A New Singular Value Decomposition Based Robust Graphical Clustering Technique and Its Application in Climatic Data

Nishith Kumar (Corresponding author)
Department of Statistics, Begum Rokeya University, Rangpur Rangpur-5400, Bangladesh
Tel: 88-019-2520-0899 E-mail: nk.bru09@gmail.com

Mohammed Nasser
Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
Tel: 88-019-1425-4010 E-mail: mnasser.ru@gmail.com

Subaran Chandra Sarker
Department of Geography and Environmental Science, Begum Rokeya University
Rangpur-5400, Bangladesh
Tel: 88-016-7143-5416 E-mail: suba.jnu23@gmail.com

Received: March 24, 2011 Accepted: April 8, 2011 doi:10.5539/jgg.v3n1p227

Abstract

An attempt is made to study mathematical properties of singular value decomposition (SVD) and its data exploring capacity and to apply them to make exploratory type clustering for 10 climatic variables and thirty weather stations in Bangladesh using a newly developed graphical technique. Findings in SVD and Robust singular value decomposition (RSVD) based graphs are compared with that of classical K-means cluster analysis, its robust version, partition by medoids (PAM) and classical factor analysis, and the comparison clearly demonstrates the advantage of SVD over its competitors. Lastly the method is tested on well known Hawkins-Bradu-Kass (1984) data.

Keywords: Singular value decomposition, Robust singular value decomposition, Principal component analysis, Clustering, K-MEANS procedure, PAM method, Factor analysis

1. Introduction

Singular value decomposition, specially its low rank approximation property is an elegant part of modern matrix theory. After its inception (1936), its two ways fascinating data reduction capacity remained unnoticed till the last quarter of last century. Since then statisticians have been showing increasing interest to SVD for principal component analysis (PCA), canonical correlation analysis (CCA), cluster analysis and multivariate outliers detection. Principal component analysis (PCA), often performed by singular value decomposition (SVD), is a popular analysis method that has recently been explored as a method for analyzing large-scale expression data (Raychaudhuri *et al.*, 2000; Holter *et al.*, 2000; Alter *et al.*, 2000). To analyze the effect of the oceans and atmosphere on land climate, Earth Scientists have developed climate indices, which are time series that summarize the behavior of selected regions of the Earth's oceans and atmosphere. In the past, Earth scientists used observations directly where as , more recently, they are using eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices (Steinbach, M. , Pang-ning Tan, 2003, Maeng-Ki Kim, Yeon-Hee Kim, 2009). For statistical climate forecasting SVD is also used (Campbell, E. 2006). The Singular Value Decomposition can be applied to the analysis of climate data to identify patterns and maximize covariation (P.S. Lucio, F. C. Conde and A. M. Ramos 2007). A quasidecadal mode is isolated using singular value decomposition (SVD) (Mark R. Jury, 2009) applied to monthly smoothed and detrended rainfall, sea surface temperature (SST) (Wallace, J. M., Smith, C. and Bretherton, C. S. (1992)), sea level pressure (SLP). SVD can reduce data in both ways–columns and rows, and is more numerically stable. As PCA can be undertaken as a by product of SVD, in modern research it is being used more frequently in place of classical PCA for data compression (Diamantaras and Kung, 1996; Wall *et al.*, 2003), clustering (Sai Jayram and Murty, 2002; Murtagh, 2002; Chipman *et al.*, 2003) and multivariate outliers detection (Penny and Jolliffe, 2001).

The conventional approach to the SVD requires that the matrix X be complete. If it has any missing elements, the calculation as sketched cannot be performed. Also there may be outliers. To solve this problem, Douglas M. Hawkins, Li Liu, S. Stanley Young (2001) have been developed Robust Singular Value Decomposition (RSVD) using Alternating L_1 regression approach. In this article we mainly develop a graphical technique on the basis of SVD and RSVD, and apply to a data set of Bangladesh that contain 10 climatic variables of 30 principal weather stations with a view to clustering them, and compare its results with that of classical K–means cluster analysis, PAM method and classical factor analysis. We see that our exploratory technique serves the purpose of both the techniques- K-means cluster analysis, and classical factor analysis have shown that this same graph with a slight change could be effectively used for outliers detection both for supervised and unsupervised learning.

The article consists of ten sections. Section 1 describes importance of SVD, and objectives and layout of the article. Section 2 delineates nature and source of data, section 3 describes the partitioning around medoid (PAM) method, section 4 illustrate the data reduction capacity of SVD and section 5, robust singular value decomposition. Sections 6, 7, 8 and 9 uphold our exploratory method for clustering, results for clustering of weather stations and climatic variables, and evaluated and advantages of our proposed method respectively. The paper concludes with some general comments about our method in section 10.

2. Nature and source of data

Source. The source of this data is the report entitled "Land Resources Appraisal of Bangladesh for Agricultural Development", BGD/81/035, Technical Report 3, volume I, @ FAO 1988. This report was prepared for the Government of the People's Republic of Bangladesh, based on the work of H. Brammer (Agricultural Development Adviser), J. Antoine (Data Base Management Expert) and A. H. Kassam & H. T. van Velthuizen (Land resources and Agricultural Consultants) of UNDP (United Nations Development Programme).

Nature. This data set contains the values of 10 climatic variables for 30 principal stations. For 30 principal stations, data sets of average values of annually means of the following parameters were collated from the unpublished and published records of the Bangladesh Meteorological Department (BMD). The climatic variables are,

1. Rainfall (RF) mm
2. Daily mean temperature (T-MEAN)0C
3. Maximum temperature (T-MAX)0C
4. Minimum temperature (T-MIN)0C
5. Day-time temperature (T-DAY)0C
6. Night-time temperature (T-NIGHT)0C
7. Daily mean water vapor pressure (VP) MBAR
8. Daily mean wind speed (WS) m/sec
9. Hours of bright sunshine as percentage of maximum possible sunshine hours (MPS)%
10. Solar radiation (SR) cal/cm2/day

The principal stations are coded by the serial number and given in Table 1.

3. Partitioning around medoids method

Partitioning Around Medoids (PAM) method is the one of the robust method of clustering. This method is developed by Kaufman and Rousseeuw (1987). K-means algorithm, attempts to minimize the average squared Euclidean distance which is very sensitive to outliers and not applicable for scales other than interval scales. We have decided in favor of the k-medoid method because it is more robust with respect to outliers and it does not only deal with interval-scaled measurements but also with general dissimilarity coefficients. Dissimilarity coefficients between objects may be obtained from the computation of distances, such as the Euclidean distance, Manhattan distance etc. The objective of this method is to find clusters, the objects of which show a high degree of similarity, while objects belonging to the different clusters are as dissimilar as possible. The first step of this method is to find the number of cluster K. After finding K, the K clusters are constructed by assigning each object of the data set to the nearest representative object. The best K is selected on the basis of average silhouette width. The silhouette width of an object is obtained by using the formula, (A-B)/max(A,B), Where B is the average distance of the object to all other objects within its cluster and A is the average distance of the object to all objects in its nearest neighboring cluster. The silhouette width lies between -1 to +1. Choose the value of K that maximizes the average silhouette width (Kaufman and Rousseeuw, 1990).

4. Data reduction capacity of SVD

C. Eckart and G. Young proved low rank approximation of SVD (1936). Singular value decomposition (SVD) has a wonderful data reduction capacity (both ”R” and ”Q” modes) with minimum recovery error. We can use SVD to perform PCA. By using SVD we can reduce both variables and observations of a data matrix.

Suppose X is $m\times n$ matrix of rank $k \leq$ min (m,n). Then by singular value decomposition we can write,

$$X = U\Lambda V^T \tag{1}$$

Where U is the column orthonormal matrix whose columns are the eigen vectors of XX^T, Λ is the diagonal matrix contain the singular values of X and V is the orthogonal matrix whose columns are the eigen vectors of $X^T X$. Suppose we approximate X by $\widetilde{X}$ whose rank is $l< k \leq$ min (m,n).

By the singular value decomposition we can write

$$\widetilde{X} = U_l \Lambda_l V_l^T \tag{2}$$

Where U_l is $m\times l$, Λ_l is a diagonal matrix of order l and V_l is $n \times l$. Now post multiply V_l in both side of (2) we have $\widetilde{X}V_l = U_l\Lambda_l$.

Here the matrix $U_l\Lambda_l$ contains the principal component scores, its first column represents the first PC, and second column represents the second PC and so on. Hence we see that X is a $m\times n$ matrix but $\widetilde{X}V_l$ is a $m\times l$. If n represents no. of variables, it then reduces data by minimizing no. of variables. On the other hand $U_l^T\widetilde{X} = \Lambda_l V_l$ reduces data by minimizing no. of observations. Both ways data reduction capacity is fully utilized in the biplot (Gabriel, 1971).

5. Robust singular value decomposition

The conventional approach to the SVD requires that the matrix **X** be complete. The calculation of SVD cannot be performed, if **X** has any missing elements. Gabriel and Zamir (1979) addressed this problem. To solve this problem and handle missing element, Hawkins D. M. , Liu L. and Young S. S (2001) proposed robust singular value decomposition (RSVD) on the basis of alternating L1 regression approach. The flowchart of the algorithm of alternating L_1 regression approach for calculating robust singular value decomposition are given in Figure 1.

6. Proposed method for clustering data

Cluster analysis identifies and classifies objects individuals or variables on the basis of the similarity of the characteristics they possess. It seeks to minimize within-group variance and maximize between-group variance. The result of cluster analysis is a number of heterogeneous groups with homogeneous contents. There are substantial differences between the groups, but the individuals within a single group are similar. Data may be thought of as points in a space where the axes correspond to the variables. Cluster analysis divides the space into regions characteristic of groups that it finds in the data. The objectives of cluster analysis are discovering types and reducing the number of cases by enabling consideration of several types instead of numerous records. Clusters may present in the data. After getting any data set we need to see whether any clusters exists or not in the data. For clustering data we have proposed an exploratory method by the help of SVD. SVD is also used for outliers detection. Since we know that the first few principal components (PC’s) account most of the variation of data. So for graphical purpose we have used the first three PC’s. The proposed methods for clustering data using first three PC’s are given below

First we construct the scatter plots of first two PC’s, and first PC and third PC. We also make six lines in X-axis also in Y-axis by the following way *median (1st PC)* ± *k* × *mad (1st PC)* in the X-axis and *median (2nd PC/3rd PC)* ± *k* × *mad (2nd PC/3rd PC)* in the Y-axis. Where *mad* = median absolute deviation. The value of k = 1, 2, 3. The position of a point with respect to the other points in the graph and the boxes made by six vertical (horizontal) lines offer us to a rough view of possible clusters. Note that *median* and *mad* are the robust location and scatter measure respectively. Our method will be more appropriate and meaningful if variables or observations are correlated.

7. Clustering weather stations and climatic variable

7.1 Weather stations

In our standardized climate data if we apply our method by using SVD then it shows Figure 2. From Figure 2 we see that Cox’s Bazar, Chittagong, Maijdi Court and Hatiya make a cluster, so we can say that the climatic behavior of these four stations is similar. The weather of Sylhet is totally different from other stations. The climatic nature of Patuakhali, Barisal, Bhola, Ishurdi and Kaptai is same because they make a cluster. Srimangal, Dinajpur, Rangpur, Sirajganj and Pabna make another cluster and the rest of the stations make a single cluster so their climatic nature is similar.

In our standardized climate data if we apply our method by using RSVD then it shows Figure 3. From the Figure 3 we see that Cox's Bazar, Chittagong, Maijdi Court and Hatiya make a cluster, so we can say that the climatic behavior of these four stations are similar. The weather of sylhet is totally different from other stations. The climatic nature of Patuakhali, Barisal, Bhola, Satkhira and Kaptai is same because they make a cluster. Dinajpur, Rangpur, Sirajganj, Ishurdi and Pabna make another cluster also the climatic behavior of Srimangal is similar to this cluster. The rest of the stations make a single cluster so their climatic nature is similar

K-MEANS procedure. By the K-MEANS procedure for $K = 5$ we get,

Cluster-1: Rajshahi, Narayangang, Brahmanbaria,Comilla, Chandpur, Jessore, Khulna, Feni and Rangamati.

Cluster-2: Ishurdi, Faridpur, Satkhira, Barisal, Bhola, Kaptai and Patuakhali.

Cluster-3: Dinajpur, Rangpur, Bogra, Jamalpur, Mymensingh, Srimangal, Sirajganj and Pabna.

Cluster-4: Dhaka, Maijdi Court, Hatiya, Chittagong and Cox's Bazar.

Cluster-5: Sylhet.

PAM method. In partitioning around medoids (PAM) method the first step is the selection of K. Where K is the no. of cluster. The average silhouette width is used for K selection. Choosing that K for which the average silhouette width is maximum. In our standardized climate data, The maximum average silhouette width is 0.30 for $K = 5$. The result of PAM method for $K = 5$ is,

Cluster-1: Rajshahi, Narayanganj, Jessore, Khulna, Satkhira.

Cluster-2: Ishurdi, Faridpur, Barisal, Bhola, Kaptai, Patuakhali.

Cluster-3: Dinajpur, Rangpur, Sylhet, Srimangal, Sirajganj, Pabna.

Cluster-4: Bogra, Jamalpur,Mymensingh, Dhaka, Brahmanbaria, Comilla, Chandpur, Feni, Rangamati.

Cluster-5:Maijdi Court, Hatiya, Chittagong, Coxs Bazar

By the help of these methods we can say that the weather of south east area of Bangladesh is similar. The climatic behavior of north east area i.e., Sylhet is totally different from other stations. The climatic nature of the hem of south Bengal is similar. The stations of north Bengal shows the similar weather. The weather of rest of the stations is similar.

7.2 Climatic variables

In our Standardized climate data if we apply our method by using SVD then it shows Figure 4. Our climatic variables *rf*, *vp*, *tmean*, *tmax*, *tmin*, *tday*, *tnight*, *ws*, *mps* and *sr* are indicated by 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10. From Figure 4 we see that the variable rainfall (*rf*) differ from all other variables and the rest of the variables form three clusters minimum temperature (*tmin*) form a cluster; Hours of bright sunshine as percentage of maximum possible sunshine hours (*mps*) and Solar radiation (*sr*) make another cluster and the rest of the variables construct another group. K-MEANS procedure is not applicable for variables.

In our Standardized climate data if we apply our method by using RSVD then the Figure 5. From Figure 5 we see that the variable rainfall (*rf*) differ from all other variables and the rest of the variables form three clusters, minimum temperature (*tmin*) form a cluster; Hours of bright sunshine as percentage of maximum possible sunshine hours (*mps*) and Solar radiation (*sr*) make another cluster and the rest of the variables construct another group.

Factor analysis based. If the variables are uncorrelated in the population then the Factor Analysis (FA) does not useful. So we first measure the sample adequacy for our climate data using Bartlett test of sphericity and Kaiser-Meyer-Olkin (KMO) statistic. Table 2 represents the SPSS results of Bartlett test of sphericity and KMO statistic.

From Table 2 we see that the value of Bartlett test statistic is 483.963 and p-value is 0.00. Hence the null hypothesis that the variables are uncorrelated in the population is rejected i.e., the variables are not uncorrelated in the population. Again the value of KMO statistics is 0.635 which is greater than 0.50, which indicate FA is appropriate. Therefore both the Bartlett test of sphericity and KMO statistic supports to conduct FA in climate data.

Normality test. Here we test the univariate normality by the Kolmogorov-Smirnov (KS) test statistic. Then the KS values and p-values are given in the Table 3. From Table 3 we see that the normality assumption of *mps*, *sr* and *rf* are rejected even at 1% level of significance i.e., *mps*, *sr* and *rf* are not univariate normal. Also the normality assumption of *mps*, *tday*, *ws*, *sr* and *rf* are rejected at 5% level of significance. Since all the variables are not univariate normal so our climate data is not multivariate normal. we know that if any data set is multivariate normal then all the variables must be univariate normal but converse is not true. Hence we can say that our climate data is nonnormal. Thus in this data set principal factor solution method of estimation is more appropriate than Maximum likelihood (MLE) method. Principal factor solution method is also denoted by PFA (principal factor analysis) method. PFA method is stable and less affected by the outliers than MLE method.

PFA method. Scree plot helps us to select the number of factors to be extracted. Figure 6 shows the scree plot of the climate data. From Figure 6 we can select four factors for our analysis of climate data. Here first four components explain about 94% of the total variation in the data. By using PFA method we get the Table 4 which describes the rotated component matrix. Also Figure 7 shows the component plot in the rotated space. From Table 4 and Figure 7 we can see that the variables *vp*, *tmean*, *tmin*, *tday* and *tnight* with high positive loadings forms a group for the first factor. The variables *tmax* and *tday* with high positive loadings but *rf* with high negative loadings for the second component. Also the variables *mps* and *sr* with high negative loadings for the third factor and *ws* with high positive loadings for the fourth factor.

By the help of these methods we can say that the characteristics of climatic variable *rf* is totally different from other climatic variables. The climatic variables *tmin* make a group. Also *mps* and *sr* make another group and rest of the climatic variables make a different cluster.

8. Evaluate our proposed method by Hawkins-Bradu-Kass (1984) data

Hawkins, Bradu and kass (Hawkins *et al*., 1984) constructed an artificial three-predictor data set containing 75 observations with 14 influential observations. Among them there are ten high leverage outliers (cases 1-10) and four high leverage points (cases 11-14) (Imon 2005). The data set is mainly well known for checking efficacy of regression diagnostic as well as robust measures. If we apply our method in this data then it shows the Figure 8. From Figure 8 we see that first two PC's show the observations 1-14 are outside our box, it also shows that there are three clusters present in the data. Observations 1-10 make 1st cluster, observations 11-14 make second cluster and the rest observations make third cluster.

9. Advantages our method

The advantages of our method from other existing methods are given below

- It is easy to understand without hard mathematics.
- It can be applied to data for both supervised and unsupervised learning.
- It is directly applied to seperate different clusters.
- It can be applied in extremely complicated data sets without any extraneous assumption.
- It can be applied to locate outliers if they present in data.

10. Conclusion

Bangladesh is located in subtropical monsoon region. In context of Bangladesh, BANGLAPEDIA (National Encyclopedia of Bangladesh) states that Bangladesh has been divided into the following seven distinct climatic zones on the basis of mainly three climatic variables rainfall, temperature and winter dew.

(i) South -eastern zone (Chittagong sub-region)

(ii) North-eastern zone (east and south Sylhet)

(iii) Northern part of the northern region (Panchagarh, Lalmonirhat etc.)

(iv) North-western zone (Rangpur , Dinajpur etc.)

(v) Western zone (greater Rajshahi)

(vi) South-western zone (Jessore , Khulna etc.)

(vii) South-central zone (Dhaka, Cumilla, Mymensingh etc.)

Actually it's an adhoc method that does not consider all the climatic variables and their correlations. On the other hand we have proposed a more scientific method and compared it with some existing clustering technique. The result of our proposed method is very much similar with the all of these methods. From our above discussion we can say that SVD is an important exploratory tool for clustering for both variables and cases. If one wishes, one may go forward for more rigorous SVD based / K-means/other methods of clustering after this exploratory graph By help of our proposed method we can clustering any data if clusters present. In climate data we see that there are five clusters present in the stations and also the four clusters present in the variables. In this article we have used S-PLUS, R, SPSS and LaTex software and its packages.

References

Alter, O., Brown, P. O. & Botstein, D. (2000). Singular Value Decomposition for Genome-wide Expression Data Processing and Modeling. *Proc Natl Acad Sci USA 2000*, 97, 10101-10106. doi:10.1073/pnas.97.18.10101, http://dx.doi.org/10.1073/pnas.97.18.10101

Campbell, E. (2006). A Review of Methods for Statistical Climate Forecasting. *Technical Report 06/134*. [Online] Available:
http://www.cmis.csiro.au/techreports/docs/TechReport_06_134-ReviewofMethodsforStatisticalClimateForecasting.pdf

Chipman, H., Hastie, T. J. & Tibshirani, R. (2003). Clustering Microarray Data. In Terry Speed (eds.). *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, pp. 194-200. [Online] Available: http://folk.uib.no/nmaja/public/speed/c3278-CH04.pdf

Diamantaras, K. I. & Kung, S. Y. (1996). *Principal Components Neural Netwarks: Theory And Applications*. John wiely & sons, Inc. N.Y, pp.45-46.

Eckart, C. & Young, G. (1936). The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1, 211-218. doi:10.1007/BF02288367, http://dx.doi.org/10.1007/BF02288367

Hawkins, D. M., Bradu, D. & Kass, G. V. (1984). Location of Several Outliers in Multiple Regression Data Using Elemental Sets. *Technometrics*, 20, 197-208. doi:10.2307/1267545, http://dx.doi.org/10.2307/1267545

Hawkins, D. M., Liu, L. and Young, S. S. (2001). Robust Singular Value Decomposition. Technical Report 122, *National Institute of Statistical Sciences*. [Online] Available: http://www.niss.org/sites/default/files/pdfs/technicalreports/tr122.pdf

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. & Fedoroff, N. V. (2000). Fundamental Patterns Underlying Gene Expression Profiles: Simplicity From Complexity. *Proc Natl Acad Sci USA 2000*, 97, 8409-8414. doi:10.1073/pnas.150242097, http://dx.doi.org/10.1073/pnas.150242097

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000). Functional Discovery Via a Compendium of Expression Profiles. *Cell 2000*; 102, 109-126. doi:10.1016/S0092-8674(00)00015-5, http://dx.doi.org/10.1016/S0092-8674(00)00015-5

Imon, A. H. M. R. (2005). Identifying Multiple Influential Observations in Linear Regression. *Journal of Applied Statistics*, 32, 929-946. doi:10.1080/02664760500163599, http://dx.doi.org/10.1080/02664760500163599

Jayram, S. A. K. V. & Murty, M. N. (2002). Clustering for Prototype Selection using Singular Value Decomposition. In Jajuga, K., Sokolowski, A. and Bock, H. (eds.). *Classification,Clustering and Data Analysis*, Springer-Verlag, Berlin, pp.81-88. ISBN: 978-3-540-43691-1

Jury, M. R. (2009). A Quasi-decadal Cycle in Caribbean Climate. *Journal of Geophysical Research*, 114, D13102, 8 PP.

Kaufman, L. & Rousseeuw, P. J. (1987). Clustering by Means of Medoids. In Y. Dodge (eds.). *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, (Berlin: Birkhäuser), 405–416.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Willy, New York. (Chapter 2), pp. 68-125.

Kim, M. K. & Kim, Y. H. (2009). Seasonal Prediction of Monthly Precipitation in China Using Large-scale Climate Indices. *Advances in Atmospheric Sciences*, Springer-Verlag, 27, 47-59. doi:10.1007/s00376-009-8014-x, http://dx.doi.org/10.1007/s00376-009-8014-x

Lucio, P. S., Conde, F. C. & Ramos, A. M. (2007). Spatial Pattern Recognition of Extreme Temperature Climatology: Assessing HadCM3 Simulations Via NCEP Reanalyses Over Europe. *Revista Brasileira de Meteorologia*, 22(2), 204-217. doi:10.1590/S0102-77862007000200006, http://dx.doi.org/10.1590/S0102-77862007000200006

Murtagh, F. (2002). Clustering in High-dimensional Data Spaces. In Jajuga, K., Sokolowski, A. and Bock, H. (eds.). *Classification, Clustering and Data Analysis*, Springer-Verlag, Berlin, pp.89-96.

Penny, K. I. & Jolliffe, I. T. (2001). A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data. *Royal Statistical Society*, 50(3), 295–307. doi:10.1111/1467-9884.00279, http://dx.doi.org/10.1111/1467-9884.00279

Raychaudhuri, S., Stuart, J. M. & Altman, R. B. (2000). Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. *Pac Symp Biocomput*, 2000(5), 452-463.

Steinbach, M., Tan, P., Kumar, V., Clooster, S. & potter, C. (2003). *Discovery of Climate Indices Using Clustering*, ACM, New York, NY, USA. pp. 446-455, Year of Publication: 2003, ISBN:1-58113-737-0.

Wall, M. E., Rechtsteiner, A. & Rocha, L. M. (2003). Singular Value Decomposition and Principal Component Analysis. In D.P. Berrar, W. Dubitzky, M. Granzow (eds.). *A Practical Approach to Microarray Data Analysis*, 91-109, Kluwer: Norwell, MA (2003). LANL LA-UR-02-4001. [Online] Available: http://arxiv.org/ftp/physics/papers/0208/0208101.pdf

Wallace, J. M., Smith, C. & Bretherton, C. S. (1992). Singular Value Decomposition of Wintertime Sea Surface Temperature and 500-mb Height Anomalies. *J. Climate*, 5, 561-576. doi:10.1175/1520-0442(1992)005<0561:SVDOWS>2.0.CO;2, http://dx.doi.org/10.1175/1520-0442(1992)005<0561:SVDOWS>2.0.CO;2

Table 1. Serial number for principal stations

1. Dinajpur	2. Rangpur	3. Rajshahi	4. Bogra	5. Jamalpur
6. Mymensingh	7. Sylhet	8. Srimangal	9. Sirajganj	10. Ishurdi
11. Pabna	12. Dhaka	13. Narayanganj	14.Brahmanbaria	15. Comilla
16. Chandpur	17. Jessore	18. Faridpur	19. Khulna	20. Satkhira
21. Barisal	22. Bhola	23. Feni	24. Maijdi Court	25. Hatiya
26. Chittagong	27. Cox's Bazar	28. Rangamati	29. Kaptai	30. Patuakhali

Table 2. Results of Bartlett Test of Sphericity and KMO statistic

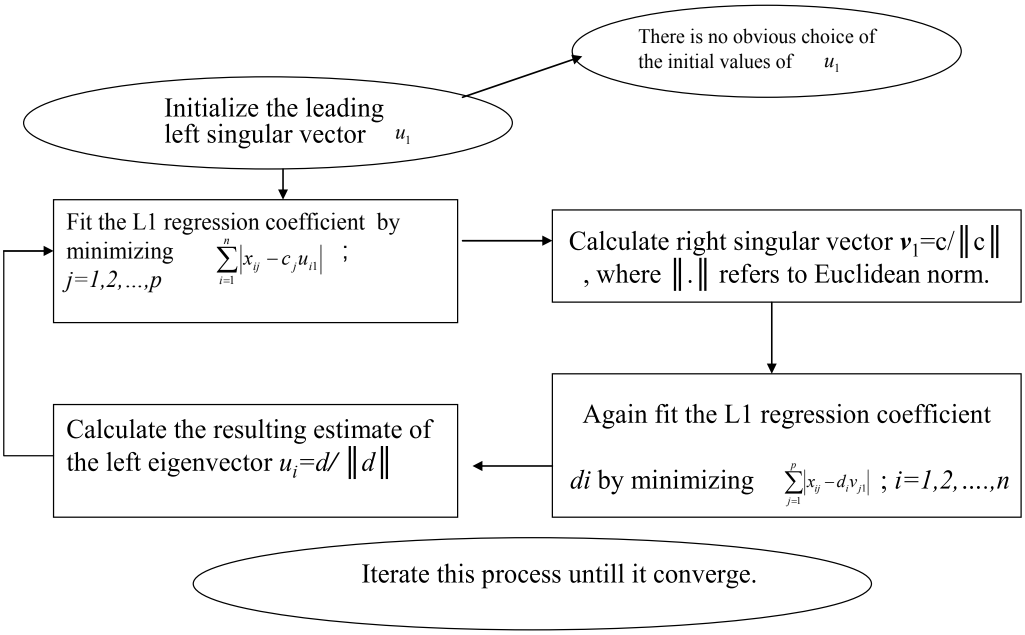
Kaiser-Meyer-Olkin (KMO) Measure of Sampling adequacy	Bartlett Test of Sphericity	
	Aprox. Chi-square	483.963
0.635	d.f.	45
	sig	0.00

Table 3. Table for KS-value and p-value

Variables	rf	vp	tmean	tmax	tmin	tday	tnight	ws	mps	sr
KS-value	0.195	0.121	0.1251	0.0845	0.1314	0.1824	0.1226	0.1649	0.2581	0.2262
p-value	0.0051	0.5	0.5	0.5	0.5	0.0121	0.5	0.0365	0	0.0004

Table 4. Table for rotated component matrix

Variables	Component			
	1	2	3	4
rf	0.090	**-0.897**	0.117	0.066
vp	**0.908**	-0.085	-0.141	0.071
tmean	**0.914**	0.302	0.070	0. 126
tmax	-0.009	**0.934**	-0.006	-0.186
tmin	**0.916**	-0.244	0.063	0.257
tday	**0.685**	**0.695**	0.068	0.037
tnight	**0.961**	0.003	0.072	0.221
ws	0.403	-0.229	0.041	**0.881**
mps	-0.078	-0.019	**0.991**	-0.042
sr	0.109	-0.075	**0.984**	0.089


There is no obvious choice of the initial values of u_1
Initialize the leading left singular vector u_1
Fit the L1 regression coefficient by minimizing $\sum_{i=1}^{n}\left|x_{ij}-c_j u_{i1}\right|$; $j=1,2,\ldots,p$
Calculate right singular vector $\boldsymbol{v}_1$=c/‖c‖ , where ‖.‖ refers to Euclidean norm.
Again fit the L1 regression coefficient
di by minimizing $\sum_{j=1}^{p}\left|x_{ij}-d_i v_{j1}\right|$; *i=1,2,….,n*
Calculate the resulting estimate of the left eigenvector u_i=*d/*‖*d*‖
Iterate this process untill it converge.

For the second and subsequent of the SVD, we replaced *X* by a deflated matrix obtained by subtracting the most recently found them in the SVD

$$X \leftarrow X - \lambda_k u_k v_k^T$$

Figure 1. Flowchart of the algorithm of alternating L_1 regression approach for Robust singular value decomposition

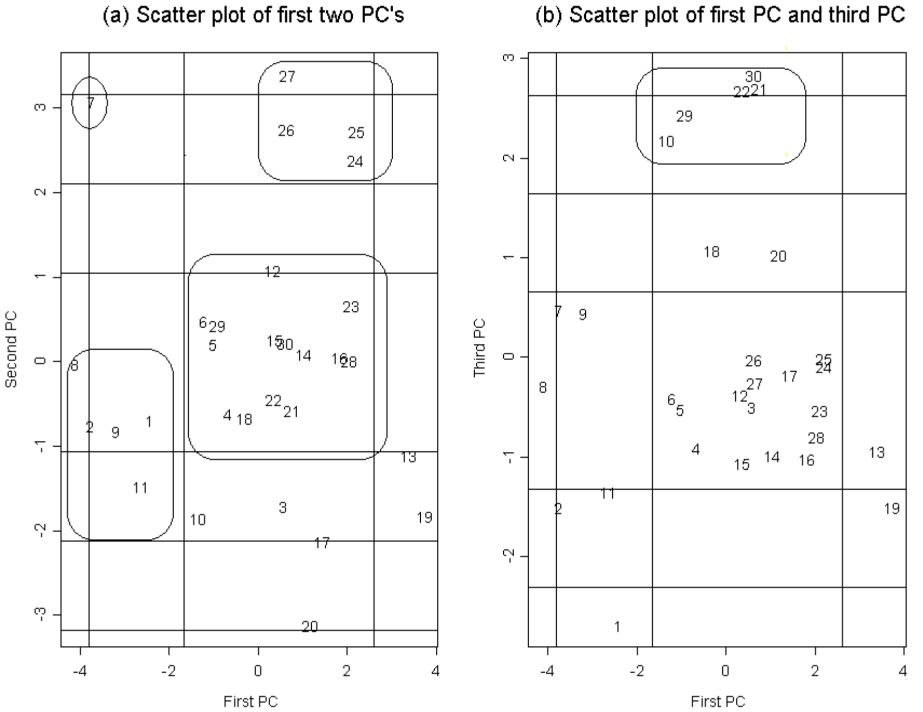


Figure 2. Scatter plot (a) scatter plot of first two PC's using SVD and (b) scatter plot of first and third PC using SVD (for stations)

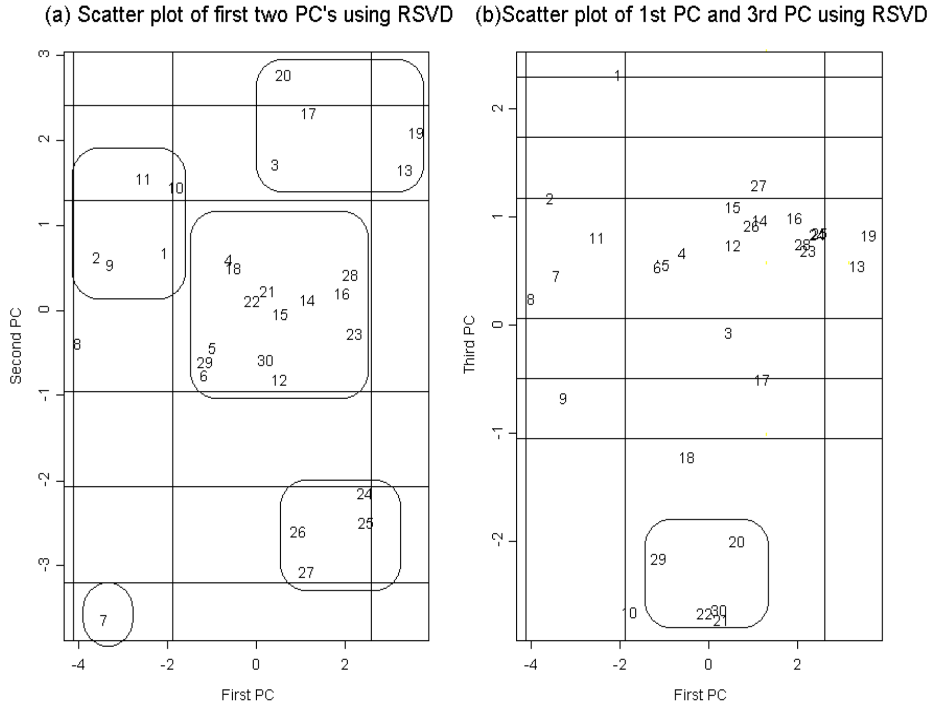


Figure 3. Scatter plot (a) scatter plot of first two PC's using RSVD and (b) scatter plot of first and third PC using RSVD (for stations)

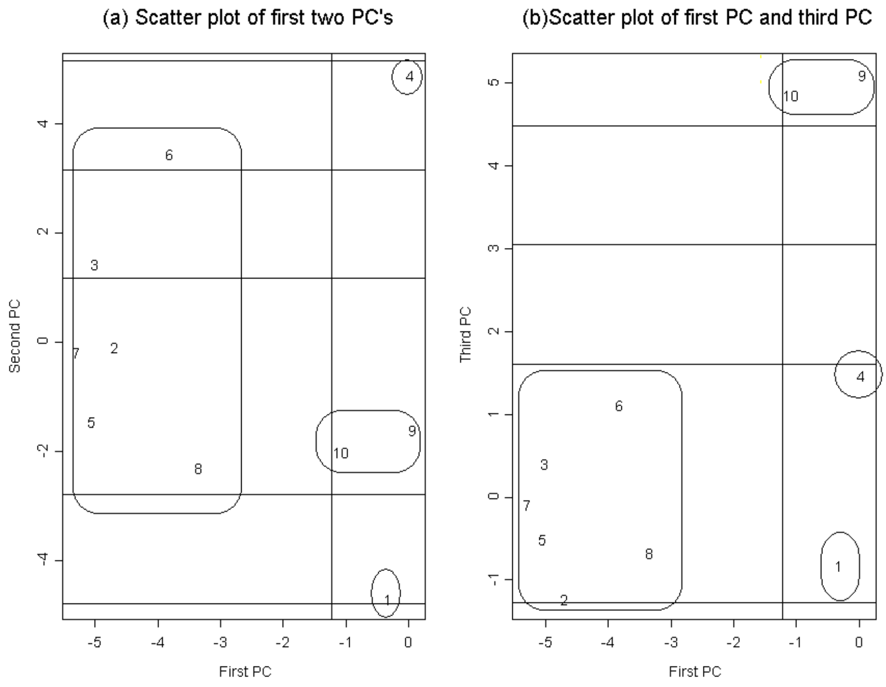


Figure 4. Scatter plot (a) scatter plot of first two PC’s using SVD and (b) scatter plot of first and third PC using SVD (for variables)

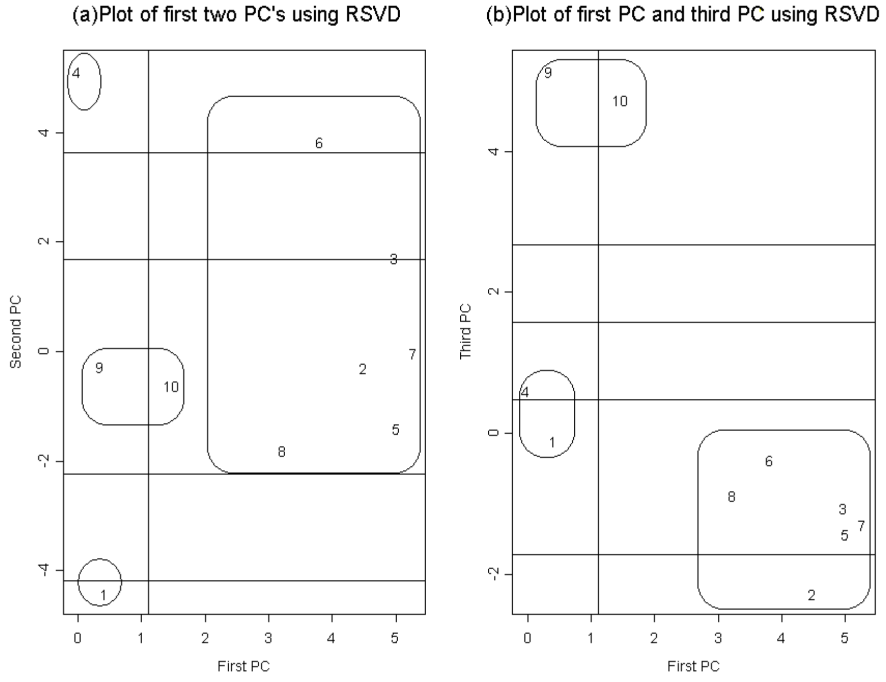


Figure 5. Scatter plot (a) scatter plot of first two PC’s using RSVD and (b) scatter plot of first and third PC using RSVD (for variables)

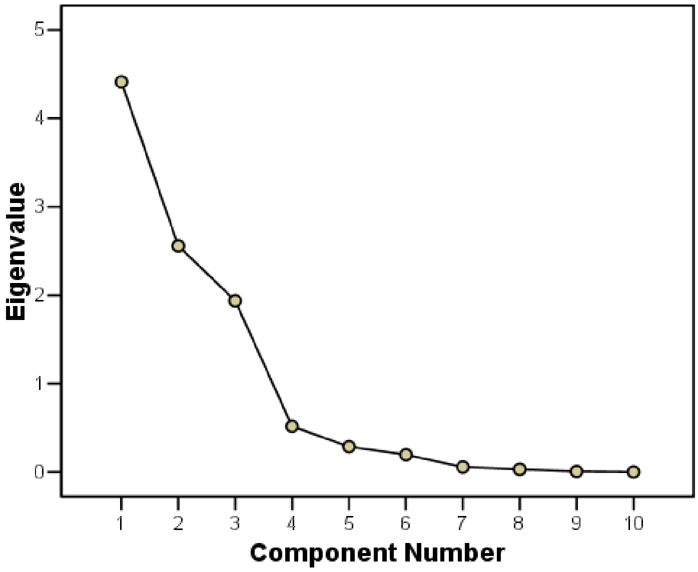


Figure 6. Scree plot of climate data

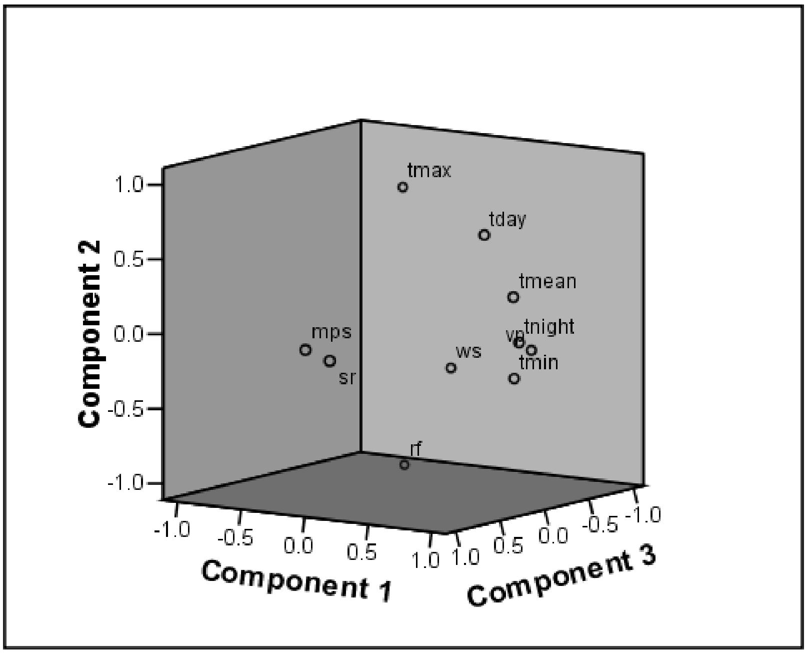


Figure 7. Component plot in the rotated space

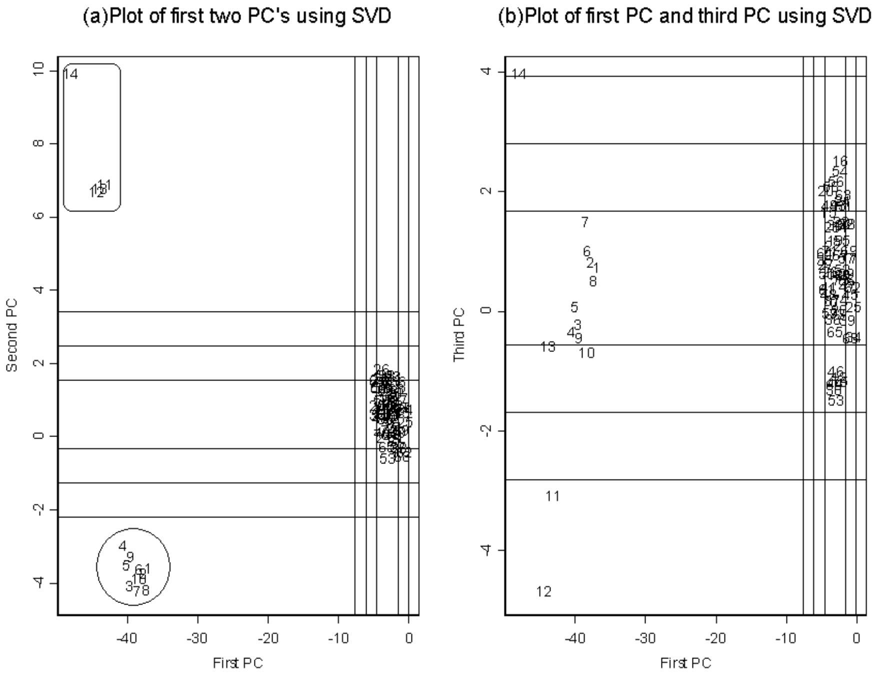


Figure 8. Scatter plot (a) scatter plot of first two PC's and (b) scatter plot of first and third PC using SVD

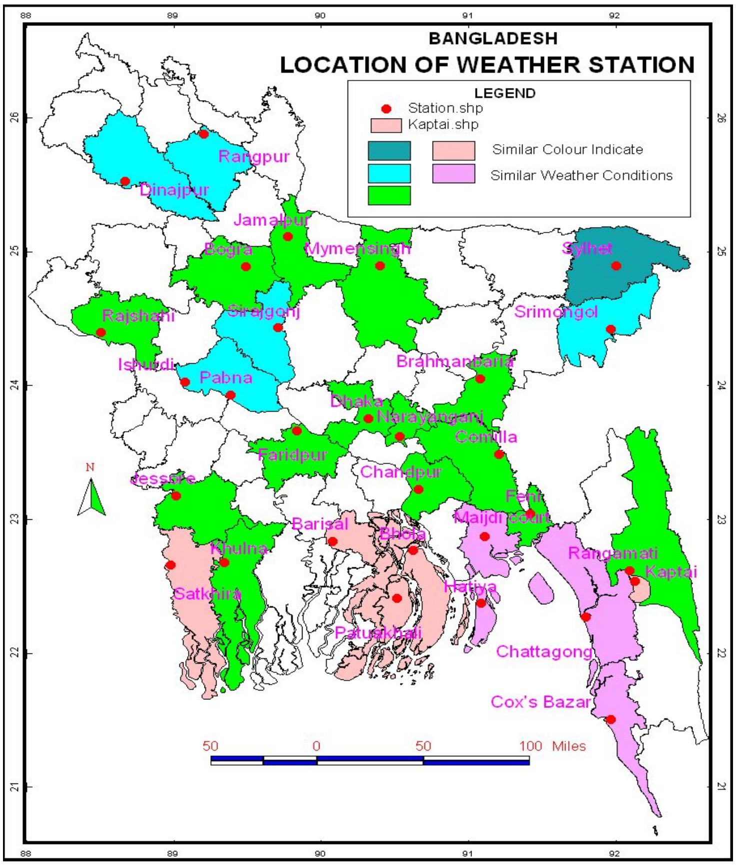


Figure 9. Clustering weather stations on map using RSVD