Tanzania Journal of Science 45(1): 1-8, 2019 ISSN 0856-1761, e-ISSN 2507-7961

A Key Note on Performance of Smoothing Parameterizations in Kernel Density Estimation

Israel U. Siloko[1], Osayomore Ikpotokin[2] and Edith A. Siloko[3]

[1]*Edo University Iyamho, Department of Mathematics and Computer Science, P.M.B. 04, Auchi, Nigeria. E-mail: siloko.israel@edouniversity.edu.ng, suzuazor@yahoo.com*

[2]*Ambrose Alli University, P.M.B. 14, Ekpoma, Department of Mathematics and Statistics, Nigeria. E-mail: ikpotokinosayomore@yahoo.co.uk*

[3]*University of Benin, Department of Mathematics, P.M.B. 1154, Benin City, Nigeria. E-mail: myakpeski@yahoo.com*

Abstract

The univariate kernel density estimator requires one smoothing parameter while the bivariate and other higher dimensional kernel density estimators demand more than one smoothing parameter depending on the form of smoothing parameterizations used. The smoothing parameters of the higher dimensional kernels are presented in a matrix form called the smoothing matrix. The two forms of parameterizations frequently used in higher dimensional kernel estimators are diagonal or constrained parameterization and full or unconstrained parameterization. While the full parameterization has no restrictions, the diagonal has some form of restrictions. The study investigates the performance of smoothing parameterizations of bivariate kernel estimator using asymptotic mean integrated squared error as error criterion function. The results show that in retention of statistical properties of data and production of smaller values of asymptotic mean integrated squared error as tabulated, the full smoothing parameterization outperforms its diagonal counterpart.

Keywords: Smoothing Matrix, Kernel Estimator, Integrated Variance, Integrated Squared Bias, Asymptotic Mean Integration Squared Error (AMISE).

Introduction

Nonparametric density estimation techniques are of wide applications with the kernel density estimator playing vital statistical roles in data analysis. Kernel estimation is a data smoothing method where inferences and conclusions are made about a set of observations. As a nonparametric method, kernel density estimation is a very useful tool for analysis and visualization of the distribution of observations (Simonoff 2012). The kernel estimator is one of the popular nonparametric techniques in density estimation. The univariate kernel estimator is of the form

$$\hat{f}(\mathrm{x}) = \frac{1}{nh_{\mathrm{x}}}\sum_{i=1}^{n} K\left(\frac{\mathrm{x}-\mathrm{X}_i}{h_{\mathrm{x}}}\right), \qquad (1)$$

where $K(\cdot)$ is a kernel function, h_{x} is smoothing parameter also called bandwidth, X_i are observations or measurements obtained from real life and n is sample size. The kernel function determines the shape of the resulting estimates while the smoothing parameter regulates the level of smoothing apply on the kernel estimator. The kernel function is a non-negative function that satisfies the following conditions.

$$\begin{cases} \int K(\mathrm{x})d\mathrm{x} = 1, \\ \int \mathrm{x}K(\mathrm{x})d\mathrm{x} = 0 \quad \text{and} \\ \int \mathrm{x}^2K(\mathrm{x})d\mathrm{x} = \mu_2(K) \neq 0\,. \end{cases} \qquad (2)$$

The first condition in equation (2) implies that the kernel function must integrate to unity, therefore most kernel functions are probability density functions; the second condition simply states that the average of the kernel is zero, while the third condition means that the variance of the kernel function denoted by $\mu_2(K)$ is not equal to zero (Scott 1992).

The bivariate kernel density estimator occupied a unique position of bridging the univariate kernel estimator and other higher dimensional kernel estimators (Duong and Hazelton 2003). In bivariate kernel density estimation, x, y are taken to be the random variables assuming values in $\mathcal{R}^2$ and they have a joint density function $f(\mathrm{x}, \mathrm{y}), (\mathrm{x}, \mathrm{y}) \in \mathcal{R}^2$ with $\mathrm{X}_i, \mathrm{Y}_i,\ i = 1, 2, \ldots, n$ being the set of observations of size n drawn from the distribution. The bivariate kernel density estimate of $f(\mathrm{x}, \mathrm{y})$ based on this sample is of the form

$$\hat{f}(\mathrm{x}, \mathrm{y}) = \frac{1}{n h_{\mathrm{x}} h_{\mathrm{y}}} \sum_{i=1}^{n} K\left(\frac{\mathrm{x} - \mathrm{X}_i}{h_{\mathrm{x}}}, \frac{\mathrm{y} - \mathrm{Y}_i}{h_{\mathrm{y}}}\right), \quad (3)$$

where $h_{\mathrm{x}} > 0$ and $h_{\mathrm{y}} > 0$ are smoothing parameters in X and Y axes and $K(\mathrm{x}, \mathrm{y})$ is a bivariate kernel function which is usually the product of two univariate kernels. The bivariate kernel density estimates are simple to understand and interpret, either as surface plots (wire frames) or contour plots. The bivariate kernel estimator in equation (3) can also be written as

$$\hat{f}(\mathrm{x}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_{\mathrm{x}}} K\left(\frac{\mathrm{x} - \mathrm{X}_i}{h_{\mathrm{x}}}\right) \frac{1}{h_{\mathrm{y}}} K\left(\frac{\mathrm{y} - \mathrm{Y}_i}{h_{\mathrm{y}}}\right). \quad (4)$$

The kernel estimator in equation (3) and equation (4) is a useful tool for data exploratory analysis and data visualization especially for bivariate data when $\hat{f}(\mathrm{x}, \mathrm{y})$ can be visualized using the familiar perspectives or contour plots (Scott 1992, Simonoff 2012, Silverman 2018). Other areas of applications of kernel density estimator are nonparametric discriminant analysis, intensity function estimation and goodness-of-fit testing (Duong and Hazelton 2003). It is generally known that the choice of smoothing parameter is very important to the performance of $\hat{f}(\mathrm{x}, \mathrm{y})$ either in the constrained form or the unconstrained form (Liu et al. 2011, Siloko et al. 2018).

Examination of performance of kernel density estimator using diagonal smoothing matrix and full smoothing matrix with emphasis on bivariate kernel employing the asymptotic mean integrated squared error as error criterion is presented in this paper. The asymptotic mean integrated squared error of the univariate and bivariate cases were discussed with the forms of parameterizations. A comparative study of forms of parameterizations with real data example was investigated and results showing the inherent statistical properties of the data and also producing smaller AMISE value with the full smoothing parameterization.

Methodology

The methodology behind the derivation of the expression of the asymptotic mean integrated squared error for kernel density estimation lie in the application of Taylor's series expansion of the kernel function. The estimate of $\hat{f}(\mathrm{x})$ in equation (1) is measured by the asymptotic mean integrated squared error. An asymptotic approximation of equation (1) using Taylor's series expansion yields the integrated variance and the integrated squared bias given by

$$\begin{cases} IV = \dfrac{R(K)}{nh} \\ ISB = \dfrac{h_{\mathrm{x}}^4}{4} \mu_2(\mathrm{K})^2 \, R(f_{\mathrm{xx}}), \end{cases} \quad (5)$$

where $R(K)$ is roughness of kernel function, $\mu_2(\mathrm{K})^2$ is variance of kernel and $R(f_{\mathrm{xx}}) = \int f_{\mathrm{xx}}(\mathrm{x})^2 d\mathrm{x}$ is the roughness of unknown probability density function (Scott 1992, Guidoum 2015). The combination of the terms in equation (5) will yield an estimate of asymptotic mean integrated squared error given as

$$AMISE = \frac{R(K)}{nh} + \frac{h_{\mathrm{x}}^4}{4}\mu_2(\mathrm{K})^2\, R(f_{\mathrm{xx}}). \quad (6)$$

The minimum of the AMISE is the solution to the differential equation

$$\frac{\partial}{\partial h} AMISE(h) = \frac{-R(K)}{nh_{\mathrm{x}}^2} + \mu_2(\mathrm{K})^2 h_{\mathrm{x}}^3 R(f_{\mathrm{xx}}) = 0.$$

Therefore, the smoothing parameter that minimizes the AMISE of the kernel estimator is given by

$$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{\mu_2(\mathrm{K})^2 R(f_{\mathrm{xx}})n}\right]^{\left(\frac{1}{4+d}\right)}, \quad (7)$$

where d is dimension of kernel and n is sample size.

Similarly, the asymptotic mean integrated squared error of the bivariate kernel is

$$AMISE = \frac{R(K)^d}{nh_{\mathrm{x}}h_{\mathrm{y}}} + \frac{h_{\mathrm{x}}^4}{4}\mu_2(\mathrm{K})^2 R(f_{\mathrm{xx}}) + \frac{h_{\mathrm{y}}^4}{4}\mu_2(\mathrm{K})^2 R\left(f_{\mathrm{yy}}\right), \quad (8)$$

where $R(K)$ is roughness of kernel, $\mu_2(\mathrm{K})^2$ is variance of kernel and $R(f_{\mathrm{xx}}) = \int\int f_{\mathrm{xx}}(\mathrm{x},y)^2 d\mathrm{x}d\mathrm{y}$, $R\left(f_{\mathrm{yy}}\right) = \int\int f_{\mathrm{yy}}(\mathrm{x},y)^2 d\mathrm{x}d\mathrm{y}$ are the roughnesses of the unknown probability density function. The choice of smoothing parameters also known as smoothing matrices in bivariate kernel is strictly based on the complexity of the underlying density and the number of parameters to be estimated. In practice, the commonest parameterizations are diagonal parameterization and full smoothing parameterization. If the product kernel is employed, then the smoothing parameters that will minimize the AMISE in equation (8) denoted by $h_{AMISE-\mathrm{x}}$ and $h_{AMISE-\mathrm{y}}$ are of the form

$$\begin{cases} h_{AMISE-\mathrm{x}} = \left[\dfrac{R(K)^d}{\mu_2(\mathrm{K})^2 R(f_{\mathrm{xx}})}\right]^{\left(\frac{1}{4+d}\right)} \times n^{-1/(4+d)} \\ \text{and} \\ h_{AMISE-\mathrm{y}} = \left[\dfrac{R(K)^d}{\mu_2(\mathrm{K})^2 R\left(f_{\mathrm{yy}}\right)}\right]^{\left(\frac{1}{4+d}\right)} \times n^{-1/(4+d)}. \end{cases} \quad (9)$$

As observed in the smoothing parameter that minimizes the AMISE of the univariate kernel, the expressions in equation (9) contain the second derivatives of the unknown density f being estimated and this will require some approximations. The order of the smoothing parameter obtained from equation (9) is $n^{-1/(d+4)}$. The full smoothing parameterization requires $\frac{d(d+1)}{2}$ smoothing parameters where d is dimension of kernel and the order is same as that of equation (9). The compact form of the smoothing parameter that minimizes the AMISE in the case of the full smoothing parameterization is of the form

$$h_{AMISE} = \left[\frac{dR(K)^d}{\mu_2(\mathrm{K})^2 R(\nabla^2 f(\mathrm{x},\mathrm{y}))n}\right]^{\left(\frac{1}{4+d}\right)}, \quad (10)$$

where

$$R(\nabla^2 f(\mathrm{x},\mathrm{y})) = \int\int \left(\frac{\partial^2 f(\mathrm{x},\mathrm{y})}{\partial \mathrm{x}^2 \partial \mathrm{y}^2}\right)^2 d\mathrm{x}d\mathrm{y},$$

$$\nabla^2 f(\mathrm{x},\mathrm{y}) = \left(\frac{\partial^2 f(\mathrm{x},\mathrm{y})}{\partial \mathrm{x}^2 \partial \mathrm{y}^2}\right)$$

is roughness of the unknown probability density function.

The choice of a kernel function is not a difficult task because most kernel functions are probability density functions. In this paper, the standard normal kernel was employed because it produced smooth density estimates and simplified the mathematical computations. The standard normal kernel function of the bivariate kernel estimator given in equation (4) is of the form

$$K(\mathrm{x},\mathrm{y}) = \frac{1}{2\pi}\exp\left(-\frac{\mathrm{x}^2+\mathrm{y}^2}{2}\right). \quad (11)$$

The matrix form of the diagonal parameterization and full smoothing parameterization of the bivariate kernel estimator given in equation (4) above are

$$H = \begin{bmatrix} h_x & 0 \\ 0 & h_y \end{bmatrix} \text{ and } H = \begin{bmatrix} h_x & h_{xy} \\ h_{xy} & h_y \end{bmatrix} \quad (12)$$

The diagonal form of smoothing parameterization considers only the elements of the leading diagonal of the smoothing matrix while the off diagonal elements are zero while the full smoothing parameterization takes into consideration all the elements as shown in equation (12). The performance of these forms of parameterizations will be compared using the asymptotic mean integrated squared error (AMISE) as the error criterion function.

Results

This section is about comparing the performance of the diagonal smoothing matrix with the full smoothing matrix using a real data example. Two data set will be examined; a univariate case and a bivariate case. The univariate case requires one smoothing parameter; hence there will be no comparison in terms of performance. The smoothing matrix that minimizes the asymptotic mean integrated squared error (AMISE) in the case of diagonal smoothing matrix is represented by $H_{D-AMISE}$, while the full smoothing matrix is represented by $H_{F-AMISE}$. It is observed that in both parameterizations, the smoothing matrices obtained are usually symmetric.

The first data set examined is the lengths of 86 spells (in days) of psychiatric treatment undergone by patients used as controls in a study of suicide risks (Silverman 2018). The data are log transformed and treated as observations on the interval (–2, 10). Figure 1 is the kernel estimate for the suicide study data and the estimate presents the data to be bimodal.

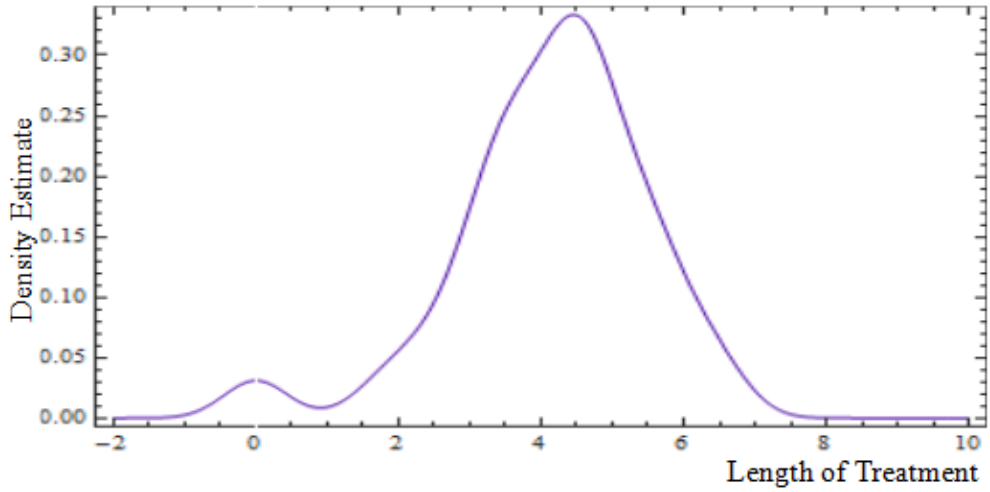

Figure 1: Kernel estimate of length of treatment data (Days).

The smoothing parameter that minimizes the AMISE and its value is in Table 1 and there is no comparison of performance since univariate kernel requires a single bandwidth.

The second data set examined is the waiting time between eruptions and the duration of the eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA (Azzalini and Bowman 1990). The data set is made up of 272 observations on two variables in which variable X represents the duration of the eruption, while variable Y represents the waiting time between eruptions. Figure 2

shows the scatterplots of the Old Faithful data while Figure 3 and Figure 4 show the kernel estimates (surface plots and contour plots) of the two forms of smoothing parameterizations using the bivariate normal kernel.

One fundamental step to observe in the examination of bivariate data set is to consider the scatterplots of the bivariate data but often times, while density estimate will reveal or highlight important features, scatterplots cannot play this vital role (Siloko et al. 2018). Scatterplots have been regarded as the most frequently used tools for graphically displaying bivariate data sets but with the serious disadvantage that the observers are only drawn to the peripheries of the data cloud, while significant structures in the main body of the data will be hidden by the high density of points (Wand and Jones 1995). In kernel density estimates, these disadvantages are eliminated because they have an advantage of presentation of information regarding the distribution of the data set. As noted from the scatterplots of the Old Faithful data, the modes were not as apparent from the scatterplots like the kernel estimates and this exemplifies the usefulness of bivariate density estimates for highlighting structure. One very important point to note from the kernel estimates of the Old Faithful data is that it is bimodal and this provides evidence in favour of eruption times and the time interval until the next eruption exhibiting a bimodal distribution.

Table 1: Variance, bias^2 and *AMISE* for treatment data set

Bandwidth	*Variance*	*Bias*^{2}	*AMISE*
0.448953	0.0073062666	0.0018265666	0.0091328332

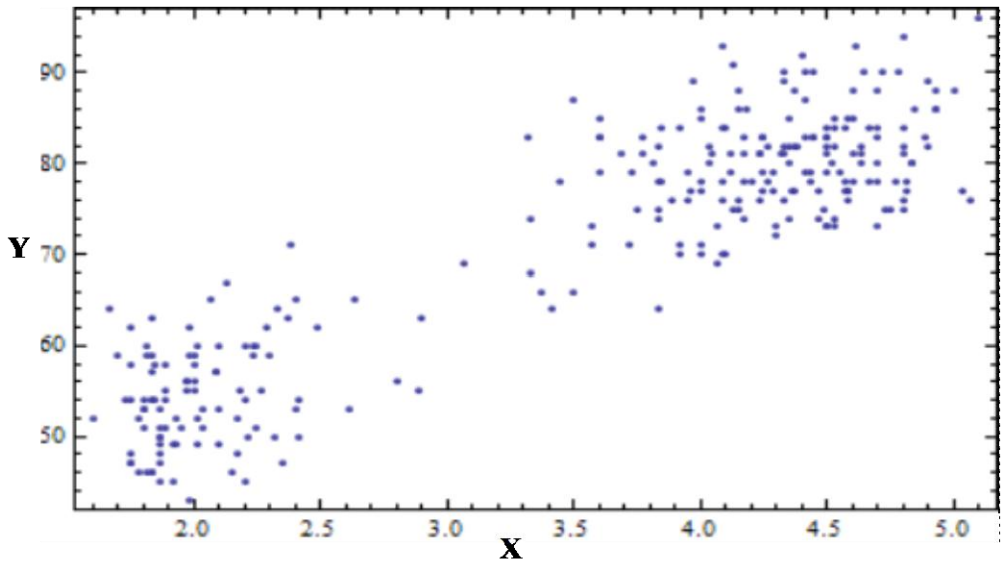

Figure 2: Scatterplot of Old Faithful data.

The scatterplots show a strong relationship between the variables and this is obvious that the time interval until the next eruption is positively correlated with the duration of the eruption. The data were standardized in order to obtain equal variances in each dimension because in most multivariate statistical analysis, the data

should be standardized in order to make sure that the difference among the ranges of variables will disappear (Sain 2002, Simonoff 2012). The smoothing matrices for the forms of parameterizations for this data are

$$H_{D-AMISE} = \begin{bmatrix} 0.431046 & 0.000000 \\ 0.000000 & 0.423014 \end{bmatrix} \text{ and } H_{F-AMISE} = \begin{bmatrix} 0.48383320 & 0.00108843 \\ 0.00108843 & 0.47481764 \end{bmatrix}$$

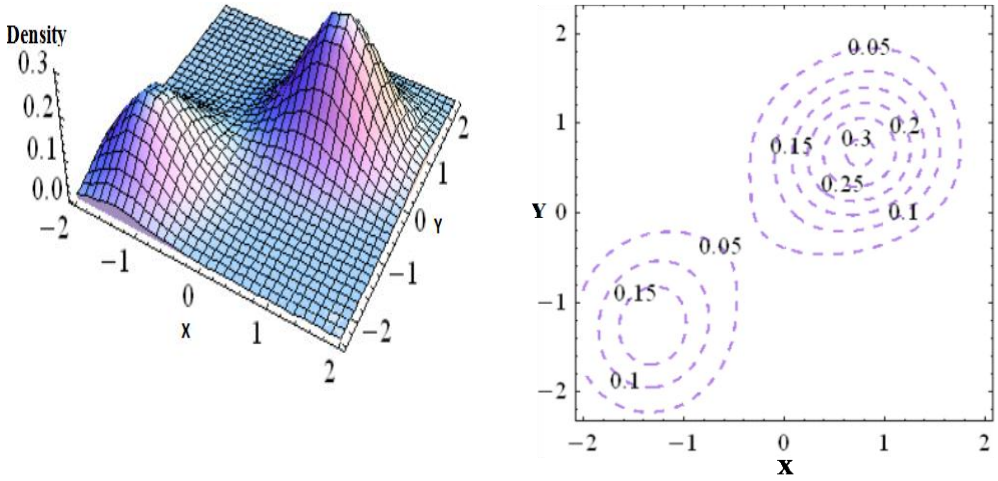

Figure 3: Kernel estimates (surface and contour plots) of H_D *smoothing* matrix.

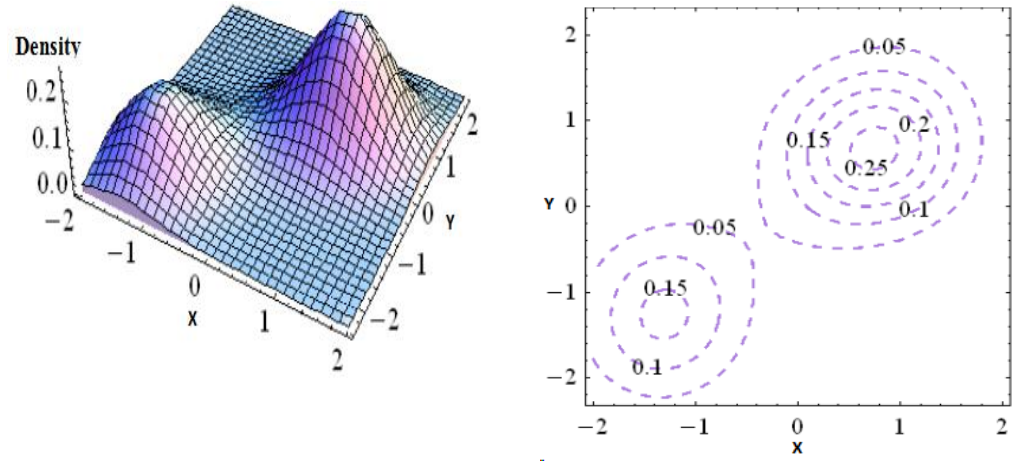

Figure 4: Kernel estimates (surface and contour plots) of H_F *smoothing* matrix.

Table 2 shows the asymptotic integrated variance (AIV), asymptotic integrated squared bias (AISB) and the asymptotic mean integrated squared error (AMISE) of the Old Faithful data. The analysis presented in Table 2 clearly shows that the full smoothing matrix did better in performance than the diagonal smoothing matrix because it produced a smaller value of the asymptotic mean integrated squared error (AMISE).

Table 2: Variance, bias^2 and *AMISE* for Old Faithful data set

Bandwidths	*Variance*	*Bias^2*	*AMISE*
$H_{D\text{-}AMISE}$	0.0016045116	0.0008023932	0.0024069048
$H_{F\text{-}AMISE}$	0.0012735059	0.0000960389	0.0013695448

As generally known, one method is better than the other one when it gives a smaller value of the asymptotic mean integrated squared error (Jarnicka 2009, Siloko et al. 2019). However; from the kernel estimates of both forms of parameterizations, the bimodality property of the distribution is retained and both parameterizations also exemplify the usefulness of the bivariate kernel density estimates for highlighting structures in a data set.

Discussion

The paper investigated the performance of smoothing parameterization of kernel density estimation with emphasis on bivariate kernel density estimator. The performance of the bivariate kernel density estimator is primarily determined by the smoothing parameter and the form of parameterization employed unlike its univariate counterpart whose performance determinant is the smoothing parameter.

The performance of kernel density estimator in relation with smoothing parameter and forms of parametrization employed is best determined by the production of minimum error value using an error criterion function. The error criterion function used in this paper is the asymptotic mean integration squared error. The full parameterization produced the minimum asymptotic mean integration squared error value when compared with the diagonal parameterization, although the two approaches produced kernel estimates whose statistical properties of the data like bimodality were retained. The smaller AMISE value of the full parameterization indicates a better choice of smoothing parameter than the diagonal parameterization.

Conclusions

The full smoothing parameterization of the bivariate kernel estimator outperformed the diagonal parameterization with the AMISE as error criterion function. However, the kernel estimates of both forms of parameterization retained the inherent feature of bimodality of the bivariate data examined for exploratory and visualization purposes. The full smoothing parameterization is therefore recommended for higher dimensions although with difficulty as the dimensions of the kernel function increases. The complexity associated with higher dimensions known as curse of dimensionality is mainly a problem with nonparametric statistics.

Acknowledgement

The authors appreciate the anonymous reviewers, Chief Editor and Technical Editor for painstakingly going through the manuscript and for their valuable comments.

References

Azzalini A and Bowman AW 1990 A look at some data on the Old Faithful geyser. *J. Royal Statist. Soc.: Series C (Appl. Statist.)* 39(3): 357-365.

Duong T and Hazelton ML 2003 Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparamet. Statist.* 15(1): 17-30.

Guidoum AC 2015 Kernel estimator and bandwidth selection for density and its derivatives. The kedd package, version 1.0.3. *Department of Probabilities and Statistics, University of Science and Technology, Houari Boumediene, Algeria.*

Jarnicka J 2009 Multivariate kernel density estimation with a parametric support. *Opuscula Mathematica* 29(1): 41-45.

Liu Q, Pitt D, Zhang X and Wu X 2011 A Bayesian approach to parameter estimation for kernel density estimation via transformations. *Annals Actuar. Sci.* 5(2): 181-193.

Sain RS 2002 Multivariate locally adaptive density estimation. *Comput. Statist. Data Anal.* 39: 165-186.

Scott DW 1992 Density estimation: Theory, practice and visualization. *The Curse of Dimensionality and Dimension Reduction*, pp.195-217, Wiley, New York.

Siloko IU, Ikpotokin O, Oyegue FO, Ishiekwene CC. and Afere BAE 2019 A note on application of kernel derivatives in density estimation with the univariate case. *J. Statist. Manage. Syst.* 22(3): 415-423.

Siloko IU, Ishiekwene CC and Oyegue FO 2018 New gradient methods for bandwidth selection in bivariate kernel density estimation. *Math. Statist.* 6(1): 1–8.

Silverman BW 2018 *Density estimation for statistics and data analysis*. Routledge, New York, 176 pages.

Simonoff JS 2012 Smoothing Methods in Statistics. Springer Science & Business Media, New York.

Wand MP and Jones MC 1995 Kernel Smoothing. Chapman and Hall/ CRC, London.