

# The Study of Different Types of Kernel Density Estimators

MingE Sha<sup>1, a</sup>, Yonggang Xie<sup>2, b</sup>

<sup>1</sup>Department of Mathematics, Kunming University, Kunming, 650214, CHINA

<sup>2</sup>School of Automatic Control and Mechanical Engineering, Kunming University, 650214, CHINA

<sup>a</sup>email: MingE\_Sha@126.com, <sup>b</sup>email:ygxie@kmu.edu.cn

**Keywords:** Kernel Density Estimation (KDE); MATLAB; Probability Density Estimation(PDE); Clustering Algorithm Construction

**Abstract.** One of the most important method of estimating and graphing the underlying density is kernel density estimation (KDE). In this paper, we present basic knowledge of KDE, and simulations were carried out which compare three bandwidth selection methods [Normal rule of thumb (NROT), Least squares cross-validation (LSCV), and Biased cross-validation (BCV)]. Four types of kernel (Standard Normal, Biweight, Laplacian, Rational Quadratic and Circular) are chosen to do the simulation. Results shows that overall LSCV performs best.

## I. Introduction

Density estimation has gained great popularity in recent 20 decades, due to the fact that it is able to apply in many fields, such as archaeology, banking, climatology, economics, genetics, hydrology, physiology and so on[1]. In statistics, normally, there are two kinds of probability density estimation techniques: semi-parametric and non-parametric techniques. Due to the outstanding performances of non-parametric probability density estimation, in this paper, we focus on kernel density estimation (KDE), which is a non-parametric way to gain understanding about the distribution of the data. It is estimating the probability density function from a random sample.

Unlike parameter estimation, such as Likelihoods, non-parameter estimation attempt to estimate the probability density directly from the data without assuming a particular form for the underlying distribution. The first published paper describing non-parameter probability density estimation was by Rosenblatt [1956], where he described the general kernel estimator.

The aim of this paper is to introduce a famous non-parametric estimation, KDE, based on the simulation. The rest of the paper is organized as follows: Section II. provides a description of the histogram estimator. Then, we describe the kernel density estimator and explain its basic properties in section III. In section IV, simulation results are presented, and we conclude this paper in section V.

## II. The Histogram

To understand kernel estimators we first need to understand histograms whose disadvantages provides the motivation for kernel estimators.

The histogram[2] is the most common and simplest form of non-parametric density estimation. It is a graphical way of summarizing and describing a data set. It divides the sample space into a number of bins and approximates the density at the center of each bin by the fraction of points in the training data that fall into the corresponding bin. In this way, a histogram visually conveys the distribution of the samples. The histogram only requires two “parameters”: the bin width and the starting position of the first bin. This is a very simple, easy and effective way to construct probability function; however, there are some drawbacks of using this method to do density estimation.

1. The curve of the probability density function is not smooth, and the discontinuities make it impossible to study the structure of the data.

2. The probability density function is largely affected by the orientation of the bins.

3. It is hard to use histogram in high dimensional situations, since the number of bins grows exponentially with the number of dimensions.

4. Data within histograms depend on the endpoints, which can cause a loss of information.

These difficulties make the histogram might be unsuitable for most of practical applications, such as generating random numbers from a fitted distribution. However, we can alleviate these problems by using kernel density estimators.

### III. Kernel Density Estimation

#### A. Different Types of KDE

Non-parametric density estimation is an important tool in the statistical analysis of data. Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. It is proposed by Rosenblatt (1955) and Emanuel Parzen (1962), and has another name Parzen window. In this section, we give brief introduction of KDE [3].

The general expression for non-parametric density estimation is:

$$P(x) = \frac{k}{NV} \text{ where } \begin{cases} V \text{ is the volume surrounding } x \\ N \text{ is the total number of examples} \\ k \text{ is the number of examples inside } V \end{cases} \quad (1)$$

We choose a fixed value of the volume  $V$  and determine  $k$  from the data. This is the method of KDE. We can see from the equation (1) that when  $N \rightarrow \infty$ , KDE converge to the true probability density, as long as  $V$  decrease with  $N$ , and that  $k$  increase with  $N$  in a proper way.

Specifically, consider the estimator of the following form[4]:

$$\hat{f}(x_0) = \frac{1}{nh} \sum_i K\left(\frac{x_i - x_0}{h}\right) \quad (2)$$

where  $h$ , is the bandwidth or smoothing parameter, the function  $K$ , which controls the weight given to the observation  $\{x_i\}$  at each point  $x_0$ , is called the Kernel, it must satisfy the condition that  $\int K(t)dt = 1$ ,  $\int tK(t)dt = 0$ , and  $0 < \int t^2 K(t)dt < \infty$ , and all odd moments are zero to ensure the estimation in equation (2) is a bona fide density estimate.

Kernel functions must be continuous, symmetric, and most preferably should have a positive (semi-) definite Gram matrix. some popular Kernels are as following:

- The Gaussian Kernel:  $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}};$  (3)

- The Epanechnikov Kernel:  $K(u) = \begin{cases} 3(1-u^2)/4, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases};$  (4)

- The Biweight Kernel:  $K(u) = \begin{cases} 3(1-u^2)^2/4, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$  (5)

In [5], the Gaussian Kernel is defined in 1-D, 2-D and N-D respectively as the following:

$$K_{1D}(x, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, K_{2D}(x, y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, K_{ND}(\vec{x}, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^N} e^{-\frac{|\vec{x}|^2}{2\sigma^2}} \quad (6)$$

where, the  $\sigma$  which is the inner scale, determines the width of the Gaussian Kernel.

- The Laplacian Kernel:  $K(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$  (7)

the Laplacian Kernel is completely equivalent to the exponential kernel, except for being less sensitive for changes in sigma parameter.

- The Rational Quadratic Kernel:  $K(x, y) = 1 - \frac{\|x - y\|^2}{\|x - y\|^2 + c}$  (8)

where the  $c$  is the intercept constant.

- The Circular Kernel: 
$$K(x, y) = \frac{2}{\pi} \arccos\left(-\frac{\|x - y\|}{\sigma}\right) - \frac{2}{\pi} \frac{\|x - y\|}{\sigma} \sqrt{1 - \left(-\frac{\|x - y\|}{\sigma}\right)^2} \quad (9)$$

if  $\|x - y\| < \sigma$ , zero otherwise.

Others kernel functions available from the existing literature are[6]: Linear Kernel, Polynomial Kernel, Gaussian Kernel, Exponential Kernel, Hyperbolic Tangent (Sigmoid) Kernel, Rational Quadratic Kernel, Multiquadric Kernel, Inverse Multiquadric Kernel, Wave Kernel, Power Kernel, Spline Kernel, Bessel Kernel and so on. The choice of a Kernel depends on the problem at hand. For example, a Gaussian Kernel allows us to model feature conjunctions up to the order of the normal distribution.

#### B. Bandwidth Selection Algorithms

Finding the optimal parameters of the KDE is extremely important in order to obtain a good estimate. Although several parameters are involved, the most crucial step is to select a proper bandwidth (smoothing parameter)[7], which controls the amount of smoothing. In this paper, we use four types of selection methods and give the equations as following:

- Normal Rule of Thumb (NROT): 
$$h_{NROT} = \left[ \frac{R(K)}{\mu_2(K)^2 R(f'')} \right]^{1/5} n^{-1/5}, \text{ when a Normal kernel is applied and } f'' = \Phi_{\sigma^2}, h_{NROT} = 1.06n^{-1/5}. [\text{Silverman (1986)}].$$
- Least Squares Cross-Validation (LSCV): 
$$h_{LSCV} = \int \hat{f}(x; h)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(x_i). [\text{Rudemo (1982) and Bowman (1984)}]$$
- Biased Cross-Validation (BCV): 
$$h_{BCV} = \frac{1}{4} h^4 \mu_2(K)^2 \left[ R(\hat{f}''(x; h)) - (nh^5)^{-1} R(K'') \right] + (nh)^{-1} R(K'').$$
 [Scott and Terrell (1987)]

### IV. Estimation and Results

In this section, we discuss different types of Kernels and present the results from simulation figures.

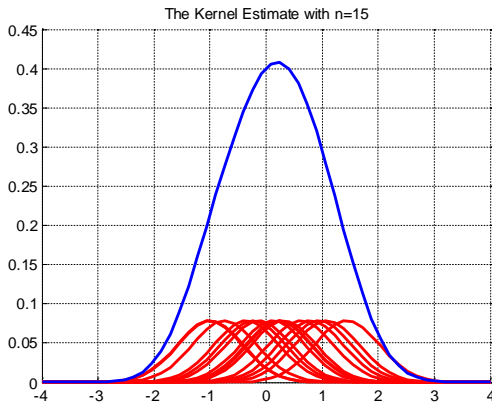


Fig. 1 The Kernel Estimate with n=15

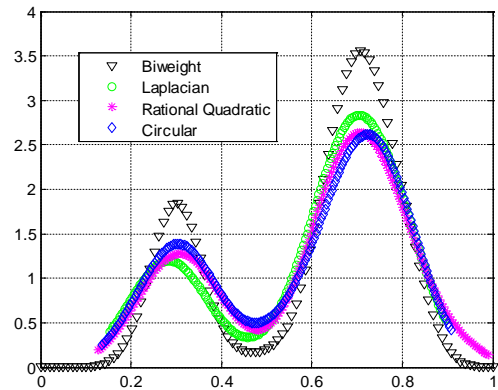


Fig. 2 The Comparison of four types of Kernel

In figure (1), we generate a standard normal kernel (indicated by the solid red lines) on each of the data points  $x_i$ . The 15 individual kernels are summed to obtain the kernel density estimate (the solid blue curve) for n=15 random variables, a weighted kernel is centered at each data point, and then taking the average of them. The data points are the rug plot on the horizontal axis, and the kernel

density estimates converge faster to the true underlying density for continuous random variables[8]. From Figure (1) we could see, there are two ‘bumps’, which we could conclude that there is a higher concentration of smaller densities.

Figure. 2 is the comparison of four types of Kernel, from the figure we can see, the Biweight Kernel has more difference compared with Laplacian, Rational Quadratic and Circular kernel.

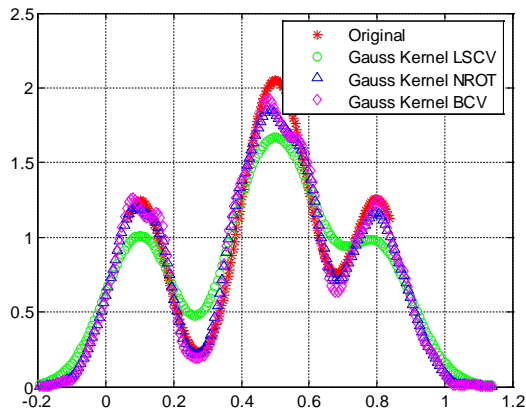


Fig. 3 The Comparison of different Gauss Kernel

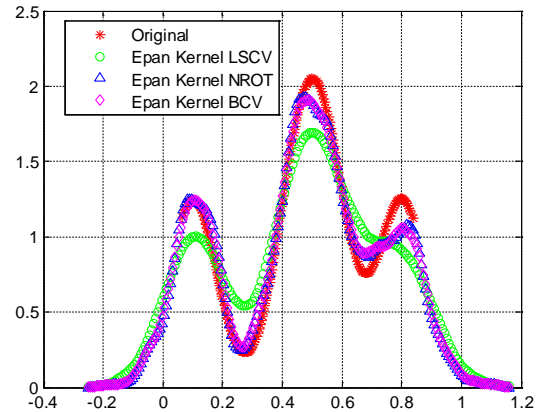


Fig. 4 The Comparison of different Epan Kernel

In figure. 3 and 4, we use different three types of bandwidth selection algorithms: LSCV, NROT and BCV, in Gaussian kernel and Epanechnikov kernel, respectively, to estimate. The results shows that LSCV has better performance at both estimation.

## V. Conclusion

KDE uses local averaging to create a smooth curve from a sample of observations, they have great utility as a tool for identifying outliers and unexpected patterns in the data not apparent from summary statistics. In this paper provides a practical description of KDE, the simulation study also conducted using different kernels and different bandwidth selection algorithms. From the simulation figures, we could see that LSCV has better performance in our examples.

## References

- [1] Simon J. Sheather, “Density Estimation”, Statistical Science 19(4) · November 2004, Vol. 19, No.4, 588-597.
- [2] Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 186: 343-414.
- [3] P Hall , M Wand. “On the accuracy of binned kernel density estimators,” Journal of Multivariate Analysis, 1994 , 56(2):165 -184.
- [4] E. Parzen. On the estimation of a probability density function and the mode. Annals of Mathematical Statistics, 1962.
- [5] Rosenblatt, M. 1956. “Remarks on some nonparametric estimates of a density function,” Annals of Mathematical Statistics, 27: pp. 832-837.
- [6] Hall, P. (1982). “The influence of rounding errors on some nonparametric estimators of a density and its derivatives,” SIAM J. Appl. Math. 42, 390-399.
- [7] <http://crsouza.com/2010/03/kernel-functions-for-machine-learning-applications/>
- [8] Jared K. Harpole, How Bandwidth Selection Algorithms Impact Exploratory Data Analysis Using Kernel Density Estimation, University of Kansas, the degree of Master’ s of Arts, 2013.

[9] [https://en.wikipedia.org/wiki/Kernel\\_density\\_estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation).