KEMENTERIAN EKONOMI
JABATAN PERANGKAAN MALAYSIA

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

PROF. DR. HABSHAH MIDI
UNIVERSITI PUTRA MALAYSIA
JAAZ SUHAIZA JAAFAR
UNIVERSITI POLY-TECH MALAYSIA
DR. NORLI ABDULLAH
UNIVERSITI OF MALAYA

PERSIDANGAN STATISTIK
KALI KE-11
"Data dan Kecerdasan Buatan: Memperkasa Masa Depan"

19 September 2024

Dianjurkan oleh:
KEMENTERIAN EKONOMI
JABATAN PERANGKAAN MALAYSIA
BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA
MALAYSIA INSTITUTE OF STATISTICS

MyStats 2024

PRESENTATION OUTLINE

01 Introduction

02 Objective

03 Methodology

04 Simulation Study & Real Example

05 Conclusion

06 References

Introduction

- High dimensional data (HDD) refers to a situation when the number of predictor variables (p) is much larger than the sample size (n),

  p >> n.

- Ex. of HDD, in gene analyses, millions of genes are measured for a single individual, tens of thousands of gene expressions values available in tumor classification using genome data, an image analysis contains thousands of resolution images in pixels with a small number of samples and many more.

Introduction

- The detection of high leverage points is very crucial, for example in a microarray data analysis to spot a malignant tumor in an MRI scan (Phillip and Foss, 2008), , and in classifying fraud detection in credit card transactions (Porwel and Mukund, 2018).

- Challenge in analyzing HDD, matrix related to some algorithm may become singular.

- The existing classical method based on the Mahalanobis distance is not applicable in HDD since the covariance matrix is not invertible.

Introduction

- Not many works have focused on detecting HLPs for HDD. Dhnn, Rana and Midi (2015, Journal of Applied Stat.), Rana, Dhhan and Midi (2018, Econ.Comput, Cybern Studies), Rashid, Midi and Dhhan (2022, Journal of Applied Stat.) developed methods based on Support Vector Regression.

- Hubert et al. (2005) proposed the use of the Robust Principal Component Analysis (ROBPCA) to diagnose bad and good leverage points in HDD. However, we discovered that the ROBPCA procedure does not perform well for outliers less than 30%.

Introduction

- Boudt et al. (2018) developed the Minimum Regularized Covariance Determinant (MRCD) technique to obtain mean and covariance matrix for HDD and used it to compute RMD to detect outliers (hereinafter referred to as HLPs)
- The method is very successful for the detection of HLPs in HDD sparse data. Nevertheless, our investigation shows that the performance of the RMD-MRCD fail to correctly detect HLPs when the dimension is more than 700.
- Zahariah and Habshah (2023, Journal of Appl Statistics) developed MRCD-PCA diagnostic method of detecting HLPs. The MRCD-PCA is very successful in the detection of HLPs with small swamping effect. The only shortcoming of this method is that its algorithm is quite cumbersome and takes longer computational running times.
- Thus, it is very important to establish an alternative reliable method of identification of HLPs in HDD by integrating MRCD in its establishment.

Objectives

- To develop a reliable method of identification of HLPs in HDD, denoted as IRPCA.
- To show that the developed method is more reliable than the existing MRCD-PCA & ROBPCA.
- To apply the methods to real data.

MINIMUM REGULARIZED COVARIANCE DETERMINANT (MRCD)

- Constraint in the MCD system to be applied to HDD. For the MCD, p must satisfy p < h for any h-subset to obtain a non-singular covariance matrix.

- An improvement to the MCD is needed to make it work for HDD. Boudt et al. (2018) formulated a new modification of the MCD, the so-called Minimum Regularized Covariance Determinant (MRCD).

MINIMUM REGULARIZED COVARIANCE DETERMINANT (MRCD)

- The fundamental objective of the MRCD is to substitute a regularized covariance estimate for the MCD subset-based covariance. H-subset of MRCD that minimizes the determinant of regularized covariance of MRCD, K(H) is as shown below,

$$H_{mrcd} = \arg \min_{H \in \mathrm{H}_h} \left(\det K(H) \right)^{1/p}$$

where *K (H)* represents a regularized covariance matrix in MRCD.

ROBUST PRINCIPAL COMPONENT ANALYSIS (ROBPCA)

- The combination of Projection Pursuit and PCA are used to project and reduce the dimension of high dimensional data into the low dimensional data set.

- Robust covariance estimator based on MCD is then applied to this low dimensional data set.

- Two distances used in the ROBPCA approach to determine outliers in PCA: robust score distance (SD) and orthogonal distance (OD). The cut-off points are employed based on the assumption that the scores are normally distributed.

ROBUST PRINCIPAL COMPONENT ANALYSIS (ROBPCA)

- The cut-off point for SD is $\sqrt{\chi^2_{A,0.975}}$ with the assumption that the PC scores are normally distributed, and the cut-off point for OD is $(\hat{\mu}_{mcd} + \hat{\sigma}_{mcd} z_{0.975})^{3/2}$, where $z_{0.975}$ is the 97.5% quantile of the Gaussian distribution.

MINIMUM REGULARIZED COVARIANCE DETERMINANT-PRINCIPAL COMPONENT ANALYSIS (MRCD-PCA)

- The method is the combination of MRCD and PCA.

- The high dimensional data is reduced into low dimension using PCA, to obtain the principal components.

- From the low dimension data, it will be reconstructed to get back the original dimension by obtaining the fitted $\hat{x}$.

MINIMUM REGULARIZED COVARIANCE DETERMINANT-PRINCIPAL COMPONENT ANALYSIS (MRCD-PCA)

- The MRCD method was performed on the fitted $x^{\hat{}}$ to determine the robust mean and robust covariance of HDD.
- The distance of each observation is computed by employing Robust Mahalanobis Distance (RMD) based on MRCD-PCA robust estimators.
- Since the distribution of MRCD-PCA is intractable, following Habshah et al.(2009, J of Applied Stat), Dhnn, Rana and Midi (2015, Journal of Applied Stat.), Rana, Dhhan and Midi (2018, Econ.Comput, Cybern Studies), Rashid, Midi and Dhhan (2022, Journal of Applied Stat.) confident bound type of cut-off points is used to identify HLPs.

THE PROPOSED IMPROVISED ROBUST PRINCIPAL COMPONENT ANALYSIS (IRPCA)

Step 1: For each observation x_{ij} , compute the centered data matrix X by subtracting the median of each column.

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

Step 2: Apply Principal Component Analysis (PCA) to the centered data to reduce the number of 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 in which the first k loadings >> 80% (Ciao, 2006).

THE PROPOSED IMPROVISED ROBUST PRINCIPAL COMPONENT ANALYSIS (IRPCA)

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 × k matrix

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

where Pp,k consists of the first k columns of Pp,p *and* $\hat{\mu}'$ *is the mean* centered data matrix.

Step 4: Estimate the robust scatter matrix of the principal component score within k-dimensional subspace using the MRCD estimator. The robust estimated mean and the covariance matrix are indicated as $\hat{\mu}_{IRPCA}$ and $\sum_{IRPCA}$, respectively.

THE PROPOSED IMPROVISED ROBUST PRINCIPAL COMPONENT ANALYSIS (IRPCA)

Step 5: Calculate Robust Mahalanobis Distance (RMD) for each observation based on the robust estimated mean and the covariance matrix of IRPCA.

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

THE PROPOSED IMPROVISED ROBUST PRINCIPAL COMPONENT ANALYSIS (IRPCA)

Step 6: Following Habshah et al.(2009, J of Applied Stat), Dhnn, Rana and Midi (2015, Journal of Applied Stat.), Rana, Dhhan and Midi (2018, Econ.Comput, Cybern Studies), Rashid, Midi and Dhhan (2022, Journal of Applied Stat.)

the cutt-off point for RMD_{IRPCA} is given by,

$$median(RMD_{IRPCA}) + 3MAD(RMD_{IRPCA})$$

where $MAD(RMD_{IRPCA}) = \frac{median|RMD_{IRPCA} - median(RMD_{IRPCA})|}{0.6745}$

for i = 1, 2, 3,…., n.

Any observations that exceeds the cut-off point are declared as HLPs.

MONTE CARLO SIMULATION

- We conducted a simulation study similar to that of Boudt et al.’s (2018), Agostenelli et al. (2015), Hubert et al. (2005), Maronna and Zamar (2002) and Zahariah and Habshah (2022) simulation designs, to show the merit of our proposed method. Boudtt et al. (2018) only considered one size of HDD matrices (200 x 400). However, in our simulation study, we generated two different sample sizes of n = 50 and n =100 with four different dimensions of data set, = 100, 200, 300, and 500 throughout 500 simulations.

MONTE CARLO SIMULATION

- Since the MRCD estimators are location and scale equivariant, following Agostenelli et al. (2015), we assume without loss of generality the mean μ=0 and the variances in diagonal elements of ∑ are all equal to 1 where ∑ is a correlation matrix.

- A clean observation was generated from $x_i \sim N_p(0, I)$ for i = 1,2,3,..,n-m. To contaminate the data set with HLPs, we generate data similar to Maronna and Zamar's (2002). We determined the smallest eigenvalue along the eigenvector direction of ∑ and denoted it as a_0 . This is the direction where the contamination is the hardest to detect.

MONTE CARLO SIMULATION

- For the contamination model, we generated $x_i \sim N_p(y_0, \delta^2 I)$ for I > n – m, where $y_0 = k\,a_0$.Following Boudt et al. (2018), we set the distance between the outliers and clean data, k = 50. Since we wanted to identify the HLPs, we considered four different contamination levels, at 5%,10%, 20%, and 30%.

- Boudt et al. (2018), in their simulation study, considered contamination levels of 20% and 40%. They evaluated their results using the mean squared error (MSE) of the scatter estimates, demonstrating that their method provides more efficient scatter estimates compared to the Orthogonalized Gnanadesikan-Kettenring (OGK) method. However, their study did not focus on identifying HLPs.

Table 1: Percentage of correct detection of HLP, masking & swamping by MRCD-PCA, IRPCA & ROBPCA FOR n=50

Contamination (%)	p	% of correct detection			% of masking			% of swamping		
		MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA
5 (3 outliers)	100	100	100	100	0	0	0	0.812	0.916	7.372
	200	100	100	100	0	0	0	0.900	0.912	7.848
	300	100	100	100	0	0	0	0.976	1.068	7.776
	500	100	100	100	0	0	0	1.200	1.020	8.428
10 (5 outliers)	100	100	100	100	0	0	0	0.464	0.636	5.784
	200	100	100	100	0	0	0	0.436	0.608	6.684
	300	100	100	100	0	0	0	0.472	0.652	7.228
	500	100	100	100	0	0	0	0.340	0.736	7.608
20 (10 outliers)	100	100	100	100	0	0	0	0.136	0.164	2.292
	200	100	100	100	0	0	0	0.176	0.180	3.508
	300	100	100	100	0	0	0	0.164	0.232	4.504
	500	100	100	100	0	0	0	0.168	0.248	5.832
30 (15 outliers)	100	100	100	100	0	0	0	0.020	0.024	0.056
	200	100	100	100	0	0	0	0.044	0.020	0.172
	300	100	100	100	0	0	0	0.032	0.028	0.064
	500	100	100	100	0	0	0	0.036	0.024	0.256

Table 2: Percentage of correct detection of HLP, masking & swamping by MRCD-PCA, IRPCA & ROBPCA FOR n=100

Contamination (%)	p	% of correct detection			% of masking			% of swamping		
		MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA	MRCD-PCA	IRPCA	ROBPCA
5 (5 outliers)	100	100	100	100	0	0	0	0.522	0.528	5.866
	200	100	100	100	0	0	0	0.496	0.544	6.618
	300	99.92	100	100	0.08	0	0	0.642	0.550	7.046
	500	99.2	100	100	0.8	0	0	0.620	0.564	7.326
10 (10 outliers)	100	100	100	100	0	0	0	0.122	0.276	4.378
	200	100	100	100	0	0	0	0.182	0.250	5.090
	300	100	100	100	0	0	0	0.160	0.182	5.636
	500	100	100	100	0	0	0	0.156	0.160	6.290
20 (20 outliers)	100	100	100	100	0	0	0	0.040	0.044	1.470
	200	100	100	100	0	0	0	0.028	0.038	2.238
	300	100	100	100	0	0	0	0.018	0.028	3.042
	500	100	100	100	0	0	0	0.028	0.034	4.550
30 (30 outliers)	100	100	100	100	0	0	0	0.004	0	0.014
	200	100	100	100	0	0	0	0	0	0.004
	300	100	100	100	0	0	0	0.004	0	0.024
	500	100	100	100	0	0	0	0.002	0.002	0.030

Table 3: Running time (in seconds) by MRCD-PCA, IRPCA & ROBPCA for n=50

Contamination (%)	p	Running time (in seconds)		
		MRCD-PCA	IRPCA	ROBPCA
5	100	1.23140	0.06425	0.089023
	200	4.23890	0.18848	0.21462
	300	10.31759	0.42762	0.51335
	500	31.09337	1.65628	1.60271
10	100	1.0386	0.0772	0.1073
	200	4.1551	0.2022	0.2249
	300	8.5156	0.3968	0.4958
	500	28.5395	1.5013	1.6135
20	100	1.0403	0.0658	0.1120
	200	4.1246	0.1686	0.1700
	300	9.0696	0.4470	0.3943
	500	26.3369	1.8042	1.8442
30	100	1.0104	0.0631	0.0808
	200	3.2407	0.1691	0.1634
	300	8.6417	0.4580	0.5107
	500	27.3748	1.5426	1.9333

Table 4: Running time (in seconds) by MRCD-PCA, IRPCA & ROBPCA for n=100

Contamination (%)	p	Running time (in seconds)		
		MRCD-PCA	IRPCA	ROBPCA
5	100	1.43447	0.06204	0.08887
	200	4.73505	0.16347	0.18677
	300	16.08993	0.42292	0.44525
	500	49.34740	1.61324	1.86286
10	100	1.70448	0.07008	0.08988
	200	5.40792	0.15996	0.22236
	300	14.28192	0.4062	0.52776
	500	38.26656	1.56276	1.53156
20	100	1.64472	0.08916	0.09996
	200	6.31812	0.17592	0.1908
	300	14.27124	0.41952	0.429
	500	38.9274	1.58388	1.90452
30	100	1.28568	0.08628	0.09432
	200	5.87064	0.17976	0.20064
	300	15.5622	0.41892	0.4254
	500	40.48008	1.57056	1.57224

Figure 1 (a) to (d) : Number of variables vs running time (in secs.) with different levels of contamination (n=50)

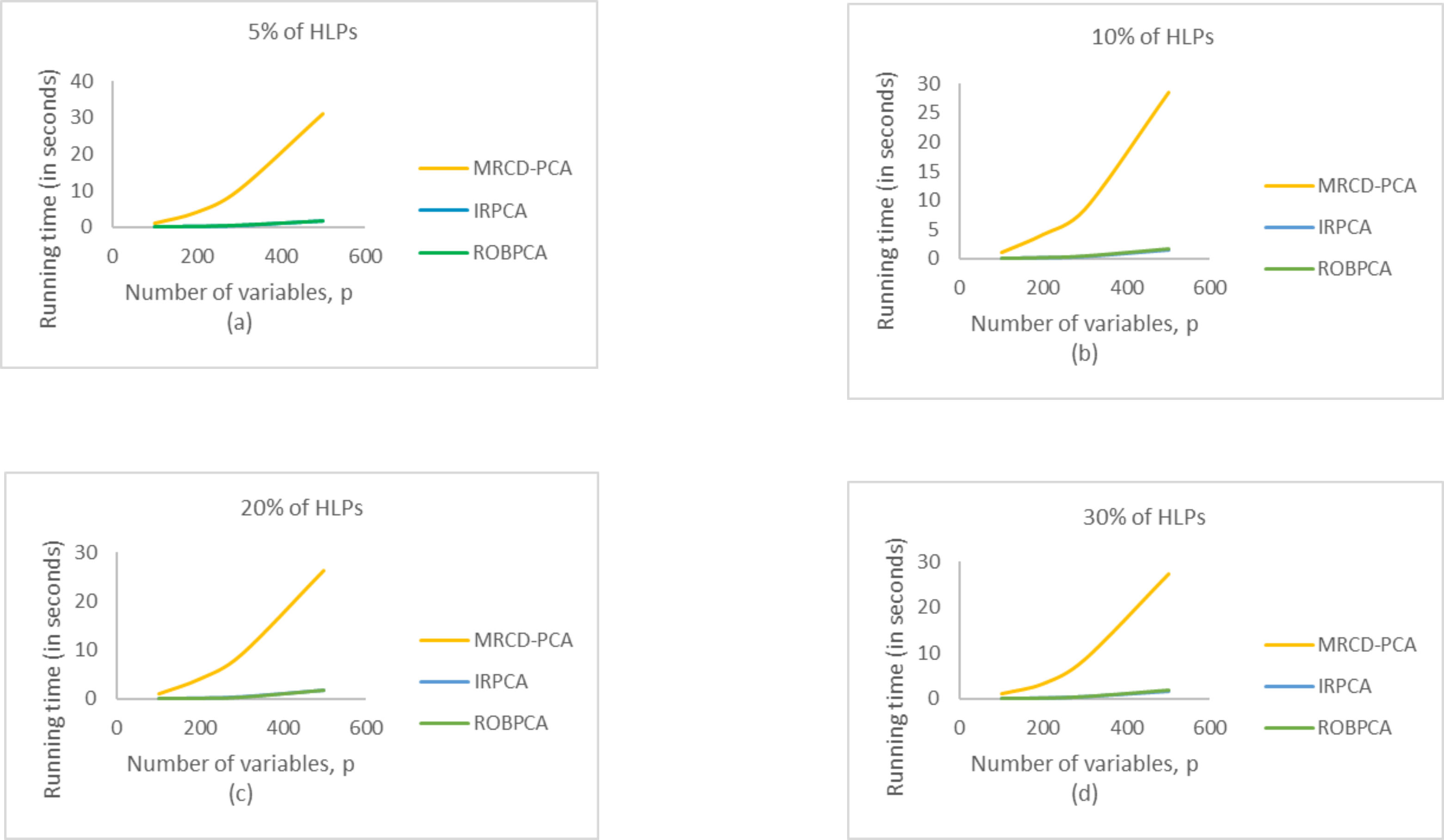
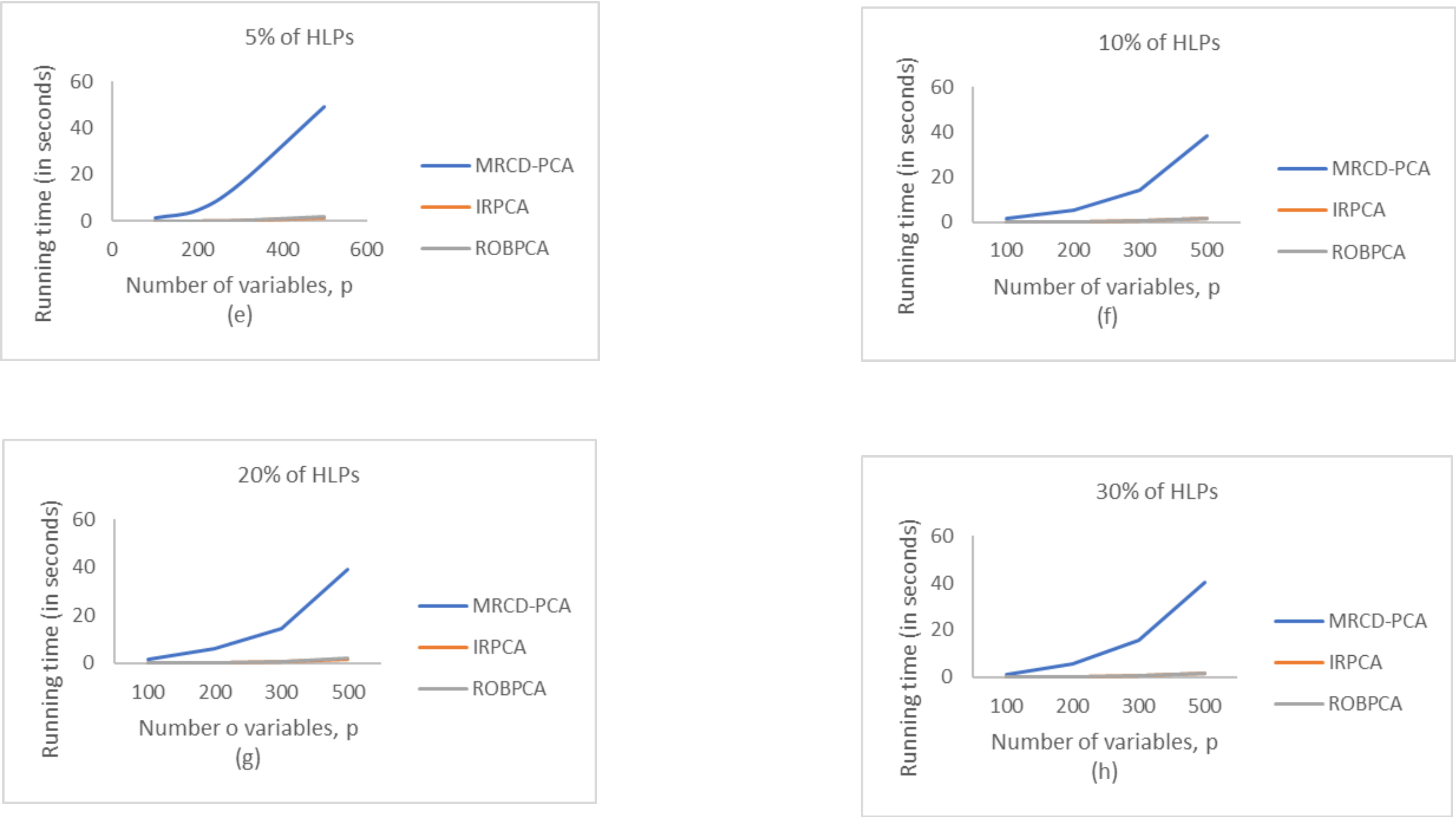


Figure 1 (a) to (d) : Number of variables vs running time (in secs.) with different levels of contamination (n=100)

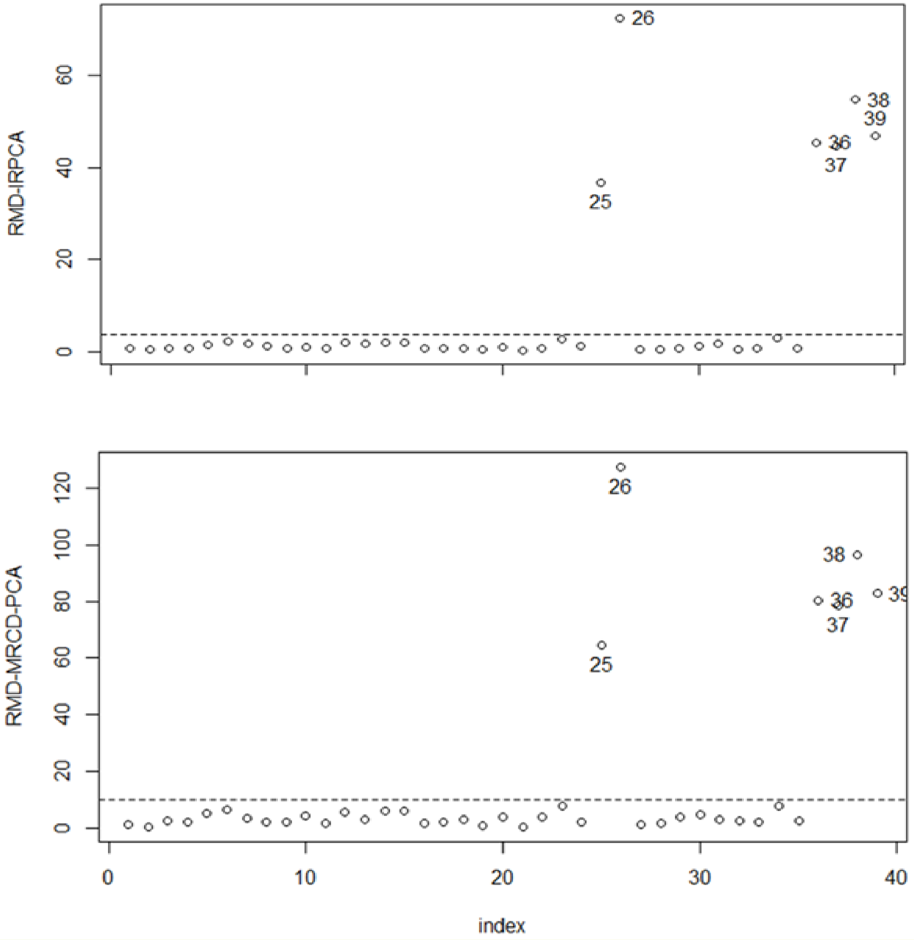


TWO REAL EXAMPLES TO ILLUSTRATE THE MERIT OR OUR METHODS

Octane data

- This dataset has been used by Hubert et al. (2005) and Boudt et al. (2018).
- It consists of near-infrared (NIR) absorbance spectra with p = 226 wavelengths and n = 39 gasoline samples.
- ROBPCA method declared six HLPs in this dataset, i.e. observation 25,26, 36, 37, 38 and 39 but also detected observation 3 & 7 as HLPs. This is caused by swamping problem.
- MRCD-PCA method successfully spots the six samples with added alcohol in the observation 25, 36, 37, 39, 38 and 26.
- IRPCA successfully identified all six observations as HLPs with the smallest computational time.

Index plot of Octane data set



- **IRPCA**

  HLP 100 % detected – Observation
  25, 26,
  36 - 39

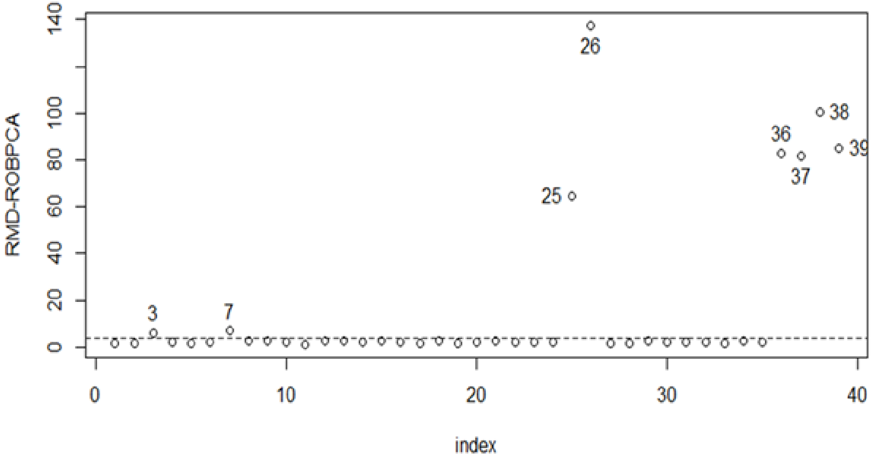
  Running Time – 6.82331 secs

- **MRCD-PCA**

  HLP 100% detected – – Observation
  25, 26,
  36 - 39

  Running Time – 21.45498 secs

Index plot of Octane data set



- **ROBPCA**

HLP 100% detected – Observation
25, 26,
36 – 39

Swamping – Observation 3 & 7

Running Time - 7.030782 secs

TWO REAL EXAMPLES TO ILLUSTRATE THE MERIT OR OUR METHODS

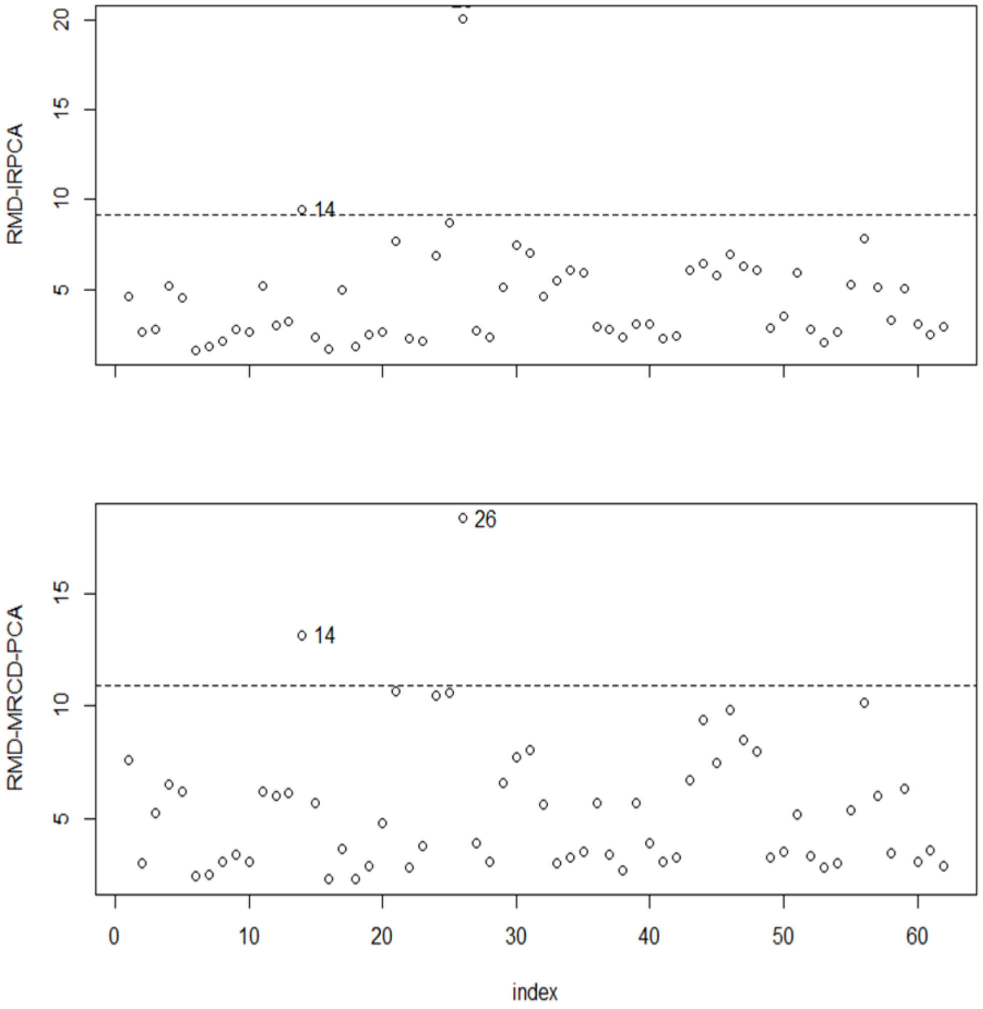
- Craniofactorial data

    - The data was collected from pediatric subjects attending the Craniofacial Clinic at the University of Malaya Medical Centre between November 2021 and December 2023.

    - The sample consists of 38 individuals with syndromic craniosynostosis (SC), & 24 individuals with normal skulls, providing a comprehensive overview of cranial variations across affected and normal subjects.

    - 92 variables of various measurements on the whole skull were treated as independent variables.

Index plot of Craniofacial data set



- **IRPCA**

  Obs. Detected as HLPs :
  14 & 26

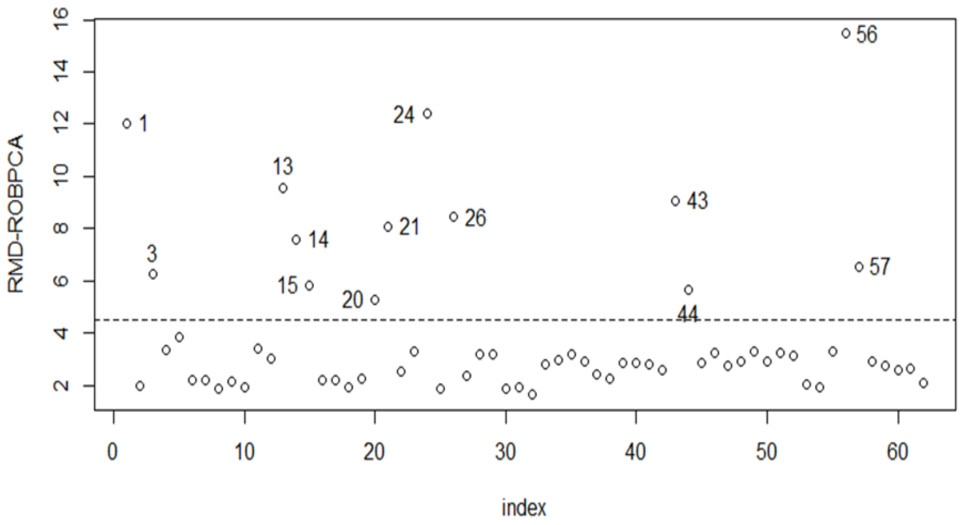
  Running Time – 3.896878 secs

- **MRCD-PCA**

  Obs. Detected as HLPs :
  14 & 26

  Running Time –6.729907 secs

Index plot of Craniofacial data set



- **ROBPCA**

Obs. Detected as HLPs :

1, 3, 13 – 15, 20 – 21, 24, 26,
43 – 44, 56 - 57

Running Time – 3.896878 secs

Conclusions

- The IRPCA methods demonstrates outstanding performance by able to detect all high leverage points (HLPs) in a high-dimensional data within a very fast computing time.

- The existing methods, MRCD-PCA and ROBPCA successfully identify high leverage points but MRCD-PCA needs longer running time while ROBPCA suffers from severe swamping problem.

- The Monte Carlo simulations and real dataset validated that our proposed method, IRPCA successfully detected HLPs with zero masking effect but with a very small swamping effect for high dimensional data with various sample sizes and number of independent variables.

References

Agostinelli, C., Leung, A., Yohai, V. J., & Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test, 24(3), 441–461. https://doi.org/10.1007/s11749-015-0450-6

Aylin, A., & Agostinelli, C. (2017). Robust iteratively reweighted SIMPLS. Journal of Chemometrics, 31(3), 1–9. https://doi.org/10.1002/cem.2881

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. Technometrics, 13(3), 469–475. https://doi.org/10.1080/00401706.1971.10488811

Asuman Seda Turkmen. (2008). Robust Partial Least Squares For Regression and Classification. Auburn University.

Bagheri, A., Habshah, M., & Imon, R. H. M. R. (2012). A novel collinearity-influential observation diagnostic measure based on a group deletion approach. Communications in Statistics: Simulation and Computation, 41(8), 1379–1396. https://doi.org/10.1080/03610918.2011.600497

Becker, C., & Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, *94*(447), 947–955. https://doi.org/10.2307/2670009

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

References

Boulesteix, A.-L. (2004). PLS dimension reduction for classification of microarray data.

Boulesteix, A. L., & Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics, 8(1), 32–44. https://doi.org/10.1093/bib/bbl016

Brereton R. G. (2015) The Mahalanobis Distance and Its Relationship to Principal Component Scores, J. Chemometrics, 29, pages 143–145. https:// doi: 10.1002/cem.2692.

Cao, L. (2006). Singular Value Decomposition applied to digital image processing. Division of Computing Studies, Arizona State University …, 1–15. Retrieved from http://www.lokminglui.com/CaoSVDintro.pdf

Chiang, J.-T. (2016). The masking and swamping effects using the planted mean-shift outliers models. International Journal of Contemporary Mathematical Sciences, 2(7), 297– 307. https://doi.org/10.12988/ijcms.2007.07024

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. Journal of the American Statistical Association, 88(423), 872–880. https://doi.org/10.1080/01621459.1993.10476352

Croux, C., & Haesbroeck, G. (1999). Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. Journal of Multivariate Analysis, 190(2), 161–190. https://doi.org/https://doi.org/10.1006/jmva.1999.1839

References

Dhhan, W., Rana, S. & Midi. H. 2015. Non-sparse ε-insensitive support vector regression for outlier detection, J. Appl. Stat. 42 (2015), pp. 1723–173

D. Peña and V.J. Yohai, *The detection of influential subsets in linear regression by using an influence matrix*, J. R. Stat. Soc. B 57 (1995), pp. 18–44.

Esbensen, K.H., Sch¨onkopf, S., and Midtgaard, T. (1994), *Multivariate Analysis in Practice.*Camo, Trondheim

Habshah, M., Norazan, M. R., & Imon, A. H. M. R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, *36*(5), 507–520. https://doi.org/10.1080/02664760802553463

Hubert, M., & Branden, K. Vanden. (2003). Robust methods for partial least squares regression. Journal of Chemometrics, 17(10), 537–549. https://doi.org/10.1002/cem.822

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. Wiley Interdisciplinary Reviews: Computational Statistics, 10(3), 1–11. https://doi.org/10.1002/wics.1421

References

Hubert, Mia, Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. Technometrics (Vol. 47). https://doi.org/10.1198/004017004000000563

Jong, S. de. (1993). SIMPLS : an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18, 251–263.

Lim, H. A., & Midi, H. (2016).Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. Computational Statistics, 31(3), 859–877. https://doi.org/10.1007/s00180-016-0662-6

Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. Technometrics, 44(4), 307–317. https://doi.org/10.1198/004017002188618509

Mehmood, T. (2016). Hotelling T2 based variable selection in partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 154, 23–28. https://doi.org/10.1016/j.chemolab.2016.03.001

References

Midi, H., Ramli, N. M., & Imon, A. H. M. R. (2009). The performance of diagnostic-robust generaliozed potential approach for the identification of multiple high leverage points in linear regression. Journal of Applied Statistics, 36(5), 1–15.

Ndaoud, M., & Tsybakov, A. B. (2020). Optimal variable selection and adaptive noisy compressed sensing. IEEE Transactions on Information Theory, 66(4), 2517–2532. https://doi.org/10.1109/TIT.2020.2965738

Peter J. Rousseeuw,A diagnostic plot for regression outliers and leverage points, Computational Statistics & Data Analysis, Volume 11, Issue 1, 1991,Pages 127-129, ISSN 0167-9473, https://doi.org/10.1016/0167-9473(91)90059-B.

Rahmatullah Imon, A. H. M. (2005). Identifying multiple influential observations in linear regression. Journal of Applied Statistics, 32(9), 929–946. https://doi.org/10.1080/02664760500163599

Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). Outlier detection for high-dimensional data.

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

References

Rousseeuw, P., & Driessen, K. (19Rousseeuw, P.J., and Van Zomeren, B.C. (1990), “Unmasking Multivariate Outliers and Leverage Points,” *Journal of the American Statistical Association*, 85, 633–651.

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

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics, 20(4), 397–405. https://doi.org/10.1080/00401706.1978.10489693

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.

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.

Terima Kasih