Kernel Density Estimation

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$. While $F(x)$ can be estimated by the EDF $\hat{F}(x)$, we cannot set $\hat{f}(x) = \frac{d}{dx}\hat{F}(x)$ since $\hat{F}(x)$ is a step function. The standard **nonparametric** method to estimate $f(x)$ is based on **smoothing** using a kernel.

While we are typically interested in estimating the entire function $f(x)$, we can simply focus on the problem where x is a specific fixed number, and then see how the method generalizes to estimating the entire function. So consider x fixed.

Definition 1 *$K(u)$ is a **kernel function** if $K(u) = K(-u)$ (symmetric about zero), $\int_{-\infty}^{\infty} K(u)du = 1$ and $\int_{-\infty}^{\infty} K(u)du = 0$.*

We will focus on the case where $K(u) \geq 0$, so that $K(u)$ is a symmetric density with zero mean. When $K(u) \geq 0$ it is called a second-order kernel and these are the most common used in applications. The kernel will be used as a weighting function.

The most common choices are the **Gaussian** kernel

$$K(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

the **Epanechnikov** kernel

$$K(u) = \begin{array}{cc} \frac{3}{4}\left(1-u^2\right), & |u| \leq 1 \\ 0 & |u| > 1 \end{array}$$

and the **Biweight** or **Quartic** kernel

$$K(u) = \begin{array}{cc} \frac{15}{16}\left(1-u^2\right)^2, & |u| \leq 1 \\ 0 & |u| > 1 \end{array}.$$

The most important choice is the **bandwidth** $h > 0$ which controls the amount of smoothing. If h is large, there is a lot of smoothing, and if h is small there is less smoothing. Let

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right).$$

Note that $K_h(u)$ is a kernel function. If $K(u)$ is a density, then so is $K_h(u)$. The difference is that the variance of K_h is that of K, multiplied by h^2. So as h gets small, the density K_h concentrates about its mean, zero.

Now consider the random variable

$$Y_h = K_h\left(X - x\right)$$

where X is the original random variable, x is a fixed number, and h is a bandwidth. Y_h has mean

$$EY_h = EK_h\left(X-x\right) = \int K_h\left(z-x\right) f(z)dz = \int K_h\left(uh\right) f(x+hu)hdu = \int K\left(u\right) f(x+hu)du$$

The second equality uses the change-of variables $u = (z-x)/h$ which has Jacobian h. The last expression shows that Y is an average of $f(z)$ locally about x.

This integral (typically) is not analytically solvable, so we approximate it using a second order Taylor expansion of $f(x+hu)$ in the argument hu about $hu=0$, which is valid as $h \to 0$. Thus

$$f(x+hu) \simeq f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2$$

and thus

$$\begin{aligned} EY_h &\simeq \int K(u)\left(f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2\right)du \\ &= f(x)\int K(u)\,du + f'(x)h\int K(u)\,udu + \frac{1}{2}f''(x)h^2\int K(u)\,u^2du \\ &= f(x) + \frac{1}{2}f''(x)h^2\kappa \end{aligned}$$

since $\int K(u)\,du = 1$, and $\int K(u)\,udu = 0$, with $\kappa = \int u^2K(u)\,du$, the variance of the kernel $K(u)$.

While for any fixed h, $EY \neq f(x)$, as $h \to 0$, $EY \to f(x)$. Thus we propose estimating $f(x)$ by the sample mean of the Y_h using a "small" value of h. The sample value of Y_h is $Y_i = K_h(X_i - x)$, with sample average

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x).$$

The is the classic nonparametric kernel density estimator of the density $f(x)$. It is the average of a set of weights. If a large number of X_i are near x, then the weights are relatively large and $\hat{f}(x)$ is larger. Conversely, if only a few X_i are near x, then the weights are small and $\hat{f}(x)$ is small. The bandwidth h controls the meaning of "near".

We derived $\hat{f}(x)$ as the estimator of $f(x)$ for fixed x. But it also is the estimator of the entire function. Interestingly, $\hat{f}(x)$ is a valid density when $K(u)$ is a density. That is, since $K(u) \geq 0$, then $\hat{f}(x) \geq 0$ for all x, and

$$\int \hat{f}(x)dx = \int \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\,dx = \frac{1}{n}\sum_{i=1}^{n}\int K_h(X_i - x)\,dx = \frac{1}{n}\sum_{i=1}^{n}\int K(u)\,du = 1$$

where the second-to-last equality makes the change-of-variables $u = (X_i - x)/h$.

We can also calculate the moments of the density $\hat{f}(x)$. The mean is

$$\begin{aligned} \int x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int xK_h(X_i - x)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int (X_i + uh)K(u)\,du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\int K(u)\,du + \frac{1}{n}\sum_{i=1}^{n} h\int uK(u)\,du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \end{aligned}$$

the sample mean of the X_i. Again we used the change-of-variables $u = (X_i - x)/h$. Note: this is the mean of the density $\hat{f}(x)$, not the expectation $E\hat{f}(x)$.

The second moment of the density is

$$\begin{aligned}\int x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int x^2 K_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int \left(X_i + uh\right)^2 K\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \frac{2}{n}\sum_{i=1}^{n}X_i h\int K(u)du + \frac{1}{n}\sum_{i=1}^{n}h^2\int u^2 K\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + h^2\kappa\end{aligned}$$

It follows that the variance of the density $\hat{f}(x)$ is

$$\int x^2\hat{f}(x)dx - \left(\int x\hat{f}(x)dx\right)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 + h^2\kappa - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2 = \hat{\sigma}^2 + h^2\kappa$$

Thus the variance of the estimated density is inflated by the factor $h\kappa$ relative to the sample moment.

We now explore the sampling properties of $\hat{f}(x)$. Specifically, we calculate the bias, variance and MSE.

The bias is easy to calculate. We have

$$E\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}EK_h\left(X_i - x\right) = f(x) + \frac{1}{2}f''(x)h^2\kappa$$

so

$$Bias(x) = \frac{1}{2}f''(x)h^2\kappa.$$

We see that the bias of $\hat{f}(x)$ at x depends on the second derivative $f''(x)$. The sharper the derivative, the greater the bias. Intuitively, the estimator $\hat{f}(x)$ smooths data local to $X_i = x$, so is estimating a smoothed version of $f(x)$. The bias results from this smoothing, and is larger the greater the curvature in $f(x)$.

The integrated squared bias (a global measure of bias) is

$$\int Bias(x)^2 dx = \frac{h^4\kappa^2 R(f'')}{4}$$

where

$$R(f'') = \int \left(f''(x)\right)^2 dx$$

is the **Roughness** of f'' or f. It is called the roughness because it indexes the amount of wiggles in f. Not surprisingly, the global bias is higher when the roughness is greater.

Furthermore, we can see that for any x and globally in x, the bias tends to zero as h tends to zero. Thus for the bias to asymptotically disappear, h must go to zero as $n \to \infty$. This is a minimal requirement for consistent estimation.

We now examine the variance of $\hat{f}(x)$. Since it is an average of iid random variables, using first-order Taylor approximations and the fact that n^{-1} is of smaller order than $(nh)^{-1}$ when $h \to 0$ as $n \to \infty$,

$$\begin{aligned} Var\left(x\right) &= \frac{1}{n} Var\left(K_h\left(X_i - x\right)\right) \\ &= \frac{1}{n} EK_h\left(X_i - x\right)^2 - \frac{1}{n}\left(EK_h\left(X_i - x\right)\right)^2 \\ &\simeq \frac{1}{nh^2}\int K\left(\frac{z-x}{h}\right)^2 f(z)dz - \frac{1}{n} f(x)^2 \\ &= \frac{1}{nh}\int K\left(u\right)^2 f\left(x + hu\right) du \\ &\simeq \frac{f\left(x\right)}{nh}\int K\left(u\right)^2 du \\ &= \frac{f\left(x\right) R(K)}{nh}. \end{aligned}$$

The integrated variance is

$$\int Var\left(\hat{f}(x)\right) dx \simeq \int \frac{f\left(x\right) R(K)}{nh} dx = \frac{R(K)}{nh}.$$

We see that for fixed x or globally, the variance tends to zero if $nh \to \infty$ as $n \to \infty$.

Together, the asymptotic mean-squared error (AMSE) for fixed x is the sum of the approximate squared bias and approximate variance

$$AMSE_h(x) = \frac{1}{4} f''(x)^2 h^4 \kappa^2 + \frac{f\left(x\right) R(K)}{nh}$$

and the mean integrated squared error (AMISE) is

$$AMISE_h = \frac{h^4 \kappa^2 R(f'')}{4} + \frac{R(K)}{nh}. \tag{1}$$

A sufficient condition for consistent estimation is that the MSE tends to zero as $n \to \infty$. This occurs iff $h \to 0$ yet $nh \to \infty$ as $n \to \infty$. That is, h must tend to zero, but at a slower rate than n^{-1}.

Equation (1) is an asymptotic approximation to the MSE. We define the asymptotically optimal bandwidth h_0 as the value which minimizes this approximate MSE. That is,

$$h_0 = \underset{h}{\operatorname{argmin}}\, AMISE_h$$

It can be found by solving the first order condition

$$\frac{d}{dh} AMISE_h = h^3 \kappa^2 R(f'') - \frac{R(K)}{nh^2} = 0$$

yielding

$$h_0 = \left(\frac{R(K)}{n\kappa^2 R(f'')}\right)^{1/5}. \tag{2}$$

This solution takes the form $h_0 = cn^{-1/5}$ where c is a function of K and f, but not of n. We thus say that the optimal bandwidth is of order $O(n^{-1/5})$. Note that this h declines to zero, but at a very slow rate.

In practice, how should the bandwidth be selected? This is a difficult problem, and there is a large and continuing literature on the subject. We see that the optimal choice is given in (2). Since n is given, and K (and thus $R(K)$ and κ) are selected by the researcher, all components are known except $R(f'')$. The obvious trouble is that this is unknown, and could take any value!

A classic simple solution proposed by Silverman has come to be known as the "reference bandwidth" or "Silverman's Rule-of-Thumb." It uses formula (2) but replacing the unknown f with the $N(0, \hat{\sigma}^2)$ distribution, where $\hat{\sigma}^2$ is an estimate of σ^2. This choice for h gives an optimal rule when $f(x)$ is normal, and gives a nearly optimal rule when $f(x)$ is close to normal. The downside is that if the density is very far from normal, the rule-of-thumb h can be fairly inefficient. Working through the integrals, the rule-of-thumb choice h is a simple function of n, depending on the kernel K being used.

Gaussian Kernel: $h_{rule} = 1.06n^{-1/5}$

Epanechnikov Kernel: $h_{rule} = 2.34n^{-1/5}$

Biweight (Quartic) Kernel: $h_{rule} = 2.78n^{-1/5}$

Unless you delve more deeply into kernel estimation theory, my recommendation is to use the rule-of-thumb bandwidth, perhaps adjusted by visual inspection of the resulting esitmate $\hat{f}(x)$. While there are other approaches, the advantages and disadvantages are delicate. I now discuss some of these choices. The **plug-in** approach is to estimate $R(f'')$ in a first step, and then plug this estimate into the formula (2). This is more treacherous than may first appear, as the optimal h for estimation of the roughness $R(f'')$ is quite different than the optimal h for estimation of $f(x)$. However, there are modern versions of this estimator which appear to work well. Another popular choice for selection of h is known as **cross-validation**. This works by constructing an estimate of the MISE using leave-one-out estimators. There are some desireable properties of cross-validation bandwidths, but they are also known to converge very slowly to the optimal values. They are also quite ill-behaved when the data has some discretization (as is common in economics), in which case the cross-validation rule can sometimes selected very small bandwidths, leading to dramatically undersmoothed estimates. Fortunately there are remedies, which are known as **smoothed cross-validation** which is a close cousin of the **bootstrap**.

Computation

Typically, we calculate $\hat{f}(x)$ in order to have a graphical representation of the density function. In this case, we start by defining a set of gridpoints $\{x_1, ..., x_g\}$ where we will calculate $\hat{f}(x)$. Some researchers set the gridpoints equal to the sample values. Others set a uniform grid between the min and max of the data or a selected quantile. At each point x_j, the density estimate

$$\hat{f}(x_j) = \frac{1}{n} \sum_{i=1}^{n} K_h (X_i - x_j)$$

is calculated. An easy way to do this is to write the computer code to loop across the x_j, and then compute $\hat{f}(x_j)$ at each point by a simple sample average of the kernel weights. (This is not an efficient computational algorithm, but ease of programming often outweighs numerical efficiency.) Once these have been all calculated, the pairs $\{x_j, \hat{f}(x_j)\}$ can be plotted.