Kernel Density Estimation

Ethan Ligon

March 19, 2023

Introduction

In general, we're interested in estimating things like:

- $\mathbb{E}(y|x)$ (Conditional expectations)
- $f(y|x)$ (Conditional pdf)
- Extends to estimating any (smooth?) function.

Note

$$\mathbb{E}(y|x = x) = \int y f(y|x) dy,$$

So if we can estimate $f(y|x)$ we can compute expectations.

Linear Model

For the linear model we've assumed $y = \alpha + \beta x + u$, with $\mathbb{E}(u|x) = 0$, so that

$$\begin{aligned}\mathbb{E}(y|x) &= \mathbb{E}(\alpha + \beta x + u|x)\\ &= \alpha + \beta x,\end{aligned}$$

so that conditional moments are linear functions of the conditioning variables. This leads us to focus on estimating the vector of *parameters* (α, β).

Non-linear Model

In contrast with what we've seen so far in this course, which focused on linear estimation, now we escape our strai(gh)t-jackets! We will aim at estimating

$$\mathbb{E}(y|x) = m(x),$$

where m is a nicely behaved (e.g., smooth, continuous, bounded) but possibly very non-linear function.

Today

Focus on estimating *unconditional* density $f(x)$. Our approach will be fully non-parametric, and will allow us to construct arbitrarily nonlinear densities.

Construction of Estimator

Suppose we have a random sample $\{X_1, X_2, \dots, X_n\}$.

Empirical Distribution Function (EDF)

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$$

We *might* think of taking the derivative of the EDF wrt x, but this would just give us a set of mass points located at the points in the sample.

Instead, *assume* density exists, and recall

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}$$

Then by analogy:

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}.$$

Note this holds h fixed!

$$
\begin{aligned}
\hat{f}(x) &= \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} \\
&= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}(x - h < X_i \leq x + h) \\
&= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}\left(\frac{|X_i - x|}{h} \leq 1\right)
\end{aligned}
$$

or

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

where

$$
k(u) = \begin{cases} 1/2 & |u| \leq 1 \\ 0 & \text{otherwise.} \end{cases}
$$

The function k is an example of a *kernel*. Note that it integrates to one.

Kernels

Lots of possible kernels. Only strict requirement is that $k(u)$ integrate to one. But there are other desirable properties:

Non-negativity $k(u) \geq 0$ for all u. (In this case we can interpret k as a probability density function.)

Boundedness $\int |u|^r k(u) du < \infty$ for all positive integers r.

Symmetry $k(u) = k(-u)$. (Note that boundedness & symmetry imply $\int uk(u)du = 0$.)

Normalized $\int u^2 k(u) du = 1$

Differentiable $\hat{f}$ will inherit differentiability of kernel, and often one prefers "smooth" estimates.

Menagerie of Kernels

See Hansen (2022) for a list of common kernels. In practice you'll most often meet:

Rectangular

$$k(u) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if } |u| < \sqrt{3} \\ 0 & \text{otherwise.} \end{cases}$$

Gaussian

$$k(u) = \frac{1}{2\pi} \exp\left(-\frac{u^2}{2}\right)$$

Epanechnikov

$$k(u) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right) & \text{if } |u| < \sqrt{5}; \\ 0 & \text{otherwise.} \end{cases}$$

Bias of $\hat{f}$

We're interested in $\mathbb{E}\hat{f}(x)$ (NB: this is for x *fixed*). In particular, we want to calculate

$$\text{Bias}(x) = \mathbb{E}\hat{f}(x) - f(x).$$

We have

$$\begin{aligned}\mathbb{E}\hat{f}(x) &= \mathbb{E}\left[\frac{1}{nh}\sum_i k\left(\frac{X_i - x}{h}\right)\right] \\ &= \mathbb{E}\left[\frac{1}{h}k\left(\frac{X - x}{h}\right)\right].\end{aligned}$$

Our next step involves a change of variable: let $u = g(v) = (v - x)/h$, so that $g^{-1}(u) = x + hu$. Then

$$\begin{aligned}\mathbb{E}\hat{f}(x) &= \int \frac{1}{h}k\left(\frac{v - x}{h}\right) f(v)dv \qquad \text{and using change-of-variable} \\ &= \int k(u)f(x + hu)du,\end{aligned}$$

which should remind you of convolutions of continuous random

Variance of $\hat{f}$

To calculate the variance of $\hat{f}(x)$ (again holding x fixed),

$$\begin{aligned}\text{Var}(\hat{f}(x)) &= \frac{1}{(nh)^2}\text{Var}\left[\sum_i k\left(\frac{X_i - x}{h}\right)\right]\\ &= \frac{1}{nh^2}\text{Var}\left[k\left(\frac{X - x}{h}\right)\right].\end{aligned}$$

Estimator of Variance

For a random sample, the quantities $k\left(\frac{X_i-x}{h}\right)$ are sometimes called the "kernel smooths"; note that there are just n of these, and our estimator $\hat{f}(x)$ is just the mean of these.

Analogy

So, we can estimate the sample variance of $\hat{f}$ by just computing the sample variance of the kernel smooths:

$$\widehat{\mathsf{Var}}(\hat{f}(x)) = \frac{1}{n}\left(\frac{1}{nh^2}\sum_i k\left(\frac{X_i - x}{h}\right)^2 - \hat{f}(x)^2\right).$$

MSE/IMSE

In general, estimates both biased and imprecise. Usual measure of this is the *Mean Squared Error*, or

$$\mathsf{MSE}(\hat{f}(x)) = \mathsf{Bias}(\hat{f}(x))^2 + \mathsf{Var}(\hat{f}(x)).$$

Note that the MSE is a *function of* x. To get a summary measure, consider the Integrated Mean Square Error, or

$$\mathsf{IMSE}(\hat{f}) = \int \mathsf{MSE}(\hat{f}(x))dx.$$

Bandwidths (asymptotics)

Idea

- Smaller bandwidths allow for more complicated estimates.
- But sample size has to increase faster than bandwidth shrinks ("effective sample size" has to increase) for asymptotic arguments to work.
- OR: To estimate more complicated things, need more data!

Bandwidths (in practice)

We don't usually get sample sizes that go to infinity, instead we usually have n fixed. So:

- We need a single fixed bandwidth.
- We can see with a fixed bandwidth model is *misspecified*, and at best only an approximation to true density.
- Increasing complexity (smaller bandwidth) holding sample size fixed tends to:
  - Increase variance
  - Decrease bias

  To balance variance vs. bias, appeal to a particular loss function (often MSE).

Bandwidth choice

So how should we go about selecting a bandwidth? The choice is often much more important than the choice of kernel.

We've seen that the MSE (and IMSE) depend on h; how about choosing h to minimize IMSE?

Silverman's rule of thumb

Silverman assumed a Gaussian kernel and that the true f was Gaussian, so he was able to compute the IMSE and find the h that minimized it:

$$h^* \approx \hat{\sigma}1.06/\sqrt[5]{n},$$

where $\hat{\sigma}^2$ is the sample variance.

Take-away

Silverman's rule of thumb is thought to be a decent choice for lots of problems. BUT: a much better general approach would be to construct an estimator of IMSE(h) —we'll later discuss how to use cross-validation to do exactly this.