11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

An Efficient Method of Identification of Outliers for High Dimensional Data for Making Accurate Statistical Inferences

Habshah Midi[1,2] * , Jaaz Suhaiza[1,3] & Norli Abdullah[4]
[1]Institute for Mathematical Research, UPM, 43400, Selangor,
[2]Department of Mathematics & Statistics, UPM, 43400, Selangor
[3]Faculty of Computing & Multimedia, Universiti Poly-Tech Malaysia,
56100 Cheras, Kuala Lumpur, Malaysia.
[4] Mathematics Division, Centre for Foundation Studies in Science, University of Malaya.

Abstract:

High Leverage Points (HLPs) are outlying observations in the *X* -directions. It is very imp ortant to detect HLPs for low and high dimensional data because the computed values of various statistical estimates are affected by their presence. In real application, many of t he data are of high dimensional. It is of vital importance to identify HLPs before making a ny statistical analysis to avoid misleading interpretation, for example to discriminate canc erous cells from non-cancerous cells. Robust Principal Component Analysis (ROBPCA) i s the popular method of identification of HLPs. Nonetheless, the weakness of the ROBP CA is that it suffers from swamping effects for less than 30% of HLPs. In this paper, we p ropose to extend the ROBPCA so that accurate number of HLPs is identified. The propo sed method is called Improved ROBPCA and denoted as IRPCA. Numerical examples h ave shown that the ROBPCA has swamping effect. The performance of the IRPCA and t he existing Minimum Regularized Covariance Determinant and PCA-based method (MRCD-PCA) is equally good. However, The MRCD-PCA algorithm is quite cumbersome and required longer computational running time. The attractive feature of the IRPCA is that it provides a simpler algorithm and it is very fast.

Keywords: High Leverage Point, minimum regularized covariance determinant, principal component analysis, robust mahalanobis distance

1. Introduction:

In real life problem, we may encounter data that are of high dimensional. The analysis of high-dimensional data has become increasingly important in many fields such as applied sciences and medicines. For instance, there are tens of thousands of gene expression values available in tumor classification utilizing genomic data; however, the number of arrays is only at the order of 10. Such high-dimensional data that refer to a situation when the number of predictor variables (*p*) is much larger than the sample size (*n*) forms a major statistical challenge in terms of data classification and other statistical analyses.

High dimensional data refers to the situations where the number of covariates or independent variables is much larger than the number of data points (i.e., $p >> n$). Dealing with this kind of data sets involves new challenging issue since it is difficult to analyze high dimensional data, due to the high correlation between variables and the risk of model overfitting. The application of conventional statistical approaches to high dimensional data tends to be ineffective, can cause serious misleading result and difficult interpretation on the pattern of the data particularly in the presence of outliers.

Among the three types of outliers; vertical outliers, residual outliers, and high leverage points, HLPs have the most detrimental effect on the computed values of various statistical estimates (Midi et al. 2021; Rashid et al. 2021; Zahariah & Habshah 2023; Habshah et al. 2023). Hence, it is imperative to identify HLPs before making any statistical analysis to avoid misleading interpretation. Accurate detection of HLPs is of vital importance in statistical analysis, as an incorrect identification of such points will substantially disrupt the standard error of estimates and give rise to a multicollinearity problem, masking and swamping of outliers, overfitting or underfitting of a model which will lead to insignificant prediction (Zahariah et. al., 2022). Swamping' refers to a situation where good observations incorrectly declared as outliers, while 'masking' refers to a situation where outliers are incorrectly declared as inliers (Rashid et al. 2021). This is the reason why the detection of outliers or HLPs is essential before making any kind of inferences.

There are many good papers in the literatures for the identification of HLPs in linear model and low dimensional data (Rousseeuw & Driessen, 1999; Lim and Midi (2016). However, only scarce papers are available in the literature for the detection of HLPs in high dimensional data. Hubert et al. (2005) developed Robust Principal Component Analysis (ROBPCA) which is the combination of projection pursuit and robust covariance estimate, i.e. MCD. ROBPCA is one of the popular method for the detection of HLPs in high dimensional data. It can successfully identify HLPs, but has serious shortcoming whereby this method suffers from swamping effects for less than 30% of HLPs (Zahariah & Habshah 2023). Moreover by making an assumption that the k-dimensional variables follow a multivariate normal distribution, the ROBPCA uses cut-off point which is based on Chi-Squared distribution. This cut-off point is inappropriate unless normality assumption is satisfied. Nevertheless, in a real situation, there is no guarantee that data would come from a multivariate normal distribution.

Robust Mahalanobis distance (RMD) is a very popular diagnostic tool used for the identification of HLPs (Hubert et al. 2012). Robust location and robust covariance matrix such as Minimum covariance determinant (MCD) are used in the computation of RMD (Rousseeuw, 1985). However, most of the robust covariance matrix is only applicable for low dimensional data because it is not invertible in high dimension cases. To rectify this problem, Boudt et al. (2018) introduced a minimum regularized covariance determinant (MRCD). Robust Mahalanobis distance which is based on the MRCD (RMD-MRCD) is then established. Nonetheless, as per Zahariah and Habshah (2023), the RMD-MRCD method suffers from serious masking effect for p>200. As a solution to this problem, Zahariah and Habshah (2023) proposed robust Mahalanobis distance (RMD) based on the combined methods of the minimum regularized covariance determinant and the principal component analysis. It is developed by incorporating the Principal Component Analysis (PCA) method in the MRCD algorithm and this method is denoted as MRCD-PCA. However this method is quite cumbersome and takes longer computational time. As a solution to this shortcoming of the MRCD-PCA, in this paper another version of diagnostic method of identification of HLPs in high dimensional data is proposed.

2. Methodology:

The Improved Robust Principal Component Analysis (IRPCA) for the Identification of High Leverage Points.

As already mentioned in the introduction section, ROBPCA can correctly identify outliers. However, it suffers from a serious swamping effect especially for high dimensional data. Thus, we propose to improve the ROBPCA so that the swamping effect can be reduced. The proposed method that we call IRPCA combines the idea of principal component analysis (PCA) and Minimum Regularized Covariance Determinant (MRCD whereby it is simple to implement and takes less computation running times. This involved transforming a high-dimensional space into a lower-dimensional subspace by using PCA and subsequently conducting our work within this newly established principal component subspace. Then, the Minimum Regularized Covariance Determinant (MRCD) is applied to this newly derived low dimensional subspace to obtain the location and scatter matrix. Instead of using robust score distance (SD) or orthogonal distance (OD) to identify outliers in HDD as suggested by Hubert et al. (2005), we propose using Robust Mahalanobis distance (RMD) to detect HLPs and suggest a confident bound type of cut-off point.

The IRPCA methodue can be summarized as follows:

Step 1 : Center the data by subtracting the median of each column x_j from each observation x_{ij}

$$x_{ij} - median(x_j) \quad (1)$$

Step 2 : Apply Principal Component Analysis (PCA) to the centered data to reduce from the original *p* variables into *k* dimensional subspace where k << p. The number of dimensions *k* retained is based on the Scree plot or Cumulative Variance for at least 80%.

Step 3 : Project the data points on the *k*-dimensional subspace and obtain the principal component score where the score are the entries of $n \times k$ matrix

$$Tn,k = (Xn,p - 1n\,\hat{\mu}')Pp,k \quad (3)$$

where $P_{p,k}$ consists of the first *k* columns of $P_{p,p}$

Step 4 : Estimate the robust scatter matrix of the principal component score within k-dimensional subspace using the Minimum Regularized Covariance Determinant (MRCD) estimator. The robust location and scatter estimates are indicated as $\hat{\mu}_{IRPCA}$ and $\sum_{IRPCA}$, respectively.

Step 5 : Calculate Robust Mahalanobis Distance (RMD) for each observation of the HDD based on the robust location and scatter estimates obtained from Step (4). The RMD of the proposed method is given by

$$RMD_i(IRPCA) = \sqrt{(x_i - \hat{\mu}_{IRPCA})^T \Sigma_{IRPCA}^{-1} (x_i - \hat{\mu}_{IRPCA})} \quad (4)$$

Step 6 : Calculate the cut-off point to identify HLPs. Since the distribution of $RMD_i(IRPCA)$ is intractable, as per Habshah et al. (2009), Rashid et al. (2022) and Zahariah & Habshah (2023) , the confident bound type of cutoff point for $RMD_i(IRPCA)$ is employed as follows,

$$median(RMD_{IRPCA}) + 3MAD(RMD_{IRPCA}) \quad (5)$$

Any observations such that its $RMD_i(IRPCA)$ exceeds the cut-off point are declared as HLPs.

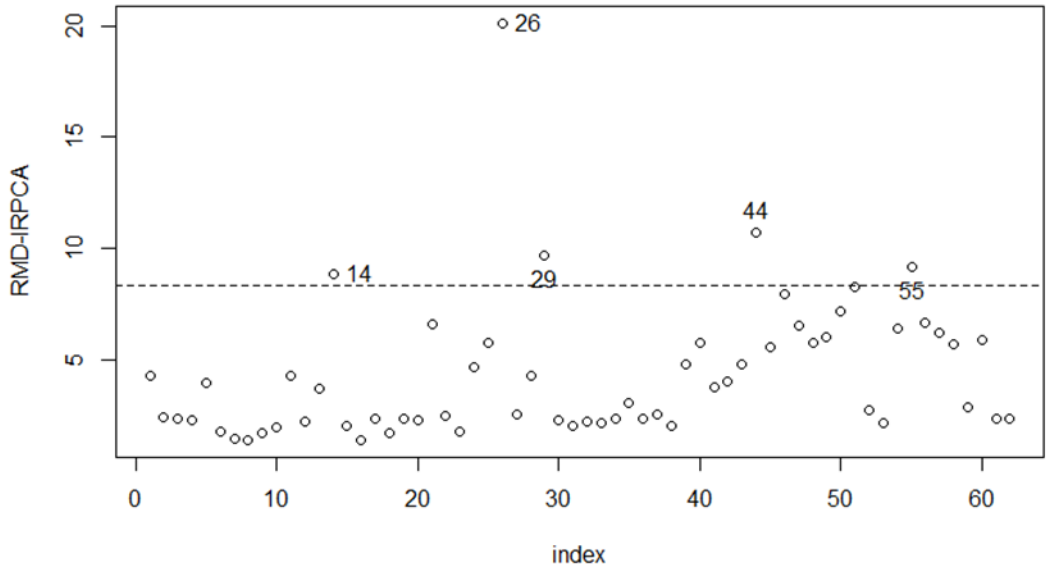
3. Result and Discussion: Simulation Study and Real Example

A simulation study similar to that of Boudt et al. (2018) and Zahariah and Midi (2023) was conducted to assess the performance of our proposed IRPCA method and compared the results with the ROBPCA and MRCD-PCA methods. The results of the simulation study is not reported here. The results show that the IRPCA and the MRCD-PCA are equally good in terms of having 100% successful in the detection of HLPs, no masking effect and a negligible swamping effect. Nonetheless the MRCD-PCA algorithm is quite cumbersome and takes the longest computational running times. On the other hand, the ROBPCA performs poorly for less than 30% HLPs. The running time for IRPCA is much faster than the MRCD-PCA and ROBPCA methods.

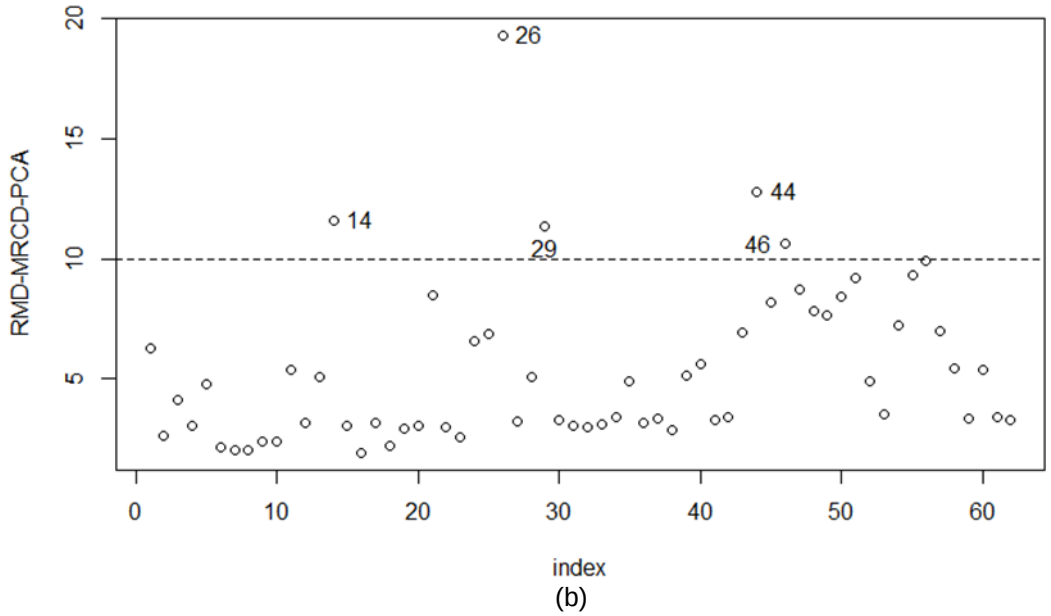
The performance of the IRPCA method is further investigated by using real data.
A real skull data set is used to evaluate the performance of our proposed IRPCA method. In this study, we considered a sample of size 62, comprising of 38 individuals with syndromic craniosynostosis (SC), and 24 individuals with normal skull who attended the Craniofacial Clinic from November 2021 to December 2023 at University of Malaya Medical Centre, Kuala Lumpur. The inclusion criteria for SC patients was paediatric subjects with SC without any previous surgical intervention and paediatric subjects with complete cranial and facial CT scans. The exclusion criteria are non-syndromic or isolated craniosynostosis (for patient group), subjects with existing or previous skull anomaly (for non-craniosynostosis group), incomplete CT scan records, and any midface hypoplasia associated with other diseases. The age range of the subjects was between 0 to 132 months. A total of 5 response measurements of maxillary and zygomatic bones were selected by the medical expert for the convenience of the skull analysis and 92 variables of various measurements on the whole skull were treated as independent variables. The IRPCA, MRCD-PCA and ROBPCA were then applied to the data. The number of HLPs detected and the computer running times by each method are exhibited in Table 1. For quick interpretation on the number of detected HLPs, one may refer to Figure 1. It is interesting to observe from the graph of Figure 1(a) and 1(b) and Table 1 that both IRPCA and MRCD-PCA detect almost the same number of observations as HLPs. The IRPCA detected (cases 14, 26, 29, 44 & 55) while MRCD-PCA detected (cases 14, 26, 29, 44 & 46). As can be expected, the ROBPCA suffers from masking effect where it detected more than 5 observations as HLPs (cases 1, 13, 14, 21, 24, 26, 29, 43, 44, 49, 50, 51, 55 & 56). It can be seen from Table 1 that the running time for IRPCA is much faster than the MRCD-PCA and ROBPCA methods.

Table 1: Number of HLPs detected by IRPCA, MRCD-PCA, and ROBPCA, and computation time (in seconds).

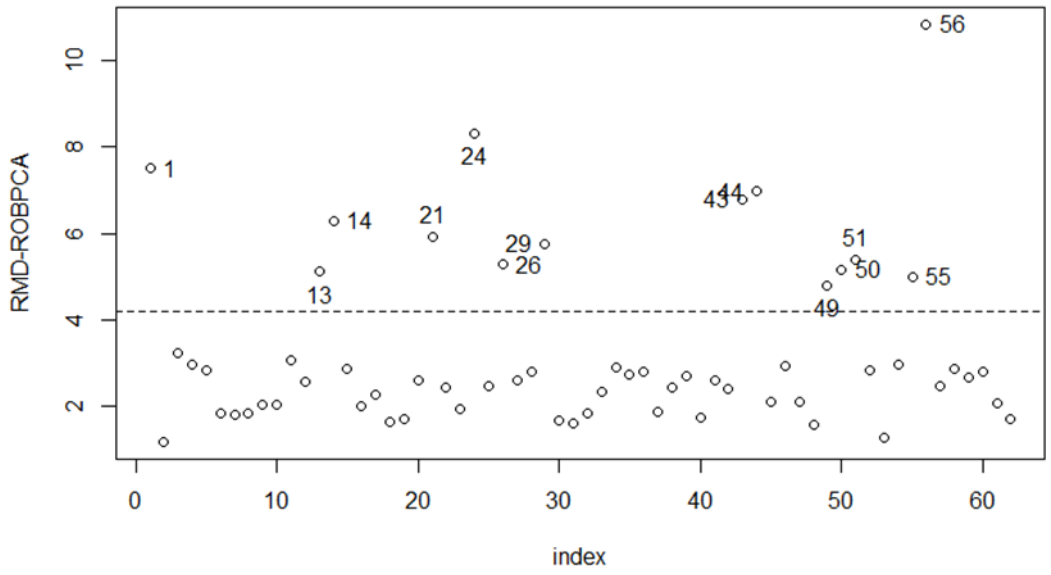
Method	Number of HLPs detected	Time (in seconds)
IRPCA	5 (14, 26, 29, 44 & 55)	3.403458
MRCD-PCA	5 (14, 26, 29, 44 & 46)	5.618928
ROBPCA	14 (1, 13, 14, 21, 24, 26, 29, 43, 44, 49, 50, 51, 55 & 56)	4.183574



(a)



(b)



(c)

Figure 1: Index plot of Craniofacial data set based on (a) eight score PC of RMD-IRPCA; (b) eight score PC of RMD-MRCD-PCA; (c) eight score PC of RMD-ROBPCA

4. Conclusion:

The proposed IRPCA is an extension work of ROBPCA. The proposed IRPCA methods and two existing methods namely the ROBPCA and MRCD-PCA are very successful in identifying HLPs. However, the ROBPCA suffers from severe swamping effect for less than 30% HLPs. The performance of IRPCA and MRCD-PCA are fairly closed to each other in terms of correct detection of outliers, no masking effects and having very small swamping effects. Nonetheless, the MRCD-PCA algorithm is not straight forward, was somewhat computationally cumbersome and, it takes very long computational running times. On the other hand, the IRPCA algorithm is quite simple and its computational running time is much faster than the MRCD-PCA.

References:

1.Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2018). The minimum regularized covariance determinant estimator. *Statistics and Computing*. https://doi.org/10.1007/s11222-019-09869-x

2.Habshah, M., Ismaeel, S. S. Arasan, J. & Mohammed, A.M. 2021. Simple and Fast Generalized –M (GM)Estimator and Its Application to Real Data. *Sains Malaysiana*, 50(3): 859-867.

3.Habshah, M., Talib, H., Jayanthi, A. & Uraibi, H.S. 2020. Fast and Robust Diagnostic Technique for the Detection of High Leverage Points. *Journal of Science and Technology.* 28 (4):1203-1220.

4.Habshah, M., Talib, H., Uraibi,H. Arasan, J. & Ismaeel, S.S. 2023. An Efficient Method of Identification of Influential Observations in Multiple Linear Regression and Its Application to Real Data. *Sains Malaysiana.* 52 (12): 3879-3892.

5.Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. Journal of Computational and Graphical Statistics, 21(3), 618–637. https://doi.org/10.1080/10618600.2012.672100

6.Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79. https://doi.org/10.1198/004017004000000563

7.Hubert, M., Reynkens, T., Schmitt, E., & Verdonck, T. (2015). Sparse PCA for High-Dimensional Data with Outliers. Technometrics, 58(4), 424–434. https://doi.org/10.1080/00401706.2015.1093962

8.Lim, H.A & Habshah, M. (2016). Diagnostic Robust Generalized Potential Based on Index Set Equal- ity (DRGP (ISE)) for the identification of high leverage points in linear model, Computational Statistics 31 (2016), pp. 859-877.

9. Rousseeuw, P.J., and Van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633–651.

10. Rousseeuw, P., & Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance. *Technometrics*, *41*(3), 212–223.

11.Rashid, A.M., Midi, H., Dhnn, W. & Arasan, J. 2021. An Efficient Estimation and Classification Methods for High Dimensional Data Using Robust Iteratively Reweighted SIMPLS Algorithm based on Nu-Support Vector Regression. *IEEE Access.* 9 (2021): 45955-45967

12.Rashid, A.M., Midi, H., Dhnn, W. & Arasan, J. 2022. Detection of Outliers in High-Dimensional Data Using Nu-Support Vector Regression. *Journal of Applied Statistics*. *49*(10): 2550-2569.

13.Zahariah, S.,Midi, H. & Mustafa, M.S. 2022. An Improvised SIMPLS Estimator Based on MRCD-PCA Weighting Function and Its Application to Real Data. *Symmetry*, 1-22.

14.Zahariah, S. & Midi, H. 2023. Minimum Regularized Covariance Determinant and Principal Component Analysis –based method for the Identification of High Leverage Points in High Dimensional Sparse Data. *Journal of Applied Statistics*. 50 (13): 2817-2835.

NOTE: THE REQUIRED NUMBER OF PAGES FOR PAPER IS SIX PAGES