European Journal of Statistics and Probability
Vol.8, No.2, pp, 1-13, September 2020
Published by ECRTD-UK
Print ISSN: 2055-0154(Print), Online ISSN 2055-0162(Online)

ON THE BANDWIDTH SELECTION

Langat Reuben Cheruiyot
University of Kabianga, P. O. BOX 2030-20200, KERICHO
School of Science & Technology
Department of Mathematics and computer Sciences

ABSTRACT: *One of the key parameters in density and regression estimation is the bandwidth. This has variously been termed as kernel width or window by various authors. It is a smoothing parameter that determines the amount of data that falls within it and therefore the amount of information that will be used to do the estimation. Under ideal situations it would be expected that there would be a bandwidth selector that does result in estimates with huge biases or variances. Unfortunately this is not the case as small bandwidths reduce the bias at the expense of huge variance while large ones has a desirable variance but unacceptably high bias. This study explores this important parameter, its optimality and influence on density and regression estimation techniques.*

KEY WORDS*: kernel function, mean integrated square error, bias-variance trade-off.*

INTRODUCTION

Density and regression estimation are essential in statistical inferences. Smoothing methods provide a powerful methodology for gaining insights into data, Jones et al (1996). This fact attests itself also in the works of (Nakarmi, 2016) and (Chen, 2018). Graphs of such estimators vary on their wiggliness because of the chosen bandwidth as well as the kernel function under use. This may have effect on the interpretation to be made. Bandwidth selection is an area of concern in kernel density estimation. These bandwidths vary with the kernel function chosen. An optimal bandwidth of one kernel function cannot be regarded as optimal for another function. Because of this many researchers have been carrying out studies aimed at determining techniques of obtaining bandwidths that minimize Mean Integrated Square Error (MISE) or Asymptotic Mean Integrated Square (AMISE) functions.

This paper has been organized as follows. In section 2, we give a brief review of literature and section 3 gives the derivation of the optimal bandwidth as well as highlighting the other bandwidths that have been in use. Comparison and discussion of the different bandwidths have been done in section 4, while conclusion has been done in section 5.

REVIEW OF LITERATURE

Estimations are considered robust if they are less sensitive to model misspecifications. The misspecifications are common with the parametric modeling. To alleviate this problem nonparametric estimation comes in handy. This type of estimation has a number of desirable characteristics. As detailed in (Hardle, 1994), it provides a versatile method of exploring the general relationship between two variables; secondly it enables one to make prediction of observations without any reference to a fixed parametric model; thirdly it is a tool for finding spurious observations by studying influence of isolated points and lastly it is a flexible method for interpolating between adjacent values of the auxiliary variable. Obviously such an approach is inevitably compelling to any researcher. Other researchers who have made contribution in the popular approach include (Hardle, 2005, Takezawa, 2006 and Tsybakov, 2009).

Nonparametric estimation can be achieved in a number of ways. These include kernel based estimation, splines and wavelets. How to apply each of these is beyond the scope of this paper. For those researchers who are interested there is vast literature available. One may see for example (Gramacki A., 2018, Ghosh S. 2018 and Blackburn et al, 2014). The focus of this paper is on the effect of one key parameter in density or regression estimation called the bandwidth. It is the parameters that play a key role in the roughness or the smoothness of a curve. Heidenreich et al (2013) has reviewed three automatic bandwidth selectors in their paper. They found out that simple plug-in and cross-validation methods produces bandwidths with a quite stable performance. In the following section we derive the optimal bandwidth and compare this with other existing selectors in the subsequent sections.

Bandwidths

Optimal bandwidth

The *MISE* which brings together the variance and the bias of an estimator is normally used to gauge the performance of an estimator.

Suppose we want to estimate the population total, *T*, using its non-parametric estimator, $\hat{T}_{np}$. Then the MISE of this estimator is given by:

$$MISE\left(\hat{T}_{np}\right) = E\left(\hat{T}_{np} - T\right)^2$$

$$= E\left(\hat{T}_{np} - E[\hat{T}_{np}] + E[\hat{T}_{np}] - T\right)^2$$

$$= E\left(\hat{T}_{np} - E[\hat{T}_{np}]\right)^2 + E\left(E[\hat{T}_{np}] - T\right)^2$$

$$+ 2E\left(\hat{T}_{np} - E[\hat{T}_{np}]\right)\left(E[\hat{T}_{np}] - T\right)$$

$$= Var\left(\hat{T}_{np}\right) + Bias^2\left(\hat{T}_{np}\right) + 0 \quad (1)$$

The two components here in equation (1) are the variance and the bias. A good estimator would be the one with the least amount of *MISE*. This implies small amounts of variance and bias respectively. One way of achieving this would be on checking for the parameter that controls the two components and adjusting it so that the least *MISE* is realized. This parameter is bandwidth size used in the kernel function. Suppose that our population estimator, $\hat{T}_{np}$, is given by:

$$\hat{T}_{np} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) \quad (2)$$

The bias is then given by:

$$Bias\left(\hat{T}_{np}\right) = E\left(\hat{T}_{np} - T\right)$$

$$= E\left(\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i)\right] - \left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} y_i\right]\right)$$

$$= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} y_i\right)$$

$$= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} m(x_i)\right)$$

$$Bias\left(\hat{T}_{np}\right) = E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} m(x)\right) \quad (3)$$

Let $K(.)$ denote a kernel function which is also twice continuously differentiable, such that:

$$\text{(a) } \int K(z)dz = 1 \quad \text{(b) } \int zK(z)dz = 0 \quad \text{(c) } \int z^2K(z)dz := K_2(K) \quad (4)$$

A non-parametric model is conventionally of the form

$$Y_i = m(X_i) + e_i \quad (5)$$

where Y_i- is the variable of interest

X_i-is the auxiliary variable

m-is an unknown function to be determined using sample data

e_i-is error term-assumed to be $N(0, \sigma^2)$

$i=1, 2, \ldots, n$

MISE should be looked at as a measure that takes *MSE* into account in a global manner. To measure precision globally we integrate the *MSE* to obtain the *MISE*. This will give cumulative error along the entire line of *m(x)* when estimated using $\hat{m}(x)$. The integral value of $MSE(\hat{m}(x))$ is given by:

$$MISE(\hat{m}(x)) = \int_{-\infty}^{\infty} MSE(\hat{m}(x))dx \tag{6}$$

It can be shown that this gives the following result.

$$MISE(\hat{m}(x)) = \int_{-\infty}^{\infty} \frac{1}{nh}\left(m(x)R(K)\right)dx + \int_{-\infty}^{\infty} \frac{1}{4} m''(x)^2 h^4 K_2{}^2(K)dx + o(h^4) + o\left(\frac{1}{nh}\right)$$

where $R(K)$ is the roughness of the kernel.

The approximate value of this, however, is the *AMISE* given by:

$$\begin{aligned} AMISE &= \int_{-\infty}^{\infty} \left(\frac{1}{nh}\left(m(x)R(K)\right) + \frac{1}{4} m''(x)^2 h^4 K_2{}^2(K) \right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{nh} R(K)m(x)dx + \int_{-\infty}^{\infty} \frac{1}{4} h^4 K_2{}^2(K) m''(x)^2 dx \\ &= \frac{R(K)}{nh} + \frac{1}{4} h^4 K_2^2(K) \int_{-\infty}^{\infty} m''(x)^2 dx \end{aligned} \tag{7}$$

It is easy to see from (7) that *AMISE* changes as a function of the bandwidth *h*, i.e small values of *h* makes the first term in (7) become large but at the same time it makes the second term small. Conversely, however, as *h* gets larger, the second term in (7) increases as the first term decreases. This, obviously, requires a balance. See Fig. 1 for an insight of this effect. It is therefore necessary to obtain an optimal value of *h* which minimizes *AMISE*. This, seemingly, would help us to address the problem arising as a result of bias-variance trade-off away from the boundary and to some extent at the boundary.

European Journal of Statistics and Probability
Vol.8, No.2, pp, 1-13, September 2020
Published by ECRTD-UK
Print ISSN: 2055-0154(Print), Online ISSN 2055-0162(Online)

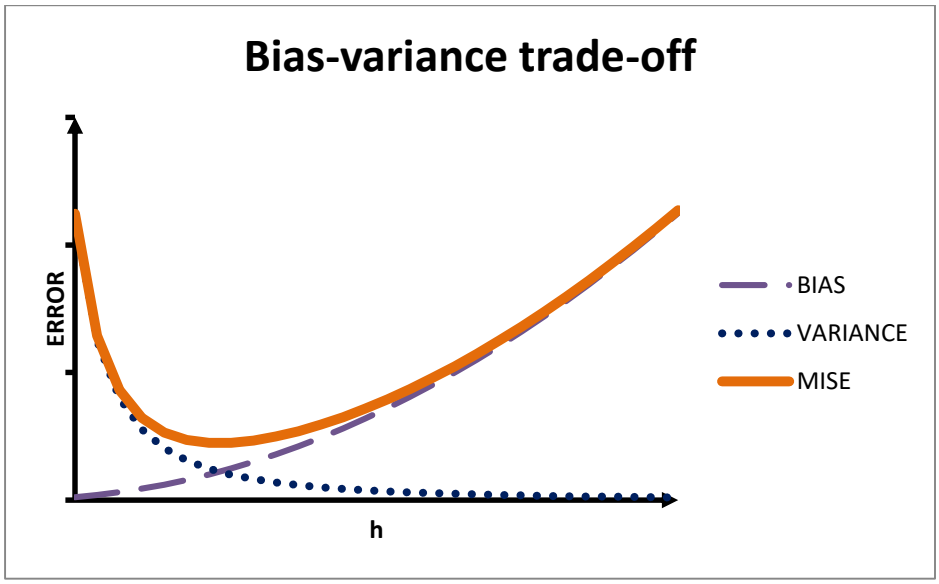

Fig. 1. The compromise between the bias and the variance relative to the bandwidth size, *h*.

We can find the expression for an optimal bandwidth by minimizing (7) with respect to *h*. The first derivative is given by:

$$\frac{dAMISE(\hat{m}(x))}{dh} = \frac{d}{dh}\left(\frac{R(K)}{nh} + \frac{1}{4}h^4 K_2^2(K)\int_{-\infty}^{\infty} m''(x)^2\,dx\right)$$

$$= h^3 K_2^2(K)\int_{-\infty}^{\infty} m''(x)^2\,dx - \frac{R(K)}{nh^2}$$

Equating this to zero and solving for *h* yields an optimal bandwidth, h_{opt}, for a given p.d.f and kernel:

$$h_{opt} = \left(\frac{R(K)}{nK_2^2(K)\int_{-\infty}^{\infty} m''(x)^2\,dx}\right)^{-\frac{1}{5}} \tag{8}$$

When h_{opt} is substituted in (7), we obtain the minimal *MISE* for the given *p.d.f* and kernel. With some little algebraic manipulation, (7) can be shown to be:

$$AMISE_{opt}\left(\hat{m}(x)\right) = \frac{5}{4}\left(\frac{\int m''(x)^2\,dx\left(\int K(z)^2\,dz\right)^4 K_2^2(K)}{n^4}\right)^{\frac{1}{5}} \tag{9}$$

It should be noted that h_{opt} in (8) above is influenced by the size of the sample, *n*, and the kernel *K* through the unknown function $m''(x)^2$. This poses a challenge and causes the inapplicability of expression (8) in practice. We highlight two common ways in which this problem can be tackled- the "plug-in" method and the cross validation method. The plug-in method simply involves the replacement of the unknown functions in the expression of interest. This is discussed in the next section.

Silverman's rule of Thumb

From the expression in (8), the idea is to replace the unknown function given by $m''(x)^2$ by an estimate. Silverman's rule of thumb computes this second derivative as if the function were the density of the normal distribution, $N(\mu,\sigma^2)$. This gives:

$$R(k) = \sigma^{-5}\int m''(x)^2\,dx = \frac{3}{8\sqrt{\pi}}\sigma^{-5} \tag{10}$$

$$K_2^2(K) = 1 \tag{11}$$

and

$$\int K(z)^2\,dz = \frac{1}{2\sqrt{\pi}} \tag{12}$$

The optimal bandwidth would therefore be given by:

$$h_{opt} = \left(\frac{1}{2\sqrt{\pi}}\frac{8\sqrt{\pi}}{3n}\sigma^5\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-\frac{1}{5}} \tag{13}$$

The standard deviation, σ, still unknown, may be estimated by its value, $\hat{\sigma}$, from the sample. This is given by:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{14}$$

Substituting this in (13) results in:

$$h_{opt} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \tag{15}$$

It has been noticed from previous researches that the value of the bandwidth obtained using expression (15) is sensitive to outliers. A single outlier may cause too large estimate of σ and hence implies a too large bandwidth, (Härdle, 2005). A more robust estimator can be obtained using the inter-quartile range calculated as follows.

$$R = X_{[0.75n]} - X_{[0.25n]} \tag{16}$$

Maintaining the assumption of normality in the true *p.d.f* i.e. $X \sim N(\mu,\sigma^2)$ and therefore the standard normal $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, we can then proceed as below:

$$\begin{aligned} R &= X_{[0.75n]} - X_{[0.25n]} \\ &= (\mu + \sigma Z_{[0.75]}) - (\mu + \sigma Z_{[0.25]}) \\ &= \sigma(Z_{[0.75]} - Z_{[0.25]}) \\ &\approx \sigma(0.67 - (-0.67)) \\ &= 1.34\sigma \end{aligned}$$

$$\Rightarrow \hat{\sigma} = \frac{R}{1.34} \tag{17}$$

This expression can thus be substituted for $\hat{\sigma}$ in (15) to give:

$$h_{rot} = 1.06\left(\frac{R}{1.34}\right)n^{-\frac{1}{5}} \approx 0.79Rn^{-\frac{1}{5}} \tag{18}$$

A 'better rule-of-thumb' bandwidth can be obtained by combining the expressions in (15) and (18). The relation obtained is:

$$h_{rot} = 1.06 \min\left\{\hat{\sigma}, \frac{R}{1.34}\right\}n^{-\frac{1}{5}} \tag{19}$$

Both expressions in (15) and (19) have been known to work well when the true densities are or are near normal. For those distributions that are substantially different from the normal, bandwidths obtained this way may not be satisfactory.

Least Squares Cross Validation Technique

This selection criterion is attributed to Rudemo (1982) and Bowman (1984). *MISE* can be decomposed as follows:

$$MISE\left(\hat{m}(x)\right) = E\left(\int\left[\hat{m}(x) - m(x)\right]^2 dx\right) \tag{20}$$

$$MISE\left(\hat{m}(x)\right) = \int \hat{m}(x)^2 dx - 2\int \hat{m}(x)m(x)dx + \int m(x)^2 dx$$

Since the third term does not depend on *h*, we can rewrite the expression as follows.

$$MISE\left(\hat{m}(x)\right)-\int m(x)^2\,dx=\int \hat{m}(x)^2\,dx-2\int \hat{m}(x)m(x)dx \tag{21}$$

This means that minimizing *MISE* is equivalent to minimizing the expression on the right hand side. The first term can be calculated directly from the data. This therefore leaves us with the second term which depends on the bandwidth, *h*, and the unknown function *m*(*x*). On a closer look at the term $\int \hat{m}(x)m(x)dx$ it can be noted that this is the expected value of $\hat{m}(X)$ where the expectation has been obtained *w.r.t.* a random variable X . This expected value can be estimated by:

$$E\left[\hat{m}(X)\right]=\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{-i}(X_i) \tag{22}$$

where

$$\hat{m}_{-i}(x)=\frac{1}{(n-1)h}\sum_{j=1,i\neq j}^{n}K\left(\frac{x-X_j}{h}\right) \tag{23}$$

is the leave-one-out estimator. This is so because the i^{th} observation is not used in the calculation of the estimator in expression (23). From (21) we can therefore have the cross validation criterion.

$$CV(h)=\int \hat{m}(x)^2\,dx-\frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}K\left(\frac{X_i-X_j}{h}\right) \tag{24}$$

But as we said earlier calculation for the integral part from the data can be obtained using the relation:

$$\int \hat{m}(x)^2\,dx=\frac{1}{n^2h}\sum_i\sum_j K*K\left(\frac{X_j-X_i}{h}\right) \tag{25}$$

where *K*K(.)* is the convolution of *K.*

Thus (24) become:

$$CV(h)=\frac{1}{n^2h}\sum_i\sum_j K*K\left(\frac{X_j-X_i}{h}\right)-\frac{2}{(n-1)h}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}K\left(\frac{X_i-X_j}{h}\right) \tag{26}$$

The value h_{cv} that minimizes this expression (26) is taken as the bandwidth.

It should be noted that the cross validation technique automatically adapts to the smoothness of the function *m*(*x*) and does not involve any assumptions of the unknown density, Hardle, (2005, p82). It is also asymptotically optimal, Stone, (1984). Even with these advantages this technique seems to pay the fairly small biases with bigger variance. The main drawback of cross-validation is that the resulting kernel density estimates tend to be highly variable and undersmooths the data, Peracchi (2011, p27).

Biased Cross Validation Technique

This technique of selecting the bandwidth was proposed by Scott and Terrell (1987). It tries to minimize *AMISE*. From (7), we have:

$$AMISE(\hat{m}(x)) = \frac{R(K)}{nh} + \frac{1}{4}h^4 K_2^2(K)\int_{-\infty}^{\infty} m''(x)^2\,dx$$

$\int_{-\infty}^{\infty} m''(x)^2\,dx$ can be estimated by:

$$\frac{1}{n^2 h}\sum_{1\neq j}^{n} K * K\left(\frac{X_i - X_j}{h}\right) \tag{27}$$

By substituting this expression into the *AMISE* function, we obtain the following *BCV*(*h*) i.e.

$$BCV(h) = \frac{R(K)}{nh} + \frac{1}{4}h^4 K_2^2(K).\frac{1}{n^2 h}\sum_{1\neq j}^{n} K'' * K''\left(\frac{X_i - X_j}{h}\right) \tag{28}$$

The estimate of the bandwidth, $\hat{h}_{BCV}$, is chosen by minimizing (27). The summary of these selectors have been given in table 1.

Table 1: Summary of bandwidth selectors

Notations in this paper	**Notations in R**	**The Bandwidth**
h_{opt}	nrd_0	The normal rule of thumb
h_{rot}	*Nrd*	Silverman's rule of thumb
h_{SJ}	*SJ*	Sheather-Jones
h_{LSCV}	*Ucv*	Least Squares /Unbiased cross-validation
h_{BCV}	*Bcv*	Biased cross validation

Varied Bandwidths and effect on the bias and the variance

In this section we assess the bandwidth effect as a smoothing parameter on estimation. To assess the effect of each bandwidth selector is not easy though. This is because there is still another parameter at play in kernel based estimation namely the kernel function. To enable as make progress in the study, the same kernel function was thereafter used to attempt to portray the bias variance trade-off. The graphs were generated using bandwidths that were varied from small ones to large ones. To point out this trade-off it will require someone with vast knowledge of the theoretical derivations made about the variance and bias. The following figures (2 - 6) show the different default bandwidths for a random sample of size *n*= 100, as used with Gaussian kernel function. Since the eye may occasionally be used to gauge the suitability of the bandwidth, four different ones were tried arbitrarily, see figures 7 & 8 for this. In many situations, it is sufficient

to subjectively choose the smoothing parameter by looking at the density estimates produced by a range of bandwidths. One can start with a large bandwidth, and decrease the amount of smoothing until reaching a ”reasonable”density estimate, Zambom and Dias (2012). The sample taken was from the faithful dataset within R.

CONCLUSION

There are many ways of bandwidth selection. A detailed study by Jones et al, (1996) gave results that have been summarized as follows:- Silverman's, h_{rot} (“Rule of Thumb”) oversmooths too often while h_{BCV} has the same tendency and is instable as well. The h_{LSCV} has an unacceptable spread and is often in the direction of undersmoothing. Sheather-Jones-solve the equation technique, h_{SJ} is a useful compromise between h_{rot} and h_{LSCV} and performs acceptably in harder to estimate densities as well.

As noted earlier larger bandwidths give smoother curves while the opposite is true for smaller ones which produce rather wiggly curves- a reality that reinforces the reasons for naming the bandwidth as a smoothing /tuning parameter. This effect was also noticeable in the figures whose smoothing parameters were taken arbitrarily.

It is also worth noting that the difference between the selectors is minimal in terms of their effects especially when default ones are taken. This can be attributed to the fact the R software adjusts it to an optimal value during the application. The effect of a kernel function is very mild hence one can choose any. Since the focus was on the bandwidth the Gaussian function was used in this study.

Within the samples taken in our study we can conclude that there was no much variation resulting from the change in selectors as seen in the figures produced.

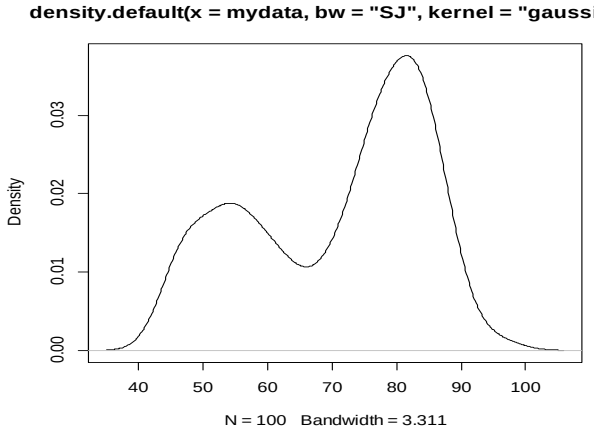

Fig. 2 SJ –Default Bandwidth

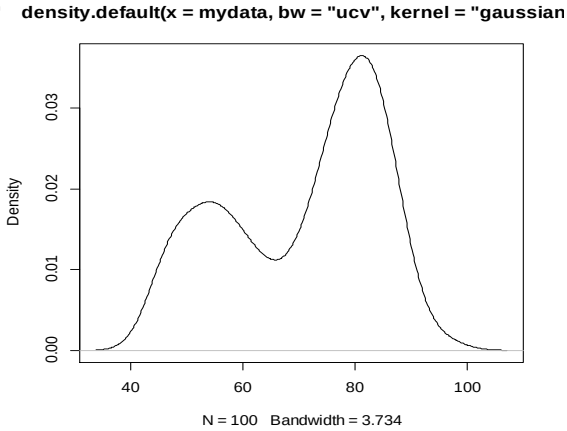

Fig. 3 ucv –Default Bandwidth

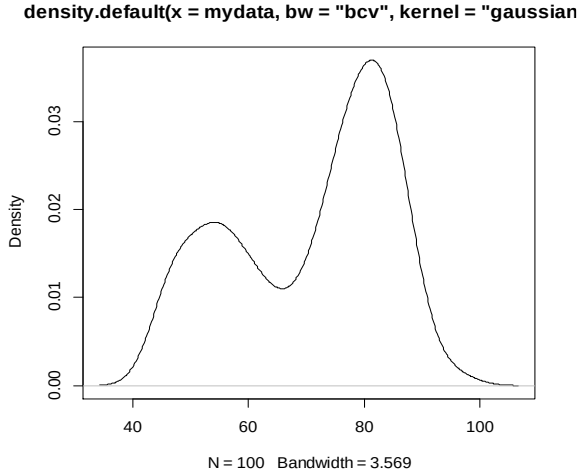

Fig. 4 bcv –Default Bandwidth

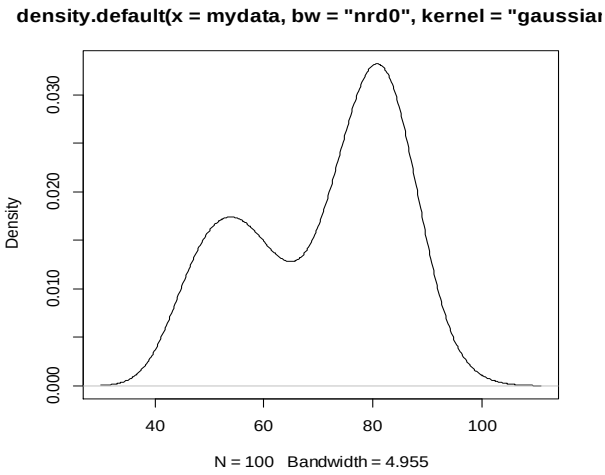

Fig. 5 nrd0 –Default Bandwidth

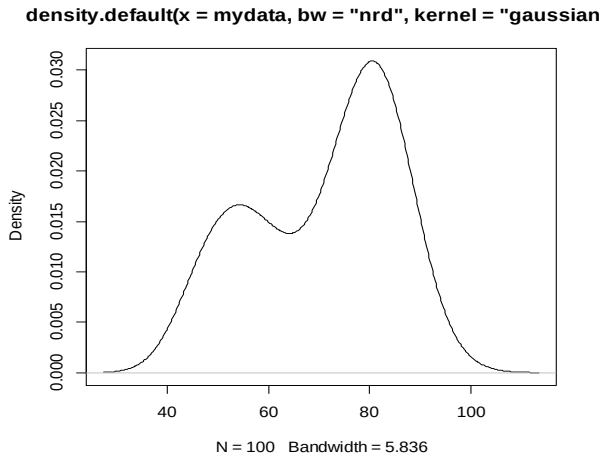

Fig. 6 nrd –Default Bandwidth

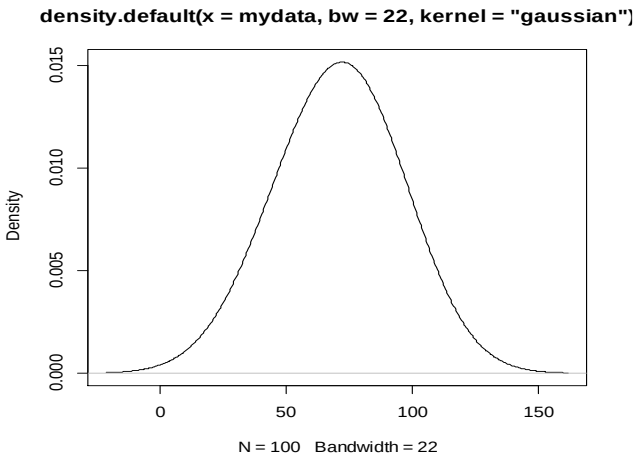

Fig. 7 Arbitrary bandwidth of 22

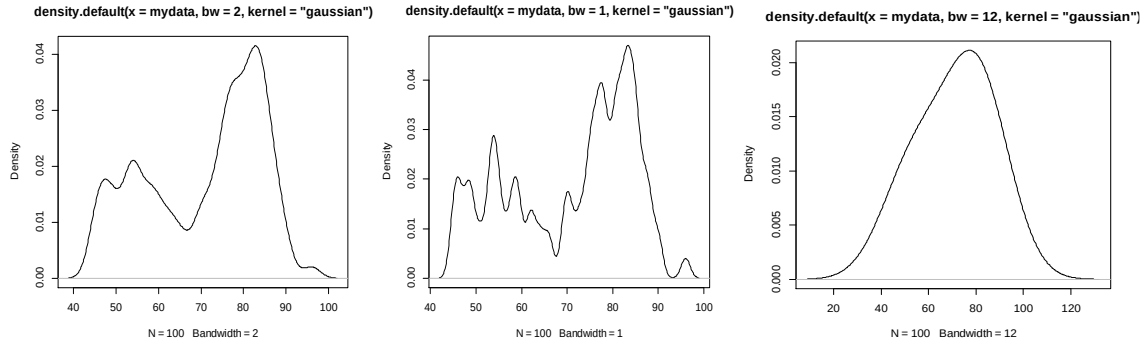

Fig. 8 Arbitrary Bandwidths of 2, 1 and 12

References

[1]. Blackburn V, Brennan S., Ruggiero J, (2014). Nonparametric Estimation of Educational Production and Costs using Data Envelopment Analysis, Springer Science+Business Media New York.

[2]. Chen, Y. C. (2018). Modal regression using kernel density estimation: A review. *WIREs Comput. Stat.*, 10, e1431.

[3]. Gramacki A., (2018), Nonparametric Kernel Density Estimation and Its Computational Aspects. Springer International Publishing AG, Gewerbestrasse 11, 6330 Cham, Switzerland

[4]. Härdle, W. (1994), *Applied Nonparametric Regression Analysis,* Cambridge: Cambridge University Press

[5]. Härdle W., Müller M., Sperlich S. and Werwatz A.(2005). *Nonparametric and Semiparametric Models-An Introduction.* Springer verlag. Berlin Heidelberg.

[6]. Heidenreich N. B., Schindler A, Sperlich S. (2013), Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis* **97**, 403–433. https://doi.org/10.1007/s10182-013-0216-y

[7]. Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association.* **91**, 401-7.

[8]. Nakarmi, Janet,(2016). "On Variable Bandwidth Kernel Density And Regression Estimation". Electronic Theses and Dissertations. 678. https://egrove.olemiss.edu/etd/678

[9]. Peracchi Franco, (2011). Nonparametric Methods University of Rome "Tor Vergata" and EIEF

[10]. Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, **9**, 65-78.

[11]. Ghosh S. (2018). Kernel Smoothing Principles, Methods and Applications, JohnWiley & Sons Ltd. 111 River Street, Hoboken, NJ 07030, USA.

[12]. Scott, D.W. and Terrell, G.R.(1987). Biased and unbiased cross-validation in density estimation. *Journal of American Statistical Association*, **82**, 1131-1146.

[13]. Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics* **12**(4): 1285–1297.

[14]. Takezawa K. (2006). Introduction to Nonparametric Regression. John Wiley Hoboken, New Jersey.

[15]. Tsybakov A. B. (2009). Introduction to Nonparametric Estimation. Springer Science+Business Media, LLC. New York.

[16]. Zambom Z. A and Dias R (2012). A Review of Kernel Density Estimation with Applications to Econometrics Universidade Estadual de Campinas.