İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION
Vol. 14, No. 2, December 2022, pp. 87–96
ISSN 1300-4077 | 22 | 2 | 87 | 96

İSTATİSTİK

A NEW COMPUTATIONAL APPROACH BASED ON DENSITY CLUSTERING FOR OUTLIER PROBLEMS IN LINEAR MODELS

Fatma Yerlikaya-Özkurt*
Department of Industrial Engineering,
Atılım University,
06830, Ankara, Turkey

Abstract: Recently, collection of huge amount of data and analysis of that much data have vital importance for human activities in many different application areas. Advanced statistical methods play crucial role for modeling of such data when the data contains outliers. Although there are number of outlier detection methods for revealing outlier observations in data, most of them may not be reasonable and appropriate for prediction purposes due to structural and requirements of modeling. In this study, density based clustering algorithm named Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is considered in order to detect the location of outlier observations effectively with respect to form of the model for given data set. Based on obtained results, the Mean Shift Outlier Model (MSOM) is constructed as a robust linear model. This newly proposed computational approach based on DBSCAN uses power of data clustering and also minimize the impact of the outlier observations by MSOM. The numerical examples are also presented to reveal the performance of the proposed approach in this study.

Key words: Outlier problem, Mean shift outlier model, Density based clustering.

1. Introduction

Recently, collection of huge amount of data and analysis of that much data have vital importance for human activities in many different application areas. Most of the statistical applications involve regression models for doing estimation and prediction. Among regression models, Linear Regression Model (LRM) is one of the most used ones by many researchers who prefer well-established form, ease of application and interpretability of the model [11]. Generally, LRM is used to investigate the relationship between a response (dependent) variable and explanatory (predictor or independent) variable(s) through estimation parameter(s). Moreover, parameter estimation of LRM is mainly based on a least squares method which can be seriously hindered by the presence of outlier observation(s) [15, 16].

Outliers occur because of changes in system behavior, human or machine error, or natural deviations in observations. In fact, these observations reduce and affect the information that we may get from the source. For this reason, it is very important to identify the existing outlier observations in given dataset [2]. Although there are number of outlier detection methods for revealing outlier observations in the dataset, most of them may not be reasonable and appropriate for prediction or estimation [9]. Thus, advanced methods play crucial role for outlier identification.

In this study, outlier observations are considered as the data points that distorting model and reducing model performance. For the detection of such outlier observations, existing statistical methods can be categorized into two which are traditional and advanced approaches. The first approach, generally, provides good result for small or relatively medium size dataset, but they fail when the dataset is high dimensional. The second approach, on the other hand, gives good performance with very low computational time on any size of dataset but especially it provides

* Corresponding author. E-mail address:fatma.yerlikaya@atilim.edu.tr

very good results on high-dimensional datasets. Therefore, advanced methods play crucial role for outlier identification [3, 6, 12, 21, 22, 23]. In this study, the new approach is proposed for outlier identification with advanced data mining tool named clustering.

There are different types of clustering algorithms such as hierarchical clustering, partitioning clustering and density based clustering. Among them density based clustering is appropriate to find outliers since it captures the data structure well with respect to regional density [8]. The most popular density based clustering algorithm named Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is preferred in this study for outlier identification [7, 24]. Based on obtained results via DBSCAN, these observations are modeled with Mean Shift Outlier Model (MSOM) which is a robust linear model.

This paper is organized as follows: Section 2 briefly reviews linear models and then MSOM is presented in detail. Section 3 presents some outlier identification methods. A background on clustering and a new outlier detection approach based on DBSCAN algorithm are also provided in the same section. Section 4 includes applications and comparisons of the new approach against existing alternative in order to illustrate the efficacy of the proposed approach. Section 5 summarizes and concludes the paper.

2. Improvements on linear model with mean shift outlier model

There are various way of modeling to handle outlier observation(s) within a dataset of interest using Linear Models (LMs). The general form of the LRM with n observations and p independent variables is given by:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{2.1}$$

where Y is the response variable and X_j $(j = 1, 2, ..., p)$ are the predictor variables. The vector of predictors is represented by $\boldsymbol{X} = (X_1, X_2, ..., X_p)^T$. The coefficient (unknown parameter) β_0 is the intercept and the rest of the unknown parameters β_j are the regression coefficients of the independent variables X_j $(j = 1, 2, ..., p)$, and ϵ is the random error term which is generally called noise [16]. If the response values (y_i) and predictor vectors $(\boldsymbol{x}_i$ $(i = 1, 2, \ldots, n))$ are inserted into the model in Eq. (2.1), the following linear system will be obtained:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.2}$$

Here, $\boldsymbol{y}$ is an $(n \times 1)$-vector of the response variable, $\boldsymbol{x}_i$ $(i = 1, 2, ..., n)$ is a $(1 \times (p+1))$ row vectors of $\boldsymbol{X}$ matrix which is a *full rank* $(n \times (p+1))$-matrix of predictor variables and $\boldsymbol{\beta}$ is a $((p+1) \times 1)$-vector of coefficients. Moreover, $\boldsymbol{\epsilon}$ is an $(n \times 1)$-vector of independently, identically distributed random errors. The corresponding mean and standard deviation are given by $E(\boldsymbol{\epsilon} \mid \boldsymbol{X}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}$. Here, σ is an unknown parameter and $\boldsymbol{I}$ is the n dimensional identity matrix. Based on the least squares estimates, $\boldsymbol{\beta}$ and σ are given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$ and $\sigma = \sqrt{\boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{H})\, \boldsymbol{y} / (n - p - 1)}$, where $\boldsymbol{H} := \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$ is called *hat operator* [17]. In order to handle outlier observation(s) for LMs, there are two main approaches called *Direct Approaches (DA)* and *Indirect Approaches (IA)*. These approaches are based on residuals from the robust regression. In general, the robust regression provides more stable results than LRM in the presence of outliers. There are three different types of outlier problems: Problems with outliers occurred in the vertical direction, problems with outliers occurred in the horizontal direction, and problems with outliers occurred at leverage points [1, 9, 18]. Figure 1 shows simple demonstration of the outliers in vertical direction (×), horizontal direction (+) and at leverage point (•). The mostly used robust regression methods to deal with outlier observation(s) in a dataset are M estimation [13], Least Trimmed Square estimation [18] and MSOM [5, 14]. In this study, the MSOM is employed in order to model the dataset consists of outliers which is describe next.

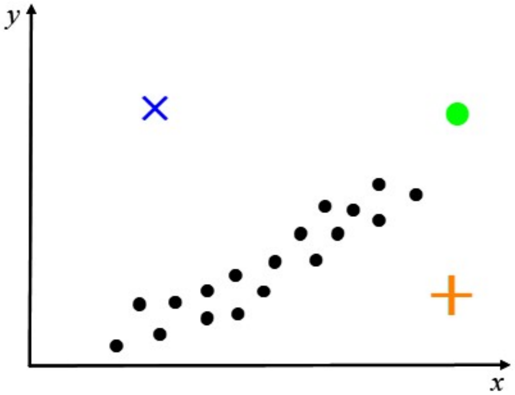

FIGURE 1. Graphical representation of outliers occurred in vertical direction, horizontal direction and at leverage point

2.1. Mean shift outlier model

The general form of the MSOM is given by:

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + \Theta\theta + \epsilon,$$

where $\Theta \in \{0,1\}$ is a constant term, and θ is the coefficient for outlier observation. In the absence of an outlier, $\Theta = 0$, and the contribution of an outlier is represented by the value θ. The linear system takes the following form after inserting all data values to the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_i\theta + \boldsymbol{\epsilon},$$

where $\boldsymbol{e}_i$ is the ith unit vector, i.e., $\boldsymbol{e}_i = (0,...,1,0,...,0)^T$ $(i = 1,2,...,n)$. In this linear system, it is assumed that either y_i or $\boldsymbol{x}_i\boldsymbol{\beta}$ deviates systematically from the model $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ by some value θ. Then, the ith observation $(y_i, \boldsymbol{x}_i\boldsymbol{\beta})$ would have a different intercept than the remaining observation, and $(y_i, \boldsymbol{x}_i\boldsymbol{\beta})$ would hence be an outlier [5, 14].

After detecting the m outliers $(m < n)$ in the dataset, the MSOM can be written as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ have same descriptions as in Eq. (2.2). On the other hand, $\boldsymbol{E}$ is an $(n \times m)$-matrix with m indicator variables, and $\boldsymbol{\theta}$ is an $(m \times 1)$-vector of the coefficients of the indicator variables. More compact form of the MSOM is rewritten as:

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\boldsymbol{X}^* = (\boldsymbol{X} \mid \boldsymbol{E})$ is an $(n \times (p+1+m))$ block matrix constructed by the matrices $\boldsymbol{X}$ and $\boldsymbol{E}$, and $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T\right)^T$ is an $((p+1+m) \times 1)$-vector constructed by the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

It should be noted that MSOM (as presented in Eq. (2.3)) gives the same residual sum of squares as the model fitted after omitting the outlier observations [20]. Therefore, MSOM is particularly convenient and preferred instead of the linear regression model in the presence of outliers.

3. Outlier identification methods

Identification of outliers is the key step before modeling with MSOM. In order to build MSOM with having good predictions, outliers should be carefully analyzed. Otherwise, the prediction model may give misleading results. Although there are various outlier detection methods, most of these methods are useless when modeling is taken into account. In this study, model based outlier identification methods are focused on. For this purposes, firstly traditional then advanced methods are introduced.

For a given dataset with n observations, the m outliers $(m < n)$ can be detected by direct approaches such as Likelihood-Ratio Test Statistic, Cooks Distance or Studentized Residuals which are described below [16].
Likelihood-Ratio Test Statistic (F_i):

$$F_i = \frac{(RSS_1 - RSS_2)/1}{RSS_2/(n-p-1)}.$$

Here, RSS_1 is the residual sum of squares obtained by using all the n observations in the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and RSS_2 is the residual sum of squares in the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_i\theta + \boldsymbol{\epsilon}$
Cooks Distance (CD_i):

$$CD_{-i} = \frac{\left(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{-i}\right)^T \left(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{-i}\right)}{p\hat{\sigma}^2},$$

where $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_{-i}$ represent the response vector and the estimated response vector after omission of the ith observation, respectively. And $\hat{\sigma}^2$ is obtained by sum of square error divided by $(n-p)$ [20].
Studentized Residuals (r_i):

$$r_i = \frac{\hat{\epsilon}_i}{\sigma_i\sqrt{(1-h_{ii})}} \qquad (i = 1, 2, ..., n),$$

where σ_i is the standard deviation of the ith residual and $\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}$, $\hat{\epsilon}_i = \boldsymbol{e}_i^T\hat{\boldsymbol{\epsilon}}$, and $\boldsymbol{e}_i^T\boldsymbol{H}\,\boldsymbol{e}_i = h_{ii}$.

An observation is defined as an outlier if it has larger Cook's distance and Studentized residual values. In order to find all potential outliers, the following steps are applied to a given dataset and repeated until all of the outlier observations are defined.

1. The LRM is constructed to fit the data.
2. The fitted values and ordinary residuals are obtained to check the better prediction.
3. The direct approaches described above are calculated to extract potential outliers.
4. The potential outlier is removed from the dataset, and the first three steps are repeated until detecting all potential outliers.

These steps are computationally slow for analyzing and detecting outlier(s) in a dataset, especially, in case of large scale dataset. These steps also contain a high error rate. Thus, it is important to have an accurate, reliable, and fast computational method for the identification of the outliers. At this stage advanced data mining tools, especially, clustering techniques play crucial role.

3.1. A background on clustering

Clustering is a data mining technique to identify the group of unlabeled data points of a data set that are similar and dissimilar to each other. Clusters are formed by assigning most similar objects (data points, entities) to the same group and dissimilar ones to the separate groups as much as possible [8].

There are different types of clustering such as hierarchical clustering, partitioning clustering and density based clustering and each preferred for different purposes. Hierarchical clustering aims to construct clusters that have an ordering from bottom to top like a tree structure. As a result it produce the hierarchical relation between the created clusters. There are two kinds of hierarchical clustering named divisive and agglomerative. Divisive clustering is splitting the single all inclusive cluster into two until having only clusters with one data point. Whereas agglomerative clustering (bottom-up approaches) is starting from the single data point as an individual cluster and merging clusters at each iteration until getting a single all inclusive cluster [24].

Partitioning clustering is based on clustering of n unlabeled data points to k clusters in which each cluster contains at least one data points. The purpose of the partitioning clustering is to minimize the distances of data points in a cluster whereas to maximize the distances between the separated clusters. In partitioning clustering, after defining number of clusters (k), the next step is to assign k random initial centers. The cluster centers are updated based on the data points assigned to a given cluster. This procedure repeatedly continues until the assigned cluster points of a sample can not be updated [24].

On the other hand, density based clustering is a clustering method that identify the arbitrarily shaped clusters in data according to the idea of a cluster is being a region with high density and separated from the other such clusters by regions of low density. Although, this type of clustering algorithms have high complexity, they can easily identify outliers in the data set. Moreover, they can handle noise and can detect the clusters automatically since they can scan the data well [8]. The most popular density-based clustering algorithm that proposed in this study to identify outliers is Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

3.1.1. DBSCAN to identify outlier

DBSCAN is a density-based clustering algorithm based on the density of the data points or closeness of the data points [7]. The points outside the dense regions are extracted and treated as outliers. This property of the DBSCAN algorithm makes it a powerful method for outlier detection. The other clustering algorithms such as k-means clustering lack this property and are very sensitive to outliers since existence of outliers can easily influence the construction of the clusters [6, 12].

DBSCAN starts with the estimation of density over a dataset with n observations. It estimates the density around each observation using epsilon neighborhood concept (eps). DBSCAN depends on the eps and a threshold value (MinPts) to detect dense regions into dataset and to classify the observation as a core, a border, or an outlier. Illustration of the concept of DBSCAN algorithm is given in Figure 2 [10].

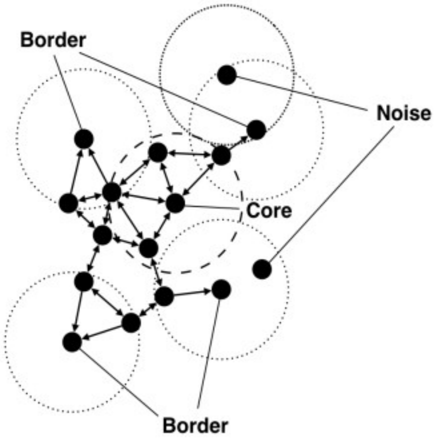

FIGURE 2. Illustration of the concept of DBSCAN algorithm given by Hahsler et. al. (2019)

The DBSCAN constructs all clusters by defining all core points which have high density and expanding each cluster to all reachable points by retrieving their epsilon neighborhood. The search continues until no more core points are found in the expanded neighborhood. End of the search, the cluster is constructed and the observations that are outside of the cluster are assigned as outliers [7, 10].

4. Applications and results

In this section, in order to apply the proposed approach, firstly, the datasets used in this study are introduced. Then, the well-known prediction performance measures are provided in the following subsections. Finally, the details of the applications and results and are presented in the last subsection. It should be noted that all computational parts of the DBSCAN, LRM and MSOM are conducted through R programming. Specifically, the R packages "dbscan" and "devtools" are installed for running DBSCAN algorithm while detecting outlier observations [10].

4.1. Data sets

4.1.1. Real world data set

The first dataset, a stack loss data, is selected from SAS Customer Support [19]. It is well-known and -studied for outlier analysis in LM. This dataset is about the operation of a plant for the oxidation of ammonia to nitric acid. It contains $n=21$ observations, $p=3$ explanatory variables which are the rate of operation (X_1), the cooling water inlet temperature (X_2), and the acid concentration (X_3). The response variable (Y) is the stack-loss. All variables' observations of this dataset are shown below:

X_1: 80 80 75 62 62 62 62 62 58 58 58 58 58 58 50 50 50 50 50 56 70

X_2: 27 27 25 24 22 23 24 24 23 18 18 17 18 19 18 18 19 19 20 20 20

X_3: 89 88 90 87 87 87 93 93 87 89 89 88 82 93 89 86 72 79 80 82 91

Y : 42 37 37 28 18 18 19 20 15 14 14 13 11 12 8 7 8 8 9 15 15

4.1.2. Simulation data set

For the simulation dataset, the data generation is based on the LM given in Eq. (2.1). The matrix of predictors is obtained from a multivariate normal distribution with zero mean and one constant $(N(0, 1))$. The random error vector is obtained from again normal distribution with $N(0, 1)$. For one dimensional vector of unknown parameter, the randomly generated value between zero and ten are preferred. In order to demonstrate the outlier identification ability of the proposed approach relatively large size dataset is chosen $(n=1000)$. After finalizing data generation, randomly 40 observations are defined and shifted with some values in order to convert them as outlier observations.

4.2. Prediction performance measures

The performance measures with their formulas used in this study are given as follows [15]:

Residual Sum of Squares (RSS)=: $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$,

Mean Squared Error (MSE)=: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$,

Root Mean Square Error (RMSE)=: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$,

Multiple Coefficient of Determination (R^2)=: $1-\left(\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)$,

Adjusted R^2 (AdjR^2)=: $1-\left(\frac{(1-R^2)(1-n)}{n-p-1}\right)$,

Correlation Coefficient (r)=: $\sqrt{R^2}$, where y_i is ith observed response value, $\hat{y}_i$ is ith fitted response value, $\bar{y}$ represents mean response value.

4.3. Results and findings

For the first dataset, after applying carefully the outlier detection steps given in Section 3, the observations 1, 2, 3, 4 and 21 are defined as outliers. It take quite long time to identify these outliers since these steps require high human intervention and they are not conducted automatically.

The linear model and related performance results obtained in the presence of all observations are given below and in Table 1, respectively.

$$Y_{LM} = -42.1062 + 0.7124\, X_1 + 1.2625\, X_2 - 0.1159\, X_3.$$

The performance results obtained after each potential observation (observations 1, 2, 3, 4 and 21) is removed from the data one by one, are presented in Table 1. In addition, the performance results obtained after removing all potential observations from the dataset are presented in the same table.

Table 1. Performance results of LM and MSOM & performance results obtained after removing each or all of the potential outlier observation(s) from the dataset

Measures	**MSE**	**RMSE**	**R^2**	**AdjR^2**	**r**
LM	8.7338	2.9553	0.9114	0.8957	0.9547
Outlier(1)	8.3682	2.8928	0.9401	0.8837	0.8620
Outlier(2)	8.9394	2.9899	0.9450	0.8930	0.8730
Outlier(3)	7.9168	2.8137	0.9514	0.9052	0.8875
Outlier(4)	7.2903	2.7001	0.9620	0.9254	0.9114
Outlier(21)	5.3198	2.3065	0.9739	0.9484	0.9387
All Outliers	1.0059	1.0029	0.97050	0.9419	0.9274
MSOM	0.7664	0.8754	0.9922	0.9870	0.9961

However, if DBSCAN algorithm is applied to given data set with eps=8 and MinPts=4 parameters, exactly same outlier observations are detected in less than a minute. The graphical representation of the clusters and outliers is given in Figure 3. In this figure, the outliers are demonstrated by black points.

After detecting all outliers for the given dataset, MSOM is built and same performance measures are calculated and given in Table 1.

$$\begin{aligned} Y_{MSOM} &= -36.6978 + 0.6658\, X_1 + 0.5673\, X_2 - 0.0103\, X_3 \\ &\quad +9.2070\, O_1 + 4.2172\, O_2 + 8.6601\, O_3 + 9.9131\, O_4 - 7.1848 O_{21}, \end{aligned}$$

where O_i for $i = 1, 2, 3, 4, 21$ represents the ith outlier observation. In addition, if the coefficients of the outlier observations are compared against the coefficients of independent variables, the contributions of outlier variables to the model are much more than independent variables. If the performance results of LM and MSOM are compared according to all performance measures, it is obvious that MSOM based on DBSCAN is much more better than LM and the LM that obtained even after removing all outlier observations.

Moreover, another application of the proposed approach is conducted by using simulated dataset which contains 40 outlier observations that randomly constructed. The DBSCAN algorithm is applied to this dataset with eps=0.3 and MinPts=10 parameters. All outliers are correctly defined and represented in Figure 4.

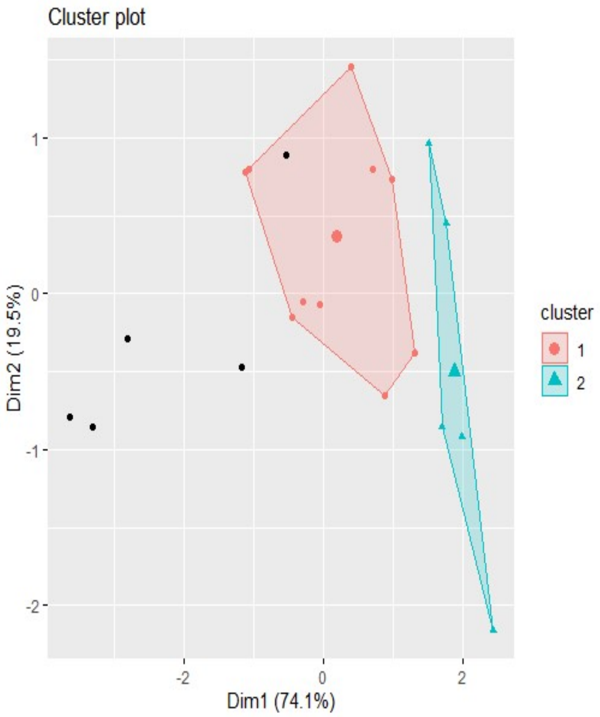

FIGURE 3. Graphical representation of outliers and clusters for stack loss dataset.

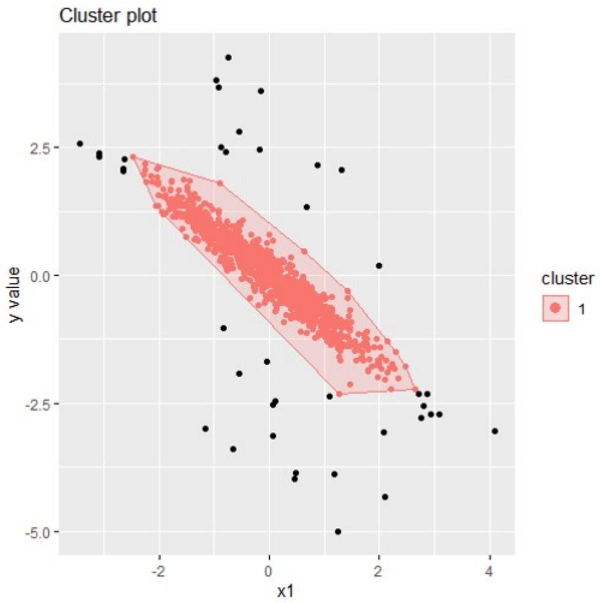

FIGURE 4. Graphical representation of outliers and clusters for simulated dataset

After detection of outliers for the simulated dataset, LM and MSOM are constructed. Performance results of the models are given in Table 2. The results in this table show that the proposed approach is still a much better than traditional approaches as the data size increases or the number of outliers in the dataset increases.

To summarize, for the computational process of outlier detection, we use DBSCAN algorithm. By using this clustering algorithm, the MSOM is improved in terms of CPU time and user effort.

TABLE 2. Performance results of LM and MSOM based on simulated dataset

Measures	**MSE**	**RMSE**	$\mathbf{R}^2$	$\mathbf{AdjR}^2$	**r**
LM	0.0473	0.2175	0.7278	0.7276	0.8531
MSOM	0.0112	0.1057	0.9357	0.9330	0.9973

5. Conclusion

Main goal of this study is proposing a new approach for a robust LM estimation within the existence of outliers. This new computational approach is based on DBSCAN and MSOM methods. DBSCAN is used for detecting the location of outlier observations effectively since it is fast, stable under perturbations on data and appropriate also for high dimensional dataset. On the other hand, MSOM is constructed as a robust linear model to overcome instability in modeling, and it also does not ignore outlier observations that are necessary to model the data adequately. The proposed method has been performed on real world and simulated datasets. It is observed that this approach performs quite well in terms of computational time and accurate detecting ability of the outlier observations than the traditional methods.

It is always possible to improve this new approach for future applications. Recommendations can be summarized as follows:

- In this study, MSOM is used as a robust model. In future, in order to capture nonlinear structure in the dataset, instead of independent variables, MSOM can be formed by using data based basis functions.
- In future applications, it is also possible to apply this new approach for classification type of datasets.
- This approach can also be effectively applied to high dimensional datasets in the existence of outliers.

References

[1] Bakar, Z.A., Mohemad, R., Ahmad, A. and Deris, M.M. (2006). A comparative study for outlier detection techniques in data mining. *In 2006 IEEE Conference on Cybernetics and Intelligent Systems IEEE*, 1-6.

[2] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* Wiley, Great Britain.

[3] Campulova, M., Michalek, J. and Moucka, J. (2019). Generalised linear model-based algorithm for detection of outliers in environmental data and comparison with semi-parametric outlier detection methods. *Atmospheric Pollution Research*, 10(4), 1015-1023.

[4] Cook, R.D. (1979). Influential observations in linear regression, *Journal of the American Statistical Association*, 74, 1691-74.

[5] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

[6] Daneshgar, A., Javadi, R. and Razavi, S.S. (2013). Clustering and outlier detection using isoperimetric number of trees. *Pattern Recognition*, 46(12), 3371-3382.

[7] Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 96, 226–231.

[8] Gan, G., Ma, C. and Wu, J. (2020). *Data Clustering: Theory, Algorithms, and Applications.* Philadelphia, PA, USA SIAM Press.

[9] Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88, 1264-1272.

[10] Hahsler, M., Piekenbrock, M. and Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1), 1-30.

[11] Hastie, T., Tibshirani, R. and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

[12] Huang, J., Zhu, Q., Yang, L., Cheng, D. and Wu, Q. (2017). A novel outlier cluster detection algorithm without top-n parameter. *Knowledge-Based Systems*, 121, 32-40.

[13] Huber, P.J. (1977). Robust covariances. *In Statistical Decision Theory and Related Topics*, 165-191.

[14] Kima, S.S., Parkb, S.H. and Krzanowskic, W.J. (1974). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35(3), 283–291.

[15] Montgomery, D.C. and Peck, E.A. (1992). *Introduction to Linear Regression Analysis.* John Wiley & Sons, New York.

[16] Rao, C.R., Toutenburg, H. and Fieger, A. (1999). *Linear Models: Least Squares and Alternatives.* Second edition, Springer.

[17] Rencher, A.C. (2000). *Linear Models in Statistics.* John Wiley & Sons, New York.

[18] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, New York.

[19] SAS Customer Support. http://support.sas.com/

[20] Taylan, P., Yerlikaya-Özkurt, F. and Weber, G.W. (2014). An approach to the mean shift outlier model by Tikhonov regularization and conic programming. *Intelligent Data Analysis*, 18(1), 79-94.

[21] Wang, Y.F., Jiong, Y., Su, G.P. and Qian, Y.R. (2019). A new outlier detection method based on OPTICS. *Sustainable Cities and Society*, 45, 197-212.

[22] Xia, J., Gao, L., Kong, K., Zhao, Y., Chen, Y., Kui, X. and Liang, Y. (2018). Exploring linear projections for revealing clusters, outliers, and trends in subsets of multi-dimensional datasets. *Journal of Visual Languages and Computing*, 48, 52-60.

[23] Xu, X., Liu, H., Li, L. and Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. *International Journal of Computational Intelligence Systems*, 11(1), 652-662.

[24] Xu, R. and Wunsch, D. (2008). *Clustering.* John Wiley & Sons, New Jersey.