Lecture 7
Problems and Solutions for Density Estimation

In this lecture we consider ways to improve density estimation.

Bias at the Boundary of the Distribution

Many distributions have *bounded support*, that is, the range of values for which $f(y) > 0$ is bounded.

Kernel density estimators have a significant weakness for estimating densities with bounded support: *edge (or boundary) effects*.

The form of the estimator generates downward bias for y within h of the boundaries, unless the density is zero there.

$$f(y) \equiv \frac{d}{dy}F(y) \equiv \lim_{h \to 0} \frac{F(y+h) - F(y-h)}{2h}$$

Intuitively this is because the estimator does not recognize the boundaries. It acts as if the density is zero outside the interval while it is in fact undefined there.

Example: Nassau County Schools

There are ways to remove this bias by using a special kernel function called a *boundary kernel* (Gasser and Müller 1979).

Practice:
the `cut` option in *density* employs boundary kernel bandwidths.

Lack of local variation in smoothing

The degree of smoothing is the same for all values of y (and controlled by h).

We often want to smooth rapidly changing parts of the density less and values with higher density more:
If $f(y)$ is large then increase h to reduce variance
If $|f''(y)|$ is large then decrease h to reduce bias

Example: Black Womens Wages

Flattening of peaks and valleys

The estimate tends to be flatter than the actual density:
proportional to $f''(y)$
flattening of peaks and troughs of the density.

We often want to smooth rapidly changing parts of the density less and values with higher density more:
If $f(y)$ is large then increase h to reduce variance
If $|f''(y)|$ is large then decrease h to reduce bias

Example: Nassau County Schools

Adjustments and Improvements of Kernel Estimates

Use Boundary kernels rather than the default ones

Use a *variable-bandwidth kernel estimator*:

$$\hat{f}(y) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{h(Y_j)} K\left(\frac{y - Y_j}{h(Y_j)}\right)$$

Use transformations: It $X = g(Y)$ then

$$f_Y(y) = g'(y) \times f_X(g(y))$$

Parametric Likelihood Density Estimation

The traditional alternative to the methods we have considered is *parametric density estimation*.

The data are assumed to come from a density function defined by a finite number of parameters. e.g.,

$$f(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{(2\pi)}} exp[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma}] - \infty < y < \infty$$

$$f(y; \mu, \sigma) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\theta_1}(1-y)^{\theta_2} \qquad 0 < y < 1$$

These distributions (i.e., parameters) can be fit to the data (i.e., $Y_1, ..., Y_m$) using a number of methods.

One method that is usually asymptotically optimal is *maximum likelihood* estimation. Assuming the Y_j has density function $f(y;\theta)$.

$$L(\theta|Y_1 = y_1, \ldots, Y_m = y_m) \equiv P(Y_1 = y_1, \ldots, Y_m = y_m; \theta) = \prod_{j=1}^{m} f(y_j;\theta)$$

Usually the log of L is considered:

$$\mathcal{L}(\theta|Y_1 = y_1, \ldots, Y_m = y_m) \equiv log[L(\theta|\cdot)] = \sum_{j=1}^{m} \log[f(y_j;\theta)]$$

A maximum likelihood estimator (MLE) for θ is a value $\widehat{\theta}$ that maximizes $\mathcal{L}(\theta|Y_1, \ldots, Y_m)$ as a function of θ.

The density estimate is then $\widehat{f}\updownarrow(y) \equiv f(y;\widehat{\theta})$.

Example:

If $f(y;\theta)$ is Gaussian with $\theta = (\mu, \sigma)$ then $\widehat{\theta} = (\bar{y},\ \frac{m-1}{m} s_Y)$.

What if we think that the true density is similar to members of $f(y;\theta)$ but not exactly a member of it?

We expect to look locally like the parametric model but vary far away.

To preserve likelihood-based estimation but to operate locally a *local log-likelihood function* is defined:

$$loc\mathcal{L}(y,\theta) \equiv \sum_{j=1}^{m} \frac{1}{mh} K\left(\frac{y-Y_j}{h}\right) \log[f(y_j;\theta)] - \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{y-u}{h}\right) \log[f(u;\theta)] du$$

Let $\widehat{\theta}(y)$ be the maximizer of $loc\mathcal{L}(y, \theta)$ for each y.
The density estimate is then $\widehat{f}_{l\mathcal{L}}(y) \equiv f(y; \widehat{\theta}(y))$.

If h is large this is close to proportional to $\mathcal{L}(\theta | Y_1 = y_1, \ldots, Y_m = y_m)$.
i.e., a parametric fit

If h is small it fits using values close to y only.

If $f(y; \theta)$ is the uniform distribution the it is very close to the ordinary kernel density estimator.

If $f(y; \theta)$ is more flexible (e.g., 2 or more parameters) the properties of the estimator are better than the ordinary kernel estimator.

Often the bandwidth h is adaptively chosen (i.e., $h(y)$) to contain exactly a proportion p of the observations closest to y

Practice:

The *locfit* library implements local likelihood density estimation.

Spline-based Density Estimation

The *logspline* library implements log-density spline density estimation.