Nonparametric Methods in Economics and Finance
Draft Version

Oliver Linton[1]

The London School of Economics and Political Science

April 27, 2006

[1]Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. e-mail: o.linton@lse.ac.uk; web page: http://econ.lse.ac.uk/staff/olinton/

Contents

Chapter 1

Notation

Define for any vector $\nu = (\nu_1, \ldots, \nu_d)^\top$ and function f: $\mathbb{R}^d \to \mathbb{R}$

$$D^\nu f(x) = \frac{\partial^{|\nu|} f(x)}{\partial x_1^{\nu_1} \ \cdots \ \partial x_d^{\nu_d}} \text{ with } |\nu| = \sum_{j=1}^{d} \nu_j.$$

For a function f define the L_2 norm

$$||f||_2^2 = \int f^2(x) dx.$$

Chapter 2

Introduction

These lectures are about nonparametric estimation. Parametric models arise frequently in economics and are of central importance. However, such models only arise when one has imposed specific functional forms on utility or production functions, like Cobb-Douglas or Leontieff. Without these ad hoc assumptions one only gets much milder restrictions on functional form like concavity, symmetry, homogeneity etc. The nonparametric approach is based on the belief that parametric models are usually misspecified and may result in incorrect inferences. By not restricting the functional form one obtains valid inferences for a much larger range of circumstances. In practice, the applicability depends on the sample size and quality of data available.

There are many cases in economics where nonparametric methods have been used and are considered important. Applications arise in cross-section and in time series cases. Many of these applications are routine estimation of nonparametric functions like densities and regression functions. For example, the density of stock returns or the income distribution of a country at a point in time; or of nonparametric regression of one variable on another. In insurance, one is often interested in the conditional hazard function, which represents the probability of dying in a small interval give that you have survived until now. The theory and methods for carrying out such estimation is well understood, and we will spend some time reviewing that in these notes. In more recent times there has been interest in a variety of models with nonparametric components that are not defined as regressions of observable variables and have more complicated structures.

In these lectures, we first discuss a number of such 'nonparametric' models. These include separable functions, program evaluation, estimation of volatility in continuous and discrete time, and

yield curve estimation. In each case there are alternative parametric models, but there is not even close to consensus on which model is correct or appropriate. Given that these models determine the range of outcomes predicted by the model, it makes sense to not limit that wherever possible. We then discuss estimation techniques. Specifically, we focus on nonparametric regression and describe the leading methods of estimation. We then discuss the asymptotic theory for the class of local linear kernel estimators. We give the pointwise asymptotic distribution, the uniform convergence rate and the local functional central limit theorem. Regarding the implementation of these estimators the choice of bandwidth is absolutely central. We define and discuss the main methods for selecting bandwidths in nonparametric models.

Chapter 3

Nonparametric Estimation

3.1 Quantities of Interest

We first discuss the main objects of interest in nonparametric estimation which are quantities derived from the distribution of a population random variable X of interest.

The cumulative c.d.f.

$$F(x) = \Pr(X \leq x).$$

This quantity is quite primitive and many other quantities can be expressed as functionals of F. The c.d.f. is bounded between zero and one and weakly increasing function. The c.d.f. is of interest for many reasons. One application is to testing for stochastic dominance. In that case also it is of interest the integrated c.d.f., $S(x) = \int_{-\infty}^{x} F(x')dx'$, and its further integral $T(x) = \int_{-\infty}^{x} S(x')dx'$ etc.

The density function $f(x)$ is defined as the Radon-Nikodym derivative of the c.d.f., i.e., it satisfies

$$F(x) = \int_{-\infty}^{x} f(x')dx'.$$

When the c.d.f. is differentiable, $f(x) = F'(x)$, but $f(x)$ is defined more generally. The density function is non-negative and integrates to one. [Also of interest are the Fisher information $I(f) = \int (f'/f)^2 f(x)dx$.]

The hazard function

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

is of interest in many areas from mortality to unemployment. The hazard function is non-negative but otherwise unrestricted. We are also interested in the cumulated hazard function

$$\Lambda(x) = \int_{-\infty}^{x} \lambda(x')dx',$$

which is a weakly increasing function. The hazard function and density function are in one to one correspondence so that $f(x) = \lambda(x)\exp(-\Lambda(x))$.

In many cases we have a vector of variables and are interested in the relationship between one variable and the others. This can be generally described by the conditional c.d.f. and density function

$$\begin{aligned} F_{Y|X}(y|x) &= \Pr(Y \leq y|X = x) \\ F_{Y|X}(y|x) &= \int_{-\infty}^{y} f_{Y|X}(y'|x)dy' \end{aligned}$$

as well as the conditional hazard function. The conditional density can be written as the ratio of joint to marginals

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

where the marginal density $f_X(x) = \int f_{Y,X}(y,x)dy$. Also of interest is the symmetric quantity

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

which measures dependence. The equivalent quantity, the conditional characteristic function is of interest too.

The regression function

$$E(Y|X = x) = \int y f_{Y|X}(y|x)dy$$

is an important quantity derived from the conditional density function. Its definition requires that $E(|Y|) < \infty$. The conditional variance $\text{var}(Y|X = x) = E(Y^2|X = x) - E^2(Y|X = x)$ is often of interest in financial applications.

The conditional quantile function is defined to be

$$Q_{Y|X}(\alpha|x) = \inf\left\{\lambda : F_{Y|X}(\lambda|x) \geq \alpha\right\}$$

for $\alpha \in (0,1)$. Lower and upper quantiles. When $F_{Y|X}(\cdot|x)$ is strictly increasing around α, $Q_{Y|X}(\alpha|x) = F^{-1}_{Y|X}(\alpha|x)$. This is defined regardless of moments but does require strict monotonicity for a simple definition.

Note that the regression function and quantile function can be defined as minimizing functionals. Specifically, consider the problem

$$E\left[\{Y-g(X)\}^2\right],$$

where g is any measurable function. Then it follows immediately that $m(x)=E(Y|X=x)$ satisfies

$$E\left[\{Y-g(X)\}^2-\{Y-m(X)\}^2\right]=E\left[\{m(X)-g(X)\}^2\right]\geq 0$$

for all g. One can also prove this result using calculus of variations techniques for the general optimization problem $Q(g)=\int\int\{Y-g(X)\}^2 f(Y,X)dYdX$, where f is the joint density.

Likewise, letting $M(x)=Q_{Y|X}(\alpha|x)$ and

$$Q(g)=E\left[\rho_\alpha(Y-g(X))-\rho_\alpha(Y-M(X))\right],$$

where $\rho_\alpha(u)=u(\alpha-1(u<0)$, we have $Q(M)\leq Q(g)$ for all g.

3.2 C.D.F. and Density Estimation

Suppose that we have an i.i.d. sample $X_1,\ldots,X_n$ drawn from the distribution F. We estimate the c.d.f. by the empirical c.d.f.

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}1\left(X_i\leq x\right).$$

This is a step function with jumps of height $1/n$ (assuming no ties, which happens with probability zero for continuously distributed data). This is CADLAG (Continue A Droite and Limite A Gauche). This estimator

Note that the density function can be interpreted as the derivative of the c.d.f.

$$f(x)=\lim_{h\to 0}\frac{1}{2h}\Pr\left[x-h\leq X_i\leq x+h\right]=\lim_{h\to 0}\frac{1}{2h}E\left[1(x-h\leq X_i\leq x+h)\right].$$

However, the density function cannot be estimated by the derivative of $F_n(x)$, since this is a discontinuous function at the sample points and constant elsewhere. However, a numerical derivative with small h would be

$$\widehat{f}(x)=\frac{1}{2h}\left[F_n(x+h)-F_n(x-h)\right].$$

This can be written in the form

$$\widehat{f}(x)=\frac{1}{2nh}\sum_{i=1}^{n}1\left(|X_i-x|\leq h\right).$$

We define now a more general class of estimators. Let h be a scalar bandwidth and $K(\cdot)$ a kernel satisfying $\int K(u)du = 1$ and $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$. A kernel K is said to be of order q if

$$\int u^j K(u)du = 0,\ j = 1, \ldots, q-1, \text{ and } \int u^q K(u)du < \infty. \tag{3.1}$$

The integrals here are over the support of the kernel which in general is some compact interval or the real line. Frequently, attention is restricted to K a probability density function symmetric about zero for which $q = 2$. In some cases we are interested in so-called boundary kernels that are functions of two arguments $K(u,t)$, where the parameter t controls the support of the kernel, thus $K(u,t)$ has support $[-1,t]$ and satisfies $\int_{-1}^{t} u^j K(u,t)du = 0,\ j = 1, \ldots, q-1$, and $\int_{-1}^{t} u^q K(u,t)du < \infty$ as for regular kernels. Then let

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

This estimate is non-negative and integrates to one in the special case where the support of X is the entire real line. Symmetric kernels. Otherwise not. If there are restrictions on the support of X it may be advisable to use a more complicated kernel that has two or more parameters, see Chen (1999).

Let $\mathcal{K}_h(x) = \mathcal{K}(x/h) = \int K_h(x')dx'$ a smooth increasing c.d.f. and let

$$\widetilde{F}_n(x) = \mathcal{K}_h * F_n = \int \mathcal{K}_h(x-y)dF_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(x - X_i)$$

be a corresponding smoothed estimator of the c.d.f., where $*$ denotes convolution. Then

$$\widehat{f}_h(x) = \widetilde{F}_n'(x) = K_h * F_n = \int K_h(x-y)dF_n(y).$$

This gives another interpretation of the kernel density and c.d.f. estimator: $\widetilde{F}_n(x), \widehat{f}(x)$ are the c.d.f. and density functions respectively of a sample of the random variables $Y_i = X_i + h\varepsilon_i$ conditional on $X_1, \ldots, X_n$, when ε_i has density K.

The tails of the kernel density estimator are like the tails of the kernel K. So for example, if K were standard normal, then the tails of $\widehat{f}(x)$ as $x \to \pm\infty$ behave likewise.

3.3 Regression Estimation

We shall for the most part assume i.i.d. sampling as would be appropriate for cross-sectional data. Also, we will concentrate on the regression problem since this is the central question of most statistical analysis in economics. Suppose that we observe a bivariate dataset $\{Y_i, X_i\}_{i=1}^n$ generated from

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.2}$$

where ϵ_i is a random error independent over observations that satisfies

$$E(\epsilon_i \mid X_i = x) = 0.$$

Then $m(\cdot)$ is the regression function of Y on X. It is usual also to assume that $\text{var}(\epsilon_i \mid X_i = x) = \sigma^2(x) < \infty$. The smoothness of m determines how well it can be estimated.

We discuss a number of estimators of $m(x)$; many of these are linear "smoothers" of the form $\sum_{i=1}^n w_{ni}(x)Y_i$, for some weighting sequence $\{w_{ni}(x)\}_{i=1}^n$ depending only on $X_1, \ldots, X_n$, but arise from different motivations and possess different statistical properties. The methods we consider are appropriate for both *random design*, where (X_i, Y_i) are i.i.d., and *fixed design,* where X_i are fixed in repeated samples. In the random design case, X is an ancillary statistic, 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 shall, for ease of exposition, present results in this form. We also quote most results only for the case where X is scalar, although we discuss the extension to multivariate data. We restrict our attention to independent sampling, but extensions to the dependent sampling case are straightforward.

Smoothing techniques have a long history starting at least in 1857 when the Saxon economist Engel found the law named after him. He analyzed Belgian data on household expenditure, using what we would now call the regressogram. Whittaker (1923) used a graduation method for regression curve estimation which one would now call spline smoothing. Nadaraya (1964) and Watson (1964) provided an extension for general random design based on kernel methods. In time series, Daniell (1946) introduced the smoothed periodogram for consistent estimation of the spectral density. Fix and Hodges (1951) extended this for the estimation of a probability density. Rosenblatt (1956) proved asymptotic consistency of the kernel density estimator. These methods have developed considerably in the last twenty five years, and are now frequently used by applied statisticians. The massive increase in computing power as well as the increased availability of large cross-sectional and high-frequency financial time-series datasets are partly responsible for the popularity of these methods.

3.3.1 Kernel Regression Estimators

Recall that

$$m(x) = \frac{\int y f(x,y)dy}{\int f(x,y)dy}, \tag{3.3}$$

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

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

We have (ignoring the limits of integration):

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

Plugging these into numerator and denominator of (3.3) we obtain the Nadaraya–Watson kernel estimate

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}. \tag{3.4}$$

The bandwidth h determines the degree of smoothness of $\widehat{m}_h$. This can be immediately seen by considering the limits for h tending to zero or to infinity, respectively. Indeed, at an observation X_i,

$$\widehat{m}_h(X_i) \to Y_i, \text{ as } h \to 0,$$

while at an arbitrary point x,

$$\widehat{m}_h(x) \to \frac{1}{n}\sum_{i=1}^{n} Y_i, \text{ as } h \to \infty.$$

These two limit considerations make it clear that the smoothing parameter h in relation to the sample size n should not converge to zero too rapidly nor too slowly. Conditions for consistency of $\widehat{m}_h$ are given in Schuster (1972). Furthermore, the kernel estimator is asymptotically normal, as was first shown in Schuster (1972).

We can interpret the kernel estimator as the minimizer of the local least squares criterion function

$$Q_n(\theta) = \sum_{i=1}^{n} K_h(x - X_i)\,(Y_i - \theta)^2\,,$$

that is, $\widehat{\theta}(x) = \arg\min_{\theta\in\mathbb{R}} Q_n(\theta) = \widehat{m}_h(x)$. The estimator is well-defined when $\sum_{i=1}^n K_h(x - X_i) \neq 0$, which happens with very high probability provided the covariate density is strictly positive at x [if $\sum_{i=1}^n K_h(x-X_i) = 0$, then define $\widehat{m}_h(x) = 0$ for example].[1] The Nadaraya-Watson estimator is linear in Y

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

where $w_{ni}(x) = K_h(x-X_i)/\sum_{i=1}^n K_h(x-X_i)$ depends only on the covariates $X_1, \ldots, X_n$. The weights satisfy $\sum_{i=1}^n w_{ni}(x) = 1$, so that if $Y_i \mapsto a + bY_i$, $\widehat{m}(x) \mapsto a + b\widehat{m}(x)$. When $K \geq 0$, the weights are probability weights since they also satisfy $w_{ni}(x) \in [0, 1]$.

In fact, one can also set-up the global objective function

$$Q_n(\theta(.)) = \int \sum_{i=1}^n K_h(x - X_i) \left(Y_i - \theta(x)\right)^2 d\mu(x),$$

where now the 'parameter' is a function $\theta(.)$ and μ is any positive measure absolutely continuous with respect to Lebesgue measure on the support of X. It can be shown that the function $\widehat{\theta}(\cdot)$ that minimizes this criterion is exactly $\widehat{m}_h(x)$ for each x.

The Nadaraya-Watson estimator can be written

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

We should mention some related estimators. First, in some cases the design density might be known in which case one can use

$$\widehat{m}_h(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\frac{Y_i}{f(x)}. \tag{3.5}$$

[1] If K has support the real line then $\sum_{i=1}^n K_h(x - X_i) \neq 0$. If K has compact support, e.g., $[-1, 1]$, then

$$\begin{aligned}
\Pr\left[\sum_{i=1}^n K_h(x - X_i) = 0\right] &= \Pr\left[\min_{1\leq i\leq n} |X_i - x| > h\right] \\
&= \Pr\left[X_i \notin [x - h, x + h] \text{ for all } i\right] \\
&= \left[1 - \{F(x + h) - F(x - h)\}\right]^n \\
&\simeq \left[1 - 2f(x)h\right]^n,
\end{aligned}$$

provided $f(x) > 0$. Thus this probability is extremely small provided $nh \to \infty$.

These estimators can be called externally normalized since the denominator factors $\widehat{f}(x)$ and $f(x)$ can be taken outside of the sum. An alternative approach is to use internal normalization, hence when f is known

$$\widehat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\frac{Y_i}{f(X_i)}. \tag{3.6}$$

The advantages of this estimator over (3.5) are discussed in Jones, Davies, and Park (1994). When f is not known, Mack and Müller (1989) consider

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

This estimator requires more computation. They are also linear with non-negative weights, provided the kernel is non-negative, but the weights do not necessarily sum to one.

3.3.2 k-Nearest Neighbor Estimators

The kernel estimate was defined as a weighted average of the response variables in a fixed neighborhood of x. The k-nearest neighbor (k-NN) estimate is defined as a weighted average of the response variables in a varying neighborhood. This neighborhood is defined through those X-variables which are among the k-nearest neighbors of a point x.

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

$$\widehat{m}_k(x) = \frac{1}{k}\sum_{i\in\mathcal{N}(x)} Y_i. \tag{3.7}$$

Connections to kernel smoothing can be made by considering (3.7) as a kernel smoother with uniform kernel $K(u) = \frac{1}{2}\mathrm{I}(|u| \le 1)$ and variable bandwidth $h = m(k)$, the distance between x and its furthest k-NN,

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^{n} K_R(x - X_i)\, Y_i}{\sum_{i=1}^{n} K_R(x - X_i)}. \tag{3.8}$$

Note that in (3.8), for this specific kernel, the denominator is equal to k/nR the k-NN density estimate of $f(x)$. Formula (3.8) provides sensible estimators for arbitrary kernels. The bias and variance of this more general k-NN estimator is given in a theorem by Mack (1981). Stute (1984) proves asymptotic normality. In contrast to kernel smoothing, the variance of the k-NN regression

smoother does not depend on f, the density of X. This makes sense since the k-NN estimator always averages over exactly k observations independently of the distribution of the X-variables. The bias constant $B_{nn}(x)$ is also different from the one for kernel estimators given in Theorem 2. An approximate identity between k-NN and kernel smoothers can be obtained by setting

$$k = 2nhf(x), \tag{3.9}$$

or equivalently $h = k/2nf(x)$. For this choice of k or h respectively, the asymptotic mean squared error formulas of Theorem 2 and Theorem 3 are identical. One can also consider the similar estimator

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

where F_n is the covariate empirical distribution. Thus nearest neighbours can be interpreted as kernel smoothing in 'rank space'.

3.3.3 Local Polynomial Estimators

The Nadaraya-Watson estimator can be regarded as the solution of the minimization problem

$$\widehat{m}_h(x) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} K_h(x - X_i) \{Y_i - \theta\}^2 . \tag{3.10}$$

This motivates the local polynomial class of estimators. Let $P_\theta(t) = \theta_0 + \theta_1 t + \ldots + \theta_p t^p/p!$ with $\theta = (\theta_0, \theta_1, \ldots, \theta_p)$ denote a polynomial. Let $\widehat{\theta}_0, \ldots, \widehat{\theta}_p$ minimize

$$\sum_{i=1}^{n} K_h(x - X_i) \{Y_i - P_\theta(X_i - x)\}^2 \tag{3.11}$$

with respect to $\theta \in \mathbb{R}^{p+1}$. Then, $\widehat{\theta}_0$ serves as an estimator of $m(x)$, while $\widehat{\theta}_j$ estimates the $j'th$ derivative of m. A variation on these estimators called $LOWESS$ was first considered in Cleveland (1979) who employed a nearest neighbor window. Fan (1992) establishes an asymptotic approximation for the case where $p = 1$, which he calls the local linear estimator $\widehat{m}_{h,l}(x)$.

The local linear estimator is unbiased when m is linear, while the Nadaraya-Watson estimator may be biased depending on the marginal density of the design. Higher order polynomials can achieve bias reduction, see Fan and Gijbels (1992) and Ruppert and Wand (1992). This class of estimators is perhaps the most popular judged by journal article counts.

3.3.4 Spline Estimators

For any estimate $\widehat{m}$ of m, the residual sum of squares (RSS) is defined as

$$RSS = \sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i)\}^2,$$

which is a widely used criterion, in parametric contexts, for generating estimators of regression functions. However, the RSS is minimized by an $\widehat{m}$ interpolating the data, assuming no ties in the $X's$. To avoid this problem it is necessary to add a stabilizer. Most work is based on the stabilizer

$$\Omega(\widehat{m}) = \int \{\widehat{m}''(u)\}^2 \, du,$$

although see Ansley, Kohn and Wong (1993) and Koenker, Ng and Portnoy (1993) for alternatives. The cubic spline estimator $\widehat{m}_\lambda$ is the (unique) minimizer of

$$R_\lambda(\widehat{m}, m) = \sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i)\}^2 + \lambda \int \{\widehat{m}''(u)\}^2 \, du. \tag{3.12}$$

The spline $\widehat{m}_\lambda$ has the following properties: It is a cubic polynomial between two successive X-values; at the observation points $\widehat{m}_\lambda(\cdot)$ and its first two derivatives are continuous; at the boundary of the observation interval the spline is linear. This characterization of the solution to (3.12) allows the integral term on the right hand side to be replaced by a quadratic form, see Eubank (1988) and Wahba (1990), and computation of the estimator proceeds by standard, although computationally intensive, matrix techniques.

The smoothing parameter λ controls the degree of smoothness of the estimator $\widehat{m}_\lambda$. As $\lambda \to 0$, $\widehat{m}_\lambda$ interpolates the observations, while if $\lambda \to \infty$, $\widehat{m}_\lambda$ tends to a least squares regression line. Although $\widehat{m}_\lambda$ is linear in the Y data, see Härdle (1990, p58-59), its dependency on the design and on the smoothing parameter 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

$$K(u) = \frac{1}{2} \exp\left(-\frac{|u|}{\sqrt{2}}\right) \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right), \tag{3.13}$$

which is of fourth order, since its first three moments are zero, while the equivalent bandwidth $h = h(\lambda; X_i)$ is

$$h(\lambda; X_i) = \lambda^{1/4} n^{-1/4} \; f(X_i)^{-1/4}. \tag{3.14}$$

One advantage of 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 smooth to pass through a particular point, see Jones (1985). Silverman (1985) discusses a Bayesian interpretation of the spline procedure.

3.3.5 Series Estimators

Series estimators have received considerable attention in the econometrics literature, following Elbadawi, Gallant and Souza (1983). This theory is very much tied to the structure of Hilbert space. Suppose that m has an expansion for all x:

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{3.15}$$

in terms of the orthogonal basis functions $\left\{\varphi_j\right\}_{j=0}^{\infty}$ and their coefficients $\left\{\beta_j\right\}_{j=0}^{\infty}$. Suitable basis systems include the *Legendre* polynomials described in Härdle (1990), and the *Fourier* series used in Gallant and Souza (1991).

A simple method of estimating $m(x)$ involves firstly selecting a basis system and a truncation sequence $\tau(n)$, where $\tau(n)$ is an integer less than n, and then regressing Y_i on $\varphi_{ti} = (\varphi_0(X_i), \ldots, \varphi_\tau(X_i))^T$. Let $\{\widehat{\beta}_j\}_{j=0}^{\tau(n)}$ be the least squares "parameter" estimates, then

$$\widehat{m}_\tau(x) = \sum_{j=0}^{\tau(n)} \widehat{\beta}_j \varphi_j(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{3.16}$$

where $W_n(x) = (W_{n1}, \ldots, W_{nn})^T$, with

$$W_n(x) = \varphi_{\tau x}^T (\Phi_\tau^T \Phi_\tau)^{-1} \Phi_\tau^T, \tag{3.17}$$

where $\varphi_{\tau x} = (\varphi_0(x), \ldots, \varphi_\tau(x))^T$ and $\Phi_\tau = (\varphi_{\tau 1}, \ldots, \varphi_{\tau n})^T$.

These estimators are typically very easy to compute. In addition, the extension to additive structures and semiparametric models is convenient, see Andrews and Whang (1990) and Andrews (1991). Finally, series estimators can adapt to the smoothness of m: provided $\tau(n)$ grows at a sufficiently fast rate, the optimal, for the smoothness class of m, rate of convergence can be established − see Stone (1982) − while fixed window order q kernel estimators achieve at best a rate of convergence of $n^{2q/2q+1}$. However, the same effect can be achieved by using a kernel estimator whose order changes

with n in such a way as to produce bias reduction of the desired degree, see Müller (1987). In any case, the evidence of Marron and Wand (1992) cautions against the application of bias reduction techniques unless quite large sample sizes are available. Finally, a major disadvantage with the series method is that there is relatively little theory about how to select the basis system and the smoothing parameter $\tau(n)$.

3.3.6 Higher Dimensions

All the above methods have generalizations to the case where $d > 1$. For example, in the kernel method we can replace the univariate K and h by multivariate kernel $\mathcal{K}$ and bandwidth matrix H, so that we replace $K((x - X_i)/h)$ by $\mathcal{K}(H^{-1/2}(x - X_i))$. A special case of this is where

$$\mathcal{K}(H^{-1/2}(x - X_i)) = K(\|x - X_i\|_H)$$

with $\|A\|_H = [\mathrm{tr}(A^\top H^{-1} A)]^{1/2}$. In practice one does not want to choose an entire matrix of bandwidths without some structure. There are several simplifying approaches. First, let $H = h\Sigma$, where Σ is some fixed symmetric positive definite matrix (in practice estimated from the data) and h is a scalar bandwidth sequence. A second approach is based on 'product kernels' where $\mathcal{K}(x) = K(x_1) \cdots K(x_d)$ and $H = \mathrm{diag}\{h_1, \ldots, h_d\}$, where again h_j reflect the scale of the j'th covariate. In the sequel we will adopt the simplest possible scheme so that we can use the same notation $K_h(x)$ for both univariate and multivariate cases. In the multivariate case, $K_h(x) = K(x/h)/h^d$.

3.3.7 Derivatives

Derivatives can be estimated by differentiating the estimate of m the required number of times. This works provided the estimate of m is itself smooth enough, which can be achieved, for example, by taking K to be smooth like the Gaussian density function. The internal kernel method is particularly convenient for estimation of derivatives because in this case

$$\widehat{m}^{(\nu)}(x) = \frac{1}{nh^{d+|\nu|}} \sum_{i=1}^{n} K^{(\nu)}\left(\frac{x - X_i}{h}\right) \frac{Y_i}{\widehat{f}(X_i)}.$$

The corresponding formula for the Nadaraya-Watson estimator is very complicated.

The local polynomial method explicitly estimates the derivatives - the parameter estimate $\widehat{\theta}_j(x)$ estimates $m^{(j)}(x)$. The problem with this method is just that in high dimensions the number of local

parameters to be estimated is very large. For d dimensions and order p polynomial we have a total of $N = \sum_{\ell=0}^{p} N_\ell$ parameters, where $N_\ell = \begin{pmatrix} \ell + d - 1 \\ d - 1 \end{pmatrix}$.

3.3.8 Some Nonlinear Estimators

The above estimators are all linear in the sense that

$$\widehat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i, \tag{3.18}$$

where $\{w_{ni}(x)\}$ only depend on the covariate $X_1, \ldots, X_n$. We now turn to some nonlinear smoothing methods.

Local Likelihood

The principle underlying the local polynomial estimator can be generalized in a number of ways. Tibshirani (1984) introduced the local likelihood procedure in which an arbitrary parametric regression function $g(x;\theta)$ substitutes the polynomial in (3.11). Fan, Heckman and Wand (1992) develop theory for a nonparametric estimator in a Generalized Linear Model (GLIM) in which, for example, a probit likelihood function replaces the polynomial in (3.11).

Suppose that $f(y|g(x))$ is the density function (or frequency function) of $Y|X$ where f is known and g is an unknown function related to the mean through a known function, i.e., for known h, $h(g(x)) = m(x)$. Then let $\widehat{\theta}_0, \ldots, \widehat{\theta}_p$ minimize

$$\ell_n(\theta) = \sum_{i=1}^{n} K_h(x - X_i) \log f(Y_i | P_\theta(X_i - x)),$$

with respect to $\theta \in \mathbb{R}^{p+1}$. Then, $\widehat{\theta}_0$ serves as an estimator of $g(x)$, while $\widehat{\theta}_j$ estimates the $j'th$ derivative of m, and $h(\widehat{\theta}_0)$ serves as an estimator of $m(x)$. This includes the standard local polynomial estimator as a special case when f is the normal density function. Suppose that Y is binary then $f(y|g(x)) = \Phi(g(x))^y [1 - \Phi(g(x))]^{1-y}$. The advantage of this method is that it imposes the restrictions implied by the data. Fan, Heckman, and Wand (1992) See also Gozalo and Linton (1999).

Local GMM

One can instead have conditional moment restrictions of the form, for some unknown function g

$$E\left[\psi\left(Y, X, g(X)\right)|X = x\right] = 0,$$

where ψ is a vector of moment conditions. For example, suppose that $E(Y|X = x) = m(x)$ and $\text{var}(Y|X = x) = m(x)$. Gagliardini and Gourieroux (), Gozalo and Linton (). Then estimation can proceed by minimizing

$$\|G_n(\theta)\| = \left\| \sum_{i=1}^{n} K_h(x - X_i)\psi\left(Y_i, X_i, P_\theta(X_i - x)\right) \right\|,$$

where $||A||$ is some vector norm, and letting $\widehat{\theta}_0$ serves as an estimator of $g(x)$.

Quantile Regression

To estimate conditional quantiles we use a version of local likelihood. The main difference is that $m(x)$ is not interpreted as the conditional mean any more but some other location parameter. Also, the criterion function need not be smooth. Let $\widehat{m}(x) = \widehat{\theta}_0$, where $\widehat{\theta}$ is any minimizer of the following criterion function

$$\sum_{i=1}^{n} K_h(x - X_i)\rho_\alpha(Y_i - P_\theta(X_i - x))$$

with $\rho_\alpha(u) = u(\alpha - 1(u < 0)$. In general the solution is easy to compute but is not unique so some additional restriction has to be imposed to obtain a well defined solution. Chaudhuri (19?)

Neural Nets

These are really nonlinear series methods or sieves of a certain kind. We minimize the least squares criterion

$$\sum_{i=1}^{n} \{Y_i - \theta(X_i)\}^2 \tag{3.19}$$

over the parameter space

$$\Theta_n = \left\{ \theta : \theta(x) = b_0 + \sum_{j=1}^{m(n)} b_j \psi(a_{j1}x + a_{j0}), \quad \left| \sum_{j=1}^{m(n)} b_j \right| \leq c_{n1}, \quad \max_{1 \leq j \leq m(n)} |a_{j0} + a_{j1}| \leq c_{n2} \right\},$$

i.e., one chooses the parameters $(b_0, \ldots, b_r, a_{10}, \ldots, a_{r0}, a_{11}, \ldots, a_{r1})$ to minimize (3.19). The parameter space increases with sample size through relaxation of the constraints $m(n)$ and c_{nj}. The function ψ is a sigmoid function − a bounded measurable 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 c.d.f. These methods have been widely used in Physics and other natural sciences. Chen and Shen (1998)

Chapter 4

Asymptotic Properties

4.1 CDF and Density Estimation

We first give the classical Glivenko-Cantelli result published in the Italian Actuarial Journal in 1933 [in the same year and journal, Kolmogorov established the limiting distribution of the normalized process].

Theorem 1 *As* $n \to \infty$,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

The only 'condition' is that X_i are i.i.d., although note that since F is a distribution function it has at most a countable number of discontinuities, and is right continuous. Note also that the supremum is over a non-compact set - in much subsequent work generalizing this theorem it has been necessary to restrict attention to compact sets. The proof of Theorem 1 exploits some special structure: specifically that for each x, $1(X_i \leq x)$ is Bernoulli with probability $F(x)$.

Let B be the Brownian Bridge process. This is a Gaussian process with covariance function

$$\text{cov}(B(s), B(t)) = \min\{s, t\} - st$$

for every s, t. We next establish the weak convergence of the empirical c.d.f. i.e., a Functional Central Limit Theorem (FCLT).

Theorem 2 *As* $n \to \infty$,

$$\sqrt{n}\,[F_n(\cdot) - F(\cdot)] \Longrightarrow B(F(\cdot)),$$

where B is the Brownian Bridge process.

The limiting process is a time changed Gaussian process. The result can be established by first establishing the result for uniform $[0,1]$ random variables and then employing the result that $X = F^{-1}(U)$, where U is a uniform random variable. Multivariate case?

We now turn to density estimation. There are several types of consistency results, either pointwise or uniform.

Theorem 3 *(Nadaraya, (1965)) Suppose that K is of bounded variation and integrates to one, that f is uniformly continuous on $\mathbb{R}$, and that $h \to 0$ and $nh^2 \to \infty$. Then*

$$\sup_{x\in\mathbb{R}} \left| \widehat{f}(x) - f(x) \right| \xrightarrow{a.s.} 0$$

This theorem places very weak assumptions on the kernel and density but somewhat stronger conditions on the bandwidth sequence. Note that the convergence is uniform over the entire real line. It is possible to establish uniform consistency of the kernel density estimator under weaker conditions on the bandwidth sequence like $nh/\log n \to \infty$ at the expense of stronger conditions on K.

Theorem 4 *(Silverman (1978)) Suppose that K is uniformly continuous with modulus of continuity w and of bounded variation, that $\int |K(x)|dx < \infty$ and $K(x) \to 0$ as $x \to \infty$, that $\int K(x)dx = 1$, and that $\int |x \log |x||^{1/2} |dK(x)| < \infty$. Suppose that f is uniformly continuous. Then, provided $h \to 0$ and $nh/\log n \to \infty$*

$$\sup_{x\in\mathbb{R}} \left| \widehat{f}(x) - f(x) \right| \xrightarrow{a.s.} 0.$$

Suppose additionally that $\int_0^1 [\log(1/u)]^{1/2} \, d\gamma(u) < \infty$, where $\gamma(u) = \{w(u)\}^{1/2}$ and that $nh(\log n)^{-2}\{\log(1/h)\} \to \infty$ and $\sum_{n=1}^{\infty} h_n^{\lambda} < \infty$ for some λ, then

$$\sup_{x\in\mathbb{R}} \left| \widehat{f}(x) - E[\widehat{f}(x)] \right| = O\left(\sqrt{\frac{\log(1/h)}{nh}} \right) \ a.s.$$

These results use arguments that are special to the univariate case.

The assumption that f is uniformly continuous is innocuous for densities of unbounded support but rather restrictive for those living on a bounded interval. For example, it rules out the uniform density. The assumption is needed for handling the bias term $E[\widehat{f}(x)] - f(x)$ and the results hold true for $\sup_{x\in\mathbb{R}} |\widehat{f}(x) - E[\widehat{f}(x)]|$ without this assumption. Recently, Giné and Guillou (2002) have shown the following

Theorem 5 *(Giné and Guillou (2002)) Suppose that the kernel K is a bounded function of bounded variation. Suppose that $h \to 0$ monotonically, such that $nh^d/|\log h| \to \infty$ and $|\log h|/\log\log n \to \infty$. Suppose further that the density f is bounded. Then*

$$\sup_{x\in\mathbb{R}} \left|\widehat{f}(x) - E[\widehat{f}(x)]\right| = O\left(\sqrt{\frac{\log(1/h)}{nh^d}}\right) \ a.s.$$

This result is quite remarkable in terms of the weakness of the conditions. To establish consistency though we also need to analyze the term $\sup_{x\in\mathbb{R}} |E[\widehat{f}(x)] - f(x)|$. This requires additional conditions. For example, one might assume that f is uniformly continuous like Silverman (1978). To establish a rate one needs stronger conditions specifically smoothness. We note that uniform continuity is an appropriate condition for densities with unbounded support but does rule out many density with compact support for example the uniform density. For those cases different conditions are appropriate. The bias term is handled by making a change of variables. We have

$$E[\widehat{f}(x)] = \int K_h(x - X)f(X)dX$$

and if we transform $X \mapsto u = (x - X)/h$ the integrand becomes $K(u)f(x - uh)du$. If the support of X is $\mathbb{R}$ then the range of integration of u is not affected. Then

$$\int K(u)f(x - uh)du = f(x)\int K(u)du - f'(x)\int K(u)udu + f''(x)\frac{1}{2}\int K(u)u^2du.$$

If the support of X is some interval $[\underline{x}, \overline{x}]$, then the range of integration of u becomes $[(x - \underline{x})/h, (\overline{x} - x)/h]$. When x is an interior point this interval tends towards $(-\infty, \infty)$, but if $x = \underline{x}$, then the interval tends towards $(0, \infty)$.

The limiting distribution of the density estimator at a point.

Theorem 6 *Suppose that K is bounded and satisfies $\int K(u)udu = 0$, $\int K(u)u^2du < \infty$ and $\int |K(u)|du < \infty$. Suppose also that f is twice continuously differentiable at x. Then*

$$\sqrt{nh}\left[\widehat{f}(x) - f(x)\right] \Longrightarrow N(b(x), v(x)),$$

where

$$v(x) = \|K\|_2^2 f(x) \text{ and } b(x) = \left(\lim_{n\to\infty}\sqrt{nh^5}\right)\frac{\mu_2(K)}{2}f''(x).$$

4.2 Regression Function

Stone (1977) gave the following result for linear estimators

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

where $\{w_{ni}(x)\}$ only depend on the covariate $X_1, \ldots, X_n$.

Theorem 7 *(Stone (1977). Let $\{w_{ni}(x)\}$ be a sequence of weights and let $X, X_1, \ldots, X_n$ be i.i.d. Suppose that the following conditions hold*
(1) There is a $C \geq 1$ such that for every nonnegative Borel function f

$$E \sum_{i=1}^{n} w_{ni}(X) f(X_i) \leq C E f(X);$$

(2) There is a $D \geq 1$ such that

$$\Pr\left[\sum_{i=1}^{n} |w_{ni}(X)| \leq D\right] = 1;$$

(3)

$$\sum_{i=1}^{n} |w_{ni}(X)| \, 1\left(||X_i - X|| > a\right) \xrightarrow{P} 0 \text{ for all } a > 0;$$

(4)

$$\sum_{i=1}^{n} |w_{ni}(X)| \xrightarrow{P} 1;$$

(5)

$$\max_{1 \leq i \leq n} |w_{ni}(X)| \xrightarrow{P} 0.$$

Then $\{w_{ni}(x)\}$ are consistent in the sense that whenever $E[|Y|^r] < \infty$,

$$E\left[|\widehat{m}(X) - m(X)|^r\right] \to 0. \tag{4.1}$$

These are quite weak conditions. Note that for many regression estimators the sequence of weights $\{w_{ni}(x)\}$ are probability weights, i.e., they lie between zero and one and sum to one. In this case conditions (1), (2), and (4) are quite natural. A consequence of condition (1) is that $E[|\widehat{m}(X)|] < \infty$ whenever $E[|Y|] < \infty$. Condition (3) is trivially satisfied for kernel estimators with

kernels of bounded support. Condition (5) is also satisfied for many estimators - for nearest neighbor estimators for example it is trivial. Stone (1977) shows that these conditions can be satisfied for a range of estimators. The standard local linear estimator does not satisfy these conditions however. He shows how to modify local linear estimators to make them probability weights and thereby to satisfy the conditions. Kohler (2000) suggests an alternative way of doing this by restricting the optimization in (3.11) to a bounded parameter space (that is allowed to expand slowly with sample size). Stone also shows how to apply these results to nonlinear estimators like conditional quantile estimators.

Devroye and Wagner (1980) showed that the kernel estimator with non-negative bounded and bounded away from zero at the origin and compactly supported kernel K satisfies (4.1) provided only $h \to 0$ and $nh^d \to \infty$.

We will use the standard theory of optimization estimators to structure our results. This theory is summarized in the appendix.

Theorem 8 *(Local Linear). Suppose that:*
(i) The marginal density f of the covariates is continuous at the interior point x and $f(x) > 0$;
(ii) The regression function is twice differentiable and $m''(x)$ is continuous at x; the variance function $\sigma^2(.)$ is continuous and positive at x;
(iii) The kernel K is continuous on its compact support;
(iv) $E(|Y|^{2+\delta}) < \infty$ for some $\delta > 0$;
(v) $h \to 0$ and $nh \to \infty$ such that $\lim_{n\to\infty} nh^5$ exists.
Then

$$\sqrt{nh}\left[\widehat{m}_{LL}(x) - m(x)\right] \Longrightarrow N(b(x), v(x)),$$

where

$$v(x) = \|K\|_2^2 \frac{\sigma^2(x)}{f(x)} \text{ and } b(x) = \left(\lim_{n\to\infty} \sqrt{nh^5}\right) \frac{\mu_2(K)}{2} m''(x).$$

Corollary 9 *(Nadaraya-Watson) Suppose that (i)-(v) above hold and that also $f''(\cdot)$ exists and is continuous at x. Then*

$$\sqrt{nh}\left[\widehat{m}_{NW}(x) - m(x)\right] \Longrightarrow N(b_{NW}(x), v(x)),$$

where

$$b(x) = \left(\lim_{n\to\infty} \sqrt{nh^5}\right) \frac{\mu_2(K)}{2} \left\{(m \cdot f)'' - m \cdot f''\right\}(x).$$

REMARKS.

1. For both estimators the mean squared error for any interior point x is $O(h^4) + O(1/nh)$, and the best rate is given by taking $h \propto n^{-1/5}$ in which case the mean squared error is of order $n^{-4/5}$. This is larger than in the parametric case where the mean squared error declines like n^{-1}.

2. The bias of the NW estimators depends on f and on its derivatives. The local linear estimator by contrast has bias uniformly of order h^2, and is "design adaptive", i.e., its bias only depends on $m''(x)$. Finally, the regularity conditions for the local linear estimator are weaker, since no derivatives of f are needed, just continuity.

3. The asymptotic distribution for both estimators has a bias and is 'nuisance parameter dependent'. To obtain correct confidence intervals we would have to estimate $m''(x), f(x)$ and $\sigma^2(x)$ in the case of the local linear estimator, and $m'(x), m''(x), f(x), f'(x)$, and $\sigma^2(x)$ in the case of the Nadaraya-Watson estimator, which in either case is an even more difficult than the problem we started out with. Therefore, in practice it is usual only to estimate the variance and to argue that the bias is of smaller order [which would be the case if $h = o(n^{-1/5})$]. This is called 'undersmoothing'.

4. Boundary bias. When the evaluation point x is at the boundary or close to the boundary, the Nadaraya-Watson Estimator suffers badly from boundary bias. Specifically, in this case the bias is $O(h)$ for any point x_n that lies within h of the boundary. This is because the change of variables argument we use to obtain the bias no longer applies. The local linear estimator does not suffer so badly in the boundary region, and its bias is the same magnitude as at interior points, namely h^2.

5. Suppose that $f(x) = 0$ or $f(x) = \infty$. Then we may still obtain consistency and asymptotic normality but at slower (faster) rates of convergence, see Hengartner and Linton (1999). Likewise with $\sigma^2(x) = \infty$ or $\sigma^2(x) = 0$.

6. Suppose that $\sigma^2(.)$ and or $f(.)$ is (boundedly) discontinuous at the point x but that m is continuous at x. Then the estimators are still consistent but the asymptotic variance changes to

$$\frac{\sigma^2(x_-)f(x_-)\int_{-\infty}^0 K(u)^2du + \sigma^2(x_+)f(x_+)\int_0^\infty K(u)^2du}{\left(f(x_-)\int_{-\infty}^0 K(u)du + f(x_+)\int_0^\infty K(u)du\right)^2}.$$

Under symmetry of the kernel this simplifies to $[\sigma^2(x_-)f(x_-) + \sigma^2(x_+)f(x_+)]/\left(f(x_-) + f(x_+)\right)^2/2$. Bias? If $m(x)$ is discontinuous at x, then the estimator converges to $(m(x_-) + m(x_+))/2$ under symmetry of the kernel.

7. Can allow the marginal density and error variance (or distribution) to vary with i and n, denoted $f_{ni}(.)$ and $\sigma^2_{ni}(.)$ provided these functions are uniformly smooth etc.. In this case the limiting

variance is

$$\|K\|_2^2 \frac{\frac{1}{n}\sum_{i=1}^n f_{ni}(x)\sigma_{ni}^2(x)}{\left[\frac{1}{n}\sum_{i=1}^n f_{ni}(x)\right]^2}.$$

An extreme example is when $X_i = i/n$, i.e., purely deterministic in which case $f_{ni}(x) \to 1$ as $n \to \infty$. The theory is as above where f can be interpreted as a limiting or average density. This would be called a fixed design. A number of estimators have different properties depending on whether the design is fixed or random but the local linear and local constant have essentially identical properties in these two cases.

8. Suppose that $E[|Y|^2] = \infty$ but $E[|Y|^{1+\alpha}] < \infty$ for some $\alpha \in (0,1)$. Then we still have consistency but with slower rates and possibly non-normal limiting distributions (this is just as in the parametric case). Stute (1986) established the almost sure convergence of the Nadaraya-Watson estimator under only weak moment conditions, namely $E(|Y|^{1+\delta}) < \infty$.

9. The compact support condition on the kernel can be weakened to just K bounded and $\int |K(u)|u^2 du < \infty$.

Some Heuristics We give some heuristics for the local constant approach. Write

$$\begin{aligned}\widehat{m}(x) - m(x) &= \frac{\widehat{r}(x) - m(x)\widehat{f}(x)}{\widehat{f}(x)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\varepsilon_i}{\widehat{f}(x)} + \frac{\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\left[m(X_i) - m(x)\right]}{\widehat{f}(x)}.\end{aligned}$$

We have shown that for each interior point x,

$$\widehat{f}(x) = f(x) + O_p(h^2) + O_p(n^{-1/2}h^{-d/2}).$$

Therefore,

$$\begin{aligned}&\widehat{m}(x) - m(x) \\ &= \frac{1}{f(x)}\left[\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\varepsilon_i + \frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\left[m(X_i) - m(x)\right]\right]\left[1 + O_p(h^2) + O_p(n^{-1/2}h^{-d/2})\right]\end{aligned}$$

assuming that $f(x) > 0$. Writing $Z_{ni}(x) = K_h(x - X_i)\left(m(X_i) - m(x)\right)$ we have

$$\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)\left(m(X_i) - m(x)\right) = E[Z_{ni}(x)] + \frac{1}{n}\sum_{i=1}^n Z_{ni}(x) - E[Z_{ni}(x)].$$

The properties of $E[Z_{ni}(x)]$ follow from a Taylor series expansion. The second term is a sum of mean zero independent and identically distributed random variables (for given n) and has variance

$$\begin{aligned}\text{var}[Z_{ni}(x)] \leq E[Z_{ni}^2(x)] &= h^{-2d} E\left[K\left(\frac{x-X_i}{h}\right)^2 (m(X_i)-m(x))^2\right] \\ &= h^{-d}\int K(u)^2 (m(x+uh)-m(x))^2 f(x+uh)du \\ &= O(h^{2-d}).\end{aligned}$$

It follows that

$$\widehat{m}(x) - m(x) = \frac{1}{f(x)}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\varepsilon_i + E[Z_{ni}(x)]\right][1+o_p(1)].$$

4.2.1 Bias reduction

When the regression function has more derivatives, it is possible to reduce the magnitude of the bias by using higher order polynomials or in the case of the Nadaraya-Watson estimator, higher order kernels, that is, kernels for which $\int u^j K(u)du = 0$, $j = 1, \ldots, q-1$ and $\int u^q K(u)du < \infty$. In either case we achieve bias of order h^q at interior points under corresponding smoothness conditions, in which case the optimal bandwidth is $h \propto n^{-1/(2q+1)}$ and the mean squared error is of order $n^{-2q/(2q+1)}$. However, the variance increases (for given bandwidth). There is no unbiased estimator. Marron and Wand.

4.2.2 Curse of dimensionality

When there are many $X's$, the variance of the Nadaraya-Watson estimator is of order $1/nh^d$, where d is the dimensions and the bandwidth matrix $H = hI_d$. The reason for this high variance is because in high dimensions, observations are more spread out. In this case, the optimal rate of convergence is obtained by setting h^{2q} the same order as $1/nh$, i.e., $h \propto n^{-1/(2q+d)}$, and the resulting mean squared error is of order $n^{-2q/(2q+d)}$.

4.2.3 Confidence Intervals

The asymptotic distribution contained in the above results can be used to calculate pointwise confidence intervals for the local constant and local linear estimators. 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 on the derivatives of the density of X. This approach can be justified when a bandwidth is chosen that makes the bias 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

$$m(x) = \widehat{m}(x) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}\,[\widehat{m}(x)]},$$

where $\widehat{\mathrm{var}}[\widehat{m}(x)]$ is a consistent estimate of the asymptotic variance of $\widehat{m}(x)$, is a valid $1-\alpha$ confidence set. To get estimates of $\mathrm{var}[\widehat{m}(x)]$ we exploit the linearity of the local constant and local linear estimates. That is, they both can be written in the form

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

where $\{w_{ni}(x)\}$ only depend on the design. We have furthermore that $\sum_{i=1}^{n} w_{ni}(x) = 1$ and so

$$\widehat{m}(x) - m(x) = \sum_{i=1}^{n} w_{ni}(x)\varepsilon_i + \sum_{i=1}^{n} w_{ni}(x)\{m(X_i) - m(x)\}.$$

We have is of smaller order. It follows that

$$\begin{aligned}\mathrm{var}[\widehat{m}(x)] \;&=\; E(\mathrm{var}[\widehat{m}(x)\,|X_1,\ldots,X_n]) + \mathrm{var}(E[\widehat{m}(x)\,|X_1,\ldots,X_n]) \\ &\simeq\; \mathrm{var}[\widehat{m}(x)\,|X_1,\ldots,X_n] = \sum_{i=1}^{n} w_{ni}^2(x)\sigma^2(X_i).\end{aligned}$$

Therefore, consider

$$\widehat{v}_1(x) = \sum_{i=1}^{n} w_{ni}^2(x)\widehat{\varepsilon}_i^2 \quad ; \quad \widehat{v}_2(x) = \widehat{\sigma}^2(x)\sum w_{ni}^2(x) \quad ; \quad \widehat{v}_3(x) = \frac{1}{nh}\frac{\widehat{\sigma}^2(x)}{\widehat{f}(x)} \cdot \int K^2(u)du,$$

where $\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i)$, $\widehat{\sigma}^2(x) = \sum_{i=1}^{n} w_{ni}(x)\widehat{\varepsilon}_i^2$, and $\widehat{f}(x)$ is the standard kernel density estimate. All three estimators are consistent, and so are the confidence intervals based on them.

Some estimators that do not satisfy $\sum_{i=1}^{n} w_{ni}(x) = 1$ have $\mathrm{var}(E[\widehat{m}(x)\,|X_1,\ldots,X_n])$ of the same magnitude as $E(\mathrm{var}[\widehat{m}(x)\,|X_1,\ldots,X_n])$ and so the asymptotics are different. One might argue that one should only care about the $\mathrm{var}[\widehat{m}(x)\,|X_1,\ldots,X_n]$ because of the ancillarity of the covariate.

Can also subtract off the mean from the residuals $\widehat{\varepsilon}_i$ since they are not guaranteed to have mean zero and do not. This does not affect the consistency of the standard error estimates but it can affect the bias.

Confidence Intervals for a general class of Smoothers We next give some heuristic calculations that allow one to provide approximate confidence intervals for a much wider class of estimators than kernels or local linear. One could argue that the conditional distribution is the appropriate framework for inference here, since the covariates are ancillary. In this case the calculations leading to the asymptotic variance are particularly easy for any linear smoother. Suppose that

$$\widehat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i, \tag{4.2}$$

where the weights $\{w_{ni}(x)\}$ only depend on the covariates $X_1, \ldots, X_n$, then

$$\operatorname{var}\{\widehat{m}(x) \,|X_1, \ldots, X_n\} = \sum_{i=1}^{n} w_{ni}^2(x)\sigma_i^2,$$

where $\sigma_i^2 = E(\varepsilon_i^2|X_i)$. Note that this is exactly true for any linear smoother of the form (4.2). If the error terms were normally distributed, then $\widehat{m}(x)$ itself is also normally distributed, conditional on the design. In general, although we may not be able to prove it, we can expect that $\widehat{m}(x)$ is asymptotically normal after location and scale adjustment. Thus we expect that under appropriate regularity conditions,

$$\frac{\widehat{m}(x) - E\{\widehat{m}(x) \,|X_1, \ldots, X_n\}}{\sqrt{\sum_{i=1}^n w_{ni}^2(x)\widetilde{\varepsilon}_i^2}} = \frac{\widehat{m}(x) - E\{\widehat{m}(x) \,|X_1, \ldots, X_n\}}{\sqrt{\sum_{i=1}^n w_{ni}^2(x)\sigma_i^2}} + o_p(1) \Longrightarrow N(0,1), \tag{4.3}$$

where $\widetilde{\varepsilon}_i = Y_i - \widehat{m}(X_i)$ are the nonparametric residuals. The result (4.3) is the basis for confidence intervals for any linear smoother. This case includes splines, series, local polynomial, nearest neighbors and the many hybrid modifications thereof. It also includes multidimensional estimates and standard estimates of derivatives.

4.2.4 Uniform Consistency

In this section we discuss the uniform consistency for kernel regression estimators, that is we look for conditions under which

$$\|\widehat{m} - m\|_p = O_p(\delta_n) \text{ or } O_{a.s.}(\delta_n)$$

for some sequence $\delta_n \downarrow 0$, where

$$\|g\|_p = \begin{cases} \left\{\int |g(x)|^p d\mu(x)\right\}^{1/p} & \text{if } p < \infty \\ \sup_{x \in C} |g(x)| & \text{if } p = \infty \end{cases}$$

with C a compact set and $\mu(\cdot)$ some measure. We shall concentrate on the L_∞ distance, which is usually the most difficult to work with. These results are especially important for the analysis of semiparametric estimators which involve averages of nonparametric estimates evaluated at a large number of points. They are also relevant for many other estimation and testing problems.

Theorem 10 *(Local Linear, Masry (1996)). Suppose that:*
(i) The marginal density f of the covariates is continuous on the compact set $\mathcal{X}$ and $\inf_{x\in\mathcal{X}} f(x) > 0$;
(ii) The regression function $m''(.)$ is Lipschitz continuous on $\mathcal{X}$;
(iii) The kernel K is Lipschitz continuous on its compact support;
(iv) For some $\delta > 0$, $E(|Y|^{2+\delta}) < \infty$;
(v) $h \to 0$ and $nh \to \infty$ such that $n^{\delta/(2+\delta)}h/\log n\{\log n(\log\log n)^{1+\epsilon}\}^{2/(2+\delta)} \to \infty$ for some $\epsilon > 0$.
Then

$$\|\widehat{m} - m\|_\infty = O\left(\left[\frac{\log n}{nh}\right]^{1/2}\right) + O(h^2) \text{ a.s.}$$

This is a simplified version of Masry (1996). By taking $h = O((\log n/n)^{1/5})$ we obtain the best possible rate of $(\log n/n)^{2/5}$. We need $n^{\delta/(2+\delta)}n^{-1/5} \to \infty$, which requires that $\delta > 1/2$.

Einhmahl and Mason (2000) establish more precise results for the stochastic part of $\widehat{m} - m$, obtaining the precise rate of almost sure convergence.

Unlike in the density estimation case we must restrict our attention to compact sets. The reason is due to the presence of the marginal covariate density in the denominator. For unbounded support, $f(x) \to 0$ as $x \to \infty$ and so $\sup_x 1/f(x) = \infty$. It is possible to extend these results to allow C_n to expand with sample size at some rate although this slows down the rate of convergence depending on the tails of the marginal distribution of the covariate. Andrews (199?). Further results for weighted norms. Law of the iterated logarithm. Results for $||\widehat{m} - m||_p$.

Functional Central Limit Theorem

Above we established an upper bound on the order of magnitude of $||\widehat{m} - m||_\infty$. We now seek to refine this result into a limiting distribution. We shall work with kernel estimates throughout and will make assumptions to guarantee that the bias term is small.

The first type of result is a local functional central limit theorem for the kernel estimator. Fix an interior point x and let

$$\nu_n(t) = \sqrt{nh}\left[\widehat{m}(x+th) - m(x+th)\right], \quad t \in [-T, T]$$

for some fixed T. Then we already have pointwise convergence in distribution of $\nu_n(t)$. Under additional conditions it can be shown that

$$\nu_n(\cdot) \Longrightarrow Z(\cdot), \tag{4.4}$$

where $Z(\cdot)$ is some Gaussian process. It follows that $\sup_{t\in[-T,T]} |\nu_n(t)|$ has the distribution of $\sup_{t\in[-T,T]} |Z(t)|$. A similar result can be established for the local in bandwidth process

$$\nu_n(t) = \sqrt{n^{1-\alpha}t}\left[\widehat{m}_t(x) - m(x)\right], \quad t \in [-T, T],$$

where $h = tn^{-\alpha}$ for some given α. We obtain likewise a functional CLT like (4.4). These results have a number of applications from establishing the efficiency of plug-in estimators to testing theory.

Let

$$T_n = \sup_{x\in C} |T_n(x)| \quad ; \quad T_n(x) = \frac{\widehat{m}(x) - m(x)}{\sqrt{\text{var}[\widehat{m}(x)]}},$$

where C is some compact set contained in the support of X, while $\text{var}[\widehat{m}(x)]$ is the asymptotic variance or conditional variance. We know that (with undersmoothing) $T_n(x)$ is asymptotically standard normal for each x, but that $T_n = O_p(\sqrt{\log n})$. We will establish that there exists increasing sequences a_n, b_n such that

$$\Pr\left[a_n(T_n - b_n) \leq t\right] \to e^{-2e^{-t}}, \tag{4.5}$$

i.e., T_n is asymptotically Gumbel. This result was first proved in Bickel and Rosenblatt (1973) for the one dimensional density case and Rosenblatt (1976) for the d-dimensional density case, and Johnston (1982) for univariate local constant nonparametric regression. Before we explore this result further, lets see why it might be important.

One application of this result is to the limiting distribution of estimates of nonparametric bounds for covariate effects in the presence of selection, see Manski (1994). The main use of (4.5) is in setting uniform confidence intervals. The confidence intervals we have provided

$$\widehat{m}(x) \pm z_{\alpha/2}\sqrt{\text{var}[\widehat{m}(x)]}$$

have been valid for a single point. However, we are usually interested in the function m at a 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 inequalities to correct the level [see Savin (1984) and Härdle (1991) for further discussion of this] and the second approach is to treat the function $\widehat{m}(\cdot)$ as a random variable and use stochastic process limit theory. In other words, we find a set of functions $\mathcal{C}(\widehat{m})$ with the property that

$$\Pr\left[m \in \mathcal{C}(\widehat{m})\right] = 1 - \alpha$$

for large n. This is provided by the limit theory (4.5) by letting

$$\mathcal{C}(\widehat{m}) = \left\{m(\cdot) : a_n(T_n - b_n) \leq c_\alpha\right\},$$

where c_α solves $\exp(-2\exp(-c_\alpha)) = 1 - \alpha$, which leads to bands of the form

$$\widehat{m}(x) \pm (b_n + \frac{c_\alpha}{a_n})\sqrt{\widehat{\text{var}}[\widehat{m}(x)]} \quad \text{all } x,$$

where $\widehat{\text{var}}[\widehat{m}(x)]$ is some estimate of $\text{var}[\widehat{m}(x)]$. This intervals has the correct coverage. In practice, these intervals do not work terribly well for the reasons discussed in Hall (1993). A better approach is based on the bootstrap, which we will cover later on.

We now present the main result.

Theorem 11 *Suppose that*

1. *The functions m, f, σ^2 are all twice continuously differentiable on C.*

2. *The kernel is symmetric about zero and differentiable with bounded support $[-A, A]$ for some A, where $K(\pm A) = 0$.*

3. *For all k, $E(|Y|^k|X = x) \leq C_k < \infty$.*

4. *$h = O(n^{-\delta})$ with $\frac{1}{5} < \delta < \frac{1}{3}$.*

Then,

$$\Pr\left[a_n(T_n - b_n) \leq t\right] \to e^{-2e^{-t}}$$

with

$$b_n = \sqrt{-2\log h} + \frac{\log \frac{C}{\pi^2}}{\sqrt{-2\log h}} \quad ; \quad a_n = \sqrt{-2\log h},$$

where $C = \|K'\|^2 / \|K\|^2$.

4.2.5 Optimality

Stone (1982) established what is the optimal rate of convergence for nonparametric regression under certain conditions. In particular for a class of distributions for (Y, X) he found the sequence b_n such that for positive constants c

$$\lim_{n\to\infty} \inf_{\widehat{m}\in\mathcal{M}} \sup_{m\in M} \Pr\left[||\widehat{m} - m||_q \geq cb_n\right] = 1.$$

Here, $q \in (0, \infty]$ and the norm is taken over a compact set $D \subseteq \mathbb{R}^d$. The set $\mathcal{M}$ includes all estimators. The set M determines the difficulty. When M includes d-dimensional functions that are p times continuously differentiable on D, the optimal rate (for $q < \infty$) is $n^{-p/(2p+d)}$. This bound is achievable if there exists an estimator $\widehat{m}$ such that

$$\lim_{n\to\infty} \sup_{m\in M} \Pr\left[||\widehat{m} - m||_q \geq c'b_n\right] = 0$$

for some constant c'. Stone exhibited a rate optimal estimator.

Fan (1993) has investigated optimality under a mean squared error criterion. Let

$$R_n(M, \mathcal{M}) = \inf_{\widehat{m}\in\mathcal{M}} \sup_{m\in M} E\left[(\widehat{m}(x) - m(x))^2\right]$$

be the pointwise MSE optimal bound for an interior point x. He showed that the (best) local linear estimator comes within 0.896^2 of the bound asymptotically when $\mathcal{M}$ is chosen to include all estimators. He also showed that when $\mathcal{M}$ is restricted to the class of estimators linear in Y, the (best modified) local linear estimator achieves the bound. In this sense the local linear estimator is Best Linear Asymptotic Minimax (BLAM). Fan modified the local linear by the inclusion of a trimming factor of order n^{-2} in the denominator to ensure that the moments existed.

We consider some non-asymptotic results. Suppose that we consider the criterion

$$Q = \sum_{i=1}^{n} E\left[(\widehat{m}(X_i) - m(X_i))^2 \,|X_1, \ldots, X_n\right]$$

otherwise known as the trace mean squared error criterion associated with the $n \times 1$ vector $\widehat{m} = (\widehat{m}(X_1), \ldots, \widehat{m}(X_n))^\top$. Suppose that $\widehat{m}$ is linear, i.e.,

$$\widehat{m} = Wy, \tag{4.6}$$

where $y = (Y_1, \ldots, Y_n)^\top$ and W is an $n \times n$ matrix just depending on the covariates. Write the regression model as

$$y = m + \varepsilon,$$

where $m = (m(X_1), \ldots, m(X_n))^\top$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, and suppose that $E[\varepsilon|X] = 0$ and $E\left[\varepsilon\varepsilon^\top|X\right] = \sigma^2 I_n$. Then

$$Q = \mathrm{tr}(WW^\top) + \mathrm{tr}(bb^\top),$$

where the bias is $b = (W - I_n)m$ and the first term is the variance. Define the symmetric matrix

$$W_c = I_n - \left[(W - I_n)^\top (W - I_n)\right]^{1/2}.$$

Cohen (1966) showed that the estimator

$$\widehat{m}_c = W_c y$$

has smaller Q - in particular, its bias is the same but its variance is smaller unless W is symmetric, specifically

$$\mathrm{tr}(W_c W_c^\top) \leq \mathrm{tr}(WW^\top).$$

This follows because

$$\begin{aligned} W_c^\top W_c - W^\top W &= I_n + \left[(W - I_n)^\top (W - I_n)\right] - 2\left[(W - I_n)^\top (W - I_n)\right]^{1/2} - W^\top W \\ &= 2(I_n - \widetilde{W}) - 2\left[(W - I_n)^\top (W - I_n)\right]^{1/2}, \end{aligned}$$

where $\widetilde{W} = (W + W^\top)/2$. Then $\mathrm{tr}((I_n - \widetilde{W})^2) \leq \mathrm{tr}((W - I_n)^\top (W - I_n))$ is equivalent to showing that $\mathrm{tr}(\widetilde{W}^2 - WW^\top) \leq 0$. This follows because we can write $\mathrm{tr}(\widetilde{W}^2 - WW^\top) = -\mathrm{tr}((W - W^\top)^\top (W - W^\top))/4 \leq 0$.

This says that any estimator of m of the form (4.6) for which W is not symmetric is inadmissible according to the trace mean squared error criterion and gives a concrete way of improving estimators. Kernel estimators have asymmetric W matrices. Only spline estimators have symmetric ones.

4.2.6 Asymptotics For Nonlinear Kernel Smoothers

There are a number of nonlinear smoothers in practical use. A leading example of a nonlinear smoother is the median kernel smoother, which may be defined as follows

$$\widehat{M}(x) = \arg\min_{\theta} \sum_{i=1}^{n} K_h\left(x - X_i\right) |Y_i - \theta| ,$$

where $K_h\left(x - X_i\right)$ are kernel weights defined above. Although the minimizer is not unique in general any rule can be applied to select $\widehat{M}(x)$ from the set of minimizers. More generally we can have that $\{w_{ni}(x)\}$ are smoother weights that satisfy $\sum_{i=1}^{n} w_{ni}(x) = 1$. Just like the usual median, the local median is a nonlinear function of the $Y's$. Note that $\widehat{M}(x)$ solves the first order condition

$$\widehat{M}(x) = \arg \operatorname{zero}_{\theta}\ G_n(\theta) = 0,$$

where

$$G_n(\theta) = \sum_{i=1}^{n} w_{ni}(x) \left\{1(Y_i - \theta > 0) - 1(Y_i - \theta \leq 0)\right\} .$$

The asymptotic properties of a general class of "GMM" estimators is presented in the appendix. What is important for that is the variance of the score function at the true parameter value and the derivative with respect to parameter values of the first order condition at the true value. The conditional variance of this score function [at the true $\theta = M(x)$] is precisely $\sum_{i=1}^{n} w_{ni}^2(x)$. Let

$$\overline{G}(\theta) = E\left[G_n(\theta) \,|X_1, \ldots, X_n\right] = \sum_{i=1}^{n} w_{ni}(x) \left\{1 - 2F_i(\theta)\right\} ,$$

where $F_i(\theta) = \Pr(Y_i \leq \theta | X_i)$. Note that

$$\overline{G}'(M(x)) = -2 \sum_{i=1}^{n} w_{ni}(x) F_i'(M(x)) \simeq -2 f_x(M(x)),$$

where $f_x = F_x'$ and $F_x = \Pr(Y \leq \theta | X = x)$, by a Taylor expansion and using the fact that $\sum_{i=1}^{n} w_{ni}(x) = 1$.

Therefore,

$$\widehat{M}(x) - M(x) = \frac{\sum_{i=1}^{n} w_{ni}(x) \{1(Y_i - \theta > 0) - 1(Y_i - \theta \leq 0)\}}{-2 f_x(M(x))} + O_p(h^2)$$

by standard arguments. Therefore, we have [with undersmoothing]

$$\frac{\widehat{M}(x) - E\left\{\widehat{M}(x) \,|X_1, \ldots, X_n\right\}}{\sqrt{\frac{\sum_{i=1}^{n} w_{ni}^2(x)}{4 f_x(M(x))^2}}} \Longrightarrow N(0,1).$$

See Jones and Marron (1990) for further discussion. This result can be extended to the class of smoothers that set

$$G_n(\theta) = \sum_{i=1}^{n} w_{ni}(x) \rho(Y_i, \theta)$$

equal to zero, where ρ is a function that satisfies $E[\rho(Y_i, \theta)|X_i = x] = 0$ if and only if $\theta = \theta_0$.

Chapter 5

Bandwidth Selection

In this lecture we describe several methods of bandwidth selection for nonparametric regression estimation. We first define some performance criteria for an estimate $\widehat{m}(\cdot)$ of the function $m(\cdot)$. In the sequel $\pi(\cdot)$ is some weighting function defined on the support of X.

1. Pointwise MSE

$$d_{MP}(\widehat{m}(x), m(x)) = E\left[\{\widehat{m}(x) - m(x)\}^2\right]$$

2. Integrated MSE

$$d_{MI}(\widehat{m}, m) = \int E\left[\{\widehat{m}(x) - m(x)\}^2\right] \pi(x)dx$$

3. Average S.E.

$$d_A(\widehat{m}, m) = \frac{1}{n}\sum_{i=1}^{n} \{\widehat{m}(X_i) - m(X_i)\}^2 \pi(X_i)$$

4. Integrated S.E.

$$d_I(\widehat{m}, m) = \int \{\widehat{m}(x) - m(x)\}^2 \pi(x) f(x)dx$$

5. Conditional MISE

$$d_C(\widehat{m}, m) = E\{d_I(\widehat{m}, m)|X_1, \ldots, X_n\}.$$

Let h_j be the bandwidth sequences that minimize the corresponding criterion d_j.

The mean squared error criteria actually have explicit formulae for the optimal bandwidth. Recall that for univariate local linear regression, the asymptotic mean squared error at the point x is

$$d_{MP}(\widehat{m}(x), m(x)) \cong \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\int K^2(u)du + \frac{h^4}{4}\left\{m''(x)\right\}^2\left(\int u^2K(u)du\right)^2$$

$$\equiv \frac{a(x)}{nh} + b(x)h^4.$$

An optimal bandwidth can be defined as one that minimizes this criterion; this bandwidth will satisfy the following first order condition

$$\frac{a(x)}{nh^2} = 4b(x)h^3,$$

which solves to give

$$h_{MP}(x) = \left[\frac{a(x)}{4b(x)}\right]^{1/5} n^{-1/5}.$$

So the optimal bandwidth depends on the unknown quantities: $\sigma^2(x)$, $f(x)$, and $m''(x)$, and changes with each point x. Frequently, people work with an Integrated mean squared error criterion $d_{MI}(\widehat{m}(x), m(x))$, in which case the optimal bandwidth is

$$h_{MI} = \left[\frac{\int a(x)\pi(x)dx}{4\int b(x)\pi(x)dx}\right]^{1/5} n^{-1/5},$$

and the optimal bandwidth depends on only averages of $\sigma^2(x)$, $f(x)$, and $m''(x)$.

We now discuss specific methods of selecting bandwidths from data.

5.0.7 Plug-in

This involves nonparametrically estimating the unknown quantities in $a(x)$ and $b(x)$ by $\widehat{a}(x)$ and $\widehat{b}(x)$, say, and then let

$$\widehat{h}_{MP}(x) = \left[\frac{\widehat{a}(x)}{4\widehat{b}(x)}\right]^{1/5} n^{-1/5} \quad ; \quad \widehat{h}_{MI} = \left[\frac{\int \widehat{a}(x)\pi(x)dx}{4\int \widehat{b}(x)\pi(x)dx}\right]^{1/5} n^{-1/5}.$$

Provided $\widehat{a}(x) \xrightarrow{P} a(x)$ and $\widehat{b}(x) \xrightarrow{P} b(x)$, then

$$\frac{|\widehat{h}_{MP}(x) - h_{MP}(x)|}{h_{MP}(x)} \xrightarrow{P} 0,$$

while if $\sup_{x:\pi(x)>0} |\widehat{a}(x) - a(x)| \xrightarrow{P} 0$ and $\sup_{x:\pi(x)>0} \left|\widehat{b}(x) - b(x)\right| \xrightarrow{P} 0$, then

$$\sup_{x:\pi(x)>0} \frac{|\widehat{h}_{MI}(x) - h_{MI}(x)|}{h_{MI}(x)} \xrightarrow{P} 0.$$

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 $m''(x)$ is of order $1/nh^5$ and the bias can be arbitrarily bad unless some additional smoothness is assumed.

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 bandwidths. In density estimation his approach yields the simple formula

$$\widehat{h} = 1.06\widehat{\sigma}n^{-1/5},$$

where $\widehat{\sigma}$ is the estimated standard deviation (this can be the sample standard deviation or the interquartile range divided by 1.3). This is based on a normal distribution model and a Gaussian kernel. In regression this approach is more complicated because one has to specify $m(.), f(.)$, and $\sigma^2(.)$. Fan and Gijbels (1996, p67) give a convenient formula for local linear regression. Suppose that we take $\pi(x) = \pi_0(x)f(x)$ and parametric regression function $m_\theta(x)$, with $m''_\theta(x) \neq 0$, and assume that the error is i.i.d. with variance σ^2. Then the optimal bandwidth can be estimated by

$$\widehat{h}_{opt} = C_{0,1}(K) \left[\frac{\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2 \pi_0(X_i)}{\frac{1}{n}\sum_{i=1}^n m''_{\widehat{\theta}}(X_i)\pi_0(X_i)} \right]^{1/5} n^{-1/5},$$

where $\widehat{\varepsilon}_i = Y_i - m_{\widehat{\theta}}(X_i)$ are the parametric residuals and $\widehat{\theta}$ is an estimate of θ, while

$$C_{0,1}(K) = \left(\frac{\int K^2(t)dt}{[\int t^2 K(t)dt]^2} \right)^{1/5}.$$

For the Gaussian kernel, $C_{0,1}(K) = 0.776$. It is convenient to take $m_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2/2$, in which case $m''_\theta(x) = \theta_2$.

Krieger and Picklands (1981) show in the context of pointwise density estimation that the resulting plug-in estimator is asymptotically efficient. Specifically, they assumed only twice continuously differentiable density and showed that for any consistent (in the relative sense) estimator $\widehat{h}_{opt}$ of h_{opt} the resulting estimator $\widehat{f}_{\widehat{h}_{opt}}(x)$ asymptotically has the same mean squared error as $\widehat{f}_{h_{opt}}(x)$. They also constructed a consistent bandwidth sequence $\widehat{h}_{opt}$. Their arguments were based on weak convergence of the local in bandwidth empirical process. See Einmahl and Mason (2005) for a recent extension of this theory.

5.0.8 Cross Validation

This approach is based on an approximation to ASE or ISE. Thus

$$\begin{aligned} d_A(\widehat{m}, m) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{m}(X_i) - m(X_i)\right\}^2 \pi(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\widehat{m}(X_i)^2\pi(X_i) - \frac{2}{n}\sum_{i=1}^{n}\widehat{m}(X_i)m(X_i)\pi(X_i) + \frac{1}{n}\sum_{i=1}^{n}m(X_i)^2\pi(X_i). \end{aligned}$$

The last term does not depend on the bandwidth, so we drop it from consideration. The first term just depends on the data and so can be computed easily. The problem arises with the second term, and in particular $\sum_{i=1}^{n}\widehat{m}(X_i)m(X_i)\pi(X_i)/n$. We clearly can't just substitute $m(X_i)$ by $\widehat{m}(X_i)$. However, we might replace it by an unbiased estimator [in the conditional distribution], which is Y_i. This is the equivalent to taking

$$R(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{m}(X_i)\right\}^2 \pi(X_i)$$

as the bandwidth criterion. Unfortunately, this method will lead us to select $h = 0$ always, because then $\widehat{m}(X_i) = Y_i$ for all i.

What has gone wrong? The problem is that $\widehat{m}(X_i)$ depends on all the $Y's$ in the sample, i.e.,

$$\widehat{m}(X_i) = \sum_{j=1}^{n} w_{ij}Y_j = w_{ii}Y_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} w_{ij}Y_j,$$

where w_{ij} are the smoother weights, so that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{m}(X_i)Y_i\pi(X_i) = \frac{1}{n}\sum_{i=1}^{n}w_{ii}Y_i^2\pi(X_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} w_{ij}Y_jY_i\pi(X_i).$$

We have

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n}w_{ii}Y_i^2\pi(X_i)|X_1,\ldots,X_n\right] &= \frac{1}{n}\sum_{i=1}^{n}w_{ii}\{m^2(X_i) + \sigma^2(X_i)\}\pi(X_i) \\ &\simeq \frac{K(0)}{nh}\frac{1}{n}\sum_{i=1}^{n}\frac{\{m^2(X_i) + \sigma^2(X_i)\}\pi(X_i)}{f(X_i)}, \end{aligned}$$

which is the same magnitude as the variance effect we are trying to pick up. Therefore, $R(h)$ is a downward biased estimator of $d_A(\widehat{m}, m)$. There are two solutions to this problem.

First, we can estimate $\sum_{i=1}^n \widehat{m}(X_i)^2\pi(X_i)/n$ and $\sum_{i=1}^n \widehat{m}(X_i)m(X_i)\pi(X_i)/n$ by

$$\frac{1}{n}\sum_{i=1}^n \widehat{m}_i(X_i)Y_i\pi(X_i) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \widehat{m}_i^2(X_i)\pi(X_i),$$

where $\widehat{m}_i(X_i)$ is the leave-out-"i" estimator. In the local constant case this is:

$$\widehat{m}_i(x) = \frac{\frac{1}{(n-1)h}\sum_{i\neq i} K\left(\frac{x-X_i}{h}\right)Y_i}{\widehat{f}_i(x)} \quad ; \quad \widehat{f}_i(x) = \frac{1}{(n-1)h}\sum_{i\neq i} K\left(\frac{x-X_i}{h}\right).$$

In conclusion, let

$$CV(h) = \frac{1}{n}\sum_{i=1}^n \left\{Y_i - \widehat{m}_i(X_i)\right\}^2 \pi(X_i).$$

Choose $\widehat{h}_{cv} \in H_n$ to minimize $CV(h)$ for some set H_n, and then let $\widehat{m}_{\widehat{h}_{cv}}(\cdot)$. An equivalent method which has some advantages computationally, is to let

$$CV(h) = \frac{1}{n}\sum_{i=1}^n \left\{Y_i - \widehat{m}(X_i)\right\}^2 \pi(X_i) + \frac{2}{n}\sum_{i=1}^n w_{ii}Y_i^2\pi(X_i).$$

This latter approach is similar in spirit to the model selection ideas of time series.

We next give a theorem due to Härdle and Marron (1985) [see also Stone (1984) for density estimation], which established the optimality of this method for local constants.

Theorem 12 *Suppose that the following assumptions are satisfied:*

1. $H_n = [\underline{h}(n), \bar{h}(n)]$

$$\underline{h}(n) \geq C^{-1}n^{\delta-1},\ \bar{h}(n) \leq Cn^{-\delta},\ C, \delta > 0$$

2. *K is Hölder continuous, i.e.,*

$$|K(x_1) - K(x_2)| \leq c|x_1 - x_2|^{\xi},\ \xi > 0$$

and $\int |u|^{\xi}K(u)du < \infty$.

3. *The regression function m and the marginal density f are Hölder continuous.*

4. *The conditional moments are bounded by constants C_i*

$$E(|Y|^i|X = x) \leq C_i \quad \text{for all } x, \text{ for } i = 1, 2, \ldots$$

5. *The marginal density $f(x)$ of x is compactly supported and is bounded from below on the support of w.*

Then the bandwidth $\widehat{h}_{cv}$ is asymptotically optimal with respect to distances d_A, d_I and d_C, in the sense that with probability one

$$\frac{d(\widehat{m}_{\widehat{h}_{cv}}, m)}{\inf_{h \in H_n} d(\widehat{m}_h, m)} \to 1 \quad ; \quad \frac{\widehat{h}_{cv}}{\widehat{h}_{opt}} \to 1.$$

The conditions of this theorem are very weak in some respects. Specifically, the amount of smoothness assumed for m and f is almost nil. This means that the bandwidth selection method is automatically adapting to the amount of smoothness. In the full proof one must take account of an general magnitude for $d(\widehat{m}_h, m)$ and of a 'parameter' set that is much larger than the one we considered, which is why the theorem is stated in this fashion. Finally, in the special case we worked with one can also establish the stronger result

$$n^{2/5} \left\{ \widehat{m}_{\widehat{h}_{cv}}(x) - \widehat{m}_{\widehat{h}_{opt}}(x) \right\} \xrightarrow{P} 0.$$

Chapter 6

The Bootstrap

The bootstrap is a very popular method for obtaining confidence intervals or performing hypothesis tests. There can be computational reasons why this method is preferred to the usual approach based on estimating the unknown quantities of the asymptotic distribution. There can also be statistical reasons why the bootstrap is better than the asymptotic plug-in approach. The bootstrap has been shown to work in a large variety of situations, we are just going to look at the simplest i.i.d. cases.

Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F. We have a statistic (root)

$$R_n(\tau; X_1, \ldots, X_n; F),$$

which is a function of the data $X_1, \ldots, X_n$ and a parameter value τ. For example R_n could be an estimator or a test statistic. Let

$$H_n(x, F) = \Pr(R_n \leq x),$$

where the probability is calculated under the true distribution F. The question is, how to estimate $H_n(x, F)$ and functions thereof.

The 'asymptotic' approach uses the fact that

$$H_n(x,\, F) \longrightarrow H(x,\, F) \text{ as } n \to \infty \text{ by CLT or other method,}$$

then estimate $H_n(x, F)$ by

$$\widehat{H}_A(x) = H(x,\, F_n),$$

where F_n is some estimate of F like the empirical distribution. For example,

$$R_n = \sqrt{n}(\bar{X} - \mu) \Longrightarrow N(0,\, \sigma^2),$$

and we approximate the distribution of R_n by $N(0, \widehat{\sigma}^2)$, where $\widehat{\sigma}^2$ is some consistent estimate of σ^2. In some cases H does not depend on F; then R_n is a pivot or asymptotic pivot.

The Bootstrap approach is based on

$$\widehat{H}_B(x) = H_n(x,\, F_n).$$

In fact, we make a further approximation by using Monte-Carlo methods to find H_n. The probability measure of the data $X_1, \ldots,\, X_n$ is denoted P_n, this is discrete with probability $1/n$ at each sample point. Let $X_1^*, \ldots, X_m^*$ be a sample from P_n and let $R_n^* = R_n(\hat{\tau}_n; X_1^*, \ldots, X_m^*; F_n)$. Then

$$\mathcal{L}(R_n^* | X_1, \ldots,\, X_n) = \widehat{H}_B.$$

Actually use T replications to approximate this distribution by an 'empirical'. Usually, $m = n$.

Note that F_n could be the empirical distribution, i.e., $F_n(x) = \sum_{i=1}^n \{X_i \leq x\}/n$ or an estimate parametric c.d.f. $F_{\widehat{\theta}}$. In the latter case, the resampling is from the distribution $F_{\widehat{\theta}}$.

Theorem 13 *(Bickel and Freedman 1986) Suppose that $X_1, \ldots,\, X_n$ are i.i.d. with finite mean μ and positive variance σ^2. Along almost all sample sequences $\{X_1, \ldots,\, X_n\}$, as $n,\, m \to \infty$*

$$\mathcal{L}\left\{\sqrt{m}(\mu_m^* - \mu_n) | X_1, \ldots,\, X_n\right\} \Longrightarrow N(0,\, \sigma^2),$$

where $\mu_m^* = m^{-1} \sum_{i=1}^m X_i^*$ *and* $\mu_n = n^{-1} \sum_{i=1}^n X_i$.

In conclusion, we have found an alternative way to approximate the distribution of $\sqrt{n}(\mu_n - \mu)$: just tabulate the distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ conditional on $X_1, \ldots, X_n$. In some cases this can be done exactly, but more often one approximates this distribution by a further step based on resampling. This idea can be used to obtain confidence intervals or to obtain critical values for tests.

6.0.9 Confidence interval for the mean

The asymptotic approach:

$$C_{n,A} = \{t : R_n(X_1, \ldots,\, X_n;\, t) \leq \widehat{H}_A^{-1}(\alpha)\}$$

$$\widehat{H}_A^{-1}(\alpha) = \sup\{x : \widehat{H}_A(x) \leq \alpha\}.$$

Then $\Pr(\tau \in C_{n,A}) \to \alpha$. For the Bootstrap, we let

$$C_{n,B} = \{t : R_n(X_1, \ldots, X_n; t) \leq \widehat{H}_B^{-1}(\alpha)\}$$

$$\widehat{H}_B^{-1}(\alpha) = \sup\{x : \widehat{H}_B(x) \leq \alpha\}$$

Then $\Pr(\tau \in C_{n,B}) \to \alpha$. To carry out a test of $\tau = \tau_0$ reject if $\widehat{\tau} \notin C_{n,B}$ or $C_{n,A}$ or let $\hat{C}_B(\alpha)$ satisfy

$$H_B^{-1}(\alpha) = \hat{C}_B(\alpha)$$

For asymptotic tests and confidence intervals you get the same result if you use $\sqrt{n}(\mu_n - \mu)$ or $\sqrt{n}(\mu_n - \mu)/s_n$, where s_n is the sample standard deviation. Not true for bootstrap. There is a difference between using pivotal or non-pivotal statistics $\sqrt{m}(\mu_m^* - \mu_n)$ or $\sqrt{m}(\mu_m^* - \mu_n)/s_n$ or $\sqrt{m}(\mu_m^* - \mu_n)/s_m^*$. In either case we find $\widehat{H}_B(x)$ and hence $\widehat{H}_B^{-1}(\alpha)$.

6.0.10 Nonparametric Density Estimation

Suppose that $X_1, \ldots, X_n$ are i.i.d. with twice continuously differentiable density f . Consider the kernel estimator

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

with bandwidth sequence $h \propto n^{-1/5}$, so that

$$n^{2/5}\{\widehat{f}(x) - f(x)\} \Longrightarrow N(b(x),\, v(x))$$

for some $b(x), v(x)$. We now investigate a bootstrap algorithm for approximating the distribution of the root $R_n = n^{2/5}\{\widehat{f}(x) - f(x)\}$. Ideally, we would like to take account of both bias and variance; the usual asymptotic approach ignores the bias.

Suppose that we resample with replacement from $\{X_1, \ldots, X_n\}$, obtaining the sample $\{X_1^*, \ldots, X_n^*\}$, where X_i^* puts mass $1/n$ at each X_i. Then let

$$\widehat{f}^*(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i^*)$$

and $R_n^* = n^{2/5}(\widehat{f}^*(x) - \widehat{f}(x))$. Following the previous work we might take $\mathcal{L}\{R_n^*|X_1, \ldots, X_n\}$ as an approximation to $\mathcal{L}\{R_n\}$. Unfortunately,

$$\begin{aligned} E\{\widehat{f}^*(x)|X_1, \ldots, X_n\} &= \frac{1}{nh}\sum_{i=1}^{n} E\left[K\left(\frac{x - X_i^*}{h}\right)\middle| X_1, \ldots, X_n\right] \\ &= \frac{1}{h}E\left[K\left(\frac{x - X_i^*}{h}\right)\middle| X_1, \ldots, X_n\right] \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \widehat{f}(x). \end{aligned}$$

In other words, $\widehat{f}^*(x)$ is a conditionally unbiased estimate of $\widehat{f}(x)$. This sounds good but since $\widehat{f}(x)$ is biased it means that $\widehat{f}^*(x)$ does a poor job of estimating that bias. However, the variance is correct, i.e.,

$$\begin{aligned} \text{var}\{\widehat{f}^*(x)|X_1, \ldots, X_n\} &= \frac{1}{n^2h^2}\sum_{i=1}^{n} \text{var}\left[K\left(\frac{x - X_i^*}{h}\right)\middle| X_1, \ldots, X_n\right] \\ &= \frac{1}{nh^2}\text{var}\left[K\left(\frac{x - X_i^*}{h}\right)\middle| X_1, \ldots, X_n\right] \\ &= \frac{1}{nh^2}\left[\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)^2 - \left\{\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right\}^2\right] \\ &= \frac{1}{nh}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)^2 - h\widehat{f}^2(x)\right] \\ &= \frac{1}{nh}f(x)\int K(u)^2 du + O(n^{-1}), \end{aligned}$$

which is the asymptotic variance of $\widehat{f}(x)$. The central limit theorem is also valid because you have independent random variables.

There are two obvious ways of correcting the bias problem:

- We can work instead with bandwidths $h = o(n^{-1/5})$ so that the bias is not present in the limiting distribution of $\widehat{f}(x)$.
- The second approach is to make an explicit bias correction to $\widehat{f}$; this requires estimation of f''.

We consider a more appealing approach that is correct but does not require explicit estimation of the higher derivatives of f. The proposal is to resample from a smoothed version of f, e.g., $\widehat{f}(x)$. Generate a sample $\{U_1, \ldots, U_n\}$ of $U[0,1]$'s, and then let $X_1^* = \widehat{F}^{-1}(U_1), \ldots, X_n^* = \widehat{F}^{-1}(U_n)$, where $\widehat{F}(x) = \int_{-\infty}^{x} \widehat{f}(z)dz$. Now let

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

as before. However, now we have

$$\begin{aligned} E[\widehat{f}^*(x)|X_1, \ldots, X_n] &= \frac{1}{h} EK\left(\frac{x - X_i^*}{h}\right) = \frac{1}{h} \int K\left(\frac{x - z}{h}\right) \widehat{f}(z)dz \\ &= \int K(u)\widehat{f}(x - uh)du \cong \widehat{f}(x) + \frac{h^2}{2}\mu_2(K)\widehat{f}''(x), \end{aligned}$$

which implies that

$$E[\widehat{f}^*(x) - \widehat{f}(x)|X_1, \ldots, X_n] \cong \frac{h^2}{2}\mu_2(K)f''(x),$$

provided $\widehat{f}''(x) \to \widehat{f}''(x)$ $a.s.$ The problem here is that for the consistency of $\widehat{f}''(x)$ we would require that $nh^5 \to \infty$, which rules out the optimal bandwidth $h \propto n^{-1/5}$. Therefore, we resample from

$$\widehat{F}_g(x) = \int_{-\infty}^{x} \widehat{f}_g(z)dz,$$

where $\widehat{f}_g(x)$ is a kernel density estimate constructed from the bandwidth g. This gives

$$E[\widehat{f}^*(x)|X_1, \ldots, X_n] = \frac{1}{n} \int K\left(\frac{x - X}{h}\right) \widehat{f}_g(x)dx \cong \widehat{f}_g(x) + \frac{h^2}{2}\mu_2(K)\widehat{f}_g''(x),$$

which includes $g = 0$ and $g = h$ as special cases. Now take $\mathcal{L}\{n^{2/5}(\widehat{f}^*(x) - \widehat{f}_g(x))|X_1, \ldots, X_n\}$ as an "estimate" for $\mathcal{L}\{n^{2/5}(\widehat{f}_h(x) - f(x))\}$. Provided f'' is continuous at x and $ng^5 \to \infty$, $\widehat{f}_g''(x) \longrightarrow f''(x)$ $a.s.$ Therefore,

$$E[\widehat{f}^*(x)|X_1, \ldots, X_n] - \widehat{f}_g(x) \cong \frac{h^2}{2}\mu_2(K)f''(x),$$

which is the same as the asymptotic mean of $\widehat{f}_h(x) - f(x)$.

6.0.11 Nonparametric Regression

Suppose that

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

where either:

MODEL 1. X_i are fixed in repeated samples but become dense on their support as sample size tends to infinity, while ε_i are i.i.d. mean zero variance σ^2.

MODEL 2. (Y_i, X_i) are i.i.d. with $m(x) = E(Y|X = x)$ and $\text{var}(Y|X = x) = \sigma^2(x)$.

The main difference is that in model 1 the errors are homoskedastic and indeed i.i.d. In model 1 we can use the following algorithm

Residual resampling

1. Calculate residuals $\hat{\varepsilon}_i = Y_i - \widehat{m}_h(X_i)$, $i = 1, \ldots, n$

2. Recenter $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \bar{\hat{\varepsilon}}$, where $\bar{\hat{\varepsilon}} = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i \neq 0$.

3. Resample $\{\varepsilon_i^*, \ldots, \varepsilon_n^*\}$ drawn with replacement from $\{\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n\}$.

4. Let $Y_i^* = \widehat{m}_g(X_i) + \varepsilon_i^*$, $i = 1, \ldots, n$. Create bootstrap observations; required that $g/h \to \infty$.

5. Calculate bootstrap nonparametric estimate

$$\widehat{m}_h^*(x) = \sum_{i=1}^n w_{ni}(x) Y_i^*$$

6. To approximate the distribution of any functional of $\widehat{m}_h(\cdot) - m(\cdot)$ use the computable conditional distribution of $\widehat{m}_h^*(\cdot) - \widehat{m}_g(\cdot)$.

Theorem 14 *Suppose that $h \propto n^{-1/5}$, $g/h \to \infty$, $g \to 0$, K is bounded support and symmetric about zero, m is twice continuously differentiable. Then,*

$$\sup_{-\infty<t<\infty} \left| \Pr[\sqrt{nh}(\widehat{m}_h^*(x) - \widehat{m}_g(x)) \leq t | Y_1, X_1, \ldots, Y_n, X_n] - \Pr[\sqrt{nh}(\widehat{m}_h(x) - \widehat{m}(x)) \leq t] \right| \to 0.$$

When the errors are not identically distributed we can use the 'wild bootstrap'. In this case you draw ε_i^* from a distribution with mean zero and variance $\tilde{\varepsilon}_i^2$. For example, a normal distribution or a discrete distribution. I

In the second model, we propose to use i.i.d. resampling with oversmoothing.

i.i.d. resampling

1. Resample (X_i^*, Y_i^*), $i = 1, \ldots, n$

2. Compute $w_{ni}^*(x)$ from X_i^*, $i = 1, \ldots, n$ in the same way as $w_{ni}(x)$ was computed from X_i, $i = 1, \ldots, n$

$$\widehat{m}_h^*(x) = \sum_{i=1}^{n} w_{ni}^*(x) Y_i^*.$$

3. To approximate the distribution of any functional of $\widehat{m}_h(\cdot) - m(\cdot)$ use the computable conditional distribution of $\widehat{m}_g^*(\cdot) - \widehat{m}_g(\cdot)$, where g is another bandwidth

Chapter 7

Additive Models

According to Luce and Tukey (1964), additivity is basic to science. It is certainly hard to think of functions that are not additive in some sense, i.e., after transformations or relabelling of variables. This simplifying structure is present in many models of economic behavior starting with Leontieff (1947); see Deaton and Muellbauer (1980) for examples. Additivity is also widely used in parametric and semiparametric models of economic data. Our purpose here is to investigate a very general class of statistical models that combine additive separability with an unrestricted functional form for the covariate effects; this general class of structures are generically called additive nonparametric regression models.

There is a large literature on estimating demand curves nonparametrically. Although there are many parametric functional forms, they are generally not needed. Instead, certain abstract properties like symmetry, homogeneity, homotheticity, separability are all that are needed for many of the implications of demand theory. A function $m(x)$ is additively separable if

$$m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$$

for some functions m_α. Estimation in these models was first discussed by Stone (1985,1986) who showed that the optimal rate for estimating $m(\cdot)$ is the one-dimensional rate of convergence e.g., $n^{2/5}$ for twice continuously differentiable functions. In the statistical literature the additive regression model has been advanced in the eighties largely by the work of Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1991). Their estimation methods, called generically backfitting, rely on iteratively computing one-dimensional smooths. Recently, Linton and Nielsen (1995), Tjøstheim

and Auestad (1994), and Newey (1994) have independently proposed an alternative procedure for estimating m_α based on integration of a standard kernel estimator. The procedure is explicitly defined and its asymptotic distribution is easily derived: it converges at the one-dimensional rate and satisfies a central limit theorem. This estimation procedure has been extended to a number of other contexts like the generalized additive model [Linton and Härdle (1996)], to dependent variable transformation models [Linton, Chen, Wang, and Härdle (1996)], to econometric time series models [Härdle and Yang (1996)], and to hazard models with time varying covariates and right censoring [Nielsen (1996)].

7.1 Model and Notation

We suppose that one observes i.i.d. observations (X_i, Y_i) for $i = 1, \ldots, n$, where the response Y_i is real valued and where the covariates $X_i = (X_{1i}, \ldots, X_{di})$ take values in $\mathbb{R}^d$. Define the regression function $m(x) = E(Y|X = x)$. Then the additive model can be written as

$$Y_i = c + m_1(X_{1i}) + \ldots + m_d(X_{di}) + \varepsilon_i, \tag{7.1}$$

where the error variables ε_i satisfy $E(\varepsilon_i|X_i) = 0$ a.s.. We shall maintain throughout that $\text{var}(\varepsilon_i|X_i) = \sigma^2(X_i) < \infty$ a.s. The functions $m_1, \ldots, m_d$ and the constant c are unknown and have to be estimated by the data. For identifiability we make the additional assumption that

$$\int m_\alpha(x_\alpha) dQ_\alpha(x_\alpha) = 0$$

for $\alpha = 1, \ldots, d$, where Q_α is a signed measure. For example, Q_α could be the marginal distribution of X_α. Without this assumption replacing e.g. $m_\alpha(x_\alpha)$ by $m_\alpha(x_\alpha) + c_\alpha$, $\alpha = 1, \ldots d$, such that $\sum_{\alpha=1}^d c_\alpha = 0$ would not change the sum $\sum_{\alpha=1}^d m_\alpha(x_\alpha)$. So the model remains unchanged although the functions m_α changed. Such arbitrariness is eliminated by our assumption. It follows that $c = \int m(x) dQ(x)$, where Q is any measure for which Q_α are the marginals.

We assume one further condition on the covariate distribution for identification of m_α. We suppose that

$$\sum_{\alpha=1}^d f_\alpha(X_\alpha) = 0 \text{ a.s.} \implies f_\alpha \equiv 0 \text{ a.s., } \alpha = 1, \ldots, d.$$

Alternatively, one can normalize by taking $m_\alpha(x_{\alpha 0}) = 0$, say, for each α. This might be convenient in some cases as particular values of the covariate might have special meaning, like zero input should produce zero output. We shall not pursue this further.

Write for each α, $x = (x_\alpha, x_{-\alpha})$, $X = (X_\alpha, X_{-\alpha})$, and $X_i = (X_{\alpha i}, X_{-\alpha i})$. We shall suppose for simplicity that X are absolutely continuous with respect to Lebesgue measure on some set $\mathcal{X}$ (usually a compact subset of $\mathbb{R}^d$) and have a density function $f(x)$, which has marginals $f_\alpha(x_\alpha)$ and $f_{-\alpha}(x_{-\alpha})$ for all α.

7.2 Estimation

7.2.1 Marginal Integration

This method is due to Linton and Nielsen (1995), who called it Marginal Integration, to Newey (1994), who called it Partial Mean, and to Tjøstheim and Auestad (1994), who called it Projection. Define

$$g_\alpha(x_\alpha) = \int m(x) dQ_{-\alpha}(x_{-\alpha}),$$

where $Q_{-\alpha}(x_{-\alpha})$ is a $d-1$ dimensional probability measure. It follows that

$$g_\alpha(x_\alpha) = c + m_\alpha(x_\alpha) + \sum_{\gamma \neq \alpha}^{d} \int m_\gamma(x_\gamma) dQ_{-\alpha}(x_{-\alpha}) \equiv m_\alpha(x_\alpha) + \mu_\alpha$$

so that $g_\alpha(x_\alpha)$ is equal to $m_\alpha(x_\alpha)$ upto an additive constant. The constants μ_α are determined by c and by the choice of measures $Q_{-\alpha}$. Since we have assumed that m_α are mean zero with respect to Q_α,

$$m_\alpha(x_\alpha) = g_\alpha(x_\alpha) - \int g_\alpha(x_\alpha) dQ_\alpha(x_\alpha). \tag{7.2}$$

We use these relations to generate estimators. In practice we have to replace m by an unrestricted nonparametric regression estimator $\widehat{m}(x)$ (and perhaps the $Q_{-\alpha}, Q_\alpha$ by estimates $\widehat{Q}_{-\alpha}, \widehat{Q}_\alpha$ when they are unknown) and approximate the integral by some method. We then let

$$\begin{aligned} \widehat{g}_\alpha(x_\alpha) &= \int \widehat{m}(x) d\widehat{Q}_{-\alpha}(x_{-\alpha}) \\ \widehat{m}_\alpha(x_\alpha) &= \widehat{g}_\alpha(x_\alpha) - \int \widehat{g}_\alpha(x_\alpha) d\widehat{Q}_\alpha(x_\alpha) \end{aligned}$$

$$\widehat{m}(x) = \widehat{c} + \sum_{\alpha=1}^{d} \widehat{m}_\alpha(x_\alpha), \quad \widehat{c} = \int \widehat{m}(x) d\widehat{Q}(x).$$

There are many choices for $\widehat{m}$ here. For example, the multivariate local constant estimator. Alternatively, one can use local polynomial estimators. There are several common choices of weighting measure $Q_{-\alpha}$:

1. $\widehat{Q}_{-\alpha}$ is the empirical distribution $\widehat{F}_{-\alpha}$ of $X_{-\alpha i}$

2. $\widehat{Q}_{-\alpha}$ is the integral of a kernel estimate $\widehat{f}_{-\alpha}$ of the density of $X_{-\alpha i}$

3. $Q_{-\alpha}$ is the integral of some fixed density $q_{-\alpha}$ defined on a subset of the support of $X_{-\alpha i}$

The common implementation of the integration method is computationally demanding. This is based on taking the empirical distribution of the covariates $X_{-\alpha}$, and involves computing

$$\widehat{g}_\alpha(x_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \widehat{m}(x_\alpha, X_{-\alpha i})$$

for each point of interest x_α. If we compute $\widehat{m}_\alpha(x_\alpha)$ at each sample observation $X_{\alpha i}$ in effect one needs to compute $\widehat{m}(X_{\alpha j}, X_{-\alpha i})$ for $i, j = 1, \ldots, n$, i.e., one has to compute n different $n \times n$ smoothing matrices. We next discuss an alternative approach.

Linton and Nielsen (1995) and Fan, Mammen, and Härdle (1998) consider the choice of optimal weighting.

7.2.2 Instrumental Variables

Letting $\eta_i = \sum_{\gamma \neq \alpha} m_\gamma (X_{\gamma i}) + \varepsilon_i$, we rewrite the model (7.1) as

$$Y_i = g_\alpha (X_{\alpha i}) + \eta_i = c + m_\alpha (X_{\alpha i}) + \eta_i, \tag{7.3}$$

which is a classical example of "omitted variable" regression. That is, although (7.3) appears to take the form of a univariate nonparametric regression model, smoothing Y on X_α will incur a bias due to the omitted variable η, because η contains $X_{-\alpha}$, which in general depends on X_α. One solution to this is suggested by the classical econometric notion of instrumental variable. That is, we look for an instrument W such that

$$E(W|X_\alpha) \neq 0 \quad ; \quad E(W\eta|X_\alpha) = 0 \tag{7.4}$$

with probability one.[1] If such a random variable exists,

$$E\left(WY|X_\alpha\right) = E\left(W|X_\alpha\right) m_\alpha\left(X_\alpha\right)$$

so that

$$g_\alpha\left(x_\alpha\right) = \frac{E(WY|X_\alpha = x_\alpha)}{E(W|X_\alpha = x_\alpha)}. \tag{7.5}$$

This suggests that we estimate the function $m_\alpha(\cdot)$ by nonparametric smoothing of WY on X_α and W on X_α. In parametric models the choice of instrument is usually not obvious and requires some caution. However, our additive model has a natural class of instruments – $f_{-\alpha}\left(X_{-\alpha}\right)/f\left(X\right)$ times any measurable function of X_α will do. Suppose that we take

$$W(X) = \frac{f_\alpha(X_\alpha) f_{-\alpha}\left(X_{-\alpha}\right)}{f\left(X\right)}. \tag{7.6}$$

We have

$$\begin{aligned} E\left(W\eta|X_\alpha\right) &= E\left(W\left(\sum_{\gamma\neq\alpha} m_\gamma(X_\gamma)\right)|X_\alpha\right) \\ &= \int \frac{f_\alpha(X_\alpha) f_{-\alpha}\left(X_{-\alpha}\right)}{f\left(X\right)}\left(\sum_{\gamma\neq\alpha} m_\gamma(X_\gamma)\right)\frac{f(X)}{f_\alpha\left(X_\alpha\right)} dX_{-\alpha} \\ &= \sum_{\gamma\neq\alpha}\int m_\gamma(X_\gamma) f_{-\alpha}\left(X_{-\alpha}\right) dX_{-\alpha} \\ &= 0. \end{aligned}$$

Furthermore, $E\left(W|X_\alpha\right) = 1$ so that $g_\alpha\left(x_\alpha\right) = E(WY|X_\alpha = x_\alpha)$.

Of course, the distribution of the covariates is rarely known *a priori*. In practice, we have to rely on estimates of these quantities. Let $\widehat{f}\left(\cdot\right), \widehat{f}_\alpha\left(\cdot\right)$, and $\widehat{f}_{-\alpha}\left(\cdot\right)$ be kernel estimates of the densities

[1] Note the contrast with the marginal integration method. In this approach one defines m_α by some unconditional expectation

$$m_\alpha(x_\alpha) = E[m(x_\alpha, X_{-\alpha})W(X_{-\alpha})]$$

for some weighting function W that depends only on $X_{-\alpha}$ and which satisfies

$$E[W(X_{-\alpha})] = 1 \quad ; \quad E[W(X_{-\alpha})m_{-\alpha}(X_{-\alpha})] = 0.$$

$f(\cdot)$, $f_\alpha(\cdot)$, and $f_{-\alpha}(\cdot)$, respectively. Then, the feasible procedure is defined by replacing the instrumental variable W_i by $\widehat{W}_i = \widehat{f}_\alpha(X_{\alpha i})\,\widehat{f}_{-\alpha}(X_{-\alpha i})/\widehat{f}(X_i)$ and computing an internally normalized one dimensional smooth of $\widehat{W}_i Y_i$ on $X_{\alpha i}$. Thus

$$\widehat{g}_\alpha^{iv}(x_\alpha) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_\alpha - X_{\alpha i}}{h}\right)\frac{\widehat{W}_i Y_i}{\widehat{f}_\alpha(X_{\alpha i})} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_\alpha - X_{\alpha i}}{h}\right)\frac{\widehat{f}_{-\alpha}(X_{-\alpha i})}{\widehat{f}(X_i)}Y_i \tag{7.7}$$

as our estimate of $g_\alpha(x_\alpha) = c + m_\alpha(x_\alpha)$. It has several interpretations in addition to the above instrumental variable estimate. First, as a version of the one-dimensional regression smoother but adjusting internally by a conditional density estimate

$$\widehat{f}_{\alpha|-\alpha}(X_{\alpha i}|X_{-\alpha i}) = \frac{\widehat{f}(X_{\alpha i}, X_{-\alpha i})}{\widehat{f}_{-\alpha}(X_{-\alpha i})},$$

instead of by a marginal density estimate. Second, one can think of (7.7) as a one-dimensional standard Nadaraya-Watson (externalized) regression smoother of the adjusted data $\widehat{Y}_i$ on $X_{\alpha i}$, where $\widehat{Y}_i = \widehat{f}_\alpha(x_\alpha)\widehat{f}_{-\alpha}(X_{-\alpha i})Y_i \Big/ \widehat{f}(X_{\alpha i}, X_{-\alpha i})$. Finally, note that $\widehat{g}_\alpha^{iv}(X_{\alpha i})$ can be interpreted as a marginal integration estimator in which the pilot estimator is a fully internalized smoother [see Jones, Davies and Park (1994)][2] and the integrating measure is the empirical covariate one

$$\widetilde{m}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\frac{Y_i}{\widehat{f}(X_i)},$$

rather than the Nadaraya-Watson: by interchanging the orders of summation, we obtain

[2] Actually, Jones, Davies and Park (1994) only considered the version where $\widehat{f}$ is replaced by f and its relation to the local linear smoother.

$$\begin{aligned}
\widehat{g}_{\alpha}^{iv}(X_{\alpha i}) &= \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{X_{\alpha i}-X_{\alpha j}}{h}\right)\frac{\widehat{f}_{-\alpha}(X_{-\alpha j})}{\widehat{f}(X_j)}Y_j \\
&= \frac{1}{nh}\sum_{j=1}^{n}\frac{K\left(\frac{X_{\alpha i}-X_{\alpha j}}{h}\right)Y_j}{\widehat{f}(X_j)}\left\{\frac{1}{nh^{d-1}}\sum_{k=1}^{n} K\left(\frac{X_{-\alpha k}-X_{-\alpha j}}{h}\right)\right\} \\
&= \frac{1}{n^2h^d}\sum_{k=1}^{n}\sum_{j=1}^{n}\frac{K\left(\frac{X_{\alpha i}-X_{\alpha j}}{h}\right)K\left(\frac{X_{-\alpha k}-X_{-\alpha j}}{h}\right)Y_j}{\widehat{f}(X_j)} \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{nh^d}\sum_{j=1}^{n} K\left(\frac{X_{\alpha i}-X_{\alpha j}}{h}\right)K\left(\frac{X_{-\alpha k}-X_{-\alpha j}}{h}\right)\frac{Y_j}{\widehat{f}(X_j)}\right\} \\
&= \frac{1}{n}\sum_{k=1}^{n}\widetilde{m}(X_{\alpha i}, X_{-\alpha k}),
\end{aligned}$$

where $\widetilde{m}(X_{\alpha i}, X_{-\alpha k})$ is an internally normalized pilot smoother.

The main advantage that the local instrumental variable method has is in terms of the computational cost. There is a convenient matrix formula for the IV estimator in the bivariate case

$$\widehat{\gamma}_1 = S_1(y. * (S_2 i)./n(S_1. * S_2)i),$$

where S_1 and S_2 are one-dimensional smoothing matrices, i is the n vector of ones, and $.*$ and $./$ denote element by element multiplication and division respectively. The marginal integration method actually needs n^2 regression smoothings evaluated at the pairs $(X_{\alpha i}, X_{-\alpha j})$, for $i, j = 1, \ldots, n$, while the backfitting method requires nr operations-where r is the number of iterations to achieve convergence. The instrumental variable procedure, in contrast, takes at most $2n$ operations of kernel smoothings in a preliminary step for estimating the instrumental variable, and another n operations for the regressions. Thus, it can be easily combined with the bootstrap method whose computational costs often becomes prohibitive in the case of marginal integration [see Kim, Linton, and Hengartner (1999)].

7.2.3 Backfitting

Hastie and Tibshirani

Consider the population problem of finding functions $m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$ to minimize the least squares criterion

$$Q = E\left[\{Y - m(X)\}^2\right], \tag{7.8}$$

where $E(Y^2) < \infty$. This can be characterized as a projection problem in Hilbert space. Let $\mathcal{H}$ be the Hilbert space of square integrable functions of X, then the regression function m is the function in $\mathcal{H}$ that minimizes Q. Define the subspace of additive functions $\mathcal{H}_{add} = \oplus_{\alpha=1}^{d}\mathcal{H}_\alpha \subset \mathcal{H}$, that is the subspace of random variables $\sum_{\alpha=1}^{d} m_\alpha(X_\alpha)$ for square integrable m_α. Then the function that minimizes Q over $\mathcal{H}_{add}$, denoted $m^*(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$, satisfies the set of equations:

$$\begin{aligned}
m_1(x_1) &= E(Y|X_1 = x_1) - m_0 - E[m_2(X_2)|X_1 = x_1] - \cdots - E\left[m_d(X_d)|X_1 = x_1\right], \\
&\vdots = \vdots \\
m_d(x_d) &= E(Y|X_d = x_d) - m_0 - E\left[m_1(X_1)|X_d = x_d\right] - \cdots - E\left[m_{d-1}(X_{d-1})|X_d = x_d\right].
\end{aligned}$$

or more compactly $P_\alpha\{Y - m^*(X)\} = 0$, $\alpha = 1, \ldots, d$, where $P_\alpha(\cdot) = E(\cdot|X_\alpha)$ is the projection operator on the subspace $\mathcal{H}_\alpha$. We can represent these first order conditions as in Hastie and Tibshirani (1990):

$$\begin{pmatrix} I & P_1 & \cdots & P_1 \\ P_2 & I & \cdots & P_2 \\ \vdots & & \ddots & \vdots \\ P_d & \cdots & P_d & I \end{pmatrix} \begin{pmatrix} m_1^* \\ m_2^* \\ \vdots \\ m_d^* \end{pmatrix} = \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_d Y \end{pmatrix}.$$

The sample analogue of the projection operator P_α is the sample smoothing matrix S_α (the n by n smoother matrix used in computing $\widehat{E}(\cdot|X_\alpha)$). Therefore, we have the corresponding sample first order condition

$$\underbrace{\begin{pmatrix} I & S_1 & \cdots & S_1 \\ S_2 & I & \cdots & S_2 \\ \vdots & & \ddots & \vdots \\ S_d & \cdots & S_d & I \end{pmatrix}}_{S:nd\times nd} \underbrace{\begin{pmatrix} \widehat{m}_1 \\ \widehat{m}_2 \\ \vdots \\ \widehat{m}_d \end{pmatrix}}_{\widehat{m}:nd\times 1} = \underbrace{\begin{pmatrix} S_1 y \\ S_2 y \\ \vdots \\ S_d y \end{pmatrix}}_{s:nd\times 1}, \tag{7.9}$$

where $y = (Y_1, \ldots, Y_n)^\top$ and $\widehat{m}_\alpha = (\widehat{m}_\alpha(X_{\alpha 1}), \ldots, \widehat{m}_\alpha(X_{\alpha n}))^\top$. The estimator $\widehat{m}$ can then be defined through $\widehat{m} = S^{-1}s$ when this inverse exists. However, in practice the inversion of S is quite difficult when n is large. Opsomer and Ruppert (1997) recommended recentering the smoothers so that we replace S_α by $S^*_\alpha = (I - ii^\top/n)S_\alpha$. In the bivariate case there is a simple solution to (7.9)

$$\begin{aligned} \widehat{m}_1 &= \left\{I - (I - S_1^* S_2^*)^{-1}(I - S_1^*)\right\} y \\ \widehat{m}_2 &= \left\{I - (I - S_2^* S_1^*)^{-1}(I - S_2^*)\right\} y \end{aligned}$$

provided the inverses exist. These only involve inverting $n \times n$ matrices.

In practice, the backfitting (Gauss-Seidel) algorithm is often used instead. This is as follows

1. For each $\alpha = 1, \ldots, d$ compute $\widehat{m}_\alpha^{[0]} = S_\alpha y$ and

2. For each $\alpha = 1, \ldots, d$ and $r = 1, 2, \ldots$

$$\widehat{m}_\alpha^{[r]} = S_\alpha \left\{ y - \sum_{\gamma<\alpha} \widehat{m}_\gamma^{[r-1]} - \sum_{\gamma>\alpha} \widehat{m}_\gamma^{[r-1]} \right\}$$

3. Repeat until some convergence criterion is satisfied like the sum of squared residuals.

Each step involves just one dimensional smoothing. The estimators are linear in y. There are some problems with this algorithm. A sufficient condition ($d = 2$) for convergence of Backfitting is if either $\|S_1 S_2\| < 1$ or if both S_1 and S_2 are symmetric e.g. cubic splines. approach is to iteratively solve empirical versions of the above equations, see Breiman and Friedman (1985), Buja, Hastie an Tibshirani (1989), and Hastie and Tibshirani (1991). Hastie and Tibshirani (1990). These estimators are computed at each observation point and so are quite computationally demanding as writ when n or d is large.

Smooth Backfitting

Mammen, Linton, and Nielsen (1998) define backfitting estimates $\widetilde{m}_\alpha$ as the minimizers of the following empirical norm

$$\|\widehat{m} - \overline{m}\|_{\widehat{f}} = \int [\widehat{m}(x) - \mu - \bar{m}_1(x_1) - \ldots - \bar{m}_d(x_d)]^2 \, \widehat{f}(x)dx, \tag{7.10}$$

where the minimization runs over all functions $\overline{m}(x) = \mu + \sum_\alpha \bar{m}_\alpha(x_\alpha)$, with $\int \bar{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)dx_\alpha = 0$. Here, $\widehat{f}(x) = n^{-1}\sum_{i=1}^n K_h(x - X_i)$ is the density estimator with marginals $\widehat{f}_\alpha(x_\alpha) = \int \widehat{f}(x)dx_{-\alpha}$ [this is the one-dimensional kernel density estimate $\widehat{f}_\alpha(x_\alpha) = n^{-1}\sum_{i=1}^n K_h(x_\alpha - X_{\alpha i})$.], while $\widehat{m}(x)$ is the unrestricted Nadaraya-Watson estimator

$$\widehat{m}(x) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i}{\sum_{i=1}^n K_h(x - X_i)} \tag{7.11}$$

A minimizer of (7.10) exists if the density estimate $\widehat{f}$ is non-negative. Equation (7.10) means that $\widetilde{m}(x) = \widetilde{m}_0 + \widetilde{m}_1(x_1) + \ldots + \widetilde{m}_d(x_d)$ is the projection in the space $\mathbf{L}_2(\widehat{f})$ of $\widehat{m}$ onto the subspace of additive functions $\{m \in \mathbf{L}_2(\widehat{f}) : m(x) = m_0 + m_1(x_1) + \ldots + m_d(x_d)\}$. This is a central point of our thesis. For projection operators backfitting is well understood (method of alternating projections, see below). Therefore, this interpretation will enable us to understand convergence of the backfitting algorithm and the asymptotics of $\widetilde{m}_\alpha$. We remark that not every backfitting algorithm based on iterative smoothing can be interpreted as an alternating projection method.

The solution to (7.10) is characterized by the following system of equations ($\alpha = 1, \ldots, d$):

$$\widetilde{m}_\alpha(x_\alpha) = \int \widehat{m}(x)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} - \sum_{\gamma \neq \alpha}\int \widetilde{m}_\gamma(x_\gamma)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} - \widetilde{m}_0 \tag{7.12}$$

$$0 = \int \widetilde{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)dx_\alpha. \tag{7.13}$$

Straightforward algebra gives

$$\int \widehat{m}(x)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} = \frac{n^{-1}\sum_{i=1}^n K_h(x_\alpha - X_{\alpha i})Y_i}{\widehat{f}_\alpha(x_\alpha)} \equiv \widehat{m}_\alpha(x_\alpha),$$

because of $\int \prod_{\ell \neq \alpha} K_h(x_\ell - X_{\ell i})dx_{-\alpha} = 1$, where $\widehat{m}_\alpha(x_\alpha)$ is exactly the corresponding univariate Nadaraya-Watson estimator. Furthermore, $\widetilde{m}_0 = \int \widehat{m}(x)\widehat{f}(x)dx$, and because of $\int \prod_{\ell=1}^d K_h(x_\ell -$

$X_{\ell i})dx_{-\alpha} = 1$, we find, as in Hastie and Tibshirani (1991), that $\widetilde{m}_0 = n^{-1}\sum_{i=1}^n Y_i$, i.e., that $\widetilde{m}_0$ is the sample mean. Therefore, $\widetilde{m}_0$ is a $\sqrt{n}$-consistent estimate of the population mean and the randomness from this estimation is of smaller order and can be effectively ignored. Note also that

$$\widetilde{m}_0 = \int \widehat{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)\ dx_\alpha \quad \text{for } \alpha = 1, \ldots, d. \tag{7.14}$$

We therefore define a backfitting estimator $\widetilde{m}_\alpha(x_\alpha)$, $\alpha = 1, \ldots, d$, as a solution to the system of equations $[\alpha = 1, \ldots, d]$

$$\begin{aligned} \widetilde{m}_\alpha(x_\alpha) &= \widehat{m}_\alpha(x_\alpha) - \sum_{\gamma \neq \alpha} \int \widetilde{m}_\gamma(x_\gamma)\frac{\widehat{f}(x)}{\widehat{f}_\alpha(x_\alpha)}dx_{-\alpha} - \widetilde{m}_0, \\ 0 &= \int \widetilde{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)dx_\alpha. \end{aligned}$$

with $\widetilde{m}_0$ defined by (7.14). Up to now we have assumed that multivariate estimates of the density and of the regression function exist for all x. This assumption is not reasonable for large dimensions d (or at least such estimates can perform very poorly). Furthermore, this assumption is not necessary. Note that (7.12) can be rewritten as

$$\widetilde{m}_\alpha(x_\alpha) = \widehat{m}_\alpha(x_\alpha) - \sum_{\gamma \neq \alpha} \int \widetilde{m}_\gamma(x_\gamma)\frac{\widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)}dx_\gamma - \widetilde{m}_0, \tag{7.15}$$

where $\widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma) = n^{-1}\sum_{i=1}^n K_h(x_\alpha - X_{\alpha i})K_h(x_\gamma - X_{\gamma i})$ is the two-dimensional marginal of the full dimensional kernel density estimate $\widehat{f}(x)$. In this equation only one and two dimensional marginals of $\widehat{f}$ are used. The integrals are computed numerically. The estimator can be computed on a grid of points in the covariate support $I_1 \times \cdots \times I_d$ so it does not need residuals as in the standard backfitting approach. This estimator has been called smooth backfitting by Nielsen and Sperlich (2005).

Upto now we have implicitly assumed that the support of X is unbounded or at least that the density approaches zero at the boundary suitably fast. We now consider a generalization of the method which takes care of the boundary effects that are present when the densities have compact support. We do not require that (7.14) holds [i.e., $\int \widehat{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)\ dx_\alpha$ may depend on α], nor that $\widehat{f}_\alpha$ be a probability density, and we allow that $\widehat{f}_\alpha$ is not the marginal density of $\widehat{f}_{\alpha,\gamma}$, i.e., it may not hold for all $\alpha \neq \gamma$ that

$$\widehat{f}_\alpha(x_\alpha) = \int \widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)\ dx_\gamma. \tag{7.16}$$

For instance this may be the case for kernel density estimates of a density with compact support. For this more general setting we want to find now an appropriate modification of (7.15). We rewrite (7.15) as

$$\widetilde{m}_\alpha(x_\alpha) = \widehat{m}_\alpha(x_\alpha) - \sum_{\gamma \neq \alpha} \int \widetilde{m}_\gamma(x_\gamma) \frac{\widehat{f}_{\alpha\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} dx_\gamma - \widetilde{m}_{0,\alpha}, \tag{7.17}$$

where $\widetilde{m}_{0,\alpha}$ is chosen such that $\int \widetilde{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)dx_\alpha = 0$ for all α. Under the assumption of (7.14), (7.16) and $\int \widehat{f}_\alpha(x_\alpha)\, dx_\alpha = 1$, this gives (7.15). In general, (7.17) can be rewritten as

$$\widetilde{m}_\alpha(x_\alpha) = \widehat{m}_\alpha(x_\alpha) - \widetilde{m}_{0,\alpha} - \sum_{\gamma \neq \alpha} \int \widetilde{m}_\gamma(x_\gamma) \left[\frac{\widehat{f}_{\alpha\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} - \widehat{f}_{\gamma,[\alpha+]}(x_\gamma) \right] dx_\gamma, \tag{7.18}$$

where for $\gamma \neq \alpha$

$$\widehat{f}_{\gamma,[\alpha+]}(x_\gamma) = \int \widehat{f}_{\alpha\gamma}(x_\alpha, x_\gamma)\, dx_\alpha \left[\int \widehat{f}_\alpha(x_\alpha)\, dx_\alpha \right]^{-1}, \tag{7.19}$$

$$\widetilde{m}_{0,\alpha} = \frac{\int \widehat{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)\, dx_\alpha}{\int \widehat{f}_\alpha(x_\alpha)\, dx_\alpha}. \tag{7.20}$$

In practice, our backfitting algorithm works as follows. One starts with an arbitrary initial guess $\widetilde{m}_\alpha^{[0]}$ for $\widetilde{m}_\alpha$; for example, $\widetilde{m}_\alpha^{[0]} = \widehat{m}_\alpha$ or $\widetilde{m}_\alpha^{[0]}$ is the marginal integration estimator of Linton and Nielsen (1995). In the α-th step of the r-th iteration cycle one puts

$$\begin{aligned} \widetilde{m}_\alpha^{[r]}(x_\alpha) = \widehat{m}_\alpha(x_\alpha) &- \sum_{\gamma < \alpha} \int \widetilde{m}_\gamma^{[r]}(x_\gamma) \left[\frac{\widehat{f}_{\alpha\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} - \widehat{f}_{\gamma,[\alpha+]}(x_\gamma) \right] dx_\gamma \\ &- \sum_{\gamma > \alpha} \int \widetilde{m}_\gamma^{[r-1]}(x_\gamma) \left[\frac{\widehat{f}_{\alpha\gamma}(x_\alpha, x_\gamma)}{\widehat{f}_\alpha(x_\alpha)} - \widehat{f}_{\gamma,[\alpha+]}(x_\gamma) \right] dx_\gamma - \widetilde{m}_{0,\alpha}, \end{aligned} \tag{7.21}$$

and the process is iterated until a desired convergence criterion is satisfied.

7.2.4 Interpretation

The backfitting method has a nice population interpretation as the projection of the given function $m(x)$ on to the space of additive functions where the norm is in terms of the expectation. This interpretation says what is happening when the additive model is not true, and it also turns out to be key in establishing efficiency properties. Here we present a similar interpretation of marginal integration. Let

$$\pi_I(m)(x) = \int m(x)dQ_{-\alpha}(x_{-\alpha})dQ_\alpha(x_\alpha) + \sum_{\alpha=1}^{d}\left[\int m(x)dQ_{-\alpha}(x_{-\alpha}) - \int m(x)dQ_{-\alpha}(x_{-\alpha})dQ_\alpha(x_\alpha)\right],$$

be the integration 'map', that takes a function $m(x)$ in the space of additive functions. It is easy to see that π_I is a linear idempotent map from $\mathcal{H}$ into itself, and moreover $\pi_I(m) = m$ if $m \in \mathcal{H}_{add}$. However, π_I is not self-adjoint, i.e., it is not an orthogonal projection with respect to the norm induced by expectation with respect to the joint distribution of the covariates. However, if we change the definition of norm on the space $\mathcal{H}$ we can find an intepretation of π_I as an orthogonal projection. Specifically, if distance is calculated by the product measure $\otimes_{\alpha=1}^{d} Q_\alpha$, then π_I is self-adjoint, and hence an orthogonal projection, Nielsen and Linton (1998). In other words we can consider $\pi_I(m)$ as the solution to the minimization problem

$$S(m) = \int \left\{ m(x) - c - \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \right\}^2 dW(x),$$

where $W = \otimes_{\alpha=1}^{d} Q_\alpha$.

Thus far, the choice between integration and backfitting is reminiscent of the choice between ordinary least squares and generalized least squares in regression. The latter estimator finds the closest linear approximation to the regression function in the covariance matrix norm, while the former method finds the closest linear approximation in the unweighted Euclidean norm, see Drygas (1970). Although generalized least squares is more efficient in the Gauss-Markov sense when the linear structure is true, the efficiency gain may not be huge and the estimator can be harder to compute.

Finally, we point out that under quite reasonable conditions, the solutions to $S(m)$ are continuous (in the supremum norm) in the weighting function W, so that small changes in weighting produce small differences in the fitted functions. In fact, there is a non-infinitesimal bound available for general weight functions. Under certain conditions,

$$\frac{\inf \int\int \left\{ m(x) - c - \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \right\}^2 dW_1(x)}{\inf \int\int \left\{ m(x) - c - \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \right\}^2 dW_2(x)} \leq \frac{(\alpha_1 + \alpha_2)^2}{4\alpha_1\alpha_2},$$

where the minimum in each case is taken over $m_\alpha \in \mathcal{H}_\alpha$, and $c \in \mathbb{R}$, while

$$\alpha_1 = \inf_m \Psi(m) \quad ; \quad \alpha_2 = \sup_m \Psi(m), \quad \text{where } \Psi(m) = \left\{ \frac{\int\int m(x)^2 dW_1(x)}{\int\int m(x)^2 dW_2(x)} \right\}^{1/2},$$

see Cleveland (1971).

7.3 Asymptotic Properties

We next discuss the asymptotic properties of these estimates. We discuss when additivity holds and when it does not hold.

Why does it work? "*But your procedure is based on a high dimensional nonparametric estimate which everyone knows works terribly, therefore the marginal integration method is doomed*" This is the fallacy of composition. Suppose we have a sample $X_1, \ldots, X_n$ i.i.d. from some population with mean $\mu = E(X)$. Then, X_i is an inconsistent estimator of μ. Does that mean that $n^{-1}\sum_{i=1}^n X_i$ is a lousy estimate also? This is the basic motivation for the integration estimator,

$$integration = averaging = variance\, reduction.$$

This is the basis for the field of semiparametric estimation. It is true that the the second order effect in the integration estimator deteriorates with dimensions [as in semiparametric problems]. However, I would argue that the second order effect in the backfitting estimator deteriorates when the number of iterations is small. Ceteris paribus, higher dimensions leads to fewer iterations.

Perhaps the more significant disadvantage of the integration method is that the curse of dimensionality does not get completely eliminated. Thus one must use bias reduction arguments to achieve the optimal rate in high dimensions and one might expect poor small sample performance relative to the asymptotics.

7.3.1 Integration Type Estimators

We first consider the estimator computed using a known density q with marginals q_α and $q_{-\alpha}$, that is,

$$\begin{aligned} \widehat{g}_\alpha(x_\alpha) &= \int \widehat{m}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \\ \widehat{m}_\alpha(x_\alpha) &= \widehat{g}_\alpha(x_\alpha) - \int \widehat{g}_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha \end{aligned}$$

$$\widehat{m}(x) = \widehat{c} + \sum_{\alpha=1}^{d} \widehat{m}_\alpha(x_\alpha), \quad \widehat{c} = \int \widehat{m}(x) q(x) dx.$$

First we argue that if the weight sequence is consistent for $\widehat{m}(x)$ in the sense of Stone (1977), then under an additional condition it is consistent for $\int m(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha}$.

Theorem 15 *(a) Suppose that* $\sup_x |\widehat{m}(x) - m(x)| = o_P(1)$, *then*

$$\sup_{x_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| = o_P(1)$$

(b) Suppose that $E[|\widehat{m}(X) - m(X)|^r] \to 0$ *and*

$$\sup_x \frac{q_{-\alpha}(x_{-\alpha}) f_\alpha(x_\alpha)}{f(x)} \leq C < \infty. \tag{7.22}$$

Then

$$E\left[|\widehat{g}_\alpha(X_\alpha) - g_\alpha(X_\alpha)|^r\right] \to 0.$$

Proof. First, we have

$$\begin{aligned} & \sup_{x_\alpha} \left| \int \widehat{m}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} - \int m(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \right| \\ \leq & \sup_x \int |\widehat{m}(x) - m(x)| \, q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \\ \leq & \sup_x |\widehat{m}(x) - m(x)|, \end{aligned}$$

which concludes the first part. Second,

$$\begin{aligned} & \int E\left[\left| \int \widehat{m}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} - \int m(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \right|^r \right] f_\alpha(x_\alpha) dx_\alpha \\ \leq & \int \int E\left[|\widehat{m}(x) - m(x)|^r\right] q_{-\alpha}(x_{-\alpha}) f_\alpha(x_\alpha) dx \\ = & \int \int E\left[|\widehat{m}(x) - m(x)|^r\right] \frac{q_{-\alpha}(x_{-\alpha}) f_\alpha(x_\alpha)}{f(x)} f(x) dx \\ \leq & C \int \int E\left[|\widehat{m}(x) - m(x)|^r\right] f(x) dx \to 0, \end{aligned}$$

which concludes the second part. ■

Condition (7.22) is quite weak and is satisfied when $f(x) > 0$ on the support of X.

The downside with these results is that they require that $\widehat{m}(x)$ be consistent in some sense. In the sequel we wish to obtain the asymptotic distribution of the marginal integration estimator and we expect that it converges at the one-dimensional rate. To establish this we need to make a more detailed analysis. A leading question is whether we can achieve optimal rates of convergence for given smoothness.

Let $\mathcal{X}^n = \{X_1, \ldots, X_n\}$. We use the following regularity conditions:

A1. *The kernel K is symmetric about zero and of order r, i.e., $\int K(u)u^j du = 0$, $j = 1, \ldots, r-1$. Furthermore, K is supported on $[-1, 1]$, bounded, and Lipschitz continuous, i.e., there exists a finite constant c such that $|K(u) - K(v)| \leq c\,|u - v|$ for all u, v.*

A2. *The functions $m(\cdot)$ and $f(\cdot)$ are r-times continuously differentiable in each direction, where $r \geq (d-1)/2$.*

A3. *The joint density $q(\cdot)$ is continuous on its compact support, which is $\mathcal{Q} = \times_{\alpha=1}^{d} [\underline{x}_\alpha, \overline{x}_\alpha]$. f is bounded away from zero on $\mathcal{Q}$*

A4. *The conditional variance $\sigma^2(x) = \mathrm{var}(Y|X = x)$ is continuous, and is bounded away from zero and infinity on $\mathcal{Q}$.*

A5. *$E[|Y|^\theta] < \infty$ for some $\theta > 5/2$.*

A6. *$h = \delta n^{-1/(2r+1)}$ for some $\delta \in (0, \infty)$.*

Define $D^r g(x_1, \ldots, x_d) = \sum_{\alpha=1}^{d} \partial^r g(x_1, \ldots, x_d) / \partial x_\alpha^r$ for any positive integer r.

Theorem 16 *Suppose that A1-A6 hold. Then,*

$$n^{r/(2r+1)} \left[\widehat{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha)\right] \Longrightarrow N\left[b_\alpha(x_\alpha), v_\alpha^2(x_\alpha)\right]$$
$$n^{r/(2r+1)} \left[\widehat{m}(x) - m(x)\right] \Longrightarrow N\left[b(x), v^2(x)\right]$$

in distribution, where $b(x) = \sum_{\alpha=1}^{d} b_\alpha(x_\alpha)$, $v^2(x) = \sum_{\alpha=1}^{d} v_\alpha^2(x_\alpha)$ *and*

$$\begin{aligned} b_\alpha(x_\alpha) &= \frac{\delta^r}{r!}\mu_r(K)\left[D^r m_\alpha(x_\alpha) - \int D^r m_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha\right] \\ v_\alpha^2(x_\alpha) &= \delta^{-1}\|K\|_2^2 \int \frac{q_{-\alpha}^2(x_{-\alpha})}{f(x)}\sigma^2(x) dx_{-\alpha} \end{aligned}$$

where $\mu_r(K) = \int K(t)t^r dt$ *and* $\|K\|_2^2 = \int K(t)^2 dt$.

REMARKS.

1. The bias reflects the recentering that goes into $\widehat{m}_\alpha(x_\alpha)$. The variance simplifies when X_α and $X_{-\alpha}$ are mutually independent, $\sigma^2(.)$ is constant, and $q_{-\alpha}$ is the covariate density $f_{-\alpha}$: in that case

$$v_\alpha^2(x_\alpha) = \delta^{-1}\|K\|_2^2 \frac{\sigma^2}{f_\alpha(x_\alpha)},$$

which is the asymptotic variance of the one-dimensional kernel estimator.

2. The individual estimates $\widehat{g}_\alpha(x_\alpha)$ are to first order uncorrelated, as are $\widehat{m}_\alpha(x_\alpha)$, which explains why the variance of $\widehat{m}(x)$ is just the sum of the individual variances.

3. The asymptotic variance depends on the choice of $q_{-\alpha}$. Linton and Nielsen (1995) obtained the optimal choice in terms of integrated variance over the support of $x_{-\alpha}$ under homoskesdasticity, but more generally it is

$$q_{-\alpha}^{opt}(x_{-\alpha}) = \tau^{-1}(x_{-\alpha}) / \int \tau^{-1}(x_{-\alpha}) dx_{-\alpha},$$

where $\tau(x_{-\alpha}) = \int \{f_\alpha(x_\alpha)\sigma^2(x)/f(x)\} dx_\alpha$ in which case the asymptotic variance constant is proportional to $1/\int \tau^{-1}(x_{-\alpha}) dx_{-\alpha}$.

Some Heuristics

First, under these conditions

$$\begin{aligned} \sup_{x\in\mathcal{Q}} |\widehat{m}(x) - m(x)| &= O_p\left(\sqrt{\frac{\log n}{nh^d}}\right) + O_p\left(h^r\right) \\ \sup_{x\in\mathcal{Q}} \left|\widehat{f}(x) - f(x)\right| &= O_p\left(\sqrt{\frac{\log n}{nh^d}}\right) + O_p\left(h^r\right). \end{aligned}$$

It follows that uniformly in x:

$$\begin{aligned}
&\widehat{m}(x) - m(x) \\
&= \frac{1}{f(x)} \left[\frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i + \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \left[m(X_i) - m(x) \right] \right] \left[1 + O_p(h^r) + O_p(n^{-1/2} h^{-d/2} (\log n)^{1/2}) \right] \\
&= \frac{1}{f(x)} \left[\frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i + \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \left[m(X_i) - m(x) \right] \right] + O_p(h^{2r}) + O_p(n^{-1} h^{-d} \log n).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int \left[\widehat{m}(x) - m(x) \right] q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \;=\; & \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int \frac{1}{f(x)} K_h(x - X_i) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \qquad (7.23) \\
& + \frac{1}{n} \sum_{i=1}^{n} \int \frac{1}{f(x)} K_h(x - X_i) \left[m(X_i) - m(x) \right] q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \\
& + O_p(h^{2r}) + O_p(n^{-1} h^{-d} \log n).
\end{aligned}$$

For the remainder terms to be of smaller order we require that $\sqrt{nh} \times n^{-1} h^{-d} \log n \to 0$, which requires that $nh^{2d-1}/(\log n)^2 \to \infty$. If $h \propto n^{-1/(2r+1)}$, this requires that $r > d - 1$.

The leading stochastic term is

$$T_{n1} = \int \frac{1}{f(x)} \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x_\alpha - X_{\alpha i}}{h} \right) w_{ni}(x_\alpha)\varepsilon_i,$$

where

$$w_{ni}(x_\alpha) = \int \frac{1}{f(x)} \prod_{\gamma \neq \alpha}^{d} \frac{1}{h} K \left(\frac{x_\gamma - X_{\gamma i}}{h} \right) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha}.$$

This term is a sum of mean zero independent random variables. By a change of variable argument we have

$$w_{ni}(x_\alpha) = \int \frac{1}{f(x_\alpha, X_{-\alpha,i})} \prod_{\gamma \neq \alpha}^{d} K(u_\gamma) \, q_{-\alpha}(X_{-\alpha i} + h u_{-\alpha}) du_{-\alpha} \simeq \frac{q_{-\alpha}(X_{-\alpha i})}{f(x_\alpha, X_{-\alpha,i})},$$

so that

$$T_{n1} \simeq \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x_\alpha - X_{\alpha i}}{h} \right) \frac{q_{-\alpha}(X_{-\alpha i})}{f(x_\alpha, X_{-\alpha,i})} \varepsilon_i,$$

which is asymptotically normal at rate $\sqrt{nh}$ and

$$\operatorname{var}(T_{n1}) \simeq \frac{1}{nh} ||K||_2^2 \int \frac{q_{-\alpha}^2(x_{-\alpha})}{f(x)} \sigma^2(x) dx_{-\alpha}.$$

For the change of variables argument one needs only some continuity conditions on $q_{-\alpha}/f$.

To handle the second term in (7.23) we write

$$Z_{ni}(x) = \frac{1}{f(x)} K_h(x - X_i)\,[m(X_i) - m(x)] = E[Z_{ni}(x)] + u_{ni}(x),$$

where $u_{ni}(x) = Z_{ni}(x) - E[Z_{ni}(x)]$ is i.i.d. and mean zero for each n. Then

$$T_{n2} = \int E[Z_{ni}(x)] q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \simeq h^r \int \left(\lim_{n\to\infty} \frac{E[Z_{ni}(x)]}{h^r} \right) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha}.$$

That is, the asymptotic bias of the integration estimator is just the integral of the asymptotic bias of the unrestricted estimator. The term

$$T_{n3} = \frac{1}{n} \sum_{i=1}^{n} \int u_{ni}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha}$$

is a sum of mean zero independent random variables with variance bounded by

$$\begin{aligned}
&\frac{1}{n} E\left[\left(\int Z_{ni}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \right)^2 \right] \\
&= \frac{1}{n} \int\int\int K_h(x_\alpha - x''_\alpha)^2 K_h(x_{-\alpha} - x''_{-\alpha}) K_h(x'_{-\alpha} - x''_{-\alpha}) \frac{[m(x'') - m(x)]\,[m(x'') - m(x_\alpha, x'_{-\alpha})]}{f(x) f(x_\alpha, x'_{-\alpha})} \\
&\quad \times q_{-\alpha}(x_{-\alpha}) q_{-\alpha}(x'_{-\alpha}) f(x'') dx_{-\alpha} dx'_{-\alpha} dx'' \\
&= \frac{1}{nh} \int\int\int K^2(v_\alpha) \frac{[m(x_\alpha + v_\alpha h, x''_{-\alpha}) - m(x_\alpha, x''_{-\alpha} + u_{-\alpha} h)]\,[m(x_\alpha + v_\alpha h, x''_{-\alpha}) - m(x_\alpha, x''_{-\alpha} + u'_{-\alpha} h)]}{f(x_\alpha, x''_{-\alpha} + u_{-\alpha} h) f(x_\alpha, x''_{-\alpha} + u'_{-\alpha} h)} \\
&\quad \times K(u_{-\alpha}) K(u'_{-\alpha}) q_{-\alpha}(x''_{-\alpha} + u_{-\alpha} h) q_{-\alpha}(x''_{-\alpha} + u'_{-\alpha} h) f(x_\alpha + v_\alpha h, x''_{-\alpha}) du_{-\alpha} du'_{-\alpha} dv_\alpha dx''_{-\alpha} \\
&= o(n^{-1} h^{-1}),
\end{aligned}$$

using a change of variables $x_{-\alpha} \longmapsto u_{-\alpha} = (x_{-\alpha} - x''_{-\alpha})/h$, $x'_{-\alpha} \longmapsto u'_{-\alpha} = (x'_{-\alpha} - x''_{-\alpha})/h$, and $x''_\alpha \longmapsto v_\alpha = (x_\alpha - x''_\alpha)/h$, and dominated convergence. ■

The basic problem with this proof method is that we require $\sup_x |\widehat{f}(x) - f(x)| = o_p(1)$, which requires $nh^d / \log n \to \infty$. If $h = O(n^{-1/5})$, then this cannot hold unless $d \leq 4$. Furthermore, even if this is true we will have remainder terms that are of the order $\left(\sup_x |\widehat{f}(x) - f(x)|\right)^2 = O(n^{-1} h^{-d} \log n)$ a.s., and these remainder terms should be smaller than $n^{-2/5}$. This requires $r > d - 1$, so if $r = 2$, we can only permit $d \leq 2$.

Hengartner and Sperlich. They propose an integration estimator with internally normalized pilot, i.e.,

$$m_\alpha(x_\alpha) = \int \widehat{m}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} - \int \widehat{m}(x) q(x) dx$$

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

where $\widehat{f}(x)$ is a standard kernel smoother. They argue that by imposing additional smoothness on the integrating density one can obtain rate optimality.

1. *The multivariate regression function $m(x) = E(Y|X = x)$ is s times continuously differentiable in x_1 and the conditional variance $\sigma^2(x) = var(Y|X = x)$ is finite and Lipschitz continuous.*

2. *The joint density of the covariates f is compactly supported, Lipschitz continuous and strictly bounded away from zero on the interior of the support.*

3. *The product measure Q has continuous density q (with respect to Lebesgue measure) bounded away from zero and infinity. Further the support of Q is contained in the support of f.*

4. *The multivariate smoothing kernel on R^d is the product of d univariate kernels K_α each of them being compactly supported, bounded, Lipschitz continuous, integrates to one. The kernels K_α are of order r_α, i.e., $\int u^j K_\alpha(u) du = 0$ for $j = 1, \ldots, r_\alpha - 1$, and $r_\alpha = s$.*

5. *The bandwidths satisfy $h_\alpha = cn^{-1/(2s+1)}$ for some c with $0 < c < \infty$, $h_\gamma = o(1)$ and $n \prod_{\gamma=1}^{d} h_\gamma \to \infty$.*

6. *The density $q_\alpha(x_\alpha) = \int q(x) dx_{-\alpha}$ has $s + 1$ continuous and bounded derivatives.*

Let

$$b_\alpha(x_\alpha) = \frac{\mu_s(K)}{s!} \left(\frac{1}{f_\alpha(x_\alpha)} D^s \eta_\alpha(x_\alpha) - \int m_\alpha(x_\alpha) D^s q_\alpha(x_\alpha) dx_\alpha \right)$$

$$v_\alpha^2(x_\alpha) = \frac{||K||_2^2}{f_\alpha(x_\alpha)} \left[\int \left\{ \sigma^2(x) + m^2(x) \right\} \frac{q_{-\alpha}(x_{-\alpha})}{f^2(x_{-\alpha}|x_\alpha)} f(x_{-\alpha}|x_\alpha) dx_{-\alpha} - \left\{ \int m(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \right\}^2 \right]$$

Theorem 17 *Suppose that A1-A6 hold. Then,*

$$n^{s/(2s+1)} \left[\widehat{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha) \right] \Longrightarrow N \left[b_\alpha(x_\alpha), v_\alpha^2(x_\alpha) \right]$$

$$n^{s/(2s+1)} \left[\widehat{m}(x) - m(x) \right] \Longrightarrow N \left[b(x), v^2(x) \right]$$

in distribution, where $b(x) = \sum_{\alpha=1}^{d} b_\alpha(x_\alpha)$ *and* $v^2(x) = \sum_{\alpha=1}^{d} v_\alpha^2(x_\alpha)$.

Oracle Estimation

Suppose that one knew $m_\gamma(\cdot)$, $\gamma \neq \alpha$. In this case, one can estimate $m_\alpha(x_\alpha)$ by a one-dimensional regression smoother $\check{m}_\alpha^{orc}$ of the partial error

$$U_{\alpha i} = Y_i - \sum_{\gamma \neq \alpha} m_\gamma(X_{\gamma i})$$

on $X_{\alpha i}$. Under standard regularity conditions, see Härdle (1990),

$$n^{2/5}\left\{\check{m}_\alpha^{orc}(x_\alpha) - m_\alpha(x_\alpha)\right\} \Rightarrow N\{b^{orc}(x_\alpha), v^{orc}(x_\alpha)\},$$

where (for local linear smoother)

$$b^{orc}(x_\alpha) = \frac{\mu_2(K)}{2} m_\alpha''(x_\alpha) \; ; \; v^{orc}(x_\alpha) = ||K||^2 \frac{\sigma_\alpha^2(x_\alpha)}{f_\alpha(x_\alpha)}, \tag{7.24}$$

where $\sigma_\alpha^2(x_\alpha) = \text{var}(\varepsilon | X_\alpha = x_\alpha)$. If one also imposed the knowledge that $E[m_\alpha(X_\alpha)] = 0$ one would recenter the estimate and the bias would be recentered, i.e., $m_\alpha''(x_\alpha) \mapsto m_\alpha''(x_\alpha) - \int m_\alpha''(x_\alpha) f_\alpha(x_\alpha) dx_\alpha$. By an application of the Cauchy-Schwarz inequality, $\check{m}_\alpha^{orc}(x_\alpha)$ has smaller variance than any integration based procedure that uses the same kernel under homoskedasticity. In that case we have

$$\begin{aligned} 1 &= \int f_{-\alpha|\alpha}^{1/2}(x_{-\alpha}|x_\alpha) \frac{q_{-\alpha}(x_{-\alpha})}{f_{-\alpha|\alpha}^{1/2}(x_{-\alpha}|x_\alpha)} dx_{-\alpha} \\ &\leq \int f_{-\alpha|\alpha}(x_{-\alpha}|x_\alpha) dx_{-\alpha} \int \frac{q_{-\alpha}^2(x_{-\alpha})}{f_{-\alpha|\alpha}(x_{-\alpha}|x_\alpha)} dx_{-\alpha} \\ &= f_\alpha(x_\alpha) \int \frac{q_{-\alpha}^2(x_{-\alpha})}{f_{-\alpha|\alpha}(x_{-\alpha}|x_\alpha)} dx_{-\alpha}. \end{aligned}$$

Efficient procedure is do one-step backfitting using the integration method to provide consistent starting values. Compare with parametric estimation: Rothenberg and Leenders (1965) and Bickel (1971). Smooth the residuals

$$\widetilde{U}_{\alpha i} = Y_i - \sum_{\gamma \neq \alpha} \widetilde{m}_\gamma(X_{\gamma i})$$

against $X_{\alpha i}$, where $\widetilde{m}_\gamma$ are preliminary, e.g., integration-based, estimators. Under various conditions,

$$n^{2/5}\left\{\check{m}_\alpha^{orc}(x_\alpha) - \widehat{m}_\alpha^{2-step}(x_\alpha)\right\} = o(1) \quad a.s.$$

See Linton (1996) and Linton, Hengartner and Kim (1997).

7.3.2 Backfitting

Opsomer and Ruppert (1997) and Opsomer (1997): Existence of a solution to the empirical equations provided smoother matrices recentred; Asymptotics bias and variance for this estimator; Not Design Adaptive, i.e., bias depends on the derivatives of the marginal density; Condition that restricts the amount of correlation between design variables.

We suppose that

$$K_h(u,v) = \mathbf{1}(u,v \in [0,1]) \frac{K_h(u-v)}{\int_0^1 K_h(w-v)\ dw}$$

with, again, $K_h(u) = h^{-1}K(h^{-1}u)$. We will suppose that the kernel K has compact support $[-C_1, C_1]$, see B1. For this reason we get that $K_h(u,v) = K_h(u-v)$ for $v \in [C_1h, 1-C_1h]$ or for $u \in [2C_1h, 1-2C_1h]$. So $K_h(u,v)$ differs from $K_h(u-v)$ only on the boundary. The norming gives that $\int_0^1 K_h(u,v)\ du = 1$. Therefore we have that $\int_0^1 \widehat{f}_{\alpha,\gamma}(x_\alpha, x_\gamma)\ dx_\gamma = \widehat{f}_\alpha(x_\alpha)$ and $\int_0^1 \widehat{f}_\alpha(x_\alpha)\ dx_\alpha = 1$.

(B1) *The kernel K is bounded, has compact support($[-C_1, C_1]$, say), is symmetric about zero, and is Lipschitz continuous, i.e., there exists a positive finite constant C_2 such that $|K(u) - K(v)| \leq C_2 |u-v|$.*

(B2) *The d-dimensional vector X has compact support $[0,1]^d$ and its density f is bounded away from zero and infinity on $[0,1]^d$.*

(B3) *For some $\theta > 5/2$, $E(|Y|^\theta) < \infty$. The conditional variance $\sigma^2(x) = \mathrm{var}[Y|X=x]$ is continuous on $[0,1]^d$.*

(B4) *The function m'' exists and is continuous. The derivative f' exists and is continuous.*

Define a constant b_0 and functions b_α on $\mathbb{R}$ [with $\int b_\alpha(x_\alpha) f_\alpha(x_\alpha)\ dx_\alpha = 0$] by

$$(b_0, b_1, \ldots, b_d) = \arg\min_{b_0,\ldots,b_d} \int [b(x) - b_0 - b_1(x_1) - \ldots - b_d(x_d)]^2 f(x)\ dx. \tag{7.25}$$

where

$$b(x) = \sum_{\alpha=1}^{d} \left[m'_\alpha(x_\alpha) \frac{\partial}{\partial x_\alpha} \log f(x) + \frac{1}{2} m''_\alpha(x_\alpha) \right] \mu_2(K)$$

is the bias function of the Nadaraya-Watson estimator. Let $\sigma^2_\alpha(x_\alpha) = \mathrm{var}[Y - m(X)|X_\alpha = x_\alpha]$.

Theorem 18 *Suppose that the additive model holds and that conditions B1-B4 hold, and that Nadaraya Watson backfitting smoothing is used, i.e., $\widehat{m}_\alpha$, $\widehat{f}_\alpha$ and $\widehat{f}_{\alpha,\gamma}$ are defined according to ? and $\widetilde{m}_\alpha$ is defined by (7.15). Suppose additionally that $n^{1/5}h \to c_h$ for a constant c_h. Then, the following convergence holds in distribution for any $x_1, \ldots, x_d \in (0,1)$,*

$$n^{2/5} \begin{bmatrix} \widetilde{m}_1(x_1) - m_1(x_1) \\ \vdots \\ \widetilde{m}_d(x_d) - m_d(x_d) \end{bmatrix} \Longrightarrow N \left(\begin{bmatrix} c_h^2 b_1(x_1) \\ \vdots \\ c_h^2 b_d(x_d) \end{bmatrix}, \begin{bmatrix} v_1^2(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & v_d^2(x_d) \end{bmatrix} \right),$$

where $b_\alpha(x_\alpha)$ was defined above and $v_\alpha^2(x_\alpha) = c_h^{-1} \, ||K||_2^2 \, \sigma_\alpha^2(x_\alpha)/f_\alpha(x_\alpha)$, $\alpha = 1, \ldots, d$. Consequently,

$$n^{2/5} \left[\widetilde{m}(x) - m(x) \right] \Longrightarrow N \left(c_h^2 \sum_{\alpha=1}^{d} b_\alpha(x_\alpha), \sum_{\alpha=1}^{d} v_\alpha^2(x_\alpha) \right).$$

The bias does not correspond to the bias of a one dimensional smoother. MLN show that for a local linear implementation one obtains the same result but with bias

$$b_\alpha(x_\alpha) = \frac{1}{2} \mu_2(K) \left[m_\alpha''(x_\alpha) - \int m_\alpha''(x_\alpha) f_\alpha(x_\alpha) dx_\alpha \right].$$

This estimator is oracle efficient.

Mammen and Park (2005) establish further expansions for the stochastic part of the estimator out to order $n^{-1/2}$.

Chapter 8

Generalized Additive and Other Separable Models

8.1 Models

Consider now the generalized additive model in which there is some known transformation G for which

$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha). \tag{8.1}$$

This arises in the context of limited dependent variable models. For example, if Y_i is binary we might take G to be the inverse of a c.d.f. F, so that the model is

$$\Pr[Y_i = 1|X = x] = F\left(c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)\right) = m(x).$$

In this example we have restrictions not just on the mean function $m(x)$ but on the entire distribution. For example, it follows that $\text{var}(Y|X = x) = m(x)(1 - m(x))$.

An alternative is the transformation model

$$\Lambda(Y) = c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + \varepsilon,$$

where Λ is a strictly monotonic transformation and ε is independent of X. The transformation can be either parametric or nonparametric.

Interaction models.

In some cases one is interested in models for both mean and variance. So for example

$$\begin{aligned} Y &= c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + \sigma(X)\varepsilon \\ \sigma^2(X) &= c_\sigma + \sum_{\alpha=1}^{d} v_\alpha(X_\alpha), \end{aligned}$$

where m_α and v_α are unknown functions. One can instead model some transformation of $\sigma^2(X)$, say $F(\sigma^2)$ as being additive.

Rigby and Stasinopoulos (1995 etc.) consider the class of Mean and Dispersion Additive Models (MADAM's) in which $E(Y|X=x) = m(x)$ and $\text{var}(Y|X=x) = \phi(x)V(m(x))$, where $G_m(m(x)) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$ and $G_\phi(\phi(x)) = \sum_{\alpha=1}^{d} \phi_\alpha(x_\alpha)$ for unknown functions m_α, ϕ_α and known functions V, G_m, and G_ϕ.

Generalized additive separability is where there is some transformation G of m that is additively separable. There are a number of different forms of separability used in economics, see Leontieff (1947), Goldman and Uzawa (1964). Let $N_1, \ldots, N_s$ be a partition of $\{1, \ldots., d\}$.

Strong Separability: For functions m_α of $x^{(j)} = \{x_l; l \in N_j\}$ and scalar argument h

$$m(x) = h\{m_1(x^{(1)}) + \ldots + m_s(x^{(s)})\}$$

Weak Separability: For s-dim argument H,

$$m(x) = H\{m_1(x^{(1)}), \ldots, m_s(x^{(s)})\}$$

Pearce Separability: For additive functions m_α

$$m(x) = H\{m_1(x^{(1)}), \ldots, m_s(x^{(s)})\}.$$

Pinkse (2001) discusses nonparametric estimation in the weakly separable case where both H and the m_α functions are unknown. Horowitz (2001) discusses nonparametric estimation in the strongly separable case.

Homothetic functions. Suppose that there exist functions h and g such that

$$m(x) = h[g(x)],$$

where g is linearly homogeneous and h is strictly monotonic on its first element. Then we say that $m(x)$ is homothetic. The difficulty is estimating g, imposing and exploiting the restriction that the estimate be linearly homogeneous. In practice, we can estimate the function m by nonparametric regression but the problem is to infer what the functions h, g must be. Tripathi and Kim (2000) discuss estimation of homogenous functions.

Let V_i be an observed scalar and Z_i and W_i be observed vectors for $i = 1, \ldots, n$. Let $R(v, z, w)$ be some function that can be nonparametrically estimated, for example, $R(v, z, w)$ could equal $E(Y_i \mid V_i = v, Z_i = z, W_i = w)$, which is estimated with observations $\{Y_i, V_i, Z_i, W_i\}$. More generally, $R(v, z, w)$ could be a density, distribution, quantile, or hazard function, or $R(v, z, w)$ could be a utility or cost function derived from a set of estimated product or factor demands. Assume there exist unknown functions h and g and known strictly monotonic functions B_1, B_2, and B_3 such that

$$R(v, z, w) = h[B_1(B_2(v)B_3(g(z)), w] \tag{8.2}$$

where h is strictly monotonic on its first element.

One leading example of equation (8.2) is when B_1 is the natural logarithm and B_2 and B_3 are exponentiation, which gives

$$R(v, z, w) = h[v + g(z), w]. \tag{8.3}$$

When R is a conditional expectation, this an example of a generalized partly linear model with unknown link function similar to Horowitz (2001) and Horowitz and Mammen (2005). Equation (8.3) also arises in the nonparametric regression context given the model $V = -g(Z) + e$ for some error term e. If we strengthen the usual nonparametric regression assumption $E(e \mid Z) = 0$ to an independence assumption $e \perp Z, W$, then we obtain equation (8.3) where $R(v, z, w)$ is the unknown conditional distribution function of Z evaluated at $Z = z$, conditional on $V = v$ and $W = w$.

Another important example of equation (8.2) is when B_1, B_2, and B_3 are the identity functions, which gives

$$R(v, z, w) = h[vg(z), w]. \tag{8.4}$$

A function $r(x, w)$ is defined to be homothetically separable in x if and only if

$$r(x, w) = h[s(x), w] \tag{8.5}$$

where h is strictly monotonic in s and s is linearly homogeneous. Let v be one element of x that never equals zero, and let z be the vector of all the other elements of x divided by v. Alternatively,

rewrite x in polar coordinates as v, z, where v is length and z is direction. Either way, $s(x)$ is linearly homogeneous if and only if $s(x) = vg(z)$ for some unrestricted function g, so a function r is homothetically separable in x if and only if it has the form of R in equation (8.4). Similarly, when w is empty, equation (8.4) is equivalent to the definition of a function that is homothetic in x. For example, if v is labor and z is the capital labor ratio, then equation (8.4) equals the definition of a homothetic production function.

In applications of homothetic separability, r may have multiple homogeneous components, that is, $r(x_0, x_1, \ldots, x_k) = h[s_1(x_1), \ldots, s_K(x_K), x_0]$ for vectors $x_0, x_1, \ldots, x_K$. In this model, each homogeneous s_k function can be estimated separately by applying the method we propose to estimate g in equation (8.4), taking $x = x_k$ and w equal to the union of all the elements in $x_0, x_1, \ldots, x_K$ except x_k. Then, given estimates of each g_k function, the function h may be estimated by nonparametrically regressing r on $g_1, \ldots, g_K, x_0$. In the same way our estimator immediately extends to models like $R(v, z_0, z_1, \ldots, z_k) = h[v + \sum_k g_k(z_k), z_0]$, where each g_k is estimated by taking $z = z_k$ and w equal to the union of all the elements in $z_0, z_1, \ldots, z_K$ except z_k.

In many applications the functions h and g are of direct interest, e.g., in equation (8.4) the returns to scale of a homothetic production function is defined as the log derivative of h with respect to g. Even when h and g are not of direct interest, our estimator will still be useful for speeding the rate of convergence and for testing whether functions are homothetic, homothetically separable, or more generally if they satisfy equation (8.2).

Homothetic and homothetically separable functions are commonly used in models of consumer preferences and firm production, e.g., $r(x, w)$ could be a utility or consumer cost function recovered from estimated consumer demand functions via revealed preference theory, or it could be a directly estimated production or producer cost function. See, e.g., Blackorby, Primont, and Russell (1978), Lewbel (1991), (1997), Matzkin (1994), Primont and Primont (1994), and Zellner and Ryu (1998).

Linear index models with $s(x) = x^\top \beta$, are a very common semiparametric specification that arises in a variety of contexts, particularly limited dependent variable models. See Powell (1994) for a survey. Replacing a linear index $x^\top \beta$ with an arbitrary linearly homogeneous function $s(x)$ is a natural generalization, particularly in contexts where economic theory gives rise to homogeneity but not necessarily linearity, such as price indices or constant returns to scale technologies.

Matzkin (1992) provides a consistent estimator for the binary threshold crossing model $y = I[s(x) + \varepsilon \geq 0]$ where $s(x)$ is linearly homogeneous and ε is independent of x. This threshold crossing

model has $E(y|x) = h[s(x)]$ where h is the distribution function of $-\varepsilon$, and so is equivalent to our framework with $r(x) = E(y|x)$ and w empty. In an unpublished manuscript, Newey and Matzkin (1993) propose an estimator of Matzkin's (1992) model. Their estimator imposes the normalization $s(x_0) = v_0 g(z_0) = 1$, estimates h using $r(x) = R(v/v_0, z_0) = h(vg(z_0)/v_0) = h(v/v_0)$, and then given essentially inverts this h function estimate to obtain $s(x)$ from $r(x) = h[s(x)]$. Advantages of our estimators are that they can include w, they converge at a faster rate, they can attain oracle efficiency for s, and they yield estimates and limiting distributions that do not depend upon a single arbitrarily chosen point x_0.

Models satisfying equation (8.5) without imposing homogeneity on s are called weakly separable. See Gorman (1959), Goldman and Uzawa (1964) and Blackorby, Primont, and Russell (1978). Pinkse (2001) provides a general nonparametric estimator of weakly separable models. Pinkse's estimator identifies $s(x)$ up to an arbitrary monotonic transformation, whereas our estimator provides the unique (up to scale) linear homogeneous $s(x) = vg(z)$ (or equivalently g up to location in equation 8.3)) and exploits this structure of $s(x)$ to obtain a faster rate of convergence than Pinkse.

Many estimators exist for strongly or additively separable models, which are models of the form $r(x) = \sum_k s_k(x_k)$ where the functions $s_k(x_k)$ are unknown, and for generalized additively separable models, defined as $r(x) = h[\sum_k s_k(x_k)]$. Those most closely resembling our model include Härdle, Kim, and Tripathi (2001), who estimate additively separable models where the $s_k(x_k)$ functions are homogeneous, and Horowitz (2001) and Horowitz and Mammen (2005) who estimate generalized additively separable models where both h and s_k are unknown functions.

Matzkin (2003) considers models of the form $y = m(x, \varepsilon)$ with an unobserved scalar ε independent of x and, as one possible identifying assumption, m being linearly homogeneous in x and ε. In contrast, our model makes no assumptions about (and provides no estimates of) the role of unobservables other than a limiting distribution theory for an estimate of R, and allows for homothetic rather than just homogeneous dependence on x.

8.2 Estimation

Estimation of $m_\alpha(x_\alpha)$ by marginal integration can be carried out in the analogous fashion, since

$$g_\alpha(x_\alpha) = \int G(m(x))dQ_{-\alpha}(x_{-\alpha}) = c + m_\alpha(x_\alpha) + \sum_{\gamma \neq \alpha}^{d} \int m_\gamma(x_\gamma)dQ_{-\alpha}(x_{-\alpha}) \equiv m_\alpha(x_\alpha) + \mu_\alpha.$$

Let $\widehat{m}(x)$ be an estimator of $E(Y|X=x)$. Then let

$$\widehat{g}_\alpha(x_\alpha) = \int G(\widehat{m}(x))dQ_{-\alpha}(x_{-\alpha}).$$

We show how the instrumental variable approach can be applied to generalized additive models. Under additivity we have

$$m_\alpha(X_\alpha) = \frac{E\left[WG(m\left(X\right))|X_\alpha\right]}{E\left[W|X_\alpha\right]} \tag{8.6}$$

for the W defined in (7.6). Since $m\left(\cdot\right)$ is unknown, we need consistent estimates of $m\left(X\right)$ in a preliminary step, and then the calculation in (8.6) is feasible.

Since v is observed and the function B_2 is known, we may without loss of generality rewrite equation (8.2) as $R(v,z,w) = h[B_1(vB_3(g(z)),w]$ by redefining v as $B_2(v)$. Next, by defining $H(B_3,w) = h(B_1(B_3),w)$ and $G(z) = B_3(g(z))$, we may again without loss of generality rewrite equation (8.2) as

$$R(v,z,w) = H[vG(z),w] \tag{8.7}$$

We start with a consistent estimator $\widehat{R}(\cdot)$ of the function $R(\cdot)$, and provide nonparametric estimators for G and H. Estimates of the original g and h can then be readily recovered from the estimates of G and H if desired.

We could have instead started with the form $R(v,z,w) = h[B_1(B_2(v) + B_3(g(z)),w]$, simplifying as above to $R(v,z,w) = H[v+G(z),w]$, but this is slightly less general, because e.g., it only includes equation (8.4) as a special case when $vg(z)$ is constrained to be positive.

We first construct an initial consistent estimator of $G(z)$ by matching. For given values v,z,z',w suppose we can find a scalar u such that $R(v,z,w) = R(vu,z',w)$, a match. Then $u = U(z,z') = G(z)/G(z')$. The function $U(z,z')$ can be estimated by finding a zero of the function $\widehat{R}(v,z,w) - \widehat{R}(vu,z',w)$, averaging over a range of values of v and w to improve convergence properties. The function $G\left(z\right)$ is then estimated using a sample analog of $G\left(z\right) = U(z,z')/E[U(Z,z')]$ (averaged over a range of values of z'), which holds given the free scale normalization $E[G(Z)] = 1$. One advantage of a scale normalization like this over more simply normalizing at a point like $G(z_0) = 1$ is that the resulting limiting distributions at every point z will then not depend upon the distribution of $\widehat{R}(v,z_0,w)$.

Given the function G, the function H can be defined as the conditional expectation

$$H(\gamma,w) = E\left[R(V,Z,W) \mid VG(Z) = \gamma, W = w\right]. \tag{8.8}$$

Therefore, given an estimate $\widehat{G}$ of G, we can estimate the function H by a regression smooth of $\widehat{R}_i$ on $V_i\widehat{G}(Z_i), W_i$.

8.3 Asymptotic Properties

Theorem 19 (Linton and Härdle (1996)) *Suppose that A1-A6 hold and that F, G are twice continuously differentiable over the relevant compact interval. Then,*

$$n^{r/(2r+1)}\left[\widehat{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha)\right] \Longrightarrow N\left[b_\alpha(x_\alpha), v_\alpha^2(x_\alpha)\right]$$
$$n^{r/(2r+1)}\left[\widehat{m}(x) - m(x)\right] \Longrightarrow N\left[b(x), v^2(x)\right]$$

in distribution, where $b(x) = \sum_{\alpha=1}^d b_\alpha(x_\alpha)$, $v^2(x) = \sum_{\alpha=1}^d v_\alpha^2(x_\alpha)$ and

$$b_\alpha(x_\alpha) = \frac{\delta^r}{r!}\mu_r(K)\left[D^r m_\alpha(x_\alpha) - \int D^r m_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha\right]$$
$$v_\alpha^2(x_\alpha) = \delta^{-1}\|K\|_2^2 \int G'(m(x))^2 \frac{q_{-\alpha}^2(x_{-\alpha})}{f(x)}\sigma^2(x) dx_{-\alpha}$$

where $\mu_r(K) = \int K(t)t^r dt$ and $\|K\|_2^2 = \int K(t)^2 dt$.

Chapter 9

Appendix

9.1 CDF and Density Estimation

PROOF OF THEOREM 1. We assume for simplicity that F is continuous and strictly monotonic. Let $x_{\alpha k}$ be the value of x that satisfies $F(x_{\alpha k}) = \alpha/k$ for integer α, k with $\alpha \leq k$. For any x between $x_{\alpha k}$ and $x_{\alpha+1,k}$,

$$F(x_{\alpha k}) \leq F(x) \leq F(x_{\alpha+1,k}) \quad ; \quad F_n(x_{\alpha k}) \leq F_n(x) \leq F_n(x_{\alpha+1,k}),$$

while $0 \leq F(x_{\alpha+1,k}) - F(x_{\alpha k}) \leq 1/k$, so that

$$F_n(x) - F(x) \;\leq\; F_n(x_{\alpha+1,k}) - F(x_{\alpha k}) \leq F_n(x_{\alpha+1,k}) - F(x_{\alpha+1,k}) + \frac{1}{k}$$

$$F_n(x) - F(x) \;\geq\; F_n(x_{\alpha,k}) - F(x_{\alpha+1,k}) \geq F_n(x_{\alpha,k}) - F(x_{\alpha,k}) - \frac{1}{k}.$$

Therefore, for any x and k,

$$|F_n(x) - F(x)| \leq \max_{1\leq\alpha\leq k} |F_n(x_{\alpha k}) - F(x_{\alpha k})| + \frac{1}{k}. \tag{9.1}$$

Since the right hand side of (9.1) does not depend on x, we can replace the left hand side by $\sup_{-\infty<x<\infty} |F_n(x) - F(x)|$.

Now let $A_{\alpha k} = \{\omega : F_n(x_{\alpha k}) \to F(x_{\alpha k})\}$. By the pointwise strong law of large numbers, $\Pr(A_{\alpha k}) = 1$. Furthermore, $\Pr(A_k) = 1$, where

$$A_k = \bigcap_{\alpha=1}^{k} A_{\alpha k} = \left\{ \omega : \max_{1 \le \alpha \le k} |F_n(x_{\alpha k}) - F(x_{\alpha k})| \to 0 \right\},$$

since

$$\Pr(A_k^c) = \Pr\left\{ \bigcup_{\alpha=1}^{k} A_{\alpha k}^c \right\} \le \sum_{\alpha=1}^{k} \Pr(A_{\alpha k}^c) = 0.$$

It follows that $\Pr(A) = 1$, where $A = \bigcap_{k=1}^{\infty} A_k$. Therefore, we can make the right hand side of (9.1) arbitrarily small with probability one. ■

PROOF OF THEOREM 2. By the triangle inequality

$$\sup_{x \in \mathbb{R}} \left| \widehat{f}(x) - f(x) \right| \le \sup_{x \in \mathbb{R}} \left| \widehat{f}(x) - E[\widehat{f}(x)] \right| + \sup_{x \in \mathbb{R}} \left| E[\widehat{f}(x)] - f(x) \right|.$$

We can write

$$\widehat{f}(x) - E[\widehat{f}(x)] = \frac{1}{h} \int K\left(\frac{x-y}{h} \right) d[F_n(y) - F(y)].$$

Since K has bounded variation and $F_n(y) - F(y)$ is a continuous from the right step function with $F_n(-\infty) - F(-\infty) = 0$ and $F_n(\infty) - F(\infty) = 1$, we can apply integration by parts (Carter and van Brunt (2000)). Let S be the set of points y where both $K\left(\frac{x-y}{h}\right)$ and $F_n(y) - F(y)$ are discontinuous (this is a subset of the sample points $X_1, \ldots, X_n$), and define $\mu_g(a) = \mu_g(a^+) - \mu_g(a^-)$. Then

$$\begin{aligned} \frac{1}{h} \int K\left(\frac{x-y}{h} \right) d[F_n(y) - F(y)] \;=\; & -\frac{1}{h} \int [F_n(y) - F(y)] dK\left(\frac{x-y}{h} \right) \\ & + \mu_{K\left(\frac{x-\cdot}{h}\right)(F_n - F)}(\mathbb{R}) + \sum_{a \in S} \mu_{K\left(\frac{x-\cdot}{h}\right)}(a) \mu_{F_n - F}(a). \end{aligned}$$

Since $F_n(\pm\infty) - F(\pm\infty) = 0$, $\mu_{K\left(\frac{x-\cdot}{h}\right)(F_n - F)}(\mathbb{R}) = 0$. The last term is also zero with probability one because the discontinuity points of $K\left(\frac{x-y}{h}\right)$ (which are countable in number by bounded variation) do not coincide with the discontinuity points of $F_n(y) - F(y)$, when X is continuously distributed. Therefore,

$$\begin{aligned} \sup_{x \in \mathbb{R}} \left| \widehat{f}(x) - E[\widehat{f}(x)] \right| \;&=\; \sup_{x \in \mathbb{R}} \left| \int K_h(x-y) d[F_n(y) - F(y)] \right| \\ &\le\; \sup_{x \in \mathbb{R}} \int |F_n(y) - F(y)| d|K_h(x-y)| \\ &\le\; \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \times \frac{1}{h} TVAR(|K|) \\ &\le\; \frac{Z_n}{n^{1/2} h} \end{aligned}$$

for some tight sequence of random variables $Z_n = \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)|$, where $TVAR$ denotes total variation. This holds in probability but also almost surely.

Then note that

$$E[\widehat{f}(x)] - f(x) = \frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy \equiv f * K_h.$$

This is called the convolution of f with K_h and is an object well treated in real analysis. In particular, there are many different results covering $f * K_h \to f$ in various senses under various conditions.

PROPOSITION (FOLLAND) *Suppose that $K \in L^1$ and $\int K = 1$.*

(a) If $f \in L^p$ $(1 \le p \le \infty)$, then $f * K_h \to f$ in the L^p norm as $h \to 0$

(b) If f is bounded and uniformly continuous, then $f * K_h \to f$ uniformly as $h \to 0$

(c) If $f \in L^\infty$ and f is continuous on an open set U, then $f * K_h \to f$ uniformly on compact sets of U as $h \to 0$

(d) Suppose that $|K(x)| \le C(1+|x|)^{-d-\epsilon}$ for some $C, \epsilon \in (0,\infty)$. If $f \in L^p$ $(1 \le p \le \infty)$, then $f * K_h \to f$ for every x in the Lebesgue set of f, in particular for every x at which f is continuous.

Suppose that f has compact support. For any M we can write $K(x) = K^A(x) + K^B(x)$, where

$$K^A(x) = K(x)1(|x| \le M) \text{ and } K^B(x) = K(x)1(|x| > M).$$

Then $f * K_h = f * K_h^A + f * K_h^B$ and

$$\int |f * K_h - f|\, dx \le \int \left| f * K_h^A - f\int K_h^A \right| dx + \int \left| f\int K_h^B \right| dx + \int |f| \int |K_h^B| dx.$$

The last two terms are bounded by $2\int |K_h^B|dx \to 0$ as $h \to 0$. Define the modulus of continuity ω_f of a function f,

$$\omega_f(t) = \sup_{x\in\mathbb{R}} \sup_{|z|\le t} |f(x+z) - f(x)|\,.$$

For a uniformly continuous function f, $\omega_f(t) \to 0$ as $t \to 0$. Then

$$\begin{aligned}\int \left| f * K_h^A - f\int K_h^A \right| dx &\le \int\int |f(x-y) - f(x)|\, K_h^A(y)dydx \\ &\le 2C\omega_f(2Mh)\int K(y)dy\end{aligned}$$

where $[-C, C]$ is the support of f. This last quantity tends to zero for M fixed as $h \to 0$.

To prove lemma for general f, approximate f by g with compact support such that $\int |f-g| < \epsilon$ for some $\epsilon > 0$. Then

$$\begin{aligned}\int |f * K_h - f|\, dx &\leq \int |f * K_h - g * K_h|\, dx + \int |f-g|\, dx + \int |g * K_h - g|\, dx \\ &\leq 2\int |f-g|\, dx + \int |g * K_h - g|\, dx.\end{aligned}$$

Then since we can take ϵ arbitrarily small the result is established. ■

9.2 General Theory for Local Nonlinear Estimators

We provide theory for a general class of estimators defined as finding zeros of an estimated moment conditions $G_n(\theta; x) \in \mathbb{R}^q$, where $G_n(\theta; x)$ is defined on the set $\widehat{\Theta} \times \widehat{\mathcal{X}}$. The population moment condition $G(\theta; x)$ is defined on the set $\Theta \times \mathcal{X}$, where $\Theta, \mathcal{X}$ are both compact and G is continuous. For a vector $a \in \mathbb{R}^q$ let $||a|| = (a^\top a)^{1/2}$. For sets A, B, let $\rho_\Delta(A, B) = \lambda(A \Delta B)$, where Δ denotes the symmetric difference, i.e., $A \Delta B = (A \cap B^c) \cup (A^c \cup B)$, and λ is Lebesgue measure. The following proposition is an extension of a standard consistency result of Pakes and Pollard (1989, Theorem 3.1). It is closely related to Lemma A1 of Newey and Powell (2003), although our results are more applicable for pointwise theory. See also Horowitz and Mammen (2004). We do not require continuity in θ or x of the estimated moment function $G_n(\theta; x)$.

PROPOSITION 1. *Suppose that $\widehat{\Theta} \subseteq \Theta, \widehat{\mathcal{X}} \subseteq \mathcal{X}$ are sets that satisfy $\rho_\Delta(\widehat{\Theta}, \Theta) \to 0$ a.s. and $\rho_\Delta(\widehat{\mathcal{X}}, \mathcal{X}) \to 0$ a.s.. Suppose further that:*

(i) *For each $x \in \widehat{\mathcal{X}}$, $\widehat{\theta}(x) \in \widehat{\Theta} \subset \mathbb{R}^p$ (with $p \leq q$) is any sequence such that*

$$\sup_{x \in \widehat{\mathcal{X}}} \left(||G_n(\widehat{\theta}(x); x)|| - \inf_{\theta \in \widehat{\Theta}} ||G_n(\theta; x)|| \right) = o_p(1).$$

(ii) *Suppose that for all $\epsilon > 0$ there exists $\delta > 0$ such that $\inf_{x \in \mathcal{X}} \inf_{||\theta - \theta_0(x)|| \geq \epsilon} ||G(\theta; x)|| > \delta$.*

(iii) *As $n \to \infty$, $\sup_{x \in \widehat{\mathcal{X}}} \sup_{\theta \in \widehat{\Theta}} \|G_n(\theta; x) - G(\theta; x)\| \xrightarrow{P} 0$.*

Then

$$\sup_{x \in \widehat{\mathcal{X}}} \left\| \widehat{\theta}(x) - \theta_0(x) \right\| \xrightarrow{P} 0. \tag{9.2}$$

PROOF OF PROPOSITION 1. First, define for all $x \in \widehat{\mathcal{X}}$,

$$\overline{\theta}_0(x) = \arg \inf_{\theta \in \widehat{\Theta}} \|G(\theta; x)\|.$$

The assumption that $\rho_\Delta(\widehat{\Theta}, \Theta) \to 0$ a.s. implies that there is a sequence of points $\theta_n(x) \in \widehat{\Theta}$ with $\theta_n(x) \to U_0(x)$ a.s. Then by the continuity of G we have that $||G(\theta_n(x); x)|| \to ||G(\theta_0(x); x)|| = 0$ a.s., whence $\theta_n(x)$ is an approximate minimizer of G over $\widehat{\Theta}$ as required. Then, by uniform continuity of G over $\Theta \times \mathcal{X}$ and the identification condition (ii) we have

$$\sup_{x \in \widehat{\mathcal{X}}} \left| \overline{\theta}_0(x) - \theta_0(x) \right| \longrightarrow 0 \ a.s.$$

$$\sup_{x \in \widehat{\mathcal{X}}} \left\| G(\overline{\theta}_0(x); x) - G(\theta_0(x); x) \right\| \longrightarrow 0 \ a.s.$$

Furthermore, we have

$$\begin{aligned}
\inf_{x \in \widehat{\mathcal{X}}} \left\| G(\widehat{\theta}(x); x) \right\| &\leq \sup_{x \in \widehat{\mathcal{X}}} \left\| G(\widehat{\theta}(x); x) \right\| \\
&\leq \sup_{x \in \widehat{\mathcal{X}}} \left\| G_n(\widehat{\theta}(x); x) \right\| + \sup_{x \in \widehat{\mathcal{X}}} \left\| G_n(\widehat{\theta}(x); x) - G(\widehat{\theta}(x); \theta, \theta) \right\| \\
&\leq \sup_{x \in \widehat{\mathcal{X}}} \left\| G_n(\overline{\theta}_0(x); x) \right\| + o_p(1) \\
&\leq \sup_{x \in \widehat{\mathcal{X}}} \left\| G(\overline{\theta}_0(x); x) \right\| + \sup_{x \in \widehat{\mathcal{X}}} \sup_{\theta \in \widehat{\theta}} \| G_n(\theta; x) - G(\theta; x) \| + o_p(1) = o_p(1),
\end{aligned}$$

using the triangle inequality, (i), and (iii). Then note that

$$\Pr \left[\sup_{x \in \widehat{\mathcal{X}}} \left\| \widehat{\theta}(x) - \theta_0(x) \right\| > \epsilon \right] \leq \Pr \left[\inf_{x \in \widehat{\mathcal{X}}} \left\| G(\widehat{\theta}(x); x) \right\| > \delta \right],$$

because $0 < \inf_{x \in \mathcal{X}} \inf_{||\theta - \theta_0(x)|| \geq \epsilon} ||G(\theta; x)|| \leq \inf_{x \in \widehat{\mathcal{X}}} ||G(\widehat{\theta}(x); x)||$, whenever $\sup_{x \in \widehat{\mathcal{X}}} ||\widehat{\theta}(x) - \theta_0(x)|| > \epsilon$. Therefore, (9.2) follows. ■

The following result is an extension of Theorem 3 of Pakes and Pollard (1989). The extension is to estimators with more general rates of convergence; it also takes account of the failure of a global stochastic equicontinuity condition for nonparametric estimators by providing a more targeted condition (iii)(a),(b), which uses a local stochastic equicontinuity condition [see Müller and Stadtmüller (1987) for comparison]. Also, condition (iv) reflects the nonparametric setting through a bias term and has a strengthening for purposes of uniformity. Notice that we do not require any smoothness on the estimated moment function $G_n(\theta; x)$ although it should be well approximated by the smooth function G.

PROPOSITION 2. *Suppose that the following conditions hold for some sequence $\alpha_n \to \infty$:*

(i) *For each* x, $\widehat{\theta}(x)$ *is any sequence such that*

$$\sup_{x\in\widehat{\mathcal{X}}}\left(||G_n(\widehat{\theta}(x);x)|| - \inf_{\theta\in\widehat{\Theta}}||G_n(\theta;x)||\right) = o_p(\alpha_n^{-1}).$$

(ii) (a) *The function* $G(\theta;x)$ *is differentiable in* U *at* $\theta=\theta_0(x)$ *with derivative matrix* $I(x)=\partial G(\theta_0(x);x)/\partial\theta$ *of full rank uniformly in* $x\in\mathcal{X}$.

(b) *The derivative matrix* $\partial G(\theta;x)/\partial\theta$ *is uniformly in* $x\in\mathcal{X}$ *Hölder continuous in* U *with exponent* $\varsigma>0$.

(iii) *There is a sequence* $\delta_n\to 0$ *such that*

(a) *For some sequence of positive numbers* $\{\epsilon_n\}$ *that converges to zero*

$$\sup_{x\in\widehat{\mathcal{X}}}\sup_{||\theta-\theta_0(x)||\le\epsilon_n}\delta_n^{-1}\left\|G_n(\theta;x)-G(\theta;x)\right\|\xrightarrow{P}0.$$

(b) *For every sequence of positive numbers* $\{\epsilon_n\}$ *that converges to zero*

$$\sup_{x\in\widehat{\mathcal{X}}}\sup_{\delta_n^{-1}||\theta-\theta_0(x)||\le\epsilon_n}\alpha_n\left\|G_n(\theta;x)-G(\theta;x)-G_n(\theta_0(x);x)\right\|\xrightarrow{P}0.$$

(iv) *For some deterministic sequence* $b_n(x)\to 0$ *with* $\limsup\alpha_n b_n(x)<\infty$,

$$\alpha_n\sup_{x\in\widehat{\mathcal{X}}}\|G_n(\theta_0(x);x)-b_n(x)-Z_n(x)\|=o_p(1),\ \text{where}$$

$$\text{for each } x,\ \alpha_n Z_n(x)\Longrightarrow N(0,V(x))$$

$$\text{for some } r\ge 1,\ \sup_{x\in\widehat{\mathcal{X}}}\|Z_n(x)\|=O_p(\alpha_n^{-1}\log^r n).$$

(v) $\theta_0(x)$ *is an interior point of* θ*for all* x.

Then

$$\alpha_n\sup_{x\in\widehat{\mathcal{X}}}\left\|\widehat{\theta}(x)-\theta_0(x)-(I^\top I)^{-1}I^\top(x)b_n(x)-(I^\top I)^{-1}I^\top(x)Z_n(x)\right\|=o_p(1),\ \text{where}$$

$$\text{for each } x,\ \alpha_n(I^\top I)^{-1}I^\top(x)Z_n(x)\Longrightarrow N(0,(I^\top I)^{-1}I^\top VI(I^\top I)^{-1}(x)).$$

Proof of Proposition 2. The proof is similar to Theorem 3 of Pakes and Pollard (1989). We first do the pointwise argument for a fixed x. Condition (ii) transfers the rate on G_n in (iii)(a) to the

same rate on $\widehat{\theta}(x) - \theta_0(x)$ because for all x, $\|G(\theta; x)\| \geq C(x) \|\theta - \theta_0(x)\|$ for θ close to $\theta_0(x)$, where $\inf_{x \in \mathcal{X}} C(x) > 0$. Therefore,

$$\delta_n^{-1} ||\widehat{\theta}(x) - \theta_0(x)|| = o_p(1) \tag{9.3}$$

for each x. Having obtained δ_n-consistency of $\widehat{\theta}(x)$, we then use condition (iii)(b) to obtain α_n-consistency along the lines of Pakes and Pollard (1989). Specifically, there exists a sequence $\epsilon_n \to 0$ such that $\Pr[\delta_n^{-1} ||\widehat{\theta}(x) - \theta_0(x)|| \geq \epsilon_n] \to 0$ and so the supremum in (iii)(a) covers $\widehat{\theta}(x)$ with probability tending to one. It follows that by the triangle inequality and (iii)(a) we have with probability tending to one

$$\begin{aligned} & \left\| G(\widehat{\theta}(x); x) \right\| - \left\| G_n(\widehat{\theta}(x); x) \right\| - \|G_n(\theta_0(x); x)\| \\ \leq & \left\| G_n(\widehat{\theta}(x); x) - G(\widehat{\theta}(x); x) - G_n(\theta_0(x); x) \right\| \leq o_p(\alpha_n^{-1}). \end{aligned} \tag{9.4}$$

Therefore,

$$\begin{aligned} \left\| G(\widehat{\theta}(x); x) \right\| &\leq \left\| G_n(\widehat{\theta}(x); x) \right\| + \|G_n(\theta_0(x); x)\| + o_p(\alpha_n^{-1}) \\ &\leq 2 \|G_n(\theta_0(x); x)\| + o_p(\alpha_n^{-1}) \leq 2 \|b_n(x) + Z_n(x)\| + o_p(\alpha_n^{-1}) \end{aligned} \tag{9.5}$$

by (i) and (iv). It follows that $||G(\widehat{\theta}(x); x)|| = O_p(\alpha_n^{-1})$ and so

$$\alpha_n ||\widehat{\theta}(x) - \theta_0(x)|| = O_p(1). \tag{9.6}$$

Let

$$L_n(\theta, x) = I(x) \cdot (\theta - \theta_0(x)) + b_n(x) + Z_n(x).$$

By similar arguments to Pakes and Pollard (1989) one shows that $||G_n(\widehat{\theta}(x); x) - L_n(\widehat{\theta}(x), x)|| = o_p(1)$. The minimizing value of $||L_n(\theta, x)||$ is $\theta^*(x) = \theta_0(x) - (I^\top I)^{-1} I^\top (x)(b_n(x) + Z_n(x))$ and it can be shown that $||G_n(\theta^*(x); x) - L_n(\theta^*(x), x)|| = o_p(\alpha_n^{-1})$. It follows that $||\widehat{\theta}(x) - \theta^*(x)|| = o_p(\alpha_n^{-1})$. The pointwise asymptotic normality of $\widehat{\theta}(x)$ then follows along the lines of their proof.

The extension to uniformity over x proceeds as follows. By the triangle inequality, we have

$$\begin{aligned} & \sup_{x \in \widehat{\mathcal{X}}} \left\| G(\widehat{\theta}(x); x) \right\| - \inf_{x \in \widehat{\mathcal{X}}} \left\| G_n(\widehat{\theta}(x); x) \right\| - \inf_{x \in \widehat{\mathcal{X}}} \|G_n(\theta_0(x); x)\| \\ \leq & \sup_{x \in \widehat{\mathcal{X}}} \left\| G_n(\widehat{\theta}(x); x) - G(\widehat{\theta}(x); x) - G_n(\theta_0(x); x) \right\| \leq o_p(\alpha_n^{-1}) \end{aligned} \tag{9.7}$$

from which we obtain that

$$\sup_{x\in\widehat{\mathcal{X}}}\left\|G(\widehat{\theta}(x);x)\right\| \leq 2\sup_{x\in\widehat{\mathcal{X}}}\|G_n(\theta_0(x);x)\| + o_p(\alpha_n^{-1}) \leq 2\sup_{x\in\widehat{\mathcal{X}}}\|b_n(x)+Z_n(x)\| + o_p(\alpha_n^{-1}).$$

Then using $\sup_{x\in\widehat{\mathcal{X}}}\|G(\theta(x);x)\| \geq (\inf_{x\in\widehat{\mathcal{X}}} C(x))\sup_{x\in\widehat{\mathcal{X}}}\|\theta(x)-\theta_0(x)\|$ for $\theta(x)$ uniformly close to $\theta_0(x)$, one obtains that $\sup_{x\in\widehat{\mathcal{X}}}||\widehat{\theta}(x)-\theta_0(x)|| = O_p(\alpha_n^{-1}\log^r n)$. By the triangle inequality

$$\begin{aligned}\sup_{x\in\widehat{\mathcal{X}}}\left\|G_n(\widehat{\theta}(x);x)-L_n(\widehat{\theta}(x),x)\right\| \;\leq\; & \sup_{x\in\widehat{\mathcal{X}}}\left\|G_n(\widehat{\theta}(x);x)-G(\widehat{\theta}(x);x)-G_n(\theta_0(x);x)\right\| \\ & +\sup_{x\in\widehat{\mathcal{X}}}\|G_n(\theta_0(x);x)-b_n(x)-Z_n(x)\| \\ & +\sup_{x\in\widehat{\mathcal{X}}}\left\|G(\widehat{\theta}(x);x)-I(x)\cdot(\widehat{\theta}(x)-\theta_0(x))\right\| \\ =\; & o_p(\alpha_n^{-1}),\end{aligned}$$

since by the mean value theorem and the H*ö*lder continuity condition (ii)(b)

$$\sup_{x\in\widehat{\mathcal{X}}}\left\|G(\widehat{\theta}(x);x)-I(x)\cdot(\widehat{\theta}(x)-\theta_0(x))\right\| = O_p((\alpha_n^{-1}\log^r n)^{1+\varsigma}) = o_p(\alpha_n^{-1}).$$

By similar arguments one shows that $\sup_{x\in\widehat{\mathcal{X}}}||G_n(\theta^*(x);x)-L_n(\theta^*(x),x)|| = o_p(\alpha_n^{-1})$. Finally, one obtains that $\sup_{x\in\widehat{\mathcal{X}}}||\widehat{\theta}(x)-\theta^*(x)|| = o_p(\alpha_n^{-1})$ by the same arguments as in Pakes and Pollard (1989).

■

9.2.1 Consistency of the Nadaraya-Watson Estimator

Consider the following first order condition

$$G_n(\theta,x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\{Y_i-\theta\}.$$

We seek to establish the consistency and asymptotic normality of the zero of $G_n(\theta,x)$, which is the Nadaraya-Watson kernel estimator. We shall suppose that x is an interior point of the support of X, that is, there is an open ball $B(x,\epsilon)$ [for some small ϵ] that is totally contained in the support of X. We take the corresponding limit to be

$$G(\theta,x) = \{m(x)-\theta\}f_X(x).$$

Then, it is clear that $\theta_0 = m(x)$ is the unique zero of $G(\theta, x)$ so that identification condition is automatically satisfied. We must show the uniform convergence of G_n to G. We have

$$\|G_n(\theta,x) - G(\theta,x)\| \leq \|G_n(\theta,x) - E_X G_n(\theta,x)\| + \|E_X G_n(\theta,x) - EG_n(\theta,x)\| + \|EG_n(\theta,x) - G(\theta,x)\|,$$

where E_X denotes expectation conditional on $X_1, \ldots, X_n$, and it suffices to work on these three terms separately. First, note that

$$E_X G_n(\theta,x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\{m(X_i) - \theta\}$$

$$EG_n(\theta,x) = EE_X G_n(\theta,x) = EK_h(x - X_i)\{m(X_i) - \theta\} = \int K_h(x - X)\{m(X) - \theta\} f(X) dX$$

by iterated expectation. Then apply a change of variables $X \longmapsto u = (x - X_i)/h$ to write

$$\int_{\underline{x}}^{\overline{x}} K_h(x - X)\{m(X) - \theta\} f(X) dX = \int_{\frac{x-x}{h}}^{\frac{\overline{x}-x}{h}} K(u)\{m(x + uh) - \theta\} f(x + uh) du,$$

where $\underline{x}$ and $\overline{x}$ are the lower and upper limits respectively of the support of X. Provided the point x is such that $\overline{x} - x > ch$ and $x - \underline{x} > ch$ for some finite c and the kernel K is of finite support, we can replace the limits of integration by those of the kernel [e.g., $-1, 1$]. In the sequel we shall assume this is this case. Therefore, by dominated convergence,

$$\sup_{\theta\in\Theta} \|EG_n(\theta,x) - G(\theta,x)\| = \sup_{\theta\in\Theta} \left\| \int_{-1}^{1} K(u)\{m(x + uh) - \theta\} f(x + uh) du - \{m(x) - \theta\} f_X(x) \right\| = o(1)$$

for any compact set Θ, provided m and f are continuous at x. Second,

$$G_n(\theta,x) - E_X G_n(\theta,x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i,$$

where ε_i is an independent sequence satisfying $E(\varepsilon_i | X_i) = 0$ with probability one. By a law of large numbers for triangular arrays of independent random variables [of the form $\sum_{i=1}^{n} Z_{ni}$], we get, provided $\sup_n E\left[|K_h(x - X_i)\varepsilon_i|\right] < \infty$, that

$$G_n(\theta,x) - E_X G_n(\theta,x) = o_p(1),$$

and since this random sequence does not depend on θ, convergence is uniform over $\theta \in \Theta$. Likewise

$$E_X G_n(\theta,x) - EG(\theta,x) = \frac{1}{n}\sum_{i=1}^{n} \{K_h(x - X_i)(m(X_i) - \theta) - E[K_h(x - X_i)(m(X_i) - \theta)]\} = \frac{1}{n}\sum_{i=1}^{n} \eta_n(X_i, \theta)$$

is a sum of independent mean zero random variables that is $o_p(1)$ by the same reasoning. The uniformity in θ comes from the linear way in which this enters and the compactness of Θ. Specifically, $\sup_n \sup_{\theta\in\Theta} E\left[|\eta_n(X_i,\theta)|\right] < \infty$.

9.2.2 Asymptotic Normality of the Nadaraya-Watson Estimator

In our case, we take $1/\varrho_n = \max\{1/\sqrt{nh}, \xi_n\}$, where

$$\xi_n = \sup_{\theta\in\Theta} \|EG_n(\theta,x) - G(\theta,x)\|.$$

Clearly,

$$\|G(\theta,x)\| = |m(x)-\theta| f_X(x) = C|\theta-\theta_0|,$$

where $C = f_X(x)$, so that (ii) is satisfied. We now turn to (iii). Let $r_\theta(z) = \{m(z)-\theta\}f(z)$ for any z. Now, provided r_θ is twice continuously differentiable at x, we have

$$\begin{aligned} EG_n(\theta,x) - G(\theta,x) &= \int K(u)\{r_\theta(x+uh) - r_\theta(x)\}du \\ &= hr'_\theta(x)\int K(u)udu + \frac{h^2}{2}\int u^2K(u)r''_\theta(x^*(u,h))du, \end{aligned}$$

where $x^*(u,h)$ lies between x and $x+uh$. Provided $\int K(u)udu = 0$, the first term drops out. By dominated convergence we then have

$$EG_n(\theta,x) - G(\theta,x) = \frac{h^2}{2}\int u^2K(u)r''_\theta(x)du\{1+o(1), \tag{9.8}$$

where $r''_\theta(x) = (mf)''(x) - \theta f''(x)$ and the error is uniform in $\theta \in \Theta$. Furthermore, provided $E[\frac{1}{h}K^2\left(\frac{x-X_i}{h}\right)\varepsilon_i^2] < \infty$, we have

$$G_n(\theta,x) - E_XG_n(\theta,x) = \frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\varepsilon_i = O_p(1/\sqrt{nh}),$$

because this random sequence is mean zero. We have

$$\begin{aligned} \frac{1}{nh}E\left[\frac{1}{h}K^2\left(\frac{x-X_i}{h}\right)\varepsilon_i^2\right] &= \frac{1}{nh}\int K^2(u)\sigma^2(x+uh)f_X(x+uh)du \\ &= \frac{1}{nh}\sigma^2(x)f_X(x)\int K^2(u)du\{1+o(1)\}. \end{aligned}$$

Furthermore,

$$E_X G_n(\theta,x) - EG_n(\theta,x) = \frac{1}{n}\sum_{i=1}^{n}\{K_h(x-X_i)(m(X_i)-\theta) - E[K_h(x-X_i)(m(X_i)-\theta)]\} = o_p(1/\sqrt{nh})$$

uniformly in $\|\theta-\theta_0\| \le \delta_n$. Again this is a sum of mean zero independent random variables and

$$\frac{1}{nh}E\left[\frac{1}{h}K^2\left(\frac{x-X_i}{h}\right)(m(X_i)-\theta)^2\right] = \frac{1}{nh}\int K^2(u)(m(x+uh)-\theta)^2 f_X(x+uh)du = o(1/nh),$$

for $\|\theta-\theta_0\| \le \delta_n$.

Finally, $G_n(\theta_0,x) = O_p(1/\sqrt{nh})$ by the same arguments.

In this case, θ is scalar and $\Gamma = -f_X(x)$. (iii) It clearly suffices to show that

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \|\varrho_n[EG_n(\theta,x) - G(\theta,x)] - \varrho_n[EG_n(\theta_0) - G(\theta_0)]\| = o(1),$$

which follows because

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \|EG_n(\theta,x) - G(\theta,x)\| = \sup_{\|\theta-\theta_0\|\le\delta_n}\left|\int u^2 K(u)\{r''_\theta(x^*(u,h)) - r''_{\theta_0}(x^*(u,h))\}du\right| = o(1)$$

and

$$\begin{aligned}&\sup_{\|\theta-\theta_0\|\le\delta_n} \|E_X G_n(\theta,x) - EG(\theta,x)\| \\ =\ &\sup_{\|\theta-\theta_0\|\le\delta_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\{K_h(x-X_i) - EK_h(x-X_i)\}(m(x)-\theta)\right\| = o_p(1/\sqrt{nh}).\end{aligned}$$

The final result combines a central limit theorem for

$$G_n(\theta_0) - E_X G_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\varepsilon_i,$$

with the bias result (9.8).

By the Lindeberg CLT for triangular arrays we have

$$\frac{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\varepsilon_i}{\sqrt{\operatorname{var}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\varepsilon_i \,|X_1,\ldots,X_n\right]}} \equiv \frac{\sum_{i=1}^{n} w_{ni}\varepsilon_i}{\sqrt{\sum_{i=1}^{n} w_{ni}^2\sigma^2(X_i)}} \Rightarrow N(0,1),$$

since the summands are independent and identically distributed for each n, provided the negligibility condition holds. Since the limit distribution does not depend on $X_1,\ldots,X_n$, it suffices to show that

this condition holds in the conditional distribution with probability one. More precisely, we have to check that with probability one

$$\frac{1}{\sum_{i=1}^n w_{ni}^2\sigma^2(X_i)}\sum_{i=1}^n E\left[w_{ni}^2\varepsilon_i^2\mathbf{1}\left(|w_{ni}\varepsilon_i| > c\sqrt{\sum_{i=1}^n w_{ni}^2\sigma^2(X_i)}\right)\right] \to 0 \tag{9.9}$$

as $n \to \infty$ for all $c > 0$. Letting $v_{ni} = w_{ni}\sigma(X_i)\Big/\sqrt{\sum_{i=1}^n w_{ni}^2\sigma^2(X_i)}$ and $\eta_i = \varepsilon_i/\sigma(X_i)$, (**??**) is bounded by

$$\max_{1\le i\le n} E\left[v_{ni}^2\eta_i^2\mathbf{1}\left(|\eta_i| > c/v_{ni}\right)\right],$$

which tends to zero provided that

$$\max_{1\le i\le n} v_{ni} = \frac{\max_{1\le i\le n}|w_{ni}\sigma(X_i)|}{\sqrt{\sum_{i=1}^n w_{ni}^2\sigma^2(X_i)}} \to 0. \tag{9.10}$$

We have already shown that $\sum_{i=1}^n w_{ni}^2\sigma^2(X_i) = O_p(1/nh)$. Provided K and $\sigma^2(\cdot)$ are bounded

$$\max_{1\le i\le n}|w_{ni}\sigma(X_i)| \le \frac{1}{nh}\max_{1\le i\le n}\left|K\left(\frac{x-X_i}{h}\right)\right|\cdot\max_{1\le i\le n}\sigma(X_i) = O(1/nh),$$

and (**??**) is satisfied. To make this hold with probability one requires a bit more work.

9.3 Bandwidth Selection Result

PROOF. We shall simplify the problem considerably. Suppose that in fact m, f are twice continuously differentiable so that the optimal bandwidth is asymptotically of the form $\theta_{opt}n^{-1/5}$ for some θ_{opt} and that

$$H_n = \left\{\theta n^{-1/5} : \theta \in [\underline{\theta}, \overline{\theta}] \text{ for some } 0 < \underline{\theta} < \overline{\theta} < \infty\right\}.$$

In this case, $d_A(\widehat{m}, m) = O(n^{-4/5})$ for any $h \in H_n$. Let $Q_n(\theta) = n^{4/5}\cdot d_A(\widehat{m}, m)$ and

$$R_n(\theta) = n^{4/5}\left[CV(\theta n^{-1/5}) - \frac{1}{n}\sum_{i=1}^n\left\{Y_i - m(X_i)\right\}^2\pi(X_i)\right].$$

The modified theorem would hold provided that

$$\sup_{\theta\in[\underline{\theta},\overline{\theta}]}|Q_n(\theta) - R_n(\theta)| \to_p 0$$

and that the probability limit of $Q_n(\theta)$ is uniquely maximized.

Note that

$$\begin{aligned} Q_n(\theta) - R_n(\theta) &= n^{4/5} \cdot \frac{1}{n} \sum_{i=1}^{n} \{\widehat{m}_i(X_i) - m(X_i)\} \varepsilon_i \pi(X_i) \\ &= n^{4/5} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{B}(X_i)}{\widehat{f}(X_i)} \varepsilon_i \pi(X_i) + n^{4/5} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{V}(X_i)}{\widehat{f}(X_i)} \varepsilon_i \pi(X_i) \\ &= T_{n1} + T_{n2}. \end{aligned}$$

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

$$n^{8/5} \cdot \frac{1}{n^2} \sum_{i=1}^{n} \left\{ \frac{\widehat{B}(X_i)}{\widehat{f}(X_i)} \right\}^2 \sigma^2(X_i)\pi(X_i) = O_p(n^{3/5} \cdot n^{-4/5}) = o_p(1)$$

uniformly in θ. The second term can not be handled similarly by splitting $\widehat{V}(X_i)$ from ε_i, and we must examine its structure more closely. The term T_{n2} is a U-statistic of the form

$$n^{4/5} \sum \sum_{i \neq i} U_{ii}(\theta) \quad ; \quad U_{ii}(\theta) = \frac{n^{4/5}}{n(n-1)h(\theta)} \sum \sum_{i \neq i} K\left(\frac{X_i - X_i}{h(\theta)} \right) \varepsilon_i \varepsilon_i \pi(X_i) \widehat{f}^{-1}(X_i)$$

whose full analysis we will postpone till later. However,

$$\begin{aligned} \text{var}\left[\sum \sum_{i \neq i} U_{ii}(\theta) \,|X_1, \ldots, X_n \right] &= \sum \sum_{i \neq i} \left[E\left\{ U_{ii}^2(\theta) \right\} |X_1, \ldots, X_n \right] \\ &= \frac{n^{8/5}}{n^2(n-1)^2 h^2(\theta)} \sum \sum_{i \neq i} K^2\left(\frac{X_i - X_i}{h(\theta)} \right) O_p(1) \\ &= O_p(\frac{n^{8/5}}{n^2 h(\theta)}) = o_p(1). \end{aligned}$$

In fact, $T_{n2}(\theta)$ is uniformly $o_p(1)$ ■

9.4 Uniform Consistency

We provide an outline theory for a general class of estimators

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

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

$$|\widehat{m}(x) - m(x)| \leq \left| \sum_{i=1}^{n} w_{ni}(x)\varepsilon_i \right| + \left| \sum_{i=1}^{n} w_{ni}(x)m(X_i) - m(x) \right|,$$

where the second term is purely deterministic conditional on $X_1, \ldots, X_n$ – by Taylor series expansion it can be shown to be uniformly $O(h^2)$. We shall establish that

$$\sup_{x \in C} \left| \sum_{i=1}^{n} w_{ni}(x)\varepsilon_i \right| = O_p\left(\sqrt{\frac{\log n}{nh}} \right) := O_p(\delta_n) \tag{9.11}$$

under suitable conditions.

We use the theorem and proof given in Müller (1988, p161).

1. $E\left[|\varepsilon_i|^s\right] < \infty$ for some $s > 2\frac{1}{2}$.

2. $\max_{1 \leq i \leq n} |w_{ni}(x) - w_{ni}(x')| \leq k_1 |x - x'|^\gamma$, $\gamma > 0$, for all $x, x' \in C$

3. $\max_{1 \leq i \leq n} |w_{ni}(x)| \geq k_2 n^{-1}$ for all $x \in C$

4. $n^{1/(s-\eta)} \max_{1 \leq i \leq n} |w_{ni}(x)| \log n \leq k_3 \delta_n$, for some η with $0 < \eta < s$ and all $x \in C$

5. $\left(\sum_{i=1}^{n} w_{ni}^2(x) \log n\right)^{1/2} \leq k_4 \delta_n$.

REMARK. It is straightforward to verify these conditions hold for the kernel weights

$$w_{ni}(x) = \frac{1}{nh} K\left(\frac{x - X_i}{h} \right),$$

which are appropriate for uniform design. For example,

$$|w_{ni}(x) - w_{ni}(x')| = \frac{1}{nh} \left| K\left(\frac{x - X_i}{h} \right) - K\left(\frac{x' - X_i}{h} \right) \right| \leq \frac{c}{nh^{1+\gamma}} |x - x'|^\gamma,$$

provided the kernel satisfies

$$|K(x) - K(x')| \leq c |x - x'|^\gamma.$$

Therefore, provided $\limsup_n nh^{1+\gamma} > 0$, this condition is satisfied. The third condition is easy to justify when the weights obey Stone's type of conditions. For example, suppose that $w_{ni}(x)$ are probability weights, then

$$\sum_{i=1}^{n} w_{ni}^2(x) \leq \max_{1\leq i\leq n} w_{ni}(x) \sum_{i=1}^{n} w_{ni}(x) = \max_{1\leq i\leq n} w_{ni}(x),$$

which gives a lower bound on $\max_{1\leq i\leq n} w_{ni}(x)$, where for kernel weights, say, it is easy to show an exact almost sure uniform bound on $\sum_{i=1}^{n} w_{ni}^2(x)$, which is pointwise of order $n^{-1}h^{-1}$.

Theorem 20 *Suppose that assumptions 1-5 hold with probability one conditional on $X_1, \ldots, X_n$. Then, (9.11) holds.*

PROOF. Let $\{B(\chi_t, n^{-\mu}),\ t = 1, \ldots, T\}$ be a cover of C, where $B(\chi_t, n^{-\mu})$ is an open ball with center χ_t and radius $n^{-\mu}$. By compactness, T can be chosen to be $O(n^{\mu})$. Let $\mu = 3/\gamma$ and $r = s - \eta$, and define $\overline{\varepsilon}_i = \varepsilon_i 1(|\varepsilon_i| \leq n^{1/r})$ and $\overline{\overline{\varepsilon}}_i = \varepsilon_i 1(|\varepsilon_i| > n^{1/r})$. Then, $\varepsilon_i = \overline{\varepsilon}_i - E(\overline{\varepsilon}_i) + \overline{\overline{\varepsilon}}_i - E(\overline{\overline{\varepsilon}}_i)$, and by the triangle inequality

$$\begin{aligned}
\sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\varepsilon_i\right| &\leq \sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\left\{\overline{\overline{\varepsilon}}_i - E(\overline{\overline{\varepsilon}}_i)\right\}\right| + \sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\left\{\overline{\varepsilon}_i - E(\overline{\varepsilon}_i)\right\}\right| \\
&= \sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\left\{\overline{\overline{\varepsilon}}_i - E(\overline{\overline{\varepsilon}}_i)\right\}\right| + \max_{1\leq t\leq T}\sup_{x\in B(\chi_t, n^{-\mu})}\left|\sum_{i=1}^{n} w_{ni}(x)\left\{\overline{\varepsilon}_i - E(\overline{\varepsilon}_i)\right\}\right| \\
&\leq \sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\left\{\overline{\overline{\varepsilon}}_i - E(\overline{\overline{\varepsilon}}_i)\right\}\right| \qquad (9.12) \\
&\quad + \max_{1\leq t\leq T}\left|\sum_{i=1}^{n} w_{ni}(\chi_t)\left\{\overline{\varepsilon}_i - E(\overline{\varepsilon}_i)\right\}\right| \qquad (9.13) \\
&\quad + \max_{1\leq t\leq T}\sup_{x\in B(\chi_t, n^{-\mu})}\left|\sum_{i=1}^{n}\left\{w_{ni}(\chi_t) - w_{ni}(x)\right\}\overline{\varepsilon}_i\right| \qquad (9.14) \\
&\quad + \max_{1\leq t\leq T}\sup_{x\in B(\chi_t, n^{-\mu})}\left|\sum_{i=1}^{n}\left\{w_{ni}(\chi_t) - w_{ni}(x)\right\}E(\overline{\varepsilon}_i)\right|. \qquad (9.15)
\end{aligned}$$

We must establish that for some $k > 0$

$$\Pr\left[\sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\varepsilon_i\right| > k\delta_n\right] \longrightarrow 0,$$

which will be true if each of the random variables (9.12)-(9.15) behaves likewise.

PROOF FOR (9.12). By the Markov inequality and crude bounding, we have

$$\begin{aligned}
\Pr\left[\sup_{x\in C}\left|\sum_{i=1}^{n} w_{ni}(x)\left\{\overline{\overline{\varepsilon}}_i - E(\overline{\overline{\varepsilon}}_i)\right\}\right| > k\delta_n\right] &\leq \Pr\left[\sup_{x\in C,1\leq i\leq n}|w_{ni}(x)|\cdot\sum_{i=1}^{n}\left\{|\overline{\overline{\varepsilon}}_i| + |E(\overline{\overline{\varepsilon}}_i)|\right\} > k\delta_n\right] \\
&\leq \frac{2\sup_{x\in C,1\leq i\leq n}|w_{ni}(x)|\cdot nE\left(|\overline{\overline{\varepsilon}}_i|\right)}{k\delta_n} \\
&\leq \frac{2\sup_{x\in C,1\leq i\leq n}|w_{ni}(x)|\cdot nE\left(|\varepsilon_i|\right)\Pr\left[\mathbf{1}(|\varepsilon_i| > n^{1/r})\right]}{k\delta_n} \\
&\leq \frac{2kn^{-1/(s-\eta)}nE\left(|\varepsilon_i|\right)E\left(|\varepsilon_i|^s\right)n^{-s/r}}{\log n} \\
&= o(1).
\end{aligned}$$

PROOF FOR (9.14) AND (9.15). In this case we have a deterministic bound:

$$\begin{aligned}
\max_{1\leq t\leq T}\sup_{x\in B(\chi_t,n^{-\mu})}\left|\sum_{i=1}^{n}\left\{w_{ni}(\chi_t) - w_{ni}(x)\right\}\overline{\varepsilon}_i\right| &\leq kn^{1+1/r}\max_{1\leq t\leq T}\sup_{x\in B(\chi_t,n^{-\mu})}|\chi_t - x|^{\gamma} \\
&= O(n^{1+1/r}n^{-\gamma\mu}) = O(n^{1/r-2}) = o(1).
\end{aligned}$$

PROOF FOR (9.13). We will require the following exponential inequality.

Lemma 21 *Suppose that Y_{ni} are independent random variables with mean zero and $|Y_{ni}| \leq M < \infty$, and suppose that $E(Y_{ni}^2) = \sigma_{ni}^2 > 0$. Then, for all $s \in [0, k/M]$,*

$$E\left[\exp\left(s\sum_{i=1}^{n} Y_{ni}\right)\right] \leq \exp\left(k\cdot s^2\sum_{i=1}^{n}\sigma_{ni}^2\right).$$

By Bonferroni's, Markov's and Exponential inequalities, we have for any positive sequence $a_t(n)$ and independent random variables $Y_{ni}(t)$

$$\begin{aligned}\Pr\left[\max_{1\le t\le T}\left|\sum_{i=1}^{n}\frac{Y_{ni}(t)}{a_t}\right|>1\right] &\le \sum_{t=1}^{T}\Pr\left[\left|\sum_{i=1}^{n}Y_{ni}(t)\right|>a_t\right]\\ &\le \sum_{t=1}^{T}e^{-a_ts_t}E\left[\exp\left(s_t\left|\sum_{i=1}^{n}Y_{ni}(t)\right|\right)\right]\\ &\le 2\sum_{t=1}^{T}e^{-a_ts_t}\exp\left(ks_t^2\sum_{i=1}^{n}\sigma_{ni}^2(t)\right),\end{aligned}$$

for any sequence $s_t(n)$ with $s_t \le k/\max|Y_{ni}(t)|$, where $\sigma_{ni}^2(t) = \mathrm{var}\left[Y_{ni}(t)\right]$. The trick is to choose a_t, s_t, and $Y_{ni}(t)$ so that the left hand side is $\Pr\left[\max_{1\le t\le T}\left|\sum_{i=1}^{n}w_{ni}(\chi_t)\left\{\overline{\varepsilon}_i - E(\overline{\varepsilon}_i)\right\}\right| > \zeta\delta_n\right]$ and the right hand side is $o(1)$.

We shall take

$$a_t(n) = \zeta\delta_n\beta_n(t) \quad \text{and} \quad Y_{ni}(t) = \beta_n(t)w_{ni}(\chi_t)\left\{\overline{\varepsilon}_i - E(\overline{\varepsilon}_i)\right\},$$

which satisfies the first requirement for any $\beta_n(t)$. We take

$$\beta_n(t) = \delta_n^{-2}n^{2/r}\max_{1\le i\le n}|w_{ni}(\chi_t)|\cdot(\log n)^2 \quad \text{and} \quad s_t = \frac{1}{\sqrt{\beta_n(t)n^{2/r}\max_{1\le i\le n}|w_{ni}(\chi_t)|}},$$

so that we can apply the exponential inequality, since

$$|Y_{ni}(t)| \le 2\delta_n^{-2}n^{3/r}\left(\max_{1\le i\le n}|w_{ni}(\chi_t)|\right)^2\cdot(\log n)^2 := M_t < \infty,$$

and

$$s_tM_t = \frac{2n^{1/r}\max_{1\le i\le n}|w_{ni}(\chi_t)|\cdot(\log n)}{\delta_n} \le k < \infty,$$

as required. Furthermore, note that $\mathrm{var}\left[\sum_{i=1}^{n}Y_{ni}(t)\right] = \beta_n^2(t)\sum_{i=1}^{n}w_{ni}^2(\chi_t)\mathrm{va}m(\overline{\varepsilon}_i)$. We therefore

obtain

$$\begin{aligned}\Pr\left[\max_{1\le t\le T}\left|\sum_{i=1}^{n} w_{ni}(\chi_t)\left\{\overline{\varepsilon}_i - E(\overline{\varepsilon}_i)\right\}\right| > \zeta\delta_n\right] &\le 2\sum_{t=1}^{T}\exp\left[k\cdot\frac{\beta_n(t)\sum_{i=1}^{n} w_{ni}^2(\chi_t)}{n^{2/r}\max_{1\le i\le n}|w_{ni}(\chi_t)|}\right.\\ &\qquad\left.-\frac{k\zeta\delta_n\beta_n(t)}{\sqrt{\beta_n(t)n^{2/r}\max_{1\le i\le n}|w_{ni}(\chi_t)|}}\right]\\ &= 2\sum_{t=1}^{T}\exp\left[k\frac{(\log n)^2\sum_{i=1}^{n} w_{ni}^2(\chi_t)}{\delta_n^2} - k\zeta\cdot(\log n)\right]\\ &= O(n^{3/\mu}n^{k_1-k_2\zeta}) = o(1)\end{aligned}$$

for large enough ζ. Here, k_1 and k_2 are constants.

■

9.5 Functional Central Limit Theorem

PROOF. The basic argument is to apply a functional central limit theorem to the random variable

$$T_n^*(x) = \frac{1}{\sqrt{\sigma^2(x)f(x)nh}}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\varepsilon_i$$

and then use the invariance principle to find the distribution of the maximum. For the FCLT to hold we need convergence of the finite dimensional distributions to normals and a tightness criterion to be satisfied [Billingsley]. Finite dimensional convergence has been established already. Unfortunately, the tightness criterion can not be satisfied here! This is that for all positive ϵ, η, there exists a $\delta > 0$ and an n_0 such that for all $n \ge n_0$

$$\frac{1}{\delta}\Pr\left[\sup_{|x-x'|\le\delta}|T_n^*(x) - T_n^*(x')| \ge \epsilon\right] \le \eta. \tag{9.16}$$

The problem is that for any distinct points x and x' with $|x - x'| > 0$, the random variables $T_n^*(x)$ and $T_n^*(x')$ are asymptotically standard normal but mutually independent. This means that knowing $T_n^*(x)$ says nothing about $T_n^*(x')$ and so the bound (9.16) cannot be satisfied for large enough n.

[However, one can establish a functional limit theorem for a reparameterized process $x \to x_0 + \theta h$ for $\theta \in [-1, 1]$ for each x_0].

The argument is as follows. First, one approximates the process $T_n^*(x)$ by

$$T_n^{**}(x) = h^{-1/2} \int K\left(\frac{x-z}{h}\right) dW(z),$$

where $W(\cdot)$ is a standard Wiener process. This follows from the weak convergence of the empirical process $F_n(y, x)$. Second, the process $T_n^{**}(h \cdot x)$ has the same distribution as

$$T_\infty(x) = \int K(x-z) dW(z),$$

since it is Gaussian and has the same mean [zero] and covariance function.

$$\begin{aligned} E\left[T_n^{**}(hx) \cdot T_n^{**}(hx')\right] &= h^{-1} E\left[\int K\left(\frac{hx-z}{h}\right) dW(z) \int K\left(\frac{hx'-z'}{h}\right) dW(z')\right] \\ &= h^{-1} \int K\left(x - \frac{z}{h}\right) K\left(x' - \frac{z}{h}\right) E\left(dW^2(z)\right) \\ &= \int K(x-z) K(x'-z) dz \\ &= E\left[\int K(x-z) dW(z) \int K(x'-z') dW(z')\right], \end{aligned}$$

by a change of variables. Let $C = \{x : x = x_0 + \theta h \text{ for } \theta \in [-1, 1]\}$. Then

$$\sup_{x \in C} |T_n^{**}(x)| = \sup_{x \in h^{-1}C} T_n^{**}(h \cdot x)$$

and finding the maximum of the process $T_\infty(x)$ over the set $h^{-1} \cdot C$ is a standard problem in extreme value theory. The solution depends on the correlation function of $T_\infty(x)$, which is

$$\begin{aligned} \operatorname{corr}(T_\infty(x), T_\infty(x')) &= \int K(x+z) K(x'+z) dz \Big/ \int K^2(z) dz \\ &\simeq 1 - (x-x')^2 \frac{\|K'\|^2}{2\|K\|^2} \end{aligned}$$

for x close to x'. In the i.i.d. case, the extreme value theory is quite simple. Suppose that we have

$$M_n = a_n \left(\max_{1 \le i \le n} X_i - b_n\right),$$

then

$$\Pr\left[M_n \leq t\right] = F^n(a_n^{-1}t + b_n) = \left(1 - \frac{n(1 - F(a_n^{-1}t + b_n))}{n}\right)^n.$$

If for some function $g(\cdot)$ we have

$$n(1 - F(a_n^{-1}t + b_n)) \to g(t), \tag{9.17}$$

then

$$\Pr\left[M_n \leq t\right] \to e^{-g(t)}.$$

For example, when X_i is standard normally distributed it is a standard result that

$$a_n = \sqrt{2\log n} \quad ; \quad b_n = \sqrt{2\log n} - \frac{1}{2}\frac{\log\log n + \log 4\pi}{\sqrt{2\log n}}$$

and $g(t) = \exp(-t)$ suffice in (9.17). See Leadbetter and Rootzén (1988). ■

9.6 An Interpretation of the asymptotics for Marginal Integration

Consider the special case

$$E(Y|X = x, Z = z) = m(x, z) = c + g(x) + h(z).$$

Suppose that $\widehat{m}(x, z)$ is a standard nonparametric local linear estimator of $m(x, z)$. Without loss of generality suppose that Z has support $[0, 1]$, and let $\{z_i\}_{i=1}^J$ be an equally space grid on $[0, 1]$, so that $|z_i - z_{i-1}| = 2h$. It follows that $J = 1/2h$ and $\widehat{m}(x, z_i), \widehat{m}(x, z_k)$ are asymptotically independent whenever $i \neq k$.

Then let

$$\widehat{d}(x) = \frac{1}{J}\sum_{i=1}^{J} \widehat{m}(x, z_i)w(z_i),$$

where $w(.)$ is a bounded smooth weighting function. Then

$$\widehat{d}(x) \xrightarrow{P} d(x) = \int m(x, z)w(z)dz = (c + g(x))\int w(z)dz + \int h(z)w(z)dz,$$

so that provided $\int w(z)dz = 1$,

$$d(x) = \int m(x,z)w(z)dz = g(x) + c'$$

with $c' = c + \int h(z)w(z)dz$ not depending on x.

We have

$$\widehat{d}(x) - d(x) = \frac{1}{J}\sum_{i=1}^{J}\left[\widehat{m}(x,z_i) - m(x,z_i)\right]w(z_i) + \frac{1}{J}\sum_{i=1}^{J} m(x,z_i)w(z_i) - \int m(x,z)w(z)dz$$

which decomposes into bias term, variance term, and integration error term. The integration error term is

$$\frac{1}{J}\sum_{i=1}^{J} m(x,z_i)w(z_i) - \int m(x,z)w(z)dz = O(J^{-1}) = O(h).$$

In practice, when $\sum_{i=1}^{J} w(z_i)/J = 1$ this bias is $\sum_{i=1}^{J} h(z_i)w(z_i)/J - \int h(z)w(z)dz$ which varies with n but not x, it can probably be eliminated by just averaging over x.

The bias term is $\sum_{i=1}^{J}\beta_n(x,z_i)w(z_i)/J$, where β_n is the bias function of $\widehat{m}(x,z_i)$. For a standard local linear estimator with product kernels K this is $(h^2/2)\int u^2K(u)du\nabla_2 m(x,z_i)$, where ∇_2 is the trace of the Hessian operator. The bias is just like the bias of the usual integration estimator.

Regarding the variance, we have by virtue of the independence

$$\begin{aligned}
\text{var}\left[\frac{1}{J}\sum_{i=1}^{J}\widehat{m}(x,z_i)w(z_i)\right] &= \frac{1}{J^2}\sum_{i=1}^{J}\text{var}\left[\widehat{m}(x,z_i)w(z_i)\right] \\
&= \frac{1}{nh^2}||K||_2^2\frac{1}{J^2}\sum_{i=1}^{J}\frac{\sigma^2(x,z_i)}{f(x,z_i)}w^2(z_i) \\
&\simeq \frac{1}{nh}\frac{1}{Jh}||K||_2^2\frac{1}{J}\sum_{i=1}^{J}\frac{\sigma^2(x,z_i)}{f(x,z_i)}w^2(z_i) \\
&\simeq \frac{2}{nh}||K||_2^2\int\frac{\sigma^2(x,z)}{f(x,z)}w^2(z)dz,
\end{aligned}$$

which is just like the marginal integration variance when $w(z) = f_Z(z)$.

Bibliography

[1] Akaike, H. (1970): "Statistical predictor information," Annals of the Institute of Statistical Mathematics 22, 203–17.

[2] Akaike, H. (1974): "A new look at the statistical model identification." IEEE Transactions of Automatic Control AC 19, 716–23.

[3] Andrews, D.W.K., (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models." Econometrica 59, 307-346.

[4] Andrews (1995). Nonparametric Kernel Estimation for Semiparametric Models, Econometric Theory

[5] Andrews, D.W.K., and Y-J. Whang (1990): "Additive and Interactive Regression Models: Circumvention of the Curse of Dimensionality," Econometric Theory 6, 466-479.

[6] Ansley, C.F., R. Kohn, and C. Wong (1993): "Nonparametric spline regression with prior information," Biometrika 80, 75-88.

[7] Bickel, P.J., and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. Annals of Statistics 1, 1071-1095.

[8] Bierens, H.J., (1987): "Kernel Estimators of Regression Functions." in Advances in Econometrics: Fifth World Congress, Vol 1, ed. by T.F. Bewley. Cambridge University Press.

[9] Blackorby, C., D. Primont and R. R. Russell, (1978), Duality, Separability, and Functional Structure: Theory and Economic Applications. New York: North Holland.

[10] Brillinger, D.R., (1980) Time Series, Data analysis and Theory. Holden-Day.

[11] Chamber, J.M., Clevelend, W.S., Kleiner, B., and P.A. Tukey (1983). Graphical Methods for Data Analysis. Duxburry Press.

[12] Chan, K.C., G. Karolyi, F. Longstaff and A. Sanders (1992). An Empirical Comparison of Alternative Models of Short-Term Interest Rate. Journal of Finance 47, 1209-1227.

[13] Chanda, K.C. (1974): "Strong mixing properties of linear stochastic process." Journal of Applied Probabilities 11, 401–408.

[14] Chen, S.X. (1999). Beta kernel estimators for density functions. Computational Statistics and Data Analysis 31, 131-145.

[15] Chen, X. and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," Econometrica, 66, 289-314.

[16] Cleveland, W.S., (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots." Journal of the American Statistical Association 74, 829-836.

[17] Cohen, A. (1966). All admissible linear estimates of the mean vector. Ann. Math. Statist. 37, 458-463.

[18] Cox, D.R., and D.V. Hinkley (1974): Theoretical Statistics. Chapman and Hall.

[19] Cox, J., J. Ingersoll and S. Ross (1985). A Theory of the Term Structure of Interest Rates. Econometrica, 53, 385-406.

[20] Craven, P. and Wahba, G. (1979): "Smoothing noisy data with spline functions," Numer. Math. 31, 377–403.

[21] Daniell, P.J., (1946): "Discussion of paper by M.S. Bartlett," Journal of the Royal Statistical Society Supplement 8:27.

[22] Devroye, L. (1981). On the almost everywhere convergence of nonparametric function estimates. Annals of Statistics 9, 1310-1319.

[23] Doukhan, P. and Ghindes, M. (1980): "Estimation dans le processus $X_n = f(X_{n-1}) + \epsilon_n$," Comptes Rendus, Académie des Sciences de Paris 297, Série A, 61–4.

[24] Einmahl, U., and D.M. Mason (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. Journal of Thoeretical Probability 13, 1-37.

[25] Elbadawi, I., A.R. Gallant, and G. Souza, (1983): "An elasticity can be estimated consistently without a priori knowledge of functional form," Econometrica 51, 1731-1751.

[26] Eubank, R.L., (1988): Smoothing Splines and Nonparametric Regression. Marcel Dekker.

[27] Family Expenditure Survey, Annual Base Tapes (1968-1983). Department of Employment, Statistics Division, Her Majesty's Stationary Office, London 1968-1983.

[28] Fan, J. (1992): "Design-Adaptive Nonparametric Regression," Journal of the American Statistical Association 87, 998-1004.

[29] Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiencies," The Annals of Statistics, 21, 196-216.

[30] Fan, J., and I. Gijbels (1996): Local Polynomial Modelling and Its Applications Chapman and Hall.

[31] Fan, J., N.E. Heckman, and M.P. Wand, (1992): "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," University of British Columbia Working paper 92-028.

[32] Fix, E. and J.L. Hodges (1951): "Discriminatory analysis, nonparametric estimation: consistency properties," Report no 4, Project no 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

[33] Gallant, A.R., and G. Souza, (1991): "On the asymptotic normality of Fourier flexible form estimates," Journal of Econometrics 50, 329-353.

[34] Garodetskii, V.V. (1977): "On the strong mixing condition for linear process," Theory of Probability and its Applications 22, 411–413.

[35] Gasser, T. and H. G. Müller (1984): "Estimating regression functions and their derivatives by the kernel method," Scandinavian Journal of Statistics 11, 171–85.

[36] Gasser, T., Müller, H. G., and V. Mammitzsch (1985): "Kernels for nonparametric curve estimation,"Journal of the Royal Statistical Society Series B 47, 238–52.

[37] Giné, W., and A. Guillou (2002). Rates of Strong uniform consistency for multivariate kernel density estimators. Ann. I. H. Poincaré 6, 907-921.

[38] Goldman, S. M. and H. Uzawa, (1964), "A Note On Separability and Demand Analysis," Econometrica, 32, 387-398.

[39] Gyorfi, L., Härdle, W., Sarda, P., and P. Vieu (1990): Nonparametric Curve Estimation from Time Series. Lecture Notes in Statistics, 60. Heidelberg, New York: Springer-Verlag.

[40] Hall, P., (1992): The Bootstrap and Edgeworth Expansion. Springer-Verlag, New-York.

[41] Hall, P., (1993): "On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation," Journal of the Royal Statistical Society Series B 55, 291-304.

[42] Hall, P. and I. Johnstone (1992): "Empirical functional and efficient smoothing parameter selection," (with discussion). Journal of the Royal Statistical Society Series B. 54, 475-530.

[43] Härdle, W. (1990). Applied Nonparametric Regression. New York: Cambridge University Press.

[44] Härdle, W. (1991). Smoothing Techniques with Implementation in S. Heidelberg, New York, Berlin: Springer-Verlag.

[45] Härdle, W. and Carroll, R. J. (1989): "Biased cross-validation for a kernel regression estimator and its derivatives," Östereichische Zeitschrift für Statistik und Informatik.

[46] Härdle, W., Hall, P. and Marron, J. S. (1988): "How far are automatically chosen regression smoothing parameters from their optimum?" (with discussion). Journal of the American Statistical Association 83, 86–99.

[47] Härdle, W., Hall, P. and Marron, J. S. (1992): "Regression smoothing parameters that are not far from their optimum" Journal of the American Statistical Association 87, 227–233.

[48] Härdle, W. and M. Jerison, (1991): "Cross Section Engel Curves over Time," CORE discussion paper 991, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

[49] Härdle, W., and O.B. Linton (1994): "Applied nonparametric methods," The Handbook of Econometrics, vol. IV, eds. D.F. McFadden and R.F. Engle III. North Holland.

[50] Härdle, W., and S. Luckhaus (1984). Uniform consistency of a class of regression function estimators. Annals of Statistics 12, 612-623.

[51] Härdle, W., and Marron, J. S. (1985): "Optimal bandwidth selection in nonparametric regression function estimation," Annals of Statistics 13, 1465-81.

[52] Härdle, W., and M. Müller (1993): "Nichtparametrische Glättungsmethoden in der alltäglichen statistichen Praxis," Allg. Statistiches Archiv 77, 9-31.

[53] Härdle, W. and D.W. Scott (1990): "Smoothing by weighted averaging of rounded points," CORE Discussion Paper 9040, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

[54] Härdle, W., and T. M. Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," Journal of the American Statistical Association 84, 986-995.

[55] Härdle, W. and P. Vieu (1991): "Kernel regression smoothing of time series," Journal of Time Series Analysis 13, 209-232.

[56] Hall, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. Journal of the Royal Statistical Society, Series B. 55, 291-304.

[57] Hart, J. and P. Vieu (1990): "Data-driven bandwidth choice for density estimation based on dependent data," Annals of Statistics 18, 873–890.

[58] Hart, D. and T. E. Wehrly (1986): "Kernel regression estimation using repeated measurements data," Journal of the American Statistical Association 81, 1080–8.

[59] Hastie, T.J., and R.J. Tibshirani (1990): Generalized Additive Models Chapman and Hall.

[60] Heckman, J., H. Ichimura, J. Smith and P. Todd (1998) "Characterization of Selection Bias Using Experimental Data" Econometrica, 66, 1017–1098.

[61] Heckman, J., H. Ichimura, and P. Todd (1998) "Matching as an Econometric Estimator" Review of Economic Studies, 65, 261–294.

[62] Hirano, K., G. Imbens, G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," NBER Technical Working Paper 251.

[63] Horowitz, J. L., (2001), "Nonparametric estimation of a generalized additive model with an unknown link function," Econometrica 69, 499-513.

[64] Johnston, G.J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. Journal of Multivariate Analysis 12, 402-414.

[65] Jones, M.C., (1985): "Discussion of the paper by B.W. Silverman," Journal of the Royal Statistical Society Series B 47, 25-26.

[66] Jones, M.C., (1989): "Discretized and interpolated Kernel Density Estimates, " Journal of the American Statistical Association 84, 733-741.

[67] Jones, M.C., Davies, S.J., and Park, B.U. (1994). "Versions of kernel-type regression estimators". Journal of the American Statistical Association, 89, 825-832.

[68] Jones, M.C., and P.J. Foster (1993): "Generalized jacknifing and higher order kernels," Forthcoming in Journal of Nonparametric Statistics.

[69] Jones, M.C., J.S. Marron, and S.J. Sheather (1992): "Progress in data-based bandwidth selection for kernel density estimation," University of New South Wales working paper no 92-04.

[70] Krieger, A.M., and J. Picklands (1981). Weak convergence and efficient density esitmation at a point. The Annals of Statistics 9, 1066-1078.

[71] Leadbetter, M.R., and H. Rootzén (1988). Extreme theory for stochastic processes. Annals of Probability 16, 431-478.

[72] Leontieff, W. (1947). Introduction to a theory of an internal structure of functional relationships. Econometrica, 15, 361-373.

[73] Li, K-C. (1985): "From Stein's unbiased risk estimates to the method of generalized cross-validation." Annals of Statistics 13, 1352–77.

[74] Linton, O.B. and J.P. Nielsen (1995): "A kernel method of estimating structured nonparametric regression using marginal integration," Biometrika

[75] Mack, Y. P. (1981): "Local properties of k-NN regression estimates," SIAM J. Alg. Disc. Meth. 2, 311–23.

[76] Mack, Y. P. and Müller (1989): "Derivative estimation in nonparametric regression with random predictor variable," Sankhya, Ser. A., 51, 59-72.

[77] Manski, C. (1994). The selection problem. In Advances in Econometrics Sixth World Congress Volume 1. Ed. C. Sims. Cambridge University Press.

[78] Marron, J.S. and D. Nolan (1989): "Canonical kernels for density estimation," Statistics and Probability Letters 7, 191-195.

[79] Marron, J.S. and M.P.Wand (1992): "Exact Mean Integrated Squared Error." Annals of Statistics 20, 712-736.

[80] Masry, E. (1996a): "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," Journal of Time Series Analysis 17, 571-599.

[81] Masry, E. (1996b): "Multivariate regression estimation Local polynomial fitting for time series," Stochastic Processes and their Applications 65, 81-101.

[82] Matzkin, R. L. (1994), "Restrictions of Economic Theory in Nonparametric Methods," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, 2523-2558, Amsterdam: Elsevier.

[83] Müller, H. G. (1987): "On the asymptotic mean square error of L_1 kernel estimates of C_∞ functions," Journal of Approximation Theory 51, 193-201.

[84] Müller, H. G. (1988): Nonparametric Regression Analysis of Longitudinal Data. Lecture Notes in Statistics, Vol. 46. Heidelberg/New York: Springer-Verlag.

[85] Nadaraya, E.A., (1964): "On estimating regression," Theory of Probability and its Applications 10, 186-190.

[86] Pagan, A.R., and Y.S. Hong (1991): "Nonparametric Estimaiton and the Risk Premium," in W. Barnett, J. Powell, and G.E. Tauchen (eds.) Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge University Press.

[87] Park, B.U., and B.A. Turlach (1992): "Practical performance of several data-driven bandwidth selectors (with discussion)," Computational Statistics 7, 251-271.

[88] Pinkse, J., (2001), "Nonparametric Regression Estimation Using Weak Separability," unpublished manuscript.

[89] Rice, J. A. (1984): "Bandwidth choice for nonparametric regression" Annals of Statistics 12, 1215–30.

[90] Robinson, P.M. (1983): "Nonparametric Estimators for Time Series." Journal of Time Series Analysis 185-208.

[91] Robinson, P. M. (1991): "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models." Econometrica 59, 1329-1364.

[92] Rosenbaum, P. and Rubin, D.B. (1983) "The central role of the propensity score in observational studies for causal effects." Biometrika, 70, pp. 41-55.

[93] Rosenblatt, M., (1956): "Remarks on some nonparametric estimates of a density function," Annals of Mathematical Statistics 27, 642-669.

[94] Ruppert, D., and M.P.Wand (1992): "Multivariate Locally Weighted Least Squares Regression," Rice University, Technical Report no 92-4.

[95] Savin, N.E. (1984). Multiple Hypothesis Testing. In The Handbook of Econometrics, vol. IV, eds. D.F. McFadden and R.F. Engle III. North Holland. Feller, Vols. II and XVI

[96] Schuster, E.F., (1972): "Joint asymptotic distribution of the estimated regression function at a finite number of distinct points," Annals of Mathematical Statistics 43, 84-8.

[97] Shibata, R.(1981): "An optimal selection of regression variables," Biometrika, 68, 45–54.

[98] Silverman, B. W. (1978): "Weak and Strong uniform consistency of the kernel estimate of a dnesity and its derivatives," Annals of Statistics 6, 177-184.

[99] Silverman, B. W. (1984): "Spline smoothing: the equivalent variable kernel method." Annals of Statistics 12, 898–916.

[100] Silverman, B. W. (1985): "Some aspects of the Spline Smoothing approach to Non-parametric Regression Curve Fitting," Journal of the Royal Statistical Society Series B 47, 1-52

[101] Silverman, B. W. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.

[102] Silverman, B. (1978). Weak and strong consistency of a kernel estimate of a density and its derivatives. Annals of Statistics 6, 177-184.

[103] Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. Annals of Statistics

[104] Stone, C.J., (1982): "Optimal global rates of convergence for nonparametric regression," Annals of Statistics 10, 1040-1053.

[105] Stute, W. (1984): "Asymptotic normality of nearest neighbor regression function estimates. Annals of Statistics 12, 917-926.

[106] Stute, W. (1986): "Conditional Empirical Processes," Annals of Statistics 14, 638-647.

[107] Tibshirani, R., (1984): "Local Likelihood estimation," PhD Thesis, Stanford University.

[108] Tikhonov, A.N. (1963): "Regularization of incorrectly posed problems," Soviet Math., 4, 1624–1627.

[109] Tripathi, G. and W. Kim, (2001), "Nonparametric Estimation of Homogeneous Functions," unpublished manuscript.

[110] Vasicek, O. (1977). An Equilibrium Characterization of the Term Structure. Journal of Financial Economics, 5(2), 177-88.

[111] Wahba, G. (1990): Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, no 59.

[112] Watson, G.S. (1964): "Smooth regression analysis," Sankhya Series A 26, 359-372.

[113] Whittaker, E.T., (1923): "On a new method of graduation," Proc. Edinburgh Math.Soc 41, 63-75.