Lecture Notes on Non- and Semiparametric Econometrics[*]

Jörg Stoye

May 1, 2025

1 Introduction

First things first: In modern statistics and econometrics usage, the term **nonparametric** does not indicate the absence of parameters or even parameterizations. It rather means that the data generating process (d.g.p.) is specified up to an **infinite dimensional** unknown quantity. Informally, this quantity may even be referred to as parameter. Conversely, a parametric model is specified up to a **finite dimensional** parameter, at least where it matters.[1]

The difference cannot be overstated. In a parametric world, as $n \to \infty$, n necessarily becomes large relative to size of the model. Also, typically every observation is at least slightly relevant for every parameter. For broad classes of well-behaved parameters, this allows for $\sqrt{n}$-consistent estimation. In contrast, since it is obviously impossible to literally estimate an infinite dimensional model from finitely many observations, nonparametric methods typically rely on a sample size dependent coarsening of the space of d.g.p.'s under consideration. In other words, the model being fitted becomes more complex with sample size, and this typically precludes $\sqrt{n}$-consistency. This is especially easy to intuit with kernel density estimation, where the density at a given point is estimated off an effective sample that diverges but at a rate slower than n. Consistency here is at a rate corresponding to the square root of effective sample size. The rate at which models are coarsened is an important tuning parameter, and figuring out how to pick it is an important part of nonparametric estimation methodology.

Why bother? The rise of nonparametric (and semiparametric, and seminonparametric...) methods is partly due to concerns with credible identification. Parametric models typically combine substantive economic assumptions, convenience assumptions, and maybe also regularity conditions. The convenience assumptions are often distributional (e.g., logit vs probit) or assume linearity of a relationship that we truly "only" believe to be smooth monotone etc. Nonparametrics can in principle avoid many of those, leading to more credible conclusions, though at a cost that will become obvious. Other factors that gave rise to nonparametrics are

[*] Do not circulate further. Borrows from lecture notes by Bruce Hansen and Alois Kneip.

[1] This coexists with older usage, where *nonparametric* would, for example, describe analysis of rank correlation. It's normally obvious which meaning is intended.

the availability of larger data sets and computing resources and the (endogenous, of course) development of appropriate theory.

In these notes, we develop the baseline theory of kernel estimation in some detail. This is partly to illustrate how nonparametric asymptotics work. The treatment later becomes more breezy, but you will notice that many ideas reappear.

2 Kernel Density Estimation

2.1 Overview

Say we want to estimate the distribution of a random variable X_i. An obvious estimator for the c.d.f. F, namely the empirical distribution, was already discussed and used in the bootstrap. The analogous estimator of the density (assuming of course that one exists) is the empirical probability mass function (p.m.f.), which consists of at most n mass points. This estimator is technically not even a density. In contrast to the empirical c.d.f., it is also often useless: The empirical c.d.f. consistently estimates the population c.d.f., but the empirical p.m.f. consistently estimates the density almost nowhere.

One fix to this is kernel density estimation, that is, the density is estimated by a weighted average of nearby mass points of the empirical p.m.f. This can be intuited as smoothing out histograms or as estimating F' by some smoothed arc slope (discrete derivative) of F_n. The basic idea is obvious enough, but there are many ways to do the smoothing. In particular, one will have to choose a **kernel** and a **bandwidth**. In practice, choice of kernel is often not too consequential, but the bandwidth is the "coarsening parameter" mentioned above and its choice tends to matter a lot.

In the scalar case, a kernel density estimator of $f(x)$ can be written as

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

where $k : \boldsymbol{R} \mapsto \boldsymbol{R}$ is the kernel function and h is the bandwidth. For a simple example, let $k(t) = 1/2 \cdot \mathbf{1}\{|t| \leq 1\}$, i.e. the **uniform kernel**, then one can write

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}\{x - h \leq X_i \leq x + h\} = \frac{1}{2h} P_n([x - h, x + h]),$$

where P_n is the empirical p.m.f.

This particular estimator is a step function. This can be avoided by using continuous kernels. But the uniform kernel makes some of the trade-offs in kernel estimation very obvious. In particular, we estimate the density at x by averaging the empirical p.m.f. over $[x - h, x + h]$. A larger choice of bandwidth h means that we average over more observations, and so the variance of our estimator will decrease. But it also means that we average over a set where the true density may be increasingly different from the true density at x, so the bias will increase. This is quite like with histograms: If we "oversmooth" (too large bin size), the thing becomes eventually constant, if we "undersmooth" (too small bin size), it becomes too wiggly and will eventually include unwanted zeros.

In practice, we attempt to choose h to intelligently resolve this bias/variance trade-off. For example, in an "oracle" situation where we know the density to be estimated, we can find the choice of h that minimizes mean square error

$$\mathbb{E}\big(\hat{f}(x) - f(x)\big)^2$$

at some x of interest. It is intuitive (we'll do algebra later) that the h optimizing this will (i) decrease with n (therefore, you will often see the notation h_n), (ii) depend on how smooth the true density is at x, with a smoother true density calling for more smoothing. If we are interested in overall performance, we could also minimize the mean integrated square error

$$\int_{-\infty}^{\infty} \mathbb{E}\big(\hat{f}(x) - f(x)\big)^2 dx,$$

and the solution will again decrease with n but also with some measure of overall smoothness of f.

The obvious catch is that we don't know f, and therefore the "oracle" bandwidth is not known in practice. This problem gave rise to a considerable literature and we will think about it much more. Also, the optimal choice of h as function of n will imply that bias and variance are of the same order in n. This is simply because we want to minimize the larger of two rates, and the maximum of two functions is typically minimized at a point of equality. So the bandwidth choice that makes for the "best" estimator also makes for an asymptotic distribution that is noncentered, much complicating inference. Therefore, it is common to choose a "too small" bandwidth so that variance dominates bias and the asymptotic distribution is centered. This is called [deliberate] **undersmoothing**.

We will next formalize these intuitions. The most important addition will be that I so far picked a very particular kernel. While bandwidth choice tends to be more consequential than kernel choice, we will formally think about both, and we will certainly generalize beyond uniform kernels (which are not common in practice).

2.2 Formal Properties of Kernels

A kernel function $k : \boldsymbol{R} \mapsto \boldsymbol{R}$ must integrate to 1: $\int_{-\infty}^{\infty} k(u)du = 1$. No other property is strictly required, but in practice kernel functions are usually symmetric: $k(u) = k(-u)$.[2] The **order** of a kernel is the index of its first nonzero moment. Thus, defining $\kappa_j(k) =$

[2] There are exceptions to this if estimation of a density near the boundaries of the random variable's support is the goal. We will ignore that.

$\int_{-\infty}^{\infty} u^j k(u)du$, the order of k equals $\min\{j > 0 : \kappa_j(k) \neq 0\}$. As we restrict attention to symmetric kernels and all odd moments of symmetric kernels are zero, any kernel that we look at has an even order that is at least 2. The most popular kernels are furthermore nonnegative, i.e. $k(u) \geq 0$ for all u. A nonnegative kernel necessarily has $\kappa_2 > 0$ and is therefore of order 2. A kernel has **higher order** if its order strictly exceeds 2. We will occasionally mention properties of higher-order kernels and you should know what the term means, but our focus will be on nonnegative kernels, which are also dominant in empirical practice.

Any nonnegative kernel can be interpreted as a probability density, and so any kernel density estimate using a nonnegative kernel is a weighted average in the everyday sense (i.e., excluding negative weights) of the empirical p.m.f. Two simple kernels are the uniform and Gaussian ones:

$$\begin{aligned} k^{uni}(u) &= \frac{1}{2}\mathbf{1}\{|u| \leq 1\} \\ k^{gau}(u) &= (2\pi)^{-1/2}\exp(-u^2/2). \end{aligned}$$

With these or any other nonnegative kernels, the kernel density estimator can be visually intuited as taking each of the n realizations of X_i and replacing it with $1/n$ times a (uniform, standard normal,...) p.d.f. centered at that realization. This makes it immediately obvious that $\hat{f}(x)$ is itself a proper p.d.f., but also that the r.v. described by $\hat{f}$ is a mean-preserving spread of the r.v. described by the empirical distribution. In contrast, with higher-order kernels, $\hat{f}(x)$ can take negative values and therefore need not be a density; but as we will see, it replicates the second (and possibly higher, depending on order of the kernel) moment of the empirical distribution.

While intuitively obvious for nonnegative kernels, we next show that $\hat{f}$ necessarily integrates to 1. To see this formally, first write

$$1 = \int_{-\infty}^{\infty} k(u)du = \int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{X_i - x}{h}\right) dx,$$

where the last step uses the change-of-variables $u(x) = \frac{X_i - x}{h}$. Next,

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{X_i - x}{h}\right) dx = 1.$$

An obvious consequence is that if k is nonnegative, then $\hat{f}(x)$ is a valid density. [3]

[3] Because this is a desirable property – you probably don't want to report negative probabilities to clients – with higher order kernels we may force it by reporting

$$\widetilde{f}(x) = \frac{|\hat{f}(x)|_+}{\int_{-\infty}^{\infty} |\hat{f}(x)|_+ dx}$$

The "estimator" of $\mathbb{E}(X)$ implied by our density estimate is what you'd expect:

$$\begin{aligned}
\int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x\frac{1}{h}k\left(\frac{X_i-x}{h}\right)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(X_i-uh)k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}X_ik(u)du-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}uhk(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i\underbrace{\int_{-\infty}^{\infty}k(u)du}_{=1}-\frac{1}{n}\sum_{i=1}^{n}h\underbrace{\int_{-\infty}^{\infty}uk(u)du}_{=0}=\overline{X},
\end{aligned}$$

where we used $u=\frac{X_i-x}{h}\Leftrightarrow x=X_i-uh$. In words, the expectation induced by the estimated density is just the sample average.

The same is not true for the estimated density's second uncentered moment and, therefore, its variance:

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2\frac{1}{h}k\left(\frac{X_i-x}{h}\right)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(X_i-uh)^2k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}X_i^2k(u)du+\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(uh)^2k(u)du-\frac{2}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}X_ihuk(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2\underbrace{\int_{-\infty}^{\infty}k(u)du}_{=1}+\frac{1}{n}\sum_{i=1}^{n}h^2\underbrace{\int_{-\infty}^{\infty}u^2k(u)du}_{=\kappa_2}-\frac{2}{n}\sum_{i=1}^{n}X_ih\underbrace{\int_{-\infty}^{\infty}uk(u)du}_{=0} \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2+h^2\kappa_2.
\end{aligned}$$

Thus, with nonnegative kernels, the density estimator implies a variance that exceeds the sample variance by $h^2\kappa_2$ (and the df-adjusted estimator by slightly more). In particular, the variance corresponding to the estimated density coincides with the sample variance for higher-order but not for nonnegative kernels. For nonnegative kernels, this is also clear from the representation mentioned earlier, i.e. $\hat{f}$ describes a mean-preserving spread of the empirical p.m.f. You can also see that higher-order kernels are variance preserving, which is of interest in theory and for very specific applications.

where $|t|_+ = \max\{t, 0\}$. The effect of this manipulation will vanish asymptotically under any assumptions that justify kernel density estimation to begin with.

2.3 Asymptotic Theory

We next develop expressions for the asymptotic bias and variance of $\hat{f}$. The goal is to (i) figure out a good choice of k and h and (ii) conduct inference. Getting there requires modest algebra.

2.3.1 Asymptotic Bias, Variance, and MSE

We start by writing

$$\begin{aligned}\mathbb{E}\hat{f}(x) = \mathbb{E}\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)\right) &= \mathbb{E}\left(\frac{1}{h}k\left(\frac{X_i - x}{h}\right)\right)\\ &= \int_{-\infty}^{\infty}\frac{1}{h}k\left(\frac{t-x}{h}\right)f(t)dt = \int_{-\infty}^{\infty} k\left(u\right)f(x+hu)du.\end{aligned}$$

The last expression has a clear intuition: $\hat{f}(x)$ effectively averages the estimated density over a neighborhood of order h of x. Averages being unbiased estimators of the corresponding expectations, $\mathbb{E}\hat{f}(x)$ is the true expectation corresponding to this average. Note that this argument was not yet asymptotic; however, without asymptotic approximations we are typically stuck here.

Next, a Taylor expansion to the order ν of our kernel k yields

$$f(x+hu) = f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2 + \cdots + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu u^\nu + o(h^\nu)$$

assuming the derivatives up to $f^{(\nu+1)}$ exist. This is not "just a regularity condition": Higher-order kernels can only be used if one is willing to assume that the true density is correspondingly smooth, and this assumption gets seriously restrictive rather quickly. Next, we use linearity of integrals to get

$$\begin{aligned}\mathbb{E}\hat{f}(x) &= \int_{-\infty}^{\infty} f(x)k\left(u\right)du + \int_{-\infty}^{\infty} k\left(u\right)f'(x)hudu + \frac{1}{2}\int_{-\infty}^{\infty} k\left(u\right)f''(x)h^2u^2du\\ &\quad + \cdots + \frac{1}{\nu!}\int_{-\infty}^{\infty} k\left(u\right)f^{(\nu)}(x)h^\nu u^\nu du + o(h^\nu)\\ &= f(x) + f'(x)h\int_{-\infty}^{\infty} uk\left(u\right)du + \frac{1}{2}f''(x)h^2\int_{-\infty}^{\infty} u^2k\left(u\right)du\\ &\quad + \cdots + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\int_{-\infty}^{\infty} u^\nu k\left(u\right)du + o(h^\nu)\\ &= f(x) + \frac{1}{2}f''(x)h^2\kappa^2 + \cdots + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu + o(h^\nu)\\ &= f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu + o(h^\nu)\end{aligned}$$

because the kernel is of order ν. This makes it immediately obvious why one might *in theory* want to use higher order kernels. For the most salient case of nonnegative kernels, we have

$$\mathbb{E}\hat{f}(x) = f(x) + \frac{1}{2}f''(x)h^2\kappa^2 + O(h^4)$$

and hence

$$\text{bias}\big(\hat{f}(x)\big) = \frac{1}{2}f''(x)h^2\kappa^2 + O(h^4).$$

We see that for a given (nonnegative) kernel, the bias is of order $O(h^2)$. Furthermore, it increases (in absolute value) with the dispersion of the kernel and also with the curvature of f at x. Its sign depends on whether f is convex or concave at x and is of lower order where the curvature of f vanishes. This is as it should be: We effectively average over a symmetric neighborhood of order h of x. As this neighborhood becomes small, f is well approximated on it as either linear or parabolic. Under the former approximation, the average becomes unbiased at the relevant rate of localization. Else, the bias is positive [negative] if f is locally convex [concave], meaning that it is locally above [below] the tangent.

Next,

$$\begin{aligned}
\text{var}\big(\hat{f}(x)\big) &= \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{nh^2}\,\text{var}\left(k\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{nh^2}\left(\mathbb{E}\left(k\left(\frac{X_i - x}{h}\right)^2\right) - \left(\mathbb{E}\left(k\left(\frac{X_i - x}{h}\right)\right)\right)^2\right) \\
&= \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{X_i - x}{h}\right)^2\right) - \frac{1}{n}\left(\mathbb{E}\left(\frac{1}{h}k\left(\frac{X_i - x}{h}\right)\right)\right)^2 \\
&= \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{X_i - x}{h}\right)^2\right) - O(1/n),
\end{aligned}$$

using that, from above, $\mathbb{E}\left(\frac{1}{h}k\left(\frac{X_i-x}{h}\right)\right) = \mathbb{E}\hat{f}(x) = f(x) + o(1)$. Plugging in from the following Taylor expansion:

$$\begin{aligned}
&\frac{1}{h}\int_{-\infty}^{\infty}\left(k\left(\frac{z-x}{h}\right)\right)^2 f(z)dz = \int_{-\infty}^{\infty}(k(u))^2 f(x+uh)du \\
&\qquad = \int_{-\infty}^{\infty}(k(u))^2(f(x)+O(h))du = f(x)\int_{-\infty}^{\infty}k(u)^2\,du + O(h) = f(x)R(k) + O(h),
\end{aligned}$$

where the last step defines the **roughness** $R(g) \equiv \int_{-\infty}^{\infty} g(t)^2\,dt$ of a function $g : \boldsymbol{R} \to \boldsymbol{R}$, we find

$$\text{var}\big(\hat{f}(x)\big) = \frac{1}{nh}\left(f(x)R(k) + O(h)\right) + O(1/n) = \frac{1}{nh}f(x)R(k) + O(1/n). \tag{2.1}$$

While we avoid writing h_n to economize on subscripts, we know that we'll send $h \to 0$ as

$n \to \infty$, so $O(1/n) = o(1/(nh))$ and the first term dominates. We also find that the variance is of order $O(1/(nh))$ and is also proportional to the density at x.

Next, the quality of our estimator is plausibly assessed by its mean squared error (MSE):

$$\text{MSE}\big(\hat{f}(x)\big) \equiv \big(\text{bias}\big(\hat{f}(x)\big)\big)^2 + \text{var}\big(\hat{f}(x)\big) = \left(\frac{1}{\nu!} f^{(\nu)}(x) h^\nu \kappa_\nu + o(h^\nu)\right)^2 + \frac{1}{nh} f(x) R(k) + O(1/n).$$

We will always choose h s.t. in the above, all terms other than the *asymptotic mean square error*

$$\text{AMSE}\big(\hat{f}(x)\big) \equiv \left(\frac{1}{\nu!} f^{(\nu)}(x) h^\nu \kappa_\nu\right)^2 + \frac{1}{nh} f(x) R(k)$$

are dominated. However, which of these terms dominates depends on our exact choice of h. Note that as long as $h \to 0$ but slow enough that $nh \to \infty$, AMSE vanishes and the estimator is thus (pointwise) consistent. These requirements make sense intuitively: Bandwidth must vanish for bias to vanish, but slowly enough so that in expectation the sample size effectively used at x diverges.

2.3.2 Asymptotically Optimal Bandwidth

We could try to specify k and h to optimize AMSE($\hat{f}(x)$). However, the solution would usually depend on x. If interest is in overall performance of the estimator, we usually integrate to get the asymptotic mean integrated square error

$$\text{AMISE}(\hat{f}) = \int_{-\infty}^{\infty} \left[\left(\frac{1}{\nu!} f^{(\nu)}(x) h^\nu \kappa_\nu\right)^2 + \frac{1}{nh} f(x) R(k)\right] dx = \left(\frac{1}{\nu!} h^\nu \kappa_\nu\right)^2 R(f^{(\nu)}) + \frac{R(k)}{nh}.$$

In the salient case of nonnegative kernels, this is

$$\text{AMISE}(\hat{f}) = \frac{1}{4} h^4 \kappa_2^2 R(f'') + \frac{R(k)}{nh}.$$

The obvious next question is optimization of kernel and bandwidth. We will first think of optimization of bandwidth given a specific kernel. Then we can write

$$\frac{d}{dh}\text{AMISE}(\hat{f}) = 2\nu \left(\frac{\kappa_\nu}{\nu!}\right)^2 R(f^{(\nu)}) h^{2\nu-1} - \frac{R(k)}{nh^2} \overset{!}{=} 0,$$

which is solved by

$$h^* = \left(\frac{R(k)}{2n\nu \left(\frac{\kappa_\nu}{\nu!}\right)^2 R(f^{(\nu)})}\right)^{1/(2\nu+1)} = \left(\frac{R(k)}{2\nu \left(\frac{\kappa_\nu}{\nu!}\right)^2}\right)^{1/(2\nu+1)} R(f^{(\nu)})^{-1/(2\nu+1)} n^{-1/(2\nu+1)}.$$

Here, the last step just isolates some terms of interest. In particular, the optimal bandwidth is proportional to $n^{-1/(2\nu+1)}$ and therefore vanishes more slowly (in terms of rate and not just constant!) as we move to higher order kernels. This comes from the fact that higher order kernels have lower order bias, so the bias-variance trade-off changes. Of course, this is in turn because higher order kernels impose more smoothness on f – no free lunch here.

Next, we can evaluate AMISE$(\hat{f})$ at the optimal bandwidth and therefore evaluate the value of our optimization problem. After simplification, this yields

$$\text{AMISE}^* = (1+2\nu)\left(\frac{R(k)^{2\nu}\kappa_\nu^2 R(f^{(\nu)})}{(2\nu)^{2\nu}(\nu!)^2}\right)^{1/(2\nu+1)} n^{-2\nu/(2\nu+1)}.$$

This is not a further approximation: The two components of AMISE are of the same order and the addition of two parts is reflected in the + sign. We already mentioned that intuitively, both terms are of the same order at the optimum, but also that this will lead to complications later.[4]

We see that the optimal bandwidth and rate of convergence depend on the order of the kernel. For nonnegative kernels, we get

$$\begin{aligned} h^*\|_{\nu=2} &= \left(\frac{R(k)}{\kappa_2^2 R(f'')}\right)^{1/5} n^{-1/5} \\ \text{AMISE}^*\|_{\nu=2} &= \frac{5}{4}\left(R(k)^4\kappa_2^2 R(f'')\right)^{1/5} n^{-4/5} \end{aligned}$$

At the other extreme, as $\nu \to \infty$, the optimal bandwidth vanishes ever more slowly and the AMISE approaches order n^{-1}. So if the density is infinitely smooth, we can in principle approximate the parametric rate! Of course, this is a purely theoretical consideration.

2.3.3 Practical Approaches to Bandwidth Choice

The obvious next idea would be to simply calculate the optimal bandwidth. We cannot do that because the answer depends on $R(f^{(\nu)})$, which we don't know. The next obvious idea would be to pre-estimate $R(f^{(\nu)})$ and then choose the bandwidth accordingly. However, and unlike with the vaguely similar step of pre-estimating a variance matrix in two-stage GMM, we encounter a regress here: Nonparametrically estimating the roughness of the 2^{nd} or higher derivative of a density is not possible at fast enough rates.

Three possible ways forward are as follows:

Silverman's Rule of Thumb The most popular approach is to resort to a parametric estimator of $R(f^{(\nu)})$ that forces f to be in a narrow parametric class, ideally one such that the implied estimator of $R(f^{(\nu)})$ is easily computed. One such class is the class of centered normal densities, parameterized only by the variance, which we can estimate by the sample

[4] We note in passing that this rate can be shown to be minimax, so it cannot in general be improved upon by a fundamentally different estimation strategy.

variance (Silverman, 1986). Some algebra reveals that the resultant bandwidth scales with the sample standard deviation $\hat{\sigma}$:

$$h^{sil} = \hat{\sigma} c_k n^{-1/(2\nu+1)},$$

where tabulations of the kernel-dependent constant c_k are widely available. Indeed, canned packages will readily implement this rule for you, often as a default option.

Cross-Validation A slightly different approach is to directly estimate the MISE and then choose the bandwidth to minimize the estimate. Write

$$\text{MISE}(h) = \int \mathbb{E}\big(\hat{f}(x) - f(x)\big)^2 dx = \mathbb{E}\left(\int \hat{f}(x)^2 dx\right) - 2\mathbb{E}\left(\int \hat{f}(x) f(x) dx\right) + \int f(x)^2 dx,$$

where for unity of notation I suppress that $\hat{f}$ depends on h. The third term on the r.h.s. does not depend on h and can be ignored. The other terms can be estimated, leading to the cross-validation criterion

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \int \hat{f}_{-i}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i) \approx \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i)$$

where $\hat{f}_{-i}$ is the **leave-one-out** estimator of f that discards observation X_i. Note the following details:

- We seem to be estimating the ISE, i.e. the realization of MISE caused by our data, but it can be shown that $\text{CV}(h)$ is unbiased and consistent for MISE; hence, this procedure is also called unbiased cross-validation.

- In the second sum above, the estimated density at X_i is computed not using X_i itself. What is the idea here? Letting f be the true density of X_i and g be some known function, it would be obvious to estimate $\int g(x)f(x)dx = \mathbb{E}g(X_i)$ through the sample analog $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$. However, in our case we have $g = \hat{f}$ and therefore the nontrivial complication that the function evaluated at X_i depends on X_i beyond taking X_i as an argument. The leave-one-out estimator shuts down that dependency. This is a trick that you will see again and again, including in neighboring fields like Machine Learning.

- Strictly speaking, this logic gives rise to the expression left of ≈ above (Bowman, 1984). However, the computationally much simpler r.h. expression (Hall, 1983) is asymptotically equivalent and is now the industry standard; e.g., it is implemented in R option `bw="bw.ucv"` in `density`. Note that the leave-one-out estimator is avoided only for computation of the integral.

The cross-validation estimator $\hat{h}$ of the optimal bandwidth is consistent under fairly general conditions and also in more general settings than considered here. Indeed, cross-validation is by now an extremely widely used tool wherever so-called tuning parameters (here: a bandwidth) have to be chosen. The rate of convergence, however, is painfully slow, with $\frac{\hat{h}-h^*}{h^*} = O_P(n^{-1/10})$ in this example.

2.3.4 Asymptotically Optimal Kernel

The choice of kernel order is principally constrained by assumptions one is willing to make on smoothness. However, and despite results about order of bias above, I do not advise to use the highest order kernel that you can maybe justify. Indeed, your default should be to use nonnegative kernels. We have more to say on the choice of kernels within a given order. In particular, the kernel affects AMISE$^*(\hat{f})$ through a positive power of $R(k)^\nu \kappa_\nu$. Let's normalize $\kappa_\nu = 1$. This is justified because optimal kernels are "identified" only up to horizontal scale; the bandwidth can be used to implicitly rescale any kernel along the u-axis and thereby freely rescale one nonzero moment. Hence, the asymptotically optimal kernel solves

$$\min R(k) \quad \text{s.t.} \quad \int_{-\infty}^{\infty} k(u)du = 1; \int_{-\infty}^{\infty} u^j k(u)du = 0, j = 1, \dots, \nu-1; \int_{-\infty}^{\infty} u^\nu k(u)du = 1,$$

where the choice of 1 in the last constraint makes the problem well-defined but is substantively w.l.o.g. (see homework). Müller (1984) showed that the solution to this problem is the Epanechnikov kernel. For $\nu = 2$, this kernel is[5]

$$k^{epa}(u) = \frac{3}{4}\left|1-u^2\right|_+ .$$

Note this kernel's simple geometry as the positive truncation of a parabola. Higher order kernels can be generated from nonnegative kernels by multiplication with certain polynomials, yielding the higher order analogs

$$\begin{aligned} k_4^{epa}(u) &= \frac{15}{8}\left(1-\frac{7}{3}u^2\right)k^{epa}(u) \\ k_6^{epa}(u) &= \frac{175}{64}\left(1-6u^2+\frac{33}{5}u^4\right)k^{epa}(u). \end{aligned}$$

The *efficiency* of any other kernel is defined as

$$\text{eff}(k) = \left(\frac{\text{AMISE}^*(k)}{\text{AMISE}^*(k_\nu^{epa})}\right)^{\frac{1+2\nu}{2\nu}},$$

[5] This presentation follows my preferred convention to have bounded kernels be supported on $[-1,1]$. This kernel has $\kappa^2 = 1/5$, and so the version that respects the normalization $\kappa_\nu = 1$ stretches it out horizontally by $\sqrt{5}$ as in the Stata manual.

where ν is the order of the kernel under consideration. This number always greater than 1 and is taken to a power that compensates the rate at which AMISE decays with n. This makes for a simple interpretation: The AMISE of the Epanechnikov kernel with n observations equals the AMISE of any other kernel k with $\text{eff}(k) \times n$ observations. The efficiency loss can therefore be interpreted as excess data needed compared to the efficient kernel.

Numerically, efficiency losses of many kernels are tiny. The simplest kernels, i.e. Gaussian and uniform, stand out as inefficient among frequently used kernels and yet their efficiencies are 1.05 respectively 1.08. This suggests – and it is also borne out in practice – that no "reasonable" kernel choice should hurt too much and that maybe bandwidth is more important. That said, it is typically easy to just use the Epanechnikov kernel, which is available in any canned kernel density estimation codes (e.g. R `density`, Matlab `fitdist`) and is frequently the default (e.g. Stata `kdensity`).[6]

2.3.5 Asymptotic Distribution and Inference

The asymptotic distribution of the estimator is straightforward to derive but, compared to the parametric case, raises its own novel problems. Recalling we assume i.i.d. data, we can use (2.1) and invoke a Central Limit Theorem to get

$$\sqrt{nh}\big(\hat{f}(x) - \mathbb{E}\hat{f}(x)\big) \overset{d}{\to} N(0, f(x)R(k)).$$

So far, so good, but note that I centered $\hat{f}(x)$ at its own expectation, not at the true value of $f(x)$. Whether this matters depends on the bandwidth. In particular, an optimal bandwidth choice implies that bias and variance are of the same order, and then this difference does matter and leads to a noncentered asymptotic distribution.

To work this out a bit, recall that up to very good approximation, the bias is $\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}$. Suppose that we optimally choose $h = Cn^{-1(2\nu+1)}$ for some constant C. Then we can solve for $\sqrt{nh}h^{\nu} = C^{\nu+1/2}$ and write

$$\sqrt{nh}\big(\hat{f}(x) - f(x)\big) \overset{d}{\to} N\left(\frac{\kappa_{\nu}C^{\nu+1/2}}{\nu!}f^{(\nu)}(x), f(x)R(k)\right),$$

in particular for nonnegative kernels

$$\sqrt{nh}\big(\hat{f}(x) - f(x)\big) \overset{d}{\to} N\left(\kappa^2 C^{3/2} f''(x)/2, f(x)R(k)\right).$$

This asymptotic bias term cannot be pre-estimated. We might of course try to estimate $f''(x)$, adjust our estimator by a plug-in estimate of its bias, and hope that we now get a centered limiting distribution. But this would require to estimate $f''(x)$ with a bias that is of

[6] These results are specific to density estimation. The Gaussian kernel is a popular default because it is applicable and "good enough" in a wide range of applications.

lower order than the above. In fact, we will later see that the bias is of the same (for the same bandwidth) or higher (for the bandwidth that is actually optimal for derivative estimation) order, and the MSE of the derivative estimator is always of higher order. So this is not going to happen.

In practice, researchers (including this lecturer) usually resolve this by assuming the bias is zero to the relevant order of approximation, thus they conduct inference as if

$$\sqrt{nh}\big(\hat{f}(x) - f(x)\big) \overset{d}{\to} N(0, f(x)R(k)).$$

This is justified by claiming that one's bandwidth is **undersmoothed**, i.e. smaller than the optimal one, which increases variance and decreases bias. Under this assumption, the asymptotic approximation is indeed formally justified. However, this trick has at least two downsides:

- The confidence interval can always be interpreted as set-valued estimator. In contrast to parametric settings, this set estimator now converges at a slower rate than the best available point estimator. So in large samples, this estimator is very large compared to the optimal estimator ± two standard errors (but of course, the latter is not a valid confidence interval!) and is also not guaranteed to contain the optimal estimator. In practice, if we report this CI, we should also report the corresponding estimator, but then of course we are sacrificing an order of estimation precision in order to be able to do inference.

- While we claim a rate for our bandwidth, in a given application, we of course choose a fixed bandwidth. In principle, we could claim for any such bandwidth that it is embedded in a sequence which converges at our rate of choice. This issue arises elsewhere too but is especially tricky here. There is no clear rule on what we're allowed to do, but for example, using a Rule-of-Thumb bandwidth and then claiming asymptotically centered estimates would be overstepping. More generally, while there is a neat theory for choosing the optimal bandwidth, I am not aware of useful guidance for choosing an optimally undersmoothed bandwidth.

Assume now that we did undersmooth and hence have available the centered normal limiting distribution. Then conducting pointwise inference is straightforward. (Uniform and simultaneous confidence bands are not! We omit them here.) In particular, we could report

$$CI = \left[\hat{f}(x) - \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\hat{f}(x)R(k)}{nh}}, \hat{f}(x) + \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\hat{f}(x)R(k)}{nh}}\right].$$

If $\hat{f}(x)$ is close to zero, this can become awkward as the left boundary may become negative.

This can be circumvented by explicitly inverting a hypothesis test:

$$CI = \left\{ f : \sqrt{nh} \frac{|\hat{f}(x) - f|}{\sqrt{fR(k)}} \leq \Phi^{-1}(1 - \alpha/2) \right\}.$$

This is computationally more burdensome as the critical value must be recomputed at each f. It also must be taken with a grain of salt because in expectation, it really matters only if the sampling distribution has considerable skewness, in which case a CLT hasn't "kicked in" yet; this undermines justification of Φ^{-1} for critical values.

Notice finally that these confidence intervals are pointwise in a very strong sense. Compared to confidence bands for f, they not only abstract from the multiple hypothesis testing problem inherent in the latter, but they also are not uniformly valid over regions of parameter space where $f(x) \to 0$ (because the variance term then vanishes). Thus, a confidence band constructed from them may not even be valid uniformly over all points, not to mention for all points simultaneously.

2.4 Extensions

We next discuss two extensions of univariate density estimation: derivatives of densities and multivariate densities. In both cases, the technical development is similar to the above except for being more tedious. We will therefore be cursory about many details. The message is not that these extensions are unimportant – on the contrary! But I hope that the preceding development gave you a good idea of how results were derived, and actually doing the derivations becomes very tedious.

An important insight will be that, while results superficially resemble the above, optimal bandwidth rates, and therefore also rates of convergence of optimal estimators, depend on the estimand. This phenomenon is typical for nonparametric estimation and inference. In the specific case of higher-dimensional density estimation, it gives rise to the **curse of dimensionality**: Rates of consistency deteriorate as the dimension of parameter space increases.

2.4.1 Multivariate Densities

Consider estimation of the density f of the random vector $\boldsymbol{X}_i \in \boldsymbol{R}^D$. Then a kernel is a function $K : \boldsymbol{R}^D \mapsto \boldsymbol{R}$. Conceptually, the only restriction on a kernel is that $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$, but we will restrict attention to kernels of the product form

$$K(\boldsymbol{u}) = k(u_1)k(u_2)\ldots k(u_D),$$

where $\boldsymbol{u} = (u_1, \ldots, u_D)$ and where k is symmetric. Thus, the kernel function is a product of D identical, component-wise kernels. If k is nonnegative, this can be thought of as the density corresponding to $(u_1, \ldots, u_D)$ being distributed independently with density k. Rather than one bandwidth, we now have a bandwidth matrix $\boldsymbol{H} = \text{diag}(h_1, \ldots, h_D)$, where we define $|\boldsymbol{H}| = h_1 \times \cdots \times h_D$. The density estimator is

$$\hat{f}(x) = \frac{1}{nh_1 \ldots h_D} \sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_{d,i} - x_d}{h_d}\right) = \frac{1}{n\,|\boldsymbol{H}|} \sum_{i=1}^{n} K\left(\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\right).$$

One can then derive

$$\begin{aligned}
\text{bias}\big(\hat{f}(\boldsymbol{x})\big) &= \frac{\kappa_\nu}{\nu!} \sum_{d=1}^{D} \frac{\partial^\nu}{\partial x_d^\nu} f(\boldsymbol{x}) h_d^\nu + o(h_1^\nu + \ldots + h_D^\nu) \\
\text{var}\big(\hat{f}(\boldsymbol{x})\big) &= \frac{f(\boldsymbol{x}) R(k)^D}{n\,|\boldsymbol{H}|} + O(1/n) \\
\text{AMISE}\big(\hat{f}\big) &= \left(\frac{\kappa_\nu}{\nu!}\right)^2 \int \left(\sum_{d=1}^{D} \frac{\partial^\nu}{\partial x_d^\nu} f(\boldsymbol{x}) h_d^\nu\right)^2 d\boldsymbol{x} + \frac{R(k)^D}{n\,|\boldsymbol{H}|}.
\end{aligned}$$

The last expression cannot be minimized in closed form. However, we observe that the univariate kernel k – which is still “identified” only up to scale – only enters through $R(k)$, so the same kernel remains optimal. Also, we can easily ascertain the rate of the optimal bandwidth: On the assumption that all components of the bandwidth are of the same order, we can set $h^{2\nu} = cn^{-1}h^{-D}$ (here, c is some constant) and solve for $h \propto n^{-1/(2\nu+D)}$, implying that AMISE* is of order $O(n^{-2\nu/(2\nu+D)})$, specializing to $O(n^{-4/(4+D)})$ if the kernel is nonnegative. This is the **curse of dimensionality**.

In order to optimize bandwidth in practice, we usually do two things:

- The D componentwise bandwidths are constrained to be of the same order, scaled only to adjust to data variability. Thus, $h_d = h\hat{\sigma}_d$, where h is common across components and $\hat{\sigma}_d$ estimates the standard deviation of $x_{d,i}$. This also means the ratios of component bandwidths are scale invariant.

- We then resort to rules of thumb or to cross-validation to determine h.

The algebra comes out somewhat differently from above, but the basic ideas are the same, and for the sake of brevity we’ll leave it at that.

2.4.2 Estimation of Derivatives

Derivatives often have interpretation, e.g. (after rescaling) as elasticities in demand estimation. So there is ample motivation for estimating them. To do that, we return to the scalar case but the estimand now is the r’th derivative $f^{(r)}$ of f.

The obvious estimator is the plug-in estimator

$$\widehat{f^{(r)}} = \hat{f}^{(r)} = \frac{1}{nh^{1+r}} \sum_{i=1}^{n} k^{(r)} \left(\frac{X_i - x}{h} \right).$$

At a minimum, that requires $k^{(r)}$ to exist, constraining our choice of kernel. Furthermore, we will see that optimal bandwidth results are yet again different from the previous chapter's reference results. Let's compute the bias:

$$\begin{aligned}
\mathbb{E}\big(\widehat{f^{(r)}}(x)\big) &= \int_{-\infty}^{\infty} \frac{1}{h^{1+r}} k^{(r)} \left(\frac{t-x}{h} \right) f(t) dt \\
&= \int_{-\infty}^{\infty} \frac{1}{h^{r}} k^{(r-1)} \left(\frac{t-x}{h} \right) f'(t) dt \\
&= \ldots = \int_{-\infty}^{\infty} \frac{1}{h} k \left(\frac{t-x}{h} \right) f^{(r)}(t) dt,
\end{aligned}$$

where we integrated by parts r times. In analogy to previous arguments, we get

$$\int_{-\infty}^{\infty} \frac{1}{h} k \left(\frac{t-x}{h} \right) f^{(r)}(t) dt = \int_{-\infty}^{\infty} k\left(u\right) f^{(r)}(x+hu) dt = f^{(r)}(x) + \frac{1}{\nu!} f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu} + o(h^{\nu}),$$

so the leading term of the bias equals $f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu} / \nu!$. Note that this argument requires existence of derivatives of f up to $r + \nu + 1$, so higher order kernels call for even stronger assumptions here. Next,

$$\begin{aligned}
\operatorname{var}\big(\widehat{f^{(r)}}(x)\big) &= \frac{1}{nh^{2+2r}} \operatorname{var}\left(k^{(r)} \left(\frac{X_i - x}{h} \right) \right) \\
&= \frac{1}{nh^{2+2r}} \left[\mathbb{E} \left(k^{(r)} \left(\frac{X_i - x}{h} \right) \right)^2 - \left(\mathbb{E} \left(k^{(r)} \left(\frac{X_i - x}{h} \right) \right) \right)^2 \right] \\
&= \frac{1}{nh^{2+2r}} \left[h \int_{-\infty}^{\infty} k^{(r)}\left(u\right)^2 f(x+hu) du - \left(h \int_{-\infty}^{\infty} k^{(r)}\left(u\right) f(x+hu) du \right)^2 \right] \\
&= \frac{f(x)}{nh^{1+2r}} \int_{-\infty}^{\infty} k^{(r)}\left(u\right)^2 du - \frac{1}{nh^{2r}} \left(\int_{-\infty}^{\infty} k^{(r)}\left(u\right) du \right)^2 \\
&= \frac{f(x) R(k^{(r)})}{nh^{2+2r}} + O\left(\frac{1}{nh^{2r}} \right),
\end{aligned}$$

Next, we can solve to get

$$\begin{aligned}
\operatorname{AMSE}\big(\widehat{f^{(r)}}(x)\big) &= \frac{f^{(r+\nu)}(x)^2 h^{2\nu} \kappa_{\nu}^2}{(\nu!)^2} + \frac{f(x) R(k^{(r)})}{nh^{1+2r}} \\
\operatorname{AMISE}\big(\widehat{f^{(r)}}(x)\big) &= \frac{R(f^{(r+\nu)}) h^{2\nu} \kappa_{\nu}^2}{(\nu!)^2} + \frac{R(k^{(r)})}{nh^{1+2r}},
\end{aligned}$$

thus the variance (the r.h. term) is much larger than before! This changes the bias-variance trade-off and we'll have to use a larger bandwidth. In particular, we find

$$\begin{aligned} h_r^* &= \left(\frac{(1+2r)\,(\nu!)^2\, R(k^{(r)})}{2\nu\kappa_\nu^2 R(f^{(r+\nu)})} \right)^{1/(1+2r+2\nu)} n^{-1/(1+2r+2\nu)} \\ \text{AMISE}^* &= (1+2r+2\nu) \left(\frac{\kappa_\nu^2}{(1+2r)\,(\nu!)^2} \right)^{\frac{1+2r}{1+2r+2\nu}} \left(\frac{R(k^{(r)})}{2\nu} \right)^{\frac{2\nu}{1+2r+2\nu}} n^{\frac{-2\nu}{1+2r+2\nu}}, \end{aligned}$$

which for nonnegative kernels specializes to

$$\begin{aligned} h_r^* &= \left(\frac{(1+2r)R(k'')}{\kappa_2^2 R(f^{(r+2)})} \right)^{1/(5+2r)} n^{-1/(5+2r)} \\ \text{AMISE}^* &= (5+2r) \left(\frac{\kappa_2^2}{4+8r} \right)^{\frac{1+2r}{5+2r}} \left(\frac{R(k^{(r)})}{4} \right)^{\frac{4}{5+2r}} n^{\frac{-4}{5+2r}}. \end{aligned}$$

As one would have hoped, the rates recover previous rate results for $r = 0$, but they also illustrate considerable penalty for trying to estimate derivatives.

The asymptotically optimal nonnegative kernels now are the biweight kernel

$$k^{bi}(u) = \frac{15}{16} \left(\left|1-u^2\right|_+ \right)^2$$

for estimating f' and the triweight kernel

$$k^{tri}(u) = \frac{35}{32} \left(\left|1-u^2\right|_+ \right)^3$$

for estimating f''. Both can be generalized to higher orders and continue to be optimal for these estimands. Note that the Gaussian kernel turns out to be quite inefficient for these tasks, and the uniform, Epanechnikov, as well as any other nondifferentiable kernels are inapplicable.

3 Nonparametric Mean Regression

We next consider estimation of

$$m(\boldsymbol{x}) \equiv \mathbb{E}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}).$$

This is called *nonparametric mean regression.* It generalizes OLS and many other models, as becomes espcially obvious upon equivalently writing

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0.$$

Note that in contrast to the linear model,

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0,$$

the stipulation $\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$ is a normalization, not a substantive assumption. We will also not assume homoskedasticity and generally treat $\sigma^2(\boldsymbol{x}) \equiv \mathbb{E}(\varepsilon_i^2|\boldsymbol{X}_i = \boldsymbol{x})$ as nonconstant. Indeed, the only substantive assumption we make is that data are i.i.d., that $m(\boldsymbol{x})$ is well-defined, and (frequently suppressed) smoothness conditions on f and σ^2. We will explicitly analyze mostly the case of scalar X_i and keep notation consistent by using f for the density of X_i. The extension to multidimensional $\boldsymbol{X}_i$ is theoretically (not computationally) routine.

3.1 Nadaraya-Watson Estimator

An intuitively obvious estimator that naturally relates to kernel density estimation is the **local constant** or **Nadaraya-Watson** estimator:

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

Very similarly to kernel density estimation, this can be intuited as smoothly generalizing the "moving window average"

$$\hat{m}^{uni}(x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}(|X_i - x| \leq h)}{\sum_{i=1}^n \mathbf{1}(|X_i - x| \leq h)}$$

which in turn specializes $\hat{m}$ to the case of uniform kernels. Note also that the estimator can be written as

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

where it is understood that $\hat{f}$ uses the same kernel and bandwidth. We will restrict attention to nonnegative kernels to avoid dealing with $\hat{f}(x) \leq 0$.

Taking a cue from the last representation of $\hat{m}$ and using $Y_i = m(X_i) + \varepsilon_i$, write

$$\begin{aligned}
\frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i k\left(\frac{X_i-x}{h}\right)}{\hat{f}(x)} &= \frac{1}{nh\hat{f}(x)}\sum_{i=1}^{n} k\left(\frac{X_i-x}{h}\right)(m(x)+m(X_i)-m(x)+\varepsilon_i) \\
&= \frac{1}{nh\hat{f}(x)}\sum_{i=1}^{n} k\left(\frac{X_i-x}{h}\right)m(x) + \frac{1}{nh\hat{f}(x)}\sum_{i=1}^{n} k\left(\frac{X_i-x}{h}\right)(m(X_i)-m(x)) \\
&\quad + \frac{1}{nh\hat{f}(x)}\sum_{i=1}^{n} k\left(\frac{X_i-x}{h}\right)\varepsilon_i \\
&= m(x) + \frac{\hat{m}_1(x)}{\hat{f}(x)} + \frac{\hat{m}_2(x)}{\hat{f}(x)},
\end{aligned}$$

where the last step defines m_1 and m_2 and observes simplification in the first term.

A simple application of the Law of Iterated Expectations yields $\mathbb{E}\hat{m}_2(x) = 0$, hence the bias of the estimator equals $\mathbb{E}(\hat{m}_1(x)/\hat{f}(x))$. In contrast, and less obviously, the contribution of $\hat{m}_2(x)$ to the variance dominates, and so in practice this is the "variance term." Write

$$\begin{aligned}
\operatorname{var}(\hat{m}_2(x)) &= \frac{1}{nh^2}\operatorname{var}\left(k\left(\frac{X_i-x}{h}\right)\varepsilon_i\right) = \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{X_i-x}{h}\right)\varepsilon_i\right)^2 \\
&= \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{X_i-x}{h}\right)^2\sigma^2(X_i)\right) = \frac{1}{nh^2}\int k\left(\frac{t-x}{h}\right)^2\sigma^2(t)f(t)dt \\
&= \frac{1}{nh}\int k(u)^2\sigma^2(x+hu)f(x+hu)du = \frac{1}{nh}\int k(u)^2\sigma^2(x)f(x)du + o(1/nh) \\
&= \frac{R(k)\sigma^2(x)f(x)}{nh} + o(1/nh),
\end{aligned}$$

where the last step used smoothness assumptions on σ^2 and f.

Next,

$$\begin{aligned}
\mathbb{E}\hat{m}_1(x) &= \frac{1}{h}\mathbb{E}\left(k\left(\frac{X_i-x}{h}\right)(m(X_i)-m(x))\right) = \frac{1}{h}\int k\left(\frac{t-x}{h}\right)(m(X_i)-m(x))f(t)dt \\
&= \int k(u)(m(x+hu)-m(x))f(x+hu)du \\
&= \int\left(k(u)\left(hum'(x)+\frac{h^2u^2}{2}m''(x)\right)(f(x)+huf'(x))\right)du + o(h^2) \\
&= \underbrace{\int k(u)udu}_{=0}\cdot hm'(x)f(x) + \int k(u)u^2du \times h^2\left(\frac{1}{2}m''(x)f(x)+m'(x)f'(x)\right) + o(h^2) \\
&= \kappa^2h^2f(x)\cdot\underbrace{\left(\frac{1}{2}m''(x)+m'(x)f'(x)/f(x)\right)}_{\equiv B(x)} + o(h^2).
\end{aligned}$$

(If you replicate this, note that the $o(h^2)$ absorbed a $O(h^3)$ cross-product term of the integral.) By similar algebra, it can be verified that $\mathrm{var}(\hat{m}_1(x)) = o(1/nh)$, so that we can use Chebyshev's inequality to get

$$\begin{aligned} & \sqrt{nh}\left(\hat{m}_1(x) - h^2\kappa_2 B(x) f(x)\right) \xrightarrow{p} 0 \\ \Leftrightarrow \quad & \sqrt{nh}\big(\hat{m}_1(x)/\hat{f}(x) - h^2\kappa_2 B(x)\big) \xrightarrow{p} 0, \end{aligned}$$

where we used $f(x)/\hat{f}(x) \xrightarrow{p} 1$ (if you feel uneasy about that, see below). In sum, and also invoking a CLT,

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2\kappa^2 B(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)\sigma^2(x)}{f(x)}\right).$$

Next, we compute

$$\mathrm{AMSE}(\hat{m}(x)) = h^4\kappa_2^2 B(x)^2 + \frac{R(k)\sigma^2(x)}{nhf(x)}.$$

It would be intuitive to just take the expectation w.r.t. X_i to get an AMISE, but that integral will not in general exist. The culprit is the $\sigma^2(x)/f(x)$ term. We therefore define a weighted integrated MSE that introduces a weighting function w:

$$\mathrm{WIMSE}(\hat{m}) \equiv h^4\kappa_2^2 \int B(x)^2 w(x)dx + \frac{R(k)}{nh}\int \frac{\sigma^2(x)}{f(x)} w(x)dx.$$

The weighting function must discount low values of f. For example, we could use $w(x) = \mathbf{1}\{f(x) \geq \delta\}$ for some tuning parameter δ. Beyond the fact that $\sigma^2(x)/f(x)$ might not be integrable, this also helps with a leap of faith in the above development: We relied on $f(x)/\hat{f}(x) \xrightarrow{p} 1$, but that is not true uniformly over small f.

We will not crank out optimized losses, but note that, by setting $h^4 = 1/nh$, we can easily solve for $h^* \propto n^{-1/5}$, and plugging in we then find that $\mathrm{WIMSE}^* \propto n^{-4/5}$, hence the optimized rate of convergence is $n^{-2/5}$. These rates should look familiar – they are the same as for density estimation. The generalization to multivariate mean regression has this same feature, including the curse of dimensionality. Furthermore, optimal kernel theory is unchanged: Normalizing $\kappa^2 = 1$, we see that WIMSE depends on our choice of kernel through $R(k)$, hence the optimal kernel is Epanechnikov.

Optimal bandwidth is a much more complex question. While rule-of-thumb bandwidths are frequently used, they were not developed for this application and in particular do not try to pre-estimate B. I rather recommend cross-validation.

The Nadaraya-Watson estimator is an important default, and it actually performs pretty well in some settings, notably in somewhat higher dimension if some covariates do not really matter. However, it has some unappealing features:

- The limit case as $h \to \infty$ is not a linear fit but a horizontal fit corresponding to the constant function $y = \overline{y}$. (Hence, "local constant estimation," turning into "global constant estimation" as $h \to \infty$.)

- The estimator will be badly biased, indeed inconsistent, at the boundaries of the support unless $m(x)$ is constant there. The reason is that at the lower boundary, observed X_i will be overwhelmingly above x, so that we have upward [downward] bias if m is increasing [decreasing], and similarly at the upper boundary.

- The estimator transforms a perfectly linear scatterplot into a nonlinear fit unless the x are perfectly evenly spaced. To see this, imagine a uniform kernel and say that the scatterplot is perfectly linearly increasing, i.e. there exist (α, β) s.t. $Y_i = \alpha + \beta X_i$ for all $i = 1, \ldots, n$. Then

$$\hat{m}(x) = \frac{\sum_i (\alpha + \beta X_i) 1\{|X_i - x| \leq h\}}{\sum_i 1\{|X_i - x| \leq h\}} = \alpha + \beta \frac{\sum_i X_i 1\{|X_i - x| \leq h\}}{\sum_i 1\{|X_i - x| \leq h\}}$$

  does *not* in general equal $\alpha + \beta x$, nor is it in general linear in x. In fact, it's a step function; that could be fixed by smoothing the kernel, but the nonlinearity cannot.

These considerations led researchers to consider generalizations of the local constant estimator and also completely different approaches. We will discuss the former in some detail and the latter very briefly.

3.2 Local Polynomial Regression

3.2.1 Description

The class of local polynomial estimators can be intuitively described as follows: At each x, fit a local (to x) polynomial approximation to $m(x)$. "Local" is operationalized by weighting all data points according to a kernel density centered at x. The intuition is again clearest with a uniform kernel, in which case we simply restrict attention to the data in window $[x - h, x + h]$. Polynomial approximation means that we regress Y_i on a constant, X_i, X_i^2, and so on. The estimator $\hat{m}(x)$ equals the fitted value of the regression at x. Operationally, other than weighting the data, we will recenter them at x, i.e. replace each X_i with $(X_i - x)$. The fitted value of the regression at x, and therefore the estimator $\hat{m}(x)$, is then just the estimated constant $\hat{\alpha}$. Conceptually, this procedure is repeated at each x; in practice, we do it on a grid.

Computationally, local polynomial regression can utilize existing routines for regression. The algorithm is as follows:

1. At each x where the regression is to be evaluated, replace each X_i with $\tilde{x}_i = X_i - x$ and assign it weight $w(\tilde{x}_i) = k(\tilde{x}_i / h)$.

2. Run weighted least squares (WLS) of Y_i on a constant, $\tilde{x}_i$, $\tilde{x}_i^2$, $\ldots$ up to desired order of polynomial. Thus, we have the closed-form regression coefficients

$$\begin{aligned}\hat{\boldsymbol{\beta}}_x &= \left(\sum_i k(\tilde{x}_i/h)\boldsymbol{Z}_i\boldsymbol{Z}_i'\right)^{-1}\sum_i k(\tilde{x}_i/h)\boldsymbol{Z}_iY_i,\\ \boldsymbol{Z}_i &= (1,\tilde{x}_i,\tilde{x}_i^2,\ldots,\tilde{x}_i^k)'.\end{aligned}$$

3. The estimator $\hat{m}(x)$ is the estimated intercept.

If we locally regress Y_i on a constant and $\tilde{x}_i$, we have local linear regression. If we regress Y_i on the constant only, we have local constant regression, i.e. we recover the Nadaraya-Watson estimator as a special case.

Similarly to the intuition already given for Nadaraya-Watson, all local polynomial approximations collapse to the corresponding global approximation as $h \to \infty$, e.g. local linear approximation collapses to OLS. While one can imagine going to relatively high polynomials with scalar X_i, the number of coefficients estimated in the local fit otherwise increases rapidly with order of polynomials (because of cross-product terms). Also taking into account that effective sample size can only grow slowly with n in higher dimension, our ability to locally fit higher order polynomials in higher dimensional covariate space is severely limited with the sample sizes that economists typically encounter. In practice, local polynomials of order greater than 1 are almost exclusively used in scalar problems.

3.2.2 Asymptotics

The derivations of asymptotic bias, variance, etc. become extremely involved. We will state without proof that for local linear estimation,

$$\sqrt{nh}\left(\hat{m}_{LL}(x)-m(x)-h^2\kappa^2m''(x)/2\right)\xrightarrow{d}N\left(0,\frac{R(k)\sigma^2(x)}{f(x)}\right).$$

Note that the bias is proportional to $m''(x)$; this is intuitive upon visually comparing local linear and local quadratic approximations.

The result can be directly compared to local constant estimation. Undoing the definition of $B(x)$ above, we get

$$\sqrt{nh}\left(\hat{m}_{LC}(x)-m(x)-h^2\kappa^2\left(\frac{1}{2}m''(x)+m'(x)f'(x)/f(x)\right)\right)\xrightarrow{d}N\left(0,\frac{R(k)\sigma^2(x)}{f(x)}\right).$$

Hence, the asymptotic distributions have the same variance and differ by a translation of $h^2\kappa_2m'(x)f'(x)/f(x)$. While we cannot in general rank the bias terms, the common sense is that "typically" the simpler bias term that does not involve m' will also be smaller. This intuition guides practice and accounts for the popularity of $\hat{m}_{LL}$, but it is not a theorem.

Note that $\hat{m}_{LL}$ tends to smooth the data a bit less, which may translate into higher finite sample variance.

With that all said, note that the order of bias has not changed, and optimal rates are still the same as for kernel density estimation with nonnegative kernels. This may be surprising, and because of higher order polynomials' ability to fit all different kinds of curves, one might expect it to change as we increase the polynomial order, though with obvious downsides in terms of finite sample performance. This is indeed so, and the order of bias goes down as we move on to 2^{nd} order polynomials.

In general, the bias and variance of higher order polynomials iterate the above pattern as follows:

- For local polynomial estimation of odd order k, the bias is $O(h^{k+1})$. The leading term is $m^{(k+1)}(x)$.
- For local polynomial estimation of even order k, the bias is $O(h^{k+2})$. The leading terms are $m^{(k+2)}(x)$ and $m^{(k+1)}(x)f'(x)/f(x)$. The second term is sometimes called "design bias."

Thus, we gain an order of bias only as we increase the order from odd to even (in increments of h^2), though we effectively lose one of the two bias terms as we increase the order from even to odd.

3.2.3 Cross-Validation

The leading way to determine bandwidths for local polynomial regression is cross-validation. The intuition is minimization of a least squares criterion, i.e. to choose h so as to minimize an estimator of $\mathbb{E}\varepsilon_i^2$. We have to be careful about two things. First, it is essential to use the leave-one-out estimator because we otherwise get a degenerate result; do you see why? Second, we disregard behavior of the estimator in regions of very low density of X_i, e.g. the tails of a normal distribution, by introducing a weighting or trimming function as with weighted integrated mean square error. The criterion then is

$$CV_{LS}(h) = \frac{1}{n}\sum_i \hat{\varepsilon}_{-i}^2 w(X_i),$$

where $\hat{\varepsilon}_{-i} = Y_i - \hat{m}_{-i}(X_i)$ is the leave-one-out residual and $\hat{m}_{-i}(X_i)$ is the leave-one-out estimator.

This is a good estimator because

$$\mathbb{E}CV_{LS}(h) = \text{WIMSE}(h) + \mathbb{E}\left(\varepsilon_i^2 w(X_i)\right),$$

and the right-hand side can be interpreted as weighted integrated squared prediction error that compounds variability of Y_i with parameter estimation error. Furthermore, since $\mathbb{E}\left(\varepsilon_i^2 w(X_i)\right)$ does not depend on h, the bandwidth that minimizes $CV_{LS}(h)$ estimates (by usual m-estimator arguments) the bandwidth that minimizes WIMSE(h).

To see the claim, write

$$\begin{aligned}\mathbb{E}CV_{LS}(h) &= \mathbb{E}\big((\varepsilon_i + m(X_i) - \hat{m}_{-i}(X_i))^2\, w(X_i)\big)\\ &= \mathbb{E}\big((m(X_i) - \hat{m}_{-i}(X_i))^2\, w(X_i)\big) + \mathbb{E}\left(\varepsilon_i^2 w(X_i)\right) - 2\underbrace{\mathbb{E}\left(\left(m(X_i) - \hat{m}_{-i}(X_i)\right)\varepsilon_i w(X_i)\right)}_{=0},\end{aligned}$$

where we used $\mathbb{E}(\varepsilon_i|X_i) = 0$. It then remains to observe that

$$\mathbb{E}\big((m(X_i) - \hat{m}_{-i}(X_i))^2\, w(X_i)|x_{-i}\big) = \int_{-\infty}^{\infty} (m(t) - \hat{m}_{-i}(t))^2\, w(t)f(t)dt$$

for any leave-one-out sample x_{-i}, hence

$$\begin{aligned}\mathbb{E}\big((m(X_i) - \hat{m}_{-i}(X_i))^2\, w(X_i)\big) &= \int_{-\infty}^{\infty} \mathbb{E}\left(m(t) - \hat{m}_{-i}(t)\right)^2 w(t)f(t)dt\\ &= \int_{-\infty}^{\infty} \mathbb{E}\left(m(t) - \hat{m}(t)\right)^2 w(t)f(t)dt = \text{WIMSE}(h).\end{aligned}$$

3.3 Semiparametric Mean Regression

Semiparametric methods combine parametric and nonparametric aspects. The aim is to get the best of both worlds: nonparametric flexibility where it matters but also avoiding the curse of dimensionality. Some authors further differentiate between semiparametrics and seminonparametrics depending on whether the quantities of interest or the nuisance parameters are nonparametrically estimated.

Classic examples of semiparametric models are as follows.

Example 3.1 (Partially Linear Model):

$$\mathbb{E}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + m(\boldsymbol{Z}_i)$$

Example 3.2 (Separable Model):

$$\mathbb{E}(Y_i|X_{1,i}, \dots, X_{K,i}) = \sum_{k=1}^{K} m_k(X_{k,i})$$

Example 3.3 (Linear Index Model):

$$\mathbb{E}(Y_i|\boldsymbol{X}_i) = m(\boldsymbol{X}_i'\boldsymbol{\beta})$$

Example 3.4 (Generated Regressors):

$$\mathbb{E}(Y_i|\boldsymbol{X}_i) = \boldsymbol{m}(\boldsymbol{X}_i)'\boldsymbol{\beta}$$

The examples differ with respect to what's typically of interest. In the partially linear model, the modal application is interested in m, and the $\mathbf{x}_i$ are controls. They are handled linearly in the hope to get "good enough" avoidance of omitted variable bias without encountering a curse of dimensionality.[7] In contrast, in the last example, it is frequently $\boldsymbol{\beta}$, or at least some components thereof, that is of interest.

In the first three examples, we can get a parametric convergence rate, as well as convergence of estimators of nonparametric objects at the rate corresponding to the argument's dimensionality, under reasonable though not innocuous conditions. We will examplarily discuss this using the partially linear model and generalizing to extremum (or m-) estimation with infinite dimensional nuisance parameters. The last case is less benign.

3.3.1 Robinson (1988) on Partially Linear Models

Write the Partially Linear Model as

$$\begin{aligned} Y_i &= \boldsymbol{X}_i'\boldsymbol{\beta} + m(\boldsymbol{Z}_i) + \epsilon_i \\ \mathbb{E}(\epsilon_i|\boldsymbol{X}_i, \boldsymbol{Z}_i) &= 0, \end{aligned}$$

noting that we do allow for heteroskedasticity, i.e. $\sigma^2(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}(\epsilon_i^2|\boldsymbol{X}_i, \boldsymbol{Z}_i = (\boldsymbol{x}, \boldsymbol{x}))$ is not restricted to be constant. The classic analysis of this model is due to Robinson (*Econometrica*, 1988). Identification is not completely trivial but we will take it for granted.

The basic ideas of estimation is to "concentrate out" m. Write

$$\begin{aligned} m_y(\boldsymbol{Z}_i) =: \mathbb{E}(Y_i|\boldsymbol{Z}_i) &= \mathbb{E}(\boldsymbol{X}_i'\boldsymbol{\beta}|\boldsymbol{Z}_i) + \mathbb{E}(m(\boldsymbol{Z}_i)|\boldsymbol{Z}_i) + \mathbb{E}(\epsilon_i|\boldsymbol{Z}_i) \\ &= \mathbb{E}(\boldsymbol{X}_i'\boldsymbol{\beta}|\boldsymbol{Z}_i) + m(\boldsymbol{Z}_i) \\ &=: \boldsymbol{m}_x(\boldsymbol{Z}_i)'\boldsymbol{\beta} + m(\boldsymbol{Z}_i) \end{aligned}$$

We therefore have

$$Y_i - m_y(\boldsymbol{Z}_i) = (\boldsymbol{X}_i - \boldsymbol{m}_x(\boldsymbol{Z}_i))'\boldsymbol{\beta} + \epsilon_i, \tag{3.1}$$

and if we knew $(m_y, \boldsymbol{m}_x)$, we could estimate $\boldsymbol{\beta}$ by OLS regression of $Y_i - m_y(\boldsymbol{Z}_i)$ on $(\boldsymbol{X}_i - \boldsymbol{m}_x(\boldsymbol{Z}_i))$. This estimator is infeasible, but it motivates a two-stage plug-in procedure: First

[7] A notable caveat is the linear model with few parameters of interest but many (relative to sample size) controls. If Machine Learning tools are used to estimate the controls, this becomes akin to a partially linear model with $\boldsymbol{\beta}$ the parameter of interest.

use the nonparametric method of choice to compute estimators $(\hat{m}_y, \hat{\boldsymbol{m}}_x)$, then compute

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))(\boldsymbol{X}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))'\right)^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))(Y_i - \hat{m}_y(\boldsymbol{Z}_i)).$$

An immediate problem is that this estimator need not even be consistent because $(\hat{m}_y, \hat{\boldsymbol{m}}_x)$ need not be consistent uniformly over the support of $\boldsymbol{Z}_i$, specifically if the density vanishes at some points. (And in practice, a close-enough-to-vanishing density would lead to bad finite sample behavior.) Robinson therefore suggests to trim the data and only use data points for which $\boldsymbol{Z}_i$ takes a value with high enough estimated density. This leads to estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} (X_i - \hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))(\boldsymbol{X}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))' \times \mathbf{1}\{\hat{f}_z(Z_i) \geq b_n\}\right)^{-1}$$
$$\sum_{i=1}^{n} (X_i - \hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))(Y_i - \hat{m}_y(\boldsymbol{Z}_i)) \times \mathbf{1}\{\hat{f}_z(Z_i) \geq b_n\},$$

where b_n is a tuning parameter that is presumed to vanish slowly. We will follow much of the literature and ignore b_n, though strictly speaking it should show up in rate results below.

Ideally, we want to claim that this estimator is "oracle efficient," that is, it asymptotically behaves like the infeasible one. (Of course, even if that holds, we should not assume that finite sample behavior is equally as good.) Thus, we would ideally get

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\overset{d}{\to} N(\mathbf{0}, \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-1}), \\
\boldsymbol{G} &= \mathbb{E}\left[(\boldsymbol{X}_i - \boldsymbol{m}_x(\boldsymbol{Z}_i))(\boldsymbol{X}_i - \boldsymbol{m}_x(\boldsymbol{Z}_i))'\right], \\
\boldsymbol{S} &= \mathbb{E}\left[(\boldsymbol{X}_i - \boldsymbol{m}_x(\boldsymbol{Z}_i))(\boldsymbol{X}_i - \boldsymbol{m}_x(\boldsymbol{Z}_i))'\sigma^2(\boldsymbol{X}_i, \boldsymbol{Z}_i)\right].
\end{aligned}$$

Strikingly, this holds true under restrictive but not "crazy" conditions. These obviously include what we would need to get OLS to work. The crucial addition is that the MSE of our nonparametric estimator has to vanish at a sufficient rate. Assuming nonnegative kernels, that condition is $\sqrt{n}(h^4 + n^{-1}h^{-d}) \to 0$. Assuming optimal bandwidth choice, this obtains if $\sqrt{n} \times n^{-4/(4+d)} \to 0$, i.e. if $d \leq 3$ because d is an integer. Note that the integer problem forces us to create some "slack" here that can be used to handle b_n. In short, this method has the desired asymptotics if the nonparametric component has up to three dimensions.[8]

In my personal view, the legacy of the paper lies primarily in the striking fact that inference on the parametric stuff is standard. This observation is prominent in the context of (causal) machine learning or "data science". However, in some applications, the quantity of true interest is m, and the covariates that enter parametrically are "controls." The idea is

[8] In principle, higher order kernels can be used to make these asymptotics "work" for arbitrarily high d. Note the scare quotes.

that linearly adjusting for their effect may be good enough. If this is the motivation, there will be a third estimation stage in which we nonparametrically regress on $\mathbf{z}_i$ after concentrating out $\mathbf{x}_i'$, i.e. nonparametric mean regression in the model

$$Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} =: \tilde{Y}_i = m(\boldsymbol{Z}_i) + \epsilon_i.$$

Here, we can ignore estimation error in $\boldsymbol{\beta}$ because the estimator converges at a faster than the nonparametric rate and therefore is "superconsistent" relative to the relevant rate of localization. Note that this final regression need not mimic the preliminary estimation of m_y in detail. Indeed, it frequently uses other values of the tuning parameters, notably if inference is desired or if eyeball criteria suggest specific bandwidths.

3.3.2 When can we estimate parametric components at parametric rate?

We next think more generally about when we can estimate parametric components of semiparametric models at a parametric rate. Indeed, in favorable cases, the asymptotics for the parametric components look like standard m-estimator asymptotics. However, "favorable" here refers to conditions that apply to some but not all interesting models and so truly must be checked on a case-by-case basis. The development follows Andrews' (*Econometrica*, 1994) work on "**MIN**imum estimation with **P**reliminary **I**nfinite Dimensional **N**uisance," but similar ideas appeared in numerous places at the time.

The idea is to develop a general theory of extremum estimation when the objective function contains an infinite dimensional nuisance parameter that must be pre-estimated. Thus, suppose that

$$\boldsymbol{\theta}_0 = \arg\min Q(\boldsymbol{\theta}, \tau_0)$$

is being estimated by

$$\hat{\boldsymbol{\theta}} = \arg\min Q_n(\boldsymbol{\theta}, \hat{\tau}),$$

where $\hat{\tau}$ is consistent for τ_0. (This is not sufficient but obviously necessary.) The question is if under reasonable conditions, we can ignore the difference between $\hat{\tau}$ and τ_0 and have standard extremum estimator asymptotics apply. Again, the answer will be a qualified "yes," i.e. such conditions are neither innocuous nor absurd and must be verified on a case-by-case basis.

To simplify expressions, we will assume m-estimation in the narrow sense, i.e. that Q_n is a sample average that approximates a corresponding expectation. This simply avoids some remainder terms in the below. It will be convenient to write

$$\frac{\partial Q_n(\boldsymbol{\theta}, \tau)}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\theta}, \tau),$$

i.e. use the shorthand $\boldsymbol{g}_i$ for the derivative of the objective function evaluated on one data point. (This is a r.v. due to a hidden argument $\boldsymbol{w}_i$, the data.) Of course, we assume $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0) = \mathbf{0}$ and also all the conditions needed for m-estimation. Our estimator is then (with probability approaching 1) characterized by

$$\mathbf{0} = \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\theta}}, \hat{\tau}) := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\theta}}, \hat{\tau}),$$

and we can replicate the standard algebra for m-estimators:

$$\begin{aligned} \mathbf{0} &= \sqrt{n}\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\theta}}, \hat{\tau}) \\ &= \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau}) + \frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) + o_P(1) \\ \Rightarrow \sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) &= -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\right)^{-1} \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau}) + o_P(1) \end{aligned}$$

As a reminder, in an "oracle" situation where τ_0 is known and used, the last line would read

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) = -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0)\right)^{-1} \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0) + o_P(1)$$

and we would recover standard m-estimator asymptotics. Since we need continuous differentiability of $\mathbb{E}\boldsymbol{g}_i$, consistency of $\hat{\tau}$, and existence of the inverse in the expression anyway, swapping in $\hat{\tau}$ for τ_0 in said inverse is not a big deal. However, we next have $\sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})$ where we would like $\sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0)$. This gap is much more difficult to handle and indeed cannot be closed without further assumptions.

The first of these assumptions is a statistical regularity condition called *stochastic equicontinuity* that is typically just assumed. To get an idea, define

$$\boldsymbol{\nu}_n(\tau) = \sqrt{n}\big(\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau)\big).$$

This is a stochastic process, i.e. a random variable that takes entire functions as realizations. Note that for any fixed value of τ, $\boldsymbol{\nu}_n$ is (under reasonable conditions) subject to a CLT. Also, it is intuitive that for τ and τ' very close to each other, $\boldsymbol{\nu}_n$ will be similarly distributed and maybe even very highly correlated. These last intuitions are formalized by assuming that

$$\hat{\tau} \xrightarrow{p} \tau_0 \Rightarrow \boldsymbol{\nu}_n(\hat{\tau}) - \boldsymbol{\nu}_n(\tau_0) \xrightarrow{p} \mathbf{0}.$$

Stochastic equicontiuity is actually a more technical condition implying this, but this will do for our ourposes. Verification of this assumption can be very tedious, yet at the same time it is not generally considered especially restrictive and can be verified in many applications. In practice, researchers who are not statisticians will likely cite such a verification or just make the assumption.

Imposing this assumption is very helpful because we assume that $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0) = \mathbf{0}$, so that (again under weak conditions) a CLT will give us

$$\boldsymbol{\nu}_n(\tau_0) = \sqrt{n}\big(\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)\big) \xrightarrow{d} N\left(0, \mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)')\right),$$

and this conclusion then extends to $\boldsymbol{\nu}_n(\hat{\tau})$. Plugging back into the m-estimator algebra, we have

$$\begin{aligned}
&\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) \\
=\ & -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\right)^{-1} \sqrt{n}\big(\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\underbrace{-\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) + \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau})}_{\text{add-and-subtract}}\big) + o_P(1) \\
=\ & -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\right)^{-1} \big(\boldsymbol{\nu}_n(\tau_0) + \sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau})\big) + o_P(1),
\end{aligned}$$

and we are good to go... if $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = o_P(1)$.

This is the second important condition, which is notably overlooked in Li and Racine's textbook.[9] It is not at all innocuous: While $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0) = \mathbf{0}$ will imply $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = o_P(1)$ under conditions we need anyway, the additional $\sqrt{n}$-factor makes the condition much more demanding. Indeed, this can be intuited as imposing a local orthogonality between the nonparametric and parametric part of the estimation problem. (If τ were a vector, it would require cross-derivatives at $(\boldsymbol{\theta}_0, \boldsymbol{\tau}_0)$ to be zero.) The condition may or may not hold in a given application and must be explicitly verified. Assuming it does hold, and under sufficient "m-estimator" regularity conditions so that we can claim $\big(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\big)^{-1} \xrightarrow{p} \big(\mathbb{E}\frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)\big)^{-1}$, we finally conclude that

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) = -\left(\mathbb{E}\frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)\right)^{-1} \boldsymbol{\nu}_n(\tau_0) + o_P(1).$$

This informally proves:

Theorem (Andrews, 1994)

Under "m-estimator" regularity conditions, stochastic equicontinuity of $\boldsymbol{\nu}_n$, and if also

[9] More precisely, they use a nonstandard definition of $\boldsymbol{\nu}_n$ in which true expectations are not subtracted. This hides the condition in the equicontinuity assumption on $\boldsymbol{\nu}_n$, but means that such equicontinuity is not implied by relevant "off-the-shelf" results. In contrast, Andrews (display 4.4) uses our definition of $\boldsymbol{\nu}_n$ and explicitly imposes the condition discussed here as Assumption N(c); observe simplification as per display 4.9.

$\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) \xrightarrow{p} \mathbf{0}$, then

$$\begin{aligned} \sqrt{n}\big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\big) &\xrightarrow{d} N(\mathbf{0},\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-1}) \\ \boldsymbol{G} &= \mathbb{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}'}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)\right) \\ \boldsymbol{S} &= \mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)'). \end{aligned}$$

3.3.3 Checking the MINPIN Assumptions in Applications

We first verify conditions under which the crucial orthogonality condition holds in partially linear regression. In that case, we have

$$\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) = \big(\boldsymbol{X}_i-\hat{\tau}_x(\boldsymbol{Z}_i)\big)\big(Y_i-\hat{\tau}_y(\boldsymbol{Z}_i)-(\boldsymbol{X}_i-\hat{\tau}_x(\boldsymbol{Z}_i))'\boldsymbol{\theta}_0\big).$$

Again, setting this equal to zero in expectation is an analog of the OLS moment condition $\mathbb{E}(\mathbf{x}_i(Y_i-\mathbf{x}_i'\boldsymbol{\beta}))=\mathbf{0}$. Now, recall from (3.1) that the true model can be written as

$$Y_i = m_y(\boldsymbol{Z}_i)+(\boldsymbol{X}_i-\boldsymbol{m}_x(\boldsymbol{Z}_i))'\boldsymbol{\theta}_0+\epsilon_i$$

with $\mathbb{E}(\epsilon_i|\boldsymbol{X}_i,\boldsymbol{Z}_i)=0$. Equating τ with $\boldsymbol{m}$ and plugging in, we get

$$\begin{aligned} \boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) &= \big(\boldsymbol{X}_i-\hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i)\big)\big(m_y(\boldsymbol{Z}_i)+(\boldsymbol{X}_i-\boldsymbol{m}_x(\boldsymbol{Z}_i))'\boldsymbol{\theta}_0+\epsilon_i-\hat{m}_y(\boldsymbol{Z}_i)-(\boldsymbol{X}_i-\hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))'\boldsymbol{\theta}_0\big) \\ &= \big(\boldsymbol{X}_i-\hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i)\big)\big(m_y(\boldsymbol{Z}_i)-\hat{m}_y(\boldsymbol{Z}_i)-(\boldsymbol{m}_x(\boldsymbol{Z}_i)-\hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))'\boldsymbol{\theta}_0+\epsilon_i\big) \end{aligned}$$

and therefore

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}|\boldsymbol{Z}_i) = \big(\boldsymbol{m}_x(\boldsymbol{Z}_i)-\hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i)\big)\big(m_y(\boldsymbol{Z}_i)-\hat{m}_y(\boldsymbol{Z}_i)-(\boldsymbol{m}_x(\boldsymbol{Z}_i)-\hat{\boldsymbol{m}}_x(\boldsymbol{Z}_i))'\boldsymbol{\theta}_0\big).$$

We can now use the Law of Iterated Expectations to write

$$\begin{aligned} & \sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) \\ = \; & \sqrt{n}\int(\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z}))\left(m_y(\boldsymbol{z})-\hat{m}_y(\boldsymbol{z})-(\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z}))'\boldsymbol{\theta}_0\right)f_z(\boldsymbol{z})d\boldsymbol{z} \\ \leq \; & \sqrt{n}\left(\sup_{\boldsymbol{z}}\|\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z})\|\times\sup_{\boldsymbol{z}}|m_y(\boldsymbol{z})-\hat{m}_y(\boldsymbol{z})|+\sup_{\boldsymbol{z}}\|\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z})\|^2\|\boldsymbol{\theta}_0\|\right). \end{aligned}$$

A sufficient and “tight” (we don’t generally expect the result otherwise) condition for this is that $\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z})=o_P(n^{-1/4})$. We therefore find that Robinson’s (1988) condition was tight. (We are loose about uniform convergence here, but the trimming step fixes that.)

In the case of generated regressors, we have $\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau)=\mathbb{E}\big(\boldsymbol{m}(\boldsymbol{X}_i)(Y_i-\boldsymbol{m}(\boldsymbol{X}_i)'\boldsymbol{\theta}_0)\big)$, and we will establish in a homework that the condition fails. Indeed, nonparametrically generated regressors are a difficult topic; see Mammen, Rothe, and Schienle (*Annals of Statistics*, 2012).

3.3.4 Summary

MINPIN theory is a reasonably general framework for thinking about semiparametric models. Again, it applies if the model in question can be thought of as m-estimation, except that the estimator of an infinite dimensional nuisance parameter enters the sample objective function. (Of course, as with partially linear models, the MINPIN estimation problem may be auxiliary to the estimation problem of true economic interest.) The theory analyzes whether the nonparametric pre-estimation step may be (first-order asymptotically) ignored. It has three core ingredients:

- We must have enough regularity so that, if the nuisance parameter τ were known, we would have a conventional m-estimation problem. Necessity of this is clear enough. Loosely speaking, the assumptions are some smoothness and uniform convergence restrictions and a CLT for the score at fixed τ, which is frequently readily available.

- The "error process" ν_n must be stochastically equicontinuous. This has been verified for important cases and is frequently just assumed.

- The "near orthogonality condition," i.e. $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = o_P(1)$, may or may not hold and must be verified. This is the hard part.

For additional intuition about what's going on here, note that without the orthogonality condition, and if τ were fixed at any particular value, we would still have consistency and even asymptotic normality of our estimator for the implied "pseudotrue" value of $\boldsymbol{\theta}$ which minimizes the population criterion function given that τ. However, if τ is being estimated, that pseudotrue value is itself a moving target. This can be ignored if it converges to our intended estimand at a fast enough rate, but this generally requires the condition (which is therefore tight).

Note also the following: If $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = O_P(1)$, our algebra suggests that we get parametric rate of convergence though with an erratic limiting distribution. One might feel this is good enough to claim subsequent ignorability of the parametric estimation step. This reasoning is dangerous because we were cavalier about the need for the nonparametric estimators to converge uniformly and also ignored b_n. The knife-edge manner in which $d = 4$ would imply that the MSE of $\hat{\boldsymbol{m}}$ is of order $O_P(n^{-1/2})$ is therefore spurious. With $d = 3$, we have enough slack in the rate to handle these issues.