{\theta\in\Theta_{k_n}}\left|Q_n(\theta)-Q_n^*(\theta)\right|\leq\right. \\ & \left. Q_n(\pi_{k_n}\theta_0)-Q_n^*(\pi_{k_n}\theta_0)+Q_n^*(\pi_{k_n}\theta_0)+O_p(\eta_n)\right) \\ \leq\ & P\left(\min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n^*(\theta)\leq Q_n^*(\pi_{k_n}\theta_0)+2\xi_{0n}+O_p(\eta_n)\right) \\ =\ & P\left(\min_{\|\theta-\theta_0\|>x\varepsilon_{0n}} Q_n^*(\theta)\leq Q_n^*(\pi_{k_n}\theta_0)+O_p\left(\varepsilon_{0n}^2\right)\right)\to 0 \end{aligned}$$

using Assumption 7.5.4. So $\left\|\hat{\theta}_n-\theta_0\right\|=o_p\left(n^{-\alpha/2}\right)$. This rate is not as fast as we want.

Next, we refine the convergence rate by exploiting the local curvature of Q_n around θ_0. Let $\xi_{1n}=1/\left(n^{\alpha+\alpha/2}\log n\right)$, $\varepsilon_{1n}=\sqrt{\xi_{1n}}=O\left(1/\sqrt{n^{\alpha+\alpha/2}\log n}\right)$, we have,

$$\begin{aligned} & P(\left\|\hat{\theta}_n-\theta_0\right\|>x\varepsilon_{1n}) \\ \leq\ & P\left(\inf_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}} Q_n(\theta)\leq Q_n(\pi_{k_n}\theta_0)+O_p(\eta_n)\right)+o(1) \\ \leq\ & P\left(\inf_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}} Q_n(\theta)-Q_n(\theta_0)\leq Q_n(\pi_{k_n}\theta_0)-Q_n(\theta_0)+O_p(\eta_n)\right)+o(1) \\ \leq\ & P\left(\sup_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}}\left|Q_n(\theta)-Q_n(\theta_0)-\left[Q_n^*(\theta)-Q_n^*(\theta_0)\right]\right|>\xi_{1n}\right) \\ & +P\left(\sup_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}}\left|Q_n(\theta)-Q_n(\theta_0)-\left[Q_n^*(\theta)-Q_n^*(\theta_0)\right]\right|<\xi_{1n}\right. \\ & \left.\inf_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}} Q_n(\theta)-Q_n(\theta_0)\leq Q_n(\pi_{k_n}\theta_0)-Q_n(\theta_0)+O_p(\eta_n)\right) \\ :\ =\ & P_3+P_4 \end{aligned}$$

where

$$P_3=P\left(\sup_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}}\left|Q_n(\theta)-Q_n(\theta_0)-\left[Q_n^*(\theta)-Q^*(\theta_0)\right]\right|>\xi_{1n}\right)\to 0 \text{ as } n\to\infty$$

by Assumption 7.5.3(ii), and

$$\begin{aligned} P_4 &= P\left(\inf_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}} Q_n^*(\theta) - Q_n^*(\theta_0) \leq Q_n^*(\pi_{k_n}\theta_0) - Q_n^*(\theta_0) + 2\xi_{1n} + O_p(\eta_n)\right) \\ &= P\left(\inf_{\varepsilon_{0n}\geq\|\theta-\theta_0\|>x\varepsilon_{1n}} Q_n^*(\theta) \leq Q_n^*(\pi_{k_n}\theta_0) + 2\xi_{1n} + O_p(\eta_n)\right) \\ &\to 0 \text{ as } n\to\infty \end{aligned}$$

by Assumption 7.5.4. This proves $\left\|\hat{\theta}_n - \theta_0\right\| = o_p\left(n^{-(\alpha/2+\alpha/4)}\right)$.

Repeating the above proof an infinite number of times, we obtain

$$\left\|\hat{\theta}_n - \theta_0\right\| = o_p\left(n^{-(\alpha/2+\alpha/4+\alpha/8+....)}\right) = o_p(\varepsilon_n)$$

as desired. ■

7.5.5 Asymptotic Normality: the Case of Smooth Functionals

We consider a special smooth functional $f(\theta) = \lambda'\beta$. We follow the same approach as in section 7.4. We first find the representer of the linear functional $f(\theta) = \lambda'\beta$ for any given λ. The norm of f is

$$\begin{aligned} \|f\|^2 &= \sup_{\{\theta\in\Theta:\|\theta-\theta_0\|>0\}} \frac{\lambda'(\beta-\beta_0)(\beta-\beta_0)'\lambda}{E\left(\frac{dm(X,\theta_0)}{d\theta}[\theta_1-\theta_2]\right)^2} \\ &= \sup_{\{\theta\in\Theta:\|\theta-\theta_0\|>0\}} \frac{\lambda'(\beta-\beta_0)(\beta-\beta_0)'\lambda}{E\left(\frac{\partial m(X,\theta_0)}{\partial\beta'}(\beta-\beta_0) + \frac{\partial m(X,\theta_0)}{\partial h}[h-h_0]\right)^2}. \end{aligned}$$

For each component β_j of β, let w_j^* denote the solution to

$$\min_{w_j\in\bar{W}} E\left(\frac{\partial m(X,\theta_0)}{\partial\beta_j} - \frac{\partial m(X,\theta_0)}{\partial h}[w_j]\right)^2$$

and define

$$w^* = \left(w_1^*, ..., w_{d_\beta}^*\right)'$$

$$\frac{\partial m(X,\theta_0)}{\partial h}[w^*] = \left(\frac{\partial m(X,\theta_0)}{\partial h}[w_1^*], ..., \frac{\partial m(X,\theta_0)}{\partial h}[w_{d_\beta}^*]\right)'$$

$$D_{w^*}(X) = \frac{\partial m(X,\theta_0)}{\partial\beta} - \frac{\partial m(X,\theta_0)}{\partial h}[w^*].$$

It is easy to show that

$$\|f\|^2 = \lambda' \left(E \left[D_{w^*}(X) D'_{w^*}(X) \right] \right)^{-1} \lambda.$$

Thus $\|f\|$ is bounded if and only if $E\left[D_{w^*}(X) D'_{w^*}(X)\right]$ is positive definite, in which case we have

$$\lambda'(\beta - \beta_0) = \langle v^*, \theta - \theta_0 \rangle$$

where $v^* = \left(\left(v_\beta^* \right)', (v_h^*)' \right)' \in \bar{V}$,

$$v_\beta^* = \left\{ E \left[D_{w^*}(X) D'_{w^*}(X) \right] \right\}^{-1} \lambda, \; v_h^* = -(w^*)' v_\beta^*.$$

Indeed, by definition

$$\begin{aligned} \frac{dm(X,\theta_0)}{d\theta}[v^*] &= \frac{\partial m(X,\theta_0)}{\partial \beta'} \left[v_\beta^* \right] + \frac{\partial m(X,\theta_0)}{\partial h}[v_h^*] \\ &= \frac{\partial m(X,\theta_0)}{\partial \beta'} \left[v_\beta^* \right] - \frac{\partial m(X,\theta_0)}{\partial h}[(w^*)' v_\beta^*] \\ &= \left(\frac{\partial m(X,\theta_0)}{\partial \beta} - \frac{\partial m(X,\theta_0)}{\partial h}[w^*] \right)' v_\beta^* \\ &= D_{w^*}(X)' v_\beta^* \end{aligned}$$

and

$$\begin{aligned} \langle v^*, \theta - \theta_0 \rangle &= E \left\{ \frac{dm(X,\theta_0)}{d\theta}[v^*] \frac{dm(X,\theta_0)}{d\theta}[\theta - \theta_0] \right\} \\ &= E \left\{ \left(v_\beta^* \right)' D_{w^*}(X) D'_{w^*}(X) (\beta - \beta_0) \right\} = \lambda'(\beta - \beta_0) \end{aligned}$$

where we have used $ED_{w^*}(X) \frac{\partial m(X,\theta_0)}{\partial h}[h] = 0$ for any h.

Given the consistency result, we can focus only on

$$N_{0n} = \left\{ \theta \in \Theta_{k_n} : \|\theta - \theta_0\| = o_p\left(n^{-1/4} \right) \right\}.$$

To exploit the first order condition for $\hat{\theta}_n$, we perturb $\hat{\theta}_n$ by small amount and examine the change of the empirical criterion function. Let $e_n = o\left(1/\sqrt{n}\right), u^* = \pm v^*, u_n^* = \pi_{k_n} u^*$ and $\tilde{\theta}_n = \hat{\theta}_n + e_n u_n^*$. Then with probability approaching one, both $\tilde{\theta}_n \in N_{0n}$ and $\hat{\theta}_n \in N_{0n}$.

Assumption 7.5.5 *Uniformly over* $\theta \in \Theta_{k_n}$, $n^{-1} \sum_{i=1}^n \left(\hat{m}(X_i,\theta) - m(X_i,\theta) \right)^2 = o_p\left(n^{-1/2} \right)$.

Assumption 7.5.6 $Q_n^*\left(\tilde{\theta}_n \right) - Q_n^*\left(\hat{\theta}_n \right) - \left[Q(\tilde{\theta}_n) - Q(\hat{\theta}_n) \right] = o_p\left(e_n/\sqrt{n} \right)$.

Assumption 7.5.7 *For any* $\bar{\theta}_n = \hat{\theta}_n + te_n u_n^*$, $t \in (0,1)$, *we have*

$$\begin{aligned} & E\left(\frac{dm\left(X,\bar{\theta}_n\right)}{d\theta}[u_n^*]\right)\left(\frac{dm\left(X,\bar{\theta}_n\right)}{d\theta}[\hat{\theta}_n - \theta_0]\right) \\ = \; & E\left(\frac{dm\left(X,\theta_0\right)}{d\theta}[u_n^*]\right)\left(\frac{dm\left(X,\theta_0\right)}{d\theta}[\hat{\theta}_n - \theta_0]\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Assumption 7.5.8 *For any* $\bar{\theta}_n = \hat{\theta}_n + te_n u_n^*$, $t \in (0,1)$, *we have*

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i,\bar{\theta}_n)\left(\frac{d\hat{m}\left(X_i,\bar{\theta}_n\right)}{d\theta}[u_n^*] - \frac{dm\left(X_i,\bar{\theta}_n\right)}{d\theta}[u_n^*]\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$$

and

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\frac{dm(X_i,\bar{\theta}_n)}{d\theta}[u_n^*]\left(\hat{m}\left(X_i,\bar{\theta}_n\right) - m(X_i,\bar{\theta}_n)\right) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[u_n^*]\left(\hat{m}\left(X_i,\theta_0\right) - m(X_i,\theta_0)\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

Assumption 7.5.9

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[v_n^* - v^*]\left(\hat{m}\left(X_i,\theta_0\right)\right) = o_p(1)$$

Assumption 7.5.10 $\sqrt{n}\langle v_n^*, \hat{\theta}_n - \theta_0\rangle = \sqrt{n}\langle v^*, \hat{\theta}_n - \theta_0\rangle + o_p(1)$.

Assumption 7.5.11 *(i)* $E[D_{w^*}(X)D'_{w^*}(X)]$ *is positive definite (ii)* $\theta_0 \in int(\Theta)$

Remark: Let

$$g(X_i,\theta_n) = \frac{dm\left(X_i,\hat{\theta}_n\right)}{d\theta}[u_n^*]\hat{m}\left(X_i,\hat{\theta}_n\right)$$

and

$$\mu_n(g(\cdot,\theta)) = \frac{1}{n}\sum_{i=1}^{n}\left[g(X_i,\theta) - Eg(X_i,\theta)\right].$$

Then

$$Q_n^*\left(\tilde{\theta}_n\right) - Q_n^*\left(\hat{\theta}_n\right) - \left[Q(\tilde{\theta}_n) - Q(\hat{\theta}_n)\right] = e_n\mu_n(g\left(\cdot,\bar{\theta}_n\right))$$

So Assumption 7.5.6 holds if $\mu_n(g(\cdot,\theta)) = o_p(1/\sqrt{n})$ uniformly over $\theta \in N_{0n}$.

Theorem 7.5.2 *Let Assumptions 7.5.5-7.5.11 and the assumptions in Theorem 7.5.1 hold, then*

$$n^{1/2}(\widehat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}(0, V_1^{-1} V_2 V_1^{-1})$$

where

$$V_1 = E[D_{w^*}(X) D'_{w^*}(X)],$$
$$V_2 = E[D_{w^*}(X) \Sigma_0(X) D'_{w^*}(X)],$$
$$\Sigma_0(X) = var(\rho(Z, \theta_0)|X).$$

Proof. Define

$$L_n(\theta) = \frac{2}{n} \sum_{i=1}^{n} m(X_i, \theta) \left(\hat{m}\left(X_i, \theta\right) - m(X_i, \theta) \right).$$

Using Assumption 7.5.5, we have

$$\begin{aligned} Q_n\left(\theta\right) &= \frac{1}{n} \sum_{i=1}^{n} \left[m\left(X_i, \theta\right) + \hat{m}(X_i, \theta) - m\left(X_i, \theta\right) \right]^2 \\ &= Q_n^*\left(\theta\right) + L_n(\theta) + \frac{1}{n} \sum_{i=1}^{n} \left(\hat{m}\left(X_i, \theta\right) - m(X_i, \theta) \right)^2 \\ &= Q_n^*\left(\theta\right) + L_n(\theta) + o_p\left(\frac{1}{\sqrt{n}} \right). \end{aligned}$$

Next, we perturb $\hat{\theta}_n$ by small amount and exploit the fact that $\hat{\theta}_n$ is the minimizer. This is a way to use the first order condition. Recall that $e_n = o\left(1/\sqrt{n}\right), u^* = \pm v^*, u_n^* = \pi_{k_n} u^*$ and $\tilde{\theta}_n = \hat{\theta}_n + e_n u_n^*$. When we move from $\hat{\theta}_n$ to $\tilde{\theta}_n$, the change in the population function is:

$$\begin{aligned} & Q(\tilde{\theta}_n) - Q(\hat{\theta}_n) \\ =\ & E\left[m^2\left(X, \hat{\theta}_n + e_n u_n^*\right) - m^2\left(X, \hat{\theta}_n\right) \right] \\ =\ & E\left\{ \left[m\left(X, \hat{\theta}_n\right) + \frac{dm\left(X, \bar{\theta}_n\right)}{d\theta} [e_n u_n^*] \right]^2 - m^2\left(X, \hat{\theta}_n\right) \right\} \\ =\ & 2E\left(\frac{dm\left(X, \bar{\theta}_n\right)}{d\theta} [e_n u_n^*] \right) \left[m(X, \hat{\theta}_n) - m(X, \theta_0) \right] + o_p\left(e_n^2\right) \quad \text{(using } m(X, \theta_0) = 0 \text{ a.s.)} \\ =\ & 2E\left(\frac{dm\left(X, \bar{\theta}_n\right)}{d\theta} [e_n u_n^*] \right) \left(\frac{dm\left(X, \bar{\theta}_n\right)}{d\theta} [\hat{\theta}_n - \theta_0] \right) + o_p\left(e_n^2\right) \\ =\ & 2E\left(\frac{dm\left(X, \theta_0\right)}{d\theta} [e_n u_n^*] \right) \left(\frac{dm\left(X, \theta_0\right)}{d\theta} [\hat{\theta}_n - \theta_0] \right) + o_p\left(\frac{e_n}{\sqrt{n}} \right) \\ =\ & 2\langle e_n u_n^*, \hat{\theta}_n - \theta_0 \rangle + o_p\left(\frac{e_n}{\sqrt{n}} \right) \end{aligned}$$

where $o_p(\cdot)$ terms follow from Assumption 7.5.7. The change of the dominating estimation error in the condition mean is

$$\begin{aligned} L_n(\tilde{\theta}_n) - L_n\left(\hat{\theta}_n\right) &= \frac{2}{n}\sum_{i=1}^{n} m(X_i,\tilde{\theta}_n)\left(\hat{m}\left(X_i,\tilde{\theta}_n\right) - m(X_i,\tilde{\theta}_n)\right) \\ &\quad -\frac{2}{n}\sum_{i=1}^{n} m(X_i,\hat{\theta}_n)\left(\hat{m}\left(X_i,\hat{\theta}_n\right) - m(X_i,\hat{\theta}_n)\right) \\ &= \frac{2e_n}{n}\sum_{i=1}^{n}\frac{dm(X_i,\bar{\theta}_n)}{d\theta}[u_n^*]\left(\hat{m}\left(X_i,\bar{\theta}_n\right) - m(X_i,\bar{\theta}_n)\right) \\ &\quad +\frac{2e_n}{n}\sum_{i=1}^{n} m(X_i,\bar{\theta}_n)\left(\frac{d\hat{m}\left(X_i,\bar{\theta}_n\right)}{d\theta}[u_n^*] - \frac{dm\left(X_i,\bar{\theta}_n\right)}{d\theta}[u_n^*]\right) \\ &= \frac{2e_n}{n}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[u_n^*]\left(\hat{m}\left(X_i,\theta_0\right) - m(X_i,\theta_0)\right) + o_p\left(\frac{e_n}{\sqrt{n}}\right) \end{aligned}$$

using Assumption 7.5.8. In the above derivation, $\bar{\theta}_n$ is a generic value of θ that is between $\hat{\theta}_n$ and $\tilde{\theta}_n$.

Now

$$\begin{aligned} -o_p\left(\varepsilon_n^2\right) &\le Q_n\left(\tilde{\theta}_n\right) - Q_n\left(\hat{\theta}_n\right) = Q_n^*\left(\tilde{\theta}_n\right) - Q_n^*\left(\hat{\theta}_n\right) - \left[Q(\tilde{\theta}_n) - Q(\hat{\theta}_n)\right] \\ &\quad + \left[L(\tilde{\theta}_n) - L\left(\hat{\theta}_n\right)\right] + \left[Q(\tilde{\theta}_n) - Q(\hat{\theta}_n)\right] \\ &= \left[L(\tilde{\theta}_n) - L\left(\hat{\theta}_n\right)\right] + \langle e_n u_n^*, \hat{\theta}_n - \theta_0\rangle + o_p\left(\frac{e_n}{\sqrt{n}}\right). \end{aligned}$$

So

$$\begin{aligned} \sqrt{n}\langle v_n^*, \hat{\theta}_n - \theta_0\rangle &= -\frac{1}{2}\sqrt{n}e_n^{-1}\left[L(\tilde{\theta}_n) - L\left(\hat{\theta}_n\right)\right] + o_p(1) \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[v_n^*]\left(\hat{m}\left(X_i,\theta_0\right) - m(X_i,\theta_0)\right) + o_p(1) \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[v^*]\left(\hat{m}\left(X_i,\theta_0\right) - m(X_i,\theta_0)\right) \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[v^*]\rho(Z_i,\theta_0) + o_p(1). \end{aligned}$$

Therefore, by assumption 7.5.10

$$\begin{aligned}
\sqrt{n}\lambda'\left(\hat{\beta}_n-\beta\right) &= \sqrt{n}\langle v^*,\hat{\theta}_n-\theta_0\rangle+o_p(1)=\sqrt{n}\langle v_n^*,\hat{\theta}_n-\theta_0\rangle+o_p(1)\\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[v^*]\rho(Z_i,\theta_0)+o_p(1)\\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_\beta^{*\prime}D_{w^*}(X)\,\rho(Z_i,\theta_0)+o_p(1)\\
&= \lambda'\left\{E\left[D_{w^*}(X)\,D_{w^*}'(X)\right]\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_{w^*}(X_i)\rho(Z_i,\theta_0)+o_p(1)\\
&\to {}^d\mathcal{N}(0,\lambda'V_1^{-1}V_2V_1^{-1}\lambda).
\end{aligned}$$

■

Remark: As before, we can exploit the first order condition directly. Ignoring the higher order term for the conditional mean estimation, we have

$$\frac{\partial\left[Q_n^*(\theta_n)+L_n(\theta_n)\right]}{\partial\theta}[v_n^*]=0.$$

Using the same argument as in the previous section, we have

$$\frac{\partial\left[Q\left(\hat{\theta}_n\right)+L\left(\hat{\theta}_n\right)\right]}{\partial\theta}[v_n^*]=-\frac{\partial\left[Q_n^*(\theta_0)+L_n(\theta_0)\right]}{\partial\theta}[\pi_{k_n}v^*]+o_p\left(\frac{1}{\sqrt{n}}\right).$$

In view of

$$\frac{\partial Q\left(\hat{\theta}_n\right)}{\partial\theta}[v_n^*]=\langle\hat{\theta}_n-\theta_0,v_n^*\rangle,$$

we obtain

$$\begin{aligned}
\langle\hat{\theta}_n-\theta_0,v_n^*\rangle &= -\frac{\partial L_n(\theta_0)}{\partial\theta}[\pi_{k_n}v^*]+o_p\left(\frac{1}{\sqrt{n}}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{dm(X_i,\theta_0)}{d\theta}[v_n^*]\left(\hat{m}(X_i,\theta_0)-m(X_i,\theta_0)\right)+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

7.5.6 Asymptotic Normality: the Case of Nonsmooth Functionals

7.6 Problems

1. Consider the partially additive IV mean regression

$$Y_1=X_1\beta_0+h_{01}(Y_2)+h_{02}(X_2)+U$$

with $\beta_0 = 1$,

$$h_{01}(Y_2) = 1/[1 + \exp\{-Y_2\}]$$

and

$$h_{02}(X_2) = \log(1 + X_2).$$

We assume that Y_2 is endogenous and

$$Y_2 = X_1 + X_2 + X_3 + R \times U + e$$

with either $R = 0.9$ (strong correlation) or 0.1 (weak correlation). Suppose that the regressors X_1, X_2, X_3 are independent and uniformly distributed over $[0, 1]$, and that e is independent of (X, U) and normally distributed with mean zero and variance 0.1. Conditional on $X = (X_1, X_2, X_3)'$, U is normally distributed with mean zero and variance $(X_1^2 + X_2^2 + X_3^2)/3$. Let $Z = (Y_1, Y_2, X_1, X_2, X_3)'$.

(i) Generate a random sample of $n = 1000$ data $\{Z_i\}_{i=1}^n$ from this design.

(ii) Given the sample $\{Z_i\}_{i=1}^n$, use the sieve MD estimator to estimate $\theta_0 = (\beta_0, h_{01}, h_{02})$. Note that the conditional moment conditions are

$$E[Y_{1i} - \{X_{1i}\beta_0 + h_{01}(Y_{2i}) + h_{02}(X_{2i})\}|X_i] = 0. \tag{7.9}$$

Please take $\Theta_n = B \times \mathcal{H}_{1n} \times \mathcal{H}_{2n}$ as the sieve space, where

$$\mathcal{H}_{1n} = \left\{ h_1(y_2) = \Pi_1' B^{k_{1,n}}(y_2) : \int [D^2 h_1(y_2)]^2 dy_2 \le c_1 \log n \right\},$$

and

$$\mathcal{H}_{2n} = \left\{ h_2(x_2) = \Pi_2' B^{k_{2,n}}(x_2) : \int [D^2 h_2(x_2)]^2 dx_2 \le c_2 \log n,\ h_2(0.5) = \log(3/2) \right\}.$$

In the above specification, $B^{k_{1,n}}(\cdot)$ and $B^{k_{2,n}}(\cdot)$ are either a polynomial spline basis with equally spaced (according to empirical quantile of Y_2 or X_2) knots, or a Hermite polynomial basis. Set $k_{1n} = k_{2n} = 5$.

As an illustration, we consider the sieve MD estimation and the series LS estimator as the $\widehat{m}(X, \theta)$ for the conditional mean function $E[\rho(Z, \theta)|X]$, thus the criterion becomes

$$\min_{\beta \in B, h_1 \in \mathcal{H}_{1n}, h_2 \in \mathcal{H}_{2n}} \frac{1}{n} \sum_{i=1}^n \{\widehat{m}(X_i, \theta)\}^2, \quad \text{with}$$

$$\widehat{m}(X, \theta) = \sum_{j=1}^n [Y_{1j} - \{X_{1j}\beta + h_1(Y_{2j}) + h_2(X_{2j})\}] p^{k_{m,n}}(X_j)'(P'P)^- p^{k_{m,n}}(X),$$

where $p^{k_{m,n}}(X)$ is taken to be the 4th degree polynomial spline sieve, with basis $\{1,\ X_1,\ X_1^2,\ X_1^3,\ X_1^4,\ [\max(X_1-0.5,0)]^4,\ X_2,\ X_2^2,\ X_2^3,\ X_2^4,\ [\max(X_2-0.5,0)]^4,\ X_3,\ X_3^2,\ X_3^3,\ X_3^4,\ [\max(X_3-0.1,0)]^4,\ [\max(X_3-0.25,0)]^4,\ [\max(X_3-0.5,0)]^4,\ [\max(X_3-0.75,0)]^4,\ [\max(X_3-0.90,0)]^4,\ X_1X_3,\ X_2X_3,\ X_1[\max(X_3-0.25,0)]^4,\ X_2[\max(X_3-0.25,0)]^4,\ X_1[\max(X_3-0.75,0)]^4,\ X_2[\max(X_3-0.75,0)]^4\}$. We note that the above criterion is equivalent to a constrained 2 Stage Least Squares (2SLS) with $k_{m,n}=26$ instruments and $\dim(\Theta_n)=1+k_{1,n}+k_{2,n}$ $(<k_{m,n})$ unknown parameters:

$$\min_{\beta\in B,h_1\in\mathcal{H}_{1n},h_2\in\mathcal{H}_{2n}}[\mathbf{Y}_1-\mathbf{X}_1\beta-\mathbf{B}\Pi]'P(P'P)^-P'[\mathbf{Y}_1-\mathbf{X}_1\beta-\mathbf{B}\Pi],$$

where

$$\mathbf{Y}_1=(Y_{11},...,Y_{1n})',\mathbf{X}_1=(X_{11},...,X_{1n})',\Pi=(\Pi_1',\Pi_2')',$$
$$\mathbf{B}_1=(B^{k_{1,n}}(Y_{21}),...,B^{k_{1,n}}(Y_{2n}))',\mathbf{B}_2=(B^{k_{2,n}}(X_{21}),...,B^{k_{2,n}}(X_{2n}))',$$

and $\mathbf{B}=(\mathbf{B}_1',\mathbf{B}_2')'$.

Since $\rho(Z,\theta)$ is linear in $\theta=(\beta,h_1,h_2)'$, the joint sieve MD estimation is equivalent to the profile sieve MD estimation for this model. We can first compute a profile sieve estimator for $h_1(y_2)+h_2(x_2)$. That is, for any fixed β, we compute the sieve coefficients Π by minimizing $\sum_{i=1}^n\{\widehat{m}(X_i,\theta)\}^2$ subject to the smoothness constraints imposed on the functions h_1 and h_2:

$$\min_{\Pi:\int[D^2h_\ell(y)]^2dy\le c_\ell\log n,\ell=1,2}[\mathbf{Y}_1-\mathbf{X}_1\beta-\mathbf{B}\Pi]'P(P'P)^-P'[\mathbf{Y}_1-\mathbf{X}_1\beta-\mathbf{B}\Pi] \tag{7.10}$$

for some upper bounds $c_\ell>0,\ell=1,2$. Let $\widetilde{\Pi}(\beta)$ be the solution to (7.10) and $\widetilde{h}_1(y_2;\beta)+\widetilde{h}_2(x_2;\beta)=(B^{k_{1,n}}(y_2)',B^{k_{2,n}}(x_2)')\widetilde{\Pi}(\beta)$ be the profile sieve estimator of $h_1(y_2)+h_2(x_2)$. Next, we estimate β by $\widehat{\beta}_{iv}$ which solves the following 2SLS problem:

$$\min_\beta[\mathbf{Y}_1-\mathbf{X}_1\beta-\mathbf{B}\widetilde{\Pi}(\beta)]'P(P'P)^-P'[\mathbf{Y}_1-\mathbf{X}_1\beta-\mathbf{B}\widetilde{\Pi}(\beta)]. \tag{7.11}$$

Finally we estimate $h_{o1}(y_2)+h_{o2}(x_2)$ by

$$\widehat{h}_1(y_2)+\widehat{h}_2(x_2)=(B^{k_{1,n}}(y_2)',B^{k_{2,n}}(x_2)')\widetilde{\Pi}(\hat{\beta}_{iv}),$$

and then estimate h_{o1} and h_{o2} by imposing the location constraint $h_2(0.5)=\log(3/2)$:

$$\widehat{h}_{2,iv}(x_2)=B^{k_{2,n}}(x_2)'\widetilde{\Pi}_2(\hat{\beta}_{iv})-B^{k_{2,n}}(0.5)'\widetilde{\Pi}_2(\hat{\beta}_{iv})+\log(3/2),$$

$$\widehat{h}_{1,iv}(y_2)=B^{k_{1,n}}(y_2)'\widetilde{\Pi}_1(\hat{\beta}_{iv})+B^{k_{2,n}}(0.5)'\widetilde{\Pi}_2(\hat{\beta}_{iv})-\log(3/2).$$

We note that although this model (7.9) belongs to the nasty ill-posed inverse problem, the above profile sieve MD procedure is very easy to compute, and in fact, $\hat{\beta}_{iv}$ and $\widetilde{\Pi}(\hat{\beta}_{iv})$ have closed form solutions. To see this, we note that (7.10) is equivalent to

$$\min_{\Pi,\lambda_\ell} (\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi)' P(P'P)^- P' (\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi) + \sum_{\ell=1}^{2} \lambda_\ell \{\Pi_\ell' C_\ell \Pi_\ell - c_\ell \log n\},$$

where for $\ell = 1, 2$, $C_\ell = \int [D^2 B^{k_{\ell,n}}(y)][D^2 B^{k_{\ell,n}}(y)]' dy$, $\Pi_\ell' C_\ell \Pi_\ell = \int [D^2 h_\ell(y)]^2 dy$ and $\lambda_\ell \geq 0$ is the Lagrange multiplier. However, we do not want to specify the upper bounds $c_\ell > 0, \ell = 1, 2$, instead we choose some small values as the penalization weights λ_1, λ_2, and solve the following problems:

$$\min_{\Pi} (\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi)' P(P'P)^- P' (\mathbf{Y}_1 - \mathbf{X}_1\beta - \mathbf{B}\Pi) + \sum_{\ell=1}^{2} \lambda_\ell \Pi_\ell' C_\ell \Pi_\ell \tag{7.12}$$

Denote

$$C(\lambda_1, \lambda_2) = \begin{bmatrix} \lambda_1 C_1 & 0 \\ 0 & \lambda_2 C_2 \end{bmatrix}$$

as the smoothness penalization matrix. The minimization problem (7.12) has a simple closed form solution:

$$\widetilde{\Pi}(\beta) = \left(\mathbf{B}' P(P'P)^- P'\mathbf{B} + C(\lambda_1, \lambda_2)\right)^- \mathbf{B}' P(P'P)^- P'[\mathbf{Y}_1 - \mathbf{X}_1\beta] = W[\mathbf{Y}_1 - \mathbf{X}_1\beta],$$

with $W = (\mathbf{B}' P(P'P)^- P'\mathbf{B} + C(\lambda_1, \lambda_2))^- \mathbf{B}' P(P'P)^- P'$. Substituting the solution $\widetilde{\Pi}(\beta)$ into the 2SLS problem (7.11), we obtain

$$\hat{\beta}_{iv} = [\mathbf{X}_1'(I - \mathbf{B}W)' P(P'P)^- P'(I - \mathbf{B}W)\mathbf{X}_1]^{-1} \mathbf{X}_1'(I - \mathbf{B}W)' P(P'P)^- P'(I - \mathbf{B}W)\mathbf{Y}_1,$$

and $\widetilde{\Pi}(\hat{\beta}_{iv}) = W[\mathbf{Y}_1 - \mathbf{X}_1\hat{\beta}_{iv}]$. Set the penalization weights $\lambda_1 = 0.005$ and $\lambda_2 = 0.000$ for simplicity. The estimated coefficients were recorded.

(iii) Repeat (ii) 500 times. Compute the mean (M) and standard error (SE) of the β_0 estimator across the 500 simulations.

(iv) Repeat (ii)-(iii) but use the naive sieve estimator that ignores the endogeneity. Compare the results with those obtained in (iii).

7.7 References

1. Ai, C., and X. Chen (2003) "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica*, 71, 1795-1843.

2. Andrews, D. (1994) "Empirical process method in econometrics", in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

3. Birgé, L., and P. Massart (1998) "Minimum contrast estimators on sieves: Exponential bounds and rates of convergence", *Bernoulli*, 4 , 329-375

4. Birman, M. and M. Solomjak (1967) "Piece-wise Polynomial Approximations of Functions in the Class W_p^α", *Mathematics of the USSR Sbornik* 73 295-317.

5. Blundell, R. and J. Powell (2003) "Endogeneity in Nonparametric and Semiparametric Regression Models", in M. Dewatripont, L.P. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications*, 2, 312-357, Cambridge: Cambridge University Press.

6. Carrasco, M., J.-P. Florens and E. Renault (2006) "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization", in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6. North-Holland, Amsterdam.

7. Chen, X. (2007): Semiparametric and Nonparametric Estimation via the Method of Sieves. *Handbook of Econometrics* vol. 6

8. Chen, X. and D. Pouzo (2009) "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals", *Journal of Econometrics*. Volume 152, Issue 1, September 2009, Pages 46-60

9. Chen, X. and X. Shen (1998) "Sieve Extremum Estimates for Weakly Dependent Data", *Econometrica*, 66,

10. Chen, X. and H. White (1999) "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators", *IEEE Tran. Information Theory,* 45, 682-691.

11. Florens, J.P. (2003) "Inverse Problems and Structural Econometrics: the Example of Instrumental Variables", in M. Dewatripont, L.P. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications*, 2, 284-311, Cambridge: Cambridge University Press.

12. Gabushin, (1967) "Inequalities for Norms of Functions and their Derivatives in the L_p Metric", *Matematicheskie Zametki*, 1, 291-298.

13. Grenander, U. (1981) *Abstract Inference*, New York: Wiley Series.

14. Hall, P. and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables", *Annals of Statistics*, 33, 2904-2929.

15. Lorentz, G. (1966) *Approximation of functions*, New York: Holt.

16. Newey, W.K. (1997) "Convergence Rates and Asymptotic Normality for Series Estimators", *Journal of Econometrics,* 79, 147-168.

17. Newey, W.K. and D. F. McFadden (1994) "Large sample estimation and hypothesis testing", in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

18. Newey, W.K. and J.L Powell (2003) "Instrumental Variable Estimation of Nonparametric Models", *Econometrica,* 71, 1565-1578. Working paper version, 1989.

19. Ossiander, M. (1987) "A central limit theorem under metric entropy with L_2 bracketing", *The Annals of Probability,* 15, 897-919.

20. Pollard, D. (1984) *Convergence of Statistical Processes.* Springer-Verlag, New York.

21. Shen, X. (1997) "On Methods of Sieves and Penalization", *The Annals of Statistics,* 25, 2555-2591.

22. Shen, X. and W. Wong (1994) "Convergence Rate of Sieve Estimates", *The Annals of Statistics,* 22, 580-615.

23. Van de Geer, S. (1995) "The method of sieves and minimum contrast estimators", *Mathematical Methods of Statistics,* 4 , 20-38.

24. Van de Geer, S. (2000) *Empirical Processes in M-estimation*, Cambridge University Press.

25. Van der Vaart, A. and J. Wellner (1996) *Weak Convergence and Empirical Processes: with Applications to Statistics,* New York: Springer-Verlag.

26. Van der Vaart, A. (1998) *Asymptotic Statistics,* Cambridge

27. Vapnik, V. (1998) *Statistical Learning Theory,* New York: Wiley Interscience.

28. Wong, W.H. and T. Severini (1991) "On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces", *The Annals of Statistics,* 19, 603-632.