NONPARAMETRIC DENSITY ESTIMATION

JIAN ZHANG
JIANZHAN@STAT.PURDUE.EDU

The *probability density function (pdf)* is a fundamental concept in statistics. Given the pdf $f(x)$ of a random variable X, probabilities associated with X can be easily computed as

$$P(a \leq X < b) = \int_a^b f(x)dx.$$

Many problems in statistics can be described as using a (random) *sample* to infer the underlying (unknown) *population*. Given a sample $x_1, \ldots, x_n \sim f$, the problem of *density estimation* is to contruct an estimate $\hat{f}$ of f based on the sample.

Parametric vs. Nonparametric

Density estimation methods can be roughly categorized into *parametric density estimation* and *nonparametric density estimation*. In parametric density estimation, f is assumed to be a member of a parametric family (such as normal with unknown μ and σ^2), and the density estimation problem is then transformed into a simpler one: find estimates of μ and σ^2 and plug into the normal density formula.

In nonparametric density estimation, we do not restrict the form of f with any parametric assumptions (We still need some smoothness assumptions in order to analyze theoretical properties). It is not easy to give a clear definition of what is "nonparametric", but sometimes it is useful to think nonparametric as "infinite parametric".

Histogram

Histogram is probably the oldest and simplest density estimator (see `hist()` in R). Given an orgin x_0 and the bin-width h, the bins of a histogram are defined to be the intervals $[x_0 + mh, x_0 + (m+1)h)$ for postive and negative integers m. The histogram estimator is defined by

$$\hat{f}(y) = \frac{1}{nh}(\#\,\mathsf{of}\ x_i\ \mathsf{in\ the\ same\ bin\ as}\ y),$$

and it is easy to verify that $\int \hat{f}(y)dy = 1$. In order to apply the histogram method, we need to choose both x_0 and h, and it is h which primarily determines the smoothness of the estimator $\hat{f}$.

Some drawbacks of histogram: (1) discontinuity of $\hat{f}$; (2) the choice of x_0 can affect the shape; (3) inefficient usage of data; (4) presentation difficulty in multivariate density estimation.

Kernel Density Estimator

Consider a *kernel function* K which satisfies that $K(y) > 0$ and

$$\int_{-\infty}^{\infty} K(y)dy = 1.$$

Usually K is taken to be some symmetric density function such as the pdf of normal. The *kernel density estimator* with kernel K is defined by

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - x_i}{h}\right)$$

where h is known as the *bandwidth* and plays an important role (see `density()` in R).

Intuitively, the kernel density estimator is just the summation of many "bumps", each one of them centered at an observation x_i. The kernel function K determines the shape of the bumps while h determines the width

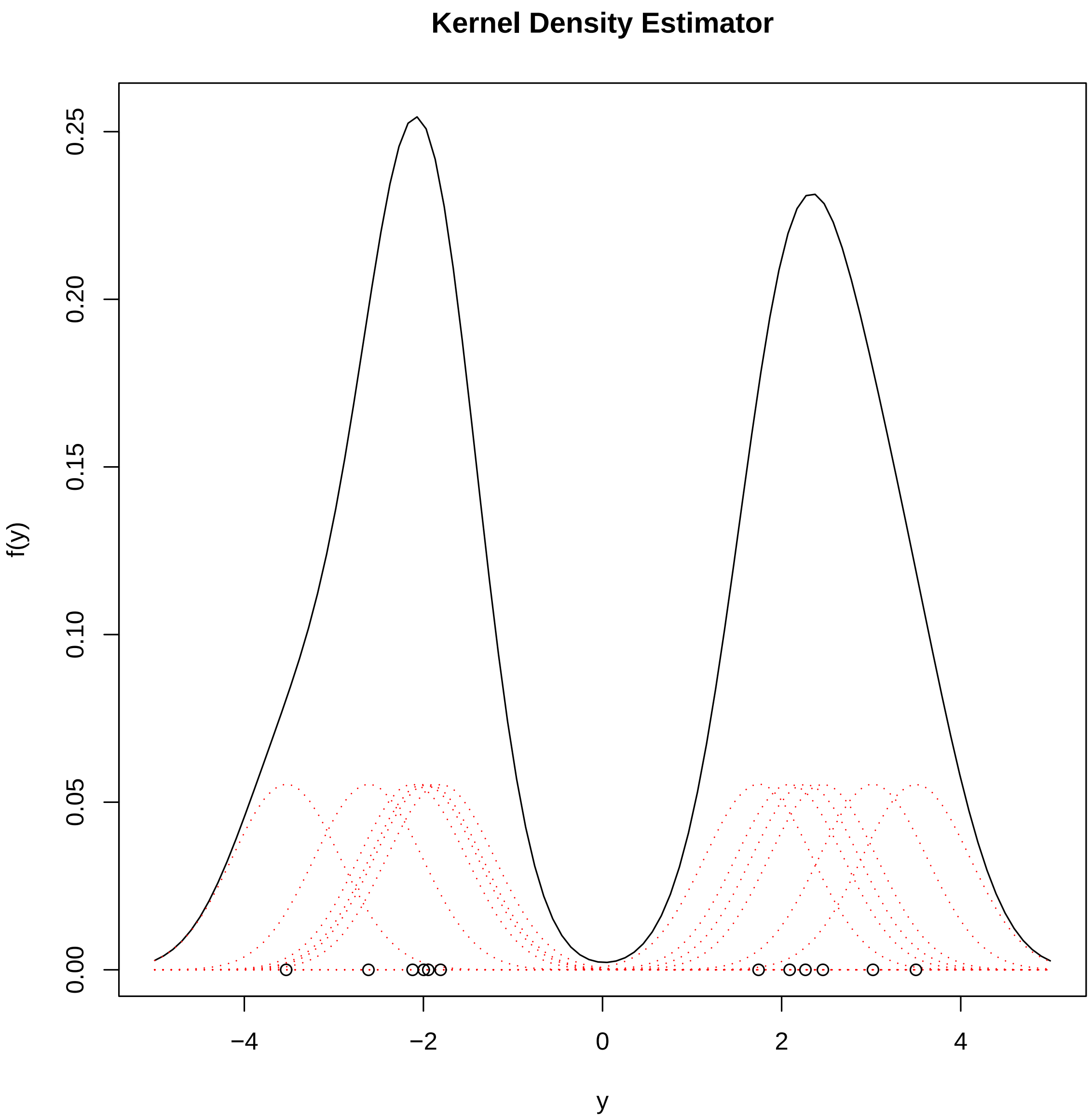

FIGURE 1. Kernel Density Estimator ($K(y) = \phi(y|0,1)$ and $h = 0.6$)

of the bumps. Figure 1 shows a kernel density estimator with $h = 0.6$ and normal kernel for 12 points sampled from a mixture density $f(y) = 0.5N(-2,1) + 0.5N(2,1)$.

FITNESS MEASURE

Once we construct an estimator $\hat{f}(y)$, we want to evaluate how good is $\hat{f}(y)$ as an approximation to the true density $f(y)$. When considering estimation at a single point, the mean square error (MSE) is often used:

$$\mathsf{MSE}_y(\hat{f}) = \mathbb{E}[(\hat{f}(y) - f(y))^2]$$

which can be decomposed into the well-known bias-variance tradeoff:

$$\mathsf{MSE}_y(\hat{f}) = (\mathbb{E}[\hat{f}(y)] - f(y))^2 + \mathbb{V}[\hat{f}(y)].$$

To measure the global accuracy of $\hat{f}$ to f, the mean integrated square error (MISE, which is the same as IMSE) is defined by

$$\begin{aligned}\mathsf{MISE}(\hat{f}) &= \mathbb{E}\left[\int (\hat{f}(y) - f(y))^2 dy\right] \\ &= \int \mathbb{E}[(\hat{f}(y) - f(y))^2] dy.\end{aligned}$$

It is important to notice that the bias

$$\mathbb{E}[\hat{f}(y)] - f(y) = \int \frac{1}{h} K\left(\frac{y-x}{h}\right) f(x)dx - f(y)$$

which does not depend on the sample size n. In other words, the bias does not go to zero as $n \to \infty$. In fact, in order for the kernel density estimator to be consistent (both bias and variance diminish), we have to let $h \to 0$ and $nh \to \infty$. To be more specific, assume that $\int K(t)dt = 1$, $\int tK(t)dt = 0$, $\int t^2 K(t)dt = k_2 \neq 0$ and $R(K) = \int K(t)^2 dt$, it can be shown that

$$\begin{aligned}\mathsf{Bias}(\hat{f}(x)) &= \frac{1}{2} h^2 f''(x) k_2 + O(h^4) \\ \mathsf{Var}(\hat{f}(x)) &= \frac{f(x)R(K)}{nh} - \frac{f(x)^2}{n} + O(h/n),\end{aligned}$$

and from which it is clear that to be consistent we need to let $h \to 0$ and $nh \to \infty$.

Bandwidth Selection

The parameter h in kernel density estimation has a very important role in controlling the smoothness of the estimator $\hat{f}$. Generally speaking, the smaller the h is, the smaller the bias and the larger the variance. Given a random sample $x_1, x_2, \ldots, x_n$, how to select a bandwidth h in kernel density estimator $\hat{f}$?

Here we introduce the idea of cross validation based on the least square criteria. Given an estimator $\hat{f}$ of f, the integrated square error is

$$\int (\hat{f}(y) - f(y))^2 dy = \int \hat{f}(y)^2 dy - 2\int \hat{f}(y) f(y) dy + \int f(y)^2 dy.$$

Since the last term does not depend on $\hat{f}$, we define the following quantity (risk)

$$R(\hat{f}) = \int \hat{f}(y)^2 dy - 2\int \hat{f}(y) f(y) dy.$$

The basic idea is to find an estimator $\hat{R}$ of $R(\hat{f})$ using the sample itself and then choose h which minimizes $\hat{R}$. Define $\hat{f}^{\setminus i}$ to be the density estimate contructed using all data except x_i:

$$\hat{f}^{\setminus i}(y) = \frac{1}{n-1} \frac{1}{h} \sum_{j \neq i} K\left(\frac{y - x_j}{h}\right)$$

and define

$$\hat{R}(h) = \int \hat{f}(y)^2 dy - \frac{2}{n} \sum_i \hat{f}^{\setminus i}(x_i).$$

It can be shown that $\hat{R}$ is in fact a good estimator of $R(\hat{f})$ since

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_i \hat{f}^{\setminus i}(x_i)\right] &= \mathbb{E}\left[\hat{f}^{\setminus 1}(x_1)\right] \\ &= \mathbb{E}\left[\int \hat{f}^{\setminus 1}(y) f(y) dy\right] \\ &= \mathbb{E}\left[\int \hat{f}(y) f(y) dy\right].\end{aligned}$$

So we have $\mathbb{E}[\hat{R}] = \mathbb{E}[R(\hat{f})]$, and then it follows that $\hat{R}(h) + \int f(y)^2 dy$ is an unbiased estimator of the mean integrated square error. As a result, we can find the h which minimizes $\hat{R}(h)$ and use it as the desired bandwidth parameter.

Other Topics

(1) The extension of kernel density estimator to multivariate case is straightforward, by replacing univariate kernel function with multivariate kernel function. One commonly used kernel function is the multivariate normal density. In the multivariate case, instead of having a single bandwidth parameter h, we need a $p \times p$ bandwidth matrix H (which is often assumed to be diagonal for simplicity).

(2) Computation can be chanllenging for multivariate kernel density estimator. If we choose a kernel function with bounded support (or truncate the multivariate normal for an approximation), then the computation can be made cheaper through the identification of nearest neighbors in $\mathbb{R}^p$.

(3) In theory people have shown that the bandwidth parameter h is very important, while the choice/shape of the kernel function has little influence on the result. Instead of using a fixed h for all $y \in \mathbb{R}$, we can consider an adaptive $h(y)$ which varies at different location.

(4) Compared to the estimation of typical parameters, density estimation is usually much more difficult. Although in theory given the density function f we can compute any interested parameter, in practice its estimation should be avoided when possible. As a simple (but illustrative) example, consider the estimation of population mean μ. An easy and good estimator is just the sample mean, and it would be strange to first construct $\hat{f}$ and then $\hat{\mu}$.

References

[1] Bernard Silverman. *Density Estimation for Statistician and Data Analysis*, Chapman & Hall/CRC, 1998.