Environment and Ecology Research 9(3): 114-118, 2021
DOI: 10.13189/eer.2021.090303 http://www.hrpub.org

A RPCA-Based Tukey's Biweight for Clustering Identification on Extreme Rainfall Data

Siti Mariana Che Mat Nor[1], Shazlyn Milleana Shaharudin[1,*], Shuhaida Ismail[2], Kismiantini[3]

[1]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
[2]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia

Received January 28, 2021; Revised April 6, 2021; Accepted May 30, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Siti Mariana Che Mat Nor, Shazlyn Milleana Shaharudin, Shuhaida Ismail, Kismiantini , "A RPCA-Based Tukey's Biweight for Clustering Identification on Extreme Rainfall Data," Environment and Ecology Research, Vol. 9, No. 3, pp. 114 - 118, 2021. DOI: 10.13189/eer.2021.090303.*

(b): *Siti Mariana Che Mat Nor, Shazlyn Milleana Shaharudin, Shuhaida Ismail, Kismiantini (2021). A RPCA-Based Tukey's Biweight for Clustering Identification on Extreme Rainfall Data. Environment and Ecology Research, 9(3), 114 - 118. DOI: 10.13189/eer.2021.090303.*

Abstract In high dimensional data, Principal Component Analysis (PCA)-based Pearson correlation remains broadly employed to reduce the data dimensions and to improve the effectiveness of the clustering partitions. Besides being prone to sensitivity on non-Gaussian distributed data, in a high dimensional data analysis, this algorithm may influence the partitions of cluster as well as generate exceptionally imbalanced clusters due to its assigned equal weight to each observation pairs. To solve the unbalanced clusters in hydrological study caused by skewed character of the dataset, this study came out with a robust method of PCA in term of the correlation. This study will explain a RPCA to be proposed as an alternative to classical PCA in reducing high dimensional dataset to a lower form as well as obtain balance clustering result. This study improved where RPCA managed to downweigh the far-from-center outliers and develop the cluster partitions. The results for both methods are compared in term of number of components and clusters obtained as well as the clustering validity. Regarding the internal and stability validation criteria, this study focuses on the cluster's quality in order to validate the results of clusters obtained for both methods. From the findings, the amount of clusters had improved significantly by using RPCA compared to classical PCA. This proved that the proposed approach are outliers resistant than classical PCA as the proposed approach made a thorough observation assessment and downweigh the ones which were distant from the data center.

Keywords Principal Component Analysis (PCA), Pearson Correlation, Tukey's Biweight Correlation, cluster analysis

1. Introduction

Hydrological extreme events are situation whereby the hydrological situation is highly extreme such as sudden increase in magnitude and frequency of high-volume rainfall, which likely brings catastrophic damage to society, economy as well as the environment. Over decades, there are numerous studies on hydrological extreme events using statistical approaches. This is due to the fact that hydrological processes such as extreme events exhibit the non-linearity and non-stationary characteristics. To address this issue, previous research applied various approaches such as frequency-analysis methods [1], [2], stochastic model [3], Covariates based models [4] and many more. Recently, Principal Component Analysis (PCA), known for its ability as dimensionality reductions tools, is regularly used as a pre-processing method in subsiding the data set dimensionality comprising diverse interrelated variables while maintaining the most variations possible in

the data set [5]. Besides, PCA was often used as a guide in the process of clustering the pattern in improving the cluster solutions' effectiveness and accurateness [6].

However, the selection of clusters number for extreme rainfall identification using classical PCA often lead to inaccuracy. This is due to the fact that classical PCA are highly sensitive to the outliers since it measures the variability through the significance of the variance based eigenvalues and eigenvectors [7]. Since extreme rainfall datasets are susceptible to outliers, the used of classical PCA may not be suitable. Robust PCA (RPCA) on the other hand is a modified version of Pearson correlation matrix in classical PCA to arrange a robust cluster partition. RPCA is a robust measure of location and scale in the PCA which incorporated using Tukey's biweight correlation to downweigh observations that were far from the data center and resistant to outlying observations, due to its characteristics. RPCA was able to overcome possible challenges clusters in a high dimensional space which may influence the partitions of cluster as well as generating excessively unbalanced. Prior to that, the initial data matrix was made standard by a scale estimator and robust location so that possible masking or swamping effects could be avoided [8]. In this paper, the classical PCA and RPCA were applied to torrential rainfall data of East Cost Peninsular Malaysia to obtain a robust cluster partitions regarding the validity indices.

2. Study Area

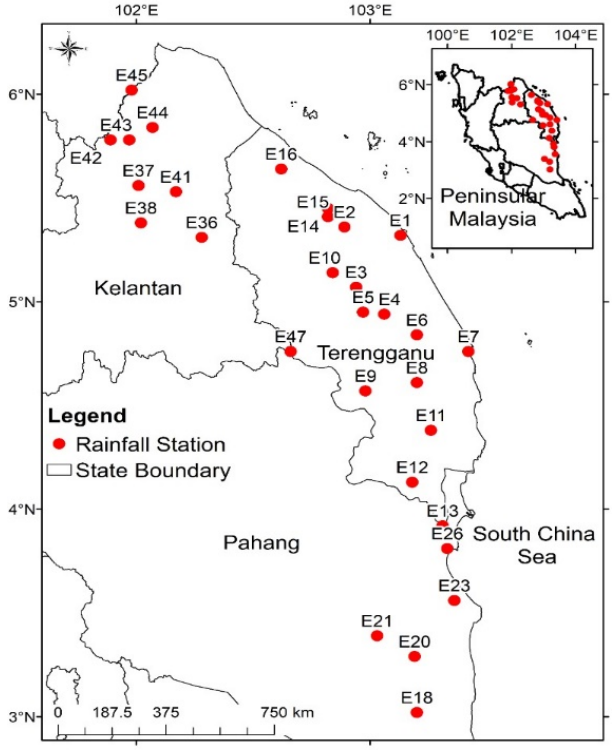

Figure 1. The location of 30 rainfall stations in east-coast Peninsular Malaysia

This study used Malaysia's Department of Irrigation and Drainage daily rainfall data with a time interval between 1987 and 2018. This study focuses on extreme hydrological event, namely the torrential rainfall, with a threshold of 60mm/day. This is a common setting applied in a tropical climate [5]. 175 days and 30 rainfall stations over east coast Peninsular Malaysia were yielded by the filtered days with rainfall exceeding 60mm in minimum 1.5% of the stations. Figure 1 shows the geographical coordinates of 48 rainfall stations chosen from east-coast of Peninsular Malaysia.

3. Methodology

Principal Component Analysis

Reducing a large dimension dataset to lower dimension while maintaining the original variability in the dataset is the purpose of the PCA [9]. An observation set of feasibly interconnected variables which transforms into a set of linearly uncorrelated ones, namely principal component (PC), helps achieve the above-mentioned purposes. The initial PC comprises the original data variations extensively. Subsequently, every succeeding component extensively comprises outstanding variations, conditional on being not correlated to prior components.

The derivative correlation matrix from the matrix contributes to the PCA in calculating its eigenvalues and eigenvectors. The related components comprising majority of the data variations is obtained as such [10-11]. A minimum of 70% of the whole variation becomes a benchmark for ideal cut-off values for cumulative percentage for extracting the number of components [12]. Once the cut-off values were determined, the newly dataset were obtained from component matrix of eigenvectors "loadings". These new datasets comprise linear transformations of the initial variables by maximizing the new principal component variance. Including the excessive principal components raises the significance of the outlier, resulting in poor pattern identification [13].

Pearson Correlation matrix

For applications like climatology and environmental sciences, Pearson correlation is generally employed in PCA to determine the eigenvectors and eigenvalues [12]. Commonly, it measures the correspondences or distances before a clustering algorithm is implemented. The Pearson correlation coefficient is defined as:

$$r_{ij} = \frac{\sum_{i=1}^{n}(X_i-\bar{X}_i)(X_j-\bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X}_i)^2 \sum_{i=1}^{n}(X_j-\bar{X}_j)^2}} \quad (1)$$

where X_i and X_j indicate the vectors of observations in matrix data $\mathbf{X}$ with n observations, with X_i and X_j indicate the vectors mean.

RPCA

Tukey's biweight correlation is dependent on the employed M-estimator in the robust correlation estimates. There is a derivative function, ψ for M-estimator that decides the assigned observational in the data set. It is capable of downweighing the observations for the purpose of reflecting the impacts from the data center [14]. The derivative function is:

$$\psi(u) = \begin{cases} u(1-u)^2 & |u| \leq 1 \\ 0 & |u| \leq 1 \end{cases} \quad (2)$$

Distinctly, if $|u|$ is adequately big, $\psi(u)$ decreases to zero. The smallest fraction of contamination in Breakdown point (BP) may cause an inaccurate result as BP is vital in measuring the resistance to the outlier data values in M-estimators [15]. For this research, Tukey's biweight with BP at 0.2, 0.4, 0.6 and 0.8 respectively are evaluated. An experiment by [16] and [9] had shown that for most conditions, the best performance was by a BP of 0.4. Such studies also found out that the results were more precise and effective than others.

There are two steps in producing the biweight estimate of correlation: first, the location estimate, $\tilde{T}$ is calculated and second, the shape estimate, $\tilde{S}$ is updated. The $(i,j)^{th}$ element of $\tilde{S}$, i.e. $\tilde{s}_{ij}$ is a resistant covariance estimate among vectors, X_i and X_j. The biweight correlation between the stated vectors could be determined by:

$$\tilde{r}_{ij} = \frac{\tilde{s}_{ij}}{\sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}} \quad (3)$$

with

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} X_i w(u_{i(k)})}{\sum_{i=1}^{n} w(u_{i(k)})} \qquad k = 0,1,2,\ldots \quad (4)$$

$$S_n^{(k+1)} = \frac{\sum_{i=1}^{n} w(u_{i(k)})(X_i - T^{(k+1)})(X_i - T^{(k+1)})^t}{\sum_{i=1}^{n} w(u_{i(k)})(u_{i(k)})} \quad (5)$$

where $T_n^{(k+1)}$ is a location vector and $S_n^{(k+1)}$ is a shape matrix such that $k = 0,1,2,\ldots$

Once the numbers of PC are determined, as a measurement, an index by [17] is applied to establish the optimal cluster partitions for the input data as implied by the maximum index value. As a consequence, a RPCA for cluster analysis exhibits higher possibilities of producing a better cluster partition as well as having higher resistance towards outlying values compared to Pearson correlation in PCA.

There are several steps regarding the proposed algorithm. First of all, the initial step is obtaining the input matrix. Secondly, the next step is standardizing the observations with the data of the medians and mean absolute deviations. Thirdly, the breakdown point is set at 0.4 for the Tukey's biweight correlation. Fourthly, the calculation of the Tukey's biweight correlation is carried out. Consequently, the selection of the principal components which are furthermost significant is done according to the overall variation accumulative percentage. After this stage is when the new set of data is derived. For this new set of data, the Calinski and Harabasz index is calculated. The significance of this step is to establish the most appropriate clusters amount. Finally, the clustering algorithm is applied.

4. Results and Discussion

Table 1 indicates a substantial dissimilarity in the sum of components as well as clusters established from both correlation measures in PCA at all levels of accumulative variation percentages. Seemingly, RPCA entails a smaller components quantity in extracting to establish, at minimum, 70% of accumulative variation percentage as a comparison to the classical PCA. As an example, 14 components were maintained with RPCA while it was 16 with classical PCA at 70% accumulative variation percentage. However, identifying rainfall patterns for the selection cumulative percentage beyond 75% was an unsuitable principal components quantity discontinuation. It is due to the result shows that cumulative percentage that extracted more than 75% variation of RPCA were obtained much number of components for rainfall dataset.

Table 1. Total components and clusters acquired according to classical PCA and RPCA from rainfall data of east-coast Peninsular Malaysia

Cumulative percentage (%)	Sum of components		Sum of clusters	
	Classical PCA	RPCA	Classical PCA	RPCA
60	22	7	2	7
65	26	10	2	4
70	30	14	2	4
75	36	18	2	4
80	42	23	2	4
85	49	29	2	4
90	58	36	2	2
95	62	45	2	2

For cluster partitions, Table 1 also demonstrates that compared to classical PCA, the sum of clusters is more influential towards RPCA depending on the sum of the retained components. The number of cluster partition become stabilize at phase 65% to 85% of the variation of the data which is 4 clusters. However, the resultant number of clusters is exactly maintained as the classical PCA when obtained more than 85% accumulative variance percentage. Moreover, the sum of clusters from the classical PCA appeared stabilizing at simply two clusters irrespective of the accumulative variation percentage employed. Predominantly in identifying rainfall patterns in hydrological studies, practically obtaining more than two cluster partitions explains the diverse rainfall patterns categories. In consequence, two clusters are observably inapplicable since the actual data structure was masked. Due to the of the clustering results sensitivity towards the acquired amount of components, the accurate amount of

components to be maintained should be classified accordingly. Importantly, the variations between clusters are indicated to be directed towards no less than a principal component [18].

Table 2. Indices to calculate clustering results quality for torrential rainfall data

	PCA-based Pearson correlation	RPCA
Connectivity	3.2790	2.9290
Dunn Index	0.2849	14.8556
Silhouette	0.6047	0.9017

For the evaluation of the cluster partitions, the clustering output at 70% accumulative variation percentage on both approaches are chosen respectively. For this study, we focus on internal and stability validation prescribed by [19], connectivity, Dunn index and Silhouette respectively. The clustering results quality obtained was demonstrated by the validity indices to signify improved internal cluster quality as well as stability. As a guideline, with a connectivity value range between 0 and infinity, and according to previous studies [20], [21], it should be minimized while the higher value of Dunn index should indicate a good quality of clusters. As for the Silhouette value, well-clustered observations have close-to-1 values while insufficiently-clustered observations have values nearing -1. Table 2 displays that RPCA shows comparatively finer clustering results for the three indices when it was contrasted against PCA-based Pearson correlation.

Table 3. Indices to calculate the clustering results stability for torrential rainfall data

	PCA-based Pearson correlation	RPCA
AD	984161.30	379054.4
FOM	74862.54	37401.03

The stability measures made comparisons of the results from the clustering according to the full data to clustering by eliminating every column, one after another [22]. According to [23], the included stability measures of clustering are the average the average distance (AD) and the figure of merit (FOM). Table 3 illustrates that RPCA also presented better clustering performance compared to PCA-based Pearson correlation. AD values for RPCA is 379054.4 which is smaller than PCA-based Pearson correlation. Smaller value of AD is preferred in stability measures of clustering. Other than that, small value of FOM for RPCA also showed better results since smaller values of FOM equaling better performance of clustering. Based on the validity indices of the clusters, conclusively, RPCA displayed better results in clustering performance in terms of internal and stability measures as a comparison to classical PCA.

5. Conclusion

Based on the findings of this research, RPCA was evidently well-performed in the clustering method compared to PCA-based Pearson correlation. Moreover, RPCA has proven its superiority over Classical PCA in terms of number of principal components as well as number of clusters. The results also showed that the number of clusters obtained using RPCA has better stability, validity indices as compared to PCA-based Pearson Correlation. This study displays significant cluster partition enhancement in dealing with imbalanced clusters existed in high dimensional data especially for hydrological data. It can be defined that the proposed RPCA is outlier resistant and able to deal extreme rainfall event in Malaysia.

Acknowledgement

The authors would like to thank Universiti Pendidikan Sultan Idris (UPSI) and the Malaysian Ministry of Education (MOE) for this study that had been produced through the Fundamental Research Grants Scheme with its referral identification of (FRGS/1/2019/STG06/UPSI/02/4).

REFERENCES

[1] D. Cooley, (2013). “Return Periods and Return Levels Under Climate Change,” in *Extremes in a Changing Climate*, New York: Springer, pp. 97–114.

[2] A. AghaKouchak, D. Easterling, K. Hsu, S. Schubert, and S. Sorooshia, (2012). “Extremes in a Changing Climate: Detection, Analysis and Uncertainty,” in *Extremes in a Changing Climate: Detection, Analysis and Uncertainty*, New York: Springer.

[3] O. G. Sveinsson, J. D. Salas, and D. C. Boes, (2005) “Prediction of Extreme Events in Hydrologic Processes that Exhibit Abrupt Shifting Patterns,” *J. Hydrol. Eng.*, vol. 10, no. 4, pp. 315–326.

[4] G. Villarini, J. A. Smith, and F. Napolitano, (2010). “Nonstationary modeling of a long record of rainfall and temperature over Rome,” *Adv. Water Resour.*, vol. 33, no. 10, pp. 1256–1267.

[5] S. M. Shaharudin, N. Ahmad, N. S. Mohamed, and N. Aziz, (2020). “Performance analysis and validation of modified singular spectrum analysis based on simulation torrential rainfall data,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 4, pp. 1450–1456.

[6] R. Indhumathi and S. Sathiyabama, (2010). “Reducing and Clustering high Dimensional Data through Principal Component Analysis.

[7] S. M. Shaharudin, N. Ahmad, and S. M. C. M. Nor, (2020). "A modified correlation in principal component analysis for torrential rainfall patterns identification," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 655–661.

[8] V. Choulakian, (2001). "Robust Q-mode principal component analysis in L1," *Comput. Stat. Data Anal.*, vol. 37, no. 2, pp. 135–150.

[9] S. M. Shaharudin and N. Ahmad, (2017). "Choice of cumulative percentage in principal component analysis for regionalization of peninsular Malaysia based on the rainfall amount," in *Communications in Computer and Information Science*, vol. 752, pp. 216–224.

[10] S. Neware, K. Mehta, and A. S. Zadgaonkar, (2013). "Finger Knuckle Identification using Principal Component Analysis and Nearest Mean Classifier," *Int. J. Comput. Appl.*, vol. 70, no. 9.

[11] Solimun, A. A. R. Fernandes, and R. A. Cahyoningtyas, (2020). "The implementation of nonlinear principal component analysis to acquire the demography of latent variable data (A study case on brawijaya university students)," *Math. Stat.*, vol. 8, no. 4, pp. 437–442. DOI: 10.13189/ms.2020.080410

[12] I. T. Jollife and J. Cadima, (2016) "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065. Royal Society of London.

[13] S. M. Shaharudin, S. Ismail, S. M. C. M. Nor, and N. Ahmad, (2019). "An efficient method to improve the clustering performance using hybrid robust principal component analysis-spectral biclustering in rainfall patterns identification," *IAES Int. J. Artif. Intell.*, vol. 8, no. 3, pp. 237–243.

[14] S. Milleana Shaharudin, N. Ahmad, and X. Yap, (2013). "The Comparison of T-Mode and Pearson Correlation Matrices in Classfication of Daily Rainfall Patterns in Peninsular Malaysia,"

[15] P. Rousseeuw, (1985). "Multivariate Estimation with High Breakdown Point," *Math. Stat. Appl.*

[16] M. Owen, (2010). "Tukey's Biweight Correlation and the Breakdown".

[17] U. Maulik and S. Bandyopadhyay, (1986). "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, 2002.

[18] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag.

[19] D. Bhalla, (2020). "Validate Cluster Analysis," [Online]. Available: https://www.listendata.com/2016/01/validate-cluster-analysis.html. [Accessed: 06-Mar-2021].

[20] Z. Mohamed and R. Rosli, (2014) "Development of A Structural Model with Multicollinearity and Outliers Problems," *Educ. - J. Sci. Math. Technol.* , vol. 1, no. 1, pp. 38–52.

[21] Y. A. Lesnussa, N. A. Melsasail, and Z. A. Leleury, (2016) "Application of Principal Component Analysis for Face Recognition Based on Weighting Matrix Using GUI Matlab," *Educ. JSMT*, vol. 3, no. 2, pp. 1–7.

[22] G. Brock, V. Pihur, S. Datta, and S. Datta, (2008). "ClValid: An R package for cluster validation," *J. Stat. Softw.*, vol. 25, no. 4, pp. 1–22.

[23] S. Datta and S. Datta, (2003). "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics*, vol. 19, no. 4, pp. 459–466.