DETECTION OF OUTLIERS IN CIRCULAR DATA USING KERNEL DENSITY FUNCTION

Hazem I. El Shekh Ahmed[1], Ali H. Abuzaid[2] and Ikhlas I. Al Awar

[1]Department of Mathematics, Al Quds Open University-Gaza, Palestine
[2]Department of Mathematics, Al-Azhar University-Gaza, Palestine

ABSTRACT

Background: Outlier detection has recently become an important problem in many industrial and financial applications. The proposal in this paper is based on detect an outlier in circular data by the local density factor (LDF). The name of local density estimate (LDE) is justified by the fact that we sum over a local neighborhood compared to the sum over the whole circular data commonly used to compute the kernel density estimate (KDE).

Methods: We discuss new techniques for outlier detection which find the outliers by comparing the local density of each point to the local density of its neighbors in circular data. In our experiments, we performed simulated two data sets generated a set of circular random variables from von Mises distribution with different sizes and each have two clusters non-uniform density and sizes, then we used (LDF) algorithm.

Results: The results show that (LDF) algorithm detect an outliers in five samples named as A, B, C, D and E using von Mises concentration parameter (k(and suitable smoothing parameter (h) for two different datasets.

Conclusion: It can be concluded from the present study that the proposed method (LDF method) can be very successful for the outlier detection task in circular data.

KEYWORDS

Circular data, Outlier detection, Kernel methods, Local density factor, Local density estimate, Kernel density estimate.

1. INTRODUCTION

Circular data can be measured either in degrees, when they are distributed in the Interval [0°,360°), or in radians, in the interval [0 , 2π). Circular data can arise in contrasting scientific fields such as earth sciences, meteorology, biology, physics, psychology, image analysis, and medicine. A special statistical measure is needed to deal with circular data. Between 5% and 10% of any set of statistical data are surprising points that are often called outliers. They may unduly affect the statistical analysis and the final outcomes [6]. Outlier detection is an important and attractive problem in knowledge discovery in large data sets. The majority of the recent work in outlier detection follow the framework of Local Outlier Factor (LOF), which is based on the density estimate theory. However, LOF has two disadvantages that restrict its performance in outlier detection. First, the local density estimate of LOF is not accurate enough to detect outliers in the complex and large databases. Second, the performance of LOF depends on the parameter k that determines the scale of the local neighborhood [5].Recently, Taylor (2008) proposed a plug-in selector for the case of circular data for the estimator with the von Mises kernel. The selector of Taylor (2008) uses from the beginning the assumption that the reference density is a von Mises to construct the AMISE.

For circular data it is appropriate the use of the von Mises function which is the circular Gaussia n [3]:

$$\hat{f}(\theta,\nu)\frac{1}{n(2\pi)I_0(\nu)}\sum_{i=1}^{n}\exp(\nu\cos(\theta-\theta_i)) \qquad (1.1)$$

Where $I_r(\nu)$ is the modified Bessel function of order r and the concentration parameter ν has the role of the inverse of the smoothing parameter h. Large values of ν lead to highly variable estimations whereas small values provide over smoothed circular densities [7].

A circular observation can be seen as a point on the unit circle, and represented by an angle $\theta \in [-\pi, \pi)$. It is periodic, i.e. $\theta = \theta + 2m\pi$ for $m \in Z$ which sets apart circular statistical analysis from standard real-line methods [11]. Recent accounts are given by Jammalamadaka and SenGupta [12] and Mardia and Jupp [13].Kernel density estimation and kernel smoothing methods in general, are classical topics in nonparametric statistics. [19] and [20] have been brought up the kernel density methodology indifferent contexts, on the other hand, [21] present some background on kernel density estimation for linear data and directional data which is focused in the issue of error measurement and expressions for the AMISE of the estimator and the exact MISE for particular cases of mixtures are obtained, in both the directional and the directional-linear context.

Beyond the linear case, kernel density estimation has been also adapted to directional data, that is, data in the q-dimensional sphere [17]. For the particular case of circular data, there are more recent works dealing with the problem of smoothing parameter selection in kernel density estimation [7] and [3]. [18] studied the kernel density estimator on the q-dimensional torus and propose some bandwidth selection methods. An alternative motivation for the kernel density estimator is provided recently in [16] by smoothing the empirical distribution function that justifies the use of asymmetric kernels while considering density estimation for non-negative random variables.

Since we do not make any assumption about the type of the density, we use a nonparametric kernel estimate to estimate the density of majority data points [14].Data points belonging to different model components may have different density without being outliers. Consequently, normal points in some model components may have lower density than outliers around points from different model components [2].In this paper, we propose a local density estimate (LDF) to compute the kernel density estimate (KDE) to detect outliers in circular data when the data follow a von Mises distribution. The name of local density estimate (LDE) is justified by the fact that we sum over a local neighborhood compared to the sum over the whole circular data commonly used to compute the kernel density estimate (KDE). The circular distance is considered as a statistic test to identify outliers in this estimation.

2. MOTIVATION FOR THE CIRCULAR KERNEL DENSITY ESTIMATOR

We consider estimation of the density for circular data, i.e. an absolutely continuous circular density $f(\theta)$ [15].

Let $f_n \subset C^*$ where C^* denotes the set of periodic analytic functions with a period 2π.
We say that f_n is an approximate identity if:

1- $f_n(\theta) \geq 0 \quad \forall \quad \theta \in [-\pi, \pi]$;

2- $\int_{-\pi}^{\pi} f_n(\theta) = 1$;

3- $\lim_{n\to\infty} \max_{|\theta| \geq \delta} f_n(\theta) = 0$ for every $\delta > 0$.

Given a random sample $\theta_1, ..., \theta_n$ for the above density, the kernel density estimator may be written as:

$$\hat{f}(\theta; h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{\theta - \theta_i}{h}) \qquad (2.1)$$

In this method, for n data samples of dimensionality $\dim$; the distribution density can be estimated as:

$$\tilde{q}(\theta_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(\theta_i)^{\dim}} K(\frac{\theta - \theta_i}{h(\theta_i)}), \qquad (2.2)$$

where K is a kernel function (satisfying non-negativity and normalization conditions)and $h(\theta_i)$ are the bandwidths implemented at angular θ_i.

If K is a probability density function, then the sample smoothing estimator $\tilde{q}(\theta_i)$ is automatically a probability density function by Terrell and Scott [14].

We consider the circular kernel density estimator using the wrapped Cauchy kernel that is given by:

$$\widehat{f_{WC}} = \frac{1}{N}\sum_{j=1}^{N} K_{WC}(\theta - \theta_j; 0, \rho) \qquad (2.3)$$

If we have the wrapped Cauchy distribution with location parameter μ and concentration parameter ρ is given by:

$$K_{WC}(\theta; \mu, \rho) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}, -\pi \leq \rho < \pi \qquad (2.4)$$

that becomes degenerate at $\theta = \mu$ as $\rho \to 1$.

In our case, we can define the function as in [2]:

$$K^*(u) = \lambda K(\lambda u),\ \lambda > 0 \quad and \quad h(\theta_i) = hd_k(\theta_i),$$

where $d_k(.)$ denotes the distance to the k th nearest neighbor of angular θ_i.

This can be used to select a scale that is appropriate for circular data.

Since we are interested in detecting outlier data samples based on comparing them to their local neighborhood, the sum in Eq. (2.1) needs only to be taken over a sufficiently large neighborhood of an angular θ. Let $\alpha N\beta(\theta)$ denotes the smallest arc lengths between two points α and β along the circumference and α is nearest neighbors of an angular θ.

Thus, from Eq. (2.1) and (2.2) we obtain the following formula for distribution density at θ_i:

$$\tilde{q}(\theta_j) \propto \frac{1}{\alpha} \sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{1}{h(\theta_i)^{\dim}} K\left(\frac{\theta_j - \theta_i}{h(\theta_i)}\right), \qquad (2.5)$$

$$= \frac{1}{\alpha} \sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{1}{(2\pi)^{\frac{\dim}{2}} h(\theta_i)^{\dim}} \exp\left(-\frac{d(\theta_j, \theta_i)^2}{2h(\theta_i)^2}\right) \qquad (2.6)$$

Here,

$$d_0(\theta_j, \theta_i) = \min(\theta_j - \theta_i, 2\pi - (\theta_j - \theta_i)) = \pi - \left|\pi - \left|\theta_j - \theta_i\right|\right| \qquad (2.7)$$

is a circular distance between two angles θ_i and θ_j.

Observe that circular distance from Eq. (2.6) may be very small if there is a neighbor θ_i very close to sample θ_j. In such a case, it is possible to misleadingly obtain a large density estimate $\tilde{q}(\theta_j)$. To prevent such issues and increase the robustness of the density estimation, following the LOF approach [8], we computer each ability distance for each sample θ_j with respect to an angular θ_i as follows:

$$rd_k(\theta_j, \theta_i) = \max(d(\theta_j, \theta_i), d_k(\theta_i)), \qquad (2.8)$$

Where $d_k(\theta_i)$ is the distance to k th nearest neighbor of an angular θ_i. Eq. (2.7) prevents the distance from θ_j to θ_i to become too small with respect to the neighborhood of an angular θ_i.

We obtain our local density estimate (LDE) by replacing the Euclidean distance in Eq.(2.5) with the reach ability distance:

$$LDE(\theta_j) \propto \frac{1}{\alpha} \sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{1}{(2\pi)^{\frac{\dim}{2}} h(\theta_i)^{\dim}} \exp(-\frac{rd_k(\theta_j, \theta_i)^2}{2h(\theta_i)^2}), \quad (2.9)$$

$$= \frac{1}{\alpha} \sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{1}{(2\pi)^{\frac{\dim}{2}} (h.d_k(\theta_i))^{\dim}} \exp(-\frac{rd_k(\theta_j, \theta_i)^2}{2(h.d_k(\theta_i))^2}), \quad (2.10)$$

Local density estimate (LDE) is justified by the fact that we sum over a local neighborhood $\alpha N\beta$ compared to the sum over the whole circular data commonly used to compute the kernel density estimate (KDE), as shown in Eq. (2.2).

3. THE PROPOSED METHOD

LDE is not only computationally more efficient than the density estimate in Eq.(2.1) but yields more robust density estimates.

LDE is based on the ratio of two kinds of distances:

1- The distance from a point θ_j to its neighbors θ_i and

2- Distances of the neighboring points θ_i to their k-th neighbors.

Namely, the exponent term in Eq. (2.9) is a function of the ratio $\frac{rd_k(\theta_j, \theta_i)}{d_k(\theta_i)}$, which specifies how is the reach ability distance from θ_j to θ_i related to the distance to the k-th nearest neighbor of θ_i. In fact, we use $d_k(\theta_i)$ as a "measuring unit" to measure the circular distance $d(\theta_j, \theta_i)$.

If $d(\theta_j, \theta_i) \le d_k(\theta_i)$, then the ratio $\frac{rd_k(\theta_j, \theta_i)}{d_k(\theta_i)}$ is equal to one (since $rd_k(\theta_j, \theta_i) = d_k(\theta_i)$), which yields the maximal value of the exponential function ($\exp(-\frac{1}{2h^2})$).

Conversely, if $d(\theta_j, \theta_i) > d_k(\theta_i)$ then the ratio is larger than one, which results in smaller values of the exponent part. The bandwidth h specifies how much weight is given to $d_k(\theta_i)$. The larger h, the more in influential are the k nearest neighbors that are further away. The smaller h, the more we focus on k nearest neighbors. Observe that we compare a given point θ_j to its neighbors in $\alpha N\beta(\theta_j)$. It is important that the neighborhood $\alpha N\beta(\theta_j)$ is not too small

(otherwise, the density estimation would not be correct). Overly large α does not influence the quality of the results, but it influences the computing time (to retrieve α nearest neighbors).
Having presented an improved local version of a nonparametric density estimate, we are ready to introduce our method to detect outliers based on this estimate. In order to be able to use LDE to detect outliers, the local density values $LDE(\theta_j)$ need to be related to the values of neighboring points. We define Local Density Factor (LDF) at a data point as the ratio of average LDE of its α nearest neighbors to the LDE at the point:

$$LDF(\theta_j) \propto \frac{\sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{LDE(\theta_i)}{\alpha}}{LDE(\theta_i) + c.\sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{LDE(\theta_i)}{\alpha}} \qquad (3.1)$$

where, c is a scaling constant.

The scaling of LDE values by c is needed, since $LDE(\theta_j)$ may be very small or evenequal to zero (for numerical reasons), which would result in very large or eveninfinity values of LDF if scaling is not performed, i.e.,

If $c = 0$ in Eq. (3.1) Observe that the LDF values are normalized on the scale fromzero to $\frac{1}{c}$. Value zero means that $LDE(\theta_j) >> \sum_{\theta_i \in \alpha N\beta(\theta_j)} \frac{LDE(\theta_i)}{\alpha}$ while value $\frac{1}{c}$ means that $LDE(\theta_j) = 0$. The higher the LDF value at a given point (closer to $\frac{1}{c}$) the more likely the point is an outlier.

Let $\sum_i$ be the covariance matrix estimated on the α data points in $\alpha N\beta(\theta_i)$. If weuse a general Gaussian kernel with covariance matrices $\sum_i$, then Eq. (2.5)becomes:

$$\tilde{q}(\theta_j) \propto \sum_{i=1}^{n} \frac{1}{h^{\dim} \left| \sum_i \right|^{\frac{1}{2}}} \exp(-\frac{d_{\sum}(\theta_j, \theta_i)^2}{2h^2}), \qquad (3.2)$$

where $d_{\sum}(\theta_j, \theta_i)^2 = (\theta_j - \theta_i)^T \sum_i^{-1} (\theta_j - \theta_i)$ is the Mahalanob is distance of vectorsθ_i and θ_j.

Let $A = diag(\lambda_1, \lambda_2, ..., \lambda_k)$ be diagonal matrix of eigenvectors, $V = [v_1, v_2, ..., v_k]$ is matrix of eigenvectors of the covariance matrix $(\sum_i)$ and (μ) is the mean of vectors in the $\alpha N \beta$ neighborhood. Then we can write $d_{\sum}(\theta_j, \theta_i)^2$ according to [2] as:

$$d_{\sum}(\theta_j, \theta_i)^2 = (\theta_j^* - \theta_i^*)^T . (\theta_j^* - \theta_i^*), \qquad (3.3)$$

where

$$\theta^* \equiv (A^T)^{\frac{-1}{2}} V^T (\theta^* - \mu), \qquad (3.4)$$

Therefore, Eq. (3.2) can be, using Eq. (3.3) and Eq. (2.6) represented in the form:

$$\tilde{q}(\theta_j) \propto \sum_{i=1}^{n} \frac{1}{h^{\dim} \left| \sum_i \right|^{\frac{1}{2}}} \exp\left(-\frac{d_{\sum}(\theta_j^*, \theta_i^*)^2}{2h^2}\right), \qquad (3.5)$$

Now, analogous to Eq. (2.9), we may generalize the LDE measure to:

$$LDE(\theta_j) \propto \frac{1}{\alpha} \sum_{\theta_i \in \alpha N \beta(\theta_j)} \frac{1}{(2\pi)^{\frac{\dim}{2}} h(\theta_i)^{\dim} \left| \sum_i \right|^{\frac{1}{2}}} \exp\left(-\frac{rd_k(\theta_j^*, \theta_i^*)^2}{2h(\theta_i)^2}\right), \qquad (3.6)$$

Equation (3.6) can be replaced within Eq. (3.1) to obtain generalized measure of the local density factor.

Now, we can summary LDE method for circular data by this algorithm:

1- Generalize the LDE measure:

$$LDE(\theta_j) \propto \frac{1}{\alpha} \sum_{\theta_i \in \alpha N \beta(\theta_j)} \frac{1}{(2\pi)^{\frac{\dim}{2}} h(\theta_i)^{\dim} \left| \sum_i \right|^{\frac{1}{2}}} \exp\left(-\frac{rd_k(\theta_j^*, \theta_i^*)^2}{2h(\theta_i)^2}\right)$$

2- Define Local Density Factor (LDF) at a data point as the following:

$$LDF(\theta_j) \propto \frac{\sum_{\theta_i \in \alpha N \beta(\theta_j)} \frac{LDE(\theta_i)}{\alpha}}{LDE(\theta_i) + c. \sum_{\theta_i \in \alpha N \beta(\theta_j)} \frac{LDE(\theta_i)}{\alpha}}.$$

where, c is a scaling constant.

3- If $c = 0$; then LDF values are normalized on the scale from zero to $\frac{1}{c}$, otherwise, the higher the LDF value at a given point (closer to $\frac{1}{c}$) the more likely the point is an outlier.

4. EXPERIMENTS

In all of our experiments, we have assumed that we have information about thenormal behavior (normal class) and rare events (outliers) in the data set. Recall that LOF algorithm [8] has been designed to properly identify outliers as data samples with small local distribution density, situated in vicinity of dense clusters.

To proposed our LDF algorithm, we created two data sets $Dataset1$ and $Dataset2$ and we generated synthetic data sets similar to those used in [2], but in [2], data sets gave these algorithm for linear data and we give LDF algorithm for circular data in our paper. $Dataset1$ is generated a set of circular random variables from the von Mises distribution $CN(180^{\circ};5)$ with $n = 100$ shown in Figure 1 with curve of Taylors bandwidth [3] and Hall, Watson and Cabrera [10] proposed LSCV and MLCV curve to maximize the cross validation loglikelihood with respect to the bandwidth parameter.

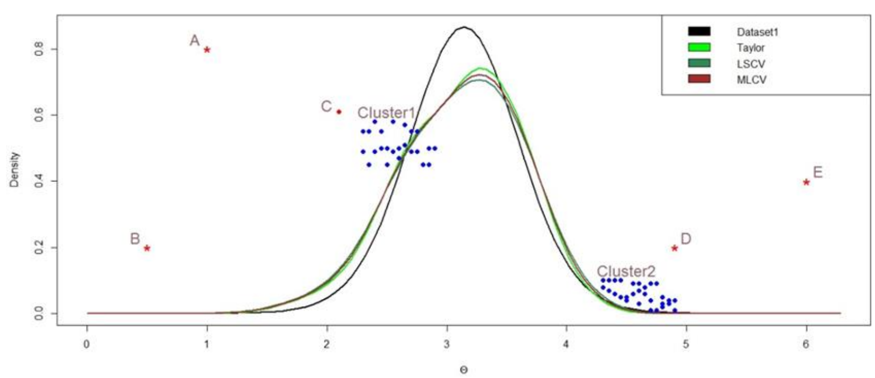

Figure 1: Simulated data sets with two clusters of Dataset1: Four outliers are marked with A, B, D and E

$Dataset1$ contains a Gaussian cluster, two clusters non-uniform density and sizes (with 24 and 29 data samples) and four clear outliers A, B, D and E. Data sample C does not belong to the $Cluster1$, but as argued in LDF is should not be regarded as an outlier, since its local density is similar to its neighbors' local densities. Although points C and D have equal distances to their closest clusters ($Cluster1$ and $Cluster2$ correspondingly), the difference in clusters density suggests that D is an outlier while C is a normal data point. As shown in Figure 2, the proposed LDF correctly identify the outliers A, B, D and E in $Dataset1$ represented as circles, these circles identify the outlier factor value when bandwidth $h = 1$ without classifying C as an outlier.

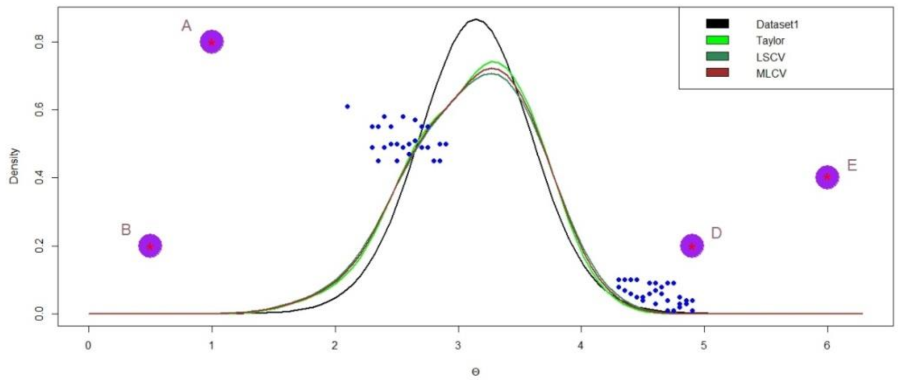

Figure 2: Results on two cluster data set in Figure 1 for k = 5 and h = 1: LDF.

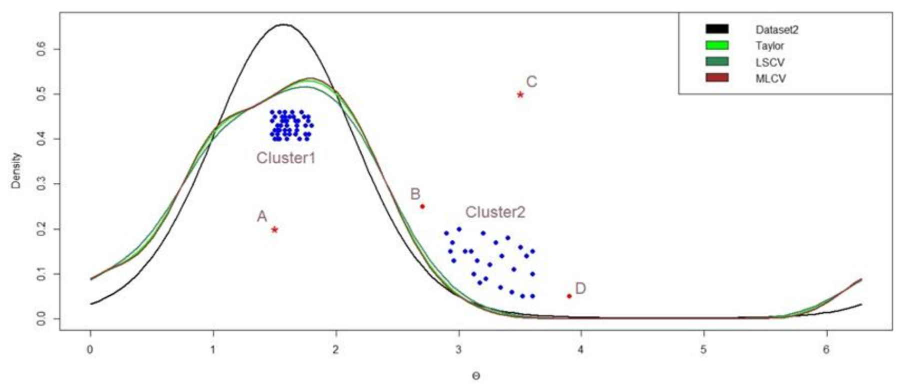

Figure 3: Simulated data sets with two clusters of Dataset2: Two outliers are marked with A and C.

We illustrate the main problem of LDF on the second data set with two clusters of different densities. $Dataset2$ is generated a set of circular random variables from the von Mises distribution $CN(90^{\circ};3)$ with $n = 125$ shown in Figure 3 with contains 44 points in the dense$Cluster1$, 25 points in sparse $Cluster2$ and two outstanding outliers $A$and C(marked with stars).

While samples $A$and$C$ are clearly outliers, we regard samples$B$ and Das outliers in analogy to sample D from $Dataset1$ (see Figure 1). Like sample$D$ in$Dataset1$, their local density is lower than the local density of their neighbors from$Cluster2$. In other words, samples$B$ and$D$ are not too far from the closet cluster to be regarded as normal data points. However, the outlier values for points$A$ and$C$ should be significantly larger than for points$B$and$D$. For$h = 5$,LDF detects two outliers$A$ and Cin Figure 4. As we can see, the proposed LDF algorithm identify samples on boundaries of clusters as outliers in different two data sets of circular data.

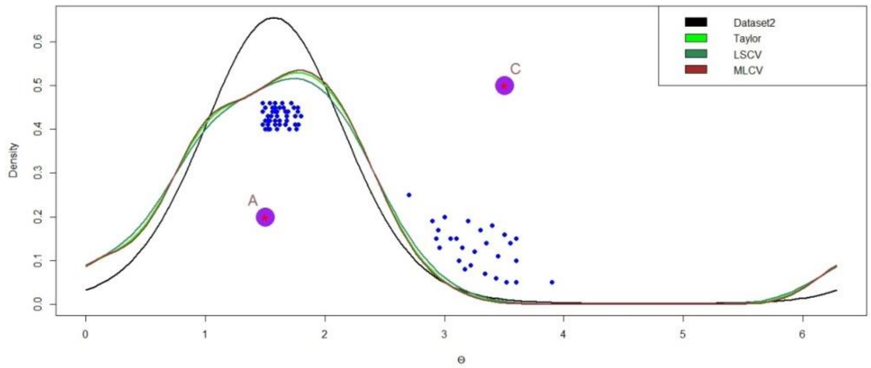

Figure 4: Results on two cluster data set in Figure 3 for k = 3 and h = 5: LDF.

5. CONCLUSION

Kernel density function is now proposed to identify outliers in the circular data. It depends on two main points: first, generalized measure of the local density factor (LDF) using a circular distance between two angles with replace this distance by reachability distance and then we created two data sets have von Mises distribution with different sample size. Every data sets have two clusters non-uniform density and sizes and the proposed method (LDF method) can be very successful for the outlier detection task.

Undoubtedly, one of the main issues in kernel estimation is the appropriate selection of the bandwidth parameter. A straightforward extension of the proposed estimator can be found in the multiple circular data, considering a multidimensional random variable. In this case, the circular part of the estimator should be properly adapted including a multidimensional kernel and possibly a bandwidth matrix.

ACKNOWLEDGMENTS

We are grateful to the Associate Editor and referees for their valuable comments on our paper who helped us improve the quality of the first draft of our paper.

REFERENCES

[1] Lima, R. (2012). Outlier detection with kernel density functions in monitoring the ISTAT LFSdata production processes, Electronic Journal of Applied Statistical Analysis,7 (1), 118-129.

[2] Latecki, L. J, Lazarevic, A. and Pokrajac, D. (2007). Outlier Detection with Kernel Density Functions, 5th International Conference, MLDM 2007, Leipzig, Germany, Proceedings. 61-75.

[3] Taylor, C. (2008). Automatic bandwidth selection for circular density estimation,Computational Statistics & Data Analysis, 52(7): 3493-3500.

[4] Terrell, G. and Scott, D. (1992). Variable kernel density estimation, The Annals of Statistic20(3): 1236-1265.

[5] Gao, J., Hu, W., Zhang, Z., Zhang, X. and Wu, O. (2011). RKOF: Robust Kernel-Based Local Outlier Detection, Springer-Verlag Berlin Heidelberg, pp. 270-283.

[6] Mahmood, E. A., Rana, S., Midi, H., and Hussin, A. G. (2017). Detection of Outliers iUnivariate Circular Data using Robust Circular Distance, Journal of Modern Applied Statistical Methods, Vol. 16, No. 2, pp. 418-438.

[7] Oliveira, M., R.M. Crujeiras and A. Rodrnguez-Casal, (2012). A plug-in rule for bandwidth selection in circular density estimation, Computational Statistics and Data Analysis, 56(2012):3898-3908.

[8] Breunig, M., Kriegel, H., Ng, R. and Sander, J. (2000). LOF: Identifying Density Based LocalOutliers, In Proceedings of the ACM International Conference on the Management Data (SIGMOD 2000), Dallas, TX, pp. 93-104.

[9] Patro, S. and Sahu, K. (2015). Normalization: A Preprocessing Stage, Department of CSE &IT, VSSUT, Burla, Odisha, India.

[10] Hall, P., Watson. G.S. and Cabrera, J. (1987). Kernel Density Estimation with Spherical Data, Biometrika, 74, 4, pp. 751-762.

[11] Taylor, C.C., Mardia, K.V., Di Marzio, M. and Panzera, A. (2012). Validating protein structure using kernel density estimates, Journal of Applied Statistics, 39 (11). 2379-2388. ISSN 0266-4763.

[12] Jammalamadaka, S. Rao and SenGupta, A. (2001). Topics in Circular Statistics, World Scientific, Singapore.

[13] Mardia, K.V. and Jupp, P.E. (1999). Directional Statistics, John Wiley, New York ,NY.

[14] Terrell, G. and Scott, D. (1992). Variable kernel density estimation, The Annals of Statistics, 20(3):1236-1265.

[15] Chaubey, Y. (2016). Smooth Kernel Estimation of a Circular Density Function: A Connection to Orthogonal Polynomials on the Unit Circle, Journal of Probability and Statistics, 2018(2):1-4.

[16] Chaubey, Y.P.; Li, J.; Sen, A. and Sen, P. K. (2012). A new smooth density estimatorfor nonnegative random variables, Journal of the Indian Statistical Association,50, 83-104.

[17] Jupp, P and Mardia, K. (1989) A Unified view of the theory of directional statistics,1975-1988, Int. Stat. Rev. 57 (3): 261-294.

[18] Di Marzio, M., Panzera, A. & Taylor, C. (2011). Density estimation on the torus, Journal of Statistical Planning and Inference. 141 (6):2156-2173.

[19] Akaike, H. (1954). An approximation to the density function, Ann. Inst. Statist. Math., Tokyo 6, 127-132.

[20] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, Ann. Math. Statist. 27:832-837.

[21] Portugues, E., Crujeiras, R. and Manteiga, W. (2013), Kernel density estimation for directional-linear data, Journal of Multivariate Analysis, 121:152-175.