Nonparametrics and Local Methods: Kernels

C.Conlon

February 28, 2023

Applied Econometrics

Big Data

- It used to be that if you had $N = 50$ observations then you had a lot of data.
- Those were the days of finite-sample adjusted t-statistics.
- Now we frequently have 1 million observations or more, why can't we use k-NN type methods everywhere?

Curse of Dimensionality

Take a unit hypercube in dimension p and we put another hypercube within it that captures a fraction of the observations r within the cube

- Since it corresponds to a fraction of the unit volume, r each edge will be $e_p(r) = r^{1/p}$.
- $e_{10}(0.01) = 0.63$ and $e_{10}(0.1) = 0.80$, so we need almost 80% of the data to cover 10% of the sample!
- If we choose a smaller r (include less in our average) we increase variance quite a bit without really reducing the required interval length substantially.

Curse of Dimensionality

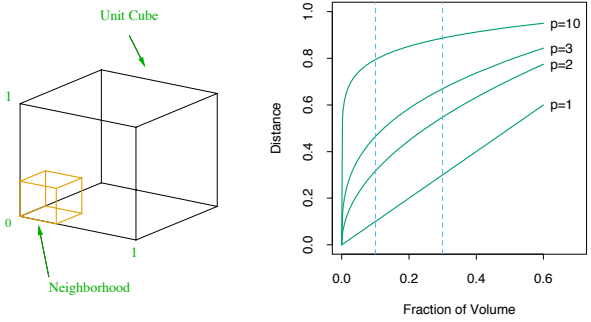

FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Curse of Dimensionality

Don't worry, it only gets worse:

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}$$

- $d(p, N)$ is the distance from the origin to the closest point.
- $N = 500$ and $p = 10$ means $d = 0.52$ or that the closest point is closer to the boundary than the origin!
- Why is this a problem?
- In some dimension nearly every point is the closest point to the boundary – when we average over nearest neighbors we are extrapolating not interpolating.

Density/Distribution Estimation

One of the more successful and popular uses of nonparametric methods is estimating the density or distribution function $f(x)$ or $F(x)$.

- Estimating the CDF is easy and something you have already done
- Q-Q plots, etc.

$$\hat{F}_{ECDF}(x_0) = \frac{1}{N}\sum_{i=1}^{N}(x_i \leq x_0)$$

- Differentiating to get density is unhelpful : $F'_{ECDF}(x) = 0$ in most places.

Density/Distribution Estimation

One of the more successful and popular uses of nonparametric methods is estimating the density or distribution function $f(x)$ or $F(x)$.

- Think about the histogram (definition of derivative):

$$\hat{f}_{HIST}(x_0) = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}(x_0 - h < x_i < x_0 + h)}{2h}$$

Density/Distribution Estimation

- Divide the dataset into bins, count up fraction of observations in each bins
- Similar to k-NN except instead of windows that vary with x_i we have fixed width bins
- Larger bin width $\longrightarrow$ More Bias, Less Variance.
- Histogram will never be smooth! (Just like k-NN).

$$\hat{f}_{HIST}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \cdot \mathbf{1}\left(\left|\frac{x_i - x_0}{h}\right| < 1\right)$$

Smooth Kernels

We can take our histogram and smooth it out:

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) \frac{1}{n} \sum_{i=1}^{n} K_h(y - y_i).$$

We call $K(\cdot)$ a Kernel function and h the bandwidth. We usually assume

(i) $K(z)$ is symmetric about 0 and continuous.

(ii) $\int K(z)dz = 1$, $\int zK(z)dz = 0$, $\int |K(z)|dz < \infty$.

(iii) Either (a) $K(z) = 0$ if $|z| \geq z_0$ for some z_0 or
(b) $|z|K(z) \longrightarrow 0$ as $|z| \longrightarrow \infty$.

(iv) $\int z^K(z)dz = \kappa$ where κ is a constant.

Kernel Smoothers

If K is C^k, then so is $\hat{f}_n$, so we can plot it nicely.

Usually we choose a smooth, symmetric K:

- $K = \phi$, density of $N(0, 1)$ (or some other symmetric density);
- K with compact support: Epanechnikov (mildly) optimal

$$K(x) = \frac{3}{4} \max(1 - x^2, 0).$$

A common nonsmooth choice: $K(x) = (|x| < 1/2)$ gives the *histogram* estimate.

Kernel Comparison

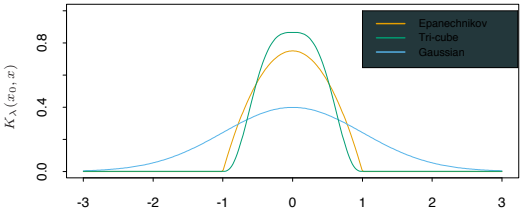

FIGURE 6.2. *A comparison of three popular kernels for local smoothing. Each has been calibrated to integrate to* 1*. The tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikov kernel has none. The Gaussian kernel is continuously differentiable, but has infinite support.*

How to Choose h

- We want both bias and variance to be as small as possible, as usual.
- In parametric estimation, it is not a problem: they both go to zero as sample size increases.

Problem with nonparametrics:

$$E\hat{f}_n(y) = \int K((y-t)/h)f(t)dt/h = \int K(-u)f(y+uh)du = f(y) + O(h^2)$$

$\longrightarrow$ bias can be made tiny by having a very concentrated kernel ($h \simeq 0$); but

$$V\hat{f}_n(y) = \frac{1}{nh^2} VK((y-Y)/h) = O\left(\frac{1}{nh}\right)$$

$\longrightarrow$ a small h gives a high variance!

Reducing h reduces bias, but increases variance; how are we to trade off?

The AMISE

- Asymptotic Mean Integrated Square Error =asymptotic approximation of a quadratic loss function

$$E\left(\hat{f}_n(y) - f(y)\right)^2 dy$$

- Simple approximate expression (symmetric kernels of order 2):

$$(\text{bias})^2 + \text{variance} = Ah^4 + B/nh$$

- **Why?** Bias in y is

$$\int K(-u)\left(f(y+uh) - f(y)\right) du \simeq h^2 \frac{f''(y)}{2} \int K(u)u^2 du.$$

Intuition: if f is close to linear around y, then averaging does not hurt us: $f''(y) \simeq 0$ and the bias is small. The bias is larger (and negative) at the mode of f.

The Variance

$$V\hat{f}_n(y) = \frac{1}{nh^2} VK((y - Y)/h)$$

The important term in

$$VK((y - Y)/h)$$

is

$$h \int K(u)^2 f(y + uh) du \simeq hf(y) \int K(u)^2 du.$$

And we end up with

$$V\hat{f}_n(y) \simeq \frac{f(y)}{nh} \int K(u)^2 du.$$

Intuition: we are really taking an average over $nhf(y)$ points. In low-density region, this induces a high *relative* imprecision:

$$\sigma(\hat{f}_n(y)) \qquad 1 \qquad \int$$

Optimal bandwidth

- The AMISE is

$$Ah^4 + B/nh$$

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

- AMISE is smallest in $h_n^* = \left(\frac{B}{4An}\right)^{1/5}$. Then,
  - bias and standard error are *both* in $n^{-2/5}$
  - and the AMISE is $n^{-4/5}$—**not** $1/n$ as it is in parametric models.
- But: A and B both depend on K (known) and $f(y)$ (unknown), and especially "wiggliness" $\int (f'')^2$ (unknown, not easily estimated). Where do we go from here?

Silverman's Rule of Thumb

- If f is normal with variance σ^2 (may not be a very appropriate benchmark!), the optimal bandwidth is

$$h_n^* = 1.06\sigma n^{-1/5}$$

- Just do it with $\sigma = s$ empirical dispersion of the y_i's , or something more robust/slightly less smooth:

$$h_n^* = 0.9 * \min(s, IQ/1.34) * n^{-1/5},\ \text{IQ=interquartile distance}$$

- Investigate changing it by a reasonable multiple.