Introduction to Applied Multivariate Analysis with R

Visualizing Multivariate Data

Kernel Density Estimators

Want to identify ***regions*** or ***clusters*** of high and low densities so we add some type of bivariate density estimate to plot

From the definition of a probability density, if the random variable X has a density f,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} \mathsf{P}(x - h < X < x + h). \tag{2.1}$$

For any given h, a naïve estimator of $\mathsf{P}(x - h < X < x + h)$ is the proportion of the observations $x_1, x_2, \ldots, x_n$ falling in the interval $(x - h, x + h)$,

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} I(x_i \in (x - h, x + h)); \tag{2.2}$$

i.e., the number of $x_1, \ldots, x_n$ falling in the interval $(x - h, x + h)$ divided by $2hn$. If we introduce a weight function W given by

Kernel Density Estimators

$$W(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{else,} \end{cases}$$

then the naïve estimator can be rewritten as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - x_i}{h}\right). \tag{2.3}$$

Unfortunately, this estimator is not a continuous function and is not particularly satisfactory for practical density estimation. It does, however, lead naturally to the kernel estimator defined by

$$\hat{f}(x) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.4}$$

where K is known as the *kernel function* and h is the *bandwidth* or *smoothing parameter*. The kernel function must satisfy the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Kernel Density Estimators

We implement the three kernel functions indicated below in R

Usually, but not always, the kernel function will be a symmetric density function; for example, the normal. Three commonly used kernel functions are

rectangular,

$$K(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{else.} \end{cases}$$

triangular,

$$K(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & \text{else,} \end{cases}$$

Gaussian,

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Bivariate Kernel Estimators

Bivariate estimator for data (x1,y1),(x2,y2),…,(xn,yn) is defined as:

$$\hat{f}(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}, \frac{y-y_i}{h_y}\right). \qquad (2.5)$$

In this estimator, each coordinate direction has its own smoothing parameter, h_x or h_y. An alternative is to scale the data equally for both dimensions and use a single smoothing parameter.

Bivariate Kernel Estimators

Bivariate Epanechnikov kernel:

For bivariate density estimation, a commonly used kernel function is the standard bivariate normal density

$$K(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.$$

Another possibility is the bivariate Epanechnikov kernel given by

$$K(x, y) = \begin{cases} \frac{2}{\pi}(1 - x^2 - y^2) & x^2 + y^2 < 1 \\ 0 & \text{else,} \end{cases}$$