Density Estimation for Statistics and Data Analysis
Chapter 1 and 2

B.W.Silverman

April 17, 2003

Introduction

Suppose we have a set of observed data points assumed to be a sample from an unknown density function. **Our goal is to estimate the density function from the observed data**. There are two approaches to density estimation,parametric and nonparametric.

• Parametric approach assumes,data is drawn from a known distribution.

• Nonparametric approach assumes that the distribution has a probability density f and then the data is used to estimate f rather than deciding beforehand if f belongs to any given parametric family.

Data sets

Two data sets have been repeatedly used in this chapter.

The first comprises the lengths of 86 spells of psychiatric treatment undergone by patients in a study of suicide risks.
The second data set contains observations of eruptions of Old Faithful geyser in Yellowstone National Park,USA.

Assumptions

We are given a sample of n real observations $X_1, ..., X_n$ whose underlying density is to be estimated. We denote the density estimator by $\hat{f}$.

Histograms

It is the oldest and most widely used density estimator. The bins of the histogram are defined as the intervals $[x_0 + mh, x_0 + (m+1)h)$,for m positive and negative integers, x_0 is the origin and h the bin width. The histogram is then defined by

$\hat{f}(x) = \frac{(\text{number of } X_i \text{ in the same bin as } x)}{nh}$.

The histogram can be generalized by allowing the bin widths to vary. Then the estimate becomes

$$\hat{f}(x) = \frac{\big(\text{number of } X_i \text{ in the same bin as } x\big)}{n\big(\text{width of bin containing } x\big)}$$

But there are certain **drawbacks** in using histograms.

• In procedures like cluster analysis and nonparametric discriminant analysis using a histogram results in inefficient use of the data.

• The histogram is not continuous so trouble arises when derivatives are required.

• Choice of origin may have an effect in the interpretation.

• Representing bivariate or trivariate data by histogram is difficult.

Naive Estimator

If the random variable X has density f,then

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

Thus,

$$\hat{f}(x) = \frac{[\text{number of } X_i's \text{ in } (x-h, x+h)]}{2hn}. \tag{2}$$

This is the naive estimator. We define it more clearly by a weight function as follows.

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } -1 < x < 1 \\ 0 & \text{if otherwise} \end{cases} \tag{3}$$

Then $\hat{f}(x)$ becomes

$$\frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right). \tag{4}$$

Now we try to connect this estimator with the histogram discussed earlier.

We may consider that the estimate is constructed by placing a "box" of width $2h$ and height $(2nh)^{-1}$ on each observation and then summing all these boxes to get the estimate. Let us now think about histograms of bin width $2h$. We assume no observation lies on the edge of a bin. If x is at the center of the bin then $\hat{f}(x)$ will be the ordinate of the histogram at x.

But this estimator also has got some **drawbacks**.

• $\hat{f}$ is not continuous but has jumps at the points $X_i \pm h$ and has zero derivative everywhere else.

Kernel Estimator

Replace the earlier weight function by K,where

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

Then

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{5}$$

is called the Kernel estimator. Here h is the smoothing parameter.

The variation in window width has an effect on the estimator. We see in the figures if h is too small then fine structure is visible and if h is too large then the important features are sometimes obscured.

Next we state the **elementary properties** of the kernel estimator.

$*$ If K everywhere is nonnegative and satisfies

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

then $\hat{f}$ will be a probability density.

$*$ The estimator $\hat{f}$ will inherit all the continuous and different properties of K.

$*$ Most commonly used estimator.

But this estimator also has **drawbacks**.

• As the window width is fixed so there is a tendency for spurious noise to appear in the tails of the estimates and if the estimates are smoothed to deal with this then the essential features of the main part of the distribution may be masked.

Nearest Neighbour Method

This is an attempt to adapt smoothing to the 'local' density of data. Here we define distance between two points on the line as

$$d(x, y) = |x - y|$$

For each t we define $d_1(t) \leq d_2(t) \leq ... \leq d_n(t)$, which are the distances from t to the points of the sample. Then the kth nearest neighbour density estimator is

$$\hat{f}(t) = \frac{k-1}{2nd_k(t)} \tag{6}$$

In the tails of the distribution the distance $d_k(t)$ will be larger than in the main part and so the problem of undersmoothing in the tails will be reduced.

Its **drawbacks** are

- The estimate is not a smooth curve.
- $d_k(t)$ will be continuous but its derivatives are not.
- From t less than the smallest data point we will have $d_k(t) = X_{(k)}(t)$ and for $t > X_{(n)}, d_k(t) = t - X_{(n-k+1)}$. Putting in $\hat{f}(t)$ we see

$$\int_{-\infty}^{\infty} \hat{f}(t)dt = \infty$$

Thus this estimate is inappropriate if an estimate of the entire density is required.

Variable Kernel Method

K is the kernel function and k is a positive integer. $d_{j,k}$ is the distance from X_j to the kth nearest point in the set of the other $(n-1)$ data points. Then the variable kernel estimate is

$$\hat{f}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h d_{j,k}} K\left(\frac{t - X_j}{h d_{j,k}}\right) \tag{7}$$

The window width is proportional to $d_{j,k}$. Therefore the data points in regions where data is few will have flatter kernels. But this estimator will be a probability density function if K is, and will inherit all the properties of K.

Orthogonal Series Estimators

Orthogonal series estimators approach the density estimation problem from a different point of view. We explain them by using an example. Suppose we are trying to estimate a density f on the unit interval $[0, 1]$. Define the sequence $\phi_\nu(x)$ by

$$\phi_0(x) = 1$$

$$\phi_{2r-1}(x) = \sqrt{2}cos2\pi rx$$

$$\phi_{2r}(x) = \sqrt{2}sin2\pi rx$$

for $r = 1, 2, ...$

Then f can be represented as $\sum_{v=0}^{\infty} f_\nu \phi_\nu$,
where for each $\nu \geq 0$,

$$f_\nu = \int_0^1 f(x)\phi_\nu(x)dx.$$

Suppose X is a rv with density f then we can write the above equation as

$$f_\nu = E\phi_\nu(X)$$

and hence a natural and unbiased estimator of f_ν based on sample $X_1, ..., X_n$ is

$$\hat{f}_\nu = \frac{1}{n}\sum_{i=1}^{n} \phi_\nu(X_i)$$

To get a good estimate of f we need to choose an integer K and estimate $\hat{f}$ by

$$\hat{f}(x) = \sum_{\nu=0}^{K} \hat{f}_\nu \phi_\nu(x). \tag{8}$$

The choice of K determines the amount of smoothing. A more general approach to smooth is by using a series of weights λ_ν, which tends to 0 when $\nu \to 0$. We can then write the estimate as

$$\hat{f}(x) = \sum_{\nu=0}^{\infty} \lambda_\nu \hat{f}_\nu \phi_\nu(x).$$

Maximized Penalized Likelihood Estimators

We now investigate if it is possible to use standard statistical techniques to do density estimation. The likelihood of a curve g as density underlying a set of independent identically distributed observations is given by

$$L(g|X_1, ..., X_n) = \prod_{i=1}^{n} g(X_i)$$

But the likelihood has no finite maximum over the class of all densities so to use in density estimation we have to place restrictions on the class of densities over which the likelihood is to be maximized.

One method is to incorporate a term which describes the roughness. Let $R(g)$ be the functional which quantifies roughness. One possible choice is

$$R(g) = \int_{-\infty}^{\infty} (g'')^2. \tag{9}$$

We now define the *Penalized log likelihood* by

$$l_\alpha(g) = \sum_{i=1}^{n} logg(X_i) - \alpha R(g) \tag{10}$$

where α is a positive smoothing parameter.

The probability density function $\hat{f}$ is said to be a *maximum penalized likelihood density estimate* if it maximizes $l_\alpha(g)$ over the class of all curves g which satisfy $\int_{-\infty}^{\infty} g = 1, g(x) \geq 0$ for all x,
$R(g) < \infty$.

General Weight Function Estimators

It is possible to define a general class of density estimators which includes several of the estimators discussed before. Suppose that $w(x, y)$ is a function of two arguments, which in most cases will satisfy the conditions

$$\int_{-\infty}^{\infty} w(x, y)dy = 1 \tag{11}$$

and

$$w(x, y) \geq 0 \text{ for all } x \text{ and } y. \tag{12}$$

The density estimator is then

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} w(X_i, t). \tag{13}$$

The above two conditions will ensure that $\hat{f}$ will be a probability density function,and $\hat{f}$ will inherit the smoothness properties of w.

The histogram can be obtained as

$$w(x, y) = \begin{cases} \frac{1}{h(x)} & \text{if } x\text{and } y \text{ fall in the same bin} \\ 0 & \text{if otherwise} \end{cases}$$

where $h(x)$ is the width of the bin containing x.

The kernel estimate is obtained as

$$w(x, y) = \frac{1}{h} K\left(\frac{y - x}{h}\right)$$

Bounded Domains And Directional Data

It is very often the case when the domain of definition is not the whole real line but an interval bounded on one or both sides. There are many possible ways to deal with this situation.

§ We can simply calculate the estimate for positive x ignoring the boundary conditions,and then set $\hat{f}(x)$ to zero for negative x.

§ We can use an a maximum penalized likelihood method by constraining $g(x)$ to be zero for negative x.

§ Another possible way is to take logarithms of the data points. Then the estimate leads to

$$\hat{f}(x) = \frac{1}{x}\hat{g}(\log x) \text{ for } x > 0.$$

§ We can augment the data by adding the reflections of all the points in the boundary,to give the set
$\{X_1, -X_1, X_2, -X_2, ...\}$. If a kernel estimate f^* is used from the data set of size $2n$ then we get

$$\hat{f}(x) = \begin{cases} 2f^*(x) & \text{for } x \geq 0 \\ 0 & \text{for} x < 0 \end{cases}$$

This estimate corresponds to a general weight function estimator, for x and $y > 0$,it is

$$w(x, y) = \frac{1}{h}K\left(\frac{y-x}{h}\right) + \frac{1}{h}K\left(\frac{y+x}{h}\right)$$

All the above discussed methods can be extended to the case where the support of the estimator is $[a, b]$. Transformation

methods can be based on transformations of the form

$$Y_i = H^{-1}\left(\frac{X_i - a}{b - a}\right)$$

where H is any cumulative pdf strictly increasing on $(-\infty, \infty)$.