Sains Malaysiana 52(5)(2023): 1595-1606
http://doi.org/10.17576/jsm-2023-5205-20

Identifying Multiple Outliers in Linear Functional Relationship Model Using a Robust Clustering Method

(Menentukan Data Terpencil Berganda bagi Model Linear Hubungan Fungsian Menggunakan Kaedah Berkelompok yang Lebih Kukuh)

Adilah Abdul Ghapor[1,*], Yong Zulina Zubairi[2], Sayed Md. Al Mamun[3], Siti Fatimah Hassan[4], Elayaraja Aruchunan[5] & Nurkhairany Amyra Mokhtar[6]

[1]*Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia*
[2]*Institute of Advanced Studies, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia*
[3]*Department of Statistics, University of Rajshahi, Bangladesh*
[4]*Centre for Foundation Studies in Science, Universiti Malaya, Kuala Lumpur, Malaysia*
[5]*Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia*
[6]*Mathematical Sciences Studies, College of Computing, Informatics and Media, Universiti Teknologi MARA, 85000 Segamat, Johor Darul Takzim, Malaysia*

Received: 12 October 2022/Accepted: 10 May 2023

ABSTRACT

Outliers are some observation points outside the usual pattern of the other observations. It is essential to detect outliers as anomalous observations can affect the inference made in the analysis. In this study, we propose an efficient clustering procedure to identify multiple outliers in the linear functional relationship model using the single linkage algorithm with the Euclidean distance as the similarity measure. A new robust cut-off point using the median and median absolute deviation for the tree heights to classify the potential outliers are proposed in this study. Experimental results from the simulation study suggest our proposed method is able to identify the presence of multiple outliers with very small probability of swamping and masking. Application in real data also shows that the proposed clustering method for this linear functional relationship model successfully detects the outliers, thus suggesting the method's practicality in real-world problems.

Keywords: Clustering; linear; measurement error; multiple outliers

ABSTRAK

Data terpencil merupakan pemerhatian data yang berada di luar corak pemerhatian data yang lain. Menentukan data terpencil adalah penting kerana pemerhatian yang luar biasa boleh mempengaruhi inferens yang dibuat ke atas analisis tersebut. Dalam kajian ini, kami mencadangkan kaedah berkelompok yang lebih kukuh untuk menentukan data terpencil berganda bagi model linear hubungan fungsian (LFRM) menggunakan satu hubungan algoritma dengan jarak Euclidean sebagai ukuran bersama. Satu nilai potongan yang kukuh dicadangkan untuk mengumpulkan data terpencil berganda dengan menggunakan median dan median sisihan mutlak bagi menentukan ketinggian pokok tersebut. Keputusan uji kaji berdasarkan simulasi menunjukkan kaedah yang dicadangkan berjaya mengesan data terpencil berganda di dalam sesebuah set data dan menunjukkan prestasi yang bagus dengan nilai 'masking' dan 'swamping' yang rendah. Aplikasi pada data sebenar juga menunjukkan kaedah berkelompok yang dicadangkan bagi model linear hubungan fungsian (LFRM) ini berjaya menentukan data terpencil, justeru, dicadangkan penggunaan kaedah ini dalam aplikasi pada data dunia yang sebenar.

Kata kunci: Berkelompok; kesilapan pengukuran; linear; terpencil berganda

Introduction

Presence of outliers is unavoidable in many fields of research, for example in the biomedical, environmental, material science, and medical field, which also includes the recent COVID-19 pandemic (Atkinson 1985; Brzezińska & Horyń 2021; Li et al. 2016; Oh & Gao 2009; O'Leary et al. 2016). Outlier in a data set can be classified as a single outlier or multiple outliers. Identifying a single outlier is quite simple from the analytical and computational side, but when there is more than one outlier, it becomes challenging as there may be masking and swamping effects. Masking happens when an outlier is unable to be detected as a true outlier, while swamping happens when 'clean' observations or inliers are falsely detected as outliers. Masking seems to be a more serious issue than swamping, but both these problems should be addressed so that appropriate analysis can be done on the data set (Sebert, Montgomery & Rollier 1998).

Many research works have been done in identifying outliers with normality assumptions, in regression modelling, where the normality assumption is used (Adnan, Mohamad & Setan 2003; Rousseeuw & Leroy 1987; Serbert et al. 1998; Toutenburg, Chatterjee & Hadi 1990). However, in real life situation, the variables cannot be exactly recorded, where in this situation, presence of errors may happen, and this is known as the Errors-in-variable-model (EIVM). There are two types of EIVM, namely the functional and the structural relationship model (Kendall 1952, 1951). Comparing the EIVM with the ordinary regression model, this EIVM is noted when the response and explanatory variables are both measured with errors, and recent work have been done on detecting outliers in EIVM as well as estimating the parameters in EIVM (Arif, Zubairi & Hussin 2022, 2020; Mokhtar et. al, 2021). Ordinary regression model on the other hand, only considers when the response variable is measured without error.

An outlier is a point or some points of observation that is outside the usual standard pattern of the observations. Outlier occurs when the data is mistakenly observed, recorded, and input into the computer system (Cateni et al. 2008). In this case, for situation when multiple outliers exist, we need to identify ways to counter this issue. Clustering technique is considered as one of the methods that is widely used to identify multiple outliers in a linear regression model (Adnan, Mohamad & Setan 2003; Loureiro et al. 2004; Serbert et al. 1998). In this paper, we will focus on the algorithm that will be able to cater for data that can be modelled by the Linear Functional Relationship Model (LFRM), where in this model, both the measurements are subjected to errors. This LFRM is important because if we ignore possible measurement error on the variables, it may lead to inconsistent estimators of the model parameters.

Earlier studies have also used clustering procedure for detecting outliers and these include clustering algorithm using the ordinary least square fit to standardize the predicted and residual values where single linkage algorithm has been used for grouping and the Euclidean distance as the similarity measure (Sebert, Montgomery & Rollier 1998). Another study proposed the least trimmed of square fit to standardize the predicted and residual values (Adnan, Mohamad & Setan 2003). However, both these techniques can only be applied in studies that are detecting outliers in only the regression model and not in the EIVM where both variables are subjected to errors. In this article, we consider the abovementioned method to detect outliers for a type of model called the linear functional relationship model (LFRM). To explain the LFRM, consider the following equation,

$$Y_i = \alpha + \beta X_i + e_i,$$

where the variables X_i and Y_i are linearly related. Parameter α is the intercept, and β is the slope parameter. For an ordinary linear regression model, the X_i variable can be observed directly. However, in reality, these two variables, X_i and Y_i cannot be observed directly as their measurements are subjected to errors. For instance, for any fixed X_i and Y_i we observe x_i and y_i from continuous linear variable subject to errors δ_i and ε_i respectively, i.e., $x_i = X_i + \delta_i$ and $y_i = Y_i + \varepsilon_i$, where the error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables,

$$\text{i.e., } \delta_i \sim N\left(0, \sigma_\delta^2\right) \text{ and } \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right).$$

In this paper, we will propose a robust method in identifying multiple outliers mainly from a Linear Functional Relationship Model, from this EIVM category. We will propose a robust clustering procedure to identify multiple outliers in the LFRM where the single linkage algorithm and the Euclidean distance will be used as the similarity measure. A new robust cut tree will be proposed using the median and the median absolute deviation (MAD) of the tree heights to classify the potential outliers.

The following section is organised as follows: First section elaborates on the materials and methods used in the clustering algorithm, while the simulation study to assess the behaviour of each level of contamination and the performance of the proposed technique is presented in the following part. The results obtained and relevant discussions are presented in the third part and the conclusions are summarized at the end.

SIMILARITY MEASURE FOR CLUSTERING ALGORITHM IN LFRM

To cluster the variables or items into their own groups, it is necessary to have a certain measurement of similarity or a measure of dissimilarity between the items. Therefore, finding the similarity measure is the first rule to cluster the items. A number of similarity measures can be found in the literature and each of them have their own strengths and drawbacks, so it is necessary to choose the best measurement that fits our model (Aldenderfer & Blashfield 1987).

The most commonly used similarity measure by using the distance measure type is the Euclidean distance, defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \ ,$$

where d_{ij} is the distance between i and j, and x_{ik} is the value of the k^{th} variable for the i^{th} observation (Wang, Zhang & Feng 2005). Euclidean distance has been applied by many researchers until today such as in the biomedical data, in clustering symptoms of COVID-19 cases, and monitoring COVID-19 infection (Ilbeigipour, Albadvi & Akhondzadeh Noughabi 2022; Kumar 2020; Ultsch & Lötsch 2022).

For this LFRM model, the Euclidean distance is used as the similarity measure. This measure is easily applied, whereby similar observations are identified by relatively small distance, while a dissimilar observation is identified by a relatively large distance.

SINGLE LINKAGE CLUSTERING ALGORITHM FOR LFRM

Next, a suitable clustering method needs to be determined. There are a few clustering techniques that can be found in the literature, and for this study, we consider the single linkage method as the calculation is mathematically straightforward and has been widely used. Single linkage algorithm uses the smallest dissimilarity between a point in the first cluster and a point in the second cluster, and also defined as using the nearest neighbour (Kaufman & Rousseeuw 1990).

The general steps for single linkage clustering algorithm in LFRM is summarised in Figure 1. We first find the smallest distance in D - $\{d_{ik}\}$ as in Step 2, and merge the corresponding objects, say U and V, to get (UV). To calculate the distances between (UV) and other clusters, W as described in Step 3 from Figure 1, we compute;

$$d_{(UV)W} = \min\left\{d_{UW}, d_{VW}\right\}$$

where d_{uw} and d_{vw} are the distances between the nearest neighbours of clusters U and W, and clusters V and W, respectively. In doing so, we obtain a distance matrix of size $n \times n$. Using the rule as in Step 4, we delete the rows and columns corresponding to the merged cluster(s). For each merged cluster, we add a single row and column. In cases when there is still more than one cluster remains, we will repeat the similar step until only one cluster is left.

A ROBUST STOPPING RULE FOR OUTLIER DETECTION IN LFRM

After the cluster is obtained from the data, the number of groups, if any, in the data set needs to be decided. The cluster tree needs to be portioned or 'cut' at a certain height. As a rule of thumb, the number of cluster groups depends upon where the tree is cut. Studies on stopping rule suggest that the difficulty is in a two clusters scenario where it is difficult to apply any feasible stopping rules (Milligan et al. 1985). Mojena's stopping rule on the other hand is widely used for linear variables (Mojena 1977). Mojena's stopping rule, or known as 'cut height' is $\bar{h} + \alpha s_h$, where $\bar{h}$ is the mean of heights for all -1 clusters, and s_h is the unbiased standard deviation of the heights which is denoted in a specified constant. The stopping rule that is based on the mean and standard deviation of the heights, however, can be easily affected in the presence of outliers, thus making the method deemed unsuitable (Hampel et al. 2011). In this paper, we will propose a new stopping rule that will be robust in the presence of outliers by using central measure of the median and the measure of spread, median absolute deviation (MAD) for the tree heights. These measures were used earlier to identify high leverage points in logistic regression model (Midi 2010).

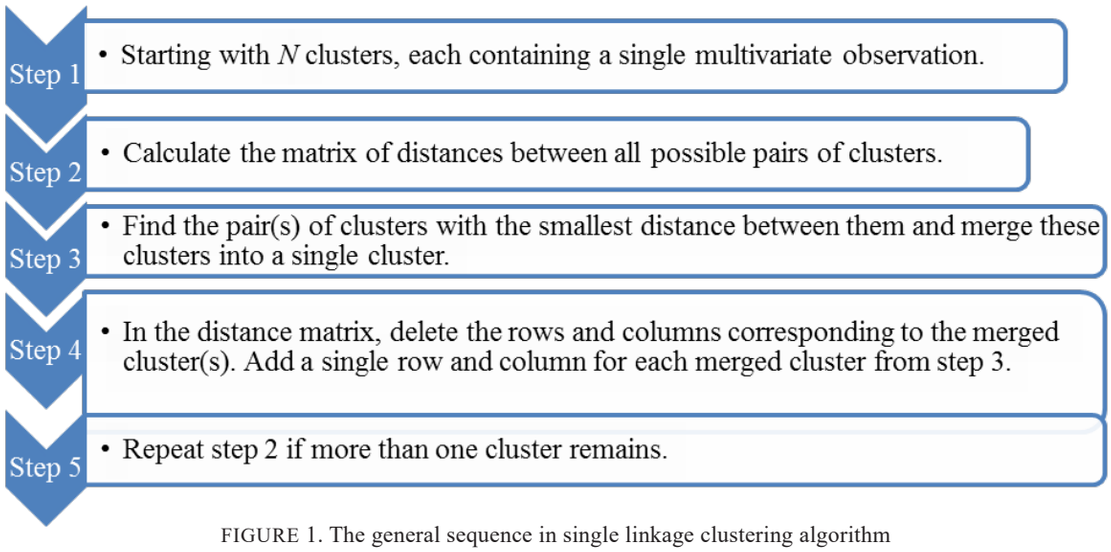

FIGURE 1. The general sequence in single linkage clustering algorithm

Thus, for this LFRM, we propose the stopping rule,

$$\tilde{h} + cMAD(h), \tag{1}$$

where h are the cluster heights; $\tilde{h}$ is the median of the heights for all $N-1$ clusters, and $MAD(h)$ is the median absolute deviation of the heights, defined by

$$MAD(h) = median\,|(h - median\ (h)|.$$

Consider $c = 3$ as suggested in the logistic regression model (Midi 2010). We may say that with 95% confidence level that the cluster groups that exceed this stopping rule will be classified as the potential outliers.

AN EFFICIENT PROCEDURE TO DETECT MULTIPLE OUTLIERS IN LFRM

Residuals are often used to measure the efficiency of a model; for instance, residuals are plotted against the corresponding predicted values to assess model adequacy. Additionally, it is also a valuable tool to identify multiple outliers where if there are no outliers present in the data, the plot of the predicted and residual values will exhibit a linear relationship (Sebert, Montgomery & Rollier 1998).

In this paper, the clustering algorithm based on the single linkage method is used to cluster the points based on the predicted values and the residual values for the LFRM. To summarize, the proposed algorithm is described in Figure 2. As described earlier, we use the Euclidean distance and single linkage method to group the observation via clustering. The cluster that exceeds the proposed robust stopping rule will be identified as the potential outliers. Generally, the cluster groups with the largest observations are considered the clean observations, and all the other observation in the small cluster are considered as outliers (He, Xu & Deng 2003).

Next, we will investigate the power of performance of the proposed clustering technique in LFRM via simulation study, and details are explained in the following section.

POWER OF PERFORMANCE FOR CLUSTERING ALGORITHM IN LINEAR FUNCTIONAL RELATIONSHIP MODEL

The power of performance of the proposed procedure is measured using the 'success' probability (*pop*), probability of masking (*pmask*), and probability of swamping (*pswamp*), respectively. Let s be the total number of simulations and *out* is the number of planted outliers in the data set. Thus, the probability of planted outliers which are correctly detected (*pop*) is

$$pop = \frac{"success"}{s},$$

where *'success'* is the number of data set that the method successfully identified all of the planted outlying observations.

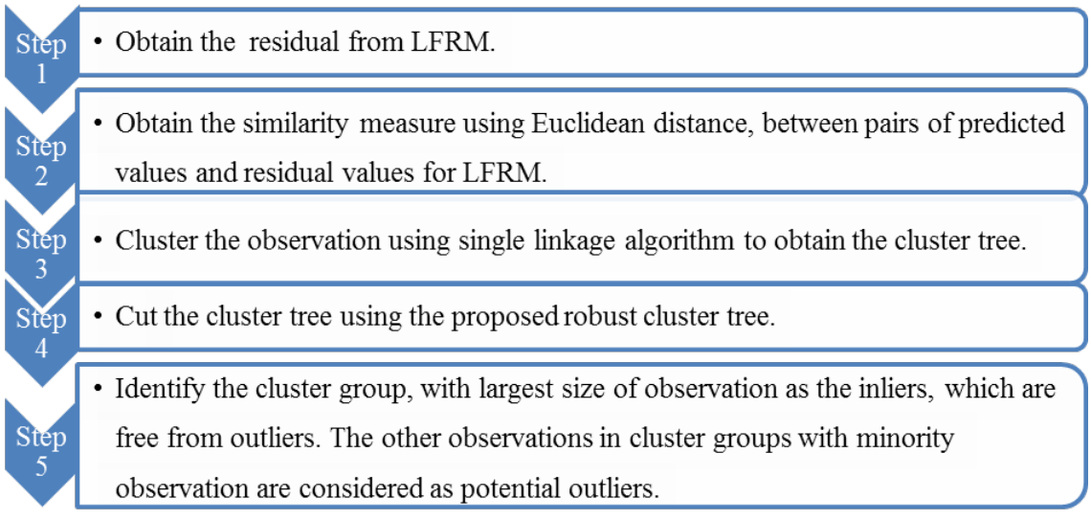

FIGURE 2. Flow chart of the steps in the proposed efficient procedure for clustering algorithm in the LFRM

As mentioned earlier, there are two main issues that needs to be highlighted in identifying the multiple outliers, namely the masking and swamping affects. The presence of masking and swamping phenomena has adverse effect in any procedure in identifying outliers. Swamping is the event of labelling normal events as anomalies or in other words, detecting clean observations as outliers. Masking on the other hand, as defined by Barnett and Lewis (1984) is 'the tendency for the presence of extreme observations not declared as outliers to mask the discordancy of more extreme observations under investigation as outliers'. This means, masking happens when outliers present in the data set are not detected, and as a result, clustering of outlying observations skews the mean and the covariance estimates, resulting in the distance measure and the outlying point from the mean is small.

The probability of masking (*pmask*) is measured by,

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

where *'failure'* is the number of outliers in the data set that is detected as inliers. The probability of swamping (*pswamp*) is measured by,

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

where *'false'* is the number of inliers in the data set that are detected as outliers.

SIMULATION STUDY

We perform a simulation study to assess how the level of contamination behaves and to obtain the power of performance for our proposed clustering technique in LFRM. We generate random sample of sizes, n = 50, 70 and 100, respectively, where the parameters are set to $\alpha = 1$, $\beta = 1$, $\sigma_\delta^2 = 0.1$, and $\lambda = 1$, respectively. The following equations for the LFRM becomes,

$$Y_i = 1 + X_i, \quad x_i = X_i + \delta_i, \text{ and } y_i = Y_i + \varepsilon_i,$$

where $X_i = 10\frac{i}{n}$, and $\delta_i, \varepsilon_i \sim N(0, 0.1)$,

where i = 1,2,…n. From the generated sample, we calculate the predicted value, $\hat{X}_i$, and the residual value, $\hat{V}_i$ from the following equations;

$$\hat{X}_i = \frac{\lambda x_i + \beta(y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2} \text{ and } \hat{V}_i = y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right), \text{ where}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + \left\{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2\right\}^{\frac{1}{2}}}{2S_{xy}}, \text{ and}$$

$$\hat{\sigma}_\delta^2 = \frac{1}{(n-2)}\left\{\sum\left(x_i - \hat{X}_i\right)^2 + \frac{1}{\lambda}\sum\left(y_i - \hat{\alpha} - \hat{\beta}\hat{X}_i\right)^2\right\},$$

where, $\bar{y} = \frac{1}{n}\sum y_i$, $\bar{x} = \frac{1}{n}\sum x_i$, $S_{xx} = \sum (x_i - \bar{x})^2$, $S_{yy} = \sum (y_i - \bar{y})^2$ and $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$.

The errors δ_i and ε_i are generated by using $\delta_i, \varepsilon_i \sim N(0, 0.1)$. In order to make some observation as outliers, we randomly contaminate the observation by replacing the mean of the contamination, $\varepsilon = 1, 2, \ldots, 10$, respectively.

For example, at point [*d*] of the response variable *y,* the observation *y*[*d*] is contaminated as

$$y^*[d] = y[d] + \varepsilon,$$

where $y^*[d]$ is the contaminated observation at position $[d]$ and ε is the degree of contamination in the range of $1 < \varepsilon < 10$. With this, it allows the outlying observation to be placed away from the inliers. In this study, for each data set, we randomly insert five outliers at certain points $[d_1, d_2, d_3, d_4, d_5]$. Then we use the clustering algorithm to identify these planted outliers for data set *n* = 50, 70 and 100, respectively. This simulation process is repeated 1000 times.

Results and Discussion

The simulation results of the power of performance for the clustering technique in LFRM with *n* = 50 are shown in Table 1. For *n* = 50, the probability of 'success' increases as the mean of contamination, ε increases. As the contamination level reaches 5, the 'success' probability shows the highest value of *probability of success*, that is equals to 1, thus suggesting a good performance. Looking at the value of *pmask,* as the level of contamination increases, the value of *pmask* decreases to a value of 0 at $\varepsilon = 5$. As for the *pswamp*, the value is almost zero. A small value of *pmask* and *pswamp* is indicative that the clustering technique is reliable and is not affected by the fundamental problem usually seen in the clustering algorithm.

As for *n* = 70 and *n* = 100, we display the results using a graphical form as given in Figures 3 and 4. Again, for both *n* = 70 and *n* = 100, the results are consistent where the 'success' probability increases as the mean of contamination, ε increases, and the value of *pmask* and *pswamp* are consistently very small as the mean of contamination, ε increases.

TABLE 1. The power of performance of the clustering method in LFRM using 'success' probability (*pop*), probability of masking (*pmask*) and probability of swamping (*pswamp*) for *n* = 50

Mean of contamination, ε	*Pop*	*Pswamp*	*pmask*
1	0.0570	0.0000	0.7366
2	0.5250	0.0000	0.2834
3	0.9510	0.0000	0.0162
4	0.9990	0.0000	0.0002
5	1.0000	0.0000	0.0000
6	1.0000	0.0000	0.0000
7	1.0000	0.0000	0.0000
8	1.0000	0.0000	0.0000
9	1.0000	0.0000	0.0000
10	1.0000	0.0000	0.0000

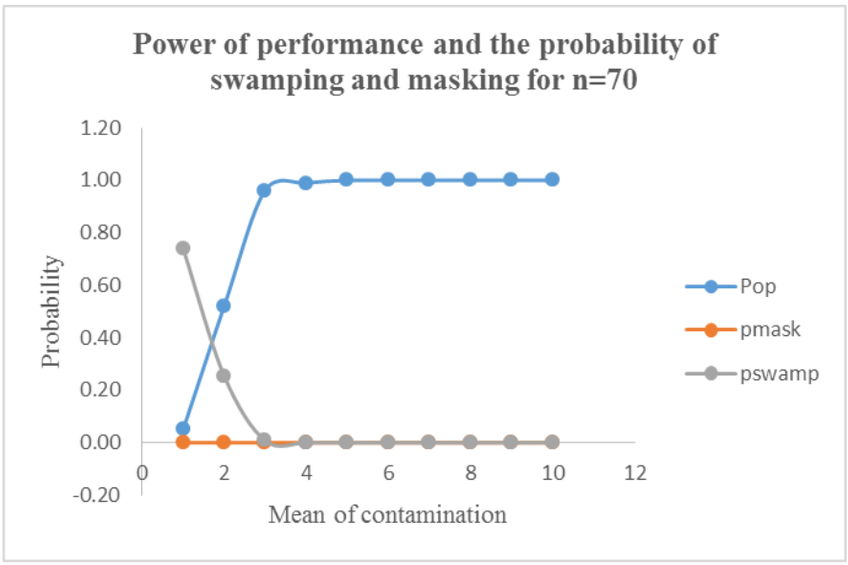

FIGURE 3. The plot of the 'success' probability (*pop*), the probability of masking (*pmask*) and also the probability of swamping (*pswamp*) for $n = 70$

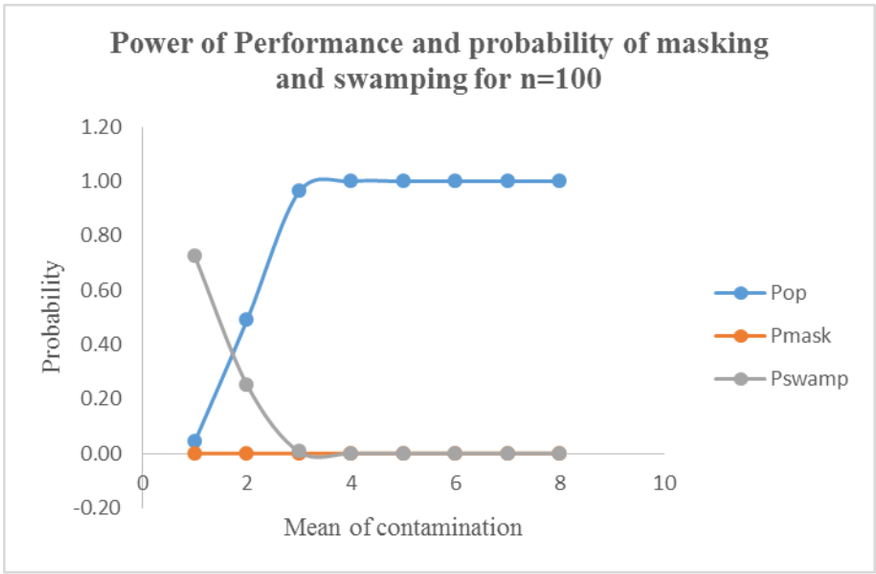

FIGURE 4. The plot of the 'success' probability (*pop*), the probability of masking (*pmask*) and also the probability of swamping (*pswamp*) for $n = 100$

From the simulation study, based on the measures of 'success' probability, probability of masking and probability of swamping, we can conclude that the proposed clustering method to identify multiple outliers in LFRM performs very well. In other words, the proposed clustering method performs the most efficient way if the outlying observation is located far from the remaining inlying observations.

APPLICATION TO REAL DATA

As an illustration, we consider two data sets to demonstrate the applicability of the proposed clustering algorithm in a LFRM, namely the Hertzsprung-Russell Stars Data and Telephone Data (Rousseeuw & Leroy 1987). The obtained data sets are often used in many multiple outlier problems in linear regression model. These data sets are usually referred to as 'classical' multiple outlier data sets. Table 2 describes the two data sets where it has been established that both have outlying observations and swamped observations.

First, we use the Hertzsprung-Russell Stars Data, where we assume measurement errors can occur at both variables and we apply the proposed clustering algorithm in LFRM. We plot the x and y variables in a scatterplot as shown in Figure 5, where x is the effective temperature at the surface of the star and y is the light intensity. From the scatterplot, there are four observations that seems to be lying away from the other observations, namely observation 11, 20, 30, and 34, respectively. In addition, observation 7 and 14 are the possible outliers. To correctly identify whether they are the outlying observations, we proceed with applying the clustering process in the LFRM and the proposed robust stopping rule to cut the tree.

Based on the proposed robust stopping rule defined in (1), the cut tree is $\widetilde{h} + 3MAD(h) = 0.4903$. From the cluster dendrogram plot as shown in Figure 6, two clusters are formed, one cluster containing the majority of the observation, and another smaller cluster containing observations 7, 11, 14, 20, 30, and 34. It can be seen that the proposed clustering technique for the data in the LFRM successfully identified outliers for observations 11, 20, 30, and 40, respectively. Observation 7 and observation 14 have been detected as the swamping observations in this study.

As another illustration, we apply the proposed method to identify outliers for the Telephone Data that can be modelled by the linear functional relationship. The scatterplot of x and y variables is shown in Figure 7, where x variable is the year from 1950 to year 1973, and y variable is the number of calls in tens of millions. From the graph, observations 15 to 24 seem to be lying away from the other observations. To confirm this, we apply the clustering process in LFRM and obtained a proposed robust stopping rule at 1.4398 as shown in Figure 8. From the cluster dendrogram plot, we observe that there are three clusters. One cluster contains the majority of the observation, and another two smaller clusters containing the outlying observation, which are observations 15 till to 24. It can be seen that the proposed clustering technique successfully identified all the outliers in the classic Telephone Data that has been modelled using LFRM.

TABLE 2. The 'classical' multiple outlier data sets

No	Data Sets	Outlying observation in the data	Outlying observations being identified	Number of observations swamped	Number of observations masked
1	Hertzsprung-Russell stars data (Rousseuw & Leroy 1987)	11,20,30 and 34	11,20,30,34,7 and 14	2	0
2	telephone data (Rousseuw & Leroy 1987)	15-24	15-24	0	0

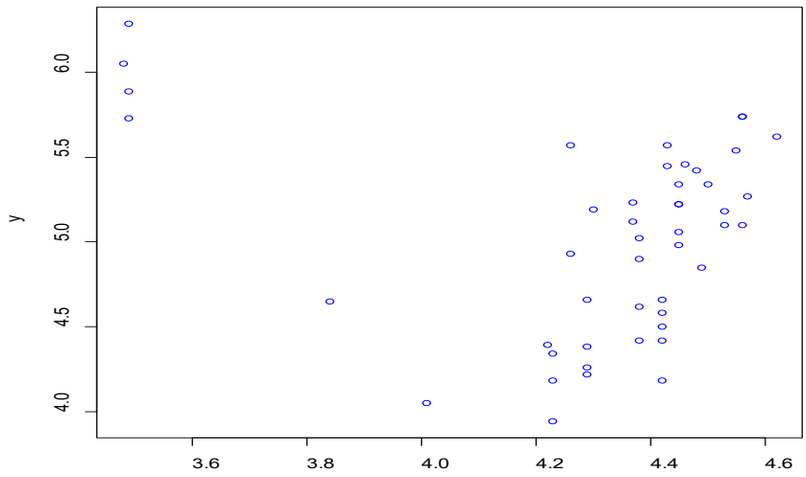

FIGURE 5. The scatterplot of Hertzsprung-Russell stars data

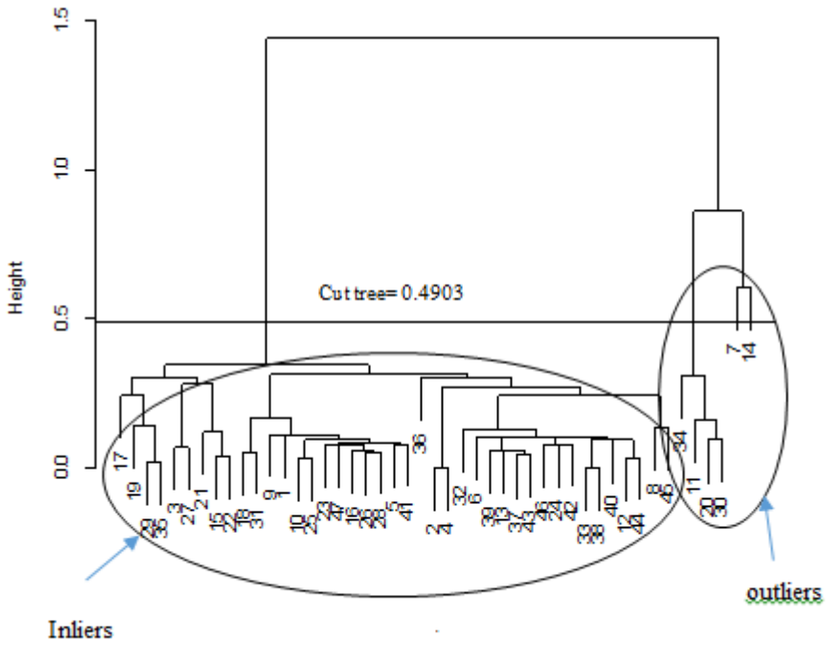

FIGURE 6. The cluster tree for Hertzsprung-Russell stars data

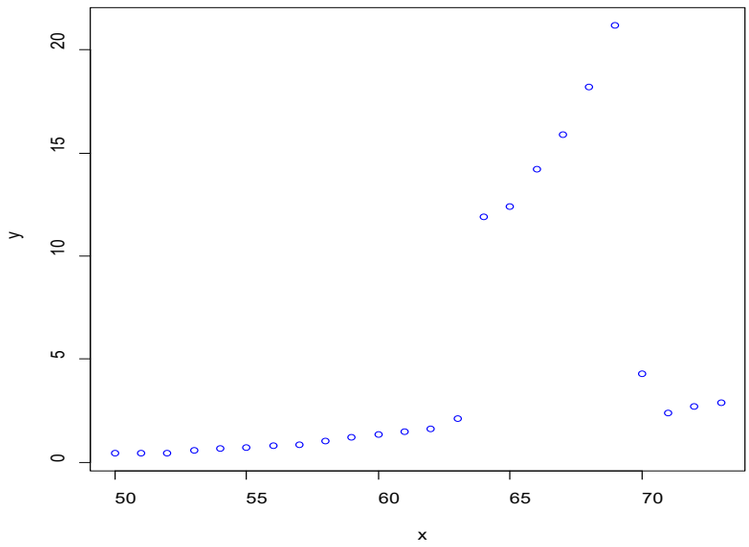

FIGURE 7. The scatterplot for telephone data

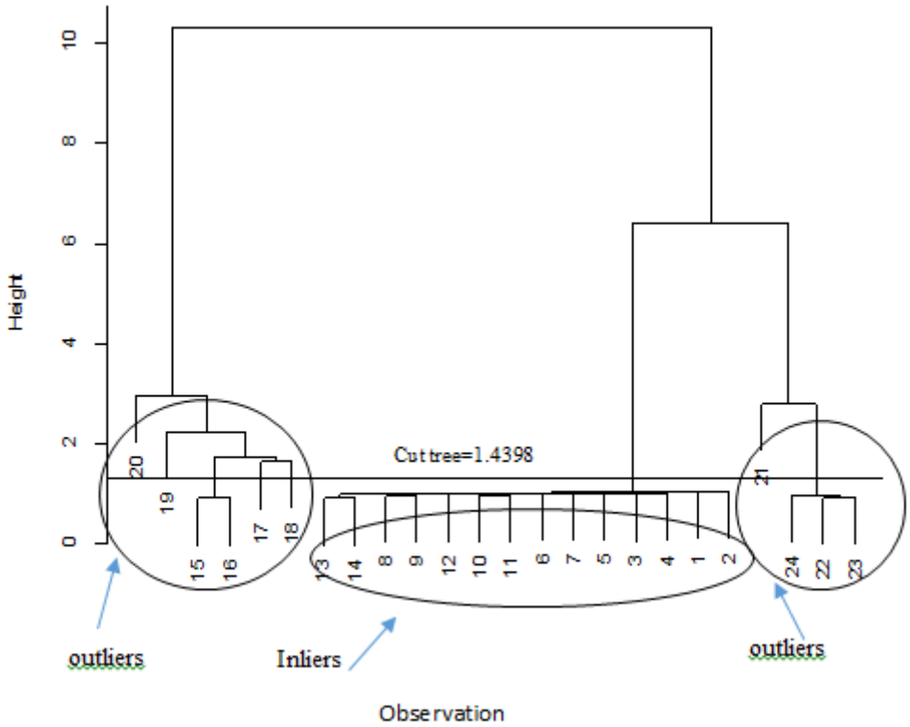

FIGURE 8. The cluster dendrogram for telephone data

Conclusions

In this study, we have proposed an efficient procedure to identify multiple outliers in the LFRM using the single linkage algorithm with the Euclidean distance as the similarity measure and a robust stopping rule based on the median and the median absolute deviation (MAD) of the tree heights. With the presence of outlier, the conventional approach by using mean and standard deviation in the cut-off rule is no longer valid as both measures may be severely affected by outliers and may produce bias measure. The novelty of our proposed method is that we use the median and MAD which are robust measures especially in situations when outliers exists. Numerical experiments using simulation study suggest that our proposed method is able to identify multiple outliers in LFRM. In other words, the probability of swamping and masking is practically small and at certain levels of contamination it is almost zero and this is good as the two main issues in multiple outlier detection are addressed. Application in real data also shows that our proposed clustering method for the LFRM successfully detects the outliers as found in other classical data.

LIMITATION AND FUTURE WORK

Some limitations in this study include the application of the proposed techniques to data that are in modelled by the LFRM. Some suggestions for future work include addressing the swamping issues and looking into the probability of swamping and masking and also addressing for a multivariate EIVM. The proposed cut-tree will be used in the application of real data sets such as clustering in business and economics as well in the future.

ACKNOWLEDGEMENTS

The authors would like to express our appreciation to the editors and reviewers for their valuable comments on this paper. This publication is partially funded by the Faculty of Business and Economics, Universiti Malaya Special Publication Fund.

REFERENCES

Atkinson, A. 1985. *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Clarendon Press.

Adnan, R., Mohamad, M.N. & Setan, H. 2003. Multiple outliers detection procedures in linear regression. *Matematika* 19: 29-45.

Aldenderfer, M.S. & Blashfield, R.K. 1984. *Cluster Analysis: Quantitative Applications in the Social Sciences*. A SAGE Publications.

Arif, A.M., Zubairi, Y.Z. & Hussin, A.G. 2022. Outlier detection in balanced replicated linear functional relationship model. *Sains Malaysiana* 51(2): 599-607. https://doi.org/10.17576/jsm-2022-5102-23.

Arif, A.M., Zubairi, Y.Z. & Hussin, A.G. 2020. Parameter estimation in replicated linear functional relationship model in the presence of outliers. *Malaysian Journal of Fundamental and Applied Sciences* 16(2): 158-160. https://doi.org/10.11113/mjfas.v16n2.1633

Barnett, V. & Lewis, T. 1984. *Outliers in Statistical Data*. 2nd ed. New York: Wiley.

Brzezińska, A.N. & Horyń, C. 2021. Outliers in COVID 19 data based on rule representation - the analysis of LOF algorithm. *Procedia Comput. Sci*. 192: 3010-3019. doi: 10.1016/j.procs.2021.09.073

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. 2011. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

He, Z., Xu, X. & Deng, S. 2003. Discovering cluster-based local outliers. *Pattern Recognition Letters* 24(9): 1641-1650.

Ilbeigipour, S., Albadvi, A. & Akhondzadeh Noughabi, E. 2022. Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making. *Informatics in Medicine Unlocked* 32: 101005. https://doi.org/10.1016/j.imu.2022.101005

Kaufman, L. & Rousseeuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc.

Kendall, M.G. 1951. Regression, structure and functional relationship, Part I. *Biometrika* 38(1/2): 11-25.

Kendall, M.G. 1952. Regression, structure and functional relationship, Part II, *Biometrika* 39(1/2): 96-108.

Kumar, S. 2020. Use of cluster analysis to monitor novel coronavirus-19 infections in Maharashtra, India. *Indian Journal of Medical Sciences* 72(2): 44-48. https://doi.org/10.25259/IJMS_68_2020

Li, Y., Jin, D.C., Bao, Z.B., Jin, H., Guo, J.W., Zhao, Y.L., Shao, J. & Yang, D. 2016. *Advances in Energy, Environment and Materials Science.* Boca Raton: CRC Press.

Mokhtar, N.A., Zubairi, Y.Z., Hussin, A.G., Badyalina, B., Ghazali, A.F., Ya'Acob, F.F., Shamala, P. & Kerk, L.C. 2021. Modelling wind direction data of Langkawi Island during Southwest monsoon in 2019 to 2020 using bivariate linear functional relationship model with von Mises distribution. *Journal of Physics: Conference Series* 1988(1): 012097. https://doi.org/10.1088/1742-6596/1988/1/012097

Milligan, G.W. & Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2): 159-179.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal* 20(4): 359-363.

Oh, J.H. & Gao, J. 2009. A kernel-based approach for detecting outliers of high-dimensional biological data. *BMC Bioinformatics* 10(4): S7.

O'Leary, B., Reiners, J.J., Xu, X. & Lemke, L.D. 2016. Identification and influence of spatio-temporal outliers in urban air quality measurements. *Science of the Total Environment* 573: 55-65.

Sebert, D.M., Montgomery, D.C. & Rollier, D.A. 1998. A clustering algorithm for identifying multiple outliers in linear regression. *Computational Statistics & Data Analysis* 27(4): 461-484.

Syaiba, B.A. & Midi, H. 2010. Robust logistic diagnostic for the identification of high leverage points in logistic regression model. *Journal of Applied Sciences* 10(23): 3042-3050.

Rousseeuw, P.J. & Leroy, A. 1987. *Robust Regression and Outlier Detection.* New York: Wiley.

Toutenburg, H., Chatterjee, S. & Hadi, A.S. 1990. Sensitivity analysis in linear regression. *Statistical Papers* 31: 232.

Ultsch, A. & Lötsch, J. 2022. Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans). *BMC Bioinformatics* 23: 233.

Wang, L., Zhang, Y. & Feng, J. 2005. On the Euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8): 1334-1339.

*Corresponding author; email: adilahghapor@gmail.com