Nonparametric Density Estimation

Econ 690

Purdue University

Density Estimation

Suppose that you had some data, say on wages, and you wanted to estimate the density of this data. How might you do this?

1. Histogram Approach

2. Assume a class of densities, say normal family, and estimate parameters via MLE.

- Both of these are reasonable, yet each can be argued to be deficient in some respects.

Density Estimation

Let's take a closer look at the histogram method:

We first divide the space into say m bins of length $2h$.

The histogram is then defined as:

$$\hat{f}(x_0) = \frac{1}{2nh}(\text{Number of } x_i \text{ in same bin as } x_0).$$

Clearly, this is a proper density estimate since:

$$\begin{aligned} \int \hat{f}(x)dx &= \frac{1}{2nh}\left[2h(\text{Number of } x_i \text{ in bin } 1) + \cdots + 2h(\text{Number of } x_i \text{ in bin } m)\right] \\ &= \frac{1}{2nh}2hn = 1 \end{aligned}$$

Density Estimation

Alternatively, we can relax the condition that the m bins form a *partition*, and instead, add up all of the x_i within a given interval of width $2h$ of x_0 as follows:

$$\hat{f}(x_0) = \frac{1}{2nh} \sum_{i=1}^{n} \left[\mathcal{I}\left(\left| \frac{x_i - x_0}{h} \right| < 1 \right) \right].$$

Density Estimation

We might think that the weight function above is not ideal in the following senses:

- Within the interval, all of the data points should not receive equal weight - those closer to x_0 should get more weight, and conversely for those farther away.
- In a similar spirit, perhaps all of the data points could potentially get some weight. (This, however, might not be warranted).

To this end, we replace the indicator function above with a continuous weighting function (or *kernel*):

$$K\left(\frac{x_i - x_0}{h}\right).$$

Density Estimation

- Though this is not always the case, we often think of K as a mean-zero symmetric density function.
- Thus, we assign the highest weight to points near zero, or x_i values closest to x_0. Points far away receive little or no weight.
- Symmetry of the kernel implies equal treatment of points the same distance below and above x_0.

The above yields the *kernel density estimator:*

$$\hat{f}(x_0) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right).$$

What if K represents the uniform density on $(-1, 1)$?

Density Estimation

Some points to note:

1. x_0 is a fixed point. Thus, to recover an entire density estimate, we need to choose a grid of points and compute the above for all points in the grid. (3σ rule, perhaps).

2. h_n is called a *bandwidth* or *smoothing parameter*. If h_n is "large" then $K(\cdot)$ is small and nearly equal for most points resulting in a "smoothed" estimate. If h_n is small, then $K(\cdot)$ assigns nearly zero weight for all points but those very close to x_0 which yields "jumpy" estimates.

Pointwise Consistency of the Kernel Estimator

We will show convergence in mean square, whence consistency. Assume:

1. $\{x_i\}$ is iid
2. $\int K(s)ds = 1$
3. $\int sK(s)ds = 0$
4. $\int |s^3 K(s)|ds < \infty$
5. $\int s^2 K^2(s)ds < \infty$
6. $f(x)$ is three times differentiable, with bounded third derivative over the support of X.
7. $h_n \to 0$, $nh_n \to \infty$

With these assumptions, we can show $\hat{f}(x_0) \xrightarrow{p} f(x_0)$ for x_0 an interior point in the support of X.

Pointwise Consistency of the Kernel Estimator

To establish pointwise consistency, we separately consider the bias and variance of the kernel estimator. As for the bias, note:

$$\begin{aligned} E[\hat{f}(x_0)] &= \frac{1}{nh_n}\sum_{i=1}^{n} E\left[K\left(\frac{x_i - x_0}{h_n}\right)\right] \\ &= \frac{1}{h_n} E\left[K\left(\frac{x - x_0}{h_n}\right)\right] \\ &= \frac{1}{h_n}\int K\left(\frac{x - x_0}{h_n}\right) f(x)dx \end{aligned}$$

Pointwise Consistency of the Kernel Estimator

Now let us make a change of variables. Let

$$s = (x - x_0)/h_n$$

which implies

$$dx = h_n d_s.$$

Noting that the limits of integration stay constant, we can write the above as:

$$\begin{aligned} E[\hat{f}(x_0)] &= \frac{1}{h_n} \int K(s) f(sh_n + x_0) h_n ds \\ &= \int K(s) f(sh_n + x_0) ds \end{aligned}$$

Pointwise Consistency of the Kernel Estimator

Now, let us take a Taylor Series Expansion of $f(sh_n + x_0)$ around x_0:

$$f(sh_n + x_0) = f(x_0) + sh_n f'(x_0) + \frac{s^2 h_n^2}{2} f''(x_0) + \frac{h_n^3 s^3}{6} f'''(\tilde{x}),$$

for some $\tilde{x}$ in between $sh_n + x_0$ and x_0. So, we can write the expectation of the kernel estimator as:

$$\begin{aligned} &= \int K(s) \left[f(x_0) + sh_n f'(x_0) + \frac{s^2 h_n^2}{2} f''(x_0) + \frac{h_n^3 s^3}{6} f'''(\tilde{x}) \right] ds \\ &= f(x_0) + \frac{h_n^2}{2} f''(x_0) \int s^2 K(s) ds + O(h_n^3) \end{aligned}$$

Pointwise Consistency of the Kernel Estimator

To explain the very last term, let a_n be a nonstochastic sequence. We write that $a_n = O(n^\alpha)$ if $|a_n| \leq Cn^\alpha$ for all n sufficiently large.

Now, consider the term in the remainder of our expansion:

$$\frac{1}{6}h_n^3 \left| \int f'''(\tilde{x}) s^3 K(s) ds \right| \leq Ch_n^3 \int \left| s^3 K(s) ds \right| = O(h_n^3).$$

Pointwise Consistency of the Kernel Estimator

Therefore, the bias of the nonparametric density estimator is given as:

$$\frac{h_n^2}{2} f''(x_0) \int s^2 K(s) ds + O(h_n^3)$$

Thus, provided $h_n \to 0$ and the above regularity conditions, we see that the bias goes to zero, which completes the first part of our proof.

Pointwise Consistency of the Kernel Estimator

Thus, it remains to show that the variance of our estimator goes to zero.

$$
\begin{aligned}
\text{Var}(\hat{f}(x_0)) &= \text{Var}\left[\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\right] \\
&= \frac{1}{n^2 h_n^2}\text{Var}\left[\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\right] \\
&= \frac{1}{n^2 h_n^2} n\text{Var}\left[K\left(\frac{x - x_0}{h_n}\right)\right] \\
&= \frac{1}{nh_n^2}\left[\underbrace{E\left[K^2\left(\frac{x - x_0}{h_n}\right)\right]}_{A} - \underbrace{E^2\left[K\left(\frac{x_i - x_0}{h_n}\right)\right]}_{B}\right]
\end{aligned}
$$

Pointwise Consistency of the Kernel Estimator

Using our previous methods for the bias of the kernel estimator, one can show that $[nh_n^2]^{-1}B \to 0$. Now, consider A:

$$\begin{aligned} A &= \frac{1}{nh_n^2}\int K^2(s)f(sh_n + x_0)h_n ds \\ &= \frac{1}{nh_n}\int K^2(s)\left[f(x_0) + sh_n f'(x_0) + \frac{s^2h_n^2}{2}f''(x_0) + R\right] ds \\ &= \frac{1}{nh_n}f(x_0)\int K^2(s)ds + \frac{1}{n}f'(x_0)\int sK^2(s)ds + \frac{h_n}{2n}f''(x_0)\int K^2(s)s^2 ds + \cdots \end{aligned}$$

Pointwise Consistency of the Kernel Estimator

Clearly, each of these terms are converging to zero, provided $nh_n \to \infty$. Thus, under the stated conditions, the kernel estimator is *consistent*.

Also, ignoring the leading terms, the variance is:

$$\text{Var}(\hat{f}(x_0)) \approx \frac{1}{nh_n} f(x_0) \int K^2(s)ds.$$

and the bias is

$$\text{Bias} \approx \frac{h_n^2}{2} f''(x_0) \int s^2 K(s)ds.$$

A few observations

1. We see an obvious tradeoff - smaller bandwiths decrease the bias of our estimator, but also increase the variance. Thus, there should be some type of optimal bandwidth selector.

2. The bandwidth choices are like *counterfactuals* - they describe how the bandwidth would behave if the sample size increased indefinitely.

3. The conditions $h_n \to 0$ and $nh_n \to \infty$ are "offsetting." The intuition is that for larger sample sizes, we look at tighter local neighborhoods, but we do not let the size of this neighborhood shrink too fast!

Optimal Bandwidth Selection

Given the bias-variance tradeoff, a natural criterion to use for bandwidth selection is Mean Squared Error:

$$\begin{aligned} \text{MSE} &= \text{Bias}^2 + \text{Variance} \\ &= \left[\frac{1}{2}h_n^2 f''(x)\int s^2 K(s)ds\right]^2 + \left(\frac{1}{nh_n}f(x)\int K^2(s)ds\right). \end{aligned}$$

Here, we have ignored the higher-order tems in the bias and variance expression. The above applies to a particular point, x. To get a global criterion, we look at *Mean Integrated Squared Error*.

Optimal Bandwidth Selection

Define

$$K_2 \equiv \int s^2 K(s) ds.$$

Then,

$$\text{MISE} = \frac{1}{4} h_n^4 K_2^2 \int (f''(x))^2 dx + \frac{1}{nh_n} \int K^2(s) ds.$$

To find a bandwidth which minimizes this criterion, we take FOC's with respect to h_n:

$$(h_n^*)^3 K_2^2 \int (f''(x))^2 dx - \frac{1}{n(h_n^*)^2} \int K^2(s) ds = 0.$$

Optimal Bandwidth Selection

This implies:

$$(h_n^*)^5 = \frac{1}{n}\left[\int K^2(s)ds\right]\left[\int (f''(x))^2 dx\right]^{-1} K_2^{-2}.$$

So, we have the optimal bandwidth:

$$h_n^* = n^{-1/5}\left[\int K^2(s)ds\right]^{1/5}\left[\int (f''(x))^2 dx\right]^{-1/5} K_2^{-2/5}.$$

Some points which emerge from the above are worth discussing:

- The optimal bandwidth depends on the unknown function $f(x)$! This is problematic, since this is the function we are trying to estimate.
- As $n \to \infty$, h_n^* gets small and $nh_n^* \to \infty$ is also satisfied.
- If f'' is "large" then the density is fluctuating rapidly, and the above tells us that h_n^* will be smaller. This is sensible - when the true density fluctuates rapidly, choose a small bandwidth to capture the local behavior.

Rule-of-thumb Bandwidth Selector

A simple rule of thumb is to assume (as in Silverman (1985)) that f and K are normal. This yields the optimal bandwidth:

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

A similar rule, argued to be more robust by Silverman, is

$$h_n^* = .9An^{-1/5},$$

where

$$A = \min\{\text{Std. Dev(x)}, \text{Interquartile Range(x)}/1.34\}.$$

Popular Kernels

Name	*DensityFunction*	*Support*
Epanechnikov	$\frac{3}{\sqrt{5}}\left(1-\frac{1}{5}x^2\right)$	$\mid x \mid \leq \sqrt{5}$
Biweight	$\frac{15}{16}(1-x^2)^2$	$\mid x \mid \leq 1$
Triangular	$1- \mid x \mid$	$\mid x \mid \leq 1$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)$	$-\infty < x < \infty$
Rectangular	$1/2$	$\mid x \mid \leq 1$

In terms of computation, it is clear that Kernels with compact support are most attractive. In this way, we can save time by only determining the weights for those points which the kernel gives non-zero weight.

Multivariate Density Estimation

Let $x_0 \in \Re^d$. The multivariate kernel density estimator is defined as

$$\hat{f}(x_o) = \frac{1}{n|H|} \sum_{i=1}^{n} K\left[H^{-1}(x_i - x_0)\right],$$

where $K : \Re^d \to \Re$ (a d dimensional kernel) and H is a $d \times d$ bandwidth / smoothing matrix.

In practice, the kernel K is often selected as a *product kernel*:

$$K(u) = \prod_{j=1}^{d} K(u_j).$$

Choosing the bandwidth Matrix

There are several popular alternatives here, and all of them are rather closely related:

1. $H = hI_d$. This choice is sensible only if all of the individual x variables are on the same scale. Otherwise, you will be choosing a constant degree of smoothing for variables with both high and low variances.

2. $H = \text{diag}\{h_1\ h_2\ \cdots\ h_d\}$. This is similar to the above. Let s_j denote the scaling factor for the j^{th} variable. Then, let $H = h\text{diag}\{s_1\ s_2\ \cdots\ s_d\}$. So, this can be regarded as an application of (1) after first transforming all variables to have unit variance.

3. $H = hS^{1/2}$, where S is an estimate of the covariance matrix of x, perhaps from the sample: $\hat{S} = 1/n \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$.

Choosing the bandwidth Matrix

For simplicity in notation, let $S^{1/2} = M$ so that $S = MM'$. (Such a factorization is possible since S is pds).
In addition, note that

$$\begin{aligned}
\text{Var}(M^{-1}x) &= E\left([M^{-1}x - E(M^{-1}x)][M^{-1}x - E(M^{-1}x)]'\right) \\
&= (M)^{-1}E[(x - E(x))(x - E(x))'](M')^{-1} \\
&= M^{-1}S(M')^{-1} \\
&= M^{-1}MM'(M')^{-1} \\
&= I_d
\end{aligned}$$

Thus the transformed data $M^{-1}x$ has unit covariance matrix. This process is often called *sphering* the data.

Choosing the Bandwidth Matrix

Now, consider applying our rule above:

$$\hat{f}(x_0) = \frac{1}{n|hM|} \sum_{i=1}^{n} K\left(\frac{1}{h} M^{-1}(x_i - x_0)\right).$$

Let $y = M^{-1}x, \Rightarrow x = My$. By a change of variables to the above,

$$\begin{aligned} f(y) &= \frac{1}{nh^d|M|}|M| \sum_{i=1}^{n} K\left(\frac{1}{h}(y_i - y_0)\right) \\ &= \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{1}{h}(y_i - y_0)\right). \end{aligned}$$

Choosing the Bandwidth Matrix

The second line is simply a multivariate density estimate for the transformed data y that results from applying $H = hI_d$ as in (1). Thus, the approach in (3) is equivalent to one that linearly transforms the data to have unit covariance matrix, applies the bandwidth selection rule in (1), and then re-transforms back to the x scale.

There remains the issue of choosing h. One popular rule-of-thumb is to again assume the use of a Gaussian kernel and that the true density is Gaussian, and derive the asymptotic mean integrated squared error.

Choosing the Bandwidth Matrix

This yields the rule:

$$H = \left(\frac{4}{d+2} \right)^{1/(d+4)} [\Sigma^{1/2}] n^{-1/(d+4)}.$$

Note that the form of H for this particular case is exactly in the form of $H = h_n S^{1/2}$.