Mathematics and Statistics 6(1): 1-8, 2018
DOI: 10.13189/ms.2018.060101 http://www.hrpub.org

New Gradient Methods for Bandwidth Selection in Bivariate Kernel Density Estimation

Siloko, I. U.[1,*], Ishiekwene, C. C.[2], Oyegue, F. O.[2]

[1]Department of Mathematical Sciences, Edwin Clark University, Nigeria
[2]Department of Mathematics, University of Benin, Nigeria

Abstract The bivariate kernel density estimator is fundamental in data smoothing methods especially for data exploration and visualization purposes due to its ease of graphical interpretation of results. The crucial factor which determines its performance is the bandwidth. We present new methods for bandwidth selection in bivariate kernel density estimation based on the principle of gradient method and compare the result with the biased cross-validation method. The results show that the new methods are reliable and they provide improved methods for a choice of smoothing parameter. The asymptotic mean integrated squared error is used as the measure of performance of the new methods.

Keywords Bandwidth, Bivariate Kernel Density Estimator, Biased Cross-validation, Gradient Method, Asymptotic Mean Integration Squared Error

1. Introduction

Kernel density estimators are widely used nonparametric estimation techniques due to their simple forms and smoothness. Kernel density estimation is the construction of a probability density estimates from a given sample with few assumptions about the underlying probability density function and the kernel function. Kernel density estimation is a popular tool for visualising the distribution of data [1]. These estimates depend on a bandwidth also known as the smoothing parameter which controls the smoothness and a kernel function which plays the role of a weighting function [2]. Bandwidth selection is a key issue in kernel methods and has attracted the attention of researchers over the years. It is still an active research area in kernel density estimation. Progress has been made recently on data based smoothing parameter selectors by some researchers [3, 4].

The importance of the bivariate kernel density estimator cannot be overemphasized because it occupies a unique position of bridging the univariate kernel density estimator and other higher dimensional kernel estimators [5]. The usefulness of the bivariate kernel density estimator is mainly in its simplicity of presentation of probability density estimates, either as surface plots or contour plots. It also helps in understanding other higher dimensional kernel estimators [2]. In bivariate kernel density estimation, $\mathbf{x}, \mathbf{y}$ is taken to be the two random variables with a joint probability density function $\boldsymbol{f}(\mathbf{x}, \mathbf{y})$. The random variables $\mathbf{X_i}, \mathbf{Y_i}, \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n}$ are the set of observations and $\boldsymbol{n}$ is the sample size. The bivariate kernel density estimate of $\boldsymbol{f}(\mathbf{x}, \mathbf{y})$ is of the form

$$\hat{\boldsymbol{f}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n h_x h_y} \sum_{i=1}^{n} \boldsymbol{K}\left(\frac{\mathbf{x}-\mathbf{X_i}}{h_x}, \frac{\mathbf{y}-\mathbf{Y_i}}{h_y}\right) \qquad (1.1)$$

where $\boldsymbol{h_x} > 0$ and $\boldsymbol{h_y} > 0$ are the smoothing parameters in the $\mathbf{X}$ and $\mathbf{Y}$ axes and $\boldsymbol{K}(\mathbf{x}, \mathbf{y})$ is a bivariate kernel function [5, 6]. The bivariate kernel density estimator in (1.1) can be written as [7]

$$\hat{\boldsymbol{f}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_x} \boldsymbol{K}\left(\frac{\mathbf{x}-\mathbf{X_i}}{h_x}\right) \frac{1}{h_y} \boldsymbol{K}\left(\frac{\mathbf{y}-\mathbf{Y_i}}{h_y}\right) \qquad (1.2)$$

Bivariate bandwidth selection is a difficult problem which may be simplified by imposing constraints on $\boldsymbol{h_x}$ and $\boldsymbol{h_y}$. For example, $\boldsymbol{h_x}$ and $\boldsymbol{h_y}$ may be restricted to be the diagonal elements of the bandwidth matrix and the advantages of imposing restrictions on $\boldsymbol{h_x}$ and $\boldsymbol{h_y}$ has been investigated [8]. One of the popular methods of bandwidth selection that is data based is the biased cross-validation method that considers the asymptotic mean integrated squared error [9]. The bivariate biased cross-validation method is based on minimizing an estimate of the asymptotic mean integrated squared error and is of the form [10]

$$\boldsymbol{BCV}(\boldsymbol{h_x}, \boldsymbol{h_y}) = \frac{1}{4\pi n h_x h_y} + \frac{1}{4n(n-1)h_x h_y} \times \sum_{i=1}^{n} \sum_{j \neq i} \left[\left(\Delta_1^2 + \Delta_2^2 \right)^2 - 8\left(\Delta_1^2 + \Delta_2^2 \right) + 8 \right] \boldsymbol{\phi}(\Delta_1)\, \boldsymbol{\phi}(\Delta_2) \tag{1.3}$$

where $\Delta_1 = \left(\frac{x - X_i}{h_x}\right)$, $\Delta_2 = \left(\frac{y - Y_i}{h_y}\right)$ and $\boldsymbol{\phi}$ is the standard normal density function.

This paper focuses on the methods of selecting bandwidth for bivariate kernel density estimator using the gradient methods. The rest of the paper is organized as follows: in section 2, we present the asymptotic mean integrated square error of the bivariate kernel density estimator; in section 3, we present the gradient methods while section 4 talks about numerical illustrations of results. Section 5 concludes the paper.

2. Asymptotic MISE Approximations

The estimate $\hat{f}(\mathbf{x}, \mathbf{y})$ in (1.2) is measured by the asymptotic mean integrated squared error (AMISE). A straightforward asymptotic approximation of (1.2) using the multivariate version of Taylor's series expansion yields the integrated variance (IV) and the integrated squared bias (ISB) as

$$\begin{cases} \boldsymbol{IV} = \frac{R(K)^d}{n h_x h_y} & \text{and} \\ \boldsymbol{ISB} = \frac{1}{4}\boldsymbol{\mu_2}(\boldsymbol{K})^2 \int \boldsymbol{tr}^2 \left(\sum_{j=1}^{2} \frac{\partial^2 f(x,y)}{\partial x^2 \partial y^2} h_x^2 h_y^2 \right) \boldsymbol{dxdy} \end{cases} \tag{2.1}$$

where $\boldsymbol{R}(\boldsymbol{K})$ is the roughness of the kernel, $\boldsymbol{tr}$ is the trace of a matrix (matrix of second partial derivatives) of $\boldsymbol{f}$, $\boldsymbol{d}$ is the dimension of the kernel and $\boldsymbol{\mu_2}(\boldsymbol{K})^2$ is the second moment of the kernel. Combining the terms in (2.1) yields an estimate of the asymptotic mean integrated squared error and is of the form

$$\boldsymbol{AMISE} = \frac{R(K)^d}{n h_x h_y} + \frac{1}{4}\boldsymbol{\mu_2}(\mathrm{K})^2 \int \boldsymbol{tr}^2 \left(\sum_{j=1}^{2} \frac{\partial^2 f(x,y)}{\partial x^2 \partial y^2} h_x^2 h_y^2 \right) \boldsymbol{dxdy} = \frac{R(K)^d}{n h_x h_y} + \frac{1}{4}\boldsymbol{\mu_2}(\mathrm{K})^2\, \boldsymbol{R}(\boldsymbol{f''}) \tag{2.2}$$

where $\boldsymbol{R}(\boldsymbol{f''}) = \int \boldsymbol{tr}^2 \left(\sum_{j=1}^{2} \frac{\partial^2 \boldsymbol{f}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^2 \partial \mathbf{y}^2} \boldsymbol{h_x^2 h_y^2} \right) \boldsymbol{dxdy}$ is the roughness of $\boldsymbol{f''}(\mathbf{x}, \mathbf{y})$.

The smoothing parameter that minimizes the AMISE of (2.2) is given by

$$\boldsymbol{H_{AMISE}} \approx \left[\frac{d R(K)^d}{\mu_2(\mathrm{K})^2\, R(f'')} \right]^{\left(\frac{1}{d+4}\right)} \times \boldsymbol{n}^{-\left(\frac{1}{d+4}\right)} \tag{2.3}$$

The $\boldsymbol{H_{AMISE}}$ yields an $\boldsymbol{AMISE} = \boldsymbol{O}\left(\boldsymbol{n}^{-\left(\frac{4}{d+4}\right)}\right)$ and the bandwidths are of order $\boldsymbol{n}^{-1/(d+4)}$. The problem with (2.3) is that it has a bias term which depends on $\boldsymbol{R}(\boldsymbol{f''})$ and cannot be evaluated if the true density function $\boldsymbol{f}$ is not known. Several suggestions have been made and one of the simplest solutions to this problem is to obtain the value of $\boldsymbol{R}(\boldsymbol{f''})$ from the normal distribution for unimodal data but for data that is multimodal (which is kernel density estimations' expectation), it behaves poorly in performance and in that situation, it serves as an initial point for other iterative methods [11, 12].

3. The Gradient Method

Gradient methods are iterative methods of optimizing functions that are at least twice continuously differentiable [13]. Gradient methods involve generating successive points in the direction of the gradient function with the desire to obtain the stationary points [14]. The gradient of a function $\boldsymbol{f}(\mathbf{x})$ denoted by $\boldsymbol{\nabla f}(\mathbf{x})$ is given by

$$\boldsymbol{\nabla f}(\mathbf{x}) = \left(\frac{\partial \boldsymbol{f}}{\partial \mathbf{x}_1}, \frac{\partial \boldsymbol{f}}{\partial \mathbf{x}_2}, \ldots, \frac{\partial \boldsymbol{f}}{\partial \mathbf{x}_n} \right)^T \tag{3.1}$$

Gradient methods are iterative techniques with the elements of the gradient being nonlinear and a positive scalar (stepsize) that determines the distance between the current point and the previous point [13]. Generally, the gradient method requires higher derivatives of the function to be approximated. In this work, we modify the gradient methods of Barzilai and Borwein and the relaxed steepest descent method of Raydan and Svaiter [15, 16].

In the application of the gradient methods, we replace the unknown quantity $\boldsymbol{R}(\boldsymbol{f''})$ in (2.3) by a suitable kernel based estimate denoted by $\widehat{\boldsymbol{R}}(\boldsymbol{f}(\mathbf{x}))$, i.e. $\boldsymbol{R}(\boldsymbol{f''}) = \widehat{\boldsymbol{R}}(\boldsymbol{f}(\mathbf{x}))$. The kernel function used is the standard normal kernel that produces smooth density estimates and simplifies the mathematical computations needed. In the case of the standard normal kernel, successive smoothing parameters will be obtained from the approximation given by

$$\boldsymbol{H_{AMISE}} \approx \left\{ \left(\frac{d R(K)^d}{\mu_2(\mathrm{K})^2 \times \hat{R}(f(\mathrm{x}))} \right)^{1/(d+4)} \right\} \times \boldsymbol{n}^{-\{1/(d+4)\}}. \tag{3.2}$$

Since all gradient methods are iterative methods and they require the choice of a starting value say $\mathbf{X_o}$ we apply a kernel based estimate as an initial point for the iteration process and is of the form

$$\mathbf{X_o} = \left[\frac{R(K)^d}{\mu_2(\mathrm{K})^2\,\sigma_j}\right]^{\left(\frac{1}{d+4}\right)} \times \boldsymbol{n}^{-\left(\frac{1}{d+4}\right)}, \qquad (3.3)$$

where $\boldsymbol{\sigma_j}$ is the standard deviation of the $\boldsymbol{jth}$ variate.

3.1. The Barzilai and Borwein Gradient Method

This method solves the problems associated with the classical gradient method with their novel approach of the stepsize selection. It requires lesser computations and the rate of convergence is also accelerated, though with the same search direction as the classical gradient method but with better performance [17]. The method is used for obtaining large sparse linear system of equations from the solution of partial differential equations and also applies in optimization theory due to its ease of computation and implementation [18, 19]. The Barzilai and Borwein gradient method is of the form

$$\mathbf{X_{t+1}} = \mathbf{X_t} + \mathbf{S_t}\,,\ \mathbf{t} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \qquad (3.4)$$

where $\mathbf{S_t} = -\frac{\mathbf{1}}{\boldsymbol{\lambda_t}}$ and $\boldsymbol{\lambda_t}$ is the stepsize. The value of the stepsize can be obtained from (3.5) as

$$\boldsymbol{\lambda_t} = \frac{g_t^{\mathrm{T}} g_t}{g_t^{\mathrm{T}} \mathrm{A}\, g_t},\ \mathbf{t} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \qquad (3.5)$$

where $\boldsymbol{g_t} = \boldsymbol{\nabla f}(\mathbf{X_t})$, $\mathbf{A}$ is the Hessian matrix of $\boldsymbol{f}$ evaluated at $\mathbf{X_t}$ and $\mathbf{A}$ must be symmetric. The value of the kernel based estimate $\widehat{\boldsymbol{R}}\big(\boldsymbol{f}(\mathbf{x})\big)$ must be positive, i.e. $\widehat{\boldsymbol{R}}\big(\boldsymbol{f}(\mathbf{x})\big) > 0$ because the smoothing parameter must be positive. The estimate $\widehat{\boldsymbol{R}}\big(\boldsymbol{f}(\mathbf{x})\big) = \mathbf{X_{t+1}}$ when substituted into (3.2) will give the smoothing parameter that minimizes the AMISE. In all the gradient methods considered, the modifications were in the method of obtaining the initial point of the iteration which is kernel based and the introduction of the third step that will be used to obtain the required solution.

ALGORITHM 1 (The Modify Barzilai and Borwein Method).

STEP1. Compute $\mathbf{X_o} = \left(\left[\frac{R(K)^d}{\mu_2(\mathrm{K})^2\,\sigma_j}\right]^{\left(\frac{1}{d+4}\right)} \times \boldsymbol{n}^{-\left(\frac{1}{d+4}\right)}\right)$, where $\boldsymbol{\sigma_j}$ is the standard deviation of the $\boldsymbol{j}\mathbf{th}$ variate for $\boldsymbol{j} = \mathbf{1}, \mathbf{2}$, $\boldsymbol{n}$ is the sample size and $\boldsymbol{d}$ is the dimension of the kernel.

STEP2. For $\boldsymbol{t} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots$,

(a) Compute the vector $\boldsymbol{g_t} = \boldsymbol{\nabla f}(\mathbf{X_t})$.

(b) Compute the step size $\boldsymbol{\lambda_t} = \frac{g_t^{\mathrm{T}} g_t}{g_t^{\mathrm{T}} \mathrm{A}\, g_t}$.

(c) Set $\mathbf{S_t} = -\frac{\mathbf{1}}{\boldsymbol{\lambda_t}}$.

(d) Update $\mathbf{X_{t+1}} = \mathbf{X_t} + \mathbf{S_t}$.

STEP3. Employ $\widehat{\boldsymbol{R}}\big(\boldsymbol{f}(\mathbf{x})\big) = \mathbf{X_{t+1}}$ to compute the smoothing parameter using (3.2) above.

STEP4. Test a criterion for stopping the iterations. If the test is satisfied, then stop. Otherwise, consider $\mathbf{t} = \mathbf{t} + \mathbf{1}$ and continue with step 2 by updating the step size,

$$\boldsymbol{\lambda_{t+1}} = \frac{g_{t+1}^{\mathrm{T}} g_{t+1}}{g_{t+1}^{\mathrm{T}} \mathrm{A}\, g_{t+1}},\ \text{where } \boldsymbol{g_{t+1}} = \boldsymbol{\nabla f}(\mathbf{X_{t+1}}).$$

3.2. The Relaxed Steepest Descent Method

Another modification of the steepest descent method was made by Raydan and Svaiter [16] and they concluded that the poor performance of the method is a function of the stepsize and not the search direction. In solving the problem of the stepsize's selection, they introduced a scalar called the relaxation parameter $\boldsymbol{\theta}$ which lies between 0 and 2 into the classical steepest method. The classical steepest descent method is of the form

$$\mathbf{X_{t+1}} = \mathbf{X_t} - \boldsymbol{\lambda_t} \boldsymbol{g_t}\,,\ \mathbf{t} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \qquad (3.6)$$

where $\boldsymbol{g_t} = \boldsymbol{\nabla f}(\mathbf{X_t})$. The introduction of the relaxation parameter $\boldsymbol{\theta}$ which is a random scalar on the interval $[\mathbf{0}, \mathbf{2}]$, resulted in the modified steepest descent method which is of the form

$$\mathbf{X_{t+1}} = \mathbf{X_t} - \boldsymbol{\theta}\boldsymbol{\lambda_t} \boldsymbol{g_t}\,,\ \mathbf{t} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \qquad (3.7)$$

where $\boldsymbol{\lambda_t} = \dfrac{\boldsymbol{g_t^{\mathrm{T}} g_t}}{\boldsymbol{g_t^{\mathrm{T}} \mathbf{A}\, g_t}}$.

Multiplying the stepsize by the relaxation parameter resulted in the improvement of the method by accelerating its rate of convergence when applied to numerical problems [20]. It should be noted that when $\boldsymbol{\theta} = \mathbf{1}$, the classical steepest descent method will be obtained and so $\boldsymbol{\theta} \neq \mathbf{1}$.

ALGORITHM 2 (The Modify Relaxation Method of Raydan and Svaiter).

STEP1. Compute $\mathbf{X_o} = \left(\left[\frac{R(K)^d}{\mu_2(\mathrm{K})^2\,\sigma_j}\right]^{\left(\frac{1}{d+4}\right)} \times \boldsymbol{n}^{-\left(\frac{1}{d+4}\right)}\right)$, where $\boldsymbol{\sigma_j}$ is the standard deviation of the $\boldsymbol{j}\mathbf{th}$ variate for $\boldsymbol{j} = \mathbf{1}, \mathbf{2}$, $\boldsymbol{n}$ is the sample size and $\boldsymbol{d}$ is the dimension of the kernel.

STEP2. For $\boldsymbol{t} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots$,

(a) Compute the gradient vector $\boldsymbol{g_t} = \boldsymbol{\nabla f}(\mathbf{X_t})$.

(b) Compute the step size $\boldsymbol{\lambda_t} = \frac{g_t^{\mathrm{T}} g_t}{g_t^{\mathrm{T}} \mathrm{A}\, g_t}$.

(c) Update $\mathbf{X_{t+1}} = \mathbf{X_t} - \boldsymbol{\theta}\boldsymbol{\lambda_t} \boldsymbol{g_t}$.

STEP3. Employ $\widehat{\boldsymbol{R}}\big(\boldsymbol{f}(\mathbf{x})\big) = \mathbf{X_{t+1}}$ to compute the smoothing parameter using (3.2) above.

STEP4. Test a criterion for stopping the iterations. If the test is satisfied, then stop. Otherwise, consider $\mathbf{t} = \mathbf{t} + \mathbf{1}$ and continue with step 2 by updating the step size,

$$\boldsymbol{\lambda_{t+1}} = \frac{g_{t+1}^{\mathrm{T}} g_{t+1}}{g_{t+1}^{\mathrm{T}} \mathrm{A}\, g_{t+1}},\ \text{where } \boldsymbol{g_{t+1}} = \boldsymbol{\nabla f}(\mathbf{X_{t+1}}).$$

In the application of this algorithm, our relaxation parameter $\boldsymbol{\theta} = \frac{\mathbf{1}}{\boldsymbol{n}}$ where $\boldsymbol{n}$ is the sample size. The

relaxation parameter is chosen as $\boldsymbol{\theta} = \frac{\mathbf{1}}{\boldsymbol{n}}$ because of the role of the sample size in the choice of smoothing parameter. Generally, it is known that the smoothing parameter depends very strongly on the sample size such that as the sample size increases, the smoothing parameter tends to be reduced [21]. The stopping criterion is $\|\mathbf{X_{t+1}} - \mathbf{X_t}\| < \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} = 10^{-5}$ where $\boldsymbol{\varepsilon}$ is the tolerance level. The results of the modify methods shall be compared with the biased cross-validation method.

4. Results and Discussion

In order to illustrate the efficiency of these methods, we compare their performances with the biased cross-validation method using the asymptotic mean integrated squared error (AMISE) as the error criterion function. The results are presented by comparing the gradient methods (Modify Barzilai and Borwein (MBB) and the Modify Relaxed Steepest Descent(MRSD)) with the Biased Cross-Validation (BCV)) for the bivariate kernel density estimator and using the asymptotic mean integrated squared error (AMISE) as the error criterion function for measuring their performance. As generally known, one method is better than the other when it gives a smaller value of the AMISE [12]. The comparison also involves the kernel estimates (graphs) of the methods considered since kernel density estimation has direct applications on data analysis such as exploratory data analysis and data visualisation [1, 2, 6].

Two sets of data were used to illustrate the results of the methods, showing in a tabular form the smoothing parameter, the Asymptotic Integrated Variance (AIV), the Asymptotic Integrated Squared Bias (AISB) and the Asymptotic Mean Integrated Squared Error (AMISE) using the bivariate standard normal kernel. An important point to note from the Tables below is that in terms of performance, the gradient methods resulted in a smaller value of the AMISE.

One very important and notable step to be taken when examining bivariate data set is to consider the Scatterplot of the data. But as it has been in most cases, while kernel density estimate will reveal or highlight important features, Scatterplot cannot play this vital role [2]. Scatterplots have been regarded as the most frequently used tools for graphically displaying bivariate data sets but with the serious disadvantage that the eye is only drawn to the peripheries of the data cloud, while structures in the main body of the data will be hidden by the high density of the points [22]. In kernel density estimates, these disadvantages of the Scatterplots are removed because they have an advantage in the presentation of information regarding the distribution of the data set. As noted from the Scatterplots of the data sets considered, the modes are not apparent from the Scatterplot vas in the kernel density estimates and this exemplify the usefulness of the bivariate kernel density estimates for highlighting structure.

The first data set examined is the Volcanic Crater data of Bunyaruguru Volcanic Field in Western Uganda [23]. It involves the Locations of Centers of Craters of 120 volcanoes in two variables in which variable X represents the first center while variable Y represents the second center. Figure 1 shows the Scatterplot of the Crater data and the Scatterplot clearly show a strong relationship between the variables with correlation coefficient $\boldsymbol{\rho} = \mathbf{0.814398}$. It is evident that the two Locations of the Centers of Craters are highly positively correlated. A significant feature of this data set that is very noticeable from the kernel density estimates (graphs) is the bimodality of the data but this is hidden as presented by the Scatterplot. We standardized the data 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 [2, 10, 24, 25]. Figures 2, 3, and 4 below show the kernel estimates of the Crater data.

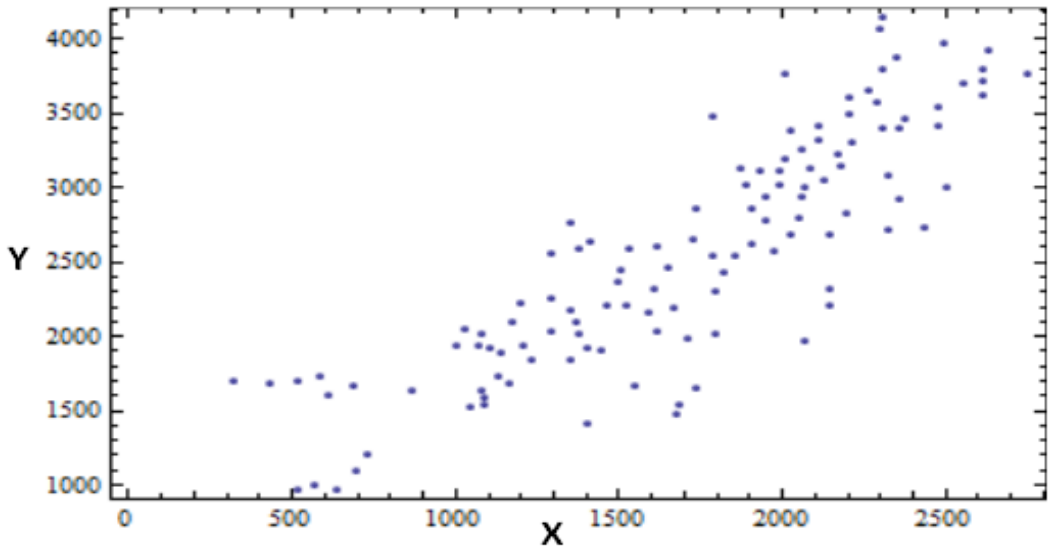

Figure 1. Scatterplot of the Volcanic Crater Data

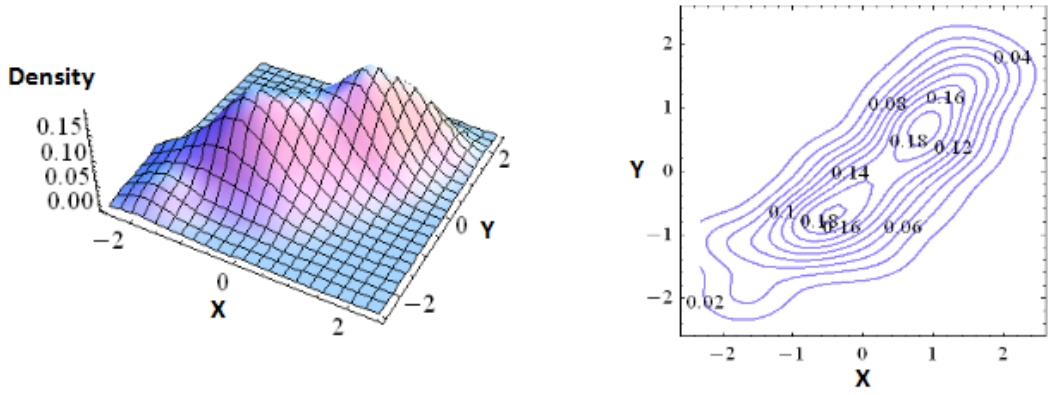

Figure 2. Kernel Estimate of the Crater Data using BCV Smoothing Parameter

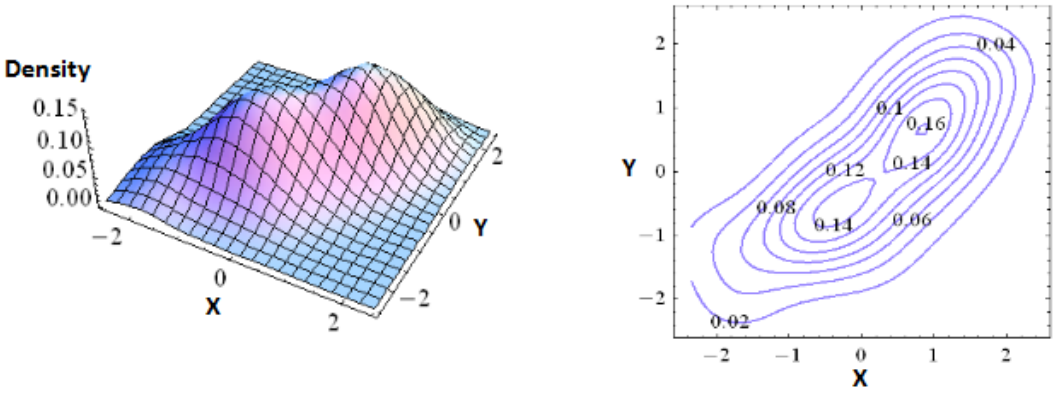

Figure 3. Kernel Estimate of the Crater Data using MBB Smoothing Parameter.

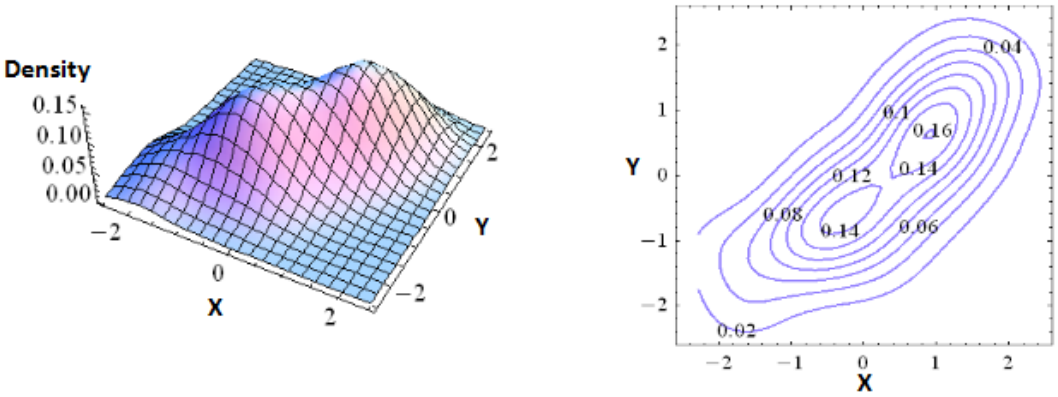

Figure 4. Kernel Estimate of the Crater Data using MRSD Smoothing Parameter

The biased cross-validation method yields smoothing parameter that produces an estimate with the bimodality being clearly present as shown in Figure 2. The gradient methods also yield smoothing parameter values that retain the bimodality of the data as shown in Figure 3 and Figure 4 respectively. The table below shows the bandwidths, the asymptotic integrated variance (AIV), the asymptotic integrated squared biased (AISB) and the asymptotic mean integrated squared error (AMISE) of the methods considered.

Table 1. Bandwidths, AIV, AISB and AMISE for the Crater Data

Methods.	h_x	h_y	*AIV*	*AISB*	*AMISE*
BCV	0.45042	0.30224	0.00487124	0.00066901	0.00554025
MBB	0.48675	0.48268	0.00282256	0.00148993	0.00431249
MRSD	0.48802	0.48393	0.00280795	0.00150548	0.00431343

From Table 1 above, it is obvious that in terms of performance, the biased cross-validation method produced the largest AMISE value. The gradient methods yield a smoothing parameter with smaller AMISE value as shown in Table 1.

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 [26]. 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. One very important point to note from the bivariate kernel estimates of this data is that the data set is bimodal and this provides very strong evidence in favour of eruption times and the time interval until the next eruption exhibiting a bimodal distribution [2]. The bivariate kernel density estimates are bimodal and it is evident that the time interval until the next eruption is highly positively correlated with the duration of the eruption. The Scatterplot of the Old Faithful data is shown in Figure 5 while Figures 6, 7 and 8 are the bivariate kernel estimates of the data for the methods considered. The Scatterplot show a strong relationship between the variables with correlation coefficient $\boldsymbol{\rho = 0.90087}$. The data is also standardized to obtain equal variances in each dimension [24, 25].

Table 2 below shows the smoothing parameters, the asymptotic integrated variance (AIV), the asymptotic integrated squared bias (AISB) and the asymptotic mean integrated squared error (AMISE) of the methods considered.

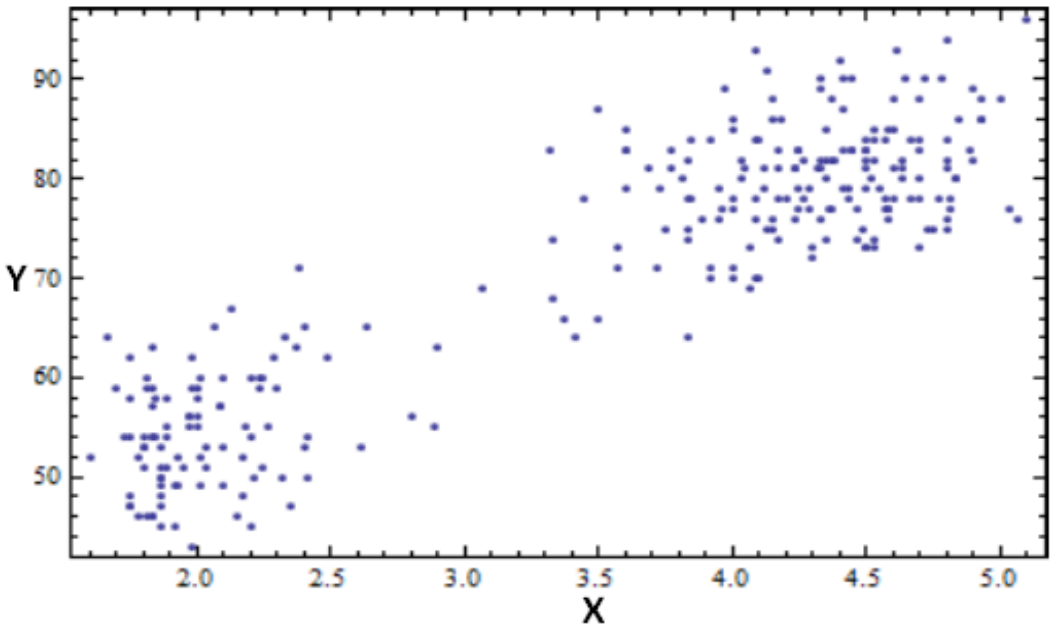

Figure 5. Scatterplot of the Old Faithful Data

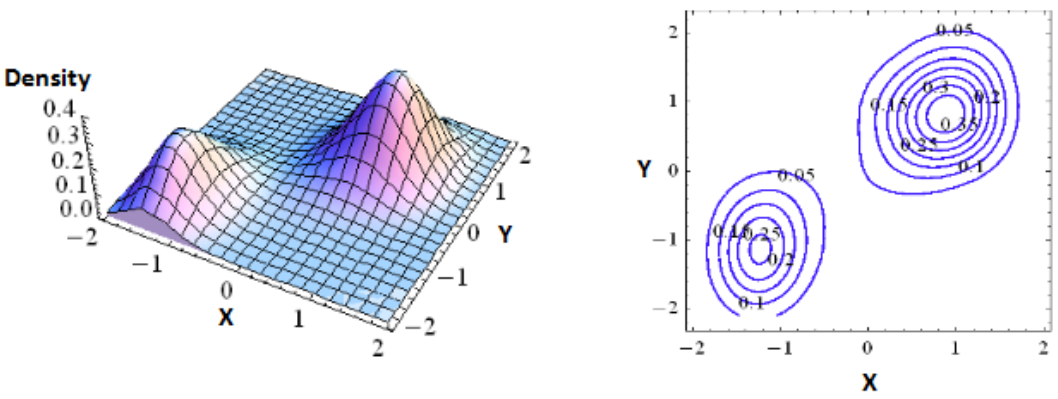

Figure 6. Kernel Estimate of the Old Faithful Data using BCV Smoothing Parameter

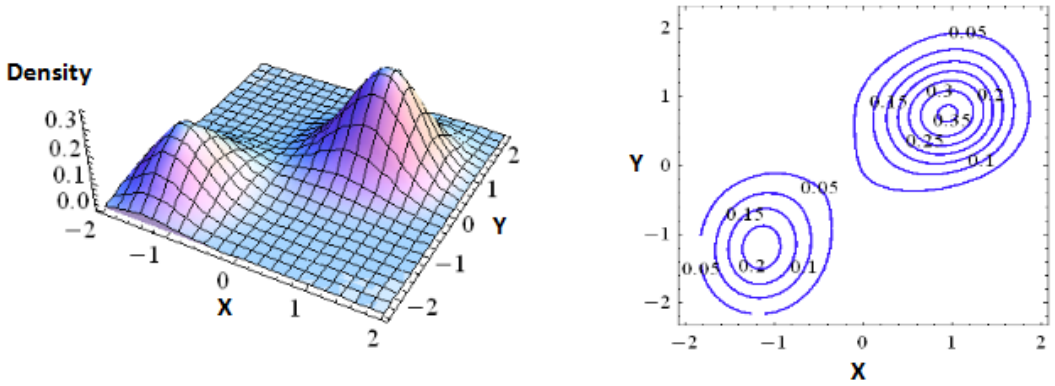

Figure 7. Kernel Estimate of the Old Faithful Data using MBB Smoothing Parameter

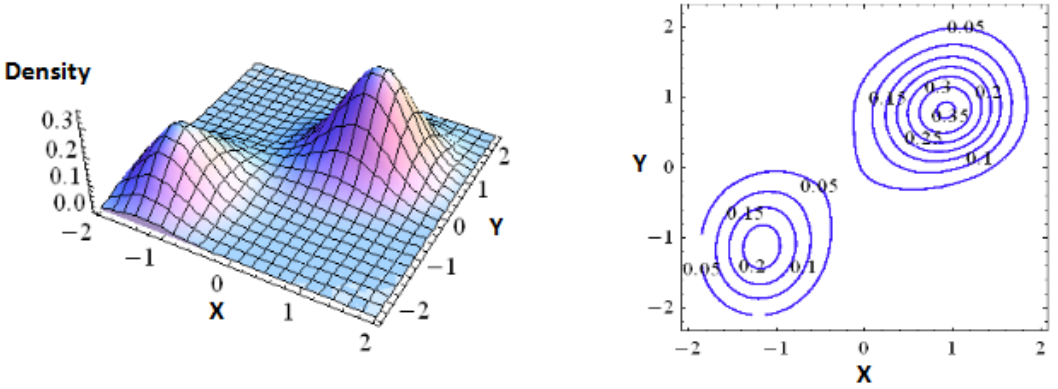

Figure 8. Kernel Estimate of the Old Faithful Data using MRSD Smoothing Parameter

Table 2. Bandwidths, AIV, AISB and AMISE for the Old Faithful Data

Methods.	h_x	h_y	*AIV*	*AISB*	*AMISE*
BCV	0.29382	0.39688	0.00250897	0.00040169	0.00291066
MBB	0.36263	0.38846	0.00207688	0.00048786	0.00256474
MRSD	0.36364	0.38955	0.00206532	0.00049335	0.00255867

Table 2 shows the performances of the biased cross-validation method and the gradient methods and from the results, the biased cross-validation method yielded larger AMISE value. Again the gradient methods yield a smoothing parameter with the smaller AMISE values as presented in Table 2.

5. Conclusions

The methods presented are compared with the biased cross-validation method because they are based on a suitable estimate of the asymptotic mean integrated squared error (AMISE). The results presented show that the new methods are reliable and they provide improved methods for a choice of smoothing parameter. An advantage of the gradient methods is that they can be easily computed provided the function $\boldsymbol{f}$ is at least twice differentiable.

As for the bivariate case that sits between the univariate and higher dimensional kernel, and that $\boldsymbol{K}$ is a standard normal product kernel, the gradient methods based on their performance is at least as competitive as the existing biased cross-validation.

REFERENCES

[1] Simonoff, J. S. Smoothing Methods in Statistics. Springer-Verlag, New York, 1996.

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

[3] Chacón, J. E. and Duong, T. Data-Driven Density Derivative Estimation, with Applications to Nonparametric Clustering and Bump Hunting. Electronic Journal Statistics, 7, 499–532 2013.

[4] Jiang, M. and Provost, S. B. A Hybrid Bandwidth Selection Methodology for Kernel Density Estimation. Journal of

Statistical Computation and Simulation, 84(3), 614–627, 2014.

[5] Duong, T. and Hazelton, M. L. Plug-In Bandwidth Matrices for Bivariate Kernel Density Estimation. Nonparametric Statistics, 15(1), 17–30, 2003.

[6] Scott, D.W. Multivariate Density Estimation. Theory, Practice and Visualisation. Wiley, New York, 1992.

[7] Zhange, X., Wu, X., Pitt, D. And Liu, Q. A Bayesian Approach to Parameter Estimation for Kernel Density Estimation via Transformation. Annals of Actuarial Science, 5(2), 181–193, 2011.

[8] Wand, M. P. and Jones, M. C. Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation. Journal of the American Statistical Association, 88, 520–528, 1993.

[9] Scott, D. W. and Terrell, G. R. Biased and Unbiased Cross-Validation in Density Estimation. Journal of the American Statistical Association, 82, 1131–1146, 1987.

[10] Sain, R. S., Baggerly, A. K. and Scott, D. W. Cross-Validation of Multivariate Densities. Journal of American Statistical Association, 89, 807–817, 1994.

[11] Zhange, X., King, M. L. and Hyndman, R. J. A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation. Computational Statistics and Data Analysis, 50, 3009–3031, 2006.

[12] Jarnicka, J. Multivariate Kernel Density Estimation with a Parametric Support. Opuscula Mathematica, 29(1), 41–45, 2009.

[13] Hansen, B. E. Econometrics. University of Wisconsin, Spring, 2013.

[14] Cameron. A. C. and Trivedi, P. K. Micro-econometrics Methods and Applications. Cambridge University Press, New York, USA, 2005.

[15] Barzilai, J. and Borwein, J. M. Two Point Step size Gradient Method. IMA Journal of Numerical Analysis, 8(1), 141–148, 1988.

[16] Raydan, M and Svaiter, B. F. Relaxed Steepest Descent and Cauchy-Barzilai-Borwein Method. International Journal of Computational Optimization and Applications, 21(2), 155–167, 2002.

[17] Raydan, M. Convergence Properties of the Barzilai and Borwein Gradient Method. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree, Doctor of Philosophy, Rice University, Houston, Texas, 1991.

[18] Farid, M., Leong, W. J. and Hassan M. A. A New Two-Step Gradient-Type Method for Large-Scale Unconstrained Optimization. Computers and Mathematics with Applications, 59, 3301–3307, 2012.

[19] Dai, Y. H. A New Analysis on the Barzilai-Borwein Gradient Method. JORC, 1, 187–198, 2013.

[20] Battaglia, J. P. The Eigenstep Method (An Iterative Method for Unconstrained Quadratic Optimization). American Journal of Operational Research, 2013, 3(2), 57–64

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

[22] Wand, M. P. and Jones, M. C. Kernel Smoothing. Chapman and Hall, London, 1995.

[23] Bailey, T. C. and Gatrell, A. C. Interactive spatial data analysis. Longman, Harlow, 1995.

[24] Cula, S. G and Toktamis, O. Estimation of Multivariate Probability Density Function with Kernel Functions. Journal of the Turkish Statistical Association, 3(2), 29–39, 2000.

[25] Sain, R. S. Multivariate Locally Adaptive Density Estimation. Computational Statistics and Data Analysis, 39, 165–186, 2002.

[26] Azzalini, A. and Bowman, A. W. A Look at Some Data on the Old Faithful Geyser. Applied Statistics, 39, 357–365, 1990.