KERNEL DENSITY ESTIMATIONS: TOTAL POINT TESTS DISTRIBUTION OF ADMITTED STUDENTS IN HIGHER EDUCATION

Markela Muça, Lule Hallaci, Llukan Puka, Klodiana Bani

Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Albania

e-mail: markela.muca@fshn.edu.al

Abstract: Histograms and kernel density estimation are the usual vehicle for representing data distribution graphically. In this paper we study some kernel density estimation in order to evaluate the percentage entrancy probability in different studies programs of the Faculty of Natural Sciences. We will summarize the techniques of cross validation methods for bandwidth choise in the kernel estimation of probability density. We would like to study the total points in the final tests distribution of admitted students in higher education. The analysis is based on a sample taken from the database of the MSH 2013 (Matura Shteterore), from the National Agency of Exams (NAE) and from the Ministry of Science and Education (MSE).All analysis were performed using KDE package in R.

Key words: *Kernel density estimation, histograms, bandwidth selection*

Introduction

Histograms are simple and easy to construct for univariate data, and are therefore used quite frequently in many application domains, but they suffer from several defects i.e., the number of bins, smoothing, the outlier points (Silverman, B.W. 2003). The outlier points are data points that lie in bins with very low frequency where all the data including in each bins are equality with the midpoint (see, figure 1). Therefore, the count for a bin does not include the point itself in order to minimize over fitting for smaller bin widths. The kernel estimation is an alternative method, which involves smoothing the data while retaining the overall structure.A histogram can be thought of as a simplistic kernel density estimation, which uses a kernel to smooth frequencies over the bins. This yields a smoother probability density function, which will in general more accurately reflect distribution of the underlying variable. The density estimate could be plotted as an alternative to the histogram, and is usually drawn as a curve rather than a set of boxes.

The words used to describe the patterns in a histogram are: "symmetric", "skewed left" or "right", "unimodal", "bimodal" or "multimodal".

A histogram is a simple nonparametric estimate of a probability distribution (Howitt, D. and Cramer, D., 2005).

For constants a_0 and h, let $a_k = a_0 + kh$ and $H_k = \#\{X_i / X_i \in (a_{k-1}, a_k]\}$ be the number of observation in the k[th] interval $(a_{k-1}, a_k]$. Then the histogram estimator of $f(x)$ is:

$$\hat{f}_{hist} = \frac{1}{hn} \sum_{k=1}^{n} H_k 1_{(a_{k-1}, a_k]}(x)$$

Histograms for the graphical presentation of bivariate or trivariate data present several difficulties; (Silverman, B.W. 2003), for example, one cannot easily draw contour diagrams to represent the data, and the problems raised in the univariate case are exacerbated by the dependence of the estimates on the choice not only of an origin but also of the coordinate direction(s) of the grid of cells. Finally, it should be stressed that, in all cases, the histogram still requires a choice of the amount of smoothing.

Though the histogram remains an excellent tool for data presentation, it is worth at least considering the various alternative density estimates that are available (Silverman B.W., 2003).

Let $\{X_1, ...X_n\}$ be a data sample, independent and identically distributed of a continuous random variable X, with density function $f(x)$. The goal is to estimate the pdf $f(x)$ from this sample.

The estimator $\hat{f}(x)$ counts the percentage of observations which are close to the point x. If many observations are near x, then $\hat{f}(x)$ is large. Conversely, if only a few X_i are near x, then $\hat{f}(x)$ is small. The bandwidth h controls the degree of smoothing.

The kernel density estimate is defined by the below equation (Hansen B, 2009)

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

where*n* is the sample size, *K* is a known density function, and *h* is a constant depending upon the size of the sample that controls the amount of smoothing in the estimate. Note that for most standard density functions *K*, where *x* is far in magnitude from any point X_i, the value of *K* (density function) will be very small. The value of the density function is dependent on the data points cluster. Some common kernel functions are display in Table 1 (Hansen B, 2009).

Table 1. Some common Kernel functions

Kernel	$k(u,r)$
Gaussian	$k(u,\infty)=\frac{1}{\sqrt{2\pi}}\exp(\frac{-u^2}{2})1_{]-\infty,\infty[}$
Epanenchnikov	$k(u,2)=\frac{3}{4}(1-u^2)1_{(\lvert u\rvert\le 1)}$
Triweight	$k(u,6)=\frac{35}{32}(1-u^2)^3 1_{(\lvert u\rvert\le 1)}$
Tricube	$k(u,9)=\frac{70}{81}(1-\lvert u\rvert^3)^3 1_{(\lvert u\rvert\le 1)}$
Biweight	$k(u,4)=\frac{15}{16}(1-u^2)^2 1_{(\lvert u\rvert\le 1)}$
Cosine	$k(u,\infty)=\frac{\pi}{4}\cos(\frac{\pi}{2}u)1_{(\lvert u\rvert\le 1)}$

The value of *K* has impact on the value of the bandwidth an on the value of the density estimate. In this paper, the true value of this bandwidth must be estimated, and there are several methods available to optimize this estimate.
These estimated methods are supported by a number of software packages which are often available in academic and other computing environments. As such, reference is made to implementations of these methods in MatLab, S-PLUS and R. Brief descriptions and sources for these and KDE package are provided by Arsalane (Arsalane Ch. G., 2015)

Methods for Selecting Bandwidth

1. *T*

   he Rule of Thumbs and the *Maximal Smoothing Principle* are two methods falling into a category, known like “Quick and dirty” methods. Rule-of-thumb bandwidths are a useful starting point, but they are inflexible and can be far from optimal (Sheather S., 2004).

2. Methods which are falling into the second category, name Plug-in methods, are based on the asymptotically best choice of bandwidth *h* (Woodrofe, 1979). Plug-in methods take the formula for the optimal bandwidth, and replace the unknowns by estimates. But these initial estimates themselves depend on bandwidths. And each situation needs to be individually studied. Plug-in methods have been thoroughly studied for univariate density estimation, but are less well developed for multivariate density estimation and other contexts.
3. Cross-validation method (Arsalane Ch. G., 2015). A flexible and generally applicable data-dependent method is cross-validation. This method attempts to make a direct estimate of the squared error, and pick the bandwidth which minimizes this estimate. There are five known methods.
    a. Unbiased cross-validation bandwidth selection (UCV), proposed by Rudemo, 1982, and Bowman, 1984. It is known like least-squares cross-validation criterion (Scott D. W. and Terrell G. R., 1987).
    b. Biased cross-validation bandwidth selection (BCV), suggested byJones and Kappenman, 1991 (Jones M. C., Marron, J. S. and Sheather S. J., 1996).
    c. Complete cross-validation bandwidth selection (CCV), suggested byJones and Kappenman, 1991.
    d. Modified cross-validation bandwidth selection (MCV), proposed by Stute, 1992 (Turlach B., 2010).
    e. Trimmed cross-validation bandwidth selection (TCV), proposed by Feluch and Koronacki, 1992, a simple modification of the unbiased (least-squares) cross-validation criterion.

Data set information. Experiment Result

We will use a sample retrieved from the database of results in year 2013, in different Universities of Tirana, in order to evaluate the percentage entrancy probability in different studies programs of Universities of Tirana. This sample

has a information for n = 400 admitted students which select in the A2 form, the study programs of this University. It contains the total points in the four exams (GRE). The Kernel density is used to estimate the total point tests distribution.

A comparison of the four midpoint sizes is provided in Figure 1, left.Note that the estimates are piecewise constant and that they are strongly influenced by the choice of bin width, where the number of bin is choise n = 7, 10, 25, 50 respectively. Figure 1, shows one of the problems with using histograms to visualize the density of points in 1D. For example, too small bin-size makes the histogram noisy and hard to interpret (the last one in Figure 1 right) and too large bin-size hide features of the underlying distribution (the first one). We use these bandwidths to build-up a so–called kernel estimator of f (x), Figure 1, right. We note that the curve changes shape, as the value of bandwidth decrease. We shown that students are classify into two groups for example:
In the first group are the students who choose archictecture and in the second group are the students who are classified as admidded at any study program of Universities of Tirana except from architecture program.

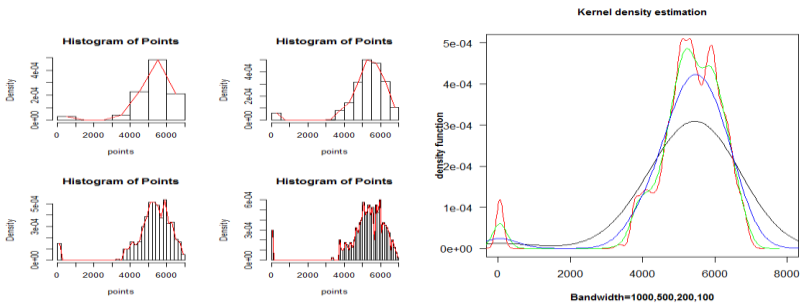

Figure 1. Histograms with different midpoint sizes and a kernel density estimation with different bandwidth.

Figure 2 shows Gaussian kernel density estimates based on six selecting methods, where we have extracted those students who are classified in the first group (admitted in architecture program). The density estimate depicted by different colour in Figure 2 is based on the five bandwidth obtained from cross-validation and one with blue line, from the plug-in methods.. The cross-validation curves are to near each other. The problem is to select the best bandwidth estimating methods and the best kernel. Also, we note that the estimated density that was computed with five different bandwidth captures three to four peaks that characterize the mode (see, grey and red curve), while the estimated density with lower values of bandwidth smoothest out these peaks.

This happens because the outliers at the distribution contribute to small values of *h* be larger than the large bandwidth values *h*. For more details on this estimator (see Silverman B. W., 1986 or Hardle W., 1991).

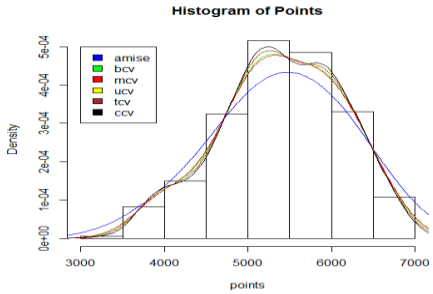

Figure 2. Histogram and Gaussian kernel density with different bandwidth selection.

Figure 3 shows that the best kernel is Gaussian kernel and the best bandwidth estimation methods is unbiased cross validation. From the experiment notes that all the kernel takes the minimal bandwidth with unbiased cross-validation method.

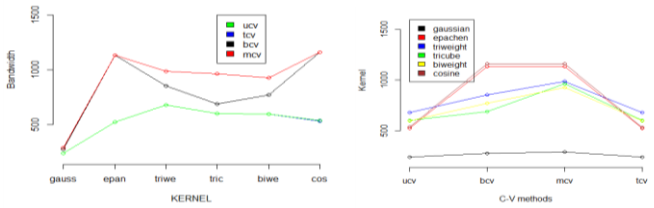

Figure 3. Different bandwidth estimation methods on the right and different kernels on the left

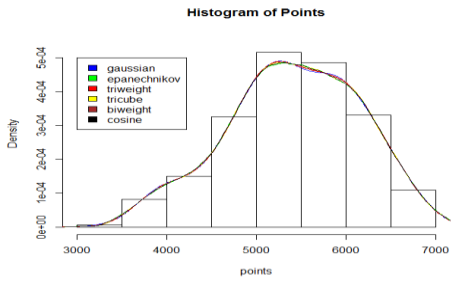

Figure 4. Histogram and some kernel density with best bandwidth methods

Table 2a, shows bandwidth value taken from least squared cross validation method (known as unbiased cross-validation. From the Table 2b, we note that:

- 25% of the students have fewer than 3796 scores and more than 6538 scores.
- A student which has 5167 scores can be on the top of the list of fewer required programs, or at the end of the list of most required programs.
- 5 students in 10 000 can have 5167 scores this because scores are calculating with three digits after the decimal point.

Table 2. Bandwidth value for some Kernel and some numerical characteristics for Gaussian kernel function.

a)

GAUSSIAN	'h' = 234.7107
EPACHENIKOV	h' = 521.6459
TRIWEIGHT	'h' = 676.1259
TRICUBE	'h' = 597.3534
BIWEIGHT	'h' = 594.9082
COSINE	'h' = 533.0223

b)

Eval. Points	**Estimation for fx**
Min: 2425	Min. : 1.500e-09
1st Qu.: 3796	1st Qu.: 5.843e-06
Median : 5167	Median : 1.295e-04
Mean : 5167	Mean : 1.820e-04
3rd Qu.: 6538	3rd Qu.: 3.622e-04
Max. : 7909	Max. : 4.897e-04

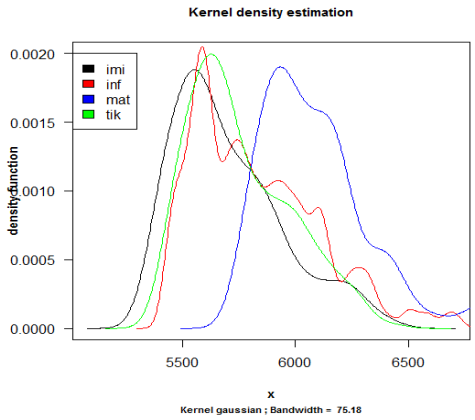

Figure 5. Gaussian kernel density for 4 samples

From Figure 5 we show that the ratio between best mathematics program students and best mathematics and computing engineering program students is the same with the ratio between worst mathematics program students and worst mathematics and computing engineering program students. But mathematics program students are better than mathematics and computing engineering students. The red curve, which display the point distribution of computer science students differ from the others, because in this branch students have maximal scores or minimal scores.

Conclusion

In this work we provide a practical description of density estimation based on kernel methods. Using a kernel density estimate in combination with a histogram adds value to the visual presentation of data. *R* software supports this task by the KDE package. This paper motivates users to familiarize themselves with the theory of kernel density estimation so that they can take control of the graphical output.

Based on this sample, we could say that, unbiased cross-validation give the minimum bandwidth value for all kernel density function, between them the best is Gaussian density estimation.

The examples show that the kernel density estimator is a useful method of representing the overall structure of the data. Some expertise and judgment is required for the selection of an appropriate value of the smoothing parameter *h*. The results show the efficiency of statistical analysis.

Literature

[1] Arsalane Chouaib Guidoum, 2015. Kernel Estimator and Bandwidth Selection for Density and its Derivatives
Department of Probabilities & Statistics. Faculty of Mathematics. University of Science and Technology Houari Boumediene. BP 32 El-Alia, U.S.T.H.B, Algeria
[2] Hansen B., 2009. Kernel Density Estimation
University of Wisconsin ,Spring 2009
[3] Hardle W., 1991. Smoothing Techniques, With Implementations in S, Springer, New York, ISBN 978-1-4612-8768-1
[5] Howitt D. and Cramer D., 2005. Introduction to Research Methods in Psychology. Harlow, Essex: Pearson Education, 354 pages, ISBN 0 131 39984-5.
[6] Jones M. C., Marron J. S. and Sheather S. J., 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc. 91 401–407.
[7] Silverman B.W., 2003. Density Estimation for Statistics and Data Analysis, London, ISBN 0-412-24620-1
[8] Sheather S.,2004. Density Estimation
Statistical Science 2004, Vol. 19, No. 4, 588–597 DOI 10.1214/088342304000000297 © Institute of Mathematical Statistics, 2004
[9] Scott D. W. and Terrell G. R., 1987. Biased and unbiased cross-validation in density estimation. J. Amer. Statist. Assoc. 82 1131–1146.
[10] Turlach B., 2010. Bandwidth Selection in Kernel Density Estimation: A review, Belgium