Estimation of the Probability Density Function by the Kernel Method

Here the goal is to obtain a smoother estimate of the probability density function (pdf) $f(x)$ than found in a histogram, kernel density estimation is a commonly used approach. To illustrate this approach, start with a histogram where the axes have been scaled to represent relative frequency. If we partition the range of the data into subintervals $a_1 < a_2 < ... < a_k$, then the density estimate for a value x in the interval $a_i < x \leq a_{i+1}$ is:

$$\widehat{f}(x) = \frac{\text{number of observations in } (a_i, a_{i+1}]}{n(a_{i+1} - a_i)}.$$

Now suppose that we use equal-spaced intervals so that $\Delta = a_{i+1} - a_i$, then an alternate way to express the histogram for a value x in the interval $a_i < x \leq a_{i+1}$ is:

$$\widehat{f}(x) = \frac{1}{n\Delta}(\text{number of observations in } (a_i, a_{i+1}]) = \frac{1}{n\Delta}\sum_{i=1}^{n} I(X_i \in (a_i, a_{i+1}]),$$

which is very similar to:

$$\widehat{f}(x) = \frac{1}{n\Delta}\sum_{i=1}^{n} I(|x - X_i| < \Delta/2) = \frac{1}{n\Delta}\sum_{i=1}^{n} w_1(\frac{x - X_i}{\Delta})$$

for $w_1(u) = I(|u| < 1/2)$. This particular function is very jagged, so the kernel method in general is:

$$\widehat{f}(x) = \frac{1}{n\Delta}\sum_{i=1}^{n} w(\frac{x - X_i}{\Delta}),$$

for a **kernel function** $w(u)$ that is typically smoother than $w_1(u)$ (for example, $w(u)$ can be the standard normal pdf) and a **bandwidth** Δ.

A key element of using kernel density estimation is the choice of bandwidth Δ. As shown in the Figures accompanying this lecture, a very small bandwidth gives essentially a 'needle plot' with small spiked areas occuring only where there is data, while a very large bandwidth gives essentially a uniform distribution as the density estimate. It turns out that using a small bandwidth gives a density estimate with low bias but large variance, while using a large bandwidth gives a density estimate with large bias but small variance. Thus, to minimize the mean-squared error of the density estimate we must reach a compromise bandwidth. The area of bandwidth selection has been, and remains, an active research topic in statistics. Our text mentions a rule suggested in Härdle (1991):

$$\Delta = 1.06 \ \min(S, \frac{IQR}{1.34}) \ n^{-1/5},$$

where S is the sample standard deviation, IQR is the sample interquartile range, and n is the sample size. For the Old Faithful Geyser data from the text, $S = 1.04$, $IQR = 2$, and $n = 107$, giving an approximate value of $\Delta = .433$. An article by Jones, Marron, and Sheather (1996) summarizes different approaches to bandwidth selection and distinguishes between what they call 'first generation' and improved (but more computationally intensive) 'second generation' methods. The recommendation above from Härdle is a first generation method, and a commonly used second generation choice is called the Sheather-Jones method.

Kernel density estimation is available in many software packages, including SAS and R. In SAS, there are at least three procedures that allow kernel density plots: Proc KDE, Proc Sgplot, and Proc Univariate. Of these procedures, Proc KDE gives the most options including the most sophisticated bandwidth selection methods (it automatically uses the Sheather-Jones method). However, apparently none of these SAS procedures outputs the value of the bandwidth estimate, althought all of them allow the user to change the selected bandwidth by use of a bandwidth multiplier. R has a density function that allows a choice of bandwidth estimation methods, choices of the kernel function, and it also allows the user to printout the bandwidth estimate.

References

Härdle, W. (1991) Smoothing Techniques. New York: Springer-Verlag.

Jones, M.C., Marron, J.S., and Sheather, S.J. (1996) A Brief Survey of Bandwidth Selection for Density Estimation. Journal of the American Statistical Association, 91: 401-407.