Athens Journal of Sciences- Volume 6, Issue 4 – Pages 287-308

Some Improvements in Nonparametric Multivariate Kernel Density Estimation

By Michael Ogbeide[*] & Joseph Osemwenkhae[‡]

A popular technique of density estimation is the kernel density estimation (KDE). It is a nonparametric estimation approach which requires a kernel function and a bandwidth (smoothing parameter H). It aid density estimation and pattern recognition. This paper presents new approaches in nonparametric density construction problem, particularly at the boundary points using the dataset and a pilot plot. However, since the main way to improve density estimation is to obtain a reduced mean squared error (MSE). When the MSE for these approaches were evaluated and compared. Some improvements were seen in two proposed approaches. These were achieved under a sufficiently smoothing technique in the existing approaches. These approaches are adaptive and they reduce under fitting and over fitting as the case may be of the data set and aid statistical inference.

Keywords: *Adaptive, bandwidths, error, kernel density estimation.*

Introduction

Data density estimation provides a nonparametric estimate of the probability function from which a set of data is drawn. Often, optimal pattern recognition algorithms require the knowledge of the underlying dataset to construct densities. Primarily, it is better to estimate the density from the data. This is not always the case, hence nonparametric approach, which has the flexibility of the model specification. One of the common nonparametric approaches is the Kernel density estimation (KDE). It has been widely regarded that the performance of the kernel methods depends largely on the smoothing parameter (window width) but depends very little on the form of the kernel. According to Scott (1992) and Osemwenkhae (2003) most times, analyses of multivariate data are more prevalent in practice than the univariate cases. The crucial problem in the multivariate kernel density estimation (MKDE) is to select the window widths (bandwidth parameters) *H*. The window widths control the smoothness of the fitted density curve. The multivariate kernel density estimator that we are going to study is a direct extension of the univarite estimator. Let $X_1,...,X_n$ denote a d-variate random sample having a density f. We shall use the notation $X_i = (X_{i1},...,X_{in})^T$ to denote the X_i and a generic vector $x \in \Re^d$ has the representation $x = (x_1,...,x_d)^T$. The d-variate

[*]Department of Mathematics and Statistics, Ambrose Alli University, Nigeria.
[‡] Department of Mathematics, University of Benin, Nigeria.

https://doi.org/10.30958/ajs.6-4-4 doi=10.30958/ajs.6-4-4

random sample $X_1,...,X_n$ drawn from f the kernel estimator evaluated at x. This according to Wand and Jones (1995) is given by;

$$\hat{f}_{h_j}(X,H) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i) \qquad (1.1)$$

where *n* is the sample size, and H is a symmetric positive definite *d* x *d* matrix called the window widths, the smoothing parameters or the bandwidth matrix. $K_H(x) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}}x)$, $|\ .\ |$ stands for the determinant of H and K is d-variate kernel satisfying $\int k(x)dx = 1$, where it is understood that the integral is over $\Re^d$ unless stated otherwise. The kernel function is often taken to be a *d*-variate probability density function.

The two common techniques for obtaining multivariate kernel from the univariate kernel k:

$K^P(x) = \prod_{i=1}^{d} k(x)$ - a product kernel.

$K^S(x) = c_{K,d}.k(x^T x),\ \ c_{K,d} = (\int k(x^T x)^{\frac{1}{2}} dx$ - a spherically symmetric kernel.

Remark: The spherically symmetric kernel corresponding to the Epanechnikov kernel when *d* = 2 is given as:

$K^S(x) = \frac{2}{\Pi}(1 - x_1^2 - x_2^2)$, $\quad x_1^2 + x_2^2 \le 1$

And the product kernel $K^P(x) = \frac{9}{16}(1 - x_1^2)(1 - x_2^2)$, $\quad x_1 \cdot x_2 \in [-1,1]$.

However, the gaussian kernel *k(u)* = $\frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$ is a popular choice among many kernels (Bowman and Azzalini 1997, Kathovnik and Shmulevich 2002).

The matrix *H* is a smoothing parameter and specifies the 'width' of the kernel around each sample point X_i. A well behaved *K* (that is a kernel bounded compactly) must satisfy the following regularity conditions:

1. $\int_{R^d} K(w)dw = 1$
2. $\int_{R^d} wK(w)dw = 0$
3. $\int_{R^d} ww^T K(w)dw = I_d$

Where I_d is a *d* dimensional identity matrix.

The first condition accounts for the fact that the sum of the kernel function over the whole region is unity. The second condition imposes the equation constraint that the means of the marginal kernel $K_i(w_i), i = 1,...,d$ are all zero. The third condition term states that the marginal kernels are all pairwise uncorrelated and each has unit variance.

However, the most important part of the estimator in (1.1) is the bandwidth matrix which contained the window sizes used for smoothing density. The fixed window size method are not sensitive to local peculiarities in the data, such as clustering/sparseness of sample value, though oversmoothing/ undersmoothing (as the case may be) tends to reduce or overcome this.

The adaptive (smoothing) methods are nonparametric density estimators that are sensitive to clustering/sparseness of sample values and other peculiarities, particularly at the tails or other peculiarities of the data set. Here the smoothing parameter *H* varies, hence the "adaptive" techniques.

This work on density estimation, is based on KDE with the analysis of the existing intersection of confidence intervals (ICI), the kernel cluster sampling approach in estimating density approaches. We modified two approaches to the multivariate kernel density estimation, and therefore, proposed the modified intersection of confidence interval ($MICI_H$) and the modified multivariate cluster sampling kernel density estimate (MMCKDE).

Literature Review

This work focuses on the kernel density estimation. There are other several density estimation approaches like the Histogram, the Scatter plots, the Orthogonal series density, the nearest neighbour method, and the Projection pursuit density estimation (Fukunaga 1990, Ogbeide et al. 2016). Clearly, in practice, one does not have access to the true density function *f(x)* which is to be estimated (Wand and Jones 1995, Wu et al. 2006). Thus, a number of approaches can be taken for finding the bandwidth that will lead to better density estimation via varying the bandwidths (Wand and Jones 1995, Katkovnik 1999, Jarnicka 2009, Ogbeide et al. 2016). Details studies on the Histograms which are the oldest density methods, the Scatter plots, the Orthogonal series density, the nearest neighbour method and the Projection pursuit density estimation approaches can be seen in (Tukey 1947, Cencov 1962, Friedman and Stuetzle 1982, Rudemo 1982, Scott and Thompson 1983, Silverman 1986, Wand and Jones 1995, Isenman 1991, Bowman and Azzalini 1997). These methods are not bonafide probability density estimates (Silverman 1986, Wand and Jones 1995). These approaches take time and they have poor visual display with poor meaning from inference. It can be difficulties to make complete inference from them. However, Scott and Thompson (1983) and Wand and Jones (1995) noted that these approaches are merely convenient presentational device in an attempt to discern features for the distribution or model underling the data.

According to Wand and Jones (1995) and Bowman and Azzalini (1997), the discontinuity in the histogram method density estimates makes it less attractive for proper adaptive uses This discontinuity leads to poor visual display which affects inference from it (Silverman 1986, Isenman 1991). We observed in Wand and Jones (1995) that Scatter plots approach is not good for most bivariate data set and beyond. Estimates from the Scatter plot obviously do not detect or highlight certain features of the data set. The Orthogonal series

density method requires a lot of evaluation before its smoothing parameters could be achieved, its application is mainly asymptotic, so there is need for better straight forward computation of the smoothing parameter and faster rate of convergence (Osemwenkhae 2003). Silverman (1986) and Katkovnik and Shmulevich (2002) noted that the nearest neighbour methods suffer under-fittings or over fittings during estimation. This makes it less attractive. The convergence rate of this method is poor when compared to the kernel density estimation (Osemwenkhae 2003). The projection pursuit density estimation (PPDE) is multivariate density estimation technique that attempt to reduce the curse of dimensionality while estimating density. It does not spring from the univariate density estimation generalization. It was developed by Friedman and Stuetzle (1982). It transforms the data set around the origin zero with a covariance matrix as the identity and determines the estimates of $\hat{f}_H$ in iterative manner. Though this approach takes time, the number of iterations determine the number of different smoothing parameters and the stopping rule is determine by balancing the bias against the variance estimate (Friedman and Stuetzle 1982, Friedman et al. 1984, Alan 1991).

Multivariate kernel density estimation (MKDE) approach is a bonafide probability density estimate. It requires a kernel function and a smoothing parameter *H* which is the window size. Wand and Jones (1995) noted that the MKDE is a well behaved density method with a kernel which satisfies the following regularity conditions. This method gives a good representation of the real data. We observed from the Scott (1992) and Bowman and Azzalini (1997) that density estimation curves either underfits or overfits as the case may be. However, this method uses a fixed value from the bandwidth matrix to estimate and smooth densities. This usually leads to over-fitting or under-fitting as the case may be (Scott 1992, Duong and Hazelton 2003). The Adaptive Multivariate kernel density estimation (AMKDE) approach requires a kernel function and smoothing parameters which are the window sizes corresponding to the data set. This varying choice of the smoothing parameter makes it adaptive. This method gives a good representation of the real data. Its density estimation curves either try to correct under fits or over fits as the case may be. This estimation method according to Scott (1992) and Wand and Jones (1995), under the regularity conditions noted that the AMKDE is a well behaved density method with varying optimal choice of $H_i \in H$. The *K* (that is a kernel bounded compactly) which must satisfy the regularity conditions. Our approach in this work is based on this technique. This approach is also called the dynamic smoothing technique or the variable kernel method. This method is called dynamic smoothing technique or the variable kernel method due to the adjustable nature of the smoothing parameter *H* (Silverman 1986, Sain 2002, Tower (2002). The variable kernel method estimates integrate to unity (Abramson 1982, Silverman 1986, Duong and Hazelton 2003). This method is a bonafide probability density estimates. It has smooth curves corresponding to the dataset adequately. According to Wand and Jones (1995) as $n \to \infty$, the mean integrated squared error $MISE \to 0$. That is the kernel density estimate converges to the mean

square error and also in probability to the true density *f*. The convergence of the method is a confirmation of the statement in the motivation of this study that the kernel methods lead to reasonably density estimation. An ideal optimal bandwidths selectors according to Bowman and Azzalini (1997) and Duong and Hazelton (2005a) should be based on every data set elements. The ideal selector which contains the unknown density function *f*, that cannot be used directly, hence the approach of varying $H_i \in H$ optimally.

The variable kernel density estimator was first proposed by Victor (1976) and Breimen et al. (1977). Victor (1976) suggested varying window sizes to enhance data density evaluation in medical decision making. This according to him is due to the fact that fitting medical data obtained at different times and rates with a single window size may not best represent the true density. Breimen et al. (1977) obtained the probability of the variable kernel density estimation. Abramson (1982) derived the square-root law for achieving higher-order bias for a kernel function. Jones (1990) clarified the main differences between the local and the variable kernel approaches. Some other works on variable kernel methodology include (Muller 1985, Hall and Park 1987, Hall 1990, Fukunaga 1990, Jones 1990, Hall 1992, Handle and Scott 1992, Scott 1992, Jones et al. 1994, Elio and Edgar 2003, Wu et al. 2008). They worked on the properties and efficiency of variable kernel method. Katkovnik and Shmulevich (2002) developed an adaptive method based on intersection of confidence intervals. The Cluster sampling approach to MKDE by Wu and Tsai (2004) and Wu et al. (2008) is an approach which utilized the idea of Breiman et al. (1977) and Abramson (1982) in determination of the density estimates. This approach uses the information matrix rows and columns to form clusters for sampling, where the cluster sizes and bandwidths factors are used to achieve the smoothing parameters. The method for multivariate density is adaptive, but Wu et al. (2008) proposed average Cluster sampling to correct deficiency with some points of discontinuities in Wu and Tsai (2004).

When we consider the studies on variable window sizes works on the multivariate cluster sampling kernel density estimate (MCKDE) and the intersection of confidence interval (ICI) methods applied to MKDE, though the methods are adaptive, one is tasked with how sensitive these methods are, and the errors committed using these methods? What are the effects when we extend them to multivariate kernel density? These questions led to the reasons for their modifications in this research. We identified points for improvements, so that the methods could be more adaptive. Recently, a wide variety of sophistication of the basic kernel estimator has been proposed, all pointing to the importance of adaptive kernel estimator (Kathovnik and Shmulevich 2002, Salgado-Ugarte and Perez-Hernandez 2003, Wu and Tsai 2004, Wu et al. 2008). The "adaptive" nature of the density estimate arises from the varying bandwidth used in the data estimation process. If *h*, the bandwidth in (1.1) above, is "fixed" during data estimation, we have the fixed kernel density estimation approach, but when it is allowed to vary all though the process of the estimation, we have the adaptive kernel method. A number of work considering the problem of kernel size selection exist (Abramson 1982, Silverman 1986, Breiman et al. 1977, Hall 1990, Jones

1990, Cao et al. 1994, Wand and Jones 1995, Simonoff 1996, Wu et al. 2007, Ogbeide et al. 2016). The main intentions are that they all aimed at improving kernel density estimation.

The most commonly used optimality criterion for selecting a bandwidth matrix is the mean integrated squared error (MISE) expressed according to Wu et al. (2006) as;

$$MISE(H) = E\{\int [\hat{f}_h(X) - f(X)]^2 dX \qquad (2.1)$$

where $\int$ is a shorthand notation for $\int_{R^n}$ and X is in n Euclidean plane R^n.

According to Horova et al. (2008), this equation (2.1) does not have a general closed- form expression, so we result to its asymptotic approximation (AMISE). Hence (2.1) could be factored as;

$$AMISE(H) \approx n^{-1}|H|^{-\frac{1}{2}} R(K) + \tfrac{1}{4} m_2(K)^2 (vec^T H)\psi_4 (vec^T H) \quad (2.2)$$

where

- $R(K) = \int K(X)^2 dX$, with $R(K) = (4\pi)^{-\frac{d}{2}}$ when K is a normal kernel.
- $D^2 f$ is $d \times d$ Hessian matrix of second order partial derivatives of f.
- $\psi_4 = \int (vecD^2 f(X))(vec^T D^2) dX$
- D is a diagonal matrix with elements $X_{11}, X_{22}, ..., X_{dd}$
- vec is the vector operator which stacks the columns of a matrix into a single vector.

We observed that the quality of the AMISE to the MISE is given according to Horova et al. (2008) by

$$MISE(H) = AMISE(H) + o(n^{-1}|H|^{-\frac{1}{2}} + trH^2) \qquad (2.3)$$

where o indicates the usual o notation. This implies that AMISE is a 'good' approximation of the MISE as $n \to \infty$. It has been shown that optimal bandwidth selector H has $H = O(n^{-\frac{2}{(d+4)}})$. According to Doung and Hazelton (2005b) substituting this into equation (2.3) yields the optimal $MISE(H)$ order as $O(n^{-\frac{4}{(d+4)}})$. The big O notation is applied element-wise. So when $n \to \infty, MISE \to 0$. This implies the kernel density estimate converges in mean squared error and so also in probability to the true density f. According to Wand and Jones (1995) and Horova et al. (2008), they asserted that it was better to estimate optimal MISE element-wise. They further asserted that the ideal optimal bandwidth selector that is point wise adaptive is given by

$$H_{AMISE} = agr \min_{h \in \mathrm{H}} \tilde{AMISE}(H) \qquad (2.4)$$

Since this ideal bandwidth selector contains the unknown density function f, that cannot be used directly. So some data density based approaches fixed the choice of bandwidth constant. However, we shall adopt point-wise adaptive bandwidth procedures in estimating densities. This implies that our bandwidth selection should be for every data 'element-wise 'adaptive to achieve the desire optimality.

The bandwidths used for the cluster approach by Wu et al. (2007) are optimal for information row/column (one dimensional) bandwidth per time in the multivariate data set. That is, it uses one bandwidth in the row or column during row/column cluster bandwidth selection. It is only row or column wise adaptive. Our approach is to make bandwidth selection to be data based on the smallest size of the row or column samples selections from the information matrix (data set). We modified the MCKDE and modified the ICI approach in estimating densities and they are presented below. The quality of the density estimates are assessed by comparing it to the density, obtained under the mean-squared error criterion. The error generated using these approach would be considered.

This work, present two novel data-driven methods that require the knowledge of pilot plot from optimal fixed window size and the variance of the estimate. This invariably reduces the amount of error at arriving at the "true density". These were achieved under a sufficiently smoothing technique in the existing approaches. They are adaptive approaches based on the data at hand. The aim is to reduce under fitting and over fitting as the case may be and improve statistical inference.

Methodology

The Modified Multivariate Cluster Sampling Kernel Density Estimation (MMCKDE)

This procedure is basically a minimization of $AMISE(H)$ with respect to H, where it is equivalent to the selection of optimal $h_{ij}\ in\{H_1, H_2,..., H_n\}$. This method is a modification of the cluster sampling approach to density estimate. The modified multivariate cluster sampling kernel density estimate (MMCKDE) is a modification of cluster sampling kernel density estimates by adjusting the amount of bandwidths using some idea from the kernel nearest neighbour estimation of the density to the multivariate data. Its smoothing parameter would be an $n \times d$ dimensional matrix obtained from forming relevant number of clusters in an information matrix. The Euclidean distance would be used to form bandwidths.

Let $h = h^* b$. According to Silverman (1986), we call b the bandwidth factor and h^* the global smoothing parameter. The common procedure is to first choose b adaptively and then h^*, by regarding b as fixed. But Wu et al. (2007) used $h = h^* b_i$, where $b = (b_1,...,b_n)$ are the bandwidths factors reflecting the average

local clusters from X_i and adopt the stabilized fixed bandwidths selector of Wu et al. (2006) to select the global smoothing parameter. This approach gives a diagonal bandwidth matrix of varying smoothing parameters h_i . In our proposed approach, we aim at element-wise adaptive density estimation for any given data set X_{ij} . Let assume

$$H = h_i^* b_{i^* j} \qquad (3.1.1)$$

where $i = 1,...,n$, $i^* = 1,...,n_i$ and $j = 1,...,d$.

with H a finite set of optimal bandwidths $H = H_1,...,H_n$ and each $H_i = h_{ij}$ .We choose our h_i^* via each information data rows' h_{MSE} . That is using the MSE approach to get each h_i^* . This is more data sensitive than any fixed h^* . We have more bandwidth factors according to the number of clusters form (starting from step 3 in the proposed algorithm) from the element wise groups from the information data rows.

Then $b_{i^* j}$ will be small if as n_{i^*} is large (that is a large number of mergers involving X_i). Basically, from the data set, the above scheme clusters are formed from the nearest nested sequence of clusters information data rows' elements with the property;

$$\{X_i\} = C_{i0} \subset C_{i1} \subset ... \subset C_{in_{i^*}} = \{X_1,..., X_n\} \qquad (3.1.2)$$

This procedure gives a full bandwidth matrix of vary smoothing parameters for possible values of data sizes for *i* rows and *j* columns. $i \le j \; and \quad i > j$.

To correct the problem of discontinuities at some points in the cluster sampling approach to MCKDE, points of discontinuity in the estimation are identified using the cluster sampling approach as a pilot guide. In this case, the use of standard techniques from cluster analysis is applied. Here, a modified sampling idea similar to Wu et al. (2006) is developed. In this case, when we consider the bandwidth factor $b_i \; to \; X_i$ according to the number of clusters form, and use the idea of density at the boundaries to choose the bandwidths *H*. Wu et al (2006) used the average cluster method which reflects the average local clustering form. In this work a proposed scheme to address points of discontinuities is suggested.

Supposed that f is a density function such that $f(x) = 0$ for $x < 0$ and $f(x) > 0$ for $x \ge 0$. We further suppose that f'' is continuous away from $x = 0$. Then, we have $\hat{f}(x;h) = \int_{-1}^{\alpha} k(z) f(x - hz) dz,$ where $0 \le \alpha \le 1$ -see Wand and Jones (1995, 46-47). Then at the boundary they obtained

$$E\,\hat{f}(0;h) = \tfrac{1}{2} f(0) + O(h) \qquad (3.1.8)$$

We use this idea base on the intuitive knowledge of kernel estimator having to find a compromise between estimating two distinct values of *f* on either side of discontinuity. We propose the use of semi inter-quartile range at

the boundary values. Since the location of the boundary of $\hat{f}(x;H)$ is usually known, we adopted this to achieve better performance in its vicinity. Suppose, we have for S number of row clusters and T number of column clusters, we have;

$$d_{ST} = \sum_{i=1}^{n_S}\sum_{j=1}^{n_T} d_{ij} \qquad (3.1.9)$$

where $d_{ij} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}(X_{ij} - X_{i+1j+1})^2}$ see Gray (1997) for lengths and distances' details. Then

$$H = H_i = \frac{H_i}{v} \text{ and } H_{i+1} \le H_i . \qquad (3.1.10)$$

where $H_i = \{h_{ij}\}$. Subjectively we adopt $v = 2$, where v is a positive real number.

The bandwidth sizes obtained are substituted into equation (1.1) above to obtain accompanying density estimates. The proposed algorithm is presented below. The modified procedures are stated below:

Algorithm 1.

Step 1: start with n clusters, each containing a single observation and an *n*x*n* symmetric matrix of distances $D = \{d_{ij}\}$.

Step 2: Search the distance matrix for the nearest pair of clusters. Let the distance between the ''nearest'' clusters S and T be $d_{ST} = \sum_{i=1}^{n_S}\sum_{j=1}^{n_T} d_{ij}$ in the case of observation *i* in the cluster S and observation *j* in the cluster T, and n_S *and* n_T are the number of observations in cluster S and cluster T, respectively.

Step 3: Merge (combine) cluster S and T. Label the newly formed cluster (ST). Update the entries in the distance matrix by (a) deleting the row's element and column's element corresponding to clusters S and T elements and (b) adding a row's element and a column's element giving the distances between cluster (ST) and the remaining clusters elements.

Step 4: Repeat steps 2 and 3 a total of $n-1$ times so that all observations will be in a single cluster at termination of the algorithm. Record the clusters that are merged and the distance levels at which the mergers take place.

Step 5: Let b_{i^*j} distance level of X_i in the dendrogram. Specifically, if n_{i*} denotes the total number of times that a cluster containing X_i is merged into a larger cluster (that is, total number of mergers that involve X_i), and $\ell_{1i^*},...,\ell_{n_i i^*}$ the distance level at which these n_i mergers take place, then $b_{i^*j} \equiv \ell_{1i^*},...,\ell_{n_i i^*}$.

Step 6: generate $H_i = h_i^* b_{i^*j}$ where h_i^* are determined via the MSE for each information data rows, and let each $H_i = h_{ij}$.

Step 7: In the case of discontinuities, begin by applying (a) $d_{ST(Opt)} = H = \frac{H_i}{2}$ and (b) $H_{i+1} \leq H_i$ in the identified points in H_i from the pilot plot. The window sizes obtained are substituted into equation (1.1) above to obtain accompanying density estimates.

The Modified Intersection of Confidence Intervals ($MICI_H$) Approach

The $MICI_H$ procedure is basically a minimization of $AMISE(H_i)$ with respect to H, where it is equivalent to the selection of optimal $h_{ij}\ in\{H_1, H_2, ...H_n\}$. Our data driven bandwidth matrix selector $\hat{H}$ is point wise data base selection approach. Its density uses a pilot plot in order to address identified problem(s).

$$\hat{H} = agr \min_{H_{D_j} \in \mathrm{H}} \tilde{AMISE}(H). \qquad (3.2.1)$$

Assuming that

$$H = \{H_1 \leq H_2 \leq ....H_n\} \qquad (3.2.2)$$

is a finite collection of window sizes, starting with a smallest $h_{ij} \in H$ and we determine a sequence of confidence intervals given by;

$$\left.\begin{aligned} &D_{ij} = [L_{ij}, U_{ij}], \quad i=1,...,n, j = 1, ..., d \\ &\bar{L}_{ij} = \hat{f}_{H_j}(X_i) - \beta.std\{\hat{f}_{H_j}(X_i)\} \\ &\underline{U}_{ij} = \hat{f}_{H_j}(X_i) + \beta.std\{\hat{f}_{H_j}(X_i)\} \end{aligned}\right\} \quad (3.2.3)$$

each h_{ij} corresponding to a value in $H_i \in H$. We assume the data at hand is normally distributed. Subjecting the data to normality, we propose $\beta = 1.06$ via normal reference rule of Silverman (1986). Then

$$H_{opt_i}(X) = [\frac{abs}{v}[\bar{L}_{ij}, \underline{U}_{ij}]] \qquad (3.2.4)$$

where $abs[\bar{L}_{ij}, \underline{U}_{ij}] = \left|\bar{L}_{ij} - \underline{U}_{ij}\right| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}\left|\bar{L}_{ij} - \underline{U}_{ij}\right|^2}$ see Gray (1997) for lengths and distances' details.

Subjectively, we adopt $v = 2$, considering pilot plots. Where v is a positive real number.

The $MICI_H$ procedure is based on consideration of the intersection of the adjusted intervals D_{ij}, $1 \leq i \leq n$ *and* $1 \leq j \leq d$. We adopt the bandwidth sizes $H_{opt_j}(X)$ to generate full bandwidths of smoothing parameters;

$$H_{opt_i}(X) = [\frac{abs}{2}[\bar{L}_{ij}, \underline{U}_{ij}]] \text{ with } H_{opt_i}(X) \leq H_{opt_{i-1}}(X) \quad (3.2.5)$$

Consequently, substituting bandwidths $H_{opt_i}(X)$ from equation (3.2.5) into the kernel density estimator in (1.1) to obtain the density estimates. Thus, the proposed algorithm is as follows:

Algorithm 2.

Step 1 $\bar{L} \Leftarrow -\infty, \bar{U} \Leftarrow \infty$

Step 2 while $(\bar{L} \le \bar{U})$ and $(i \le J)do$

Step 3 $\bar{L}_{ij} \Leftarrow \hat{f}_{H_i}(X_i) - \beta.std\{\hat{f}_{H_i}(X_i)\}$

Step 4 $\bar{U}_{ij} \Leftarrow \hat{f}_{H_i}(X_i) + \beta.std\{\hat{f}_{H_i}(X_i)\}$

Step 5 $\bar{L}_{ij} \Leftarrow \max[\bar{L}, \bar{L}_{ij}], U_j \Leftarrow \min[\bar{U}, U_{ij}]$

Step 6 $i \Leftarrow i+1$

Step 7 $H_{opt_i}(x) = [\frac{abs}{2}[\bar{L}_{ij}, \bar{U}_{ij}]]$

Step 8 do $i \Leftarrow i+1$

Step 9 $H_{opt_i}(X) \le H_{opt_{i-1}}(X)$

Step 10 compute $h_{ij}\ in\ H_i \in H$

Step 11 end while $(i = n)$.

Results

Application/ Results

Here we use the data of Little and Rubin (2002, Pg 310, exercise 14.7) with missing observations of a survey of 20 graduates of a university class five year after graduation with missing data of race (White or Others) and income (in Dollar), with estimates based on mode related expectation adaptive maximization (MEAM) imputation. 1 represents male, 2 represents female. 1 represents white race, 2 represents other race. The results are presented in Table 1. The obtained bandwidths from the MMCKDE and MICI$_H$ approaches are substituted into equation (1.1) to get the resulting density presented in Table 2. The Mode-related Expectation Adaptive Maximization (MEAM) based on the Expectation Maximization (EM) approach for the data in missing data experiment showed reduced mean squared error and faster rate of convergence compared to some other approaches, hence it use for imputation (see Ogbeide 2018).

Table 1. *The Estimates of Data Set with Missing Observations in Little and Rubin (2002,Pg 310) using the MEAM Approach*

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Sex	1	1	1	2	2	2	2	2	2	2	2	1	1	2	2	1	1	1	2	2
Race	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	-	-	-	-	-
$MEAM_{Race}$	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	1	1
Income	25	46	31	05	16	26	08	10	02	-	-	20	29	-	32	-	-	38	15	-
$MEAM_{Income}$	25	46	31	05	16	26	08	10	02	11.1666	11.2380	20	29	37.292	32	34	34.6875	38	15	11.6005

Table 2. *Estimated Densities for the Multivariate Cluster Sampling Kernel Density Estimation (MCKDE), the Modified Multivariate Cluster Sampling Kernel Density Estimation (MMCKDE) and MICI$_H$ Approaches from the Data Set with Missing Observation in Little and Rubin (2002) (2002, Pg 310)*

Data point	Density estimates from various bandwidths approaches							
X	Fixed H density $_{Race}$	MCKDE $_{Race}$	MMCKDE $_{Race}$	MICI$_H$ Race	Fixed H $_{Income}$	MCKDE $_{Income}$	MMCKDE $_{Income}$	MICI$_H$ $_{Income}$
1	0.0414	0.0414	0.0414	0.0414	0.0543	0.0543	0.0543	0.0543
2	0.0414	0.0414	0.0414	0.0414	0.099	0.0981	0.099	0.0993
3	0.0414	0.0414	0.0414	0.0414	0.0674	0.0660	0.0674	0.0674
4	0.0414	0.0414	0.0414	0.0414	0.0109	0.0109	0.0163	0.0171
5	0.0414	0.0414	0.0414	0.0414	0.0348	0.0348	0.0370	0.0382
6	0.0414	0.0414	0.0414	0.0414	0.0565	0.0565	0.0770	0.0830
7	0.0414	0.0414	0.0414	0.0414	0.0177	0.0174	0.0174	0.0172
8	0.0414	0.0414	0.0414	0.0414	0.0301	0.0331	0.0329	0.0331
9	0.0414	0.0414	0.0414	0.0414	0.0042	0.0043	0.0043	0.0044
10	0.0414	0.0414	0.0414	0.0414	0.0231	0.0279	0.0279	0.0281
11	0.0482	0.0488	0.0499	0.0501	0.0267	0.0312	0.0324	0.0332
12	0.0820	0.0820	0.0820	0.082	0.0431	0.0435	0.0554	0.0556
13	0.0820	0.0820	0.0820	0.0820	0.0621	0.0630	0.0630	0.0640
14	0.0820	0.0820	0.0820	0.0820	0.0846	0.0853	0.0867	0.0872
15	0.0820	0.0820	0.0820	0.0820	0.0693	0.0695	0.0695	0.0699
16	0.0414	0.0414	0.0414	0.0414	0.0414	0.0401	0.0267	0.0269
17	0.0414	0.0414	0.0418	0.0421	0.0826	0.0826	0.0826	0.0791
18	0.0414	0.0414	0.0414	0.0414	0.0825	0.0825	0.0825	0.0831
19	0.0414	0.0414	0.0414	0.0414	0.0341	0.0345	0.0347	0.0334
20	0.0414	0.0414	0.0414	0.0414	0.0279	0.0279	0.0257	0.0250
Density sum	**0.9972**	**0.9978**	**0.9993**	**0.9998**	**0.9523**	**0.9601**	**0.9927**	**0.9995**

https://doi.org/10.30958/ajs.6-4-4 doi=10.30958/ajs.6-4-4

Below are the table of the calculated bandwidth selections errors and convergence rate from the data set with missing observations in Little and Rubin (2002, Pg 310).

The relative errors, h^* (which is the error in relation to the fixed optimal bandwidth value), $AMISE^*$ and the convergence rates of methods are given in Table 3.

Table 3. *Table of Bandwidth Selection Errors and Convergence Rate from the Estimated Bandwidths for the Race and Income Using the Multivariate Cluster Sampling Kernel Density Estimation (MCKDE), the Modified Multivariate Cluster Sampling Kernel Density Estimation (MMCKDE) and the $MICI_H$ Approaches from the Data Set with Missing Observation in Little and Rubin (2002)*

Approach	Relative error v	Variance	δ	h^*	$AMISE^*$	Convergence rate
MCKDE (Race)	0.3000	0.2812	0.5302	0.1637	6.5021×10^{-2}	0.4071
MMCKDE (Race)	0.1000	0.1875	0.4330	0.1091	2.3555×10^{-2}	0.7411
$MICI_H$ (Race)	**0.0080**	**0.0072**	**0.0848**	**0.0041**	5.4365×10^{-3}	**0.9763**
MCKDE (Income)	-0.0097	8.003	2.8289	4.6596	8.9928×10^{-2}	1.0029
MMCKDE (Income)	-0.2085	7.7639	2.7863	4.5204	5.7629×10^{-2}	1.8675
$MICI_H$ (Income)	**-0.2313**	**6.9157**	**2.6297**	**4.0265**	5.5502×10^{-2}	**1.9995**

Table 3 showed that there are reduced relative errors, h^* (which is the error in relation to the fixed optimal bandwidth value) and $AMISE^*$ in the proposed methods. The proposed methods have faster convergence rates compared to their original versions. That is, the $MICI_H$ have lower error propagation and faster convergence rates when used to estimates the Little and Rubin (2002) data with fixed optimal H, MCKDE and the MMCKDE approaches respectively.

The estimated bandwidth selection errors and convergence rates from the data set with missing observation in Little and Rubin (2002, Pg 310) data, via the various methods favour the use of the MICI approach over the other approaches. This is because its bandwidth errors are smaller as well as having higher convergence rate. The MMCKDE has some improvement over the MCKDE approach. These can be seen in Tables 2 and 3. Generally, the AMISE shows the difference between the "true density" and the estimated density. The AMISE for $MICI_H$ is smaller than that of MMCKDE and MCKDE approaches. The graphical densities displays of the data are given below (Figures 1 and 2).

Figure 1. *Density for Race using the Fixed H, MCKDE, MMCKDE and MICI$_H$ Approaches*

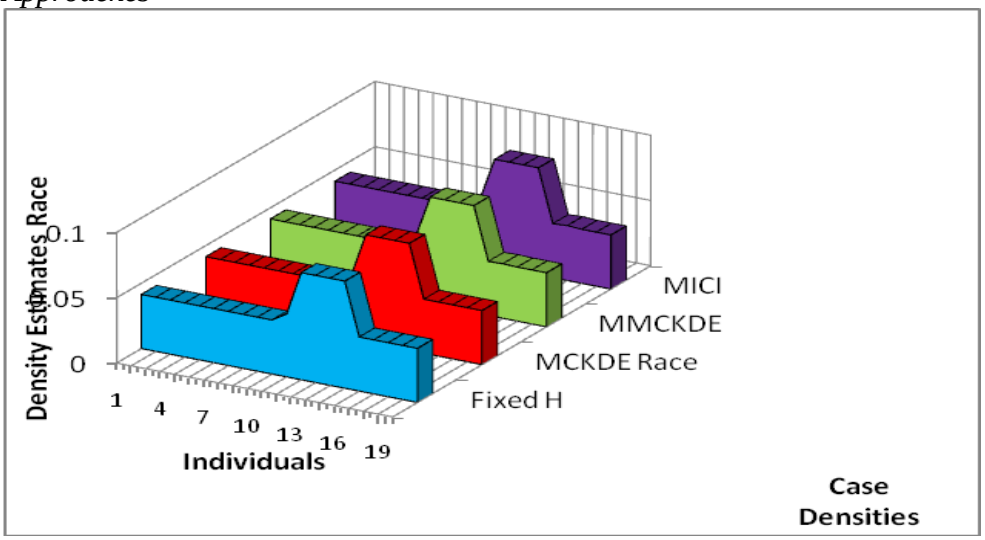
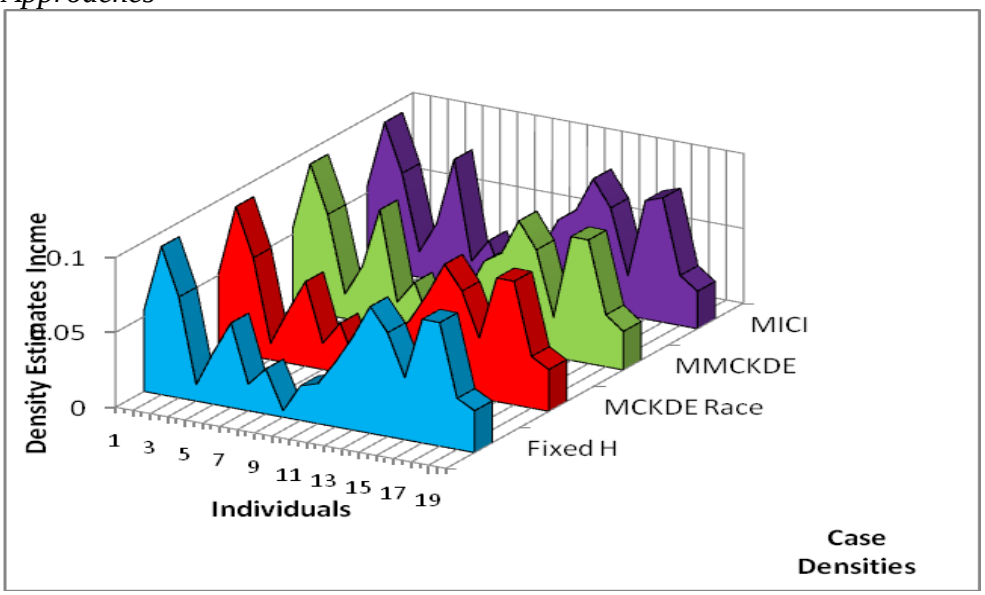


Figure 2. *Density for Income using the Fixed H, MCKDE, MMCKDE and MICI$_H$ Approaches*



The surface plots are given below.

Figure 3a. *Graphical Density Estimates for Race Data using the Fixed H Approach*

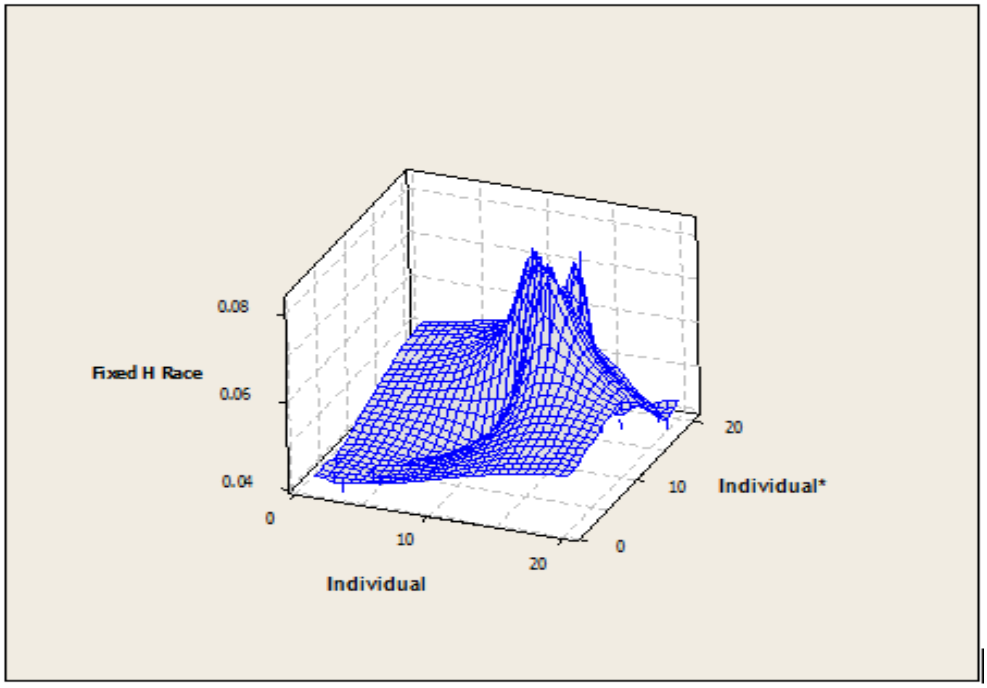


Figure 3b. *Graphical Density Estimates for Race Data using the MCKDE Approach*

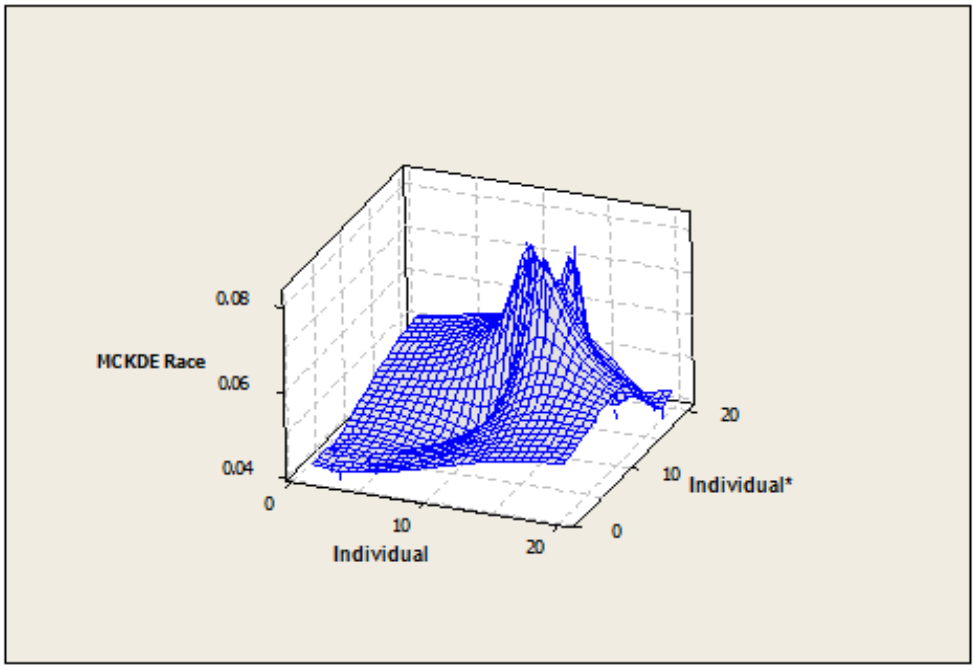


Figure 3c. *Graphical Density Estimates for Race Data using the MMCKDE Approach*

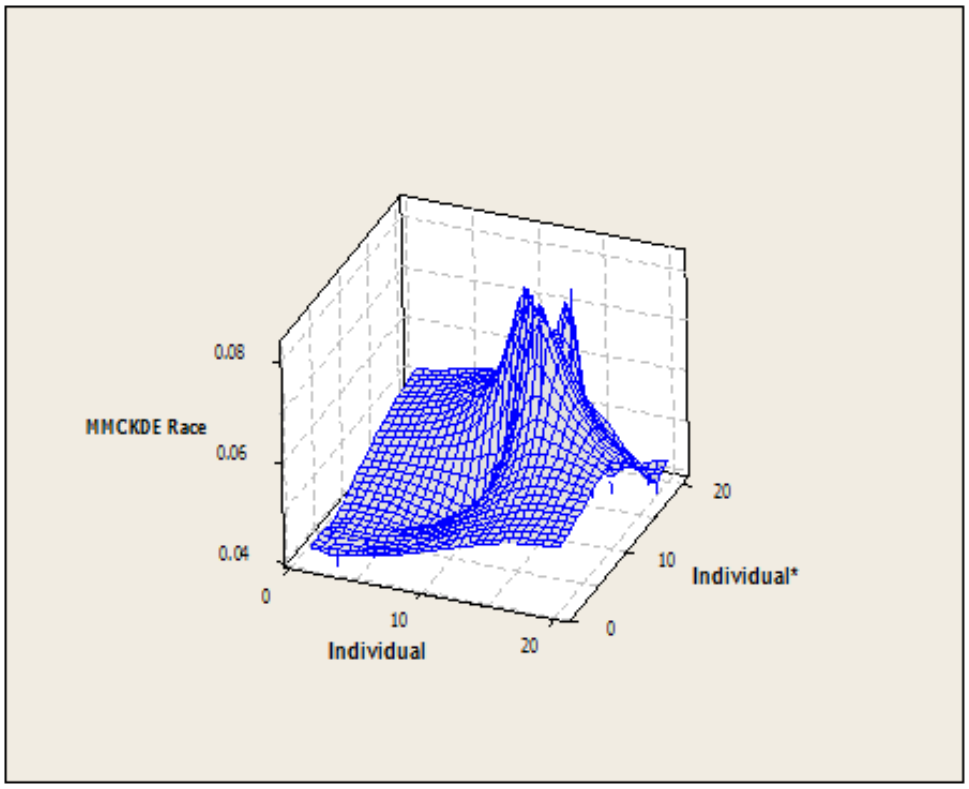


Figure 3d. *Graphical Density Estimates for Race Data using the $MICI_H$ Approach*

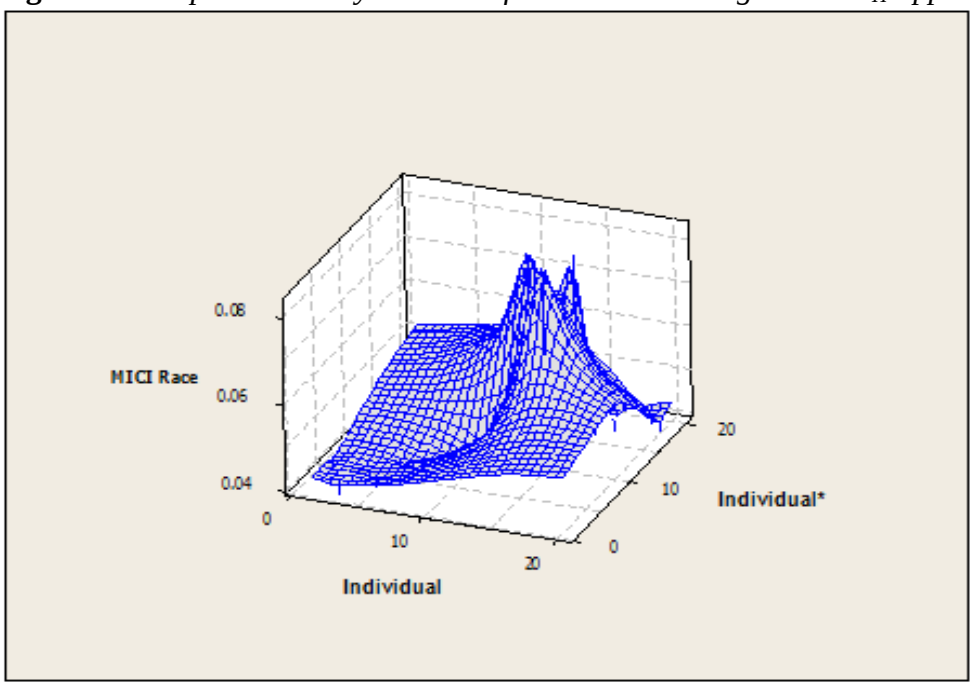


Figure 4a. *Graphical Density Estimates for Income using the Fixed H Approach*

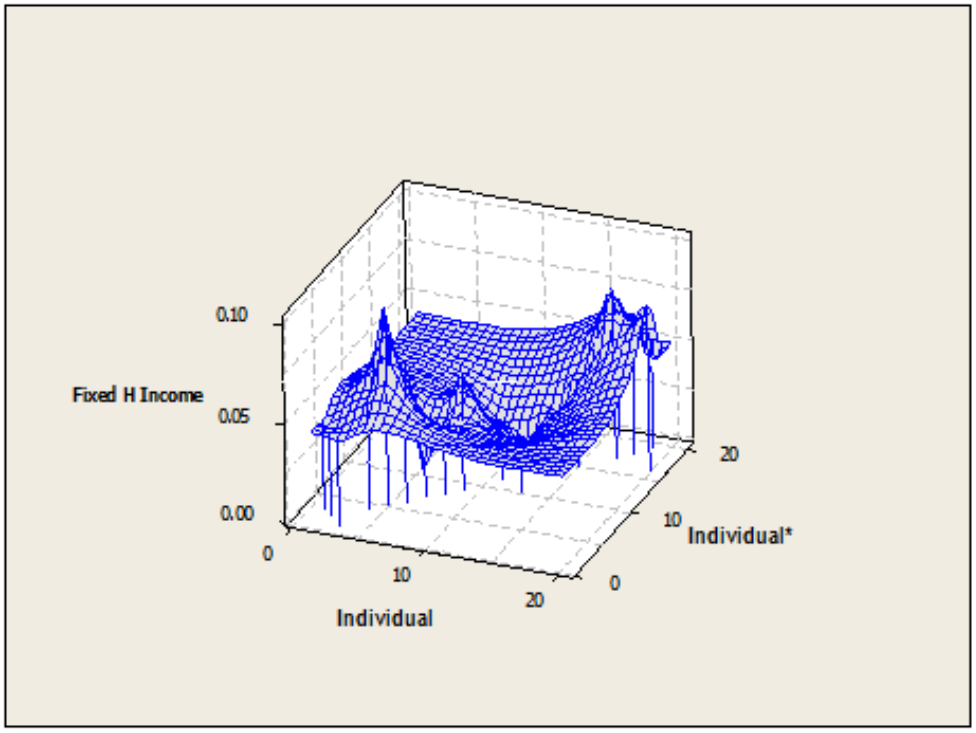


Figure 4b. *Graphical Density Estimates for Income using the MCKDE Approach*

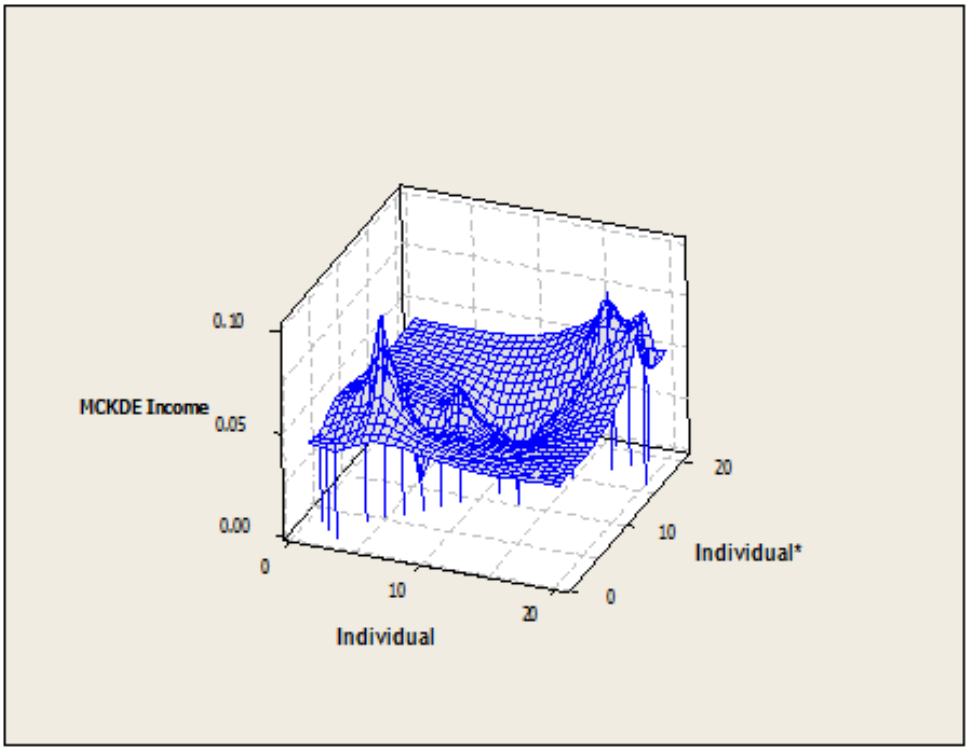


Figure 4c. *Graphical Density Estimates for Income using the NMCKDE Approach*

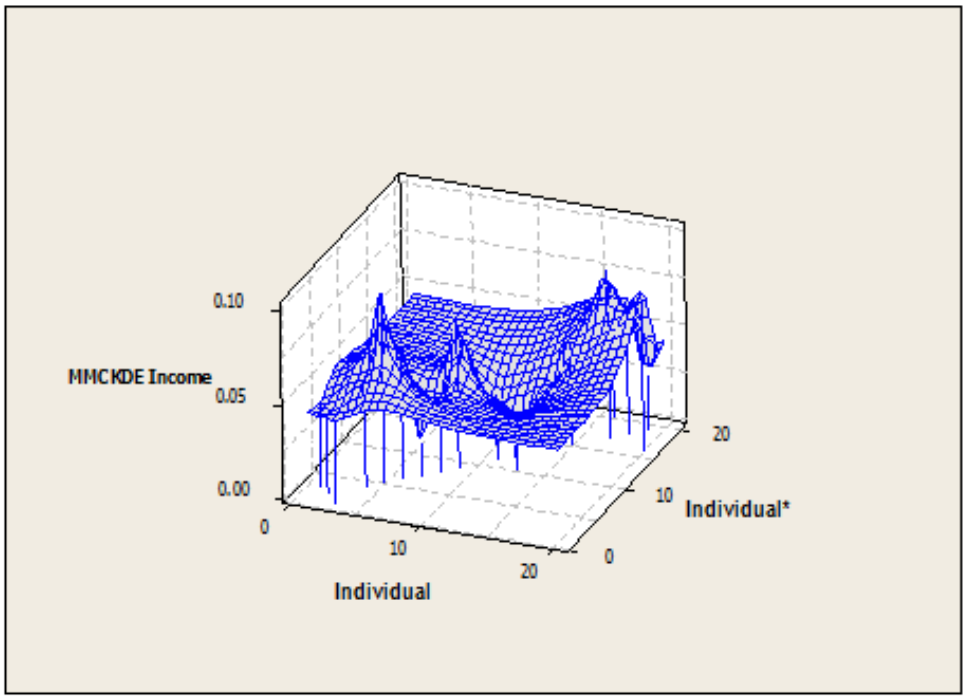
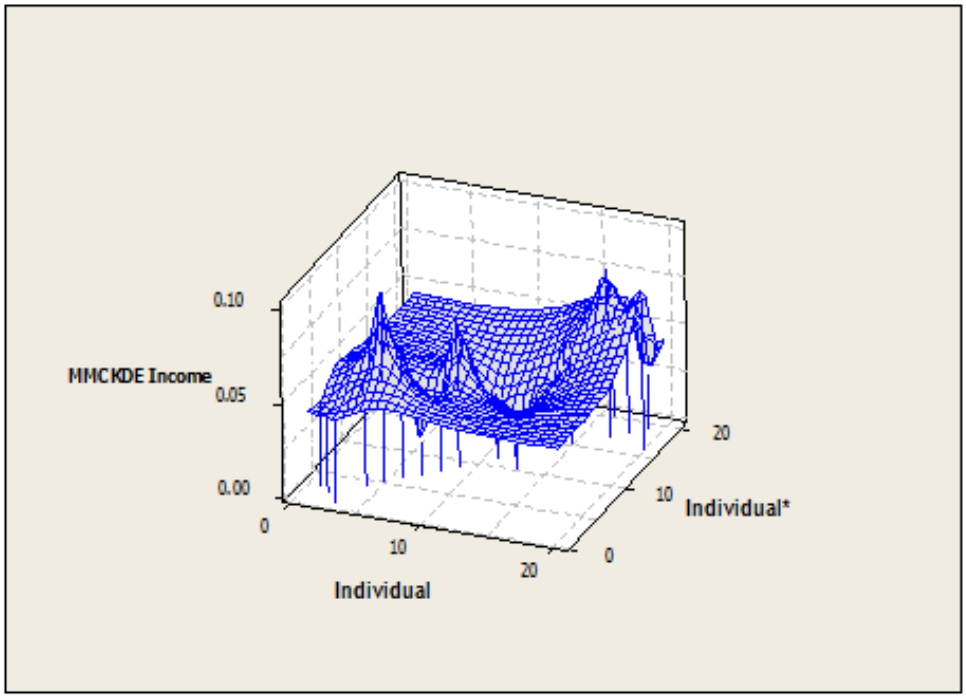


Figure 4d. *Graphical Density Estimates for Income using the $MICI_H$ Approach*



The various approaches have identifiable differences from Figures 3a-4d, using the fixed H, MCKDE, MMCKDE and $MICI_H$ for the dataset in Little and Rubin (2002).

The MMCKDE corrects identified cluster sampling points of discontinuities in the multivariate kernel nearest neighbourhood density estimate. The $MICI_H$ method which is based on the ICI rule, produces smaller but optimal smoothing parameters extended to the multivariate data set. This is an attempt to achieve reduced error and show more hidden features of the density (see Marrion and Tsybakov 1996).

The modified Intersection of confidence interval ($MICI_H$) approach in estimating density show better improvements over the other approaches presented. These are seen in the quality of the density estimates assessed by comparing them with the density obtained using the mean-squared error criterion in Table 3 and in Figures 3a-4d.

In practise, the smaller the variance of the estimate, the better will its contribution to the overall density estimation, as we do not know the true density $f(x)$ -Silverman (1986), Wand and Jones (1995), Katkovnik and Shmulevich (2002). We have reduced variances and bias (from *AMISE*) in our proposed approaches (see Table 2 and Table 3).

According to Silverman (1986), Scott (1992), Cao et al. (1994), Wand and Jones (1995), Katkovnik and Shmulevich (2002) one way of evaluating the method of adaptive window size selection is to compare it to the optimal fixed window size (this is a pilot plot). Our new approaches behave in quite a similar manner. The other approaches are to aim at reducing the AMISE rate in the bandwidth selection method and better convergence rate these were also achieved as seen in Table 3.

Conclusions

We propose two new varying bandwidths approaches in order to achieve adaptive multivariate kernel density estimation. The quality of the proposed approaches estimates have shown some improvements when assessed and compared with the estimates obtained using existing approaches. These are seen in the errors generated via the AMISE using these proposed approaches, and the convergence rates compared to some other known approaches when applied to some data sets. The MMCKDE and the $MICI_H$ methods are adaptive approaches to data distribution. The MMCKDE corrects identified points of discontinuities in the MCKDE. The $MICI_H$ is based on the intersection of adaptive confidence intervals. Like in every other improved methods, $MICI_H$ requires only simple two additional steps when compared to the ICI approach. These additional procedures are in the choice and application of the smoothing parameters in the multivariate density estimation. The $MICI_H$ and MMCKDE generate full bandwidth matrices. The cost of these steps brings about the adaptive density constructed. The performance of these approaches with this available data shows that these approaches will perform significantly well for very large size dataset.

References

Abramson IS (1982) Arbitrariness of the Pilot Estimate in Adaptive Kernel Methods. *Journal of Multivariate Analysis* 12(4): 562 -567.

Alan JI (1991) Recent Developments in Nonparametric Density Estimation. *Journal of the American Statistical Association* 86(413): 205-221.

Bowman AW, Azzalini A (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford: Clarendon Press.

Breiman L, Meisel W, Purcell E (1977) Variable Kernel Estimates of Multivariate Density. *Technometrics* 19(2): 135-144.

Cao R, Cuevas A, Manteiga WG (1994) A Comparative Study of Several Smoothing Methods in Density Estimation. *Computational Statistics and Data Analysis* 17(2): 153-176.

Cencov NN (1962) Evaluation of an Unknown Distribution from Observations. *Soviet Math* 3: 1559-62.

Duong T, Hazelton ML (2003) Plug-In Bandwidth Matrix for Bivariate Kernel Density Estimation. *Nonparametric Statistics* 15(1): 17-30.

Duong T, Hazelton ML (2005) Convergence Rates for Unconditional Bandwidth Matrix Selector for Multivariate Kernel Density Estimation. *Journal of Multivariate Analysis* 93(2005): 417-433.

Elio L, Edgar A (2003) *Parallel Computation of Kernel Density Estimates Classifiers and their Ensembles*. Proceedings of International Conference on Computer Communications and Control Tecnnologies.

Fukunaga K (1990) *Statistical Pattern Recognition* (2nd Edition). New York: Academic Press.

Friedman JH, Stuetlzle W (1982) Projection Pursuit Regression Analysis. *Journal of the American Statistical Association* 76(376): 817-823.

Friedman JH, Stuetlzle W, Schneider T (1984) Tool for Viewing Multi-Dimension Surfaces. *SIAM Journal of Science and Statistical Computing*.

Gray A (1997) *The Intuitive Idea of Distance on Surfaces in Modern Geometry of Curves and Surfsces with Mathemtica*. (2nd Edition). Boca Raton, FL: CLC Press, 341-345.

Hall P (1990) On the Bias Variable Bandwidth Curve Estimation. *Biometrika* 77(3): 527-536.

Hall P (1992) On the Global Properties of Variable Band Width Density Estimator. *The Annals of Statistics* 20(2): 762-78.

Hall P, Park BU (1987) Extend to which Least-Squares Cross-Validation Minimises Integrated Squared Error in Non-Parametric Density Estimation. *Probability Theory and Related Fields* 92(1): 1-20.

Hardle W, Scott DW (1992) Smoothing by Weighted Averaging of Rounded Points. *Computational Statistics* 7: 97-128.

Horova I, Kolacek J, Zelinka J, Vopatova K (2008) *Bandwidth Choice for Kernel Density Estimates*. Yokohama, Japan: IASC.

Isenman AJ (1991) Recent Developments in Nonparametriuc Density Estimation. *Journal of the American Statistical Association* 86(413): 205-224.

Jarnicka J (2009) Multivariate Kernel Density Estimation with Parameter Support. *Opuscula Matmtica* 29(1): 41-55.

Jones MC (1990) Variable Kernel Density Estimates and Variable Kernel Density Estimators. *Australian Journal of Statistics* 32(3): 36-71.

Jones MC, Mckay IJ, Hu TU (1994) Variable Location and Scale Density Estimation. *Annals of the Institute of Statistical Mathematics* 46(3): 521-535.

Katkovnik V (1999) A New Method for Varying Bandwidth Selection. *IEEE Transaction in Signal Process* 47(9): 2567-2571.

Katkovnik V, Shmulevich I (2002) Kernel Density Estimation with Adaptive Varying Window Size. *Pattern Recognition Letters* 23(14): 1641-1076.

Little RJA, Rubin DB (2002) *Statistical Analysis with Missing Data* (2nd Edition). New Jersey. USA: Wiley and Sons Publisher.

Marrion JS, Tsybakov A (1996) Visual Error for Qualitative Smoothing. *Journal of American Statistical Association* 90(43): 499-507.

Muller HG (1985) Empirical Bandwidth Choice for Non-Parametric Kernel Regression by means of Pilot Estimators. *Statistical Decisions Supplement* 2: 193-206.

Ogbeide EM (2018) A New Iterative Imputation Method based on Adaptive Expectation Maximization. SAU Science-Tech Journal 3(1): 133-142.

Ogbeide EM, Osemwenkhae JE, Oyegue FO (2016) On a Modified Multivariate Cluster Sampling Kernel Approach to Multivariate Density Estimation. *Journal of Nigerian Association of Mathematical Physics* 34: 123-132.

Osemwenkhae JE (2003) Higher Order Forms in Kernel Density Estimation. PhD Thesis. Nigeria: Department of Mathematics, University of Benin.

Rudemo M (1982) Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics* 9(2): 65-78.

Sain RR (2002) Muiltivariate Locally Adaptive Density Estimation. *Computational Statistics & Data Analysis* 39(2): 165-186.

Salgado-Ugarte IH, Perez-Hernandez MA (2003) Exploring the Use of Variable Bandwidth Kernel Density Estimators. *Stata Journal* 3(2): 1-15.

Scott DW (1992) *Multivariate Density Estimation*. New York: John Wiley.

Scott DW, Thompson JR (1983) Probability Density Estimation in Higher Dimensions. In JE Gentle (Ed.), *Computer Science and Statistics Proceedings of the Fifteen Symposium on Interface, Amsterdam, Holland*, 173-179.

Silverman BW (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simonoff JS (1996) *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Tower S (2002) *Kernel Probability Density Estimation Method*. State University of New York at Stony Book: Seminar Reports on Particular Physics.

Tukey JW (1947) Non Parametric Estimations II. Statistically Equivalent Blocks and Tolerances - The Continuous Case. *AMS* 18: 29-539.

Victor N (1976) Nonparametric Allocation Rules. In FT Dombai, F Gremy (Eds.), *Decision Making and Medical Care: Can Information Help?* North-Holland, Amsterdam, 515-529.

Wand MP, Jones MC (1995) *Kernel Smoothing* London: Chapman and Hall/CRC.

Wu TJ, Tsai MH (2004) Root *n* Bandwidths Selectors in Multivariate Kernel Density Estimation. *Probability Theory and Related Fields* 129(4): 537-558.

Wu TJ, Chen CF, Chen HY (2006) A Variable Bandwidths Selectors in Multivariate Kernel Density Estimation. *Statistics and Probability Letters* 77(4): 462-467.

Wu KL, Shan K., Yung K, Miin-Shen Y, Yuan C (2007) Mean Shift-based Clustering to KDE. *Journal of Pattern Recognition* 40(11): 3035-3052.

Wu D, Tian Y, Datt A. (2008) Analysis of the Stochastic Interplay between Object Maintenance and Churn. *Computer Communications* 31(2): 220-239.