A modified correlation in principal component analysis for torrential rainfall patterns identification

Shazlyn Milleana Shaharudin[1], Norhaiza Ahmad[2], Siti Mariana Che Mat Nor[3]
[1,3]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
[2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Article Info

Article history:

Received Jul 18, 2020
Revised Sep 20, 2020
Accepted Oct 10, 2020

Keywords:

Cluster analysis
Principal component analysis
Tukey's biweight

ABSTRACT

This paper presents a modified correlation in principal component analysis (PCA) for selection number of clusters in identifying rainfall patterns. The approach of a clustering as guided by PCA is extensively employed in data with high dimension especially in identifying the spatial distribution patterns of daily torrential rainfall. Typically, a common method of identifying rainfall patterns for climatological investigation employed T mode-based Pearson correlation matrix to extract the relative variance retained. However, the data of rainfall in Peninsular Malaysia involved skewed observations in the direction of higher values with pure tendencies of values that are positive. Therefore, using Pearson correlation which was basing on PCA on rainfall set of data has the potentioal to influence the partitions of cluster as well as producing exceptionally clusters that are eneven in a space with high dimension. For current research, to resolve the unbalanced clusters challenge regarding the patterns of rainfall caused by the skewed character of the data, a robust dimension reduction method in PCA was employed. Thus, it led to the introduction of a robust measure in PCA with Tukey's biweight correlation to downweigh observations along with the optimal breakdown point to obtain PCA's quantity of components. Outcomes of this study displayed a highly substantial progress for the robust PCA, contrasting with the PCA-based Pearson correlation in respects to the average amount of acquired clusters and indicated 70% variance cumulative percentage at the breakdown point of 0.4.

Corresponding Author:

Shazlyn Milleana Shaharudin
Department of Mathematics, Faculty of Science and Mathematics
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak, Malaysia
Email: shazlyn@fsmt.upsi.edu.my

1. INTRODUCTION

For climatologist and hydrologist, identifying the pattern of rainfall for spatial torrential is substantial specifically for the classification of hydrologic events. This process simplifies the hydrologic convolution. Hence, the amount of rainfall concerning the records of time series was monitored at a number of stations for rain gauge with an extensive data documentation would be analyzed.

According to [1], principal component analysis (also known as PCA) has been extensively employed in atmospheric science, climatology, meteorology and other scientific fields that use large datasets. In Malaysia, many studies had investigated on fitting the distribution of rainfall, either hourly, daily or annually [2-5]. PCA is sensitive to the outliers since it measures the variablility through the significance of the variance based on the computation of the eigenvalues as well as the eigenvectors of the covariance or correlation matrix in the dataset. In order to resist the distortion caused by the outliers, a robust PCA methods need to be developed.

In this case, the challenge in classifying the patterns of daily rainfall of high dimensional dataset may include high degree of immaterial and infomation redundance which has potential to lower supplementary analysis' performance. A possible way to curb abovementioned concern is to decrease the dimensions by applying PCA, followed by an analysis of cluster to classify the Peninsular Malaysia's patterns of rainfall [6]. The analysis of cluster is established for the observation segmentations into similar and dissimilar patterns of its respective cluster [7].

A common PCA approach, by building on the Pearson correlation matrix, needs the entities configuration points between data columns and rows [8]. Normally, the employment of Pearson correlation is done via the derivation of T-mode correlation to calculate the likeness of day-to-day rainfalls [9]. Since the weight of each observed pair is equal, Pearson correlations could be susceptible on non-Gaussian distributed data. This might implicate slanted observations like the outlying values [10]. Consequently, the use of PCA-based Pearson correlation on the data of torrential rainfall is capable of modifying the subsets of cluster and produce exceedingly unbalanced clusters within the context of a high dimensional space.

Observations weighting is applied to act as a measure of resistant in this study via presenting a Tukey's biweight correlation matrix as a subtitute to Pearson correlation matrix in PCA to arrange a robust cluster partition. Tukey's biweight correlation is contingent on Tukey's biweight function which is dependent on M-estimators employed in the estimation of robust correlation [11]. In addition, the different phase of breakdown points is proposed in such robust approach where it is a substantial part to acquire the components quantity for the extraction in PCA [12].

In present study, modified correlation in PCA for torrential rainfall patterns identification is presented by imputing Tukey's biweight correlation into PCA replacing Pearson correlation. This study might be helpful for hydrologist in identifying spatial cluster patterns of rainfall help to analyze environmental models and improve evaluations on climate change.

2. DATA

The rainfall data between 1975 and 2007 as recorded by Malaysia's Department of Irrigation and Drainage, i.e. *jabatan pengairan dan saliran (JPS)* was acquired. It comprised information of 75 stations across Peninsular Malaysia with distinct four regions' geographical coordinates (i.e. west, northwest, east and southwest). This current study emphasizes on the incidence of events of torrential rainfall which is also explained as extreme rainfall. Necessarily, some criteria need to be selected to lead the threshold formation. This would clearly separate the factors which institute torrential rainfall day in the studied regions of Peninsular Malaysia. For climates in the tropics, the highest employed threshold is 60 mm/day for this goal. The days that were filtered and taken into account were the ones which the rainfall which exceeded 60 mm, and this applied to at least 2% of the overall number of the stations. From this filter, there were 250 days considered valid for the data, and in 15 rainfall stations, sufficient for the demonstration of the main torrential centers as represented in Figure 1.

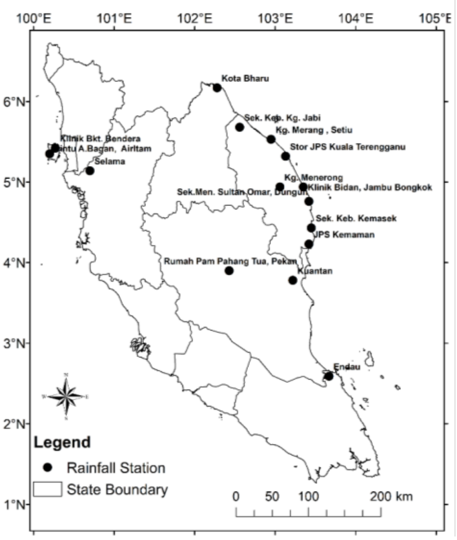

Figure 1. Rainfall stations that represent the main torrential centers in Peninsular Malaysia

3. RESEARCH METHOD

3.1. Principal component analysis

PCA is constructed for dimensional decrease of huge matrix of data to lower ones through keeping almost every initial data variability [13]. By transforming an observation set of variables which are hypothetically interrelated into a set of variables known as the principal components which are not directly interrelated, it has the achievability potential. The primary principal component represents the same amount as the original data variations. Consequently, every subsequent component represents the same amount as the rest of the variation subject while it is uncorrelated to the prior component.

In PCA, the derivation of correlation matrix or covariance from the data matrix were vital to compute eigenvalues and eigenvectors to find the associated components which illustrates the highest amount of data variations [14]. In this research, correlation matrix was employed. The common benchmark proposed to remove the eigenvalues from a huge data for the extraction of the sum of components is 70% of accumulative percentage of the overall variation [15]. The component matrix of eigenvector “loadings” that outline newer variables comprising linear transformation of the initial variables which, in another axes, amplifies the variance, to the maximum is known as reduced matrix.

Six steps were used in the PCA algorithm. First, the input data matrix was obtained. The second step was by centering the data matrix via the substraction of all means of the observation. Next, the third step was to obtain the T-mode decomposition by transposing the data. Subsequently, step four was to compute the correlation matrix of T-mode. This was followed by step five where the eigenvalues along with the eigenvectors for T-mode correlation matrix in the PCA. Finally, in step six, the matrix of component loadings was computed for the benefit of further analysis.

3.2. Tukey’s biweight correlation

Tukey's biweight correlation is contingent on Tukey's biweight function which depends on the M-estimators applied in robust correlation estimation. A derivative function, ψ of M-estimate that ascertains the weights was assigned to the observations in the set of data. It is capable of downweighting observations to exhibit the effects from the data centre [16]. Below is the derivation of the derivative function:

$$\psi(u) = \begin{cases} u(1-u)^2 \; |u| \leq 1 \\ 0 \; |u| > 1 \end{cases} \tag{1}$$

Apparently, if $|u|$ is sufficient, then $\psi(u)$ diminishes to zero. The breakdown point is a significant feature to calculate the resistance to the M-estimator’s values of outlying data. A breakdown point is the least amount of contamination portion that might lead to inaccuracy in the results [15]. For this study, using breakdown points at 0.2, 0.4, 0.6 and 0.8, theTukey's biweight were evaluated with the use of the simulated data and the best performance is at the breakdown point of 0.4. As stated by [17], generally, a breakdown point of 0.4 functions more effectively for majority of the situations. It gives more accuracy and efficiency than the lower ones.

The correlation biweight estimation is generated through primarily computing the estimation of location, T˜ and later through the update of the estimate of shape, S˜. The $(i,j)^{th}$ element of S˜, i.e. $\tilde{s}_{ij}$ functions as an estimation of the resistance of covariances between both vectors, X_i and X_j. The vectors biweight correlation are determined as below:

$$\tilde{r}_{ij} = \frac{\tilde{s}_{ij}}{\sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}} \tag{2}$$

With

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} X_i w\left(u_{i(k)}\right)}{\sum_{i=1}^{n} w\left(u_{i(k)}\right)} k = 0,1,2,\ldots \tag{3}$$

$$S_n^{(k+1)} = \frac{\sum_{i=1}^{n} w\left(u_{i(k)}\right)(X_i - T^{(k+1)})(X_i - T^{(k+1)})^t}{\sum_{i=1}^{n} w\left(u_{i(k)}\right)(u_{i(k)})} \tag{4}$$

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

Therefore, a PCA based Tukey’s biweight correlation for K-means cluster analysis has higher tendency in generating a partition of cluster which is better and has higher resistance on the outlying values compared to Pearson correlation in PCA.

3.3. Robust principal component analysis

Since Pearson correlation has higher sensitivity towards non-Gaussian distributed data, Tukey’s biweight correlation in PCA on the torrential rainfall set of data is recommended. Primarily, the data matrix would be standardized by an estimator of scal and robust location to prevent the effects of masking or swamping [18].

The reduced set of data is further used in k-Means cluster analysis to acquire the partitions of cluster. The method of k means needs stipulation of the number of clusters before applying the algorithm. For such concern, as an act of control, the Calinski and Harabasz Index [19] was applied to decide the optimum quantity of partition of cluster for the input data as signified through the maximum value of index.

There are ten steps for the employment of the suggested algorithm. Initially, the input matrix was determined. Secondly, the observation of the study was standardized by using median as well as mean absolute deviation (MAD), for instance such that x_{ij} signifies the input matrix elements.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}}{median(|x_{ij} - median(x_{ij})|)} \tag{5}$$

The third step was the data T-mode decomposition arrangement via data arrangement. The next step which is for the purpose of Tukey's biweight correlation, the breakdown point of 0.4 was set. Consequently, the Tukey's biweight correlation matrix was calculated. This was followed by the calculation of the correlation matrix eigenvalues as well as eigenvectors. The seventh step was the most substantial components were chosen depending on the cumulative percentage from the overall variation. The following step was the calculation of the component loadings’s matrix. The Calinski and Harabasz index was, afterwards, computed contingent on the component loadings’s matrix. This was for the purpose of determining the most appropriate cluster amount. Finally, k-means was employed to the component loadings’s matrix.

4. RESULTS AND DISCUSSION

Table 1 shows the sum of components that were attained using robust PCA-based Tukey’s biweight correlation extracted from the data which was simulated. For this approach, tt showed how breakdown point selection had effects on the amounts of extraction for the components. As shown in Table 1, the flagging of less significance components for extraction were led by higher breakdown point (r=0.8). Meanwhile, for r=0.4 breakdown point, it has stability to extract some components. In this case, enough components were maintained, or specifically there were 12 components. Additionally, it is unfavavorable in hydrological data for excessive components to be extracted because it could suggest low frequencies variation or else, insignificant ispatial scales [20]. Hence, according to [21], the breakdown point selection was utterly substantial the contecxt of Tukey’s biweight correlation that was based on robust PCA.

As illustrated in Figure 2, in comparison to Pearson, the Tukey's biweight correlation needs fewer components for extraction and achieving a minimum of 70% from the variation's cumulative percentage. As an example, there were 28 components maintained with the Tukey's in comparison to the Pearson's at variations cumulative percentage of 80%.

Table 1. The sum of components formed in a number of values of breakdown point on 70% variance cumulative percentage

Breakdown Point, r	Sum of Components
0.2	9
0.4	12
0.6	6
0.8	3

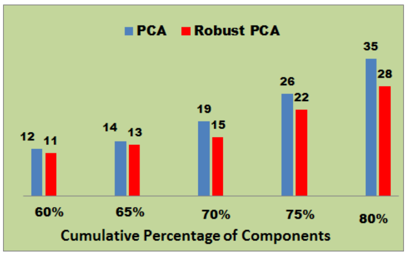

Figure 2. Bar chart depicts the amount components obtained using two approaches

For cluster partitions, Figure 3 indicates that theTukey's biweight correlation has even higher susceptibility towards the clusters total formed on the number of components maintained. This was different from the Pearson's. The total clusters based on the PCA-based Pearson correlation appeared to be stabilizing at two clusters irrespective of the use of the variation's accumulative percentage. However, PCA-based Tukey's biweight correlation evidently displays distinguishing designs for the clusters amount generated at a distinct accumulative percentage of variations. Due to the clustering results' sensitivity towards the amounts of components which need to be obtained, the right components amount to maintain must be defined appropriately. In climatology studies especially the identification of patterns of rainfall, practically a higher number than merely two cluster partitions should be obtained for the purpose of describeing the patterns of rainfall variations [22]. Accordingly, two clusters were inadequate as it masked the actual data structure.

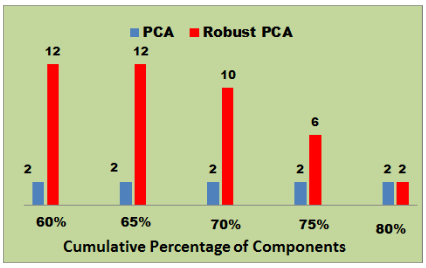

Figure 3. Bar chart depicts the amounts of clusters acquired from two approaches

For the evaluation of the solutions of cluster, the clustering output at 70% variation accumulative percentage on PCA-based Tukey's biweight correlation as well as Pearson correlation which are 10 clusters and 2 clusters correspondingly were selected. The evaluation of all of the clusters were evaluated contingent fundamental criteria of quality cluster that was recommended by [23]. The three criteria are external, internal and relative that are respectively through the Davies-Bouldin Index, Rand Index and Silhouette Index as represented in Table 2. For a parameter, a cluster of a high quality would have a lower Davies-Bouldin index value and a bigger Rand and Silhouette index value [24-25]. The Tukey's biweight correlation is illustrated in Table 2 which shows enhanced results of clustering, relatively, for three of the indices in comparison with the Pearson correlation.

Table 2. Measurement index for the clustering resuts quality

Correlation	Rand Index	Silhouette Index	Davies-Bouldin Index
Tukey's biweight	0.55	0.1	2.02
Pearson	0.53	0.04	4.78

5. CONCLUSION

This paper presents a modified correlation in PCA for torrential rainfall patterns identification. The employment of Tukey's biweight correlation that is contingent on PCA is recommended for leading the cluster solution of k-means clustering method in the identification of Peninsular Malaysia's torrential patterns of rainfall. It aims to present an alternative correlation matrix because of the concerns developed when managing non-Gaussian distributed data, specifically since the character of the data is skewed. This research illustrates that with PCA based Tukey's biweight correlation, the cluster partition displayed a significant compared to the Pearson's to prevent inaccuracy and unbalanced clusters within a high dimensional space.

ACKNOWLEDGEMENTS

This study was produced under the Fundamental Research Grants Scheme through FRGS Vote No. 2019-0132-103-02, (FRGS/1/2019/STG06/UPSI/02/4) offered by the Malaysian Ministry of Education.

REFERENCES

[1] I. T. Jollife and J. Cadima, "Principal Component Analysis: A review and recent developments", *Philosophical Transactions Society*, 374, 2016.

[2] S. M. Shaharudin et al., "Fitting statistical distribution of extreme rainfall data for the purpose of simulation", *Indonesian Journal of Electrical Engineering and Computer Science*, 18(3), pp 1367-1374, 2019.]

[3] S. Dan'azumi, S. Shamsudin, A. Aris, "Probability Distribution of Rainfall Depth at Hourly Time-Scale", *World Academy of Science, Engineering and Technology*. 4(12), pp. 670-674, 2010.

[4] A. A. Jemain and J. Suhaila, "Fitting the Statistical Distribution for Daily Rainfall in Peninsular Malaysia based on AIC criterion", *Journal of Applied Science Research*. 4(12), pp. 1846-1857, 2008.

[5] S. Yue, M. Hashino, "Probability Distribution of Annual, Seasonal and Monthly Precipitation in Japan", *Hydrological Science Journal*. 52(5), pp. 863-877, 2007.

[6] S. M. Shaharudin, "Spatial and temporal torrential rainfall guided cluster pattern based on dimension reduction method", Thesis, 2017.

[7] S. M. Shaharudin et al., "An efficient method to improve the clustering performance usin hybrid robust principal component analysis-spectral biclustering in rainfall patterns identification", *IAES International Journal of Artificial Intelligence (IJ-AI),* 8(3), pp. 237-243, 2019.

[8] S. M. Shaharudin et al., "The comparison of T-Mode and pearson correlation matrices in classification of daily rainfall patterns in Peninsular Malaysia", *Malaysian Journal of Industrial and Applied Mathematics*, EISSN 0127-9602, vol. 29, 2013.

[9] S. M. Shaharudin and N. Ahmad, "Improved Cluster Partition in Principal Component Analysis Guided Clustering", *International Journal of Computer Applications*, 75(11), pp. 1162-1167. 2013.

[10] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, New York, 1958.

[11] S. M. Shaharudin et al., "Identification of rainfall patterns on hydrological simulation using Robust Principal Component Analysis", *Indonesian Journal of Electrical Engineering and Computer Science,* 11(3), pp. 1162-1167, 2018.

[12] S. M. Shaharudin and N. Ahmad, "Cgoice of cumulative percentage in principal component analysis for regionalization of Peninsular Malaysia based on the rainfall amount", Asian Simulation Conference (AsiaSim 2017), *Modelling, Design and Simulation of systems,* pp. 216-224, 2017.

[13] S. Neware et al., "Finger knuckle identification using principal component analysis and nearest mean classifier", *International Journal of Computer Applications*, 70(9), 2013.

[14] S. M. Shaharudin and N. Ahmad, "Modeling, design and simulation systems", CCIS, 752: 216-224, Springer, Singapore. doi: 10.1007/978-981-10-6502-6_19. 2017.

[15] J. Hardin, "A robust Measure of correlation between two genes on a microarray", *BMC Bioinformatics*. 8(220). 2007.

[16] P. J. Rousseeuw and A. M. Leroy, "Robust regression and outlier detection", New Jersey: John Wiley & Sons, Inc. 2003.

[17] M. Owen, "Tukey's biweight correlation and the breakdown", Thesis, Pomona College, 2010.

[18] V. Choulakian, "Robust q-mode principal component analysis in L1", *Computational Statistics & Data Analysis,* 37, pp. 135-150, 2001.

[19] U. Maulik, "Performance evaluation of someclustering algorithms and validity indices", *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 24(12), 2002.

[20] G. M. Mimmack et al., "Choice of distance matrices in cluster analysis: defining regions", *Journal of Climate*, 14, pp. 2790-2797, 2002.

[21] M. S. Barrera et al., "PCA based on multivariate MM-estimators with fast and robust bootstrap", *Journal of American Statistical Association*, 101, pp. 1981-1211, 2006.

[22] J. A. Awan et at., "Identification and trend analysis of homogeneous rainfall zones over the East Asia monsoon region", *International Journal of Climatology*, 35, pp. 1422-1433, 2015.

[23] B. S. Everitt et al., *Cluster Analysis*, London: Arnold Publisher, 2001.

[24] A. Abraham et al., “Validation guideline for small scale classification result in medical domain”, *Hybrid intelligent system: 17th International Conference on Hybrid*, pp. 272-281, 2018.

BIOGRAPHIES OF AUTHORS

Shazlyn Milleana was born in Johor Bahru, Malaysia, in 1988. She is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI). She graduated with a bachelor science degree in Industrial Mathematics from Universiti Teknologi Malaysia, in 2010. Upon graduation, she began her career as an Executive in banking institution. In the following year, she received an offer to continue her study as a fast-track PhD student at the same university. During her PhD journey, she developed an interest in multivariate analysis, specifically in finding patterns which deals with big data. Her research focuses on the area of dimension reduction methods specifically in climate informatics which involves analysis on huge climate-related datasets based on techniques in Data Mining. She had published her research in Scopus indexed journal and presented her work in various local and international conferences. She completed her PhD thesis at the end of 2016 and was conferred a doctorate degree in 2017.

Norhaiza is a senior lecturer at the Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia (UTM). She graduated with an honors degree in Mathematics, Statistics and Operational research from the University of Manchester, in 1996. She joined UTM in August 2000. In the following year, she continued her studies at the University of Sheffield for her master’s degree. In Sheffield, she developed an interest in multivariate analysis, specifically in finding patterns which lead her to pursue a PhD degree at the University of Kent. She completed her PhD thesis at the end of 2007 and was conferred a doctorate degree in 2008. Finding patterns in any data have always been her research interests. She started the interests in profiling data – finding statistically distinctive and significant groups and features in the object of interest whilst at Sheffield. Currently, her research interests revolve around hydroinformatics particularly in investigating the streamflow variability of the local rivers.

Siti Mariana is a graduate of Bachelor Degree in Education (Mathematics) from Universiti Pendidikan Sultan Idris (UPSI) in 2018. She is currently pursuing her studies in Masters of Science Statistics and working on to publish papers in her scope of field. Her research focuses on the area of missing data and biclustering specifically in hydrology which involves analysis on huge climate-related datasets.