HUDM 6026

Density Estimation

Final project

- Introduction: a brief, half to one page, description of the method/data, source/references, types of applications, and what you want to achieve.
- Main part: implement the method. Include graphs, if applicable, and summary of your results.
- Discussion: Discuss any issues you may have encountered. If applicable compare to other methods. Evaluate performance.

Grading the project

- Clarity, 15%: If I scratch my head and ask myself, “what the heck are they trying to say?” several times when reading the paper, then its probably not very clear.
- Thoroughness, 55%: Did you perform all applicable methods? Was the analysis adequate? Is there something in the data/method that you failed to discuss?
- Summary and comparison, 25%: Did you use the methods correctly and did you compare them.
- The wow factor, 5%: Extremely well-written papers will be rewarded. Did the student go beyond the call of duty in their analysis? If method was covered in class this is a mandatory step.

Density Estimation

- Density estimation is a collection of methods for constructing an estimate of a probability density, as a function of a sample of data.
- We have used density estimation informally: a histogram is a type of density estimator. The R function `density` is another type of density estimator which uses kernel methods.
- Here we discuss nonparametric density estimation.

Univariate Density Estimation

Here we discuss histogram, frequency polygon, and kernel density estimators.

Histograms

The histogram is a piecewise constant approximation to the function. The main issue about histograms is selecting the class boundaries. Here we present the Sturges and Scott methods.

Histograms

Suppose that a random sample X_1, … , X_n is observed. To construct the frequency or probability histogram the data must be sorted into bins. The class boundaries for the bins can be any numbers within the range of the sample, but some choices are more reasonable than others. Bins' widths can be different but here we discuss uniform bin width. Given the class intervals of width *h* are provided, the histogram density estimate based on a sample of size *n* is:

$$\hat{f}(x) = \frac{\nu_k}{nh}, t_k \leq x < t_{k+1}$$

Where ν_k is the number of sample points in the class interval $[t_k, t_{k+1})$.

Sturges’ Rule

Sturges’ rule tends to oversmooth the data, but it is the default in many statistical packages. According to Sturges, the optimal width of the class intervals is given by

$$\frac{R}{1+\log_2 n}$$

Where R is the sample range.

Scott’s Rule

Scott shows that the optimal choice of bin width is

$$h_n^* = \left(\frac{6n}{\int f'(x)^2 dx} \right)^{1/3}$$

However *f* is unknown so the above can’t be computed. It is estimated by:

$$\hat{h} = 3.49 \hat{\sigma} n^{-1/3}$$

Frequency Polygon

Histograms are piecewise continuous but discontinuous at the endpoints of the bins. A frequency polygon provides the a continuous density estimate from the same frequency distribution use to produce the histogram by connecting the dots between the midpoints of each class interval.

Optimal frequency polygon bin width is

$$\hat{h} = 2.15\hat{\sigma}n^{-1/5}$$

Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric technique for density estimation in which a known density function (the kernel) is averaged across the observed data points to create a smooth approximation.

KDE – Smoothing the Histogram

Let $X_1, \ldots, X_n$ be a random sample taken from a continuous, univariate density *f*. The kernel density estimator is given by,

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\{(x - X_i)/h\}$$

- K is a function satisfying $\int K(x)\,dx = 1$
- The function *K* is referred to as the *kernel*.
- *h* is a positive number, usually called the *bandwidth* or *window width*.

Kernels

- Gaussian
- Epanechnikov
- Rectangular
- Triangular
- Biweight
- Cosine

Refer to Table 10.2 Rizzo, page 299.

… *most unimodal densities perform about the same as each other when used as a kernel.*

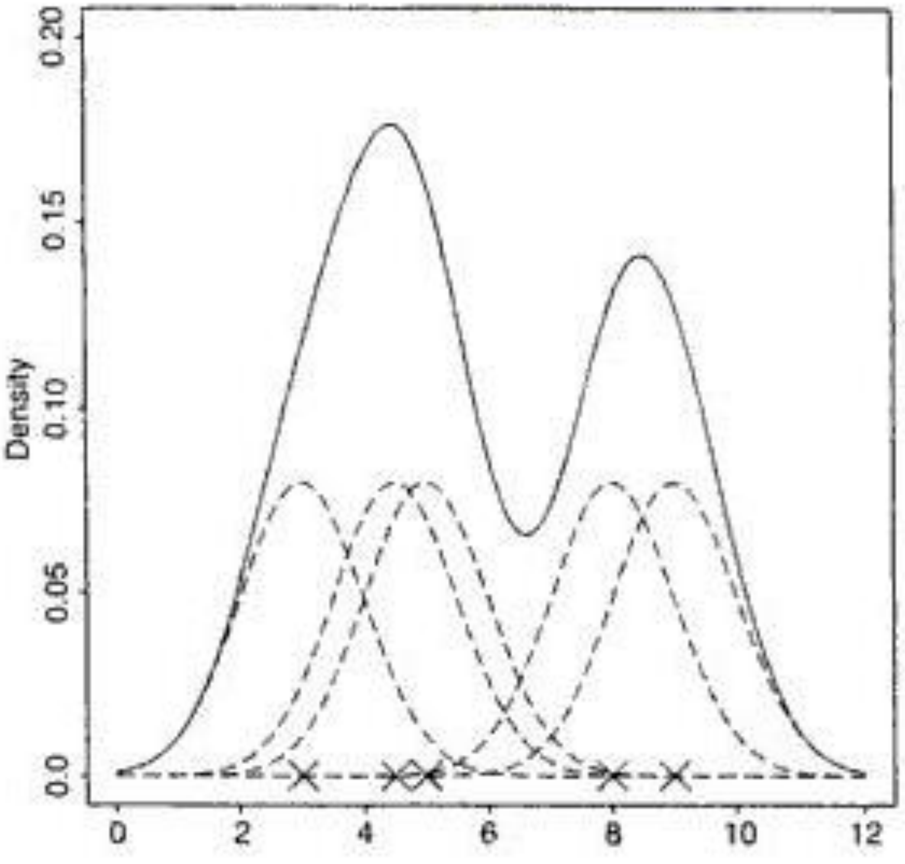
- Typically *K* is chosen to be a unimodal PDF.
- Use the **Gaussian** kernel.

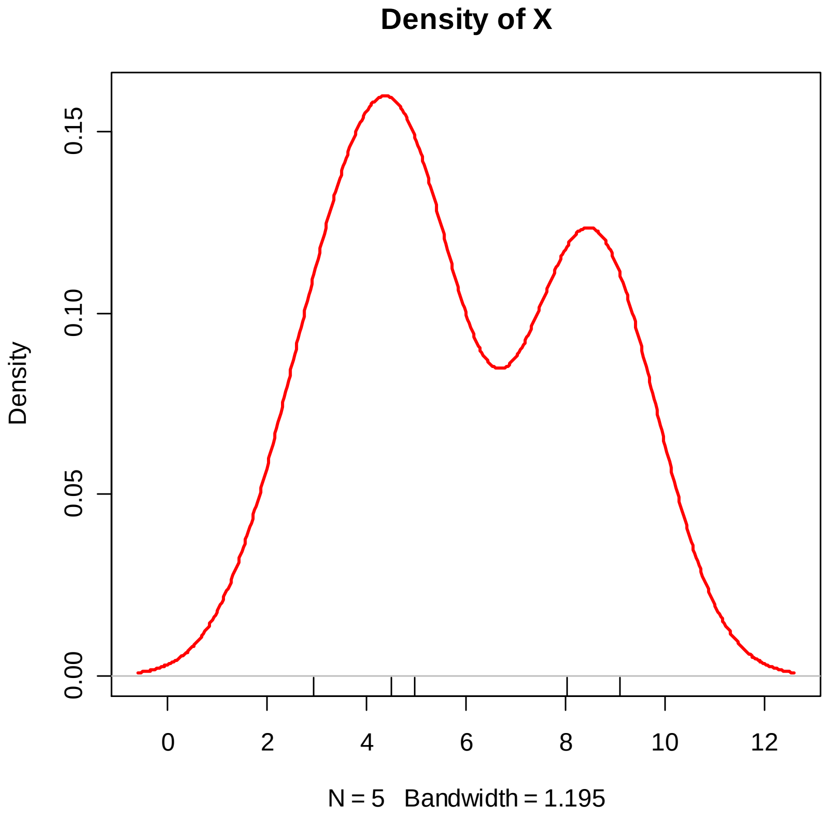
Wand M.P. and M.C. Jones (1995), *Kernel Smoothing*, Monographs on Statistics and Applied Probability 60, Chapman and Hall/CRC, 212 pp.

KDE – Based on Five Observations

Kernel density estimate constructed using five observations with the kernel chosen to be the *N*(0,1) density.

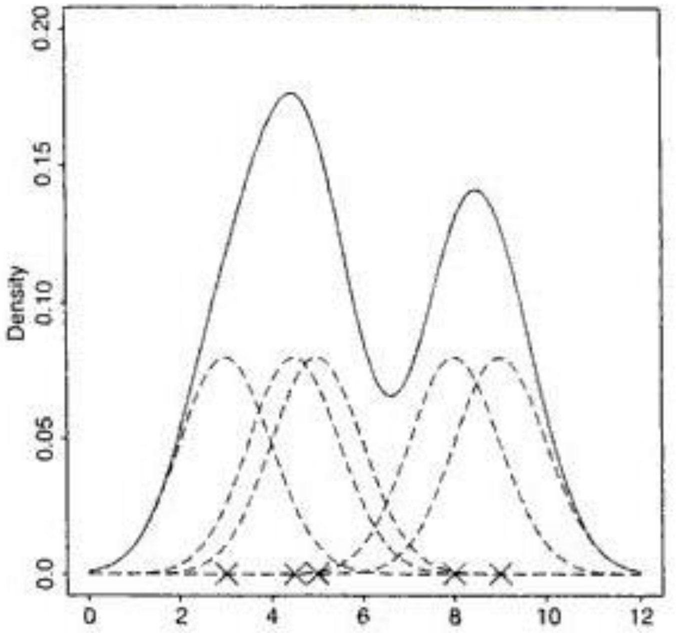
x=c(3, 4.5, 5.0, 8, 9)





KDE – Numerical Implementation

```
"kde" <- function(x,h)
{
npt=100
r <- max(x) - min(x); xmax <- max(x) + 0.1*r; xmin <- min(x) - 0.1*r
n <- length(x)
xgrid <- seq(from=xmin, to=xmax, length=npt)
f = vector()
for (i in 1:npt){
  tmp=vector()
  for (ii in 1:n){
    z=(xgrid[i] - x[ii])/h
    density=dnorm(z)
    tmp[ii]=density
  }
  f[i]=sum(tmp)
}
f=f/(n*h)
lines(xgrid,f,col="grey")
} #end function
```



$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\{(x - X_i)/h\}$$

Bandwidth Estimators

- Optimal Smoothing
- Normal Optimal Smoothing
- Cross-validation
- Plug-in bandwidths

Bandwidth Estimators

- For a Gaussian kernel the bandwidth *h* that optimizes the IMSE is

$$h = 1.06\sigma n^{-1/5}$$

- Silverman rule:

$$h = 0.9\mathrm{min}(S, \frac{\mathrm{IQR}}{1.34})n^{-1/5}$$