Notes on Nonparametric Estimation/Inference for Econometrics II

Tim Armstrong

last updated: April 22, 2020

1 Introduction

- In nonparametric estimation, we seek to relax parametric functional form assumptions.
- Let $m(x)$ denote the conditional mean of y_i given x_i and $\sigma^2(x_i)$ the conditional variance. Then

$$y_i = m(x_i) + e_i, \quad E(e_i|x_i) = 0, \quad E(e_i^2|x_i) = \sigma^2(x_i).$$

- In linear regression, we had $m(x) = x'\beta$. We also covered nonlinear least squares estimation for models of the form $m(x) = m(x, \beta)$ where $\beta \in \mathbb{R}^k$. In nonparametric theory, we allow for approximation error in $m(x)$ and use tools from approximation theory (e.g. Taylor approximations).
- These notes cover estimation of $m(x_0)$ with x_0 given, using derivative smoothness conditions based on Taylor approximations.
- First, we cover finite sample theory, following Sacks and Ylvisaker (1978). We then consider some asymptotic results, following Fan (1993), Fan and Gijbels (1996) and Cheng et al. (1997). This mirrors the approach we took to linear regression: finite sample results such as Gauss-Markov, then asymptotic approximations. This differs from some texts, which do not cover finite sample theory.

2 Finite Sample Theory

2.1 Approximately Linear Models

- One approach to nonparametric estimation is the approximately linear model of Sacks and Ylvisaker (1978), in which we assume

$$m(x) = \psi(x)'\beta + r(x) \text{ where } |r(x)| \leq M(x)$$

  where

  - $\psi(x) = (\psi_1(x), \ldots, \psi_p(x))'$ is a vector of approximating functions given by the researcher (e.g. $(1, x, x^2, \ldots, x^5)'$).
  - $r(x)$ is approximation error
  - $M(x)$ is a (for now, known) bound on the approximation error.

- When $M(x) = 0$, this reduces to linear regression (on the $\psi_j(x_i)$'s, rather than x_i itself).

2.2 Estimation of the Conditional Mean at a Given Point

- Focus on case where x_i's are scalar-valued.

- If we wish to estimate $m(x_0)$, $m'(x_0)$, $m''(x_0)$, etc., a convenient setup is to base this on the Taylor approximation at x_0:

$$m(x) = \sum_{j=1}^{p} (x - x_0)^{j-1}\beta_j + r(x) \text{ where } |r(x)| \leq \frac{C}{p!}|x - x_0|^p$$

  I.e. we take $M(x) = C|x - x_0|^p/p!$ and $\psi(x) = (1, x - x_0, (x - x_0)^2, \ldots, (x - x_0)^{p-1})'$. This holds if the pth derivative of $m(x)$ is bounded by C.

- Here $\beta_j = m^{(j-1)}(x_0)/(j-1)!$. Thus, to estimate the jth derivative, we need an estimate of β_{j+1} ($\beta_1 = m(x_0)$ is the intercept).

- For simplicity, let us focus on estimating $\beta_1 = m(x_0)$.

- Consider a linear estimator:

$$\hat{m}(x_0) = \sum_{i=1}^{n} w_i y_i.$$

  - $w_i = w_i(X)$ can depend on $X = (x_1, \ldots, x_n)'$ but not the y's.
  - For linear regression, the Gauss-Markov theorem shows that OLS (or, under heteroskedasticity, WLS) is minimum variance among unbiased (conditional on the x's) linear estimators.
  - Because of the "specification error" $r(x)$, unbiasedness is too much to ask. However, we can minimize the variance subject to a bound on the bias. By varying the bound on the bias, we trace out a bias-variance tradeoff.

2.3 Variance and Worst-Case Bias

- thm. For a linear estimator $\hat{m}(x_0) = \sum_{i=1}^{n} w_i y_i$, the bias (conditional on the x's) can be arbitrarily large unless

$$\sum_{i=1}^{n} w_i = 1 \text{ and } \sum_{i=1}^{n} w_i \cdot (x_i - x_0)^{j-1} = 0,\ j = 2, \ldots, p \tag{1}$$

  If (1) holds, the bias varies from $-\overline{\text{bias}}(\hat{m}(x_0))$ to $\overline{\text{bias}}(\hat{m}(x_0))$ where

$$\overline{\text{bias}}(\hat{m}(x_0)) = \sum_{i=1}^{n} |w_i| C |x_i - x_0|^p / p!$$

  pf.: The bias is given by

$$\begin{aligned} E(\hat{m}(x_0) - m(x_0)|X) &= \sum_{i=1}^{n} w_i m(x_i) - m(x_0) \\ &= \beta_1 \sum_{i=1}^{n} w_i + \sum_{j=2}^{p} \beta_j \sum_{i=1}^{n} w_i \cdot (x_i - x_0)^{j-1} + \sum_{i=1}^{n} w_i r(x_i) - \beta_1. \end{aligned}$$

  It can be seen by inspection that, if (1) does not hold, this can be made arbitrarily

large by making the β_j's large. If (1) holds, then the bias reduces to

$$\sum_{i=1}^{n} w_i r(x_i),$$

which is maximized by taking $r(x_i) = \text{sign}(w_i) \cdot C|x_i - x_0|^p/p!$. This gives the maximum bias claimed in the theorem, and the minimum bias follows by symmetry.

- The (conditional) variance is given by

$$\text{var}(\hat{m}(x_0)|X) = \sum_{i=1}^{n} w_i^2 \sigma^2(x_i).$$

- Thus, the minimum variance linear estimator with bias bounded by B is characterized by the problem

$$\min_{w_1,\ldots,w_n} \sum_{i=1}^{n} w_i^2 \sigma^2(x_i) \text{ s.t. } \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i \cdot (x_i - x_0)^{j-1} = 0, \; j = 2, \ldots, p,$$
$$\sum_{i=1}^{n} |w_i| \cdot C|x_i - x_0|^p/p! \leq B.$$

- Solution characterized by Sacks and Ylvisaker (1978) (elementary argument using Lagrangian).

- How we trade off bias and variance depends on our objectives. One possibility is to minimize the worst-case mean squared error (MSE):

$$E[(\hat{m}(x_0) - m(x_0))^2|X] = [E(\hat{m}(x_0)|X) - m(x_0)]^2 + var(\hat{m}(x_0)|X)$$
$$\leq \overline{\text{bias}}(\hat{m}(x_0))^2 + \text{var}(\hat{m}(x_0)|X).$$

2.4 Local Polynomial Estimators

- The resulting optimal weighting can be unintuitive and cumbersome to compute. A popular alternative is a local polynomial estimator.

- The local polynomial estimator of $m(x_0)$ of order $p-1$ with kernel $k(\cdot)$ and bandwidth h_n is the estimate of the intercept in the WLS regression of y_i on 1, $x_i - x_0$, $(x_i - x_0)^2$,. . . ,

$(x_i - x_0)^{p-1}$ with weights $k((x_i - x_0)/h_n)$: letting

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^n (y_i - q(x_i - x_0)'\beta)^2 k((x_i - x_0)/h_n) = \text{argmin}_\beta \sum_{i=1}^n (y_i - q_i'\beta)^2 k_i$$

where

$$q(t) = (1, t, t^2, \ldots, t^{p-1})', \quad q_i = q(x_i - x_0), \quad k_i = k((x_i - x_0)/h_n),$$

the local polynomial estimate of $m(x_0)$ is $\hat{m}(x_0) = \hat{\beta}_1$.

- Popular choices of kernel: uniform ($k(u) = I(|u| \leq 1/2)$), triangular ($k(u) = \max\{1 - |u|, 0\}$)
- The bandwidth h can be used to trade off bias and variance: increasing h means using observations further away from x_0, which increases bias but reduces variance.

- By the usual least squares algebra (rearrange FOCs), it follows that

$$\hat{m}(x_0) = e_1' \left(\sum_{i=1}^n q_i q_i' k_i \right)^{-1} \sum_{i=1}^n q_i k_i y_i$$

where $e_1 = (1, 0, \ldots, 0)'$. Thus, it takes the form $\hat{m}(x_0) = \sum_{i=1}^n w_i y_i$ with

$$w_i = e_1' \left(\sum_{j=1}^n q_j q_j' k_j \right)^{-1} q_i k_i. \tag{2}$$

- Lemma: The local polynomial estimator satisfies (1).

  pf.: Note that (1) can be written as

$$\sum_{i=1}^n w_i q_i' = \sum_{i=1}^n w_i q(x_i - x_0)' = e_1'.$$

  For the weights w_i given above for the local polynomial estimator, we have

$$\sum_{i=1}^n w_i q_i' = e_1' \left(\sum_{j=1}^n q_j q_j' k_j \right)^{-1} \sum_{i=1}^n k_i q_i q_i' = e_1',$$

which gives the result.

- Thus, the worst case bias of the local polynomial estimator is given by

$$\begin{aligned}\overline{\text{bias}}(\hat{m}(x_0)) &= \sum_{i=1}^{n} |w_i| C |x_i - x_0|^p / p! \\ &= \sum_{i=1}^{n} \left| e_1' \left(\sum_{j=1}^{n} q_j q_j' k_j \right)^{-1} q_i k_i \right| C \frac{|x_i - x_0|^p}{p!}\end{aligned}$$

- Popular special cases include $p = 1$ (Nadaraya-Watson) and $p = 2$ (local linear).

3 Confidence Intervals

- We will go over a simple form of confidence intereval called a fixed-length confidence interval (FLCI), which goes back to Knafl et al. (1982) and has been shown by Donoho (1994) and Armstrong and Kolesár (2018) to have certain optimality properties.

- Local polynomial estimate takes the form

$$\hat{m}(x_0) = \sum_{i=1}^{n} w_i y_i \text{ where } w_i = e_1' \left(\sum_{j=1}^{n} q_j q_j' k_j \right)^{-1} q_i k_i.$$

as defined in (2).

- Conditional on X, worst-case bias is $\overline{\text{bias}}(\hat{m}(x_0)) = \sum_{i=1}^{n} |w_i| C |x_i - x_0|^p / p!$. Variance is $\text{se}(\hat{m}(x_0))^2 = \sum_{i=1}^{n} w_i^2 \sigma^2(x_i)$.

- Let $\widehat{\text{se}}(\hat{m}(x_0))$ be an estimate of $\text{se}(\hat{m}(x_0))$. For example, we can take

$$\widehat{\text{se}}(\hat{m}(x_0))^2 = \sum_{i=1}^{n} w_i^2 \hat{u}_i^2$$

where $\hat{u}_1, \ldots, \hat{u}_n$ are the residuals from the local polynomial regression.

- To form a CI, note that

$$\frac{\hat{m}(x_0) - m(x_0)}{\text{se}(\hat{m}(x_0))} = \frac{\hat{m}(x_0) - E[\hat{m}(x_0)|X] + \text{bias}(\hat{m}(x_0))}{\text{se}(\hat{m}(x_0))} \overset{d}{\approx} Z + \frac{\text{bias}(\hat{m}(x_0))}{\text{se}(\hat{m}(x_0))}$$

where $Z \sim N(0,1)$ and $\text{bias}(\hat{m}(x_0)) = E[\hat{m}(x_0) - m(x_0)|X]$.

- Let

$$\text{cv}_{1-\alpha}(t) = \ 1-\alpha \text{ quantile of } |Z+t| \text{ where } Z \sim N(0,1).$$

We would like to use $\text{cv}_{1-\alpha}(\text{bias}(\hat{m}(x_0))/\text{se}(\hat{m}(x_0)))$ as the critical value, but cannot since $\text{bias}(\hat{m}(x_0))$ is unknown. However, we can get an upper bound on this critical value by using $\text{cv}_{1-\alpha}(\overline{\text{bias}}(\hat{m}(x_0))/\text{se}(\hat{m}(x_0)))$. Plugging in the variance estimate, this gives the approximate CI

$$\hat{m}(x_0) \pm \text{cv}_{1-\alpha}(\overline{\text{bias}}(\hat{m}(x_0))/\widehat{\text{se}}(\hat{m}(x_0))) \cdot \widehat{\text{se}}(\hat{m}(x_0)).$$

- To compute this CI, we need to know C (bound on pth derivative) in order to compute $\overline{\text{bias}}$. In practice, it is common to ignore bias (i.e. act as if $\overline{\text{bias}}$ is equal to 0) and compute the interval as $\hat{m}(x_0) \pm \text{cv}_{1-\alpha}(0) \cdot \widehat{\text{se}}(\hat{m}(x_0) = \hat{m}(x_0) \pm z_{1-\alpha/2} \cdot \widehat{\text{se}}(\hat{m}(x_0)$. This is formally justified by making an "asymptotic promise" to take $h_n \to 0$ quickly enough so that $\overline{\text{bias}}(\hat{m}(x_0))/\text{se}(\hat{m}(x_0)) \to 0$, which is called undersmoothing. We will discuss this in Section 5.2. Unfortunately, to know whether h is small enough that we are "undersmoothing," we need to know whether $\overline{\text{bias}}(\hat{m}(x_0))/\text{se}(\hat{m}(x_0))$ is small for the sample size at hand, which gets us back to having to know C! So, undersmoothing isn't really a solution to not knowing C.
  - The interval $\hat{m}(x_0) \pm \text{cv}_{1-\alpha}(\overline{\text{bias}}(\hat{m}(x_0))/\widehat{\text{se}}(\hat{m}(x_0))) \cdot \widehat{\text{se}}(\hat{m}(x_0))$ is sometimes called "bias-aware" to distinguish it from the interval $\hat{m}(x_0) \pm z_{1-\alpha/2} \cdot \widehat{\text{se}}(\hat{m}(x_0))$. Bias-aware CIs are particularly attractive in settings where the "asymptotic promise" of undersmoothing seems unappealing, such as when x_i has a discrete distribution (we will see in the next section that the asymptotics used for undersmoothing require a continuously distributed x_i). See Armstrong and Kolesár (2018), Kolesár and Rothe (2018) and Imbens and Wager (2019) and Noack and Rothe (2019) for recent applications along these lines.
- It can be shown formally that confidence intervals must depend a priori knowledge of C (and p) in this setting (this insight goes back to Low (1997); see Armstrong and Kolesár (2018) for a recent discussion and results). One approach is to place some auxiliary assumptions, such as assuming that C can be estimated using a global polynomial (see

Armstrong and Kolesár, 2020, for discussion).

4 Asymptotics for Nadaraya-Watson and Local Linear

- To gain additional insight, let us see what happens as $n \to \infty$.

4.1 Nadaraya-Watson Estimator ($p = 1$)

- When $p = 1$, we get the Nadaraya-Watson estimator

$$\hat{m}(x_0) = \frac{\sum_{i=1}^n k((x_i - x_0)/h) y_i}{\sum_{i=1}^n k((x_i - x_0)/h)}.$$

The associated weights w_i are

$$w_i = \frac{k((x_i - x_0)/h)}{\sum_{j=1}^n k((x_j - x_0)/h)}.$$

The associated smoothness class bounds the approximation error of a 0th order (constant) Taylor approximation at 0: $|m(x) - m(x_0)| \leq C|x - x_0|$. This is called a Lipschitz condition.

- The worst-case bias over functions $m(\cdot)$ satisfying this condition is

$$\begin{aligned}\overline{\text{bias}}(\hat{m}(x_0)) &= \sum_{i=1}^n |w_i| \cdot C|x_i - x_0| = C\frac{\sum_{i=1}^n |k((x_i - x_0)/h)| \cdot |x_i - x_0|}{\sum_{i=1}^n k((x_i - x_0)/h)} \\ &= hC\frac{\frac{1}{nh}\sum_{i=1}^n |k((x_i - x_0)/h)| \cdot |x_i - x_0|/h}{\frac{1}{nh}\sum_{i=1}^n k((x_i - x_0)/h)}.\end{aligned}$$

Suppose that x_i has a continuous density $f(\cdot)$ and $k(\cdot)$ is bounded with finite support (i.e. $k(t) = 0$ for $|t|$ large enough). Then, as $h = h_n \to 0$,

$$\begin{aligned}&E\left[\frac{1}{h}|k((x_i - x_0)/h)| \cdot |x_i - x_0|/h\right] = \frac{1}{h}\int |k((x - x_0)/h)(x - x_0)/h| f(x)\, dx \\ &= \int |k(u)u| f(x_0 + uh)\, du \to f(x_0)\int |k(u)u|\, du\end{aligned}$$

where we use the substitution $u = (x - x_0)/h$. Similarly, $E\left[\frac{1}{h}\sum_{i=1}^n k((x_i - x_0)/h)\right] \to$

$f(x_0)\int k(u)\,du$. Thus, assuming regularity conditions hold for a SLLN for inid sequences, we will have

$$\begin{aligned}\frac{1}{h}\cdot\overline{\text{bias}}(\hat{m}(x_0)) &= C\frac{\frac{1}{nh}\sum_{i=1}^n |k((x_i-x_0)/h)|\cdot|x_i-x_0|/h}{\frac{1}{nh}\sum_{i=1}^n k((x_i-x_0)/h)}\\ &\to C\frac{\int |k(u)u|\,du}{\int k(u)\,du} \quad \text{a.s.}\end{aligned}$$

- The variance (conditional on the x_i's) is

$$\sum_{i=1}^n w_i^2\sigma^2(x_i) = \frac{\sum_{i=1}^n k((x_i-x_0)/h)^2\sigma^2(x_i)}{\left(\sum_{i=1}^n k((x_i-x_0)/h)\right)^2} = \frac{1}{nh}\frac{\frac{1}{nh}\sum_{i=1}^n k((x_i-x_0)/h)^2\sigma^2(x_i)}{\left(\frac{1}{nh}\sum_{i=1}^n k((x_i-x_0)/h)\right)^2}$$

Assuming $\sigma^2(\cdot)$ is continuous at x_0, we will have

$$\begin{aligned}&E\left[\frac{1}{h}k((x_i-x_0)/h)^2\sigma^2(x_i)\right] = \frac{1}{h}\int k((x-x_0)/h)^2\sigma^2(x)f(x)\,dx\\ &= \int k(u)^2\sigma^2(x_0+uh)f(x_0+uh)\,du \to \sigma^2(x_0)f(x_0)\int k(u)^2\,du.\end{aligned}$$

From this and similar calculations for the denominator, it follows that, assuming regularity conditions hold for a SLLN,

$$\begin{aligned}nh\cdot\text{var}(\hat{m}(x_0)|X) &= \frac{\frac{1}{nh}\sum_{i=1}^n k((x_i-x_0)/h)^2\sigma^2(x_i)}{\left(\frac{1}{nh}\sum_{i=1}^n k((x_i-x_0)/h)\right)^2}\\ &\to \frac{\sigma^2(x_0)\int k(u)^2\,du}{f(x_0)\left(\int k(u)^2\,du\right)^2} \quad \text{a.s.}\end{aligned}$$

- Thus, the bias is $\mathcal{O}(h)$ and standard deviation is $\mathcal{O}(1/\sqrt{nh})$. If we set h so that these two terms are of the same order of magnitude, this gives us

$$\begin{aligned}&\mathcal{O}(h) = \mathcal{O}(1/\sqrt{nh})\\ &\implies \mathcal{O}(h^{3/2}) = \mathcal{O}(n^{-1/2})\\ &\implies h = \mathcal{O}(n^{-1/3}).\end{aligned}$$

With this choice, bias and standard deviation are both of order $n^{-1/3}$. Thus, we get a $n^{-1/3}$ rate of convergence (slower than the usual $n^{-1/2}$).

- Optimal h depends on C and on how we trade off bias and variance.

4.2 Local Linear Estimator

- The local linear estimator is $\hat{\beta}_1$ where

$$(\beta_1, \beta_2) = \text{argmin}_\beta \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2(x_i - x_0))^2 k((x_i - x_0)/h_n)$$
$$= \text{argmin}_\beta \sum_{i=1}^{n} (y_i - \beta_1 - (h\beta_2)(x_i - x_0)/h)^2 k((x_i - x_0)/h_n).$$

Since minimizing over (β_1, β_2) gives the same β_1 as minimizing over $(\beta_1, h\beta_2)$, we can use the least squares solution to the latter problem:

$$\hat{m}(x_0) = e_1' \hat{Q}_{qq}^{-1} \hat{Q}_{qy}$$

where

$$\hat{Q}_{qq} = \frac{1}{nh} \sum_{i=1}^{n} k((x_i - x_0)/h) \begin{pmatrix} 1 & (x_i - x_0)/h \\ (x_i - x_0)/h & [(x_i - x_0)/h]^2 \end{pmatrix}$$

and

$$\hat{Q}_{qy} = \frac{1}{nh} \sum_{i=1}^{n} k((x_i - x_0)/h) y_i \begin{pmatrix} 1 \\ (x_i - x_0)/h \end{pmatrix}.$$

- This gives the weights w_i as

$$w_i = \frac{1}{nh} e_1' \hat{Q}_{qq}^{-1} \begin{pmatrix} 1 \\ (x_i - x_0)/h \end{pmatrix} k((x_i - x_0)/h) = \frac{1}{nh} k_n^*((x_i - x_0)/h)$$

where

$$k_n^*(u) = e_1' \hat{Q}_{qq}^{-1} \begin{pmatrix} 1 \\ u \end{pmatrix} k(u)$$

is called the equivalent kernel.

- The associated smoothness class bounds the approximation error of a 1st order (linear) Taylor approximation at 0: $|m(x) - m(x_0) - m'(x_0)(x - x_0)| \leq C(x - x_0)^2/2$.

- The worst-case conditional bias over this class is

$$\overline{\text{bias}}(\hat{m}(x_0)) = (C/2)\sum_{i=1}^{n} |w_i|(x_i - x_0)^2 = (C/2)\sum_{i=1}^{n} \left|\frac{1}{nh}k_n^*((x_i - x_0)/h)\right| (x_i - x_0)^2$$
$$= h^2(C/2) \cdot \frac{1}{nh}\sum_{i=1}^{n} |k_n^*((x_i - x_0)/h)| \, [(x_i - x_0)/h]^2 \equiv h^2 b_n$$

- The conditional variance is

$$\text{var}(\hat{m}(x_0)|X) = \sum_{i=1}^{n} w_i^2 \sigma^2(x_i) = \sum_{i=1}^{n} \left[\frac{1}{nh}k_n^*((x_i - x_0)/h)\right]^2 \sigma^2(x_i)$$
$$= \frac{1}{nh} \cdot \left[\frac{1}{nh}\sum_{i=1}^{n} k_n^*((x_i - x_0)/h)^2 \sigma^2(x_i)\right] \equiv \frac{1}{nh} v_n.$$

- Under regularity conditions (which include x_0 being on the interior of the support of x_i), we will have, assuming $k(\cdot)$ is symmetric around zero,

$$\hat{Q}_{qq} \xrightarrow{a.s.} Q_{qq} \equiv f(x_0) \int k(u) \begin{pmatrix} 1 & u \\ u & u^2 \end{pmatrix} du = f(x_0) \begin{pmatrix} \int k(u)\, du & 0 \\ 0 & \int k(u) u^2 \, du \end{pmatrix},$$

  so that

$$k_n^*(u) \xrightarrow{a.s.} e_1' Q_{qq}^{-1} \begin{pmatrix} 1 \\ u \end{pmatrix} k(u) = \frac{k(u)}{f(x_0) \int k(u)\, du}$$

  and

$$b_n \xrightarrow{a.s.} (C/2)\frac{\int |k(u)| u^2 \, du}{\int k(u)\, du} \equiv b_\infty, \quad v_n \xrightarrow{a.s.} \frac{\sigma^2(x_0) \int k(u)^2 \, du}{f(x_0) \left(\int k(u)\, du\right)^2} \equiv v_\infty.$$

- Worst-case bias is of order h^2 and standard deviation is of order $(nh)^{-1/2}$. To set them equal, h must decrease like $n^{-1/5}$. This gives a $n^{2/5}$ rate of convergence.

5 Other Issues

5.1 Optimal Bandwidth

- Optimal bandwidth for CIs of the form given in Section 3 minimizes CI length. For local linear estimators ($p = 2$):

$$\begin{aligned} h_{CI}^* &= \operatorname{argmin}_h 2 \cdot \operatorname{se}(\hat{m}(x_0)) \cdot \operatorname{cv}_{1-\alpha}\left(\frac{\overline{\operatorname{bias}}(\hat{m}(x_0))}{\operatorname{se}(\hat{m}(x_0))}\right) \\ &\approx \operatorname{argmin}_h 2 \cdot \frac{1}{\sqrt{nh}}\sqrt{v_\infty} \cdot \operatorname{cv}_{1-\alpha}\left(\sqrt{nh^5}b_\infty/\sqrt{v_\infty}\right). \end{aligned}$$

- For estimation, MSE criterion is often used. Optimal bandwidth for local linear estimators is then

$$\begin{aligned} h_{MSE}^* &= \operatorname{argmin}_h \overline{\operatorname{bias}}(\hat{m}(x_0))^2 + \operatorname{se}(\hat{m}(x_0))^2 \\ &\approx \operatorname{argmin}_h h^4 \cdot b_\infty + nh \cdot v_\infty \end{aligned}$$

- Both h_{CI}^* and h_{MSE}^* set h to be of order $n^{-1/5}$, so that bias and standard deviation are balanced. In fact, they turn out to be close to each other asymptotically in this case: h_{CI}^*/h_{MSE}^* converges to a constant that is close to 1 (see Armstrong and Kolesár, 2020).

5.2 Undersmoothing

- Note that $\lim_{t\to 0} \operatorname{cv}_{1-\alpha}(t) = \operatorname{cv}_{1-\alpha}(0) = z_{1-\alpha/2}$. Thus, if $\frac{\overline{\operatorname{bias}}(\hat{m}(x_0))}{\operatorname{se}(\hat{m}(x_0))} \to 0$, the CIs constructed in Section 3 will be asymptotically equivalent to using $z_{1-\alpha/2}$. This is called undersmoothing.

- For local linear, $\frac{\overline{\operatorname{bias}}(\hat{m}(x_0))}{\operatorname{se}(\hat{m}(x_0))} = \mathcal{O}(h^2 \cdot \sqrt{nh}) = \mathcal{O}(\sqrt{nh^5})$, so undersmoothing corresponds to $nh^5 \to 0$.

- Note that getting the optimal rate of convergence required h to decrease like $n^{-1/5}$. Undersmoothing means $h \to 0$ faster than optimal rate, leading to CI shrinking at slower than optimal rate.

- Undersmoothing is popular in practice, since we don't need to know C (the bound on $m''(x)$) to form the critical value. However, it has some disadvantages:
  - It is not clear how small h has to be in practice to say that we are "undersmoothing:" for any finite n, how can we tell if $nh^{1/5} \to 0$ or not? In practice, we need $\frac{\overline{\text{bias}}(\hat{m}(x_0))}{\text{se}(\hat{m}(x_0))}$ to be small, but this gets us back to having to know C.
  - It is suboptimal: the optimal bandwidth for constructing CIs of this form (the one that minimizes CI length) is of order $n^{-1/5}$.
- In practice, it is more common to form CIs based on $z_{1-\alpha/2}$ than $\text{cv}_{1-\alpha}(\overline{\text{bias}}/\text{se})$. This leads to some undercoverage. However, if h is chosen optimally for MSE, it turns out that undercoverage is not that bad (see Armstrong and Kolesár, 2020).

5.3 Choice of C and p

- In practice, the optimal h and order of the polynomial requires knowledge of C and p. Estimators that are "close to" optimal for any C and p without knowledge of these constants are called "adaptive."
- Adaptive estimators can be very different depending on the criterion one uses when defining "optimal."
  - For MSE of $\hat{m}(x_0)$ for a particular x_0, one can use Lepski's method. See Sun (2005) for an application to regression discontinuity.
  - For the integrated mean square error (IMSE) criterion:

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

    one can choose h and the order of the polynomial adaptively using cross validation (see Section 11.6 in Hansen's text).
  - For CI construction, adaptation is severely limited for smoothness assumptions like those considered here (Low, 1997). To get around this, we must make stronger assumptions.
    - Shape restrictions: Cai and Low (2004), Armstrong (2015). Adaptivity is limited by the nature of the shape restrictions, however.

* "Self-similarity" restrictions: Giné and Nickl (2010), Chernozhukov et al. (2014).

5.4 Pointwise-in-m Asymptotics

- We took the approach of (1) assuming a bound on $m^{(p)}(x)$ (2) deriving finite sample results (variance and bounds on bias) (3) deriving asymptotic approximations to to the variance and bound on bias derived in (2). We might call this "uniform-in-$m(\cdot)$" asymptotics, since the asymptotic bound on bias holds uniformly over $m(\cdot)$ satisfying our conditions.

- Another approach is to fix $m(\cdot)$ and derive the limit of the bias under some smoothness assumptions. This usually leads to an expression similar to b_∞, but with C replaced by $m^{(p)}(x_0)$.

- We might call this a "pointwise-in-$m(\cdot)$" approach: we fix $m(\cdot)$ when doing asymptotics, rather than getting approximations that work uniformly over some class of functions $m(\cdot)$.

- Unfortunately, this can lead to estimators that perform poorly in practice because they use asymptotic bounds on bias that don't work well in finite samples.

- For example, it is often proposed to estimate the leading bias term using an estimate of $m^{(p)}(x_0)$ and use this, e.g., to estimate the "optimal" bandwidth (see, e.g., Imbens and Kalyanaraman, 2012). However, as our analysis shows, the optimal bandwidth depends on a bound for $m^{(p)}(x)$ over <u>all</u> x (or, at least, all x in some neighborhood of x_0), not on $m^{(p)}(x_0)$. For example, if $m(x) = (x - x_0)^3$, "pointwise-in-$m(\cdot)$" asymptotics suggests that bias of local linear is 0 (since $m''(x_0) = 0$ in this case). This is clearly not true in finite samples!

- Furthermore, we need $p + 1$ derivatives to estimate $m^{(p)}(x_0)$, so if we really believed this assumption, we would use a higher order local polynomial estimator in the first place.

- Overall, "pointwise-in-$m(\cdot)$" asymptotics can be nice since they give another way of thinking about the problem and often lead to similar results. However, when they lead us to estimators that don't work well under "uniform-in-$m(\cdot)$" asymptotics, we should be suspicious. See Chapter 1.2.4 in Tsybakov (2009) for more on these issues.

References

Armstrong, T. (2015): "Adaptive testing on a regression function at a point," *The Annals of Statistics*, 43, 2086–2101.

Armstrong, T. B. and M. Kolesár (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86, 655–683.

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 11, 1–39, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1199.

Cai, T. T. and M. G. Low (2004): "An Adaptation Theory for Nonparametric Confidence Intervals," *The Annals of Statistics*, 32, 1805–1840.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On automatic boundary corrections," *The Annals of Statistics*, 25, 1691–1708.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Anti-concentration and honest, adaptive confidence bands," *The Annals of Statistics*, 42, 1787–1818.

Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22, 238–270.

Fan, J. (1993): "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, CRC Press.

Giné, E. and R. Nickl (2010): "Confidence bands in density estimation," *The Annals of Statistics*, 38, 1122–1170.

Imbens, G. and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79, 933–959.

Imbens, G. and S. Wager (2019): "Optimized Regression Discontinuity Designs," *The Review of Economics and Statistics*, 101, 264–278, publisher: MIT Press.

Knafl, G., J. Sacks, and D. Ylvisaker (1982): "Model robust confidence intervals," *Journal of Statistical Planning and Inference*, 6, 319–334.

Kolesár, M. and C. Rothe (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 108, 2277–2304.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25, 2547–2554.

Noack, C. and C. Rothe (2019): "Bias-Aware Inference in Fuzzy Regression Discontinuity Designs," *arXiv:1906.04631 [econ, stat]*.

Sacks, J. and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *The Annals of Statistics*, 6, 1122–1137.

Sun, Y. (2005): "Adaptive Estimation of the Regression Discontinuity Model," SSRN Scholarly Paper ID 739151, Social Science Research Network, Rochester, NY.

Tsybakov, A. B. (2009): *Introduction to Nonparametric Estimation*, New York: Springer.