Data Analysis Using Principal Component Analysis

Shruti Sehgal[1] Harpreet Singh[2] Mohit Agarwal[3] V. Bhasker[4] Shantanu[5]

[1, 3, 5] Electronics and Communication Engineering
Amity School of Engineering & Technology
Noida (U.P)-201303, India
shrutiaquarius90@gmail.com

[2]SSPL, Ministry of Defence
R & D Organization
Delhi-110054, India
harpreetgujral86@yahoo.co.in

[4]Department of Physics and Astrophysics
University of Delhi
Delhi-110007, India
bhasker84@gmail.com

***Abstract*— In this paper, we have evaluated an algorithm using Principal Component Analysis (PCA) for its application in data analysis. In the research field, it is very difficult to understand the large amount of data and is very time consuming too. Therefore, in order to avoid wastage of time and for the ease in understanding we have scrutinized a PCA algorithm that can reduce the huge dimension of the data into 2-dimensional. The method of PCA is used to compress the maximum amount of information into first two columns of the transformed matrix known as the principal components by neglecting the other vectors that carries the negligible information or redundant data. The main objective of the paper is to separate two compounds say A and B having different concentrations for all four sensors and identifies which sensors have the similar or different concentration with the help of various plots that explains the correlation between the different variables.**

***Keywords*— Data Analysis, Eigen, PCA**

I. INTRODUCTION

In Principal Component Analysis (PCA) is one of the pattern recognition techniques and one of its applications is to analyse the high dimensional data that is not easy to understand by just looking at the large amount of data. For data analysis, I need to reduce the high dimension of the data into low dimension and then making a plot and interpret the results. PCA is used to present the important information into few simple plots namely score plot and loading plot. In the field of research, it is very difficult to analyze large amount of data. PCA algorithm is used to compute relation between the huge correlated data set [1] [2].

In linear algebra, PCA has its mathematical algorithm that explains the correlation between the data containing the variables as columns and observations or samples as rows. The goal of the PCA algorithm is to reduce the large correlated variables into small number of variables. These correlated variables are called as principal components [3]. The main motive is to establish a matrix that contains the maximum amount of information in first two columns and then project the data using 2-dimensional plot in MATLAB software [4].

II. ALGORITHM

We have studied an algorithm in which the various variables are having different correlated variables and the main objective is to separate two different compounds say A and B having different concentrations for four sensors. The step-by-step PCA algorithm given in fig 1 is implemented in MATLAB software. We have felt that to interpret the huge dimensional data is not easy to plot and for the same reason we have to reduce the dimension of the data first from 11-dimensional to 2-dimensional and then plot the principal components in 2-dimensional loading plot to understand the relation between concentrations of four different sensors [5] [6].

The steps of the PCA Algorithm are given below:

- We start with the data set 'A' which is in the form of matrix of dimension m x n, where m rows represent the variables whereas n columns represent the samples i.e. observations. We will now linearly transform this matrix into another matrix 'B' of the same dimension m x n, so that for some matrix Z given by equation (1).

$$B = Z * A \tag{1}$$

Where, Z is another matrix of dimension m x m.

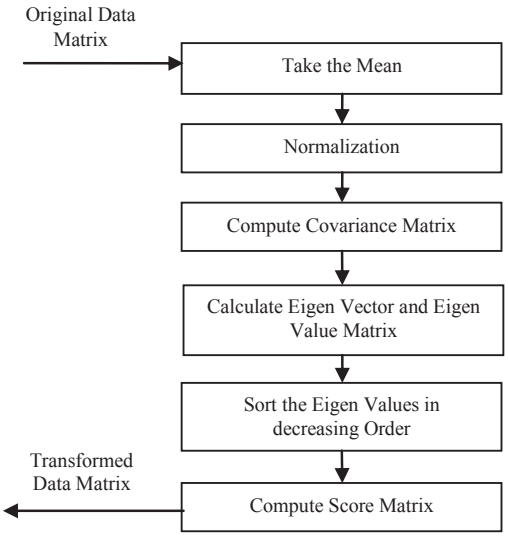

Fig. 1 Flow chart for PCA Algorithm

- Normalization is the important part of the algorithm in which we need to calculate the mean of the original data matrix and subtract off the mean for finding principal components as given in equation (2) [7].

$$Mean\ (m) = 1/N \sum_{n=1}^{N} A[m,n] \quad (2)$$

- Compute Covariance matrix of A which will be of dimension m x m in equation (3).Here, each element of covariance matrix C_A represents all possible pair of covariance. In fact all diagonal elements represent variance and the non diagonal elements of the matrix are covariance.

$$C_A = A * \frac{A^T}{(n-1)} \quad (3)$$

- We are required to decide some features that the transformed matrix 'B' should exhibit which relates to the features of corresponding covariance matrix C_B. It should have minimum covariance and maximum variance. Small variance may be redundant data, therefore we need to maximize the variance and minimize the covariance.
- Therefore, we have to decide the value of Z such that covariance matrix, C_B becomes a diagonal matrix. Therefore, in equation (4) (5),

$$C_B = B * \frac{B^T}{(n-1)} = \frac{(ZA)(ZA)^T}{(n-1)} = \frac{(ZA)(A^T * Z^T)}{(n-1)} \quad (4)$$

$$C_B = \frac{Z(AA^T)Z^T}{(n-1)} = \frac{ZY * Z^T}{(n-1)} \quad (5)$$

Where, $Y = A$. Here Y is of dimension m x m.

- We can represent the matrix Y in the form given in equation (6). Since every square matrix is orthogonally (orthonormally) diagonalizable.

$$Y = EDE \quad (6)$$

- Now E is an m × m orthonormal matrix whose columns represent orthonormal Eigen vectors of Y, and D is a diagonal matrix which has the Eigen values of Y as its (diagonal) entries.
- Here, we decide the value of the transformation matrix Z. We make the rows of the Z to be the Eigen vectors of Y. Therefore, in equation (7),

$$Z = E^T \quad (7)$$

- In equation (8), (9), using the value of Z, we get

$$C_B = \frac{ZY * Z^T}{(n-1)} = \frac{E^T(ED * E^T)E}{(n-1)} \quad (8)$$

$$C_B = \frac{D}{(n-1)} \quad (9)$$

- The Eigen values are arranged in the descending order, since the largest value of Eigen value tells the relative importance of the corresponding principal component as shown in fig 3.
- Similarly, Eigen vector in the matrix E must be arranged according to their respective Eigen values in the diagonal matrix given in matrix in section III.
- Our motive is to plot only first two columns or principal components of the final matrix that carries the maximum amount of information using MATLAB Software.

III. DATA ANALYSIS

The goal of our paper work is to analyse the data of dimension 11x4 as given in the matrix below in Table I. We have 4 different sensors having different concentrations say C1, C2, C3.....and C11 as plotted in fig 4. There are two compounds say A and B, we have taken 4 different concentrations for Compound A and 7 different concentrations for Compound B. Our objective is to separate two different compounds depending on their value of concentrations.

Table I Original data used for data analysis

Different Compounds Concentration		Four Different Sensors			
		Sensor1	Sensor2	Sensor3	Sensor4
Compound A	C1	503	55	22	41
	C2	672	58	38	63
	C3	427	33	32	48
	C4	161	18	15	4
Compound B	C5	638	15	45	20
	C6	103	1	22	9
	C7	106	6	23	4
	C8	116	10	16	2
	C9	108	1	16	4
	C10	635	63	17	8
	C11	312	25	19	8

The second step is to evaluate the normalized matrix of dimension 11x4 by subtracting the mean from the original matrix of dimension 11x4 as given below in equation (10). The data is normalized so that we can easily compute the variance.

$$\begin{bmatrix} 159.2727 & 29.0909 & -2.0909 & 21.8182 \\ 328.2727 & 32.0909 & 13.9091 & 43.8182 \\ 83.2727 & 7.0909 & 7.9091 & 28.8182 \\ -182.7273 & -7.9091 & -9.0909 & -15.1818 \\ 294.2727 & -10.9091 & 20.9091 & 0.8182 \\ -240.7273 & -24.9091 & -2.0909 & -10.1818 \\ -237.7273 & -19.9091 & -1.0909 & -15.1818 \\ -227.7273 & -15.9091 & -8.0909 & -17.1818 \\ -235.7273 & -24.9091 & -8.0909 & -15.1818 \\ 291.2727 & 37.0909 & -7.0909 & -11.1818 \\ -31.7273 & -0.9091 & 5.0909 & -11.1818 \end{bmatrix} \quad (10)$$

The third step gives the reduced covariance matrix of dimension 4x4 that can also be computed by simply multiplying the original data matrix with the transpose of the original data matrix given below in equation (11).

$$\begin{bmatrix} 5.6779 & 0.4505 & 0.1531 & 0.3392 \\ 0.4505 & 0.0537 & 0.0043 & 0.0316 \\ 0.1531 & 0.0043 & 0.0099 & 0.0138 \\ 0.3392 & 0.0316 & 0.0138 & 0.0457 \end{bmatrix} \quad (11)$$

The Eigen vector matrix computed as given below in equation (12) that shows the relation between the uncorrelated variables. The diagonal matrix is carried out and sorted in the decreasing order in equation (13). The diagonal values are the Eigen values that are taken out of the matrix placed in a

column and each Eigen value carries the how much of the variance is contained by the principal components and arrange it in the decreasing order.

$$\begin{bmatrix} 0.9947 & -0.0873 & -0.0316 & 0.0441 \\ 0.0792 & 0.3857 & 0.8030 & -0.4473 \\ 0.0268 & 0.0553 & -0.5078 & -0.8593 \\ 0.0598 & 0.9168 & -0.3103 & 0.2442 \end{bmatrix} \quad (12)$$

$$\begin{bmatrix} 5.7383 & 0 & 0 & 0 \\ 0 & 0.0275 & 0 & 0 \\ 0 & 0 & 0.0211 & 0 \\ 0 & 0 & 0 & 0.0004 \end{bmatrix} \quad (13)$$

The final step of the algorithm gives the calculation of final score matrix in equation (14) that has the maximum information contained in the first two columns known as principal components PC1 and PC2 that are arranged according to their amount of variance in the decreasing order. And hence the last two columns say PC3 and PC4 can be neglected that has a very small amount of information that can be a redundant data [8].

$$\begin{bmatrix} 161.9812 & 17.2050 & 12.6208 & 1.1337 \\ 332.0678 & 24.6636 & -5.2609 & -1.1351 \\ 85.3280 & 22.3238 & -9.8944 & 0.7406 \\ -183.5372 & -1.5209 & 8.7494 & -0.4123 \\ 292.4600 & -27.9905 & -28.9309 & 0.0825 \\ -242.0897 & 1.9557 & -8.1754 & -0.1590 \\ -238.9818 & -0.9061 & -3.2115 & -4.3435 \\ -229.0252 & -2.4569 & 3.8598 & -0.1653 \\ -237.5759 & -3.3966 & -3.7354 & 3.9959 \\ 291.8089 & -21.7631 & 27.6515 & -0.3893 \\ -32.4361 & -8.1140 & 6.3269 & 0.6519 \end{bmatrix} \quad (14)$$

IV. RESULTS & DISCUSSIONS

The data is analysed using PCA algorithm in which we can interpret that out of four sensors, one sensor is having less association as compared to other three sensors as clearly shown in fig 2.

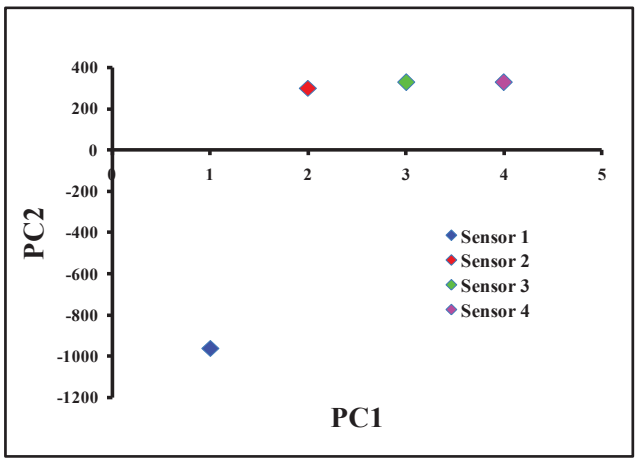

Fig. 2 Loading plot for 4 different sensors

Fig 2 depicts the *loading plot* for the four different sensors that is a column vector 2-dimensional plot used to present the information about the sensors. As we can see that sensor 2, sensor 3 and sensor 4 are close to each other and have very high correlation as compared to sensor 1. The sensor 1 is on the negative side of the origin have negative correlation. Also, the sensors on the same side of origin are having similar compound concentrations. We have identified the variables that are close to each other having a very high correlation and also the variables that are far apart from each representing the negative correlation [9].

In fig 3, we have plotted the cylindrical graph for the Eigen values known as the *Eigen value spectrum* that provides a relationship between the Eigen values and Eigen vector number. Eigen vector number is the total number of Eigen values that are four in number for the given data matrix. All Eigen values are greater than one [10].

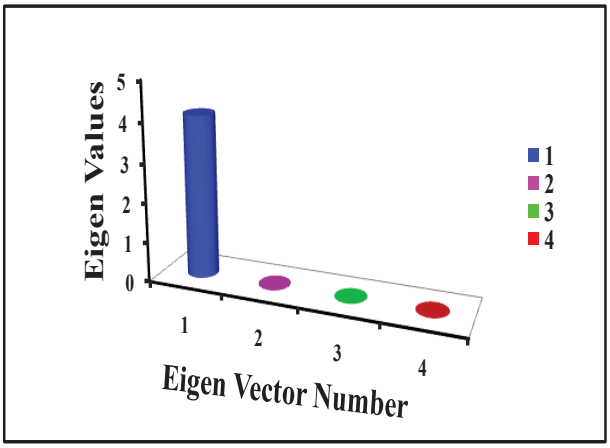

Fig. 3 Eigen value Spectrum

In this paper, we can understand the information more precisely by plotting the simple graphs. Like in fig 4, we have drawn a 2D *Score plot* that has clearly separated out the two different compounds say, A and B.

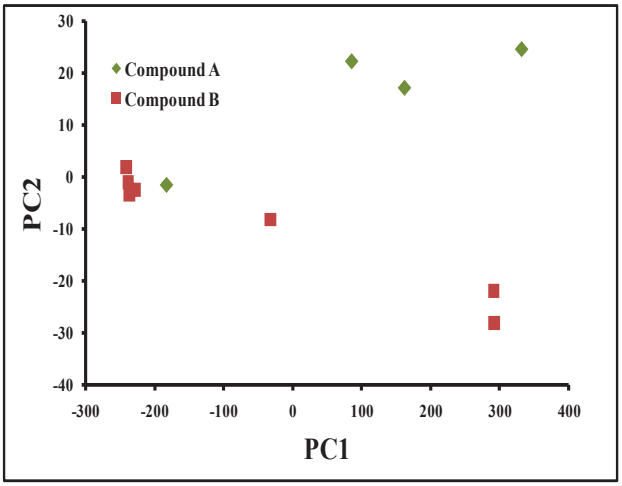

Fig 4 2D Score plot

We can see that the concentrations of Compound A are having their concentrations apart from each other for C1, C2, C3 and C4 explaining that the variables are having different concentrations and they are dissimilar. Whereas, the few concentration of Compound B for C6, C7 and C9 are showing similar trend as shown in Table I. And, also C5 and C10 are giving the same correlation [11].

V. CONCLUSION

We proposed an algorithm that is used to reduce the dimension of the original data carried out for data analysis from 11-dimensions to 2-dimensional data set. The goal of the algorithm is to limit the maximum information only in the first two columns called as principal components and neglect the rest of the columns carrying the negligible amount of information. To reduce the dimensionality of the data, I have used PCA that gives better results. Moreover, I can have a clear vision of the correlation between the different variables when they are represented in the 2D plot in MATLAB software [12] [13].

VI. FUTURE PLANS

Our future work is to learn other method of pattern recognition like Artificial Neural Network (ANN) for analysing the large observations. There are various applications of PCA that can be implemented for future prospects using Statistical Package for the Social Sciences (SPSS) software [14] [15].

ACKNOWLEDGMENT

Authors would like to thankfully acknowledge Mr. M. U. Sharma and Surface Acoustic Wave (SAW) group team members from Solid State Physics Laboratory, DRDO where the work has been carried out and to our friends who are always there for the help and kind support.

This is to acknowledge that this paper contains no confidential information related to the work at SSPL, DRDO.

REFERENCES

[1] Stojanovic, Branka, and Aleksandar Neskovic. "Impact of PCA based fingerprint compression on matching performance." In Telecommunications Forum (TELFOR), 2012 20th, pp. 693-696. IEEE, 2012.

[2] Beltran, Luis A. "Nonparametric multivariate statistical process control using principal component analysis and simplicial depth." PhD diss., University of Central Florida Orlando, Florida, 2006.

[3] Shinde, R. L., and K. G. Khadse. "Multivariate process capability using principal component analysis." Quality and Reliability Engineering International 25, no. 1 (2009): 69-77.

[4] Bell, Anthony and Sejnowski, Terry. (1997) “The Independent Components of Natural Scenes are Edge Filters.” Vision Research 37(23), 3327-3338.

[5] Levinson, William A. Statistical process control for real-world applications. CRC Press, 2011.

[6] PCA and LDA in DCT domain ,Weilong Chen, Meng Joo Er *, Shiqian Wu Pattern Recognition Letters 26 (2005) 2474–2482

[7] Roweis, S. "EM Algorithms for PCA and SPCA." In Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems. Vol.10 (NIPS 1997), Cambridge, MA, USA: MIT Press, 1998, pp. 626–632.

[8] Rafael Gonzalez and Richard Woods. Digital Image Processing. Addison Wesley, 1992.

[9] W.J Krzanowski. Between-Groups Comparison of Principal Components Journal of American Statistical Association, 74(367):703-707, 1979.

[10] Krzanowski, W. J. Principles of Multivariate Analysis. Oxford University Press, 1988.

[11] Seber, G. A. F. Multivariate Observations. Wiley, 1984.

[12] Jackson, J. E., A User's Guide to Principal Components, John Wiley and Sons, 1991, p. 592.

[13] Ilin, A., and T. Raiko. "Practical Approaches to Principal Component Analysis in the Presence of Missing Values."J. Mach. Learn. Res.. Vol. 11, August 2010, pp. 1957–2000.

[14] Bishop, Christopher ‘Neural Networks for pattern Recognition’, Clarendon, Oxford, UK, 1996.

[15] Chakra Chennubhotla and Allan D. Jepson. Sparse PCA: Extracting multi-scale structure from data International Conference on Computer Vision, Vancouver, July 2001.