THE OPTIMAL BANDWIDTH FOR KERNEL DENSITY ESTIMATION OF SKEWED DISTRIBUTION: A CASE STUDY ON SURVIVAL TIME DATA OF CANCER PATIENTS

Netti Herawati, Khoirin Nisa, Eri Setiawan
Department of Mathematics University of Lampung
Jl. Prof. Dr. Soemantri Brodjonegoro No. 1 Gedung Meneng Bandar Lampung
e-mail: netti.herawati@fmipa.unila.ac.id, khoirin.nisa@fmipa.unila.ac.id, eri.setiawan@fmipa.unila.ac.id .

ABSTRACT

In this paper, optimal bandwidth selection for skewed distribution is studied through data simulation. The data are generated from Exponential (1), Exponential (5), Gamma (1,6), Gamma (1,9), Weibull (1,5), and Weibull (1,10) distributions having parameter(s) that produces a skewed density function with n = 100. The Gaussian kernel density functions of the generated distributions are estimated using Scott (Nrd), Silverman's rule of thumb (Nrd0), Silverman's Long-Tailed distribution (Silverman-LT), Biased Cross validation (BCV) and Sheater-Jones (SJ) bandwidth methods. The kernel density estimates are compared to the corresponding probability density functions. The selected optimal bandwidth then is applied to kernel density estimation of survival time data of cancer patients. Result indicates that, overall, Silverman's Rule of Thumb (Nrd0) method outperformed the other methods.

Keywords:*kernel density estimation,optimal bandwidth, survival timedata*

1. INTRODUCTION

Information about data distribution and its probability density function (PDF) is important in various statistical analysis. However, in some cases the information about the probability density function is unknown. One approach to density estimation is parametric. Another approach is non parametricinthat less rigid assumptions will be made about the distribution of the observed data. The objective of density estimation in nonparametric approach is to obtain the density function curve which is a smooth curve with minimum sampling variance and has importantinformation of the data. The simplest way to estimate the density estimationis using histogram. But it has a weakness in its shape which is influenced by the selection of the starting point and the width of the class interval. Different starting points will produce different histograms, as well as different interval classes will result in different histogram shapes.

In this study, we use nonparametric kernel density estimation to estimate the probability density function of skewed distributions. The most important part of kernel density estimation is the selection of kernel functions and the selection of bandwidth [1,3]. Bandwidth is a scale factor that controls how large the probability of spreading point on the curve. Selection of bandwidth will determine whether the obtained density function curve will be undersmoothing or oversmoothing. The value of the bandwidth that is too small will produce an undersmoothing density function curve, and vice versa, the value of the bandwidth that is too large will produce an oversmoothing density function curve. Therefore an optimal bandwidth is required to obtain a density function curve corresponding to the actual data distribution. To illustrate this issue, we provide an example of undersmoothing and oversmoothing density functions presented in Figure 1.

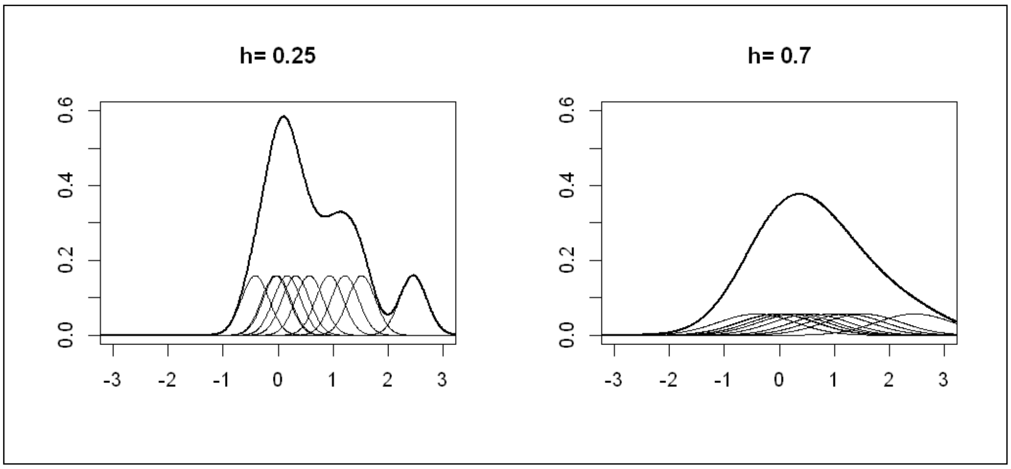

Figure 1.Undersmoothing (left) and Oversmoothing (right).

There are various well known methods available for obtaining optimal bandwidth, for example the Asymptotically first-order optimal (AFO) bandwidths[5], Unbiased Cross Validation (UCV) method, the Biased Cross Validation (BCV) method, Silverman's rule of thumb (Nrd0) method, Silverman's Long-Tailed distribution (Silverman-LT), Scott (Nrd) method, and Sheater-Jones (SJ) method. Each of these methods is known as the optimal bandwidth but produces different bandwidth values. The bandwidth selection in kernel density estimation becomes an interesting topic to study and has been investigated by many authors.

Studies on the optimal bandwidth selection have been done by many authors, one can see e.g. [2]-[6] for various techniques and issues related to bandwidth selections. In this paper we empirically study the optimal bandwidth estimations for skewed distribution by data simulation. We consider several optimal bandwidths mentioned above. The selected one then is applied to real data on survival time of lung cancer patients. The rest of the paper is organized as follow, in Section 2 we review the kernel method for density estimation and present the available optimal bandwidths. In Section 3 we describe our research methodology. The simulation results are presented in Section 4 and the application on survival data is given in Section 5.

2. KERNEL DENSITY ESTIMATION

Kernel Density Estimation is a method to estimate the frequency or probability function of a given value given a random sample. Given a set of observations (*xi*) with $1 \le i \le n,$ it isassumed that the observations are a random sampling of a probability distribution *f*. The kernel estimator of *f* is given as follow:

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where:

- X_i is an identical independently distributed random variable
- K: $R^p \rightarrow R$ is the kernel, a function centered on 0 and that integrates to 1.

- *n* is the number of observations
- *h* is the bandwidth, a positive valued smoothing parameter that would typically tend to 0 when the number of samples tend to ∞

The kernel estimator depends on two parameters, i.e. the kernel function *K* and the bandwidth *h.* There are several kernel functions *K(.)* which can be used for density estimation, some of the most used kernel functions are Epanechikov, biweight, triangular, Gaussian and rectangular kernels. For more details on the kernel function one can see e.g. [1] and [9]. However, the selection of kernel function *K(.)* used for the estimation does not really effect the accuracy of the estimation, furthermore the bias of density estimation using kernel estimator does not rely on the sample size but it depends only on the bandwidth *h* choice. [1]

Selecting an appropriate bandwidth for a kernel density estimator is of crucial importance, and the purpose of the estimation may be an influential factor in the selection method. In many situations, it is sufficient to subjectively choose the smoothing parameter by looking at the density estimates produced by a range of bandwidths. One can start with a large bandwidth, and decrease the amount of smoothing until reaching a ”reasonable” density estimate. However, there are situations where several estimations are needed, and such an approach is impractical. An automatic procedure is essential when a large number of estimations are required as part of a more global analysis.

The problem with using the optimal bandwidth is that it depends on the unknown quantity *f''* which measures the speed of fluctuations in the density *f*, i.e., the roughness of *f*. Many methods have been proposed to select a bandwidth that leads to good performance in the estimation. The followings methods are some optimal bandwidths selection methods available in R software for kernel density estimation:

A. Scott (Nrd) bandwith method

A bandwidth that optimize the Integrated Mean Square Error (IMSE) introduced by Scott is given in the following simple formula :

$$h = 1.06\,\sigma\, n^{-1/5},$$

Where σ is population standard deviation (estimated by the sample standard deviation) and *n*is thesample size. The Scott bandwidth isusually used for normal symmetric and unimodal data [7].

B. Silverman’s rule of thumb (Nrd0) bandwith method

If the data is unimodal but not symmetric, then the following Silverman’s rule of thumb bandwidth will optimize the IMSE :

$$h_{opt} = 0.9\min\{S, IQR/1.34\}n^{-1/5},$$

where S is the sample standard deviation, IQR is the Inter Quartile Range (Q3 - Q1) [7].

C. Silverman's Long-Tailed distribution (Silverman-LT)

Silverman [1] introduced a bandwidth estimator for skewed and long-tailed distribution given in the following formula:

$$h = 0.79\,(\mathrm{IQR})\,n^{-1/5}\;.$$

D. Unbiased Cross Validation(UCV)bandwith method

The UCV bandwidth (h_{UCV}) is the bandwidth h that minimize the following function

$$\mathrm{UCV}(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2h\sqrt{\pi}} \sum\sum_{1\le i<j\le n} \left[\exp\left(\frac{-(x_i-x_j)^2}{4h^2}\right) - \sqrt{8}\exp\left(\frac{-(x_i-x_j)^2}{2h^2}\right)\right].$$

The h_{UCV} valueis obtained iteratively [8].

E. Biased Cross validation (BCV)bandwith method

The BCV bandwidth (h_{BCV}) is the bandwidth h that minimize the following function

$$\mathrm{BCV}(h) = \frac{R(K)}{nh} + \frac{\mu_2(K)^2}{2n^2h} \sum\sum_{1\le i<j\le n} \int K''(w)K''\left(w + \frac{(x_i-x_j)}{h}\right)dw.$$

The h_{BCV}value is also obtained iteratively [8].

F. Sheater-Jones(SJ)bandwith method

Sheater-Jones bandwidth methodis given as follow

$$h = \left[\frac{R(K)}{\mu_2(K)^2\hat{S}_{\mathrm{D}}\left(a_2(h)\right)}\right]^{1/5} n^{-1/5},$$

where $a_2 = \hat{c}_1 h^{5/7}$ with *c*is an appropriate constant [4].

3. RESEARCH METHOD

A simulation study using R program was conducted to compare the several optimal bandwith selection methods: Scott (Nrd), Silverman's rule of thumb (Nrd0), Silverman's Long-Tailed distribution (Silverman-LT), Unbiased Cross Validation(UCV), Biased Cross validation (BCV) and Sheater-Jones (SJ) of n= 100 which were artificially repeated from skewed distributions: Exponential (1), Exponential (5), Gamma (1,6), Gamma (1,9), Weibull (1,5), and Weibull (1,10). Gaussian kernel with selected optimal bandwidth selection method is applied to estimate the density function of survival time data of lung cancer patients consisting of 62 patients measured in days, ie patient duration ranging from 100 days before treatment until the patient died. This data is part of the research data of Veteran Administration USA [10].

4. RESULTS AND DISCUSSION

In this section we present and discuss the result from our data simulation. The simulation results of the generated skewed distributions (Exponential (1), Exponential (5), Gamma (1,6), Gamma (1,9), Weibull (1,5), and Weibull (1,10)) using Nrd0, Nrd, UCV, BCV, SJ, and Silverman-LT bandwidth method scan been seen clearly in Figure 2-7.

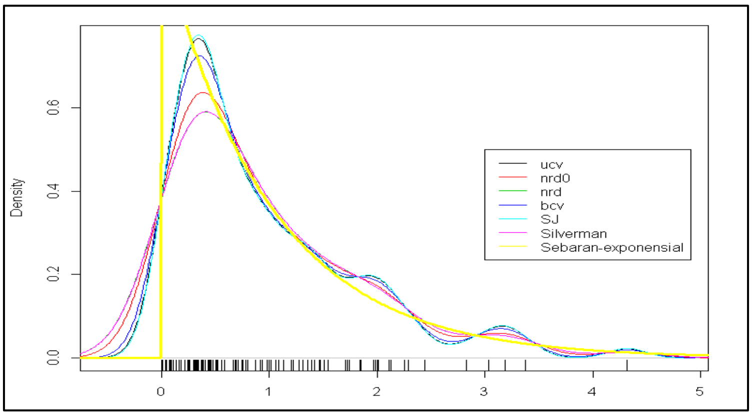

Figure 2. Density estimation curves of Exponential(1) distribution with different bandwidth methods

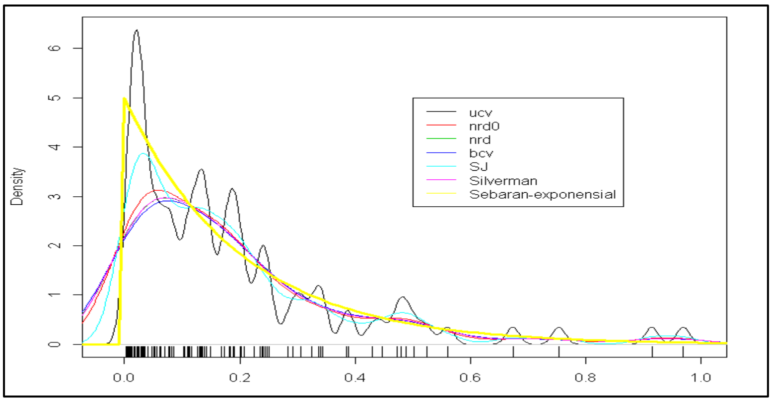

Figure 3. Density estimationcurve of Exponential(5) distribution with different bandwidth methods**.**

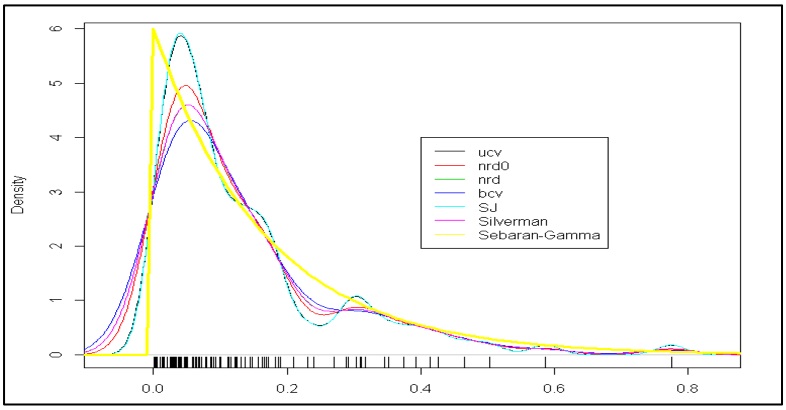

Figure 4. Density estimationcurve of Gamma(1,6) distribution with different bandwidth methods.

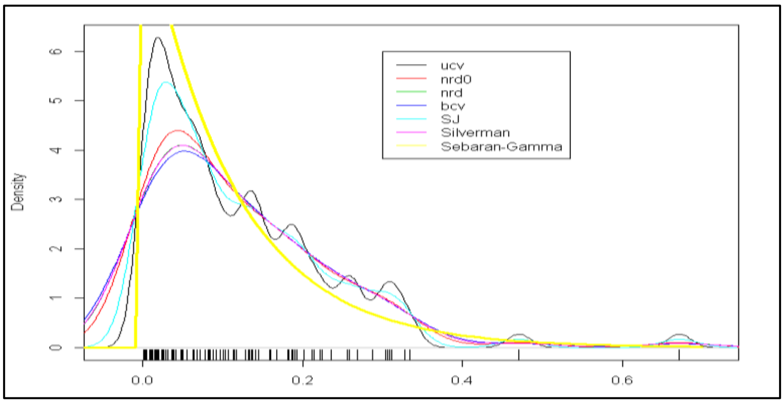

Figure 5. Density estimationcurve of Gamma(1,9) distribution with different bandwidth methods.

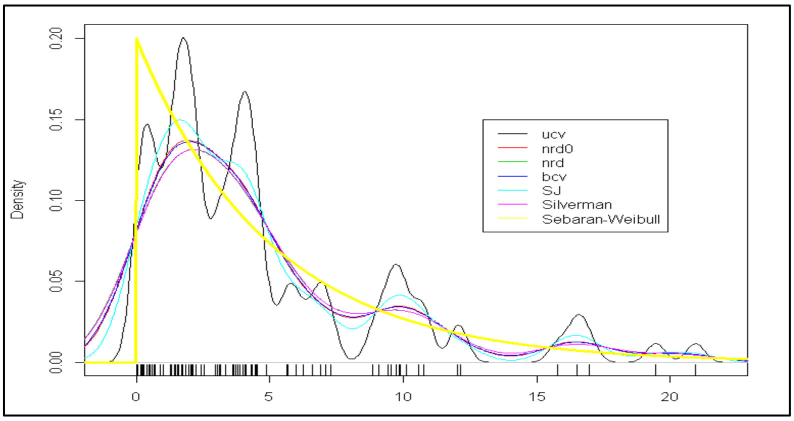

Figure 6. Density estimationcurve of Weibull (1,5) distribution with different bandwidth selection methods.

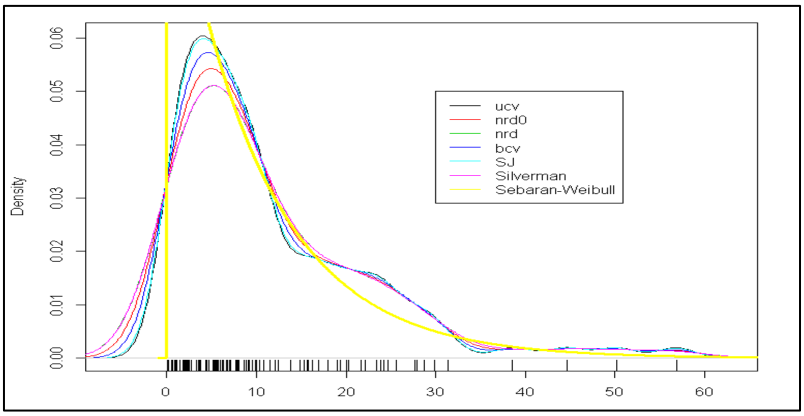

Figure 7. Density estimation curve of Weibull (1,10) distribution with different bandwidth selection methods.

It is apparent from Figure 2-7 that UCV and SJ bandwidths provide unfavorable results for estimating the density curve of the data since the curves are significantly distorted from the real distribution (PDF) curve (ie, yellow curves), particularly in Fig. 3, Fig. 5 and Fig. 6. The kernel estimation using UCV bandwidth is the worst one followed by SJ bandwidth. The bandwiths behaviour of Nrd0, Nrd, Silverman-LT and BCV methods are seen to meet the real PDF curve. The curve of Nrd0 method has a shape similar to the curve of Silverman-LT, while the curve of the Nrd method has a shape similar to the BCV curve. From these two groups, it is obvious that the Nrd0 andSilverman-LT curves approximate the actual data curve better than the Nrd and BCV curves.In order to ensure the best bandwidth estimation methods, specifically we provide the curve of the density of the two bandwidth selection methods (Nrd0 and Silverman-LT) for Exponential (1), Exponential (5), Gamma (1,6), Gamma (1,9), Weibull (1,5), and Weibull (1,10) distributions as shown Fig. 8.

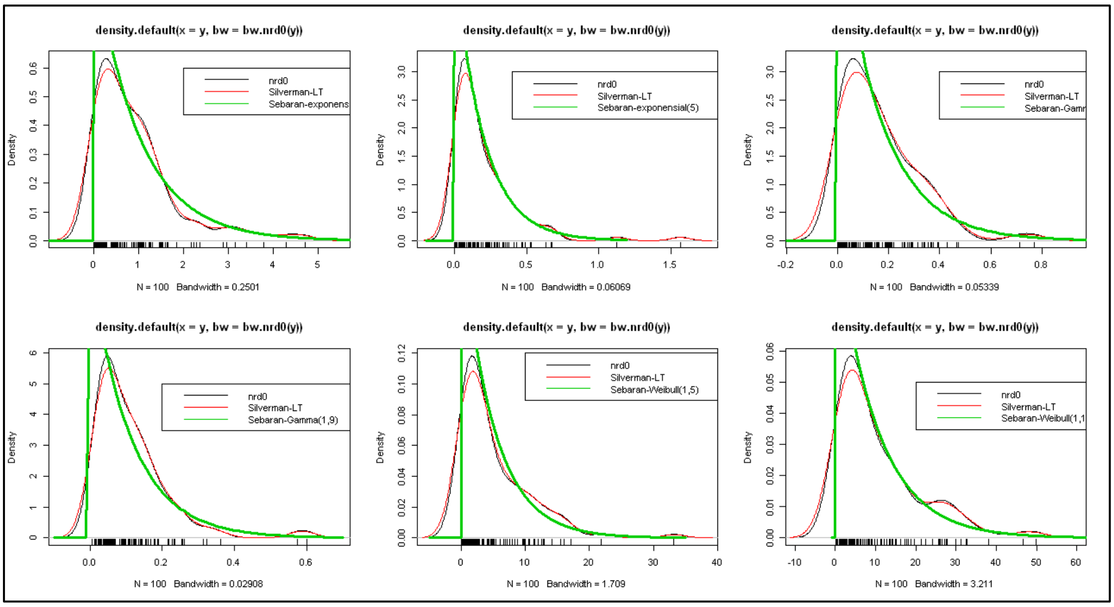

Figure 8. Comparison of Nrd0 and Silverman-LT bandwidth selection methods for Exponential (1), Exponential (5), Gamma (1,6), Gamma (1,9), Weibull (1,5), and Weibull (1,10) distributions.

Figure 8 shows that the density curves of the Nrd0 and Silverman-LT methods are similar. However, at its peak, the density curve of the Nrd0 bandwidth method approximates to the real data distribution (PDF) curve better than the density curve of the Silverman-LT method. Based on the above description, it can be concluded that the Nrd0 bandwidth estimation method gives the best density curve estimation among other methods.

5. APPLICATION ON SURVIVAL DATA OF CANCER PATIENTS

To see the performance of the selected bandwidth method resulting from the simulation study above, we apply it to the survival time data of cancer patients. First, we predict the data distribution using histogram and polygon as shown in Figure 9.

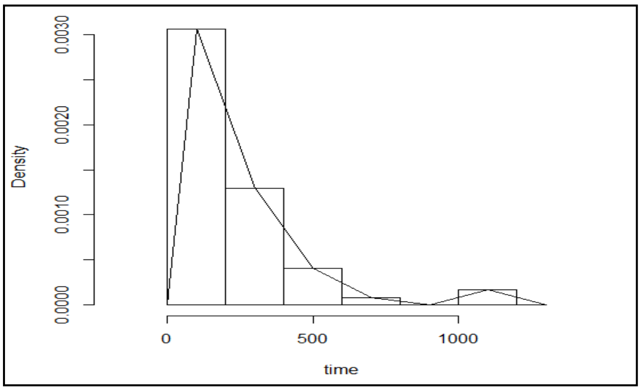

Figure 9. Histogram of survival time data of cancer patients

The histogram above shows that the data is skewed to the right. The values of all optimal bandwidths discussed in previous section are presented in Table 1.

Table 1. Optimal bandwith values of all bandwith selection methods

Methods	Bandwidth
Silverman's Rule of Thumb (Nrd0)	34.35
Scott (Nrd)	40.46
Unbiased Cross Validation (UCV)	14.37
Biased Cross Validation (BCV)	44.85
Sheater – Jones (SJ)	20.39
Silverman Long-Tailed distributions (Silverman-LT)	36.72

We obtained the bandwidth values of the two best methods (Nrd0 and Silverman-LT) are 34.35 and 36.72 respectively. We present the plot of the polygon curve and kernel density estimates using Nrd0 and Silverman-LT bandwidths in Figure 10.

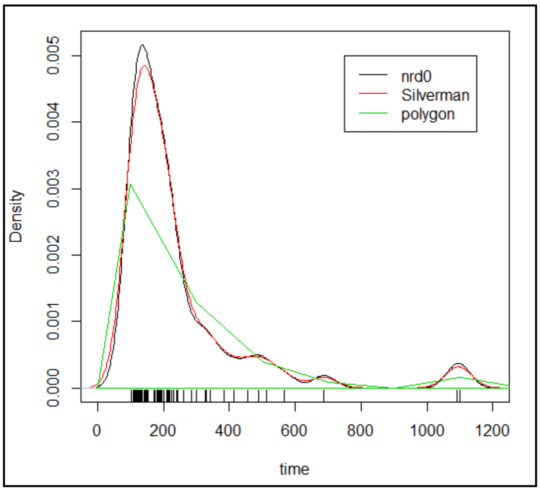

Figure 10. Comparison kernel density estimates using Nrd0 and Silverman-LT bandwidths and polygon of survival time data of cancer patients.

From Figure 10, it can be seen that Nrd0 and Silverman-LT bandwidth methods yields the density curves that best suits the distribution of survival time data of cancer patients. The Nrd0 and Silverman-LT density curves look similar. However, at the peak and tail of the densities are quite different. For final decision, we choose the best bandwidth to fit the data in accordance to the result of data simulation, i.e. the Nrd0 bandwidth. This results is similar to the skewed distributions simulation result as shown in Figure 8.

6. CONCLUSION

In this paper, we have shown that one can estimate the density curve using kernel density estimation method using various bandwith selection methods. For skewed distribution considered in this paper, overall, Silverman's Rule of Thumb (Nrd0) bandwith outperformed all other methods. It provides the best density estimates curvefor both skewed distribution and survival time data of cancer patients compared to others.

.

REFERENCES

[1] Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[2] Samworth, R.J. & Wand M.P. 2010. Asymptotics and Optimal Bandwidth Selection for Highest Density Region Estimation. *The Annals of Statistics*, **38**(3), 1767–1792.

[3] Chen., S. 2015. Optimal Bandwidth Selection for Kernel Density Functional Estimation. *Journal of Probability and Statistics*, **2015**:1-21.

[4] Sheather, S.J. &Jones,M.C. 1991. A Reliable Data-Based Bandwith Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society*, **53** (3), 683-690.

[5] Arai, Y. & Ichimura, H. 2013. Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to Sharp Regression Discontinuity Design. GRIPS discussion paper. National Graduate Institute for Policy Studies, Tokyo Japan.

[6] Hansen, B.E. 2004. *Bandwidth Selection for Nonparametric Distribution Estimation. Research supported by the National Science Foundation*. Department of Economics University of Wisconsin, Madison.

[7] Rizzo, M.L. 2008. *Statistical Computing with R*. Chapman & Hall/CRC. Boca Raton.

[8] Vrahimis, A. 2010. Smoothing Methodology With Applications To Nonparametric Statistics. thesis submitted to the University of Manchester, School of Mathematics.

[9] Sheather, S.J. 2004. Density Estimation. *Statistical Science*, **19**(4), 588-597.

[10] Prentice, R.L. 1973. Exponential Survival with Censoring and ExplanatoryVariables. *Biometrika*, **60**, 279-288.