Bandwidth selection algorithm and performance of variable window kernel density estimators: A Monte Carlo simulation study

Submitted By:

Tariq Majeed

Registration No. **PIDE2015FMPHIL-ETS12**

Supervisor

Dr. Atiq-ur-Rehman

Assistant Professor, PIDE

Department of Econometrics and Statistics

Pakistan Institute of Development Economics (PIDE) Islamabad

Pakistan Institute of Development Economics

CERTIFICATE

This is to certify that this thesis entitled: **"Bandwidth Selection Algorithm and Performance of Variable Window Kernel Density Estimators: A Monte Carlo Simulation Study"** submitted by Mr. Tariq Majeed is accepted in its present form by the Department of Econometrics and Statistics, Pakistan Institute of Development Economics (PIDE), Islamabad as satisfying the requirements for partial fulfillment of the degree in **Master of Philosophy in Econometrics**.

Supervisor:

Dr. Atiq-ur-Rehman
Assistant Professor
PIDE, Islamabad.

External Examiner:

Dr. Ateeq-ur Rahman Irshad
Assistant Professor
International Business & Marketing, NBS
NUST, Islamabad

Head,
Department of Econometrics and Statistics:

Dr. Amena Urooj

بسم الله الرحمن الرحيم

DECLARATION

I hereby solemnly declare that the work ***"Bandwidth selection algorithm and performance of variable window kernel density estimators: A Monte Carlo Simulation study"*** present in the following thesis is my own effort, except where otherwise acknowledged and that the project is my own composition. No part of the thesis has been previously presented for any other degree.

Tariq Majeed

Registration No. PIDE2015FMPhil-ETS12

Dear Past Thank You For All The Lessons.

Dear Future I Am Ready…

This Humble Effort is Dedicated to My Beloved Parents and My Brother Fazal Majeed

Acknowledgment

I have no word to thank Almighty Allah who bestowed me with the courage to complete this research thesis successfully in time from such a well reputed institution PIDE, Islamabad. I am extremely thankful to the Almighty who enabled me to pursue each and every bit of relevant information, meticulously and diligently which are the pre-requisites for success in any enterprise.

I would like to express my special acknowledgements to my honorable supervisor Dr. Atiq Ur Rehman, Assisstant professor PIDE Islamabad, for his valuable guidance, encouragement and cooperation during the completion of this thesis. It would not have been possible without his constructive criticism and friendly advices.

A word of thanks is also extended to Dr. Amna Urooj, Head of the department of Econometrics and Statistics, who gave me fruitful ideas and supported my research. I would also like to thank to Dr. Uzma Zia, Dr Hafsa Hina who was always willing to help and give her best suggestions in the fulfillment of this research thesis.

Words cannot express how grateful I am to my Parents, for their encouragement and support during my research and in whole life. Their prayers for me were what sustained me thus far.

Last but not least I would like to thanks my brother, whole family, friends especially M. Ilyas, M. Hanif, Zia uddin, Ali Gohar and Wajid Khan Wazir, and to all those who provided me the possibility to complete this thesis.

At last once again I am very indebted to Dr Hafsa Hina for putting me on the way to success in PIDE and to my beloved maternal uncle Wazir Zada, whose love and care make me stay here in PIDE and to wear the crown of philosophy.

Tariq Majeed

Abstract

Kernel density estimation has number of applications in different fields such as Econometrics, Economics, Engineering, Agriculture, Signal processing and identifying accident hot spots etc. Kernel density estimation procedures need a number of decision such as choice between variable kernel density estimation (VKDE) and fixed kernel density estimation (FKDE). It has been proven that VKDE performs better than FKDE. However among the VKDE, one has to make choice of bandwidths selection algorithm and kernel functions. There are four general classes of bandwidth selection algorithm i.e. *rule of thumb*, *Classical*, *Plug in* and *Bootstrap*. The most popular algorithm one from each of these classes are Silverman rule of thumb (SRT), Least square cross validation (LSCV), Improve plug in (IPI), Exact bootstrap (EB) and one cannot find appropriate guideline for choice between these algorithms. In addition to bandwidth selection algorithm the VKDE also depend on kernel function. There are nine different type of kernel functions i.e. Epanechnikov, Bi-weight, Tri-weight, Gaussian, uniform, Triangular, Tri-cube, Cosine and Sigmoid. This study is aimed to help in the choice of kernel function and bandwidth selection algorithm. We compare four kernel function and four bandwidth selection algorithm via Monte Carlo simulation for ten different types of normal mixture distribution. Our results show that IPI bandwidth and epanechnikov kernel function is the best choice for Gaussian, kurtotic unimodal, tri-modal and double claw distribution. For the remaining six distributions the EB bandwidth performed well.

Keywords: Kernel density estimation, Bandwidth selection algorithms, kernel function, SiZer, SiCon

Contents

Chapter 1

Introduction

Density estimation is a process employed to construct an estimate of probability density functions based on observed data. It frequently used for identification and description of the structure of data set on the basis of the random sample. There are two general classes for density estimation; namely parametric class and non-parametric. Like many other econometric procedures the parametric estimators depends on strict distributional assumptions which are usually not compatible with data. In parametric approach to density estimation the parameters are identified by applying the Maximum likelihood or Bayesian Methods. Then these parameters determine uniquely the distribution and density function of data set. Provided that the underlying assumptions are correct, the main advantage of parametric approach is ease of inference, efficiency and the absence of bias problems. However, the disadvantage of parametric approach is that it depends on more rigid distributional assumptions. If the distribution is not exactly known then it provides inconsistent estimates and the associated asymptotic bias provides invalid inference.

Non-parametric density estimation methods are used to model the data without making any assumption about the distribution of data. This approach is more flexible and depends only on smoothness assumption of density function. Therefore, this approach “let the data speak for themselves” impose without imposing any distribution assumptions. The non-parametric approach is well suited for exploratory data analysis and for analysis where density function cannot be clearly specified. A variety of non-parametric methods are used to estimate density such as histogram, data clustering techniques and kernel density estimation. This study concerned with kernel density estimation (KDE) which is a non-parametric density estimation method.

Kernel density estimator (KDE) introduced by Hodges (1951) and developed by Rozenblatt (1956) and Parzen (1962) as an alternative and improved non-parametric approach for density estimation. KDE has a wide scope. It has been applied in many fields, including Economics (DiNardo,et.al. 1995), Archaeology (Baxter,et.al. 2000), Banking (Tortosa, 2002), Genetics (Segal and Wiemels, 2002), Hydrology (Kim and Heo, 2002) and identifying Accident hot spots (Toran.et.al. 2015), etc. to investigate the features of data.

There are two types of KDE; fixed kernel density estimation (FKDE) and variable (adaptive) kernel estimation (VKDE). In FKDE the bandwidth is fixed for all data points while in VKDE the bandwidth varies at each data point. It is generally known that VKDE is superior to FKDE (Lemke D et al. 2015), however, with the VKDE, one has to make a variety of other specification decision for estimation of density: these include choice of bandwidth selection algorithm and choice of kernel function.

Bandwidth is the most crucial parameter for kernel density estimation. It controls the overall smoothness of the density. If we choose a large value of bandwidth then it over-smooth the data and local modes might be missed in the center of density estimate. Similarly, choosing a small bandwidth under-smooth the data and spiky estimator appears. So it is important to choose appropriate bandwidth to get the real structure and more information about data. There are four general classes of data driven bandwidth selection methods i.e. *Rule of thumb, Classical, Plug In* and *bootstrap* that are needed for KDE. The popular algorithms of these classes are *Silverman rule of thumb* (SRT), *Least square cross validation* (LSCV), *Improve plug in* (IPI) and *Exact Bootstrap* (EB) respectively. There is no appropriate guideline to decide between these algorithms for VKDE. So, one aim of this study is to compare these four selection algorithms for VKDE.

Given any method of choice of bandwidth, one has to choose the appropriate kernel function. Kernel function determines the shape of the mass assigned to each data point and the contribution of each data point to the estimated density. In KDE a Kernel function is placed on each data point and then averaging these kernel functions, which are placed on each data point, gives a resulted KD estimate. There are more than ten different kinds of kernel function used of which we take four kernel functions. The second goal of the study is the comparison of kernel functions.

If we chose optimal bandwidth algorithm and kernel function it gives a single density estimate with a number of peaks and trough. Mostly for noisy data a single density curve provides misleading information about the peaks and troughs. In order to know that which peaks are really there and which are spurious we have to move towards SiZer and SiCon. The significance of zero crossings of the derivative (SiZer) and significant convexity (SiCon) approach introduced by Chaudhry and Marron (1999) which is based on family of smooth approach provides a clear and direct answer to the problem of which peaks or modes are really there in the data and which are spurious. Both SiZer and SiCon depends on bandwidth selection algorithm and kernel function. Therefore the third purpose of this study is to compare the performance of bandwidth selection algorithms using SiZer and SiCon for significance of modes and curvatures of real data set.

1.2 Objective of the study

- To compare the performance of selected bandwidth algorithms for VKDE using Monte Carlo simulation and real data sets
- To find the best kernel function for VKDE
- To check the significance of peaks, troughs and curvatures in estimated density curve for real data set

1.3 Significance of the study

The KD estimators have the appropriate qualities of directly producing a density estimate, and the effects of grid size and placement cannot influence it (Silverman1986). Furthermore, it has the capability to accurately estimate densities of any shape provided that the bandwidth is selected appropriately. VKDE had shown its superiority over the FKDE procedure but still has the problems of choosing optimal bandwidth for pilot KDE. In literature very little is known about performance of the choices of bandwidth and kernel function for VKDE. Our study will help researchers to make the choices which optimize performance of VKDE. It would also be helpful for the practitioner to dig out the true structure of data through SiZer and SiCon.

1.4 Organization of the study

The remaining of the study organized as, chapter 2 review of literature in which we review different types of KDE, their drawbacks, bandwidth selection methods, and comparison of KDE. Chapter 3, methodology comprises step by step derivation and explanation of different bandwidth selection algorithms, VKDE and AMISE. Chapter 4 discusses background, implementation and explanation of SiZer and SiCon. Chapter 5 contains our finding and discussion. And finally we concluded our study with conclusion in chapter 6.

Chapter 2

Literature Review

Density estimate usually used to informally investigate the properties (such as skewness, uni-modality, multimodality etc.) of a given data set. The two general methods to estimate density are parametric and non-parametric methods. The parametric methods assume that the data are drawn from a particular parametric family of distributions such as Weibull, Gamma and Gaussian distribution etc. In this approach the parameters are identified by Maximum likelihood or Bayesian method which in turn used to estimate distribution function and also the density function. The parametric approach fit exactly if prior assumptions about the distribution are true. However, there are many situations, particularly in the social science, where these assumptions are in fact not met. In such situation it is better to used non-parametric density estimation. The oldest and widely used nonparametric density estimator is the histogram but because of scale, origin and non-continuity problems this is not good technique for density estimation (Silverman, 1986). The second method for density estimation is the naive method which is constructed by placing a box of width 2h and height $(1/2nh)$ on each observation. Summing these boxes a density estimate is obtained. But this method also has a lot of problems so its usage is not satisfactory for density estimation.

Fix and Hodge (1951) used the difference quotient of sample distribution function method (also called the naive method) for density estimation. The fundamental idea was that the proportion of data points fall in the neighborhood of any fixed value of grid point[1] may be used to estimate the probability of data points in that region. Taking ratio of the estimated probability and the

[1] Grid point is a point at which we want to estimate density

[2] K-NNE/ balloon estimator can be used as a point wise estimator as well as global estimator. Point wise balloon

measure of neighborhood the density estimate is obtained. But this method has the problem of selecting regions in which the data points fall. If these regions are small then the number of data points fall in these regions is too small. So this will not be an accurate estimate of the probability of data points fall in these regions while taking the regions too large this is still not a good approximation of probability density. The next method for density estimation is the kernel function method which is an improved form of the naïve method. There are two types of kernel function density estimator. We will discuss these types shortly in the coming section.

2.1 Fixed Kernel density estimation

KD estimation is also known as the Parzen-Rosenblatt window method introduced by Rosenblatt (1956) and Parzen (1961). Parzen-Rozenblatt density estimation is an alternative and most prominent approach for density estimation. It has a faster convergence rate than all other non-parametric approaches for density estimation. The fundamental idea is to estimate the density function at grid point x using neighboring observations. For estimating density function two parameters i.e. kernel function and bandwidth is required. Kernel function is required to smooth out the contribution of each data point to density estimate in the neighborhood grid point, at which we want to estimate density, while bandwidth control the overall smoothness of density. Parzen and Rozenblatt obtained FKD estimate by selecting suitable bandwidth, through optimization of mean integrated square error (MISE), and averaging all kernel functions placed on each data point. Woodroof (1968) and Nadaraya(1972) introduced and used a two stage method to estimate density given the kernel function. In this FKDE methods two initial gauss for bandwidth is made in order to obtain rough estimates of density and first non-vanishing moment of density on the bases of which a new bandwidth is computed. This new bandwidth is then used to estimate density in the usual fashion. Both the method of Woodroof (1968) and Nadaraya(1972) are same but was bases on different optimality criterion.

Several authors {Rosenblatt (1956), whittle (1958) and Parzen (1961)} used FKDE for univariate density estimation. The next problem was how to estimate multivariate density. The first steps for multivariate FKDE was taken by Cacoullos in 1966 and after that Epanichnikov in 1969.

2.1.1 Problems with fixed kernel density estimators

Kernel density estimation is the most widely used technique for density estimation. However, the main problem concerned with it is the fixed bandwidth. By selecting smaller window size every data point gets its own density and spurious noise appear in the tail of the distribution, while selecting large window size the density become smooth and much of the detail of the data masked in the center. Minnotte (1998), showed that FKDE have trouble with multimodal data. According to him, fixing a single window size that adequately differentiates between distinct peaks and troughs is very difficult. Selecting a large window size over smooth the density and much of the significant modes disappear. While small window size leads to spurious modes by under smoothing. The solution for this problem is the variable kernel density estimation.

2.2 Variable kernel density estimators

Variable window KDE can be divided into two groups: balloon estimators and sample point estimators. In both cases the window size varies. In case of balloon estimators' window size vary only at each estimation point. In contrast to balloon estimator the sample point estimator use different bandwidth for each data point.

2.2.1 Balloon estimators/ VKDE

Loftsgaarden and Quesenberry (1965) introduced balloon estimator, known as K-nearest neighbor estimator (K-NNE), for density estimation. The balloon estimators adopt the bandwidth to the local density of data. The NNE of Loftgarrden and Quessenbarry is given by

$$f_n^{\wedge}(z) = \left\{\left(\frac{k_{(n)} - 1}{n}\right)\right\}\left\{\frac{1}{v_{rk(n)_{.z}}}\right\}$$

In this estimator, $k_{(n)}$ is a specific number of data points chosen from n observations. Similarly $v_{rk(n)_{.z}}$ is the volume of hyper sphere and r the radius of hyper sphere. $rk(n)_{.z}$ is the distance from z to $k_{(n)}th$ nearest x_i to z which is determine by the Euclidian distance function from estimation point to the k_{th} nearest point. The main difference of NN-estimators and kernel estimators is that in NN-estimator case a specific number of observations are selected and its distance from estimation points is calculated. On other hand in kernel estimator, such is defined by Perzen (1962), the distance is specified from estimation point and the number of observations falling in that specified region is count. The benefit of NN-estimator is that it is always positive in data sparse area and performed elegantly for high dimensional data.

2.2.1.1 Problem with balloon estimators

The problems associated with balloon estimators are the larger bias in the tail of distribution as well as it is not a proper density function. Secondly it is a worse method for univariate and bi-variate density estimation (Terrell and scot 1992). The first problem was investigated by Mack and Rosenblatt (1979). They noticed that in the tail of distribution in spite of self-adjusting characteristic the NN-E has larger bias. Hall (1983) disentangled the problem of high bias of

NN-E in the tail of distribution by utilizing a generalized NN-E. But the problem of non-density function, when consider as a global estimate[2], still remained unsolved (Jones, 1990).

2.2.2 Sample point estimator/ VKDE

To improve the resulting density estimate Breiman et al. (1977) introduced sample point estimator, for which bandwidth changed at each data point. Their estimator is given by

$$\hat{f_{(x)}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B^{di}}\ k(\frac{x-x_i}{B_i})$$

Where B_i denote the Euclidian distance from one sample point to the other nearest sample point and d_i is the dimension. They were interested in entire density estimation. Their simulation study showed that such an estimator performed well in bi-dimensional case for bivariate normal and normal mixture densities but not quite well for univariate case.

For any dimension size Abramson in 1982 used square root law. The Abramson VKDE at target argument, $x = 0$, is given by

$$f_n(0) = n^{-1}\sum_{i=1}^{n} B^{-p}\ c(X_i)^p k\ B^{-1}\ c(X_i)X_i)$$

The window sizes B^{-1} depend on both n and scalar function c. Where c determined by the local behavior of density f only and is given by $c(x) = f^{-}(0)^{\frac{1}{p-\frac{1}{2}}}\ .f^{-}(x)^{1/2}$).

Abramson chooses bandwidth as the function of negative square root of density $B_i \propto f(x)^{-1/2}$.

---

[2] K-NNE/ balloon estimator can be used as a point wise estimator as well as global estimator. Point wise balloon estimator worked as a fixed kernel estimator.

In practice, a pilot estimate of density is required to estimate local bandwidth for each sample point. Abramson was interested in point wise estimation but latterly Hall and Marron (1988) used this method for global density estimation. Silverman in 1986 introduced a three steps adoptive VKDE. The first step is for pilot density estimator construction using global bandwidth B. In second step he define local variable bandwidth factors, l_{vB} as

$$l_{vB} = \left(\frac{geomean\left(f^{\wedge}(x)\right)}{f^{\wedge}(x)} \right)^{\frac{1}{\alpha}}$$

Where, $geomean$ is the geometric mean of the pilot density estimate $f^{\wedge}(x)$ and α is the sensitivity parameter. The final step is the VKDE which is define below

$$f^{\wedge}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_{(B)}} k\left(\frac{X_i - y}{V_{(B)}}\right)$$

$V_{(B)} = B * l_{vB}$ is the variable bandwidth vector at each data point X_i. The local bandwidth l_{vB} used to adjust the global bandwidth B for each data point in VKDE. The intuitive appealing of sample point VKDE is that it is a probability density function and have resolved the problems of fixed KDE by allowing the bandwidth to vary at each data point. It increase bandwidth in areas of low data densities and decrease in region of high data points and thus eliminate noise where data is sparse and recover detail in data enrich region. The simulation study of Silverman (1986) showed that sample point estimator performed well especially for small and moderate sample sizes in both one and two dimension case. It is because for small sample size the bias contribution from tail is negligible.

2.2.2.1 Problems with sample point estimator

Terrell and Scott (1992) identify the non-locality problem associated with sample point density estimator. The non-locality problem means that the sample point estimator is affected by the data point very far away and not just by nearby.

2.3 Bandwidth selection methods

KDE rely on optimal bandwidth and kernel function. Some studies argue that kernel function has not significant influence on density estimates while the bandwidth is the most crucial parameter for density estimation (Baszczynska.A 2005). There are four general classes of bandwidth selection algorithm i.e. *Rule of thumb*, *Plug in*, *Classical* and *Bootstrap* available in the literature to estimate the optimal bandwidth.

2.3.1 Rule of thumb algorithm

Silverman (1986) used rule of thumb approach for choosing optimal bandwidth. In this method he replaced the unknown roughness of density by its value for a normal distribution. For Gaussian kernel function and normally distributed dataset the robust Silverman rule of thumb bandwidth is given by

$$B = .9n^{-1/5}\min(\sigma^2(data), iqr(data)/1.34)$$

2.3.2 Classical bandwidth selection algorithm

The first well known class of automatic (data driven) bandwidth selection algorithm is the classical algorithms. It comprises a number of algorithms such as biased cross validation (BCV) of Terrell and Scott (1985) and Terrell (1990), likelihood cross validation (LCV) of Habbema et al. (1974), Cao, R. & Manteiga, W. G. (1994), indirect cross validation (ICV) of Savchuk et al. (2010) and least square cross validation (LSCV) of Rudemo (1982) and Bowman (1984) etc. The detail review of classical bandwidth selection algorithms is available in Zambom A.Z and Dias (2012). According to Zambom A.Z and Dias (2012) LSCV is the most popular and readily implemented classical bandwidth selection algorithm for KDE. Sain et.al (1994) showed that LSCV algorithm has certain appeal for VKDE as an unbiased estimate of integrated square error (ISE). The LSCV algorithm used "leaves one out"[3] estimators for the cross term of ISE to construct an unbiased estimate of optimal bandwidth. ISE is the distance between the true density and estimated density. Minimizing the estimated ISE with respect to B we obtain optimal bandwidth. The minimized LSCV score function often have a common local minima and empirically show better performance for local minimizer than that of global (Hall and Marron, 1991a)

2.3.3 Plug in algorithms

The plug in approach proposed by Woodroof (1970) is used to estimate the roughness of density in first step, and then plug in this estimate in the optimal formula of bandwidth. In this method the optimal bandwidth for estimation of roughness of density is different from optimal bandwidth used for density estimation. There is a lot of plug in bandwidth selection algorithms such is Scott et al (1977), Hall and Marron (1987), Park and Marron (1990), Sheather and Jones (1991), Hall et al (1991) and improve plug in (IPI) of Botev et al (2010). Park and Marron (1990) compared the data driven bandwidth selecting algorithms in term of their asymptotic convergent rate. It has been realized that the *plug in* method is the robust method when the underlying density is sufficiently smooth.

The weaknesses of plug in algorithms is discuss in detail by Loader, C.R. (1999). According to him, plug in algorithms dependent on arbitrary specification of pilot bandwidth and fail when

[3] an estimator define by using data set except the data point at which we want to estimate density

this specification is wrong. Secondly the plug in approach over-smooths the density and missing important features of the data. The IPI bandwidth algorithm introduced by Botev (2010) solved these problems. It is independent of pilot bandwidth and worked well even when the actual density is not smooth.

2.3.4 Bootstrap bandwidth selection methods

Bootstrap is a statistical methodology introduced by Efron (1992) for estimating standard errors and confidence interval of statistics/estimates. Bootstrap bandwidth selection algorithm devised and proposed by Taylor (1989) for density estimation. In this method the MISE is estimate and then minimize for obtaining optimal bandwidth. The Taylor algorithm does not capture the bias component of MISE and thus flops to direct estimate MISE. To solve the problem of Taylor algorithm, Faraway and Jhun (1990) introduced a Smooth bootstrap bandwidth algorithm which is based on initial/ pilot density estimate. But this method is also not free from problems. The main issue with their algorithm is the production of error due to Monte Carlo resampling. J C. Miecznikovski et.al. (2010) derived the Exact bootstrap estimator of MISE for kernel density estimation by using the work of Huston and Erust (2000).[4] Minimizing the Exact bootstrap estimated AMISE with respect to bandwidth; they obtained Exact bootstrap (EB) optimal bandwidth for KDE. The Miecznikovski EB bandwidth algorithm solves all the problems associated with Taylor and Faraway algorithms.

2.4 Drawback of kernel function density estimators

The traditional methods of KDE have some major drawbacks in showing the complete structure of the data. The most crucial among these drawbacks is the non-consensus of researchers in finding the optimal bandwidth. The second drawback is that which features are really there in

[4] Huston and Erust (2000) derived the exact bootstrap estimator of mean and variance of L-estimator.

data set and which are sampling artifacts. A new approach named significance of zero crossing of derivative (SiZer) and significant convexity (SiCon) introduced by Marron and Chung (1999) provided prodigious solution to these problems. The detail of this topic is given in chapter 4.

2.5 Comparison of KDE's using different bandwidth selection algorithms

In this section we are reviewing the comparison of different KDE's available in literature.

2.5.1 Comparison of fixed kernel density estimators

Sheather (2004) compare FKDE's with *least square cross validation* (LSCV) and *Sheather-Jone (*S-J) *plug in* bandwidth algorithms using Gaussian kernel function for Gaussian distribution and PGA Gulf tour real dataset. He concluded that for densities having widely varying curvature or roughness of density, $f''(x)$, the FKDE with LSCV bandwidth perform well while in the same case S-J plugs in bandwidth over smooth the density. Rodchuen and Suwattee (2010) compared FKDE's in term of AMISE for different simulated datasets by using solve the equation (STE), direct plug in (DPI) and Silverman rule of thumb (SRT) bandwidth selection algorithms. They showed that FKDE with DPI bandwidth algorithm perform better than FKDE with STE and SRT bandwidths for Gaussian distributed data set. For skewed unimodal and asymmetric claw distributions the FKDE with SRT bandwidth comparatively does well. The FKDE with STE bandwidth for kurtotic unimodal, separated bimodal or multimodal distribution have the lower AMISE as compare with AMISE of FKDE with SRT and DPI bandwidths. They also show that as the sample size increase the AMISE become close to zero and give an accurate estimate of density.

2.5.2 Comparison of fixed vs Balloon/KNN kernel density estimators

Mack and Rosenblatt (1979) compared theoretically the performance of FKDE and K-nearest neighbor estimator (KNN-E) for densities having a simple exponential form or inverse polynomial decay in the tail of distribution. They showed that both methods have the same variances but KNN-E has much larger bias than FKDE. They have concluded that for the aforementioned types of densities the KNN-E performs much worse than FKDE.

Terrell and Scott (1992) compared fixed and balloon kernel density estimators in term of asymptotic mean integrated square error (AMISE). Their simulation studies suggest that for small sample size and univariate and bivariate cases the balloon estimator performed poorly than FKDE. However, for multivariate case the balloon estimator become competitive. Giovanna Menardi (2014) carried out a simulation study for comparison of balloon, fixed (with Scheater & Jones bandwidth) and sample point /VKDE (with Silverman square root law bandwidth) estimators using multimodal normal distribution. He noted that the balloon estimator outperform the fixed and sample point estimators and that balloon estimator do not over smooth the density around the modes.

2.5.3 Comparison of Fixed vs variable kernel density estimators

Abramson (1982) developed VKDE and investigate that the convergence rate is marginally improved from $O(n^{\frac{-2}{(p+4)}})$ to $(n^{\frac{-2}{(p+4)}})$. Hall and Marron (1988) achieved comparatively faster convergence rate of $O(n^{\frac{-4}{(p+8)}})$ for VKDE than FKDE.

Terrell and Scott (1992) carried out a simulation study for comparison of VKDE and FKDE in term of AMISE. Their simulation study showed that VKDE perform better than FKDE for small and moderate sample size. Katkovnik, and Shmulevich (2002) proposed and developed a new technique for VKDE which is based on intersection of confidence interval (ICI) rule. The ICI rule require only the knowledge of density estimate and its variance. However, the variance depends on the unknown density. So a pilot estimate of the density is required in order to estimate variance. They have used the *Sheather- Jone plug in* bandwidth for pilot KDE. Simulation studies were carried out for symmetric, left skewed and right skewed Gaussian kernel. The results exposed that VKDE outperformed the fixed bandwidth method. Hazelton (2003) used the VKDE method of Sain and Scott but instead of zero order spline log-bandwidth function he used cubic spline log-bandwidth function. He made a comparison of his own VKDE with Sain & Scott and FKDE (using *Sheather - Jones plug in* bandwidth) by taking five targeted densities, unimodal normal, skewed unimodal, kurtotic unimodal, symmetric tri-modal and Asymmetric bimodal. The simulation study disclosed that both VKDE's outperform the FKDE for kurtotic unimodal, asymmetric bimodal and symmetric tri-modal. However, FKDE worked well than both VKDE's for skewed unimodal and normal densities. The study also indicated that his VKDE with cubic spline bandwidth function outperform the Sian & Scott kernel with zero order spline. Shimazaki & Shinomoto (2010) compared the performance of fixed and variable KDE for real biological data set. The methods were applied to the spike data of MT neuron. They observed that fixed kernel method choose a small bandwidth while the variable kernel method select a wider bandwidth in the period in which spike are not abundant. Ferdosi et al. (2011) studied the performance of four kernel density estimation techniques, the KNN-E, adoptive Gaussian KDE, adoptive Epanichnikov KDE, and modified Breman KDE. These density estimators were applied on six simulated and three astronomical real datasets. The comparisons were made in term of MISE and Kullback-Liebler divergence. The modified

Breman estimator which is the variable kernel density estimator performed better than other methods.

2.6 Literature Gap

In this study we will compare the performance of VKDE with most widely used traditional and newly introduced data driven bandwidth selection algorithms and kernel functions. VKDE used in different fields of knowledge to estimate the realistic structure of data and provide more information. Researchers made comparison of FKDE vs FKDE, FKDE vs VKDE and FKDE vs Balloon kernel density estimators. But in literature no one compared VKDE vs VKDE using the most recent and traditional bandwidth selection algorithms and kernel functions. In literature researchers used SiZer and SiCon for checking significance of modes and curvatures. We have used SiZer and SiCon for the same purpose as well as for confirmation of our simulation results. So that is the contribution of our study to literature.

Chapter 3

Methodology

This study is carried out in order to compare the performance of VKDE's by using *SRT*, *IPI*, *LSCV* and *EB* bandwidth selection algorithms. The following four types of kernel function will be used as weight function for VKDE's which is

- Epanechnikov kernel: $k_{Epn} = 0.75(1 - g^2)$ Tri-weight kernel: $k_{Tri} = 1.09(1 - g^2)^3$
- Bi-weight kernel: $k_{bi} = 0.937(1 - g^2)^2$ Gaussian kernel $k_{Gau} = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-g^2}{2}\right)$

The grid point $g,$ is a point at which we want to estimate density. In this study ten different kinds of Gaussian mixture models introduced by Marron and Wand (1992) used for Monte Carlo simulation. The accuracy of VKDE is measured by asymptotic mean integrated square error (AMISE) criteria. The estimator with small AMISE value is considered as the best estimator of density.

3.1 Algorithm of variable kernel density estimator and AMISE

Step 1: Calculate optimal bandwidth B by using SRT, LSCV, IPI and EB bandwidth selection algorithms

Step 2: Calculate the pilot kernel density estimate by using optimal bandwidth B obtained through step 1.

Step 3: The local variable bandwidth vector can be calculated by utilizing the pilot KDE and its geometric mean.

$$l_{vB} = \left(\frac{geomean\left(f^{\wedge}(x)\right)}{f^{\wedge}(x)}\right)^{\frac{1}{\alpha}}$$

Step 4: The variable bandwidth vector is obtained by multiplying the optimal bandwidth B with

local variable bandwidth vector $$V_{(B)} = B\left(\frac{geomean\left(f^{\wedge}(x)\right)}{f^{\wedge}(x)}\right)^{\frac{1}{\alpha}}$$

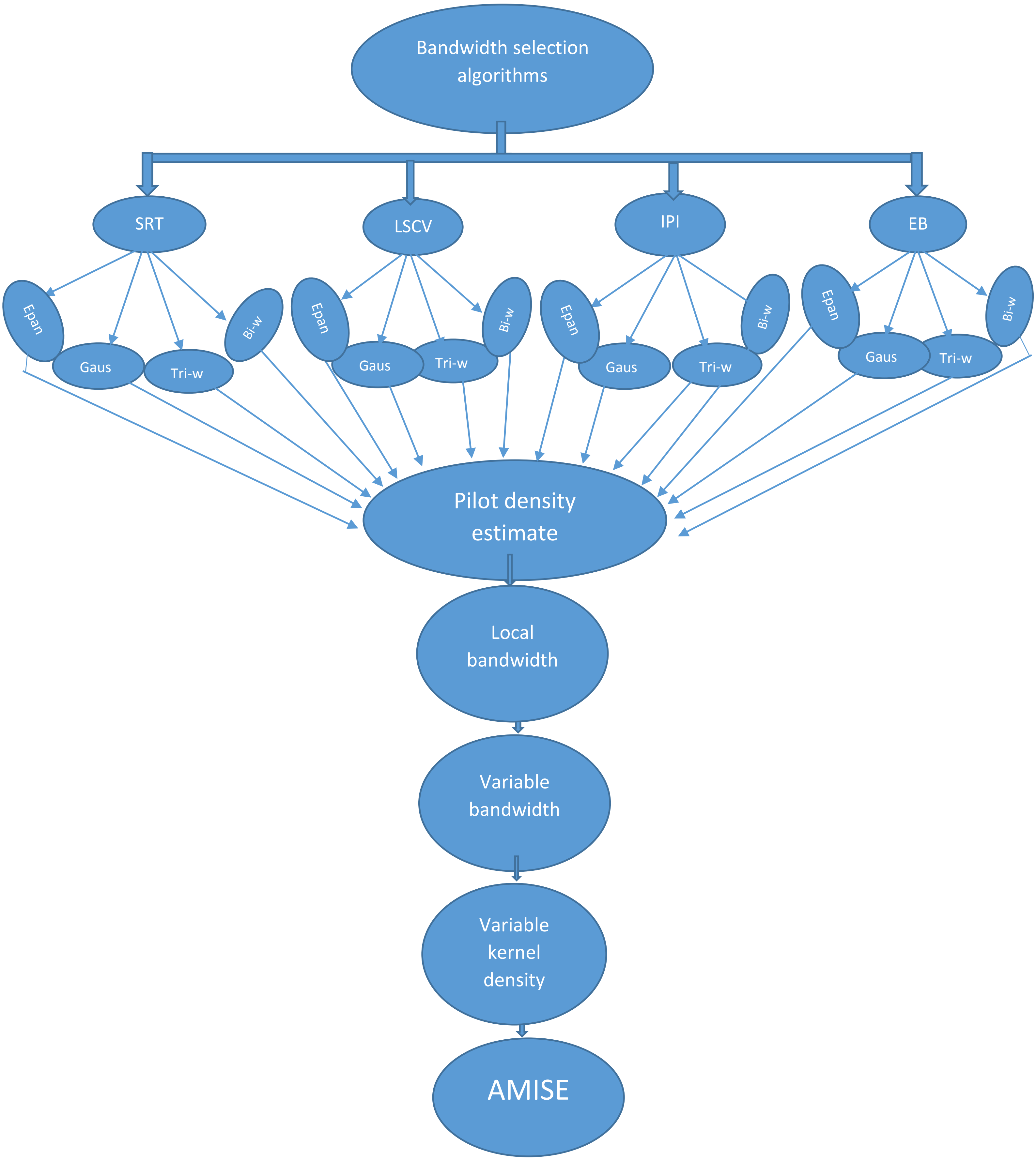
Step 5: Variable kernel density estimate is obtained by putting $V_{(B)}$ in VKDE formula

$$f^{\wedge}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_{(B)}}k\left(\frac{X_i - y}{V_{(B)}}\right)$$

Step 6: In final step the estimate of AMISE can be obtain by utilizing the above information.

$$AMISE = \left(\frac{1}{p!}R(f^{(p)})V_{(B)}{}^{p}k_p(k)\right)^2 + \frac{R(K)}{nV_{(B)}}$$

Flow chart of VKDE and AMISE



3.2 Data generating process

We are generating artificial data from the following known normal mixture distributions introduced by Marron and Wand (1992) for Monte Carlo simulation. The concept behind using these distributions is that, these distributions can closely approximate any kind of density. Secondly we believe that any real data situation can be effectively modeled by these distributions.

1. Gaussian $N(0,1)$,
2. Skewed unimodal $\frac{1}{5}N(0,1)+\frac{1}{5}N\left(0.5,\left(\frac{2}{3}\right)^2\right)+\frac{3}{5}N\left(\frac{13}{12},\left(\frac{5}{9}\right)^2\right)$,
3. Kurtotic unimodal $\frac{2}{3}N(0,1)+\frac{1}{3}N\left(0,\left(\frac{1}{10}\right)^2\right)$,
4. Separated bimodal $\frac{1}{2}N\left(-\frac{3}{2},\left(\frac{1}{2}\right)^2\right)+\frac{1}{2}N\left(\frac{3}{2},\left(\frac{1}{2}\right)^2\right)$,
5. Tri-modal $\frac{9}{20}N\left(-\frac{6}{5},\left(\frac{3}{5}\right)^2\right)+\frac{9}{20}N\left(\frac{6}{5},\left(\frac{3}{5}\right)^2\right)+\frac{1}{10}N\left(0,\left(\frac{1}{4}\right)^2\right)$,
6. Double claw $\frac{49}{100}N\left(-1,\left(\frac{2}{3}\right)^2\right)+\frac{49}{100}N\left(1,\left(\frac{2}{3}\right)^2\right)+\sum_{l=0}^{6}\frac{1}{350}N\left(-\frac{(l-3)}{2},\left(\frac{1}{100}\right)^2\right)$,
7. Asymmetric claw $\frac{1}{2}N(-0,1)+\sum_{l=-2}^{2}\frac{2^{1-l}}{31}N\left(\left(l+\frac{1}{2}\right),\left(\frac{2^{-l}}{10}\right)^2\right)$,
8. Smooth comp $\sum_{l=0}^{5}\frac{2^{5-l}}{63}N\left(\left(\frac{65-96*2^{-l}}{21}\right),\left(\frac{32}{63*2^{l}}\right)^2\right)$,
9. Strongly skewed $\sum_{l=0}^{7}\frac{1}{8}N\left(3\left\{\left(\frac{2}{3}\right)^l-1\right\},\left(\frac{2}{3}\right)^{2l}\right)$
10. Outlier distribution $\frac{1}{10}N(0,1)+\frac{9}{10}N\left(0,\left(\frac{1}{10}\right)^2\right)$.

3.2.1 Optimal bandwidth

Optimal bandwidth is obtained by minimizing AMISE with respect to B . The AMISE is given below

$$AMISE = \int_{-\propto}^{\propto} AMSE\,(f^{\wedge}(x))dx$$

It is a global measure of precision and can be decompose into two parts, integrated square bias and Variance.

Bias

Bias of density estimator is the difference between the estimated density and actual density as given by $bias\left(f^{\wedge}(x)\right) \;=\; E\left(f^{\wedge}(x) - f(x)\right)$

The asymptotic integrated bias of kernel density estimator, $f^{\wedge}(x)$, is given by P^{th} order Taylor approximation as

$$= f(x) + \frac{1}{p!} f^{p} B^{p} k_{p}(k) - f(x)$$

$$= \frac{1}{p!} f^{p} B^{p} k_{p}(k) \qquad 3.1$$

Variance

Variance is the square of deviation of an estimate from its mean value. It is given by the following formula

$$var\left(f^{\wedge}(x)\right) = E\left(f^{\wedge}(x) - \overline{f^{\wedge}}(x)\right)^{2}$$

For variance of kernel density estimate the 1^{st} order Taylor approximation is used[5].

$$var\left(f^{\wedge}(x)\right) \;=\; \tfrac{1}{n} E k_{B}(X_{i} - x)^{2} - \tfrac{1}{n}(E k_{B}(X_{i} - x))^{2}$$

[5] The remained of first order Taylor approximation is very small so that’s why we used first order Taylor approximation

$$\cong \frac{1}{nB^2}\int_{-\infty}^{\infty} k(\frac{z-x}{B})^2 f(z)dz - \frac{1}{n}f(x)^2$$

$$= \frac{1}{nB}\int_{-\infty}^{\infty} k(u)^2 f(x+Bu)du$$

$$= \frac{f(x)\int_{-\infty}^{\infty} k(u)^2 du}{nB}$$

$$= \frac{R(K)}{nB} \qquad 3.2$$

The term $f(x)\int_{-\infty}^{\infty} k(u)^2 du = R(K)$ is the roughness of kernel.

Asymptotic mean integrated square error

The asymptotic mean integrated square error (AMISE) is the sum of square bias (3.1) and variance (3.2) so

$$AMISE = \left(\frac{1}{p!}R(f^{(p)})B^p k_p(k)\right)^2 + \frac{R(K)}{nB} \qquad 3.3$$

The $k_p(k)$ is the variance of kernel, $R(K)$ the roughness of kernel, $R(f^{(p)})$ the p[th] order derivative or roughness of unknown density $R(f^{(p)})$ and B is the bandwidth. Minimizing the AMISE with respect to B the optimal bandwidth is obtained

$$B_{opt} = \frac{d}{dB}\left(\frac{R(f^{(2p)})B^{2p}k_p^2(k)}{(p!)^2} + \frac{R(k)}{nB}\right)$$

$$B_{opt} = R\left(f^{(2p)}\right)^{\frac{-1}{(2p+1)}} \left(\frac{(p!)^2 R(k)}{2pk_p^2(k)}\right)^{\frac{1}{(2p+1)}} (n)^{\frac{-1}{(2p+1)}} \qquad 3.4$$

Equation 3.4 shows optimal bandwidth for KDE. We see that the optimal bandwidth depend on the roughness of kernel $R(k)$, variance of kernel $k_p^2(k)$ and roughness of density $R\left(f^{(2p)}\right)$. The first two can be easily determine by kernel function. But the last one is the fly in ointment in way of determining the optimal bandwidth. To estimate the roughness of density and calculate the optimal bandwidth a lot of bandwidth selection procedures have been proposed. We are using the most popular of them.

Silverman rule of thumb bandwidth algorithm

Silverman introduced the rule of thumb approach to replace the unknown quantity $R\left(f^{(2p)}\right)$ in (3.4) by reference density $R\left(g_{\sigma^{\wedge}}^{(p)}\right)$. The reference density (g_{σ}) used is a normal density with $N(0,\ \sigma^{\wedge 2})$. The Silverman bandwidth selection algorithm gives optimal bandwidth in case when the true density is normal. We are using normal mixture distributions in our simulation study so it would be the best choice to select for comparison.

For normal density and Gaussian kernel the optimal global bandwidth is given by

$$B_{opt} = 2\sigma^{\wedge}\left(\frac{\pi^{\frac{1}{2}}(p!)^3 R(k)}{2p(2p)!\, k_p^2(k)}\right)^{\frac{1}{(2p+1)}} (n)^{\frac{-1}{2p+1}} \qquad 3.5$$

Here p is the order of kernel, $\sigma^{\wedge}$ sample standard deviation and the value in parenthesis is constant but different for different kernel function. The optimal bandwidths for different kernel function by using *SRT* algorithm has given by Hansen (2016) in his book

Gaussian $B_{opt} = 1.06\sigma^{\wedge} n^{\frac{-1}{5}}$ Epanechnikov $B_{opt} = 2.34\sigma^{\wedge} n^{\frac{-1}{5}}$

Bi-weight $B_{opt} = 2.78\sigma^{\wedge} n^{\frac{-1}{5}}$ Tri-weight $B_{opt} = 3.15\sigma^{\wedge} n^{\frac{-1}{5}}$

Improve plug in algorithm

The improved plug in bin size introduced by Botev (2010) is completely non-parametric and independent of normal rule of thumb approach. The selected method does not require numerical optimization and approximately as fast method as the normal reference rule method. As we know from the above equation (3.5) that the optimal bandwidth depend on the curvature of the true density which is unknown. So it is necessary to estimate the curvature before the estimate of density. Botev derived the improved plug in scale function from solve the equation (STE) bandwidth selection function. The STE bandwidth assumes the true density as a Gaussian density in order to compute the estimate of curvature. This assumption leads to not a good estimate of bandwidth when in reality the true density is not Gaussian. Botev found a solution to the non-linear equation of the STE bandwidth derived by Jone, et.al. (1996), as given below,

$$B_{sj=\Psi\Upsilon^{s}(B)} \qquad 3.6$$

for some stage s using either fixed point repetition or the method of Newton's with $B = 0$ intially. In the above equation $\Psi = \left(\frac{6\sqrt{2}-3}{7}\right)^{\frac{2}{5}} \approx 90$ and Υ^{s} is the plug in stages.

The algorithm of fixed point repetitions are given by

Step1: execute with $w_n = \varepsilon$ where ε is the machine precision and $n = 0$

Step 2: set $w_{n+1} = \Psi\Upsilon^{s}(w_n)$

Step 3: if $|w_{n+1} - w_n| < \varepsilon$, then stop and set optimal bandwidth $B^{*} = w_{n+1}$; otherwise set $n = n + 1$ and repeat from second step again.

Step 4: used the optimal bandwidth B^{*} in third step to estimate KDE and $B_1^{*} = \Upsilon^{s+1}w_{n+1}$ as the optimal bin size for the curvature estimation.

The fixed point repetition algorithm is a succeeded algorithm for finding the root of equation

$$B_{=\Psi\Upsilon^{s}(B)}$$

Secondly it gives a unique root. The third advantage of this algorithm is that solution to the equation $B_{=\Psi\Upsilon^5(B)}$ and $B_{sj=\Psi\Upsilon^{s+5}(B)}$, for any $s > 0$ do not differ practically. Alternatively we can say that by increasing the stage of bandwidth selection rule beyond $s = 5$ no gain is achieved. Using the discrete cosine transform this procedure provides us a fast computational estimate of bandwidth but here we do not go to the theoretical derivation and linkage of cosine transform.

Least square cross validation bandwidth selection algorithm

LSCV method is an automatic data driven method of choosing the optimal smoothing parameter. Sain et.al (1994) showed that LSCV method has certain appeal for variable kernel density estimation as an unbiased estimate of L-2 error (ISE). The ISE for this method is given by

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

$$= \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) f(x) dx + \int f(x)^2 dx \qquad 3.7$$

ISE is unknown so cross validation used to replace it with an unbiased estimate. As we know that the third term of above expression (3.7) does not depend on B so we ignore it. The first term can be calculated directly from estimated density by using pilot bin size. The second cross term of 3.7 can be calculated by “leave one out” estimator as

$$\hat{f}_{-i}(X_i) = \frac{1}{(n-1)B} \sum_{j \neq i}^{n} k\left(\frac{X_j - X_i}{B}\right)$$

“Leave one out” estimate is an estimate of density at $x = X_i$, computed without observation X_i. So the unbiased estimator of $\int \hat{f}(x) f(x) dx$ is

$$\frac{1}{n}\hat{f}_{-i}(X_i) = \frac{1}{n(n-1)B}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k\left(\frac{X_j - X_i}{B}\right)$$

By combining the estimators of first and second cross terms of 3.7 we get the LSCV criteria, the unbiased estimator of ISE, given below

$$LSCV = \frac{1}{n^2B}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{K}(X_i - X_j) - \left(\frac{2}{n(n-1)B}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k\left(\frac{X_j - X_i}{B}\right) - \frac{k(0)}{nB}\right) \quad \dots\dots 3.8$$

In order to estimate the first term of (3.8) we have need of pilot bandwidth. Most often in literature normal reference method used to construct pilot bandwidth so we also used this technique. The first term than become as

$$First\ term = \frac{1}{n^2B\sqrt{2}B_p}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{K}(X_i - X_j)}{\sqrt{2}B_p}\right)$$

Put in (3.8) the required estimate of LSCV score function obtained.

$$ISE(B) = LSCV = \frac{1}{n^2B\sqrt{2}B_p}\left(\frac{\sum_{j=1}^{n}\bar{K}(X_i - X_j)}{\sqrt{2}B_p}\right) - \left(\frac{2}{n(n-1)B}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k\left(\frac{X_j - X_i}{B}\right) - \frac{k(0)}{nB}\right) \quad \dots\ 3.9$$

Optimizing the above function with respect to h we will get LSCV classical optimal data driven bandwidth.

Further solving the LSCV expression we get

$$CV_{f(0)} = \left(\frac{k_1^2}{4}\right)\left[\int \left(f^{(2)}x^2\right)dx\right]B^4 + \frac{k}{nB} \qquad 3.10$$

Where $k_1 = \int v^2 k(v)dv$ is the variance of kernel function and $k = \int k^2(v)dv$ is the roughness of kernel function. Minimizing 3.10 with respect to B, we get the optimal bandwidth empirically as below

$$\frac{d(AISE)}{dB} = \frac{d}{dB}\left(\left(\frac{k_1^2}{4}\right)\left[\int \left(f^{(2)}x^2\right)dx\right]B^4 + \frac{k}{nB}\right)$$

$$B_{lscv} \ = k^{\frac{1}{5}}k_1^{-2/5}\left\{\left[\int f^{(2)}(x)\right]^2\right\}^{\frac{-1}{5}} \qquad 3.11$$

Exact Bootstrap bandwidth selection algorithm

There are a lot of bootstrap bandwidth selection algorithms in literature. We are using the Miecznikovski, et al (2010) newly derived bootstrap bandwidth procedure. For derivation of bootstrap bandwidth they have expressed the kernel density estimator as L-estimator. An L-estimator is an estimator that is equal to a linear combination of order statistics of the measurements.

The L-estimator form of kernel density function is given

$$f^{\wedge}(x;B) = \sum_{j=1}^{n} \psi_j Z_{j:n}$$

Where ψ_j is equal to $\frac{1}{nB}$ and $Z_j = k\left(\frac{x-X_i}{B}\right)$.

The order statistics of sample Z_j is $Z_{1:n} \leq Z_{2:n} \leq Z_{3:n} \ldots . Z_{n:n}$ where $j = 1, \ldots, n$

With this L-estimator framework of kernel density estimator they obtained exact mean and variance of kernel density estimator.

The exact bootstrap mean of KDE is given by

$$\overline{f_J^b}(x;B) = \frac{1}{nB}\sum_{j=1}^{n}\sum_{i=1}^{n} g_{i(j)} Z_{i:n}$$

Where the weights $g_{i(j)}$ are $g_{i(j)} = b_{c,n-c+1}\left(\frac{j}{n}\right) - b_{c, \quad n-c+1}\left(\frac{j-1}{n}\right)$

The exact bootstrap estimator for the variance of kernel density estimator is given by

$$\sigma^2_{f^\wedge(x)} = \left(\frac{1}{nB}\right)^2 \left(\sum_{j=1}^{n} \sigma^{\wedge 2}_{j:n} + 2\sum_{j<k}^{n} \sigma^{\wedge}_{jk:n}\right)$$

Where the variance is equal to $\sigma^{\wedge 2}_{j:n} = \sum_{j=1}^{n} g_{j(c)}\left(Z_{j:n} - u^{\wedge}_{c:n}\right)^2$ and $u^{\wedge}_{c:n} = \sum_{j=1}^{n} g_{j(c)} Z_{j:n}$

The covariance is given by

$$\sigma^{\wedge}_{jk:n} = \sum_{j=2}^{n}\sum_{i=1}^{j-1} g_{ij(cs)}\left(Z_{i:n} - Z_{j:n} - u^{\wedge}_{s:n}\right) + \sum_{j=1}^{n} \epsilon_{j(cs)}(Z_{j:n} - \varpi^{\wedge}_{c:n})(Z_{j:n} - \varpi^{\wedge}_{s:n})$$

The weight $g_{ij(cs)}$ and $\epsilon_{j(cs)}$ are given by

$$\epsilon_{j(cs)} = \int_{(j-1)/n}^{j/n} \int_{(j-1)n}^{u_s} f_{cs}\left(\varpi_{c,}\varpi_s\right) du_c du_s,$$

$$g_{ij(cs)} = \int_{(j-1)/n}^{j/n} \int_{(i-1)n}^{i/n} f_{cs}\left(\varpi_{c,}\varpi_s\right) d\varpi_c d\varpi_s,$$

The bootstrap exact estimator of asymptotic mean integrate square error is given below

$$\widehat{MISE} = \int_{-\infty}^{\infty} \sigma^2_{f^\wedge(x:B)} + \left(\varpi^{\wedge}_{f^\wedge(x:B)} - f^\wedge(x)\right)^2 dx \qquad 3.10$$

Minimizing the $\widehat{MISE}$ with respect to B we obtain the bootstrap bandwidth for kernel density estimation.

3.2.2 Pilot density estimation

Pilot density estimate is a fixed bandwidth density estimate obtain by putting the value of optimal global bandwidth in kernel density estimator

$$f^\wedge(x) = \frac{1}{nB_{opt}} \sum_{i=1}^{n} k\left(\frac{X_i - x}{B_{opt}}\right) \qquad 3.11$$

3.2.3 Local bandwidth for variable kernel density estimation

The optimal local bandwidth vector for variable kernel density estimation is inversely related to the pilot estimate of density as define by Abramson (1982)

$$lvB_{.} = \left(\frac{g}{f^{\wedge}(x)}\right)^{\frac{1}{\alpha}} \qquad 3.12$$

$f^{\wedge}(x)$ is the Pilot estimate of density function with fixed bandwidth and g the geometric mean of pilot density.

3.2.4 Variable bandwidth vector for kernel density estimator

The variable bandwidth vector can be calculate by scaling the global optimal bandwidth B by local bandwidth which is given below

$$vB_{(Xi)} = B\left(\frac{g}{f^{\wedge}(x)}\right)^{\frac{1}{\alpha}}$$

3.2.5 Variable kernel density estimation

Putting the value of variable bandwidth in variable kernel density estimator we obtain variable kernel density estimate. $f^{\wedge}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{vB_i^d}k\left(\frac{X_i - y}{vB_i}\right)$

3.2.6 Mean integrated square error of variable kernel density estimation

Substituting variable bandwidth in AMISE as given below

$$AMISE = \left(\frac{1}{p!}R(f^{(p)})vB_{(Xi)_{.}}^{\ p}k_p(k)\right)^2 + \frac{R(K)}{nvB_{(Xi)_{.}}}$$

We get the estimated AMISE of SRT, LSCV, IPI and EB methods for comparison of variable kernel density estimators.

Chapter 4

Significant zero crossing of derivative (SiZer)

As we know the problem associated with histogram that is the problem of origin has been solved by kernel density estimation. But the issue of bandwidth is still under discussion in kernel density estimation. The various optimal data driven bandwidth methods provide a single bandwidth which lead to estimation of single true density curve. Even the optimal bandwidth vector, one value for each data point, in variable kernel density estimator also gives an estimated line. But this single curve does not show the complete structure of data set. So for that reason a new approach which is the family of smooth approach has been introduce by Marron and Chung in 2001[6]. In this, family of smooth approach, a vector of bandwidth instead of one is used and gives a family of smooth KDE curves. The bandwidth vector used in this approach contains both small and large bandwidth values which give more information about the structure of data. The problem of family approach is that it does not show which features in data are signals and which sampling artifacts. As well as it cannot be applied evocatively in case where the data change with change in time (Skrovseth.et al 2012). To solve these problems Choudhry and Marron (1999) introduced SiZer exploratory data analysis. In SiZer technique each location affects both past and future value of estimated density curve (Skrovseth.et al 2012). SiZer is based on family of smooth approach and originated from the scale space theory of Lindeberge (1994) in computer science. It uses a color scheme to analyze the visible features in data over location and scale. The blue color used to show the significantly increasing of the curve and red significantly decreasing. The purple color indicates a region where the curve is neither increasing nor decreasing. Finally

[6] The family of smooth approach was introduced by Marron and Chung in 1997 and published in 2001 in journal of computational statistic.

the gray color in map shows the region where the data is too sparse. The peak in density estimation is shown by the density curve going up on left side and coming down on right side. SiZer map shows this up going portion of curves by blue color and downward portion by red and provide statistical significance to these ups and down. When a bump observed, on the right side of the bump the derivative is negative on the left it is positive and on top derivative is equal to zero. When a trough is observed, to the lift of trough the derivative is significantly negative to the right it is positive and at the minimum it is zero. So top and trough both are indicated by zero crossing of the derivative/slope.

4.1 Testing hypothesis in SiZer

SiZer is totally dependent on family of smooth curves. Statistically the smooth curves are given below

$$\left\{f_b^{\wedge}(y) : \in [Bin_{mi}, Bin_{mx}]\right\}$$

In this approach we estimate a number of curves depend on a vector of bandwidth V_B. Each bandwidth in V_B is responsible for a single density curve. So if we have n bandwidths in vector then a family of n curves will estimate. The selected bandwidths in V_B is from minimum to maximum. The minimum bandwidth is equal to $B_{mi} = 2B$, while the maximum bandwidth is the range of data y.[7]

$$B_{mx} = y_{mx} - y_{mi}$$

[7] Where B is the optimal bandwidth calculated with any of bandwidth selection procedure, i.e. IPI, LSCV, etc.

A small bandwidth gives a wiggly curve while a large bandwidth provides a smooth line. The density curves based on V_B shows different structure in data. So we are indifferent to decide that which peaks are really there and which are due to sampling variability.

Hypothesis H_0: $\frac{\partial^m \mathrm{E}(\hat{f_B}(y))}{\partial y^m} = 0$

H_A: $\frac{\partial^m \mathrm{E}(\hat{f_B}(y))}{\partial y^m} \neq 0$

If the null hypothesis is rejected then the derivative is either positive or negative depending upon the sign of $\frac{\partial^m \mathrm{E}(\hat{f_B}(y))}{\partial y^m}$ and showing significant increasing or decreasing in curve. The hypothesis is tested independently at each location in scale space.[8] The $\hat{f_B}'(y)$ is an unbiased estimate of the density derivative at each point/location in scale space. We assume that $\hat{f_B}'(y)$ follow normal distribution and performed hypothesis testing with a proper estimate of standard deviation

$$sd = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[k'\,(y_t - y) - \hat{f_B}'(y)\,\right]^2}$$

As we assumed that the estimated density derivative is normally distributed so for this reason enough data points is required within the kernel window.

Making confident interval for hypothesis testing quantile of the data is necessary to calculate. There are four different choices available to calculate the quantile.

1) Point wise Gaussian quantile $q_1(B) = \varphi^{-1}\left[1 - \frac{\alpha}{2}\right]$

[8] Location is the data point and scale is the bandwidth in scale space theory of Lindeberge (1994).

2) Simultaneous over y Gaussian quantile

3) Bootstrap simultaneous over y quantile

4) Bootstrap simultaneous over y and B quantile

The first one is not good choice because sometime it incorrectly shows spurious clusters as significant. This problem is solved by adjusting the length of confidence interval to do simultaneous inference. The second choice of quantile for simultaneous confidence band is based on the fact that when the locations are far from each other than the estimated density at these location are independent which implies that the derivative is also independent at the given locations. So in this case the problem associated with simultaneous confidence interval is approximated by the m independent confidence limits problem. The m independent confidence interval is calculated through effective sample size for each scale space

$$E_{ss}(y, B) = \frac{\frac{1}{B}\sum_{1=1}^{n} k\left(\frac{y_i - y}{B}\right)}{\frac{1}{B}k_{(0)}}$$

k is the kernel function, most often Gaussian kernel function is used in SiZer map building. So we will also take this mass function in our analysis. The m independent confidence interval, denoted by $m(B)$ in the following equation, is the number of independent blocks of average size from a set of n observation.

$$m(B) = \frac{n}{avg\ E_{ss}(y, B)}$$

The effective sample size $E_{ss}(y, B)$ can also be used to show where the smooth is based on shatter data by highlighting the region by gray color in SiZer map where $E_{ss} \leq n_0$.[9] Chaudhuri and Marron (1999) suggested $n_0 = 5$. Therefore to avoid the problem of small effective sample size $E_{ss}(y, B)$ they also modify $m(B)$ to $m(B)'$. where $m(B)'$is given by

$$m(B)' = \frac{n}{avg_{y \in D_B} E_{ss}(y,B)}$$

Where D_B is the location with large data points. Assuming that $m(B)$ blocks of data are independent the simultaneous quantile for 95% confidence band is given by

$$q_2(B) = \varphi^{-1}\left[\frac{1+(1-\alpha)^{1/m(B)}}{2}\right]$$

The φ^{-1} term in above expression is the inverse of standard Gaussian distribution[10]. The third and fourth choice can be obtained by bootstrapping over either y or $y\ and\ B$. But these choices are time costly, computer intensive and less informative as compare to the independent blocks approach (skrovseth.S.O.et al. 2012). The confidence interval for estimated derivative is given in the following equation

$$C.I = \left\{f_B^{\wedge}(y) - q * sd\left(f_B^{\wedge}(y)\right); \left(f_B^{\wedge}(y)\right) + q * sd\left(f_B^{\wedge}(y)\right)\right\}$$

4.1.1 Decision rule on the basis of confidence interval

When zero lies below or above the confidence limits then we reject the null hypothesis and conclude that the smooth is either increasing or decreeing depending on the sign of derivative. In

[9] Gray color show sparse of data

[10] In hypothesis testing we have assumed that density derivative follow normal distribution so for this reason while calculating quantile we used, φ^{-1}, inverse of standard normal distribution.

case if zero lies within the confidence limits then we cannot reject the null hypothesis and there is either peak or trough.

4.2 Significant convexity (SiCon)

SiCon used to check the significance of curvatures in estimated curve. Not like SiZer, SiCon map is based on second derivative of smooth curve. In SiCon, color map, the area of significant curvature are depicted by cyan and orange color. Cyan color show significant concavity while orange color indicates convexity. The green color in SiCon map shows no significant curvature.

Chapter 5

Results and discussion

Monte Carlo simulation study is carried out to compare the performance of VKDE using different kernel functions and bandwidths. The AMISE is taken as the performance criteria. The effects of various kernel functions and bandwidths on the performance of VKDE are considered for different sample sizes. Random samples of size 50, 100 and 200 are drawn from each distribution with 1,000 time replication. For normal mixture distributions of Marron and Wand (1992) using kernel functions and bandwidths the $\overline{AMISE}\ (\widehat{f})$ values of VKDE are depicted in Tables 5.1 to 5.10. In each table for VKDE the best kernel function and bandwidth is shown by bold number and lowest value of $\overline{AMISE}\ (\widehat{f})$ respectively. Simulation is performed with the help of Matlab programming language.

5.1 Simulation results for Gaussian distribution

The $\overline{AMISE}\ (\widehat{f})$ of VKDE for Gaussian distribution with 50,100 and 200 sample sizes have indicated that epanechnikov kernel function is the best choice among the other kernel functions as shown by bold numbers in Table 5.1. Thus our results of comparing kernel functions for Gaussian distribution support the result of Rodchuen M et al (2010). Similarly for the same distribution the IPI bandwidth is comparatively the best bandwidth with lowest $\overline{AMISE}\ (\widehat{f})$ values. Therefore, we conclude that VKDE with IPI bandwidth and epanechnikov kernel function perform better than VKDE with the remaining methods.

Table.5.1 $\overline{AMISE}\ (\widehat{f})$ for Gaussian distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$
50	Epanechnikov	**0.16**	**0.199**	**0.1088**	**0.0797**
	Gaussian	0.1908	0.1626	0.1342	0.107
	B-weight	0.2762	0.2629	0.4515	0.2307
	T-weight	0.5774	0.3054	0.5942	0.3023
100	Epanechnikov	**0.1494**	**0.0988**	**0.1793**	**0.0958**
	Gaussian	0.1889	0.1074	0.215	0.0962
	B-weight	0.2749	0.2283	0.2425	0.2202
	T-weight	0.5317	0.3049	0.6283	0.2931
200	Epanechnikov	**0.1293**	**0.0976**	**0.1148**	**0.0937**
	Gaussian	0.1607	0.1062	0.159	0.103
	B-weight	0.3048	0.2203	0.2303	0.2135
	T-weight	0.4581	0.3005	0.6205	0.3131

5.2 Simulation results for kurtotic unimodal distribution

For kurtotic unimodal distribution the $\overline{AMISE}\,(\widehat{f})$ of VKDE with IPI bandwidth is comparatively lower than VKDE with SRT, LSCV and EB bandwidths. So IPI is the best bandwidth selection method for VKDE in kurtotic unimodal case. Similarly for the same distribution the epanechnikov kernel function perform well as shown by bold number in Table 5.2. These results also have similarities with the results of Rodchuen M et al (2010) for kernel function.

Table 5.2 $\overline{AMISE}\hat{f}_{(x)}$ for Kurtotic unimodal Distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$
50	Epanechnikov	**0.144**	**0.1029**	**0.2896**	**0.0994**
	Gaussian	1.6113	0.3579	0.3827	0.3568
	B-weight	0.2391	0.2084	0.3054	0.2271
	T-weight	0.4212	0.3165	0.6748	0.3085
100	Epanechnikov	**0.1115**	**0.0959**	**0.2886**	**0.0958**
	Gaussian	0.6308	0.3416	0.3602	0.2901
	B-weight	0.2152	0.2072	0.3017	0.2091
	T-weight	0.3165	0.3108	0.338	0.3026
200	Epanechnikov	**0.0984**	**0.0956**	**0.2878**	**0.0942**
	Gaussian	0.4418	2.9647	0.3408	0.2687
	B-weight	0.2148	0.2101	0.3009	0.2017
	T-weight	0.314	0.3104	0.3351	0.3014

5.3 Simulation results for outlier distribution

For sample generated from outlier distribution the VKDE with EB bandwidths have lower $\overline{AMISE}\ (\widehat{f})$ than VKDE with SRT, IPI and LSCV bandwidth. The results in Table 5.3 also show that the epanichnikov kernel function is the best of all other selected kernel functions.

Table 5.3 $\overline{AMISE}\hat{f}_{(x)}$ for Outlier distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$
50	Epanechnikov	**1.1843**	**0.6069**	**0.3011**	**0.6026**
	Gaussian	5.5265	4.6176	0.3223	4.3998
	B-weight	1.2274	1.227	0.3724	1.2269
	T-weight	1.803	1.8029	0.4331	1.8031
100	Epanechnikov	**1.1756**	**0.6194**	**0.2925**	**0.618**
	Gaussian	4.9351	4.801	0.3221	4.7227
	B-weight	1.2447	1.2448	0.3585	1.2444
	T-weight	1.8207	1.8199	0.4156	1.8167
200	Epanechnikov	**1.1748**	**0.6273**	**0.2536**	**0.6266**
	Gaussian	5.0761	4.9412	0.2681	4.9183
	B-weight	2.1345	1.2538	0.3396	1.2537
	T-weight	1.8297	1.8298	0.3989	1.8298

5.4 Simulation results for bi-modal distribution

For data sampled from bi-modal distributed population the VKDE using EB bandwidth perform well with the lowest $\overline{AMISE}\ (\hat{f})$ as compared to VKDE with LSCV, SRT and IPI bandwidth algorithms. Epanechnikov kernel function with IPI, LSCV and EB bandwidth algorithms perform better than Gaussian, bi-weight and tri-weight. For SRT bandwidth the Gaussian kernel function outperforms the other.

Table 5.4. $\overline{AMISE}\hat{f}_{(x)}$ **For Bi-Modal distribution**

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$
50	Epanechnikov	0.7796	**0.7795**	**0.704**	**0.7676**
	Gaussian	**0.7212**	1.3507	0.7197	1.1128
	B-weight	2.2138	1.5481	0.893	1.5436
	T-weight	2.2131	2.213	0.9637	1.2896
100	Epanechnikov	0.8046	**0.8045**	**0.7039**	**0.7995**
	Gaussian	**0.7525**	1.261	0.7426	1.2397
	B-weight	1.3188	1.0836	0.7839	1.0835
	T-weight	2.2565	2.2566	0.9757	1.3188
200	Epanechnikov	0.8171	**0.8173**	0.7025	**0.8168**
	Gaussian	**0.7213**	1.2901	**0.6869**	1.3059
	B-weight	1.3332	1.0975	0.6921	1.6046
	T-weight	2.2777	2.2782	0.9813	1.3332

5.5 Simulation results of skewed unimodal distribution

Table 5.5, shows that VKDE with Gaussian kernel function and EB bandwidth give the lowest L2-error as compared to the SRT, LSCV and IPI bandwidth methods for all sample of sizes 50, 100 and 200. For the SRT, LSCV and IPI bandwidths the epanichnikov kernel function is the best among the others. These results are also according to the results of Rodchuen M et al (2010). We also observed from table that as the sample size increase the $\overline{AMISE}\ (\hat{f})$ decrease.

Table 5.5 $\overline{AMISE}\hat{f}_{(x)}$ for Skewed unimodal distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$
50	Epanechnikov	**0.7663**	**0.7639**	0.6711	**0.7518**
	Gaussian	1.0715	0.9817	**0.3659**	0.93
	B-weight	1.1769	1.2159	0.8436	1.1766
	T-weight	1.6103	1.6639	0.981	1.6092
100	Epanechnikov	**0.7658**	**0.7632**	0.3807	**0.7509**
	Gaussian	1.0736	1.0324	**0.3651**	1.002
	B-weight	1.203	1.2045	0.4396	1.2023
	T-weight	1.646	1.6459	0.4978	1.6445
200	Epanechnikov	**0.7619**	**0.7617**	0.3801	**0.7504**
	Gaussian	1.0699	1.0299	**0.3624**	1.0013
	B-weight	1.2001	1.2017	0.4216	1.2003
	T-weight	1.633	1.6371	0.45	1.6328

5.6 Simulation results for strongly skewed distribution

For data generated from strongly skewed distribution the EB bandwidth outperformed the IPI, LSCV and SRT bandwidth. VKDE with EB bandwidth and bi-weight kernel function appears comparatively with the lowest $\overline{AMISE}\ (\widehat{f})$ as shown in Table 5.6. For the remaining three bandwidths Gaussian kernel worked well.

Table 5.6 $\overline{AMISE}\hat{f}_{(x)}$ for Strongly Skewed distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$
50	Epanechnikov	60.1175	60.1264	2.6062	60.125
	Gaussian	**18.3199**	**18.363**	2.6048	**18.319**
	B-weight	60.4568	60.4532	**2.6045**	60.4518
	T-weight	61.1112	61.112	2.6066	61.1059
100	Epanechnikov	324.0054	324.0159	2.6105	323.9981
	Gaussian	**34.87**	**34.8766**	2.6130	**34.01**
	B-weight	326.0062	326.0188	**2.6085**	326.0013
	T-weight	327.3548	327.355	2.6101	327.3419
200	Epanechnikov	281.9892	281.9895	2.6123	281.9873
	Gaussian	**175.1921**	**175.1845**	2.6105	**174.483**
	B-weight	283.7447	283.7592	**2.6083**	283.7441
	T-weight	284.9709	284.9629	2.612	284.9611

5.7 Simulation results for tri-modal distribution

Table 5.7, for data generated from tri-modal distribution shows that VKDE using IPI bandwidth and epanichnikov and Gaussian kernel functions have smaller $\overline{AMISE}\ (\hat{f})$ than VKDE with LSCV, SRT and EB bandwidth. But by using the bi-weight and tri-weight kernel functions the situation become different. If bi-weight and tri-weight kernel functions is used then VKDE with EB bandwidth perform well.

Table 5.7 $\overline{AMISE}\hat{f}_{(x)}$ for Tri-modal distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\mathrm{BOOTSTRAP}}\ (EB)$	h_{IPI}
		$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$	$\overline{AMISE}\ (\hat{f})$
50	Epanechnikov	**1.1264**	1.1068	**1.5409**	1.1065
	Gaussian	1.6939	**0.8437**	2.4124	**0.8376**
	B-weight	1.8166	1.7873	1.575	1.7872
	T-weight	2.3836	2.3846	1.6282	2.3834
100	Epanechnikov	**1.1323**	**1.1237**	1.5408	**1.1236**
	Gaussian	1.6904	2.0204	2.4505	1.3818
	B-weight	1.8113	1.8114	1.5805	1.8112
	T-weight	2.3125	2.4131	**1.4195**	2.2924
200	Epanechnikov	1.1335	**1.1333**	**1.5487**	**1.1331**
	Gaussian	1.68055	2.5751	2.3456	1.2175
	B-weight	1.8043	1.8242	1.5834	1.8242
	T-weight	2.2691	2.4273	1.6393	2.2661

5.8 Simulation results for double claw distribution

For double claw distributed sample of any size, the $\overline{AMISE}\ (\hat{f})$ of VKDE using IPI bandwidth and epanechnikov kernel function is lower than the $\overline{AMISE}\ (\hat{f})$ of the remaining three algorithms. All bandwidth selection algorithms worked well with epanechnikov kernel function.

Table 5.8 $\overline{AMISE}\hat{f}_{(x)}$ Double claw distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$
50	Epanechnikov	**1.0534**	**1.0528**	**1.7972**	**1.0527**
	Gaussian	1.404	1.8397	1.45	1.2345
	B-weight	1.7807	1.7941	1.8198	1.7806
	T-weight	2.4198	2.4202	2.496	2.4109
100	Epanechnikov	**1.0653**	**1.0634**	1.8123	**1.0623**
	Gaussian	6.5675	6.0868	2.1629	1.2959
	B-weight	1.7941	1.7812	**1.4366**	1.3942
	T-weight	2.4337	2.4337	2.5096	2.4332
200	Epanechnikov	**1.0735**	**1.0739**	**1.8004**	**1.0733**
	Gaussian	36.5256	35.2181	12.1537	3.6771
	B-weight	1.8068	1.8071	1.831	1.5069
	T-weight	2.447	2.4474	2.5041	2.4469

5.9 Simulation results for asymmetric claw

For asymmetric claw distribution, the VKDE with EB bandwidth and Gaussian kernel function have the smallest $\overline{AMISE}\ (\widehat{f})$ as compared to VKDE with IPI, LSCV and SRT bandwidths as shown in Table 5.9. The SRT and LSCV bandwidths worked well with epanechnikov kernel function while for EB and IPI bandwidths the Gaussian kernel function is the best choice. The $\overline{AMISE}\ (\widehat{f})$ for VKDE continuously fall with increasing sample size.

Table 5.9 $\overline{AMISE}\hat{f}_{(x)}$ Asymmetric Claw distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$
50	Epanechnikov	**1.3857**	**1.385**	0.9697	1.3758
	Gaussian	7.585	17.2398	**0.0969**	**1.2555**
	B-weight	2.0695	2.0693	0.5352	2.0687
	T-weight	2.7302	2.7309	1.3889	2.73
100	Epanechnikov	**1.353**	**1.3531**	0.5833	1.3511
	Gaussian	4.5833	2.5628	**0.0968**	**0.9199**
	B-weight	2.0246	2.025	0.4585	2.0244
	T-weight	2.672	2.6725	0.7203	2.6716
200	Epanechnikov	**1.3466**	1.3465	0.5536	1.3345
	Gaussian	4.4576	**0.9455**	**0.0833**	**0.8452**
	B-weight	2.0021	2.0018	0.4255	2.0017
	T-weight	2.6253	2.5554	0.7152	2.4734

5.10 Simulation results for smooth comb distribution

For smooth comb distributed sample data, VKDE using EB bandwidth and tri-weight kernel function perform well than VKDE with IPI, SRT and LSCV bandwidths methods as shown in Table 5.10. The $\overline{AMISE}\ (\widehat{f})$ for VKDE goes on diminishing with increasing sample size.

Table 5.10 $\overline{AMISE} f^{\wedge}_{(x)}$ For Smooth Comb distribution

N	Kernel function	h_{SRT}	h_{LSCV}	$h_{\text{BOOTSTRAP}}$	$h_{\text{improved plug in}}$
		$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$	$\overline{AMISE}\ (\widehat{f})$
50	Epanechnikov	1.2685	1.2703	0.3108	1.271
	Gaussian	**0.3791**	**0.3617**	0.2968	**0.3579**
	B-weight	1.6074	1.6388	0.2973	1.6128
	T-weight	1.8922	1.9395	**0.2959**	1.88
100	Epanechnikov	1.2615	1.4288	0.3044	1.4347
	Gaussian	**0.4069**	**0.4853**	0.2946	**0.4197**
	B-weight	1.8958	1.9294	0.2822	1.917
	T-weight	2.3072	2.3213	**0.2817**	2.3138
200	Epanechnikov	1.5367	1.5266	0.2918	1.5326
	Gaussian	**0.4262**	**0.4833**	0.2923	**0.4864**
	B-weight	2.1067	2.1116	0.2695	2.1084
	T-weight	2.5888	2.6068	**0.2639**	2.605

5.11 Real data analysis

Our objective here is to compare the VKDE's of real data set using different bandwidth selection algorithms and to check the significance of modes. Therefore we take a real data set. For real data analysis, to see realistic structure of data, we have taken daily opening data of Karachi stock exchange 100 index from 2010 to 2017 with log transform. The reason behind log transformation is to reduce the scale of data. As we know the stock market data has a number of modes so it would be really helpful for us to compare the VKDE, SiZer and SiCon for modes significances with different selected bandwidth methods. For simplicity and necessary for SiZer and SiCon we have taken only Gaussian kernel function.

Figure 5.1 shows the VKDE of KSE-100 opening data using SRT bandwidth and Gaussian kernel function. Usually SRT select large bandwidth and lead to over smooth fit. So, that's why figure 5.1 shows over smooth density of KSE-100 opening data with two long modes. From this figure we infer that the KSE-100 comes from bimodal population.

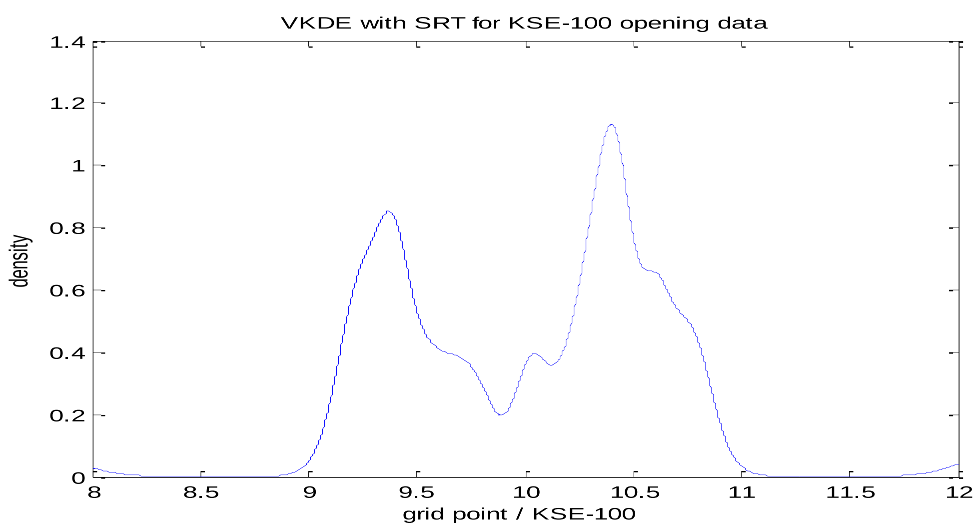


FIG. 5.1 VKDE with Silverman rule of thumb bandwidth algorithm for KSE-100 index opening data

Similarly by contrast to SRT the LSCV algorithm select a small bandwidth and under smooth density curve is obtained. Figure 5.2 depict the VKDE with LSCV bandwidth. This figure is under smooth and indicates that KSE-100 has two large modes, at 9.6 and 10.3, and five small modes.

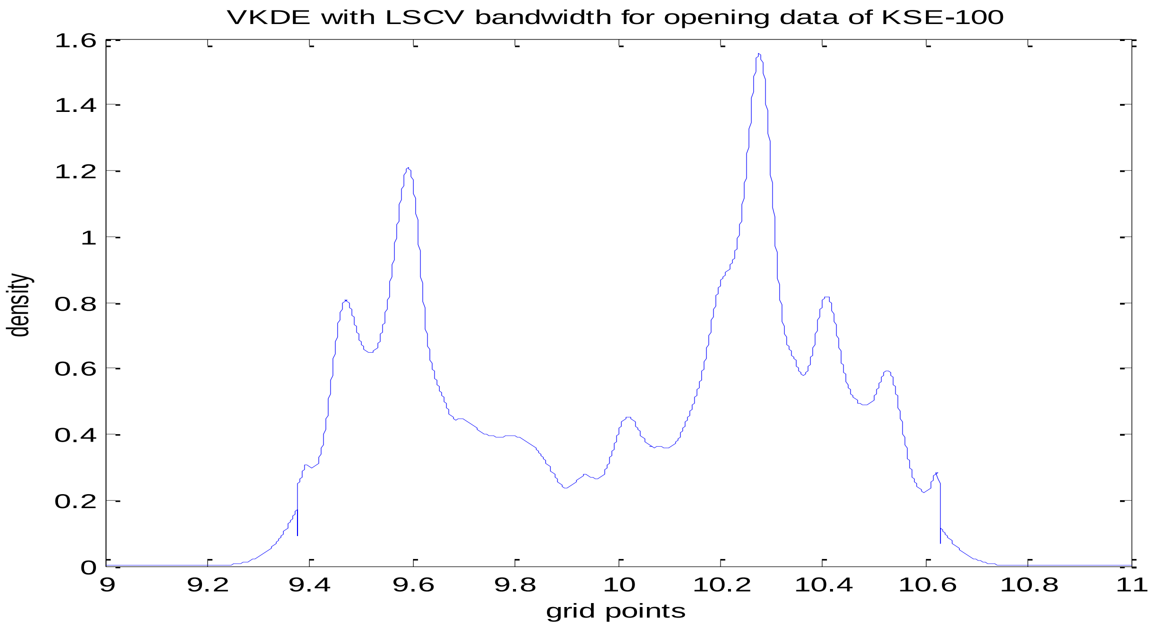


FIG. 5.2 figure 5.2 shows VKDE with LSCV algorithm for KSE-100 index opening data.

Figure 5.3 shows VKDE with IPI (improved plug in) bandwidth selection algorithm. The IPI bandwidth algorithm provides a little bit smaller bandwidth for KSE-100 opening data which lead to under smooth density curve. The density curve shows a large number of big and small modes. We will confirm through SiZer and SiCon, whether these modes are really there or spurious.

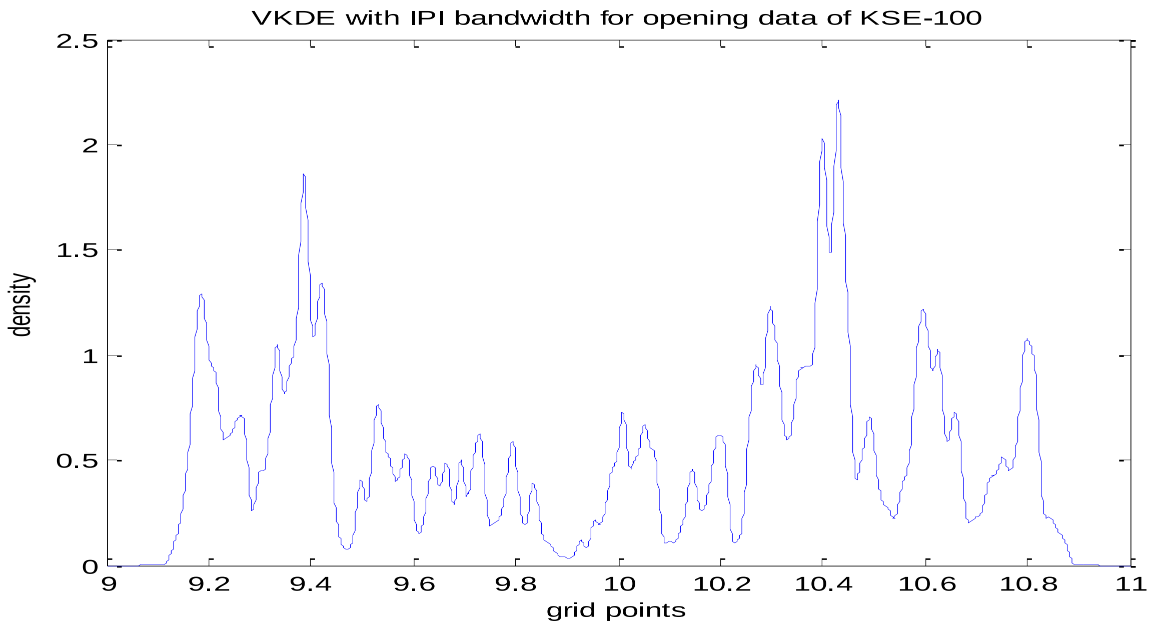


FIG. 5.3 indicate VKDE using IPI bandwidth algorithm for KSE-100 index opening data

VKDE using EB bandwidth for KSE-100 data is given below in figure 5.4. The EB algorithm selected the smallest of all bandwidths for KSE-100 opening data. The variable density fit with EB bandwidth is very under smooth and shows approximately a density on each data point, as given below

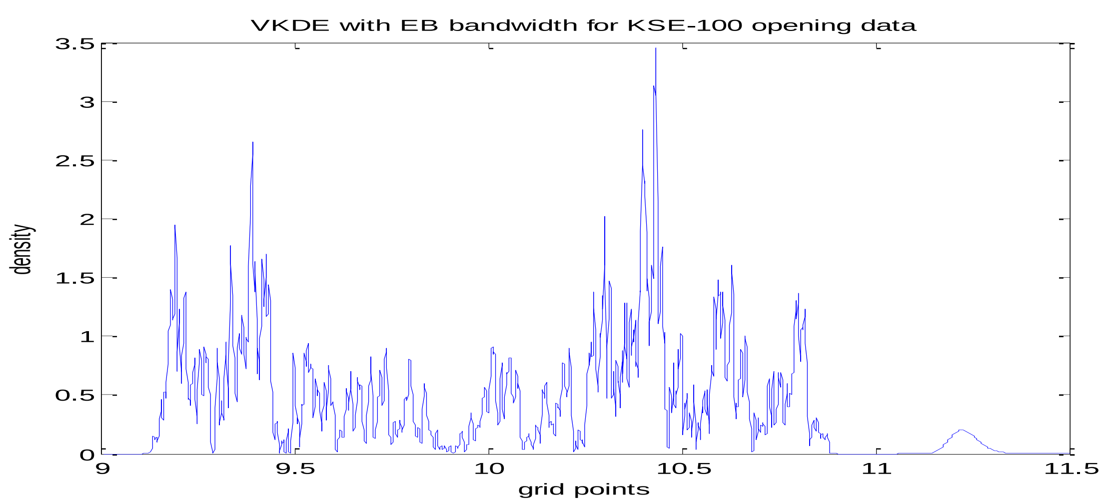


Fig. 5.4 depict VKDE of KSE-100 index opening data using EB bandwidth selection algorithm

Now, in order to check that which bandwidth algorithms does well for VKDE of KSE-100 opening data i.e. which one of the above figures show the realistic structure of KSE-100. It is very arduous to conclude because we cannot use AMISE or empirical distribution as a performance criterion[11]. So to identify that, we move toward the color comparison. Actually the color comparisons are SiZer and SiCon, which provide information about the significance of modes and curvature respectively. SiZer and SiCon maps will clearly show that which of these modes and curvatures are really there and which are noise/ spurious.

5.12 Checking significant modes and curvatures using SiZer and SiCon

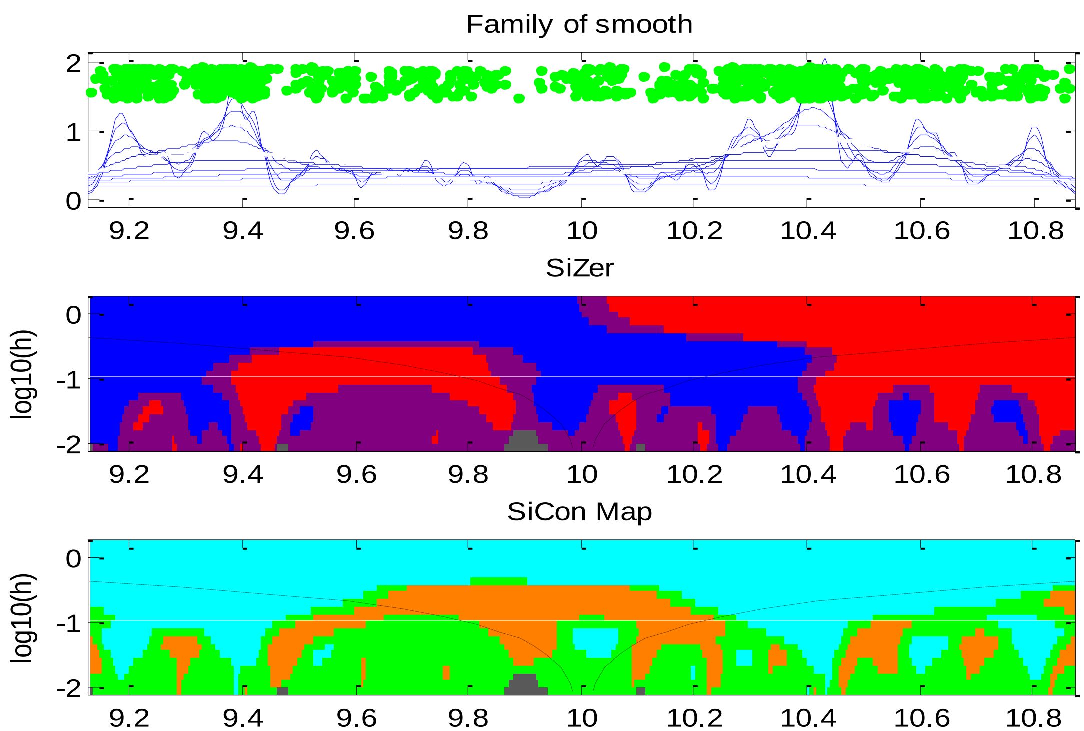
In this section of the study we will show which of the bandwidth selection algorithms perform better using SiZer and SiCon maps for real data set of KSE-100 index. The bandwidth used in SiZer and SiCon map will be consider the best one if it shows more information/ significant modes and curvatures.

5.12.1 SiZer and SiCon maps using SRT bandwidth

The family of smooth figure at the top of map 5.1 showed a number of kernel density curves obtained using a range of bandwidth selected through SRT algorithm. At very small bandwidths there are very large numbers of modes in KSE-100 opening data. At medium size bandwidths it shows five strong modes at 9.2, 9.4, 10.4, 10.6 and 10.8 while at large scales/bandwidth it show only two peaks and one trough i.e. at 9.4, 10.4 and 9.9 respectively. Looking at the results of family of smooth we are completely flummox that which structure of the KSE-100 is the realistic structure, whether the data really has two, five or more than five modes. So to avoid this confusion we move toward SiZer map of map 5.1 using SRT bandwidth. The highlighted white

[11] For real data set it is impossible to estimate AMISE because we don't know the true density of data. We can also not use the empirical distribution the reason is that our data set consist of distinct observations. So for this purpose we compare the performance of selected bandwidths for VKDE using SiZer and SiCon maps.

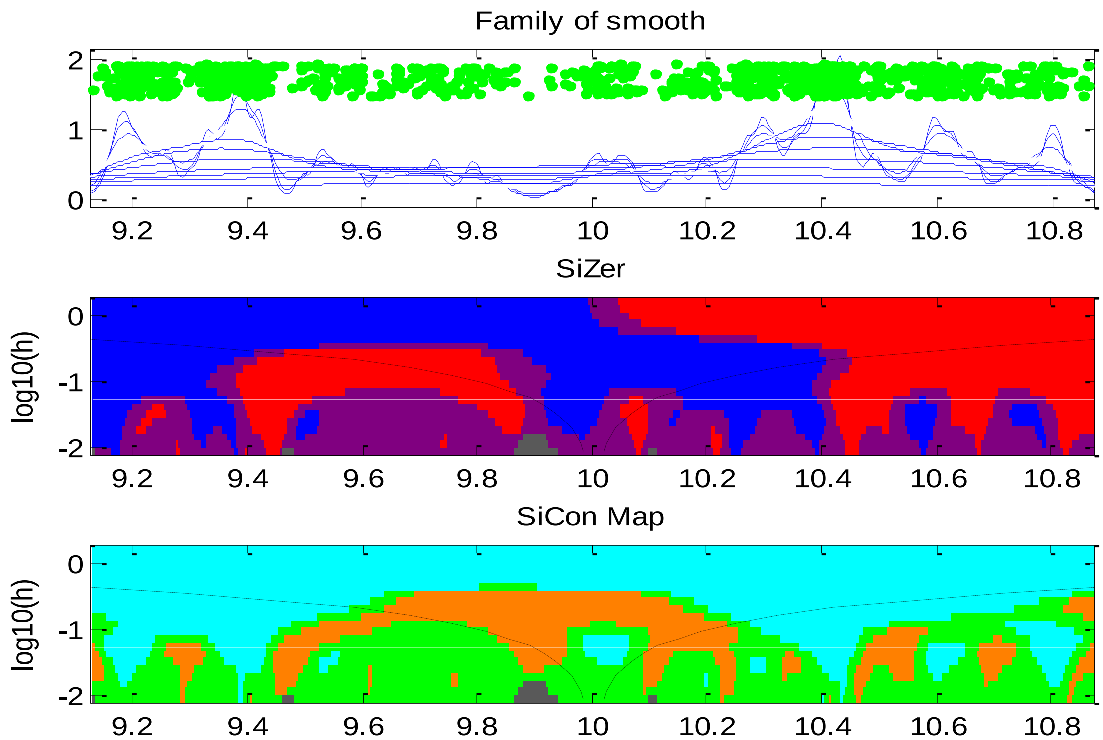
line in SiZer map pass through blue and red colors over points 9.4 and 10.4. The blue color on left and red on right of these points indicate that the peaks are really there. Similarly the same line cross the border of red color and inter into blue over 9.9, means that this trough is also significant. The SiCon, an alternate to SiZer which give more information than SiZer, shows the significant curvature which depend on second derivative. In SiCon map white line pass through cyan-orange- orange-orange-cyan-cyan colors, indicating concavity-convexity-convexity-convexity-concavity-concavity. So SiCon indicate that the structure at 9.4, 9.9 and 10.4 are really there with additional convexity at 10.2 and concavity at 10.8. In net shell we conclude that KSE-100 index opening data with SRT bandwidths range have two significant peaks and one trough.



Maps 5.1 Map 5.1 reflect the family of smooth, SiZer and SiCon maps for KSE-100 index opening data using SRT bandwidth algorithm

5.12.2 SiZer and SiCon using IPI bandwidth

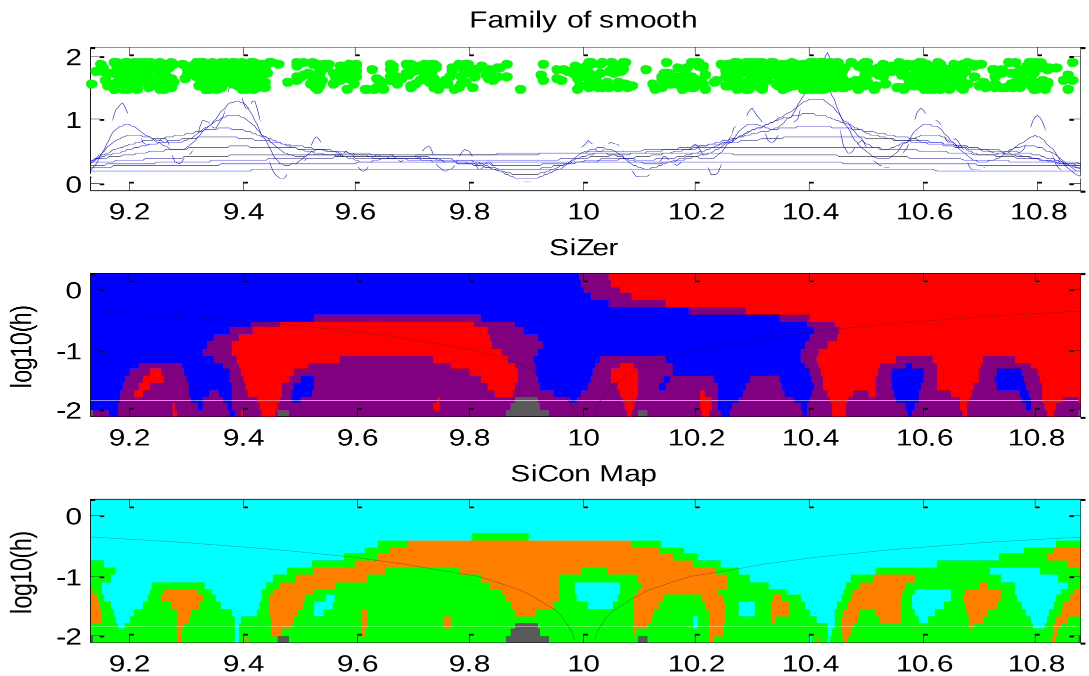
This map 5.2 shows the family of smooth, SiZer and SiCon for KSE-100 opening data by utilizing IPI bandwidth range. Here the smooth family with small resolution/ bandwidths show a wide range of small and high modes while moderate size bandwidth depict seven modes i.e. at 9.2,9.4, 10.1, 10.3, 10.4, 10.6 and at 10.8. The large bandwidths indicate only two peaks at points 9.3 and 10.4 respectively. For confirmation of these feature we jumped to SiZer map. The IPI bandwidth SiZer showed that the peaks at 9.4, 10.1, 10.4 and 10.6 are really there while the remaining are just sampling artifact. The SiZer also show a significant trough at point 9.9. The SiCon map highlighted Orange-cyan-orange-cyan-orange-orange-cyan-orange-cyan-orange-cyan-orange-cyan-orange Pattern, which indicate convexity-concavity-convexity-concavity-convexity-convexity-concavity-convexity-concavity-convexity-concavityconvexity-concavity-convexity. SiCon give us more information about data than SiZer.



Map 5.2. SiZer and SiCon of KSE-100 index opening data with application of IPI bandwidth algorithm

5.12.3 SiZer and SiCon for LSCV bandwidth

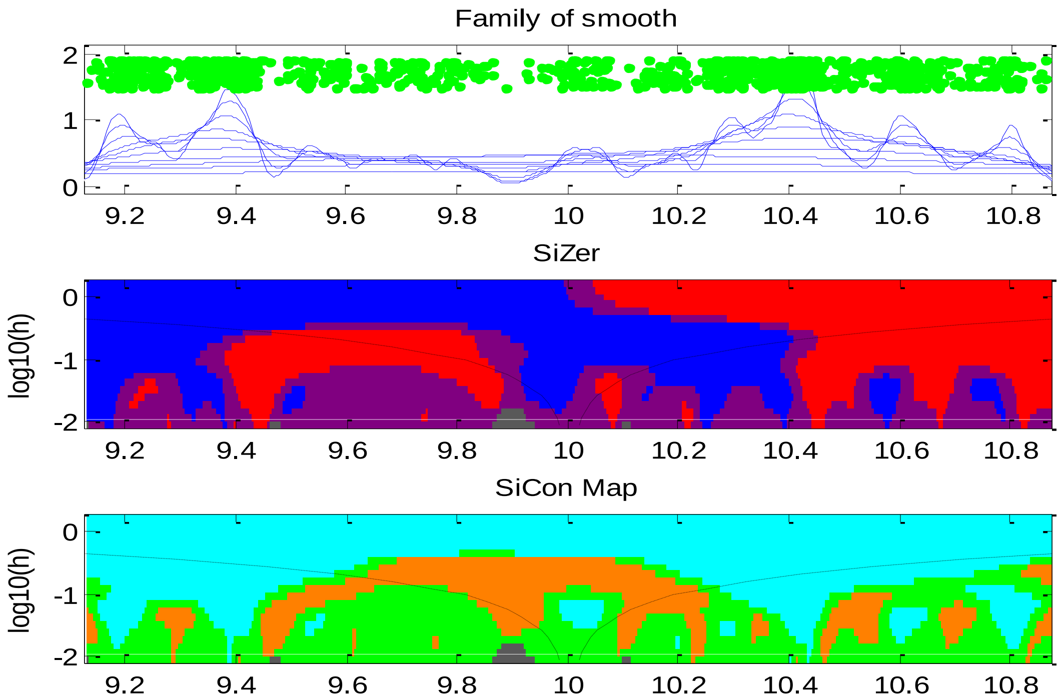
The map 5.3 given below shows the SiZer and SiCon map of KSE-100 opening data using bandwidths range selected through LSCV algorithm. Like IPI SiZer, the SiZer using LSCV bandwidth showed that the modes at 9.4, 10.4, 10.6 and 10.8 are significant but contrary to IPI method the mode at 10.1 and trough at 9.9 is insignificant. The SiCon map showed convexity at 9.3, 9.48, 9.61, 10.1, 10.21, 10.47, 10.83 and concavity at 9.39 and 10.43.



Map 5.3. Color scheme analysis of KSE-100 index opening data using LSCV algorithm

5.12.4 SiZer and SiCon for EB bandwidth

The SiZer and SiCon maps for KSE-100 with EB scales method are given below in map 5.4. SiZer depict three modes over 9.4, 10.4 and 10.6 with a small trough over 10.2. Similarly the more informative SiCon showed upward curvature (convexity) over 9.3, 10.22, 10.43, 10.82 and downward curvature over 9.39 and 10.45.



Map 5.4. Color scheme analysis of KSE-100 index opening data utilizing EB algorithm

Among the SiZer and SiCon maps the one having a large number of significant modes and curvatures is the best one because it gives more information about the structure of data. Analyzing all of the above maps we came to know that the SiZer and SiCon with IPI bandwidth is more informative about the structure of KSE-100 data because it gives more significant modes and curvature as compare to SRT, LSCV and EB algorithms. The KSE-100 data is multimodal data for which the IPI bandwidth worked well as shown by the SiZer and SiCon map. In our simulation study for tri-modal and double claw/ multimodal distribution the IPI bandwidth performed well than other selected bandwidth methods. So these results support our simulation as well as the result of Botve et al (2010). On the basis of these results, we can conclude that IPI algorithm is the best one for multimodal data density estimation.

Chapter 6

Summary, Conclusion and Recommendation

6.1 Summary

Kernel density estimation (KDE) is a well-known and frequently used method for density estimation. Density estimation is used for finding the structure of data. As we mentioned earlier that KDE has two types FKDE and VKDE. VKDE outperform the FKDE and Balloon estimator but still it needs a variety of decisions about the choice of bandwidth and kernel function. There are a large number of bandwidth selection algorithms and kernel functions but the researchers have no general consent over which one bandwidth and kernel function is the best for VKDE. This study focus on the comparison of VKDE using SRT, LSCV, IPI and EB bandwidth selection algorithms for Gaussian mixture models of Marron and Wand (1992). This study also focuses on the performance of tri-weight, bi-weight, Gaussian and epanechnikov kernel functions for VKDE. AMISE is taken as base of comparison. The VKDE with smallest value of AMISE is considered as the best. VKDE is the most attractive way for density estimation but it is a little bit misers in showing which modes and trough are really there in the data set and which are spurious. Therefore, this study too focuses on SiZer and SiCon to find out the significance of modes, trough and curvatures in our real data set.

6.2 Conclusion

From the simulation results of our study we conclude that for Gaussian and kurtotic unimodal distribution the IPI algorithm with epanichnikov kernel function is the best of all selected algorithms. In case of outlier distribution the performance of all methods are not satisfactory but relatively the performance of EB bandwidth with epanechnikov kernel function is appealing. The EB bandwidth does comparatively better than others for bi-modal, skewed and strongly skewed distribution with epanechnikov, Gaussian and Bi-weight kernel function respectively. For sample from tri-modal distribution with different sample sizes the IPI bandwidth using epanechnikov kernel function does batter. While in the same case using bi-weight kernel function the EB bandwidth give better result. The performance of IPI bandwidth for double claw distribution with epanechnikov kernel function is very well compared to others selected bandwidths and thus confirmed the results of Botev et.al (2010).[12] For sample from asymmetric claw and smooth comp distribution the EB bandwidth with Gaussian and bi-weight kernel function performed well than that of other bandwidths respectively. For Gaussian, kurtotic unimodal, skewed unimodal and asymmetric claw distribution the $\overline{AMISE}\,(\widehat{f})$ decrease as the sample sizes increase. Generally for small sample size the variability in $\overline{AMISE}\,(\widehat{f})$ is very high. The IPI bandwidth work well for multimodal data as studied by Botev (2010) is confirmed by our simulation and real data results. Our real data set, KSE-100 opening data, has a number of modes which is confirmed by VKDE, SiZer and SiCon with IPI bandwidth. With the reamaing selected bandwidths (SRT, LSCV and EB) VKDE's, SiZer and SiCon provided us less information about the structure of real data set.

6.3 Recommendation

As we found that the IPI bandwidth work comparatively well for multimodal data therefore, we recommend this algorithms for multimodal variable kernel density estimation and modes significance. Like other studies in literature our study also showed that epanecnikov kernel function is the most efficient and appealing mass function for kernel density estimation.

[12] Kroese et.at (2011) showed that IPI bandwidth selection algorithm is the best one for multimodal data.

Therefore, the researchers have to use it as a weight function in their analysis of density estimation.

As we have mentioned earlier in literature and found empirically that there is no one bandwidth section algorithms which outperformed in each and every case. So, further study should be done in order to find such a bandwidth algorithm which worked well in all situations. Moreover, the sample size can affect the performance of SiZer and SiCon so future research could also be made to analyze the impact of sample size on SiZer and SiCon maps.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, 1217-1223.

Baszczyńska, A. (2005). Some Remarks on the Choice of the Kernel Function in Density Estimation. Acta Universitatis Lodziensis. Folia Oeconomica, (Acta Universitatis Lodziensis. Folia Oeconomica nr 194/2005).

Baxter, M. J., Beardah, C. C., & Westwood, S. (2000). Sample size and related issues in the analysis of lead isotope data. *Journal of Archaeological Science*, *27*(10), 973-980.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. The annals of Statistics, 38(5), 2916-2957.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 353-360.

Breiman, L., Meisel, W., & Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics, 19*(2), 135-144.

Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, *18*(1), 179-189.

Cao, R., Cuevas, A., & Manteiga, W. G. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, *17*(2), 153-176.

Chaudhuri, P., & Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, *94*(447), 807-823.

Chaudhuri, P., & Marron, J. S. (2000). Scale space view of curve estimation. *Annals of Statistics*, 408-428.

DiNardo, J., Fortin, N. M., & Lemieux, T. (1995). *Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach* (No. w5093). National bureau of economic research.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics* (pp. 569-593). Springer New York.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153-158.

Faraway, J. J., & Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association*, *85*(412), 1119-1122.

Ferdosi, B. J., Buddelmeijer, H., Trager, S. C., Wilkinson, M. H. F., & Roerdink, J. B. T. M. (2011). Comparison of density estimation methods for astronomical datasets. *Astronomy & Astrophysics*, *531*, A114.

Fix, E., & Hodges Jr, J. L. (1951). *Discriminatory analysis-nonparametric discrimination: consistency properties*. California Univ Berkeley.

Habbema, J. D. F. (1974). A stepwise discriminant analysis program using density estimetion. In Compstat (pp. 101-110). Physica-Verlag.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *The Annals of Statistics*, 1156-1174.

Hall, P., & Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, *6*(2), 109-115.

Hall, P., & Marron, J. S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. The Annals of Statistics, 163-181.

Hall, P., & Marron, J. S. (1988). Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields*, *80*(1), 37-49.

Hall, P., & Marron, J. S. (1991). Local minima in cross-validation functions. Journal of the Royal Statistical Society. Series B (Methodological), 245-252.

Hall, P., Sheather, S. J., Jones, M. C., & Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 263-269.

Hansen, B. E. (2016). „„Econometrics (University of Wisconsin Department of Economics Manuscript)"".

Hazelton, M. L. (2003). Variable kernel density estimation. *Australian & New Zealand Journal of Statistics*, *45*(3), 271-284.

Hutson, A. D., & Ernst, M. D. (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(1), 89-94.

Jones, M. C. (1990). Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, *32*(3), 361-371.

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.

Katkovnik, V., & Shmulevich, I. (2002). Kernel density estimation with adaptive varying window size. *Pattern recognition letters*, *23*(14), 1641-1648.

Kim, K. D., & Heo, J. H. (2002). Comparative study of flood quantiles estimation by nonparametric models. *Journal of hydrology*, *260*(1), 176-193.

Lemke, D., Mattauch, V., Heidinger, O., Pebesma, E., & Hense, H. W. (2015). Comparing adaptive and fixed bandwidth-based kernel density estimates in spatial cancer epidemiology. International journal of health geographics, 14(1), 15.

Lindeberg, T. (1994). Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, *21*(1-2), 225-270.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in?. *Annals of Statistics*, 415-438.

Loftsgaarden, D. O., & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, *36*(3), 1049-1051.

Mack, Y. P., & Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, *9*(1), 1-15.

Marron, J. S., & Chung, S. S. (2001). Presentation of smoothers: the family approach. *Computational Statistics*, *16*(1), 195-207.

Marron, J. S., & Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, 712-736.

Miecznikowski, J. C., Wang, D., & Hutson, A. (2010). Bootstrap MISE estimators to obtain bandwidth for kernel density estimation. *Communications in Statistics-Simulation and Computation*, *39*(7), 1455-1469.

Minnotte, M. C. (1998). Achieving higher-order convergence rates for density estimation with binned data. *Journal of the American Statistical Association*, *93*(442), 663-672.

Nadaraya, E. A. (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theory of Probability & Its Applications*, *19*(1), 133-141.

Park, B. U., & Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, *85*(409), 66-72.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065-1076.

Rodchuen, M., & Suwattee, P. (2010). Comparison of Kernel Density Estimators. *Thailand Statistician*, *8*(2), 167-181.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832-837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 65-78.

Sain, S. R., & Scott, D. W. (1996). On locally adaptive density estimation. *Journal of the American Statistical Association*, *91*(436), 1525-1534.

Sain, S. R., Baggerly, K. A., & Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, *89*(427), 807-817.

Savchuk, O. Y., Hart, J. D., & Sheather, S. J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, *105*(489), 415-423.

Scott, D. W., & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the american Statistical association*, *82*(400), 1131-1146.

Scott, D. W., Tapia, R. A., & Thompson, J. R. (1977). Kernel density estimation revisited. *Nonlinear Analysis: Theory, Methods & Applications*, *1*(4), 339-372.

Segal, M. R., & Wiemels, J. L. (2002). Clustering of translocation breakpoints. *Journal of the American Statistical Association*, *97*(457), 66-76.

Sheather, S. J. (2004). Density estimation. *Statistical Science*, *19*(4), 588-597.

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

Shimazaki, H., & Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation. *Journal of computational neuroscience*, *29*(1-2), 171-182.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Skrøvseth, S. O., Bellika, J. G., & Godtliebsen, F. (2012). Causality in scale space as an approach to change detection. *PloS one*, *7*(12), e52253.

Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, *76*(4), 705-712.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, *85*(410), 470-477.

Terrell, G. R., & Scott, D. W. (1980). On improving convergence rates for nonnegative kernel density estimators. *The Annals of Statistics*, 1160-1163.

Terrell, G. R., & Scott, D. W. (1985). Oversmoothed nonparametric density estimates. Journal of the American Statistical Association, 80(389), 209-214.

Terrell, G. R., & Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, 1236-1265.

Toran Pour, A., Moridpour, S., & Rajabifard, A. (2015). Using kernel density estimation in GIS to identify accident hot spots in Melbourne tram stops. In *Road Safety and Simulation International Conference* (pp. 62-69). Road Safety and Simulation International Conference.

Tortosa-Ausina, E. (2002). Financial costs, operating costs, and specialization of Spanish banking firms as distribution dynamics. *Applied Economics*, *34*(17), 2165-2176.

Whittle, P. (1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 334-343.

Woodroofe, M. 1967. On the maximum deviation of the sample density. Annals of Math. Stat. 38:475-481.

Woodroofe, M. (1970). On choosing a delta-sequence. *The Annals of Mathematical Statistics*, *41*(5), 1665-1671.

Zambom, A. Z., & Dias, R. (2012). A review of kernel density estimation with applications to econometrics. arXiv preprint arXiv:1212.2812.