Canadian Journal of Pure and Applied Sciences
Vol. 11, No. 2, pp. 4255-4262, June 2017
Online ISSN: 1920-3853; Print ISSN: 1715-9997
Available online at www.cjpas.net

THE EFFECT OF SMOOTHING PARAMETER IN KERNELS AGGREGATION

*Siloko, I U[1], Ishiekwene, C C[2] and Oyegue, F O[2]
[1]Department of Mathematical Sciences, Edwin Clark University, Kiagbodo, Nigeria
[2]Department of Mathematics, University of Benin, Benin City, Nigeria

ABSTRACT

Kernel density estimation depends on appropriate smoothing parameter selection in its implementation since the method is mainly for data exploration and visualization purposes. While considering the effect of the smoothing parameter, the form of aggregation employed will determine the size of the smoothing parameter required for better performance. This paper considered two aggregating methods with respect to the asymptotic mean integrated squared error (AMISE) as the criterion function by introducing the multiplier factor that regulates the selection of smoothing parameter in the multiplicative aggregation. The results of the forms of aggregation considered were compared using real life data.

Keywords: Kernel density aggregation, kernel density estimator, smoothing parameter, multiplier factor, asymptotic mean integrated squared error (AMISE).

INTRODUCTION

The Gaussian Mixture Model (GMM) and the Kernel Density Estimators (KDE) are the most popular nonparametric density estimation techniques with practical applications (Kobos and Mandziuk, 2009). In kernel density estimation, there is the kernel function that averages the observations to produce a smooth approximation (Silverman, 1986). The kernel function is a probability density function and also a standardized weighting function. The smoothness of any kernel density estimate is majorly determined by the smoothing parameter. The smoothing parameter is regarded and interpreted as a resolution of observations viewed such that the technique of viewing the set of observations with different resolutions will give better interpretation of the structures in the observations under consideration (Kobos and Mandziuk, 2009). Researchers are providing solution to the problem of smoothing parameter selection with some recent works which includes Chacón and Duong, 2010; Zhang *et al*., 2011; Chacón and Duong, 2013, Chacon and Duong, 2015; Jiang and Provost, 2014).

In statistical learning environment, different aggregation methods were considered especially in classifications and regression problems such as Bagging (1996a), Stacking (Breiman, 1996b), Boosting (Schapire, 1977) and Random forests (Breiman, 2001). Stacking was further extended to density estimation (Smyth and Wolpert, 1999). Originally, the boosting idea was designed for classification problems but it was extended to kernel density estimation (Marzio and Taylor, 2005) and was further revisited (Ishiekwene, 2008), where it was regarded as a bias reduction strategy.

This paper presents a solution to the problem of smoothing parameter selection associated with the kernel density estimation particularly in aggregating. We have shown by numerical examples that aggregations could produce better results in terms of performance provided the smoothing parameter is chosen rightly while aggregating. We shall discuss briefly the existing aggregation algorithms and describe the role of the smoothing parameter in the aggregations that will be considered. Finally, we shall introduce the multiplier factor for the multiplicative aggregation and compare its results with the additive aggregation method.

Aggregation Algorithms

Aggregation algorithms in density estimation can be grouped into two major categories depending on the aggregation form employed (Bourel and Ghattas, 2013). The first group known as the additive model with the form of linear combination is given by

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m g_m(\mathbf{x}) \qquad (1)$$

Where β_m are the coefficients of the model and g_m is a parametric or nonparametric density family. The values of m could be different parameters that can be evaluated in the case of parametric density estimation, different kernels or different smoothing parameters for a kernel

*Corresponding author e-mail: suzuazor@yahoo.com

function in the case of kernel density estimation (Bourel and Ghattas, 2013). The model in Equation (1) has been applied in classification and regression problems like boosting, bagging and other topics in parametric regularization (Ridgeway, 2002; Rosset, 2003).

If the additive model in Equation (1) is taken to be a nonparametric density family, particularly the kernel density estimators with smoothing parameter, then we have

$$\hat{f}(\mathbf{x}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - X_i}{h}\right) \tag{2}$$

Where $K(.)$ is a kernel function and h is the bandwidth (Silverman, 1986). The major problem that is confronting the implementation of the kernel density estimator since its introduction is the choice of the smoothing parameter also known as the bandwidth. The choice of smoothing parameter is often critical and crucial during implementation but not the choice of the kernel function since most kernel functions are probability density function (Silverman, 1986). The kernel function in Equation (2) is a non-negative function that must satisfy the conditions

$$\begin{cases} \int K(\mathbf{x})d\mathbf{x} = 1, \\ \int \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0 \quad \text{and} \\ \int \mathbf{x}^2 K(\mathbf{x})d\mathbf{x} = k_2(K) \neq 0 \end{cases} \tag{3}$$

The multidimensional kernel density estimator of Equation (2) using the product approach is

$$\hat{f}(\mathbf{x}) = \left(n\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_1 - \mathbf{X}_{i1}}{h_1}, \ldots, \frac{\mathbf{x}_d - \mathbf{X}_{id}}{h_d}\right) \tag{4}$$

Where K is the kernel function with variance $\mu_2(\mathbf{K}) = \int \mathbf{x}^2 K(\mathbf{x})\, d\mathbf{x}$ and h_j are the smoothing parameters for each dimension (Epanechnikov, 1969; Sain *et al*., 1994).

The second form of aggregation is the multiplicative model that was introduced (Marzio and Taylor, 2004) and was also extended to the multidimensional case in kernel density estimation (Marzio and Taylor, 2005). The multiplicative aggregation is a bias reduction technique and it uses the kernel density estimator. This aggregation is of the form

$$f(\mathbf{x}) = \prod_{m=1}^{M} \beta_m g_m(\mathbf{x}) \tag{5}$$

The multivariate form of the multiplicative aggregation also known as boosting algorithm is a systematized algorithm where each step m is computed using Equation (6) given as (Bourel and Ghattas, 2013)

$$\hat{f}_m(\mathbf{x}) = \left(n\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_m(i) K\left(\frac{\mathbf{x}_1 - \mathbf{X}_{i1}}{h_1}, \ldots, \frac{\mathbf{x}_d - \mathbf{X}_{id}}{h_d}\right) \tag{6}$$

Where K is a fixed kernel, h_j are the smoothing parameter(s) and $\mathbf{W}_m(i)$ is the weight of observation i at step m (Bourel and Ghattas, 2013). The weight of each observation is then updated as(Marzio and Taylor, 2005)

$$\mathbf{W}_{m+1}(i) = \mathbf{W}_m(i) + \log\left(\frac{\hat{f}_m(\mathbf{x}_i)}{\hat{f}_m^{(-i)}(\mathbf{x}_i)}\right) \tag{7}$$

Where $\hat{f}_m^{(-i)}(\mathbf{x}_i)$ is the leave-one-out estimator of the multivariate product kernel that uses different smoothing values for its axes and is given by

$$\hat{f}_m^{(-i)}(\mathbf{x}_i) = (n-1)^{-1}\left(\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_m(i) K\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{h_j}\right) \tag{8}$$

Also $\hat{f}_m(\mathbf{x}_i)$ is of the form given by

$$\hat{f}_m(\mathbf{x}_i) = n^{-1}\left(\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_m(i) K\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{h_j}\right) \tag{9}$$

Boosting in kernel density estimation involves the weights being updated at each step and with the final estimator being a product of all the density estimates that integrates to unity (Marzio and Taylor, 2005; Marzio and Taylor, 2004). The algorithm given below is for the multidimensional case in which the product kernel that uses different smoothing parameter was employed.

STEP1. Given that $i = 1, 2, \ldots, n$, initialise the weights of the observations

$$W_1(i) = 1/n$$

STEP2. Select $H = h_1, h_2, \ldots, h_d$ the smoothing parameters.

STEP3. For $m = 1, \ldots, M$.

(i) Obtain a weighted kernel estimate

$$\hat{f}_m(\mathbf{x}) = n^{-1}\left(\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} W_m(i) K\left(\frac{\mathbf{x}_j - X_{ij}}{h_j}\right)$$

(ii) Update the weights according to

$$\mathbf{W}_{m+1}(i) = \mathbf{W}_m(i) + \log\left(\frac{\hat{f}_m(\mathbf{x}_i)}{\hat{f}_m^{(-i)}(\mathbf{x}_i)}\right)$$

STEP4. Provide as output

$$C\prod_{m=1}^{M}\hat{f}_m(\mathbf{x}),$$

Where C is the normalization constant such that $\hat{f}_m(\mathbf{x})$ integrates to unity.

Smoothing Parameter Selection in Kernel Aggregation

Appropriate selection of the smoothing parameter is often critical to the process of kernel aggregation in kernel density estimation because its performance is based on its right selection. The quality of the estimates $\hat{f}(\mathbf{x})$ in Equation (4) and Equation (6) is measured by the asymptotic mean integrated squared error defined as

$$\begin{aligned}\text{AMISE}\left(\hat{f}(\mathbf{x};h)\right) &\approx \frac{R(K)^d}{nh_1h_2,\dots,h_d} + \frac{1}{4}h_j^4\mu_2(K)^2\int tr^2\left(\nabla^2 f(\mathbf{x})\right)d\mathbf{x} \\ &= \frac{R(K)^d}{nh_1h_2,\dots,h_d} + \frac{1}{4}h_j^4\mu_2(K)^2R\left(\nabla^2 f(\mathbf{x})\right) \qquad (10)\end{aligned}$$

Where $j = 1, 2, \dots, d$, $R(K) = \int K^2(\mathbf{x})d\mathbf{x}$ represents the roughness of the kernel, $\mu_2(K)^2$ is the variance, $R\left(\nabla^2 f(\mathbf{x})\right) = \int tr^2\left(\nabla^2 f(\mathbf{x})\right)$ is the roughness of the function, tr is the trace of a matrix, n is the sample size, $h_1, h_2, \dots, h_d$ are the smoothing parameters to be determined, d is the dimension of the kernel and $\nabla^2 f(\mathbf{x})$ is the Hessian array of f (Sain, 2002). The smoothing parameter that minimizes Equation (10) above is given by

$$H_{AMISE} \approx \left[\frac{dR(K)^d}{\mu_2(\mathbf{K})^2 R\left(\nabla^2 f(\mathbf{x})\right)}\right]^{\left(\frac{1}{d+4}\right)} \times n^{-\left(\frac{1}{d+4}\right)} \qquad (11)$$

This smoothing parameter is of order $n^{-1/(d+4)}$ with $AMISE = O\left(n^{-\left(\frac{4}{d+4}\right)}\right)$. There is no generally acceptable rule for selecting smoothing parameter in the additive case but different rules are available. The multiplicative aggregation that is known to be bias reducing approach demands larger smoothing parameter in its implementation (Marzio and Taylor, 2005; Ridgeway, 2002; Siloko and Ishiekwene, 2016).

The multiplicative aggregation known as boosting in kernel density estimation leads to a reduction in the bias component of the AMISE but with its major problem being the smoothing parameter required for its iterations (Marzio and Taylor, 2005). The reduction in the bias component resulted in a reduction in the AMISE. In solving the problem of smoothing parameter selection in multiplicative aggregation, we introduced a multiplier known as the bandwidth multiplier which is denoted by β. Therefore the smoothing parameter required for the multiplicative aggregation in kernel estimation is

$$H_m = \beta^{1/2} \times (h_1^*, h_2^*, \dots, h_d^*) \qquad (12)$$

Where

$$\begin{cases} m = 2, 3, \dots, M \\ \\ \beta = 2m \end{cases} \qquad (13)$$

In Equation (12) above, $h_1^*, h_2^*, \dots, h_d^*$ are the smoothing parameters obtained from Equation (11), m represents the number of iterations in the aggregation and $2m$ denotes the order of the kernel. The bandwidth multiplier factor helps in the selection of smoothing parameters for the multiplicative aggregation and has solved curse of dimensionality problem that is associated with the multivariate kernel density estimation (Marzio and Taylor, 2004; Sain, 2002). In the application of Equation (12), we excluded the case when $m = 1$ in Equation (13) because it will result in second order kernel which is the same as the additive aggregation form.

RESULTS AND DISCUSSION

In this section, we will compare the estimates of the additive and the multiplicative aggregations. The results of the additive aggregation in terms of performance were compared with the multiplicative aggregation in a tabular form and the later displaying better results than the former in terms of performance using two data sets. The performances of these two forms of aggregation will be measured using the AMISE as the error criterion. While the additive aggregation tend to retain some desirable features of the data set such as multimodality, the multiplicative aggregation at times may smooth out the multimodality feature as a result of using larger smoothing parameters but with a reduction in the AMISE. In all the cases considered, 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 (Sain *et al*., 1994; Sain, 2002; Simonoff, 1996; Cula and Toktamis, 2000).

The first data set examined is the Volcanic Crater data of Bunyaruguru Volcanic Field in Western Uganda (Bailey and Gatrell, 1995). 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.

A significant feature of this data set that is very noticeable from the kernel estimates of the additive aggregation is the bimodality of the data but this is hidden as presented by the multiplicative aggregation due to the usage of larger smoothing parameter and the multiplication

A significant feature of this data set that is very noticeable from the kernel estimates of the additive aggregation is the bimodality of the data but this is hidden as presented by the multiplicative aggregation due to the usage of larger smoothing parameter and the multiplication involve. Figure 1 is the kernel estimate of the additive aggregation while Figure 2 and Figure 3 are the kernel estimates of the multiplicative aggregation.

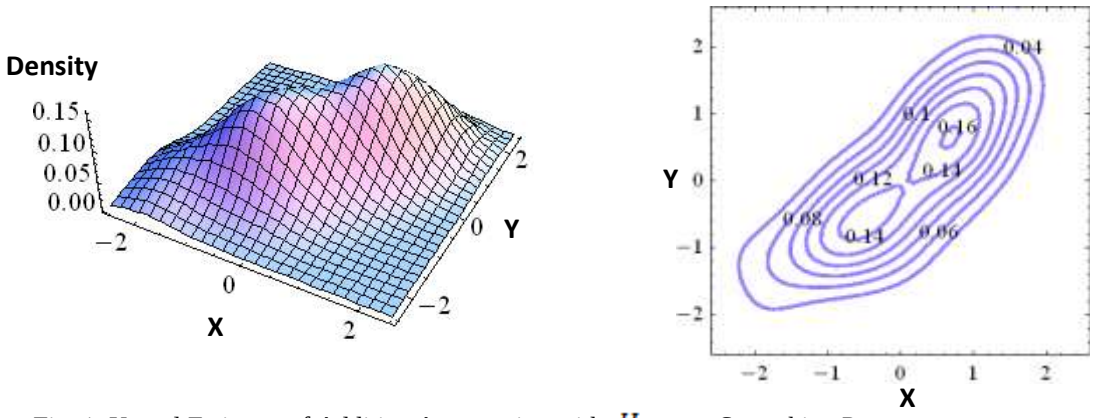

Fig. 1. Kernel Estimate of Additive Aggregation with H_{AMISE} Smoothing Parameter.

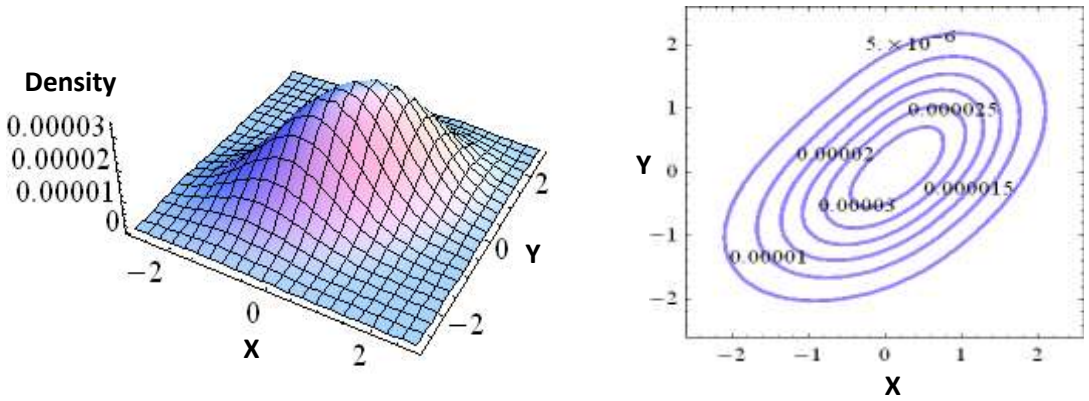

Fig. 2. Kernel Estimates of the First Step of the Multiplicative Aggregation.

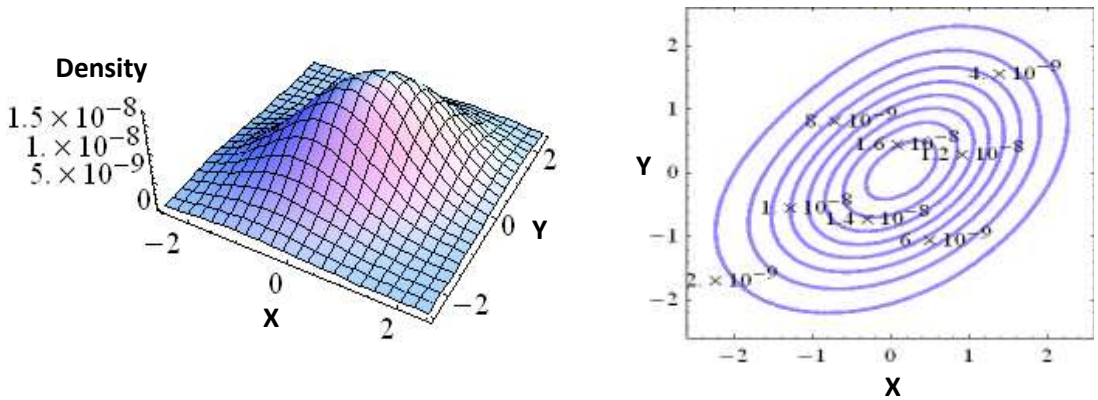

Fig. 3. Kernel Estimates of the Second Step of the Multiplicative Aggregation.

As generally known, one method is better than the other one when it gives a smaller value of the AMISE (Jarnicka, 2009). The multiplicative aggregation (higher order bias reduction techniques) yielded AMISE with smaller values when compared with the additive method (second order kernels). This shows that the multiplicative aggregation performed better than the additive aggregation in terms of reduction in the AMISE but is capable of smoothening out the inherent features of the set of observations that could be of great statistical significance.

Table 1 and 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 forms of aggregation considered with the AMISE as the error criterion function for measurement of performance.

Table 1. Analysis of Additive Aggregation for the Crater Data.

Method	h_x	h_y	***AIV***	***AISB***	***AMISE***
H_{AMISE}	0.48018	0.48053	0.00287398	0.00143694	0.00431092

Table 2. Analysis of Multiplicative Aggregation for the Crater Data.

Steps.	h_x	h_y	***AIV***	***AISB***	***AMISE***
1	0.96036	0.96106	0.000718496	0.000059200461	0.000777696461
2	1.17620	1.17705	0.000478997	0.000000418095	0.000479415095

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.

One very important point to note from the 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 (Silverman, 1986). Figure 4 is the kernel estimate of the additive aggregation while Figure 5 and Figure 6 are the kernel estimates of the multiplicative aggregation.

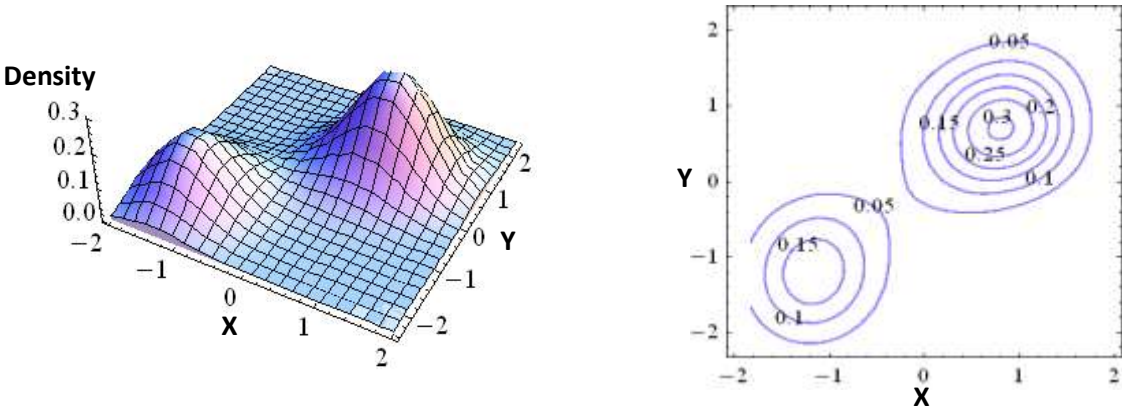

Fig. 4. Kernel Estimate of Additive Aggregation with H_{AMISE} Smoothing Parameter.

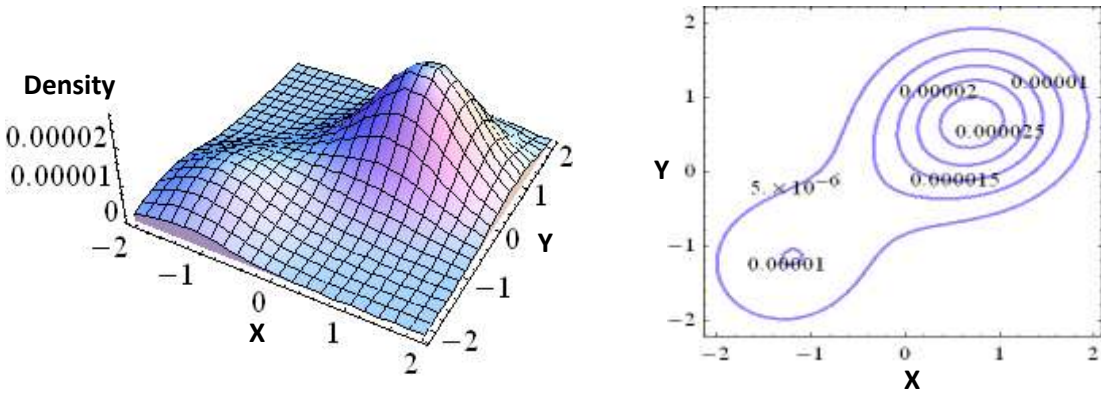

Fig. 5. Kernel Estimates of the First Step of the Multiplicative Aggregation.

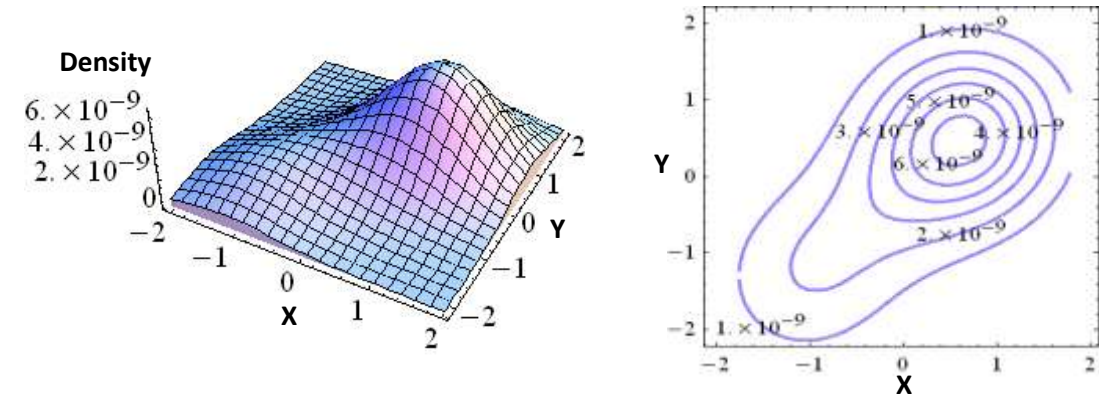

Fig. 6. Kernel Estimates of the Second Step of the Multiplicative Aggregation.

Table 3 and Table 4 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 forms of aggregation considered with the AMISE as the error criterion for performance evaluation. The analysis of the performances of the additive and multiplicative aggregations is presented in Table 3 and Table 4.

The results of the analysis show that the multiplication aggregation also known as boosting in kernel estimation resulted in smaller values of the AMISE when compared with the additive aggregation. The additive aggregation retained the bimodality of the data but this tends to disappear at the second step of the multiplicative aggregation.

Table 3. Analysis of Additive Aggregation for the Faithful Data.

Method	h_x	h_y	***AIV***	***AISB***	***AMISE***
H_{AMISE}	0.43105	0.42301	0.00160451	0.00080239	0.00240690

Table 4. Analysis of Multiplicative Aggregation for the Faithful Data.

Steps.	h_x	h_y	***AIV***	***AISB***	***AMISE***
1	0.86210	0.84602	0.000401128	0.000013524623	0.000414652623
2	1.05585	1.03616	0.000267419	0.000000178189	0.000267597189

CONCLUSION

The additive and multiplicative aggregations were considered with the later displaying better results than the former in terms of performance using the AMISE as the error criterion function. The multiplicative aggregation technique targets reduction in the AMISE without considering features such as multi-modality that might be present in a given data set due to the principle of over smoothing which means using larger smoothing parameters. As observed from the data sets considered, the additive aggregation retained the features of the data set such as bimodality while the multiplicative aggregation at times may smooth out this feature as a result of using larger smoothing parameters but with a reduction in the AMISE. The smoothing parameter used for the multiplicative aggregation was obtained using the bandwidth multiplier.

REFERENCES

Azzalini, A. and Bowman, AW. 1990. A Look at Some Data on the Old Faithful Geyser. Applied Statistics. 39:357-365.

Bailey, TC. and Gatrell, AC. 1995. Interactive spatial data analysis. Longman, Harlow.

Bourel, M. and Ghattas, B. 2013. Aggregating Density Estimators: An Empirical Study. Open Journal of Statistics. 3(5):344-355.

Breiman, L. 1996[a]. Stacked Regression. Machine Learning. 24(1):49-64.

Breiman, L. 1996[b]. Bagging Predictors. Machine Learning. 24(2):123-140.

Breiman, L. 2001. Using Iterated Bagging to Debias Regressions. Machine Learning. 45(3):261-277.

Chacón, JE. and Duong, T. 2010. Multivariate Plug-In Bandwidth Selection with Unconstrained Pilot Bandwidth Matrices, Test. 19:375-398.

Chacón, JE. and Duong, T. 2013. Data-Driven Density Derivative Estimation, with Applications to Nonparametric Clustering and Bump Hunting. Electronic Journal of Statistics. 7:499-532.

Chacón, JE. and Duong, T. 2015. Efficient Recursive Algorithms for Functionals Based on Higher Order Derivatives of the Multivariate Gaussian Density. Statistical Computing. 25:959-974.

Cula, SG. and Toktamis, O. 2000. Estimation of Multivariate Probability Density Function with Kernel Functions. Journal of the Turkish Statistical Association. 3(2): 29-39.

Epanechnikov, VA. 1969. Nonparametric Estimation of a Multivariate Probability Density. Theory Probab. Appl. 14:153-158.

Freund, Y. and Schapire, R. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 55(1):119-139.

Ishiekwene, CC. 2008. Bias Reduction Techniques in KDE. An Unpublished Ph.D. Thesis. School of Postgraduate Studies, University of Benin, Benin City, Nigeria.

Jarnicka, J. 2009. Multivariate Kernel Density Estimation with a Parametric Support. Opuscula Mathematica. 29(1):41-45.

Jiang, M. and Provost, SB. 2014. A Hybrid Bandwidth Selection Methodology for Kernel Density Estimation. Journal of Statistical Computation and Simulation. 84(3): 614-627.

Kobos, M. and Mańdziuk, J. 2009. Classification Based on Combination of Kernel Density Estimators. Lecture Notes in Computer Science. 5769. Springer. 125-134.

Marzio, DM. and Taylor, CC. 2004. Boosting Kernel Density Estimates: A Bias Reduction Technique? Biometrika. 91:226-233.

Marzio, DM. and Taylor, CC. 2005. On Boosting Kernel Density Methods for Multivariate Data: Density Estimation and Classification. Statistical Methods and Applications. 14:163-178.

Ridgeway, G. 2002. Looking for Lumps: Boosting and Bagging for Density Estimation. Computational. Statistics and Data Analysis. 38(4):379-392.

Rosset, S. 2003. Topics in Regularization and Boosting. A Dissertation Submitted to the Department of Statistics and the Committee on Graduate Studies of Stanford University.

Sain, RS. 2002. Multivariate Locally Adaptive Density Estimation. Computational Statistics and Data Analysis. 39:165-186.

Sain, RS., Baggerly, AK. and Scott, DW. 1994. Cross-Validation of Multivariate Densities. Journal of American Statistical Association. 89:807-817.

Siloko, IU. and Ishiekwene, CC. 2016. Boosting and Bagging in Kernel Density Estimation. Nigerian Journal of Science and Environment. 14(1):32-37.

Silverman, BW. 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Simonoff, JS. 1996. Smoothing Methods in Statistics. Springer, New York, USA.

Smyth, P. and Wolpert, D. 1999. Linearly Combining Density Estimators via Stacking. Machine Learning. 36:59-83.

Zhang, X., Wu, X., Pitt, D. and Liu, Q. 2011. A Bayesian Approach to Parameter Estimation for Kernel Density Estimation via Transformation. Annals of Actuarial Science. 5(2):181-193.

Received: Feb 23, 2017; Revised: April 10, 2017;
Accepted: April 15, 2017