ISSN No:-2456-2165

A Comparative Study of the Rule of Thumb, Umbiased Cross Validation and the Shearther Jones-Direct Plug-in Approaches of Kernel Density Estimation using Real Life Data

E.C. Ogwu and H.I. Ojarikre
Department of Mathematics, Faculty of Science,
Delta State University Abraka, Nigeria

Abstract: An important measure of the performance of any statistical method is how well it performs in practice especially using real life data. This work has compared the performance of different approaches of kernel density estimation (KDE) for several real data sets, the rule of thumb (NRD0), unbiased cross validation (UCV), and the Sheather Jones direct plug-in (SJ-DPI) approaches were considered. The data set examined includes, the daily closing prices of the Nigerian exchange (NGE), Coca-cola (KO) and NASDAQ stock data for a period of 1year, starting from October 31st 2021 to September 30th 2022. In this work we obtained the kernel density estimation of the data sets using the UCV, NRD0 and SJ-DPI approaches, we further compared the performances of each of the above approaches in terms of the resulting KDE plots and the integrated square error (ISE). We found out that the SJ-DPI approach produced the best KDE plot for the studied data sets and that the SJ-DPI and NRD0 approaches has the best performance for small and large samples sizes respectively in terms of ISE for the studied data sets. The study therefore suggests that the rule of thumb, unbiased cross validation and the plug-in approaches of KDE should be applied to other real data sets to compare the performance of these approaches. We further suggested that, other approaches of KDE such as the Bayesian approach, the solve the equation approach and the biased cross validation approaches be applied to NASDAQ, NGE, and COCA-COLA data sets in other to compare and investigate the performance of these approaches.

***Keywords**:- Kernel density estimation, smoothening parameter, rule of thumb, unbiased cross validation, direct plug-in,integrated square error.*

I. INTRODUCTION

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function (PDF) of a random variable. KDE primarily is a data smoothing problem where inferences about the population are made, based on a finite data sample. In some areas of specialization such as signal processing and econometrics it is also called the Parzen–Rosenblatt window method, after Rosenblatt (1956) and Parzen(1962). Who are usually credited with independentlycreating it in its present form.One of the well-known applications of KDE is in estimating the class-conditional marginal densities of data when using a naive Bayes classifier, which can improve its prediction accuracy.

KDE is one of the widely used non-parametric estimation techniques for estimating the probability density function of a random variable. For a univariate random variable X with unknown density$f(x)$, if we draw a sample of n independent and identically distributed observations $x1, x2 \dots\dots\dots\dots xn$ then the kernel density estimator is given by (Wand and Jones, 1995)

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{h}k(\frac{x-x_i}{h})$$

Where h is the bandwidth that controls the amount of smoothness, and k (·) is the kernel function which is usually chosen to be a symmetric density function.

Bandwidth selection plays an unavoidable role in nonparametric density estimation. An appropriate bandwidth can help produce an estimated density that is close to the true density; however, a poorly chosen bandwidth can make the true underlying features of the density to be hidden. Thus, a careful choice of bandwidth is suggested. A range of alternatives are now available for practitioners to select bandwidths, the most common being data-driven and plug-in methods. Several data-driven approaches exist which choose the bandwidth by minimizing the distance between the true and estimated density. In the continuous data setting these methods are shown to converge slowly and displays an unpredictable finite sample performance. Unlike data-driven methods, plug-in methods require *a priori* assumptions about the unknown distribution of the data and then seek to minimize the asymptotic mean integrated square error (AMISE) of a density estimator$\hat{f}(x)$.

Jones, Marron, and Sheather (1996) noted that there has been a major progress in recent years in data-based bandwidth selection for kernel density estimation. Some "second generation" methods, like the plug-in and smoothed boot strap techniques, have been developed which are far more superior to the well-known "first generation" methods, such as rules of thumb, least squares cross-validation, and biased cross-validation. The literature about bandwidth selection is very rich. Apart from the rule-of-thumb which often leads the choice of the bandwidth h among practitioners, we can cite more relevant methods of selection, such as the minimization of AMISE, evaluated for instance with cross validation or a plug-in technique Mathieu et al (2022).

Some of the smoothening parameter selection techniques which have become very popular are the cross-validation methods; plug in methods and rules of thumb. A Bandwidth is a data smoothing parameter, which helps to controls the smoothing of the kernel. It is the parameter responsible for the appearance of the bumps in a kernel. If the Bandwidth is large, the kernel is smooth in appearance. A large value of h will producea large bias, a low variance and over-smoothing of the curve. Kimari, Adem and Kiti(2015). This is the reason why some of the features may be concealed (Zambom, 2013). A small value of h results in a low bias and an increased variance which results in a curve with many sharp projecting points. This type of curve is not appealing, though it brings out most of the details in the data. A bias-variance trade-off needs to be considered in the choice of the kernel and bandwidth (Silverman, 1986; Bert, 1992). A Bandwidth chosen using various Bandwidth selection methods helps in the accurate choice of the estimator (Mugdadi&Jetter, 2010).

In finance and economics, non-parametric methods are not limited to the estimation of a PDF. We can cite for example their use to estimate the impact of market events, as with the non-parametric news impact curve in econometric volatility models. The rationale of such an approach is that linear impact models misestimate the reaction of markets to extreme events. Similarly, parametric models do not often describe accurately enough the tails of the PDF of price returns. Extreme events may also lead to other methodological choices in addition to the non-parametric approach. Thus in this study we want to investigate the kernel density estimation of selected stock data sets. This will be done using the rule of thumb, cross validation, and direct plug-in approaches. Particularly, we want to compare the performance of these approaches in terms of the resulting KDE plots and their performance in terms of ISE. This is a purely non parametric approach for estimating the pdf of a sampled data.

II. MATERIALS AND METHOD

A. Data collection methods

The real life data sets used for our investigation includes Globsl X MSCI Nigeria ETF (NGE) data set, National Association of Security Dealers Automated Quotations (NASDAQ) data set, and the Coca-Cola company (KO) data set.Secondary data were collected for these stock markets.The official international record of closing prices were used in this work. All the historical data in this category were downloaded from www.yahoofinance.com. The website contains an international repository of all the prices including open, high, low, and close prices, for each stock market.

B. Data analysis method

The following methods were adopted in the analysis of our data. All computations are coded in R-studio software

➢ *Kernel density estimation*

Let X_1, X_2, ...,$X_n$be randomly chosen sample from a population with unknown probability density function $f(x)$. The KDE for density function for any estimation point x is given as

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

Where K is a kernel function which is assumed to satisfy the following properties

$$\int_{-\infty}^{\infty} k(u)du = 1, \int_{-\infty}^{\infty} uk(u)du = 0, \int_{-\infty}^{\infty} u^2k(u)du = \mu_2(k) < \infty$$

And $h = h_i$ is a sequence of bandwidths or smoothening parameters that converges to zero with increasing value of n (Demir, 2019). For our purpose, we let K be a symmetric continuous probability density function, precisely the Gaussian kernel function.

➢ *Unbiased cross validation method*

According to Serdar (2019) an automatic method, LSCV which is also called unbiased cross-validation (UCV) is a flexible and easily computable method. In LSCV, the optimal bandwidth

$$\hat{h}_{LSCV} = arg\min_{h} LSCV(h)$$

Which minimizes the following cross-validation function $LSCV\,(h)$ over h is given as follows

$$LSCV(h) = \int \hat{f}_h^2\,(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h(i)}\,(X_i)$$

Where

$$\int \hat{f}_h^2\,(x)dx = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}(k * k)(\frac{X_i - X_j}{h})$$

➢ *Sheather and Jones direct plug-in method*

According to Opathalage(2021), the function φ_r associated with the normal density with variance σ^2 can be shown to be

$$\varphi_r = \frac{(-)^{\frac{r}{2}}r!}{(2\sigma)^{r+1}\left(\frac{r}{2}\right)!\,\pi^{\frac{1}{2}}}$$

When r is even, and $\varphi_r = 0$ for r odd. Making use of this, the following four stage direct plug-in was proposed by Sheather and Jones (1991) using a second order kernel K.

first estimate φ_8 using normal scale rule by using (1), leading to

$$\hat{\varphi}_8{}^{NS} = \frac{105}{\left(32\pi^{\frac{1}{2}}\hat{\sigma}^9\right)}$$

Where $\hat{\sigma}$ is an estimate of scale? Next φ_6 can be estimated using the kernel estimator $\hat{\varphi}_6(g1)$ where,

$$g1 = \left[\frac{-2K^{(6)}(0)}{\mu_2(K)\hat{\varphi}_8{}^{NS}}\right]^{\frac{1}{9}} n^{\frac{-1}{9}}$$

Then φ_4 is estimated using the kernel estimator $\hat{\varphi}_4(g2)$ where

$$g2 = \left[\frac{-2K^{(4)}(0)}{\mu_2(K)\hat{\varphi}_6(g1)}\right]^{\frac{1}{7}} n^{\frac{-1}{7}}$$

Finally, the selected plug-in bandwidth can be obtained as,

$$\hat{h}_{DPI,4} = [\frac{R(K)}{\mu_2(K)^2\hat{\varphi}_4(g2)}]^{\frac{1}{5}} n^{\frac{-1}{5}}$$

- *Silverman's rule of thumb (Nrd0) bandwidth method*

We use the rule of thumb formula as stated by Bhaveshkumar ((2015) which is the Silverman's rule of thumb that estimates h assuming $f(x)$ being Gaussian Silverman (1986). For a Gaussian pdf $R(f^{II}) = \frac{3\sigma^{-5}}{8\sqrt{\pi}}$ and for a Gaussian kernel

$R(K) = \frac{1}{2\sqrt{\pi}}$ Accordingly

$$h_{rot} = 1.0592\sigma N^{\frac{-1}{5}}$$

Where
$\sigma\ is\ the\ standard\ deviation\ of\ f(x)$Silverman (1986).

- *Integrated square error (ISE), mean integrated square error (MISE) and Asymptotic mean integrated square error (AMISE)*

  The method used to calculate the ISE, MISE, and AMISE in this work are stated below. According to Silverman (1986),

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2$$

$$MISE(h) = E\left[\int (\hat{f}_h(x) - f(x))^2 dx\right]$$

$$AMISE(h) = \frac{R(K)}{nh} + h^4 AR(f^{II})(\int \frac{x^2k}{2})^2$$

III. RESULTS AND DISCUSSIONS

Each of the data sets collected (see method of data collection above) wereanalyzed and all computations are coded on R studio software. We represent the results of our analysis below.

A. Kernel density estimation of Nasdaq stock data

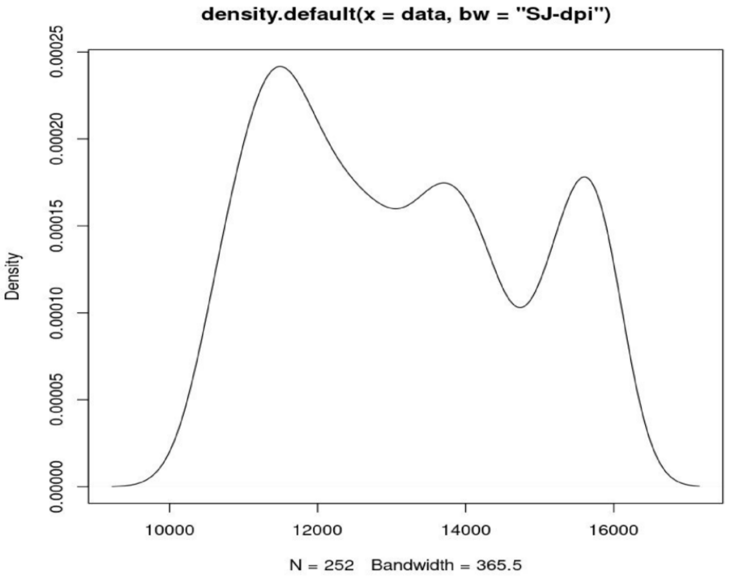

Fig. 1: Graphical representation of the KDE of NASDAQ stock data using the Shealther Jones plug-in bandwidth

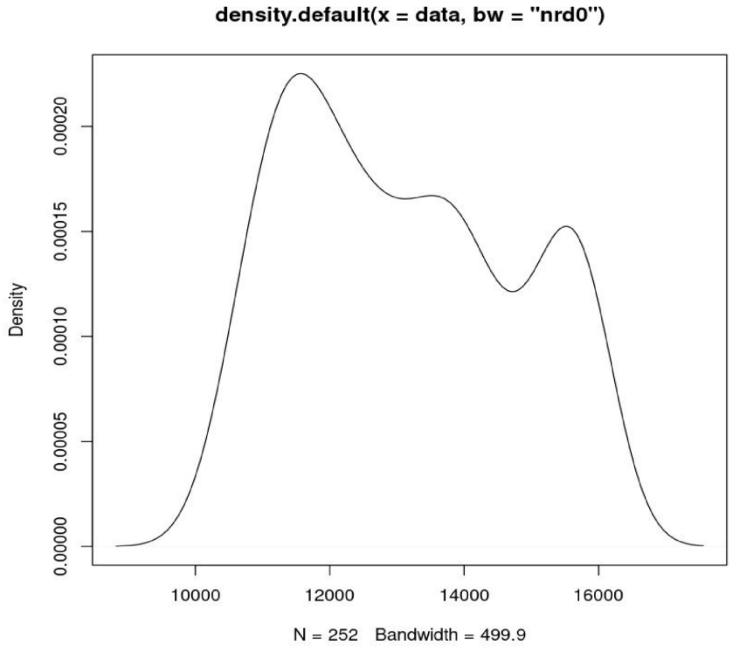

Fig. 2: Graphical representation of the KDE of NASDAQ stock data using the unbiased cross validation bandwidth

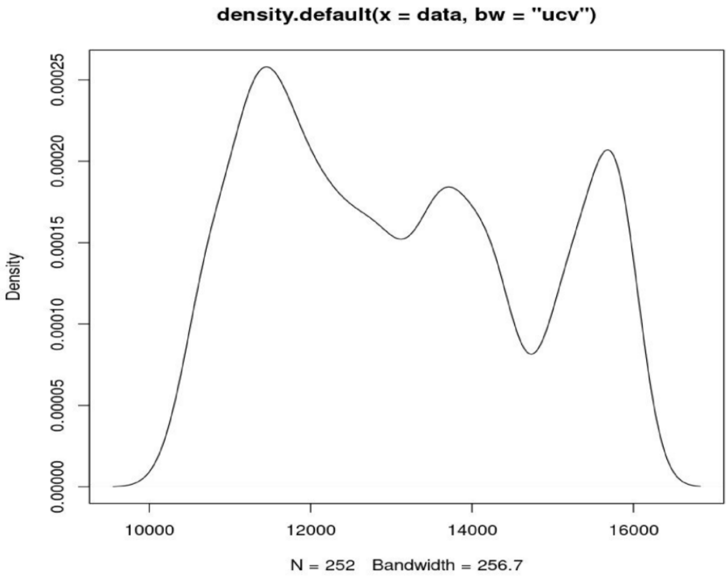

Fig. 3: graphical representation of the KDE of NASDAQ stock data using the rule of thumb bandwidth

Method	**Sample size**					
	50		**100**		**273**	
	ISE	**Bandwidth**	**ISE**	**Bandwidth**	**ISE**	**Bandwidth**
NRDO	0.00093	146.1	0.00027	170.3	0.00016	499.9
UCV	0.00087	115.4	0.00034	360	0.00018	256.7
SJ-DPI	0.00092	122.1	0.00031	238.9	0.00017	365.5

Table 1: Integrated square error and optimal bandwidth of the NRD0, SJ-DPI, and UCV KDE approaches for NASDAQ data

B. *Kernel density estimation of COCACOLA stock data*

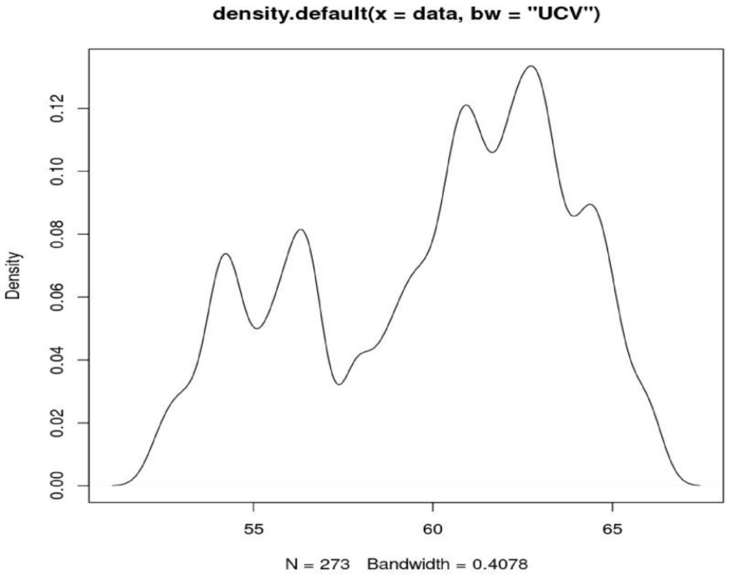

Fig. 4: graphical representation of the KDE of COCA-COLA stock data using the least square cross validation bandwidth

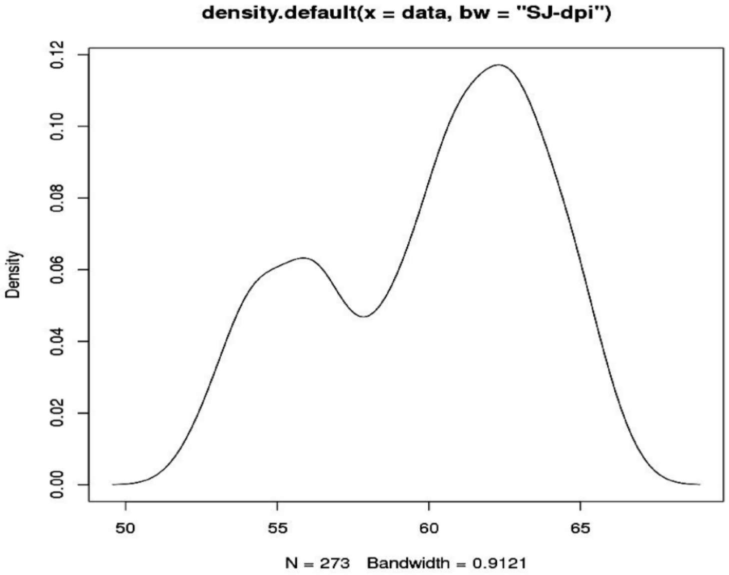

Fig. 5: graphical representation of the KDE of COCA-COLA stock data using the Shearther Jones direct-plug-in bandwidth

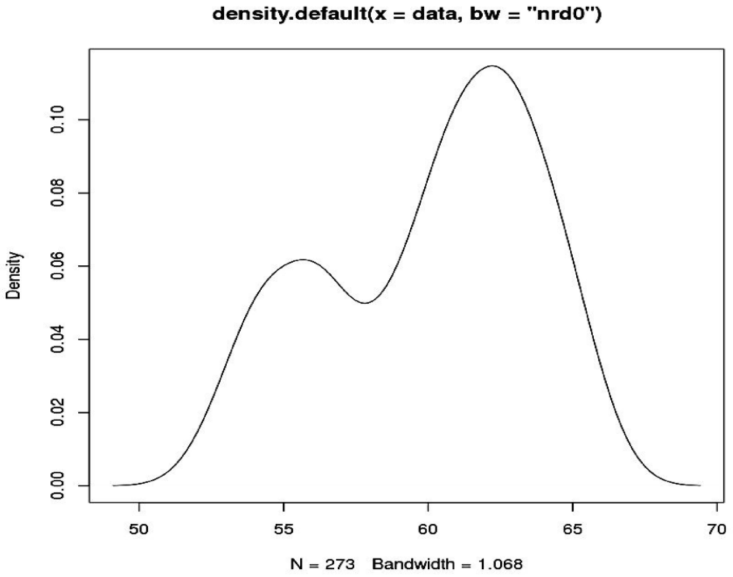

Fig. 6: graphical representation of the KDE of COCA-COLA stock data using the rule of thumb bandwidth

Method	Sample size					
	50		100		273	
	ISE	Bandwidth	ISE	Bandwidth	ISE	Bandwidth
NRDO	0.2014	0.5188	0.1142	0.8546	0.0763	1.068
UCV	0.2593	0.2684	0.1360	0.3144	0.0035	0.408
SJ-DPI	0.2327	0.3574	0.1151	0.8183	0.0783	0.912

Table 2: Integrated square error and optimal bandwidth of the NRD0, SJ-DPI, and UCV KDE approaches for COCA-COLA stock data

C. Kernel density estimation of NGE stock exchange

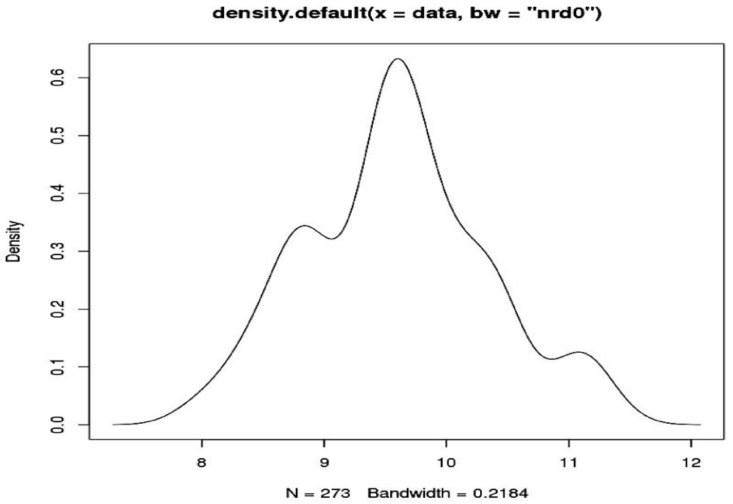

Fig. 7: graphical representation of the KDE of NGE stock data using the rule of thumb bandwidth

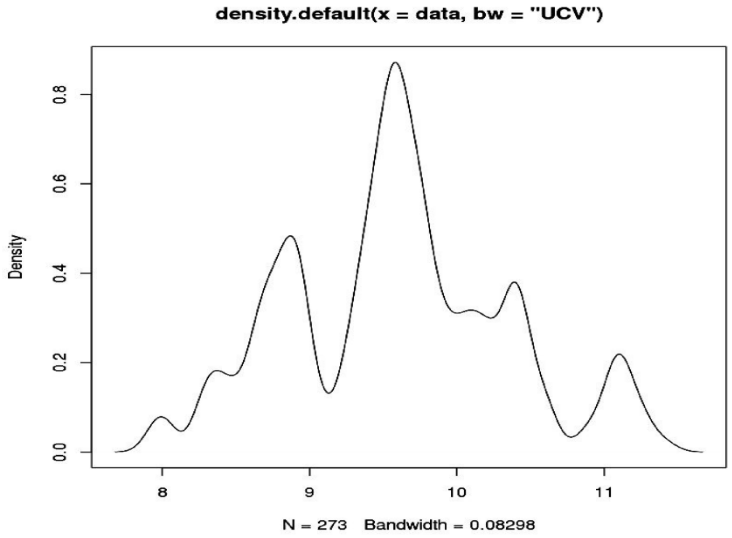

Fig. 8: graphical representation of the KDE of NGE stock data using the unbiased cross validation approach

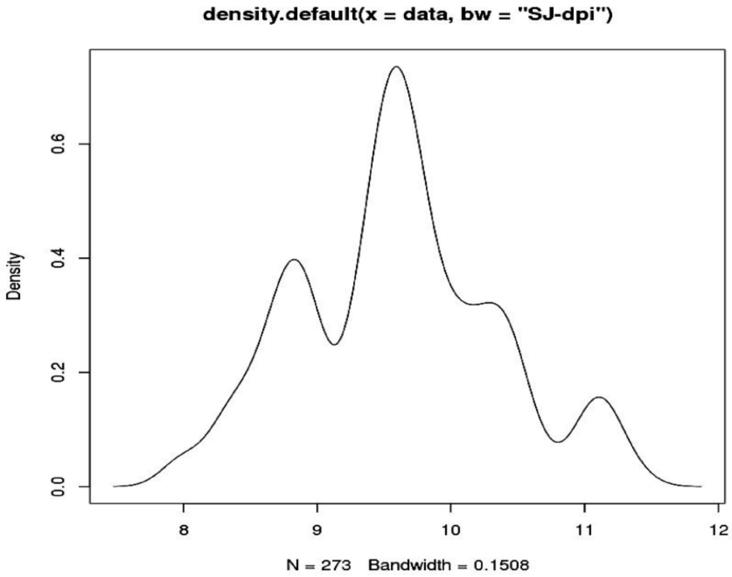

Fig. 9: graphical representation of the KDE of NGE stock data using Shearther Jones direct plug-in bandwidth bandwidth

Method	Sample size					
	50		100		273	
	ISE	Bandwidth	ISE	Bandwidth	ISE	Bandwidth
NRDO	0.7153	0.158	0.5080	0.1865	0.3720	0.2184
UCV	0.2311	0.031	0.6630	0.0438	0.4366	0.0830
SJ-DPI	0.8677	0.102	0.5423	0.1388	0.3969	0.1506

Table 3: Integrated square error and optimal bandwidth of the NRD0, SJ-DPI, and UCV KDE approaches for NGE data

IV. DISCUSSION OF RESULTS

A. *Comparing the performance of the UCV, NRD0 and SJ-DPI approaches using The KDE plots of the NGE, COCA-COLA and NASDAQ data sets*

By observing the KDE plot for NASDAQ stock data in figures 1, 2 and 3 above, we can observe firstly, the visual evidence of three nodes. We can also see that the depression between the nodes for the UCV approach is deeper than that of the SJ-dpi approach. Furthermore, the KDE plot of the NRD0 bandwidth has a very mild depression between the peaks hence we conclude that NRD0, UCV and the SJ-DPI kde plots are normally smoothed for the NASDAQ stock data set, with the SJ-dpi approach performing better than the UCV and the NRD0 approaches. This is in line with Jones et al (1996) who noted that “The density estimate based on the bandwidth UCV is severely under-smoothed, there are many spurious bumps, which make it hard to understand the structure of the data”. Thus the SJ-DPI approach has the best performance among the three approaches followed by the NRD0 approach and then the UCV approach for the NASDAQ data set.

Furthermore, by observing carefully the KDE plot for the coca-cola stock data set as shown in figure 4, 5 and 6 above, we noticed that the KDE plot for the SJ-DPI and the nrd0 approaches are almost similar, exhibiting bimodality. However, the NRD0 kde plot made the first node almost insignificant this depicts over smoothening. Furthermore, the KDE plot of the UCV approach is rough having five nodes this shows under smoothening. Hence the SJ-DPI approach produced the best KDE plot for coca-cola stock data set because it revealed accurately the true nature of the data set with a normal smoothness.

By observing the respective KDE plots of the NGE stock data using the three approaches as shown in figure 7, 8 and 9 above,We observed a visual evidence of four nodes depicting multimodality with one high peak. Moreover, the KDE plot for the UCV approach again is very rough showing under smoothening, on the other hand, the NRD0 approach made the third node almost invisible, thus the approach with the best performance is the SJ-DPI approach which clearly shows the four nodes with a normal level of smoothness.

From the above analysis we then draw a conclusion that the SJ-DPI approach has the best performance in terms of KDE plot for the studied data sets.

B. *Comparison of the ISE of each bandwidth selection method of the KDE for different sample sizes of the NGE, KO, and NASDAQ data set*

As noted by Grund et al. (1994), there is no important practical difference between ISE and MISE in assessing the performance of bandwidth selectors. Thus we compare the performance of each approach using the ISE as our error criterion function. For our comparison the sample size of 25 and 50 are considered as small sample sizes. While sample size of 100 and 200 are considered as large sample sizes the rule guiding our comparison and rating of performance using the ISE is that “ any smoothening parameter selector with a lower ISE value is considered to perform better than its counterpart with a higher ISE value for the same sample size of the data set”.

The ISE and bandwidth of NASDAQ stock data is presented in table 1. As expected, we can see that for the three bandwidth selectors, the ISE tend to reduce as the sample size increased. The UCV approach shows lowest ISE for sample size $n = 50$ whereas the nrd0 outperforms the ucv approach when the sample size reaches $n = 100$, we can also see that the SJ-dpi bandwidth outperforms UCV approach for sample size $n = 100$ and $n = 273$. Also for sample size $n = 273$ the nrd0 approach outperforms the other two hence the nrd0 has the best performance for large sample sizes of the NASDAQ data.

For COCA-COLAstock data, as shown in table 2, the nrd0 approach has the best performance for sample size $n = 50$ and $n = 100$ however the ucv approach has the best performance for sample size $n = 273$.

For the NGE stock data in table 3, we observed that the UCV approach has the best performance for sample size $n = 50$, and worst performance for larger sample sizes. The nrd0 approach has the best performance for sample size $n = 50$ and $n = 273$ respectively. Also the SJ-DPI has the poorest performance for sample size 50 but performed better than the UCV approach for sample size $n = 100$ and n=273 respectively. This findings is in line with Heidenreichet al (2011) who noted that“CV bandwidth leads to undersmoothing and is known to hardly stabilize for large data sets (they often just choose the smallest possible value among all bandwidths), whereas plug-in depends on prior information and typically works badly for small samples”.

From the above discussion of our results we conclude that the approaches with the best performance in terms of ISE for the various data sets under investigation are the SJ-DPI for small sample sizes and the NRD0 approach for large sample sizes

V. CONCLUSION

We have carefully carried out a comparative studyof the rule of thumb, unbiased cross validation and the Shearher Jones direct-plug approaches of kernel density estimation with real data sets. We have used the NASDAQ, NGE, and COCA-COLAdata sets as the subject of our investigation. We therefore conclude that the results of kernel density estimation for the real data sets used are similar with the results obtained from the kernel density estimation for data set generated from normal distribution. We observed that, as the sample sizes increased, the values of the ISE and bandwidth for the real data sets reduced, we also found out that the SJ-DPI approach produced the best kde plot for the studied data sets and finally that the SJ-DPI and NRD0 approaches has the best performance for small and large samples sizes respectively in terms of ISE for the studied data sets.

REFERENCES

[1.] Bert, V.E. (1992). Asymptotics for least squares cross-validation Bandwidths in non- smooth cases. The Annals of Statistics, 20(3), 1647 – 1657.

[2.] Bhaveshkumar C. D (2015) The Gram-Chalier A series based extended Rule of thumb for bandwidth selection in univirate and multivariate kernel density estimation DhirubaiAmbani Institutes of technology (DATICT) Gadhingagar, India - 382001

[3.] Demir. S (2019) Adaptive kernel density estimation with generalized least square cross-validation, *Journal of Mathematics and Statistics*: Volume 48(2), 616–625

[4.] Grund, B. &Polzehl, J(1997). Bias corrected bootstrap bandwidth selection. J. Nonparametr. Stat. 8, 97–126

[5.] Heidenreich .N, Schindler. A, and Sperlich .S (2011) Bandwidth selection for kernel density estimation: a review of fully automatic selectors StAAdv Stat Anal (2013) 97:403–433 DOI 10.1007/s10182-013-0216-y

[6.] 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, 401– 407. [10].

[7.] Kimari F, Adem A. &Kiti L(2015) Efficiency of various Bandwidth Selection Methods across Different Kernels, *IOSR Journal of Mathematics* (IOSR-JM) e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 15, Issue 3, PP 55-62.

[8.] Matthieu G, Jules K. & Sana L (2022) Estimation of time-varying kernel densities and chronology of the impact of COVID-19 on financial markets. HAL Id: hal-02901988 retrieved from https://hal.science/hal-02901988v2

[9.] .

[10.] Mugdadi,A. R. &Jetter,J. (2010).A simulation study for the Bandwidth selection in the kernel density estimation based on the exact and the Asymptotic MISE. Pak. J. Statist, 26(1), 239 – 265.

[11.] Omar M. E, Marie .M & Baker. M. H (2010). A Comparative Study for Bandwidth Selection in Kernel Density Estimation: *Journal of Modern Applied Statistical Methods*

[12.] Opathalage .S (2021) -driven smoothing parameter selection in density estimation: a thesis submitted to the faculty of graduate studies of the university of Manitoba

[13.] Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode". *The Annals of Mathematical Statistics.* 33 (3): 1065–1076. doi:10.1214/aoms/1177704472. JSTOR 2237880.

[14.] Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function". *The Annals of Mathematical Statistics.* 27 (3): 832–837. doi:10.1214/aoms/1177728190.

[15.] Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, 53, 683 – 690.

[16.] Sheather, S.J. (2004) .Density estimation. Statistical Science, 19, 588 – 597.

[17.] Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

[18.] Wand, M.P. & Jones, M.C(1995). Kernel smoothing. Monographs on Statistics and Applied Probability, vol. 60. Chapman and Hall, London

[19.] Wand, M.P., Marron, J.S., Ruppert, D(1991). Transformations in density estimation. J. Am. Stat. Assoc. 86, 343–353

[20.] Zambom, A. Z. & Dias, R. (2013). A review of kernel density estimation with applications to econometrics.IntEconom Rev (IER), 5(1), 20 – 4