Nonparametrics and Local Methods: Advanced Topics

C.Conlon

February 27, 2021

Applied Econometrics

Local Bandwidths

If you only care about $f(y)$ at some given point, then

$$A = f''(y)^2 \left(\int u^2 K \right)^2 /4 \text{ and } B = f(y) \int K^2.$$

So in a low-density region, worry about variance and take h larger. In a curvy region, worry about bias and take h small.

Higher-Order Kernels

- K of order r iff $\int x^j K(x)dx = 0$ for $j < r$ and $\int x^r K(x)dx \neq 0$. Try $r > 2$?
- The beauty of it: bias in h^r if f is at least $C^r \ldots$ so AMISE can be reduced to $n^{-r/(2r+1)}$, almost $\sqrt{n}$-consistent if r is large.
- But gives wiggly (and sometimes negative) estimates $\rightarrow$ leave them to theorists.

Back to the CDF

Since now we have estimated the density with

$$\hat{f}_n(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right),$$

a natural idea is to integrate; let $\mathcal{K}(y) = \int_{-\infty}^{y} K(t)dt$, try

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}\left(\frac{y - y_i}{h}\right)$$

as a reasonable estimator of the cdf in y. Very reasonable indeed:

- when $n \longrightarrow \infty$ and h goes to zero (at rate $n^{-1/3} \ldots$) it is consistent at rate $\sqrt{n}$
- it is nicely smooth and accords well with the density estimator
- $\ldots$ it is a much better choice than the empirical cdf.

What if y is of dimension $p_y > 1$?

"Easy": use p_y-dimensional K (often a p_y-product of 1-dim kernels) and bandwidth h, and do

$$\hat{f}_n(y) = \frac{1}{nh_y^p} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right).$$

- **1st minor pitfall:** the various dimensions may have very different variances, so use $(h_1, \ldots, h_{p_y})$.
- **2nd minor pitfall:** they may be strongly correlated; then sphericize first.
- **Major problem:** next slide...

The Curse of Dimensionality

- Computational cost increases exponentially.
- *Much worse:* to achieve precision ϵ in dimension p_y, the number of observations you need increases as
$$n \simeq \epsilon^{-(2+p_y/2)}.$$
- The *empty space* phenomenon: if $(y_1, \ldots, y_{p_y})$ all are iid uniform on $[-1, 1]$, then only $n/(10^{p_y})$ observations on average have all components in $[-0.1, 0.1]$. Bias still in h^2, but variance in $1/nh^{p_y}$ now.

Silverman's Table

Silverman (1986 book) provides a table illustrating the difficulty of kernel estimation in high dimensions. To estimate the density at 0 of a $N(0,1)$ with a given accuracy, he reports:

Dimensionality Required	Sample Size
1	4
2	19
5	786
7	10,700
10	842,000

Not to be taken lightly... in any case convergence with the optimal bandwidth is in $n^{-2/(4+p_y)}$ now—and Silverman's rule of thumb for choosing h_n^* must be adapted too.

Usually we care about conditional densities

That is: we have covariates x, we want the density $f(y|x)$. Again, "easy":

1. get a kernel estimator of the joint density $f(y, x)$;
2. and one of of the marginal density $f(x)$;
3. then define

$$\hat{f}_n(y|x) = \frac{\hat{\hat{f}}_n(y,x)}{\hat{\hat{f}}_n(x)} = \frac{\frac{1}{nh_y^{p_y}h_x^{p_x}}\sum_{i=1}^n K\left(\frac{y-y_i}{h_y}\right)K\left(\frac{x-x_i}{h_x}\right)}{\frac{1}{h_y^{p_y}}\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)}.$$

But the joint density is $(p_x + p_y)$ dimensional... and the curse strikes big time.

Nonparametric Regression

Data $(y_i, x_i)_{i=1}^n$ now, we are after $E(g(y,x)|x) = m(x)$ for some function g.

- Best-fit approach, quite unbiased:
  - if $x = x_i$ then $\hat{m}_n(x) = g(y_i, x_i)$; otherwise ... whatever.

  But: very jagged estimate; variance independent of n, so not consistent.
- Better and most usual: Nadaraya-Watson, inspired from kernel idea:

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n g(y_i, x_i) K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}.$$

  again, bias in h^2 and variance in $1/nh$ if $p_x = 1$.
  Pitfall 1: very unreliable where $f(x)$ is small.
  Pitfall 2: the formal for the optimal bandwidth h is very ugly.

Choosing h

- Plug-in estimates work badly.
- Fortunately, cross-validation amounts to

$$\min_h \sum_{i=1}^{n} \frac{(g(y_i, x_i) - \hat{m}_n(x_i; h))^2}{1 - k_i(h)}$$

  where $k_i(h) = K_h(0) / \sum_{j=1}^{n} K_h(x_i - x_j)$.
- So not that hard, and can be done on a subsample and rescaled.

Local Linear Regression

- The Nadaraya-Watson estimator in x can be obtained very simply by regressing $g(y_i, x_i)$ on 1, weighting each observation by $K((x - x_i)/h)$.
- We could also regress on 1 and $(x - x_i)$ (going to higher terms has problems) instead;
  Advantages:
  - the bias becomes 0 if the true $m(x)$ is linear.
  - the coefficient of $(x - x_i)$ estimates $m'(x)$.
  - behaves better in "almost empty" regions.

  Disadvantages: hardly any, just do it!

Local Linear

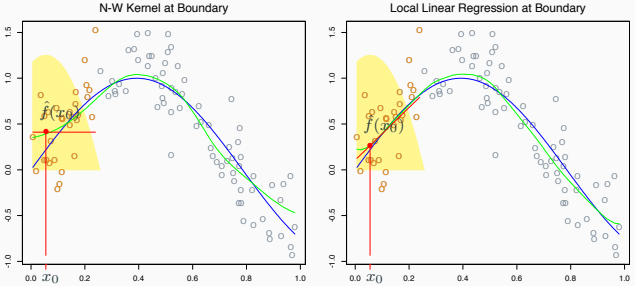

FIGURE 6.3. *The locally weighted average has bias problems at or near the boundaries of the domain. The true function is approximately linear here, but most of the observations in the neighborhood have a higher mean than the target point, so despite weighting, their mean will be biased upwards. By fitting a locally weighted linear regression (right panel), this bias is removed to first order.*

Local Linear

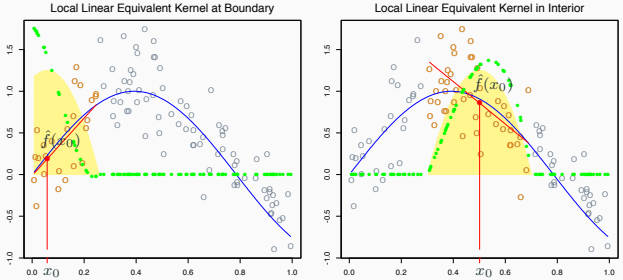

FIGURE 6.4. *The green points show the equivalent kernel $l_i(x_0)$ for local regression. These are the weights in $\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0) y_i$, plotted against their corresponding x_i. For display purposes, these have been rescaled, since in fact they sum to 1. Since the yellow shaded region is the (rescaled) equivalent kernel for the Nadaraya–Watson local average, we see how local regression automatically modifies the weighting kernel to correct for biases due to asymmetry in the smoothing window.*

Local Quadratic

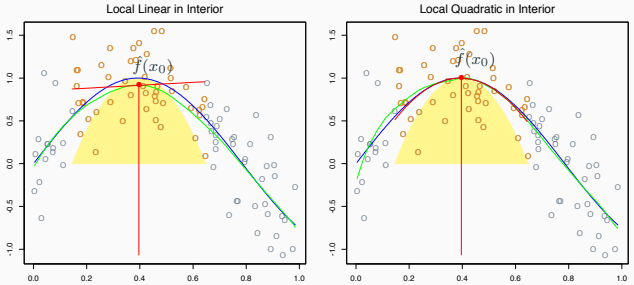

FIGURE 6.5. *Local linear fits exhibit bias in regions of curvature of the true function. Local quadratic fits tend to eliminate this bias.*

Nonparametric Regression, summary, 1

Nadaraya–Watson for $E(y|x) = m(x)$

$$\hat{m}(x) = \frac{\sum_i y_i K_h(x - x_i)}{\sum_i K_h(x - x_i)}$$

- bias in $O(h^2)$, variance in $1/(nh^{p_x})$
- optimal h in $n^{-1/(p+4)}$: then bias, standard error and RMSE all converge at rate $n^{-2/(p+4)}$
- to select h, no rule of thumb: cross-validate on a subsample and scale up.

Nonparametric Regression, summary, 2

Nadaraya–Watson=**local constant regression**: to get $\hat{m}(x)$,

1. regress y_i on 1 with weight $K_h(x - x_i)$
2. take the estimated coeff as your $\hat{m}(x)$.

Better: **local linear regression**

1. regress y_i on 1 and $(x_i - x)$ with weight $K_h(x - x_i)$
2. take the estimated coeffs as your $\hat{m}(x)$ and $\hat{m}'(x)$.

To estimate the standard errors: bootstrap on an *undersmoothed* estimate (so that bias is negligible.)

What if my distribution is discrete-continuous?

Very often in microeconometrics some covariates only take discrete values (e.g. gender, race, income bracket...). Say the only discrete variable is gender, we care about the density of income of men.

- The kernel approach adapts directly: we separately estimate a density for men (on the corresponding subsample).
- *Better:* mix the two subsamples! Add women, but **with a small weight** w.
- Intuition: by doing so we increase the bias (the density for women is probably different than for men) $\to$ bad, in w^2 but we reduce the variance, by $O(w)$; and this dominates for small w. (cf Li-Racine).

Review: What was the point?

- OLS is lowest variance among linear unbiased estimators.
- But there are nonlinear estimators and potentially biased estimators.
  - Everything faces a bias-variance tradeoff.
  - Nearly anything can be written as Kernel.