.

The effect of different distance measures in detecting outliers using clustering-based algorithm for circular regression model

Nur Faraidah Muhammad Di and Siti Zanariah Satari

Citation: AIP Conference Proceedings **1842**, 030016 (2017); doi: 10.1063/1.4982854
View online: http://dx.doi.org/10.1063/1.4982854
View Table of Contents: http://aip.scitation.org/toc/apc/1842/1
Published by the American Institute of Physics

The Effect of Different Distance Measures in Detecting Outliers using Clustering-based Algorithm for Circular Regression Model

Nur Faraidah Muhammad Di[1, a)] and Siti Zanariah Satari[2, b)]

[1,2]*Faculty of Industrial Sciences & Technology,*
Universiti Malaysia Pahang, 26300 Gambang,
Pahang Darul Makmur, Malaysia.

[a)] corresponding author: nurfaraidah@gmail.com
[b)]zanariah@ump.edu.my

Abstract. Outlier detection in linear data sets has been done vigorously but only a small amount of work has been done for outlier detection in circular data. In this study, we proposed multiple outliers detection in circular regression models based on the clustering algorithm. Clustering technique basically utilizes distance measure to define distance between various data points. Here, we introduce the similarity distance based on Euclidean distance for circular model and obtain a cluster tree using the single linkage clustering algorithm. Then, a stopping rule for the cluster tree based on the mean direction and circular standard deviation of the tree height is proposed. We classify the cluster group that exceeds the stopping rule as potential outlier. Our aim is to demonstrate the effectiveness of proposed algorithms with the similarity distances in detecting the outliers. It is found that the proposed methods are performed well and applicable for circular regression model.

INTRODUCTION

Cluster analysis is defined as a technique to create groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite dissimilar. While outliers are the set of objects that are significantly deviates or dissimilar from the remainder of the data set. Clustering and outlier detection share a well-known complementary relationship. In clustering, the goal is to partition the points into dense subsets, whereas in outlier detection, the goal is to determine points which do not seem to fit naturally in these dense subsets [1].

The clustering-based outlier detection is often categorized as the classification problem where the main concern is to find both clusters and outliers [2]. As a result of the classification, the detected outliers can be removed, thus can produce more reliable clustering. Numerous study discovered that cluster-based method in outlier detection produced good results. Ott et al. [3] stated the advantages of combining clustering and outlier selection include: (i) the resulting clusters tend to be compact and semantically coherent (ii) the clusters are more robust against data perturbations and (iii) the outliers are contextualized by the clusters and more interpretable. A good clustering method should have the ability to tolerate noise and detect outliers in the data set. In fact, the clustering method also should have the algorithms that use the same functionality to cluster the data and detect outliers at the same time.

Numerous works have been done by researchers in proposing the best clustering algorithms to cluster the data. The algorithms then can be divided into several categories such as hierarchical clustering, fuzzy clustering, density-based clustering, and etc. To date, there are many clustering algorithms that compromise the outlier detection [2][3][4][5]. In this study, we proposed an outlier detection procedures based on hierarchical clustering. Hierarchical clustering can further be divided into agglomerative and divisive [6].

The 3rd ISM International Statistical Conference 2016 (ISM-III)
AIP Conf. Proc. 1842, 030016-1–030016-13; doi: 10.1063/1.4982854
Published by AIP Publishing. 978-0-7354-1512-6/$30.00

- Agglomerative algorithms – Starts with every single object in a single cluster. Then, it repeats merging the closest pair of clusters according to some similar criteria until all the data merge in one cluster.
- Divisive algorithms – Starts with all objects in one cluster of all data points and recursively split the clusters into non-overlapping clusters.

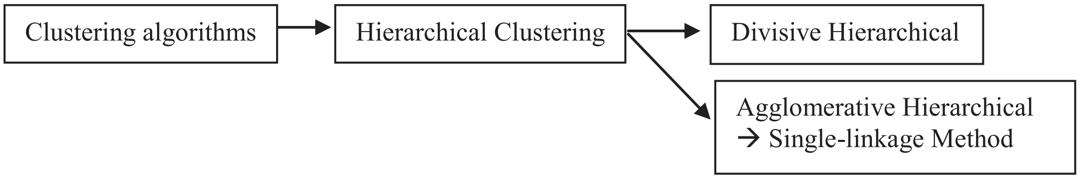


FIGURE 1: The proposed procedures using Single linkage method

Hierarchical clustering methods produce a cluster tree also known as dendogram. A dendrogram is a tree diagram that used to represent the results of a cluster analysis. Fig. 1 displays the proposed clustering algorithms used in this study. We used single-linkage agglomerative hierarchical clustering since it is the simplest clustering technique that merged observations with shortest distance. Thus, the last cluster can be indicated as the outliers and will be cut using stopping rule at certain point.

The aim of this study is to construct new clustering algorithms using Euclidean distance to detect multiple outliers in circular regression model. Then, the performance of the procedure will be compared with existing method that utilized City-block as distance measure in detecting outliers under different conditions.

DISTANCE MEASURES

In clustering, to cluster the items into their own cluster, it is necessary to have a measure of similarity or a measure of dissimilarity between the items. Similarity measure is defined as the distance between various data points. The performance of many algorithms depends on the selection of suitable good distance function. Basically, these measures are the metric function, i.e. Euclidean distance, City-block distance and Minkowski distance. In this study, we aim to use Euclidean distance and City-block distance as basis to formulate a new measure for circular data. The choice of the distance measure depends upon the application, and there is no universal solution of which measure should be used. The Euclidean distance and City-block distance are well-known methods for distance measurement, which are mostly used in clustering algorithms.

Euclidean distance is commonly used distance to compute a measure of distance. Euclidean distance is a straight-line distance between two points. It computes the root of square differences between coordinated of a pair of subjects:

$$d_{ij} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \tag{1}$$

where d_{ij} is the distance between observation i and j, d is the number of variables, x_{ik} is the value of the kth variable for the ith observation and x_{jk} is the value of the kth variable for the jth observation where i = 1,2,…,d and j = 1,2,…,d.

City-block distance is also called Manhattan distance represents distance between points in a city road grid. It computes the absolute differences between coordinated of a pair of subjects:

$$d_{ij} = \sum_{k=1}^{d} | x_{ik} - x_{jk} | \tag{2}$$

where d_{ij} is the distance between observation *i* and *j*, *d* is the number of variables, x_{ik} is the value of the *k*th variable for the *i*th observation and x_{jk} is the value of the *k*th variable for the *j*th observation where $i = 1,2,\ldots,d$ and $j = 1,2,\ldots,d$.

Euclidean distance and City-block distance are part of Minkowski metric with different values of *p*. The distance between observation *i* and *j* is given by:

$$Dis_{ij} = p\sqrt{\sum_{k=1}^{d} | x_{ik} - x_{jk} |^{p}} \tag{3}$$

Note that if *p*=1, the distance becomes the city-block distance, and when *p*=2 it becomes Euclidean distance.

PROPOSED CLUSTERING ALGORITHMS

There are five steps in the proposed procedures:

- Step 1: Obtain the predicted and residual values from Down and Mardia circular regression model.
- Step 2: Obtain the circular distance based on Euclidean distance between pairs of predicted values (*i*) and residuals (*j*) from Step 1.
- Step 3: Cluster the observation using single linkage clustering algorithm and obtain the cluster tree.
- Step 4: Cut the cluster tree.
- Step 5: Identify the cluster group with the largest size of observation as inliers (clean subset). The remaining cluster groups with minority observations are considered outliers.

Down and Mardia Circular-circular Regression Model

In 1998, Sebert et al [7] proposed a new clustering-based procedure for multiple outlier identification that utilizes the predicted and residual values obtained from a least squares fit of the data. However, this procedure is only applicable to the linear data. Then, Satari [8] extended their worked to investigate the applicability of clustering technique in circular regression models using Down and Mardia [9] circular-circular regression model.

To understand the Down and Mardia model, let *α* and *β* be an angular location parameters and *ω* is a slope parameter in the closed interval [-1,1], where *u* and *v* are fixed independent angle and the dependent random angle respectively. The DM model is given by:

$$\tan\frac{1}{2}(v-\beta) = \omega\tan\frac{1}{2}(u-\alpha) \,, \tag{4}$$

which has unique solution

$$v = \beta + 2\tan^{-1}\{\omega\tan\frac{1}{2}(u-\alpha)\}, \tag{5}$$

defines a one-to-one relationship between *u* and *v* provided *ω* is not zero. If the dependent random angle, *v* in Equation (5) is replaced by *μ*, the mean direction for *v* given *u*, then the regression curve is given by:

$$\tan\frac{1}{2}(\mu-\beta) = \omega\tan\frac{1}{2}(u-\alpha), \tag{6}$$

which has unique solution

$$\mu = \beta + 2\tan^{-1}\{\omega\tan\frac{1}{2}(u-\alpha)\}. \tag{7}$$

Assume that *v* given *u* has the von Mises distribution with mean direction *μ* and concentration parameter *κ*. Since *μ* is a function of *u* with parameters *α*, *β*, and *ω*, then,

$$v \mid u \sim M(\mu(u:\alpha,\beta,\omega),\kappa), \tag{8}$$

where,

$$\mu(u:\alpha,\beta,\omega)=\beta+v(u-\alpha:\omega), v(u-\alpha:\omega)=2\tan^{-1}\{\omega\tan\frac{1}{2}(u-\alpha)\}. \tag{9}$$

Therefore, the probability density function (pdf) and the angular error are given in the Equation (10) and (11) respectively:

$$f(v)=\frac{1}{2\pi I_0(\kappa)}\exp\{\kappa\cos(v-u)\}, \tag{10}$$

$$e=v-\mu(u:\alpha,\beta,\omega). \tag{11}$$

The angular error has the von Mises distribution with the mean direction 0 and nonnegative error concentration parameter κ.

Euclidean circular distance and City-block distance

Jammalamadaka and SenGupta [10] defined circular distance as the measure between two observation that take the smaller of the two arc lengths between the points along the circumference, i.e., for any two angles α and β,

$$d_0(\alpha,\beta)=\pi-|\pi-|\alpha-\beta||. \tag{12}$$

Based on equation above, the matrix of distance between all possible pairs of variables are calculated by insert the similarity distance (Euclidean and city-block). Satari [8] utilized the city-block distance based on circular distance. The distance is given by:

$$d_{ij}=\sum_{k=1}^{d}(\pi-|\pi-|\theta_{ik}-\theta_{jk}||), \tag{13}$$

where d_{ij} is the distance between i and j, d is the number of variables, and θ_{ik} is the value of kth variable for the ith observation where $i = 1,2,\ldots,d$ and $j = 1,2,\ldots,d$.

We proposed new distance measure for circular data based Euclidean distance as follow:

$$d_{ij}=\sqrt{\sum_{k=1}^{d}(\pi-|\pi-|\theta_{ik}-\theta_{jk}||)^2} \tag{14}$$

Single Linkage Clustering Algorithms

In single linkage clustering, the distance between two clusters is smallest between any single data point in the first cluster and any single data point in the second cluster. At each step, two clusters that contain the closest pair of observation are combined.

The algorithm is composed of the following steps:

i. Start with N clusters, each cluster containing a single multivariate observation.
ii. Calculate the matrix of distances between all possible pairs of cluster. The distances are computed using Euclidean or City-block distance that based on circular distance.
iii. Find the pair(s) of clusters with the closest members of the two clusters. The closest members are calculated by finding smallest distance in $D=\{d_{ik}\}$ and merge the corresponding objects, let say U and V, to get (UV). The distance between (UV) and the other cluster W can be computed using

$$d_{(UV)w} = \min\{d_{UW}, d_{VW}\}. \tag{15}$$

where d_{UW} and d_{VW} are distances between the nearest neighbors of clusters U and W and clusters V and W, respectively.
iv. In the distance matrix, the rows and columns responding to the merged cluster(s) are deleted. Then, a single row and column for each merged cluster from (iii) is added.
v. If more than one cluster remain, go back to step (ii).

The single linkage algorithm basically will produce clusters tree, where the branches in the tree are represent clusters.

Stopping Rules

After the cluster tree were obtained from the single linkage algorithms, the cluster tree must be portioned or "cut" at a certain height. In this study, we used stopping rule proposed by Satari [8] to cut the tree. The stopping rule is given by:

$$\bar{h} + 2.06 s_h \tag{16}$$

where $\bar{h}$ is the average heights of the cluster tree for all $N-1$ clusters, and s_h is the circular standard deviation of the heights:

$$s_h = \sqrt{-2\log \bar{R}_h} \tag{17}$$

where $\bar{R}_h$ is the mean resultant length of the height for $N-1$ clusters.

Outlier detection

Three outliers were planted in each of data set at certain point [d_1, d_2, d_3]. These multiple outliers are detected using three method; the power performance using "success" probability (*pout*), masking error (*pmask*) and swamping error (*pswamp*). The power of performance, *pout*, *pmask*, *pswamp* are carried out by simultaneous simulation using the same data set.

pout – the probability that all the planted outliers are successfully detect

$$pout = \frac{"success"}{s} \tag{18}$$

where "*success*" is number of data set that the method successfully identified all the planted outliers, and s is the total number of simulations.

pmask – the probability that the planted outliers are falsely detected as inliers

$$pmask = \frac{"\,failure\,"}{(out)(s)} \tag{19}$$

where "*failure*" is the number of outliers in all data set that detected as inliers, and out is the number of planted outliers, *out* = 3.

pswamp – the probability of clean observations detected as outliers

$$pswamp = \frac{"\,false\,"}{(n - out)s} \tag{20}$$

where "false" is the number of inliers in all data set that detected as outliers, and n is the number of sample.

Simulation Study

We generate two sample sizes, n = 30 and n = 100 for independent circular variable (u) and circular error (e) from von Mises distribution. The values of u are assumed fixed and generated from $VM(\pi/2, 2)$. Whereas, the values of e are generated from $VM(0, \kappa)$, where the concentration parameter are κ = 5, κ = 10. Then, based on the generated random samples u and e, the value of response variable (v) were calculated using Down and Mardia circular-circular regression model with fixed values of α = 1.5, β = 1.5 and ω = 0.5. We also set four contamination level of λ = 0.0, λ = 0.4, λ = 0.6, λ = 0.8, and λ = 1.0 to the simulated data. The simulation study is done using SPlus statistical package.

RESULTS AND DISCUSSION

The results are displayed separately in different table based on the method of detecting outliers which are, *pout, pmask and pswamp*. Then, in each table, the results are categorized according to different distance measures (Euclidean and City-block), number of sample (n), the value of kappa (κ), and the level of contamination (λ). There are 20 conditions investigated for each procedures.

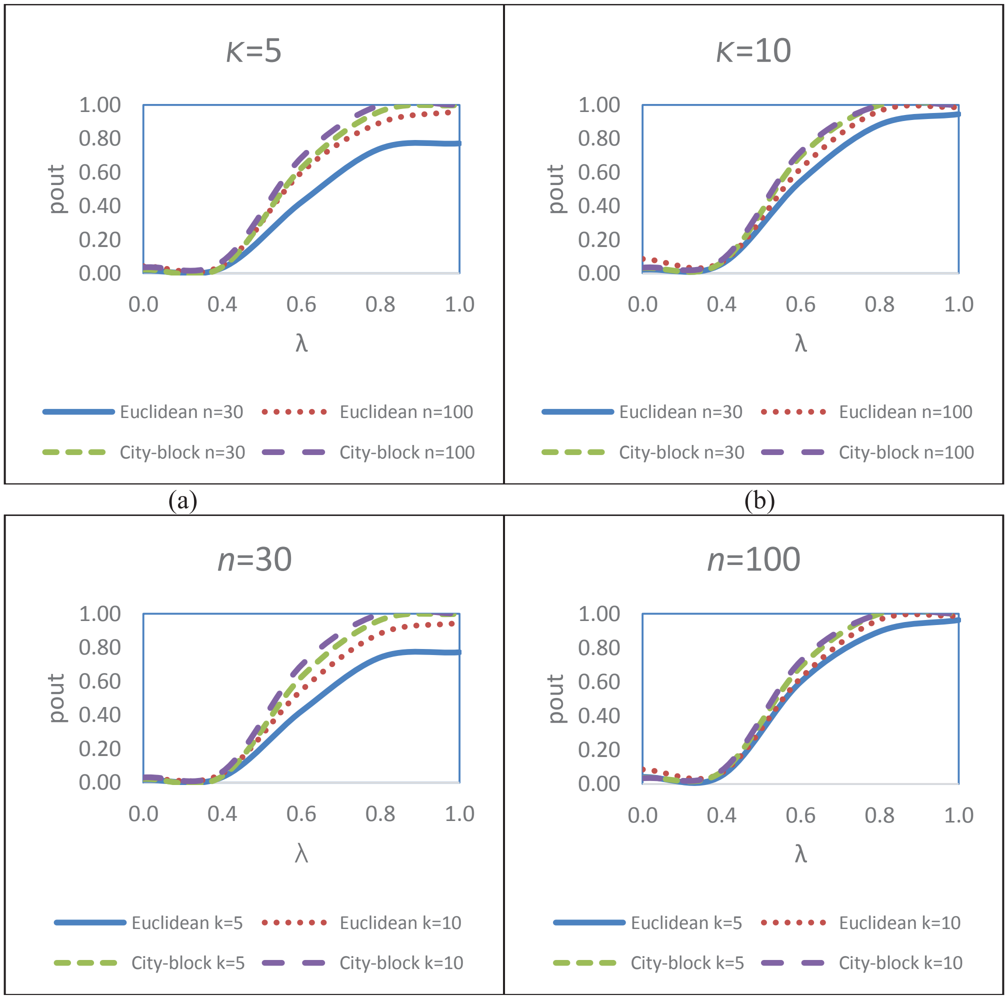
Table 1 shows the *pout* results for procedures of both distance measures. The values indicated the probability of the method successfully identified all the planted outliers. The probability of getting the highest "success" will be achieved as the value approaching 1.0. As displayed on the table, the City-block distance produced higher values compare to the Euclidean distance for almost conditions. The procedure with City-block distance show the values of *pout* approximately equal to 1.0 for λ = 0.8 and λ = 1.0. On the other hand, the Euclidean distance also managed to detect outliers especially when κ = 10.

The procedure with Euclidean distance produce similar pattern across the level of contamination. However, the overall performance for procedure with Euclidean distance is fall behind in terms of *pout* values. The pattern can also be seen between *pout* values of κ = 5 and κ = 10 for both procedures. The power of performance for κ = 10 produced higher *pout* values compare to κ = 5.

Fig. 2 (a) and (b) shows the plot of success probability (*pout*) versus level of contamination with fixed value of κ = 5 and κ = 10. From the pattern, we can say that the values of *pout* increase significantly as the level of contamination increase. The curve also shows that Euclidean distance is slower to reach high *pout* values especially when κ = 5 compare to Euclidean with κ = 10.

TABLE 1: *pout* values for proposed procedures based on Euclidean and City-block distance

λ	n	Euclidean		City-block (Satari)	
		$\kappa = 5$	$\kappa = 10$	$\kappa = 5$	$\kappa = 10$
0.00	30	0.0160	0.0220	0.0220	0.0310
	100	0.0420	0.0860	0.0360	0.0340
0.40	30	0.0330	0.0550	0.0392	0.0650
	100	0.0490	0.0670	0.0720	0.0798
0.60	30	0.4230	0.5470	0.6260	0.6960
	100	0.6000	0.6220	0.6880	0.7220
0.80	30	0.7420	0.8833	0.9630	1.0000
	100	0.8960	0.9650	1.0000	1.0000
1.00	30	0.7720	0.9460	1.0000	1.0000
	100	0.9630	0.9880	1.0000	1.0000



(a) (b)

(c) (d)

FIGURE 2. (a) Plot of success probability (*pout*) versus level of contamination with concentration parameter $\kappa = 5$. (b) Plot of success probability (*pout*) versus level of contamination with concentration parameter $\kappa = 10$. (c) Plot of success probability (*pout*) versus level of contamination with sample size, $n = 30$. (d) Plot of success probability (*pout*) versus level of contamination with sample size, $n = 100$.

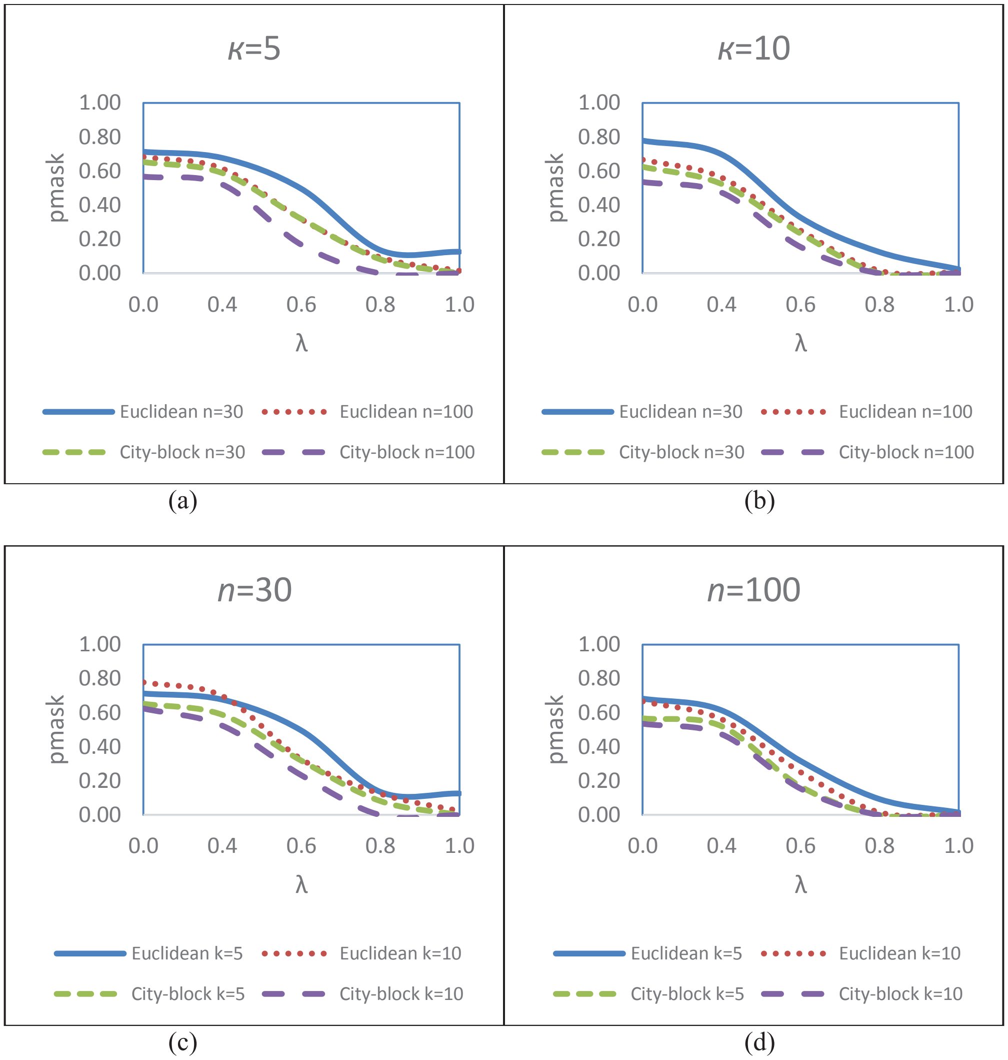
Similar pattern can be seen from Fig. 2 (c) and (d), where the plot indicate the *pout* values versus level of contamination with sample size, n = 30 and n = 100 respectively. Generally, we can see that as the sample size increase, the pout values are gradually increase. Although the City-block distance was perform better, the Euclidean distance appeared to be as good as City-block distance with higher sample sizes and higher level of contamination.

The results of masking error are illustrated in Table 2 and Fig. 3 below. Masking error is defined as the probability that the planted outliers are falsely detected as inliers. The lowest value of *pmask* (*pmask* = 0.0) indicate the zero probability that the planted outliers are falsely detected. The lowest value of *pmask* are shown in 7 conditions (refer Table 2); City-block distance with all λ = 0.8 - 1.0, κ = 5 – 10, n = 30 – 100. City block distance with λ = 0.6, , κ = 10 and n = 30 also managed to detected all the planted outliers.

TABLE 2: *pmask* values for proposed procedures based on Euclidean and City-block distance

λ	n	Euclidean		City-block (Satari)	
		$\kappa = 5$	$\kappa = 10$	$\kappa = 5$	$\kappa = 10$
0.00	30	0.7133	0.7790	0.6533	0.6253
	100	0.6833	0.6670	0.5678	0.5360
0.40	30	0.6770	0.6980	0.5882	0.5241
	100	0.6144	0.5620	0.5211	0.4738
0.60	30	0.4960	0.3277	0.3188	0.2334
	100	0.3172	0.2514	0.1676	0.1546
0.80	30	0.1358	0.1248	0.0821	0.0000
	100	0.0933	0.0157	0.0000	0.0000
1.00	30	0.1267	0.0240	0.0000	0.0000
	100	0.0153	0.0050	0.0000	0.0000

Fig. 3. (a) and (b) show the plot of masking error (*pmask*) versus level of contamination with concentration parameter κ = 5 and κ = 10. The figures show the decreasing values of *pmask* as the level of contamination increase. At the fixed value of κ, the *pmask* values are decrease gradually except for City-block distance with n = 100. This procedure is managed to avoid masking error faster as compared to Euclidean with n = 30. Likewise, at the fixed value of n, the *pmask* values are also decrease gradually for κ = 5 and κ = 10. For n = 100, both procedures produced similar pattern. However, when n = 30, the *pmask* values for Euclidean distance become constant when λ = 0.8. This is due to set of simulation data and it is expected at λ = 1.0, the *pmask* value approaching 0.00.



(a) (b)

(c) (d)

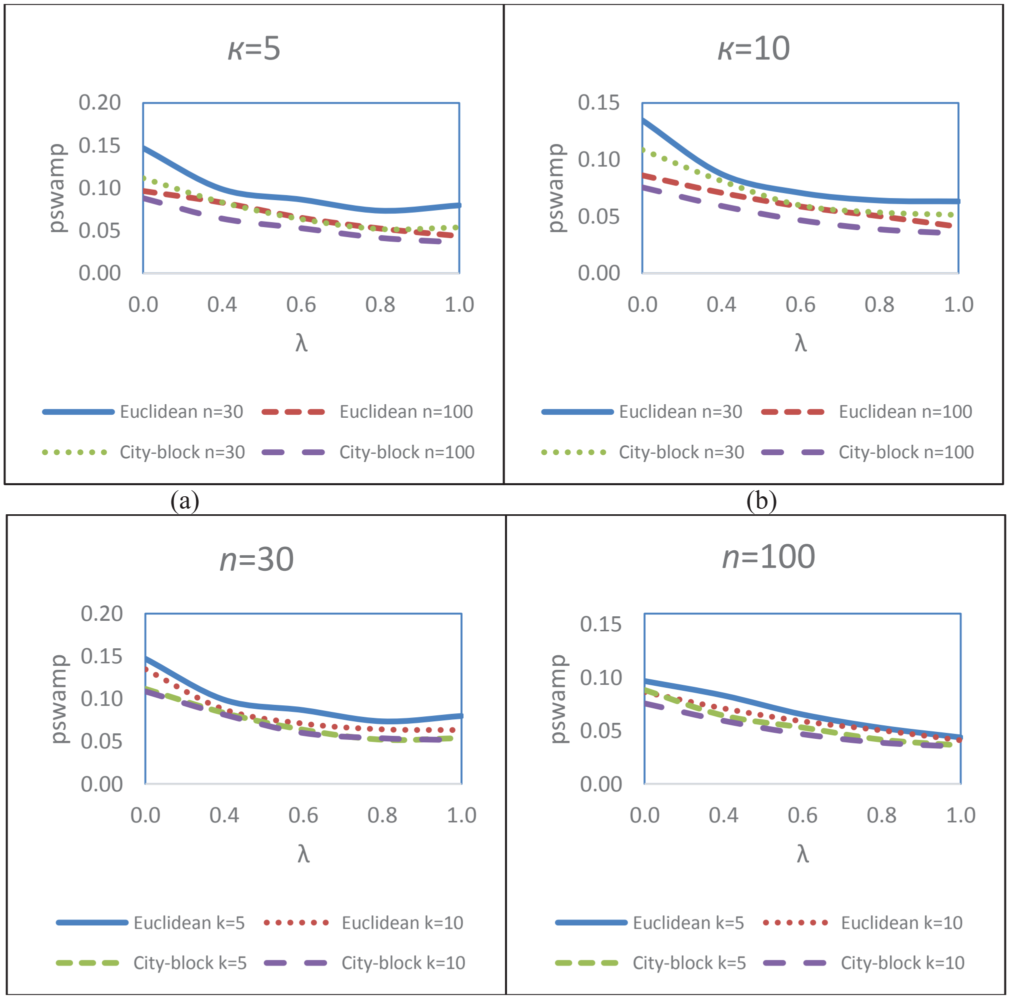
FIGURE 3. (a) Plot of masking error (*pmask*) versus level of contamination with concentration parameter κ = 5. (b) Plot of masking error (*pmask*) versus level of contamination with concentration parameter κ = 10. (c) Plot of masking error (*pmask*) versus level of contamination with sample size, n = 30. (d) Plot of masking error (*pmask*) versus level of contamination with sample size, n = 100.

To study the existence of swamping error, Table 3 displays the probability of swamping error for both proposed procedures. The *pswamp* values indicated the probability that the clean observations (inliers) are detected as outliers. The values of *pswamp* should be as low as possible. From the table, the *pswamp* values for both procedures did not touch zero for all cases suggesting the presence of swamping errors even at highest conditions (n = 100 and κ = 10). A future research need to be conducted with higher values in sample size and concentration parameter.

Fig. 4 (a) and (b) illustrates the plot of swamping error for both procedures at fixed values of κ = 5 and κ = 10. In general, the *pswamp* values decrease gradually as the level of contamination increase. Similarly, the values of *pswamp* decrease gradually at both fixed value of n (refer Fig. 4 (c) and (d)).

TABLE 3: *pswamp* values for proposed procedures based on Euclidean and City-block distance

λ	n	Euclidean		City-block (Satari)	
		$\kappa = 5$	$\kappa = 10$	$\kappa = 5$	$\kappa = 10$
0.00	30	0.1470	0.1347	0.1117	0.1087
	100	0.0967	0.0863	0.0883	0.0757
0.40	30	0.0988	0.0875	0.0834	0.0814
	100	0.0831	0.0710	0.0643	0.0594
0.60	30	0.0867	0.0709	0.0637	0.0599
	100	0.0653	0.0589	0.0531	0.0468
0.80	30	0.0736	0.0642	0.0521	0.0537
	100	0.0527	0.0504	0.0417	0.0388
1.00	30	0.0799	0.0634	0.0540	0.0515
	100	0.0438	0.0411	0.0363	0.0351



(a) (b)

(c) (d)

FIGURE 4. (a) Plot of swamping error (*pswamp*) versus level of contamination with concentration parameter $\kappa = 5$. (b) Plot of swamping error (*pswamp*) versus level of contamination with concentration parameter $\kappa = 10$. (c) Plot of swamping error (*pswamp*) versus level of contamination with sample size, $n = 30$. (d) Plot of swamping error (*pswamp*) versus level of contamination with sample size, $n = 100$.

APPLICATION USING SIMULATED DATA

To assess the performance of the proposed procedure, we test the proposed procedure using simulated data. The data were generated from random sample for the independent circular variable, *u* and circular error, *e* with sample size of *n* = 70. The data of *u* and *e* are generated from *u*~*VM* (0,2) and *e*~*VM* (0,20). We planted three outliers at point of [d_1,d_2,d_3] with contamination level of λ = 0.4, λ = 0.6 and λ = 1.0. Fig. 5 shows the outliers that have been planted to the data, where each outliers are set with different level of λ; outlier 1 with λ = 0.4, outlier 2 with λ = 0.6 and outlier 3 with λ = 1.0.

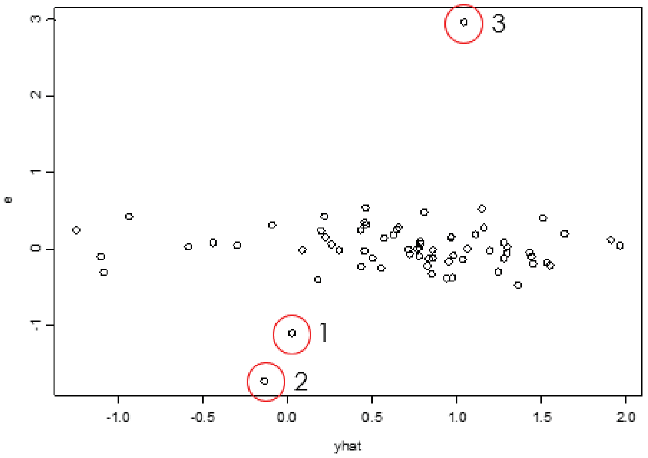
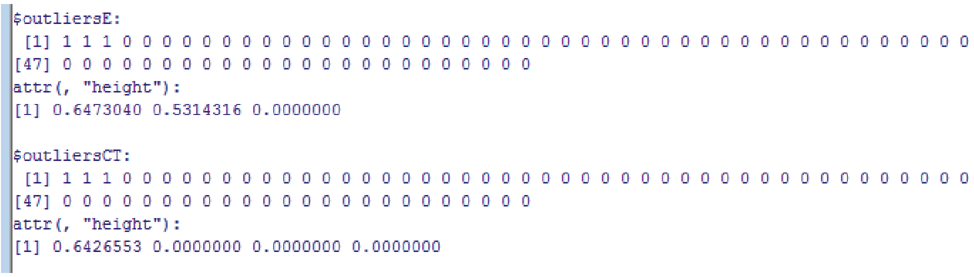


FIGURE 5. The scatter plot of predicted and residual values for the simulated data from the Down and Mardia circular-circular regression model fit.

The results of the application on simulated data can be seen from Fig. 6 and Fig. 7. The value 1 in Fig. 6 indicate the outlier, while the value 0 indicate the inlying observation. From the result, all the planted outliers are detected using both procedures with no existence of masking and swamping error.



```
$outliersE:
 [1] 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[47] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
attr(, "height"):
[1] 0.6473040 0.5314316 0.0000000

$outliersCT:
 [1] 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[47] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
attr(, "height"):
[1] 0.6426553 0.0000000 0.0000000 0.0000000
```

FIGURE 6. The result of outlier detection using proposed procedures Euclidean ($outliersE) and City-block ($outliersCT).

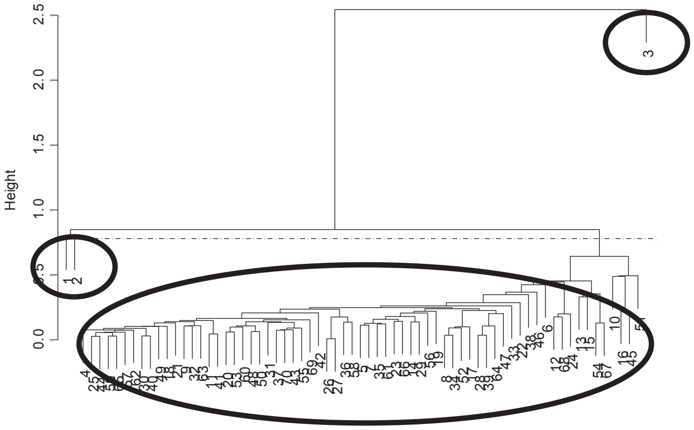


FIGURE 7. The cluster tree and corresponding cut height for simulated data set

The cluster tree and corresponding cut height is shown in Fig. 7. From the cluster tree, we can see that there are three groups formed. The largest group is considered as inlier, while the other two groups consists of observation 1, 2 and 3 are considered as outliers.

CONCLUSION

From the overall results, the procedure with City-block distance appear to outperform the procedure with Euclidean distance with the higher probability of detecting multiple outliers in circular regression model. The procedure also has the lower probability of encounter the masking and swamping error. On the other hand, the Euclidean distance is on par with City-block distance in several condition, for example in bigger sample sizes, higher level of contamination and concentration parameter. In addition, from the results, we can see that there are patterns in each of the conditions for different procedures, for example the contamination level, the concentration parameter and the sample sizes. Generally, as the level of contamination increase, the *pout* values are increase, and the masking and swamping errors are decrease. For fixed values of concentration parameter and sample sizes, the similar pattern also can be seen.

REFERENCES

1. C. C. Aggarwal and P. S. Yu, Outlier detection for high dimensional data. ACM Sigmod Record **30**(20), 2001.
2. M. H. Marghny and A. I. Taloba, Outlier Detection using Improved Genetic K-means. International Journal of Computer Applications **11**(28), 2011.
3. L. Ott, L. Pang, F. T. Ramos, and S. Chawla, "On integrated clustering and outlier detection," in *Advances in neural information processing systems*, 2014. pp. 1359-1367.
4. V. Kumar, S. Kumar and A. K. Singh, Outlier Detetction: A Clustering-Based Approach. International Journal of Science and Modern Engineering **1**(7), 2013.
5. P. Kasture and J. Gafge, Cluster based Outlier Detection. International Journal of Computer Applications **58** (10), 2012.
6. G. Guojun, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications* (Siam, Philadelphia, 2005).
7. D. M. Sebert, D. C. Montgomery and D. A. Rollier, D.A, A clustering algorithm for identifying multiple outliers in linear regression. Computational Statistics and Data Analysis **27**, 461-484 (1998).

8. S. Z. Satari, "Parameter Estimation and Outlier Detection for Some Types of Circular Model," Ph.D. thesis, University of Malaya, 2015.
9. T. D. Down, and K. V. Mardia, Circular regression. Biometrika **89**(3), 683-697 (2002).
10. S. R. Jammalamadaka, and A. Sengupta, *Topics in Circular Statistics* (World Scientific, Singapore, 2001).