International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Choice of the Bandwidth in Kernel Density Estimation

Onrina Chandra

Department of Statistics, University of Calcutta, India, *onrinachandra37[at]gmail.com*

Abstract: ***Given a set of observations, the knowledge of the underlying probability density function that generates the sample is often of interest. Kernel Density Estimation is a nonparametric method used to guess the underlying density function using the sample observations. Although arguably the most popular method of density estimation, KDE is not free from drawbacks. This method of estimation varies greatly with the choice of the smoothing parameter used to estimate the density. This paper gives an overview of the KDE and discusses some statistical properties of the ideal estimator used to guess the unknown density. An outline of some existing methods of choosing a smoothing parameter are discussed. Here we only consider estimation under the univariate setup. The idea of KDE can easily be generalized to a multivariate dataset.***

Keywords: Kernel density estimation, bandwidth, smoothing parameter, kernel, least squares cross validation, mean integrated square error

1. Introduction[1,2,3,4,5,16,17,18]

Given a set of observations, the underlying density function believed to produce the dataset, can be estimated using two approaches. In the parametric approach it is assumed that the sample is drawn from a particular density. The parameters are then estimated on the basis of the sample. Following this, the estimates are plugged in as the values of unknown parameters of the assumed density. However, in real life situations the sample may not always belong to a well-defined family of distributions.

The nonparametric approach paves the way to make inferences regarding the data without making any such rigid assumptions about the underlying density function.

Among all nonparametric methods the Kernel Density Estimation is the most widely used method of estimation.

Suppose we are given a sample of iid random variables $X_1, X_2, \ldots, X_n$ drawn from the unknown probability density f. Using the Kernel Density Estimator (KDE) $\hat{f}$ we approximate f.

The kernel density estimator at point x is denoted by

$$\hat{f}(x,h) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)$$

Where K is the kernel and h>0 is called the bandwidth or smoothing parameter

K is generally a smooth, symmetric function which satisfies $\int k(x)\,dx = 1$.

Intuitively the kernel estimator can be considered to be a sum of bumps (smooth functions) at the observations. K and h smooth each data point Xi into small density bumps .The shape of the bumps depends on K while their width depends on h. Therefore, h controls the amount of smoothing in the estimate of the density. The individual density bumps are then added up to obtain the final density estimate $\hat{f}$.

2. Effect of Kernel and Bandwidth on Estimation[1,2]

Usually a non negative kernel is chosen for estimation which makes both the kernel and the corresponding estimator $\hat{f}$ density functions. However in some situations a kernel may also be chosen which it takes both negative and positive values.

Some common choices of a kernel are given in Table 1

Table 1: Some common choices of the kernel

Kernel	K(t)
Uniform	$\frac{1}{2}I(\lvert t\rvert \leqq 1)$
Triangle	$(1-\lvert t\rvert)I(\lvert t\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-t^2)I(\lvert t\rvert \leq 1)$
Biweight	$\frac{15}{16}(1-t^2)^2 I(\lvert t\rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$

Although the Gaussian kernel is the popular choice, the Epanechikov kernel is the most efficient kernel[1].However under most common choices of the kernel we get a fairly good approximation of the true density.

The quality of the estimator depends more on the choice of the bandwidth than the choice of the kernel . If a very small value of the bandwidth is taken then the features of the data become heightened while a very large value of h results in an overly smooth estimate which obscures prominent features of the dataset as is illustrated by the following example.

Licensed Under Creative Commons Attribution CC BY

3. A Practical Example

We consider the inbuilt dataset faithful from R for the following example. We plot the data faithful$ eruptions and superimpose a density curve. The data plot is shown in Figure1.

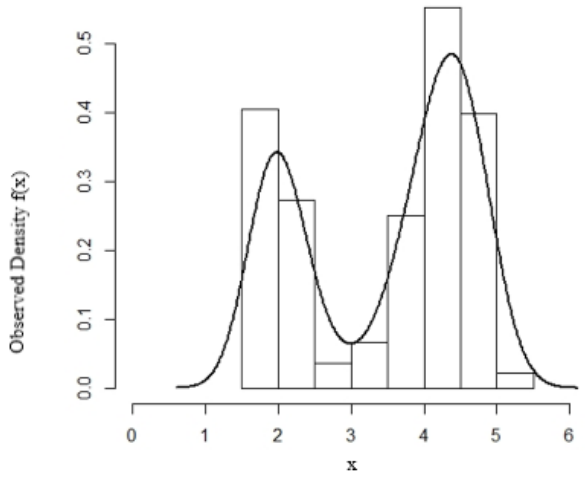

Figure 1: Histogram and Density Plot of the Data set.

From the histogram above it is observed that the dataset is bimodal with maximum density around the values 1.75 and 4.5. We assume each observation is drawn from a Normal Distribution centered around the observation. This is equivalent to assuming a Gaussian Kernel for the data set with mean x_i for the ith observation and the common standard deviation h for all observations. The sum of all Kernels gives the final estimate of the density. This is demonstrated in the following figures. For convenience the kernel density for 6 out of all 272 observations are plotted. The black curves depict the Gaussian kernel centered at the 6 observations. The combination of all 272 of such curves gives the estimate of the density shown in red.

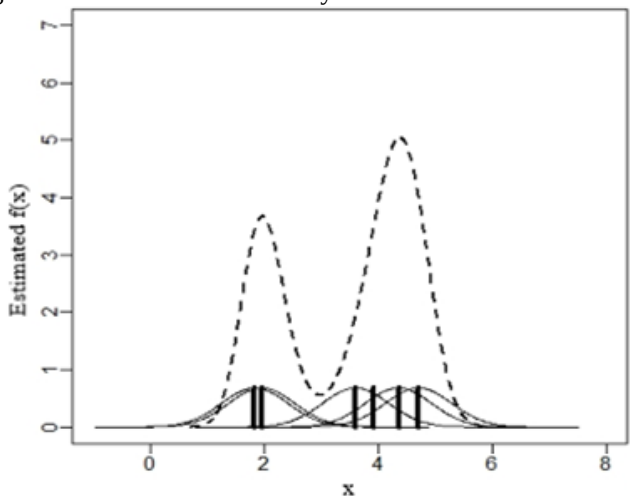

Figure 2: Density Estimation using a Gaussian Kernel with h=0.3

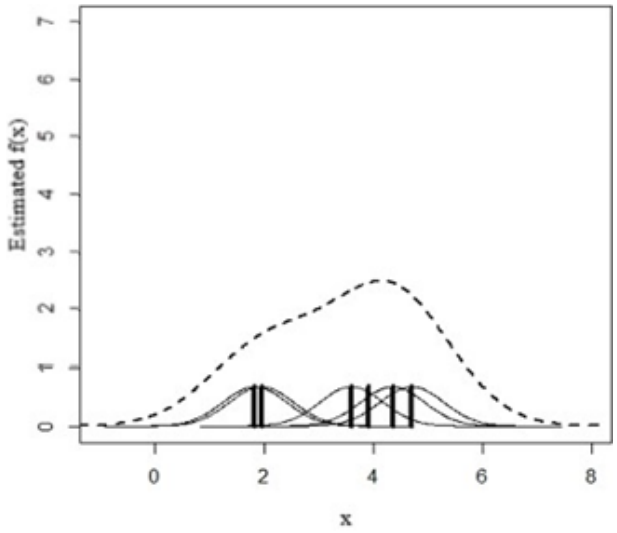

Figure 3: Density Estimation using a Gaussian Kernel with h=1

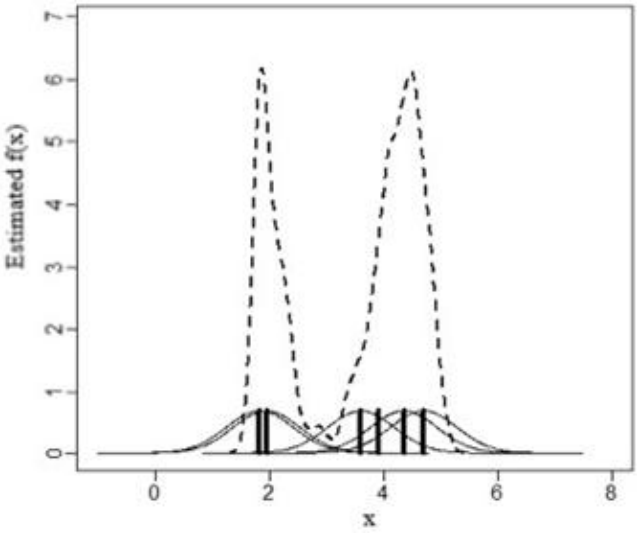

Figure 4: Density Estimation using a Gaussian Kernel with h=0.1

From Figures 2,3,4 it is observed the choice h=0.3 gives a density estimate which is closest to the true underlying density function(Figure-1). A bandwidth value taken too close to zero such as h=0.1 results in a graph with more fluctuations than is present in the actual data while a large bandwidth value h=1 causes a smoothed out graph which obscures the distinct bimodal feature of the unknown density function.

4. Criteria for the Optimal Choice of the Bandwidth [1,2,3,4,5,6,7,8,9]

The most common measure of efficiency of the estimate $\hat{f}$ is given by the mean integrated square error MISE
It is denoted by

$$\mathrm{MISE}\left(\hat{f}\right) = E\int (\{\hat{f}(x) - f(x)\}2)dx$$

The MISE can be seen as the mean squared error MSE used for estimating more than one single value.Not unlike the MSE, MISE too can be expressed in terms of the bias and the variance of $\hat{f}$ as below[1]:

Licensed Under Creative Commons Attribution CC BY

$$\begin{aligned}\mathrm{MISE}\left(\hat{f}\right) &= E\int \{\hat{f}(x) - f(x)\}^2\, dx = \int \mathrm{MSE}_x\, dx \\ &= \int \{E(\hat{f}(x)) - f(x)\}^2\, dx \\ &+ \int E\{\hat{f}(x) - E(\hat{f}(x))\}^2\, dx \\ &= \int bias_x^2(x)\, dx + \int var\hat{f}(x)\end{aligned}$$

The first term of the RHS is called the integrated square bias and the second term is called integrated variance. Substituting the expression of $\hat{f}$ in terms of the kernel we get

$$E\hat{f}(x) = E\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) = \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)\, dy$$

And $$nvar\hat{f}(x) = \int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)\, dy - \left\{\frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy\right\}^2$$

From the above equations we can infer that both the expectation and variance of the estimator depend on h and k. Moreover, the expectation of $\hat{f}$ as seen above is actually a convolution of the kernel K and the unknown density f,h playing the role of the scaling parameter.Hence $\hat{f}$ estimates a true density which has been smoothed out by the kernel.
The exact value of the expectation and variance can be complicated and that's why a common practice is to resort to asymptotic approximations of the MISE.

With $E\hat{f}(x)$ as given above, the $bias_x$ is given by:

$$bias_x = \int \frac{1}{h} f(y)K\left(\frac{x-y}{n}\right) dy - f(x)$$

We make the change of variable $y = x - hs$. Assuming k is symmetric about zero and $\int s^2K(s)\, d(s) = u_k$, as k is a density which integrates to 1,we get

$$\begin{aligned}Bias_x &= \int f(x - hs)k(s)\, ds - f(x) \\ &= \int \{f(x - hs) - f(x)\}k(s)\, ds \\ &= -hf'(x)\int sK(s)\, d(s) + \frac{1}{2}h^2 f''(x) \int s^2K(s)\, d(s) + \cdots \\ &= \frac{1}{2}h^2 f''(x)u_k + o(h^2)\end{aligned}$$

As seen above as $h \to 0$, the bias decreases at a rate $o(h^2)$.

The above equation expresses the bias of the KDE as a function of the curvature of f, as denoted by $f''(x)$.

Therefore we conclude that the bias will be very large if the curve changes frequently. This is a reasonable interpretation since the KDE, as mentioned above, provides a smoothed out version of the unknown density. Hence the bias in the estimate will increase with the rapidity of changes of the curve.

Integrating over the range of x,

$$\int bias_x^2(x)\, dx \approx \frac{1}{4}h^4 u_k{}^2 \int f''(x)dx$$

Similarly transforming y to $x - hs$

$$var\hat{f}(x) \approx \frac{1}{nh}\{f(x) - hsf'(x) +$$

…K2s+O1n≈1nhfx∫K2s𝑑s [1-2]

Integrating over the range of x, as $\int f(x)\, dx = 1$

$$\int var\hat{f}(x) \approx \frac{1}{nh}\int K^2(s)\, ds$$

The optimal choice of h is one which reduces the Mean Integrated Square Error. From the approximations above, we conclude that a small value of h reduces the bias while a large value of h reduces the variance. As the variance increases, so does the error in estimation and we get spurious and highly fluctuating estimates of the density. On the other hand as bias increases we get an estimate which may smooth out important features of the density .To obtain the minimum MISE a compromise is needed in between lower bias and lower variance.

Using the above equations, the MISE is can be simplified as,

$$\mathrm{MISE}\left(\hat{f}\right) = \frac{1}{nh}\int K^2(s)\, ds + \frac{1}{4}h^4 u_k{}^2 \int f''(x)dx + O\left(\frac{1}{n}\right) + o(h^4)$$

The dominating part of the MISE is given by $\frac{1}{nh}\int K^2(s)\, ds + \frac{1}{4}h^4 u_k{}^2 \int f''(x)dx$ which is called the Asymptotic Mean Squared Error or AMISE.

$$\mathrm{AMISE}\left(\hat{f}\right) = \frac{1}{nh}\int K^2(s)\, ds + \frac{1}{4}h^4 u_k{}^2 \int f''(x)dx$$

The optimum value of h, h_0 is obtained by differentiating $\mathrm{AMISE}\left(\hat{f}\right)$ and is obtained as

$$h_0 = u_k^{-2/5} n^{-1/5}\left\{\int K^2(s)\, ds\right\}^{1/5} \left\{\int \left(f''(x)\right)^2\, dx\right\}^{-1/5}$$

As seen above, h itself is inversely proportional to a positive power of $\left(f''(x)\right)^2$,which denotes the change of slope of the unknown density. Hence if the true density is expected to vary rapidly a small value of h should be chosen for estimation. [1-2]

5. Selection of Optimal Bandwidth [9,10,11,12,13,14,15]

An appropriate choice of h is imperative for a good estimation of the unknown density. As the AMISE depends directly on the unknown density it cannot be readily used in practice to obtain the optimal bandwidth.

One way out would be to plug in an estimate of $\left(f''(x)\right)^2$ from some standard distribution which is discussed below. This method works well for unimodal densities but tends to oversmooth the data in multimodal cases. Therefore, it becomes crucial to look for other methods of obtaining an optimum value. Overviews of a few common methods are given below:

6. Reference to a Standard Distribution

A simple way of obtaining an optimal bandwidth would be to estimate $\left(f''(x)\right)^2$ using some standard family of distributions. The most common approach is to assume f follows a $N(\mu,\sigma^2)$ distribution. Using this estimate of f and taking a Gaussian kernel we get
$h_0 = 1.06n^{-1/5}\,\sigma$.

Licensed Under Creative Commons Attribution CC BY

Here sigma is estimated by sample standard deviation usually. This substitution of σ works well for univariate distributions. However, it provides smoothed out estimates in case of bimodal distributions. A greater drawback is that this procedure gives inaccurate estimates if the true density is long tailed and skewed. An alternative which is more robust to outliers is the Interquartile Range R. Under the normality assumption, using R the optimum value of h is obtained as

$$h_0 = 0.79Rn^{-1/5}$$

Being robust $\hat{R}$ may be considered ideal for the univariate case but it fails to estimate multimodal distributions as accurately as $\hat{\sigma}$.

A compromise between both the above procedures in to take $h_0 = 1.06\min\left(\hat{\sigma}, \hat{R}/1.34\right)n^{-1/5}$ as an estimate of optimal h.

This works fairly well in case of both multimodal distributions and skewed distributions.

Least Squares Cross Validation

Here, we minimize the Integrated Square Error(ISE) instead of the AMISE .We first use the sample to calculate KDE and then we use it again to validate how well the obtained KDE estimates f.

The Integrated Square Error (ISE) is given by

$$ISE = \int \{\hat{f}(x) - f(x)\}^2 dx = \int \hat{f}^2(x)\,dx - 2\int \hat{f}(x)\,f(x)\,dx + \int f^2(x)\,dx$$

Since the first term of the RHS is independent of h minimizing ISE would require the minimization of the last two terms.

$$D = \int \hat{f}^2(x)\,dx - 2\int \hat{f}(x)\,f(x)\,dx$$

The basic idea behind cross validation is to obtain a value of D on the basis of the sample. The optimal h is the value for which this estimate $\hat{D}(h)$ is minimized. $\int \hat{f}(x)\,dx$ is estimated from the estimate of $\hat{f}$.

We define $\hat{f}_{-j}$ as the estimate of density obtained from all density points except X_j .

In mathematical terms,

$$\hat{f}_{-j}(x) = \frac{1}{(n-1)h} \sum_{y \neq j}^{K} \left(\frac{x - x_y}{h}\right)$$

We minimize the least square cross validation function

$$M(h) = \int \hat{f}^2 - \frac{2}{n}\sum_{j} \hat{f}_{-j}(x_j)$$

to obtain an optimal value of the bandwidth.

It is important to note $E\{(M(h)) + \int f^2\,dx\}$ for all h equals the MISE. Thus minimizing $E(M(h))$ brings us back to the obtaining an unbiased estimate of the MISE. An advantage of this method is that it is asymptotically optimal. The calculation of M(h) is not very simple and often calls for numerical methods .

An offshoot of the least squares cross validation method is the likelihood cross validation method which minimizes the function $C(h) = \frac{1}{n}\sum_{j=1}^{n} \log \hat{f}_{-j}(X_j)$ to obtain optimal h.

The Test Graph Method

Assuming kernel k is symmetric and twice differentiable under certain regularity conditions the best possible h is the value of h which results in the most rapid convergence of $\sup|\hat{f} - f|$ to zero.This ensures that the estimate of the density uniformly converges to true density implying the estimates of the $f''(x)$ would be close to the true value.

For the optimum h we would get

$$\frac{\sup|\hat{f}'' - Ef''|}{\sup|Ef''|} \to P \quad \text{where P is a function of the kernel only.}$$

The numerator depends on the random error (fluctuations) in estimation of $\hat{Ef}''$ while the denominator depends on the trend of $\hat{f}''$ .Thus for good estimation it is expected that the random fluctuation will be much less than the trend. Therefore to obtain the optimal value of h test graphs of the function $\hat{f}''$ for various values of h are drawn. The optimal h should be the one which represents a graph that despite having random fluctuation has a clear trend.

Apart from the above mentioned three, there are a variety of methods of bandwidth selection but no method is universally accepted.

7. Conclusion[1,2,3,4,5,6,7,8]

Despite the wide applicability of Kernel Density Estimation there are many issues regarding its practical performance. Bandwidth selection forms the chief issue in the framework of Kernel Density Estimation. While the optimal bandwidth provides estimates which are very close to the true density, a bandwidth selected without proper consideration may provide crude estimates. The selection of bandwidth depends on the purpose for which estimation is needed, whether a general idea of the true density is required, in which case a reference to standard method would be sufficient; or whether the data is to be studied to make inferences regarding the population, for which more complicated methods such as cross validation is used. Each method has its own set of advantages and for a large sample possesses asymptotic properties which provide estimates close enough to the true density in the long run.

Licensed Under Creative Commons Attribution CC BY

References

[1] Silverman, B. W. (1986). Density estimation for statistics and data analysis ,CRC press, pp 1-72.ISBN:0-412-24620-1
[2] Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: A review. *In CORE and Institut de Statistique*,38(3), 1-33
[3] Devroye, L., & Gyorfi, L. (1985). Nonparametric density estimation:The L1 view ,John Wiley, New York, pp. 191-232 .ISBN: 0471-81646-9
[4] Sheather, S. J. (2004). Density estimation. *Statistical science*, 19(4), 588-597
[5] Marron, J. S.; Wand, M. P. (1992). Exact Mean Integrated Squared Error,*Ann. Statist.*, 20(2), 712-736.
[6] Loader, C. (1999). Bandwidth Selection: Classical or Plug-In? *The Annals of Statistics*, 27(2), 415-438.
[7] Wand, M. P., & Jones, M. C. (1994). Kernel smoothing,Crc Press, pp 1-85.ISBN:0-412-55270-1,
[8] Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9(2), 65-78.
[9] Jones, M. C., & Kappenman, R. F. (1992). On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*,19(4), 337-349.
[10] Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91(433), 401-407.
[11] Stone, C. (1984). An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *The Annals of Statistics*, 12(4), 1285-1297.
[12] Jones, M. C. (1991). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters*, 12(1) ,51–56.
[13] Devroye, L., & Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimates. *The Annals of Statistics*,24(6), 2499-2512.
[14] Yen-Chi Chen (2017).A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1), 161-187
[15] Chu, C. Y., Henderson, D. J., & Parmeter, C. F. (2015). Plug-in bandwidth selection for kernel density estimation with discrete data. *Econometrics*, 3(2), 199-214.
[16] Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. *Information Sciences*, 512, 1192-1201.
[17] Duong, T. (2020). ks: Kernel density estimation for bivariate data, University of Western Australia, Australia
[18] Devroye, Luc. (1987),A Course in Density Estimation. Progress in Probability and Statistics, Birkhäuser, Boston ,pp 1-35. ISBN:978-0-8176-3365-3

Licensed Under Creative Commons Attribution CC BY