World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:8, 2017

Principle Components Updates via Matrix Perturbations

Aiman Elragig, Hanan Dreiwi, Dung Ly, Idriss Elmabrook

***Abstract*—This paper highlights a new approach to look at online principle components analysis (OPCA). Given a data matrix $X \in \mathbb{R},^{m\times n}$ we characterise the online updates of its covariance as a matrix perturbation problem. Up to the principle components, it turns out that online updates of the batch PCA can be captured by symmetric matrix perturbation of the batch covariance matrix. We have shown that as $n \to n_0 \gg 1$, the batch covariance and its update become almost similar. Finally, utilize our new setup of online updates to find a bound on the angle distance of the principle components of X and its update.

***Keywords*—Online data updates, covariance matrix, online principle component analysis (OPCA), matrix perturbation.

I. Introduction

ONLINE learning has become an urge in many applications [5], [6]. As opposed to batch settings, in online learning we do not have an access to the whole data we are intended to learn from. Instead, we receive data points sequentially, one by one, and the ultimate goal is to determine an adaptive model capturing the overall trend of the whole data.

Many reserachers have dedecated much of their work to develop online algorithms, using different approaches, to cope with such sequential updates of data [9]. However, most of the algorithms developed in the context of online learning are essentially based on tweaking the already known models for batch data, i.e., they become able to learn from data on-the-fly [7], [8], [18]–[20]. Despite of all evident successes in developing online models [1], [12], there is, however, a need to a rigorous theory which can be used to reveal some aspects, the exiting techniques fails to clarify.

One of the most important online learning problems, which has received tremendous amounts of investigation, is the online principle components analysis (OPCA) [2], [10]–[12]. Rougly speaking, PCA aims to extract the main modes of varition of the data around their mean through computing new variables called the principle compoeents [9]. PCA has found applications in various fields such as face regonition [12], lating sememtic indexing [16], sentiment analysis [13], industerial process modeling [14], astronomy [15], to name but few. Expensive computation is one of the main major challenges in using PCA technique as it often requires approximation which resulting in prediction errors.

This paper is a step forward to better understand online PCA update mechanism providing a new view to look at online updates of data. This paper is organised as follows. Section II will give some required mathematical concepts. Section III presents a detailed description of PCA as a technique with some exploration of the currant related literatures. Section IV shows a formulation of online updates of data as a symmetric matrix perturbation. Section V provides the main result of the paper. In Section VI we use some matrix perturbation results to illustrate the significance of our setup.

II. Mathematical Preliminaries

In this section we will briefly some basic mathematical concepts. For a vector $v \in \mathbb{R}^n$, its Euclidean norm is given by $||v|| := \sqrt{v^T v}$. For a matrix $A \in \mathbb{R}^{m\times n}$, the he spectral norm of A is defined as $||A|| = \max\{||Av|| : ||v|| = 1\}$. The transpose of A is another matrix obtained by exchanging its row with its columns and its often denoted by A^T. The matrix A is said to be symmetric if, and only if $A = A^T$. The singular value decomposition of the matrix A is given by

$$A = U\Sigma V^T,$$

where U is an $m \times m$ orthonormal matrix , Σ is diagonal matrix, and V is an $n \times n$ orthonormal matrix. The diagonal entries of the diagonal matrix Σ. $\sigma_1, \sigma_2, \ldots, \sigma_n$ are known as the singular values of A. For a matrix $S \in \mathbb{R}^{n\times n}$, the scalar λ is an eigenvalue of S if there exists a non-zero vector v such that $Av = \lambda$. The vector v is referred to as the eigenvector of S corresponding to the eigenvalue λ. If S is symmetric the we can write

$$S = U\Lambda U^T,$$

where U is an $m\times m$ orthonormal matrix and Λ is a diagonal matrix whose entries are called the eigenvalues of the matrix S. Conventionally we often list the eigenvalues of S in a non-increasing order namely, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. It is worth to add here that

$$||S|| = \sigma_1 = \lambda_1.$$

III. Principle Component Analysis

Principle components analysis (PCA) is a very common method for reducing the dimensionality of data [23]–[26]. To have an idea about PCA we first discuss a related

A. Elragig. Aiman is with the Faculty of SciTech., Bournemouth University (UK), Also he is a lecturer at the University of Benghazi, Libya (e-mail: aelragig@bournemouth.ac.uk).

H. Dreiwi is a lecturer with the Department of Mathematics, faculty of science, University of Benghazi, Libya (e-mail: ahs2412006@yahoo.com).

D. Ly, is a postdoctorate and research assistant with the Faculty of SciTech., Bournemouth University, UK (e-mail: dly@bournemouth.ac.uk).

I. Elmabrook is a Professor in Applied Mathematics with the Department of Mathematics, faculty of science, University of Benghazi, Libya.

concept namely, the covariance matrix. Given the data set of m-dimensional data points x_i. We can defined the matrix

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}.$$

Lets assume that each columns in the matrix X corresponds to a data point, whereas each row refers to all measurement of a particular value x_i. Now the covariance matrix of the data X can be defied as follows.

$$\mathrm{Cov}(X) = \frac{1}{n-1} X X^T$$

Its easy to see the following properties of Cov(X)

- The main diagonal entries of Cov(X) show the variance of particular measurement features.
- The off-diagonal entries of Cov(X) represent the covariance between the measurement types.
- Cov(X) is square symmetric $m \times m$ matrix

The matrix Cov(X) gives all correlations between all pairs of data in X. The principle components of the data X are the eigenvectors of the covariance matrix Cov(X). In other words they are the singular vectors of the matrix X.

Often, PCA is performed in batch mode in whcih all the training data are ready for processing (Computing PCA). In the batch setting there will be no information to be used in laearning once all the whole available data are trained [27]. In contrast, in online learning data arrive sequentially and in every time instance we experience new data and that require an adaptive PCA algorithm to cope with such setting.

Typically PCA is implemented by using either the eigenvalue decomposition of the covariance matrix Cov(X) or the SVD of the data matrix X, after centring the data.

Its worth to mention here that computing SVD requires $O(nm\min(n, m)$ floating point operations [22]. In small sized data this will be of a reasonable accepted complexity, however, batch PCA is indeed infeasible in many setting with

- massive datasets for which n and m are, say, in the thousands or millions
- datasets that change rapidly and may need to be processed on the fly for instance, databases and streaming data.

In dimensionality reduction, for instance, we require to project the data into a low dimensional subspace and this can be easily handled using the SVD of the data matrix X. The dimension of the choose subspace, say k, should be small enough so that it reduces the data significantly and in the same time it should be large enough to retain the main variation characteristics of the data. In mathematical terms the best rank-k (X_k) representation of the matrix X is given by

$$X_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Updating the above decomposition as new data arrived is computationally exepensive given the exponential growth of data in modern applications. Hence effecint (i.e., accurate and fast) PCA algorithms are in high demand. Over the years a large number of solutions has emerged from fields. A numerical resolution to what is so- callked *secular equations* is one of the ekigant cinoutation of PCA coputations [28]–[30]. In this approach it has been shown that updating the PCA is equivalent of finding the zeros of a rational function. In [31], PCA updates has been formulated as a sum of the eigenvalue decomposition of a diagonal matrix and a rank-1 matrix. Incremental SVD is one of the most significant approach for having an approximate low rank PCA [32]. Incremental PCA (IPCA) has been around for more than a decades and have shown efficient performance in various applications [2], [32], [34]. The candid covariance-free incremental PCA (CCPCA) developed in [2] is one of the most recent and effective PCA algorithm which can compute the principle components of a stream of samples incrementally without estimating the covariance matrix. For a detailed literture on online pca we refer the readre to [2], [9]

IV. Data Streams as Perturbations

Let's assume that X be a data matrix with m rows and n columns. Also assume that X is formed by stacking data points column by column. In online setting, X will be updated with new data points, i.e., new columns. So, we assume that X is updated with the data point (vector) $x_{m\times 1}$. Accordingly, we have the following new data matrix

$$\bar{X} = \begin{pmatrix} X_{m\times n} & x_{m\times 1} \end{pmatrix}_{m\times(n+1)}$$

Now the covariance matrix of the new data matrix can be obtained as follows

$$\frac{1}{n}\bar{X}\bar{X}^T = \frac{1}{n}\begin{pmatrix} X & x \end{pmatrix}\begin{pmatrix} X \\ x \end{pmatrix}^T = \frac{n-1}{n}\left(\frac{1}{n-1}XX^T\right) + \frac{1}{n}xx^T$$

In other words we can write

$$\mathrm{Cov}(\bar{X}) = \frac{n-1}{n}\mathrm{Cov}(X) + \frac{1}{n}xx^T \tag{1}$$

Looking at Equation (1), it is straightforward to see that

- The matrices $\frac{n-1}{n}$Cov(X) and Cov(X) share the same eigenvectors.
- if λ is an eigenvalue of Cov(X) then, $\frac{n-1}{n}\lambda$ is an eigenvalue of Cov(X). Moreover,

  $$\lim_{n\to n_0 \gg 1} \sigma(\mathrm{Cov}(\bar{X})) = \sigma(\mathrm{Cov}(X)),$$

  where σ stands for a matrix spectrum.
- $\frac{1}{n}xx^T$ is symmetric matrix with a norm equals to $\dfrac{||x||^2}{n}$

To make a concrete insight for Equation (1), we consider the case when $m = 2$, i.e., a two dimensional feature space. Figure 1 depicts Equation (1) through its action on an arbitrary vector v (which can be an eigenvector of Cov(X)). Its clear that the eigendata of Cov$(\bar{X})$ eventually converges to the ones of Cov(X) as we are exposed to a sufficient amount of the data stream. The word "sufficient" suggests imposing certain threshold on the number of data above which new data arrivals

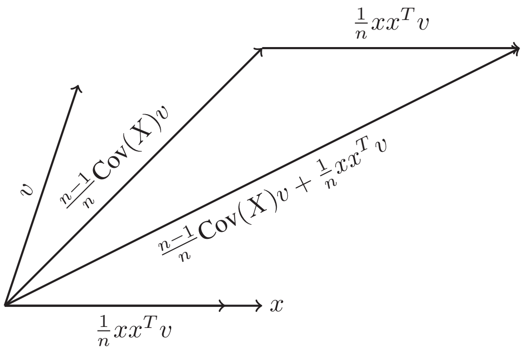
$\frac{1}{n}xx^Tv$
v
$\frac{n-1}{n}\text{Cov}(X)v$
$\frac{n-1}{n}\text{Cov}(X)v + \frac{1}{n}xx^Tv$
x
$\frac{1}{n}xx^Tv$

Fig. 1 The action of Cov($\bar{X}$) on v (the hypotenuse) is a decomposition of the action of the batch covariance added to the projection of v in the direction of the new data vector x, weighted by the reciprocal of the position of x on the queue of the data stream

would be ignored. This in turn could reduce the time and space complexity of any algorithm using PCA as one of its underling components.

V. The Main Result

Given a data matrix $X \in \mathbb{R}^{m\times n}$ and a date point x then

1. *Up to the principle components, the online updates of the batch data Cov(X) with the data stream point $x_{(m\times 1)}$ can be viewed as*

$$Cov(\bar{X}) = Cov(X) + \Delta, \quad where \quad \Delta = \frac{1}{n}xx^T. \qquad (2)$$

2. *We have*

$$\lim_{n\to n_0\gg 1} Cov(\bar{X}) = Cov(X).$$

VI. Some Applications of Proposition V

In this section we make use of some results from symmetric perturbation theory to characterise some bounds on the inclination of the updated first principle component of Cov($\bar{X}$), and its corresponding eigenvalue. The analysis is identical for other principle components.

1. Bound on the Eigenvalues
Following Weyl Theorem (See, Appendix **B**), and considering equation (1) we have

$$\lambda_1\left(\text{Cov}(\bar{X})\right) = \lambda_1\left(\text{Cov}(\bar{X}) = \frac{n-1}{n}\text{Cov}(X) + \Delta\right)$$

$$\leq \frac{n-1}{n}\lambda_1\left(\text{Cov}(X)\right) + \lambda_1(\Delta)$$

$$= \frac{n-1}{n}\lambda_1\left(\text{Cov}(X)\right) + \frac{||x||^2}{n}.$$

As the smallest eigenvalue of xx^T is zero, we can also show that

$$\lambda_1\left(\text{Cov}(\bar{X})\right) \geq \frac{n-1}{n}\lambda_1\left(\text{Cov}(X)\right),$$

thus we have

$$\frac{n-1}{n}\lambda_1\left(\text{Cov}(X)\right) \leq \lambda_1\left(\text{Cov}(\bar{X})\right) \leq \frac{n-1}{n}\lambda_1\left(\text{Cov}(X)\right) + \frac{||x||^2}{n}$$

The above inequity says that the principle components of the updated data will be always greater or equal to the batch one!

2. Bound on Eigenspaces Inclination
Assume that $\mathbf{V}_1$ and $\mathbf{V}_2$ are the subspaces generated by any corresponding principle components of Cov(X) and Cov($\bar{X}$). Therefore, based on Davis-Kahan Theorem (See, Appendix **B**), the angle distance θ between the eigenspace generated by the principle components of X and $\bar{X}$ is bounded by some value depending on the length of the new added data point x. Namely, we have

$$\sin 2\theta \leq \frac{2\|\Delta\|}{\delta} = \frac{2||x||^2}{n\delta}$$

where δ is gap between the eigenvalues of Cov(X).
It is clear that for large enough n, the two principle components will coincide, meaning that there will be no significant change in the maximum direction of variation of data. This observation is really an interesting! As it shows that the deviation of the existing principle component will not exceed the value $\frac{||x||^2}{n\delta}$. As we have mentioned in Section IV this could be used to infer a threshold while ruining PCA-based algorithms, that is, we only process certain amount of the data and ignore the rest.

VII. Experiments

In this section we have used the *iris* data set (see, [17]) to show the convergence of the dominant eigenvalue of Cov($\bar{X}$) to the one of Cov(X) as we experience large amount of data (i.e., $n \gg 1$). In Figure 2, the left subplot shows this convergence, whereas the right subplot depicts the bound on the deviation of the first components.

VIII. Conclusion

In this paper we contextualized online updates of data as a matrix perturbation problem. Up to principle components, we showed that online updating of data is equivalent to perturbing a scalar multiple of the batch covariance matrix by a symmetric matrix. It has been shown that as $n \to n_0 \gg 1$, the updated covariance matrix will become almost similar to the batch one. Our setup encourages the use of a wide range of matrix perturbation theoretic results in the context of online learning. Symmetric perturbation, in particular, is one of the simplest classes of matrix perturbations which has characteristics that makes it very useful to reveal many aspects of online learning.

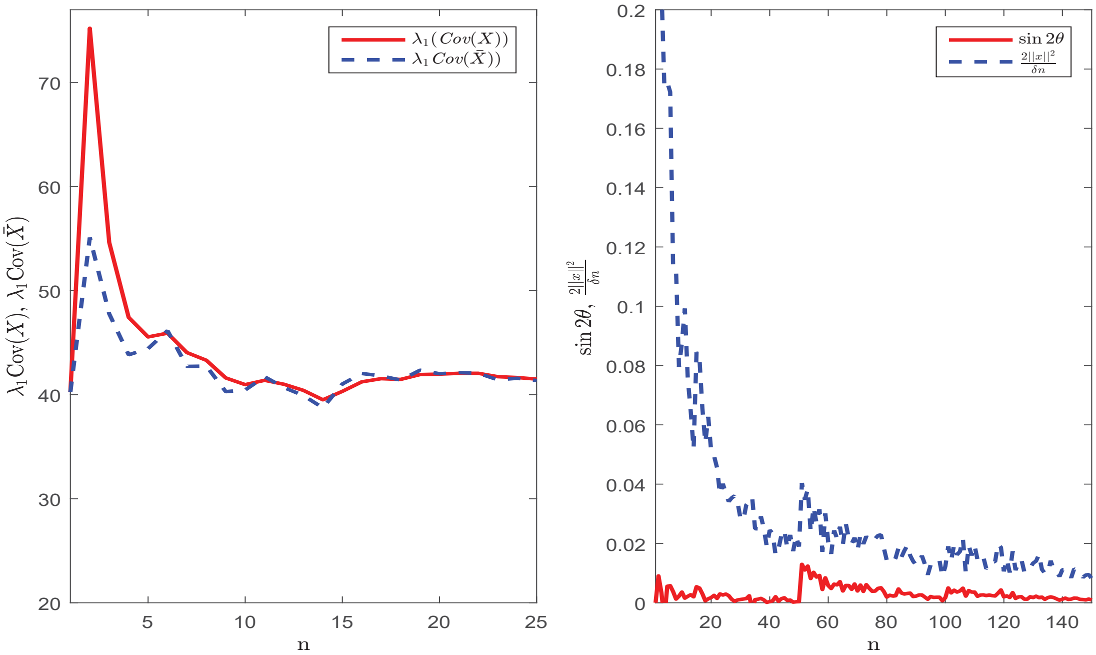

Fig. 2 The left subplot show the convergence of λ_1 Cov$(\bar{X})$ (dashed blue) to λ_1 Cov(X) (solid red) while the right shows the bounds on the deviation of principle components as we experience more data

Appendix

A. Weyl's Theorem [4]

Assume that A and B are two Hermitian matrices. Assume that their eigenvalues are ordered in a increasing order as follows.

$$\lambda_{min} = \lambda_n(A) \leq \lambda_{n-1}(A) \leq \lambda_{n-2}(A)$$

$$\ldots \leq \lambda_2(A) \leq \lambda_1(A) = \lambda_{max},$$

$$\lambda_{min} = \lambda_n(B) \leq \lambda_{n-1}(B) \leq \lambda_{n-2}(B)$$

$$\ldots \leq \lambda_2(B) \leq \lambda_1(B) = \lambda_{max},$$

$$\lambda_{min} = \lambda_n(A+B) \leq \lambda_{n-1}(A+B) \leq$$

$$\lambda_{n-2}(A+B) \ldots \leq \lambda_2(A+B) \leq \lambda_1(A+B) = \lambda_{max},$$

then for each $k = 1, 2, \ldots n$, we have

$$\lambda_k(A) + \lambda_n(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_1(B). \qquad \square$$

B. Davis-Kahan, [3]

Let X and Δ be symmetric matrices. Assume that $X = Q\Lambda Q^T = Qdiag(\lambda_i)Q^T$ and $\bar{X} = X + \Delta = \tilde{Q}diag(\tilde{\lambda}_i)\tilde{Q}^T$. Also let's write

$$Q = (q_1, q_2, \ldots, q_n) \quad \text{and} \quad \tilde{Q} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_n).$$

Let θ_i denote the acute angle between the q_i and $\tilde{q}_i$ then,

$$\sin 2\theta_i \leq \frac{2\Delta}{\min_{i\neq j}|\lambda_i - \lambda_j|}$$

provided that $\min_{i\neq j}|\lambda_i - \lambda_j| \neq 0$.

References

[1] J. Weng, Y. Zhang, W.S. Hwang, Candied covariance-free incremental principle component analysis. *IEEE Trans. Pattern Anal. Mach. Intel. Res* 9 (2003) 2287-2320.

[2] Y. Zhang, J. Wang, Convergence Analysis of Complementary candid incremental principle analysis. *Technical Report MSU-CSE-01-23, Department of Computer Science and Engineering, Michigan State University, East Lansing, MI.*

[3] G. Stewart, J. Sun, Matrix Perturbation Theory. *New York: Academic Press*,1990

[4] R. A. Horn and C. Johnson. Matrix Analysis. *Cambridge University press*, 1990.

[5] Fontenla-Romero, scar, et al. *Online machine learning." Efficiency and Scalability Methods for Computational Intellect* 27 (2013).

[6] Provatas, Spyridon. *An online machine learning algorithm for heat load forecasting in district heating systems.* (2014).

[7] H. Wang, Pi. Daoying, S. Youxian, Online SVM regression algorithm-based adaptive inverse control. *Neurocomputing* 70(4) (2007) 952-959.

[8] W. Barbakh, C. Fyfe, Online clustering algorithms. *International Journal of Neural Systems* 18(03) (2008) 185-194.

[9] H. Cardot, D. Degras, Online Principle Component Analysis in High Dimension: Which Algorithm to Choose?,Technical report *arXiv:*1511.03688 (2015).

[10] Z. Karnin, E. Liberty, Online PCA with Spectral Bounds, *JMLR: Workshop and Conference Proceedings* 40(2015)1-12.

[11] A. Balsubramani, S. Dasgupta, Y. Freund, The fast convergence of incremental pca. *Advances in Neural Information Processing Systems* 26(2013)3174-3182.

[12] H. Zhao, P. Chi, J.T. Kwok, A novel incremental principle components analysis and its application for face recognition. *IEEE Transactions on Systems. Man. Cybernetics-Partt B: Cybernetics* 36(4)(2016) 873-886.

[13] A. D. Iodice, A. Markos, Low-dimensional tracking of association structures in categorical data. *Stat. Comput.* 25(5)(2015)1009-1022

[14] J. Tang, W. Yu, T, Chai, L. Zhao, Online principle component analysis with applications to process modeling. *Neurocomputing* 82(2012)167-178.

[15] T. Budavari, V. Wild, A. Dobos, C. Yip, Reliable eigenspectra for new generation surveys. *Numer. Math.*, 394(2009)1496-1502

[16] H. Zha, H. D. Simon, On upadating problems in latent sementing indexing, *AIAM J. Sci. Comput,*, 21(2)(1999)782-791.

[17] R. A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7 (2) (1936) 179188.

[18] J Nie, W Kotlowski, MK Warmuth, Online PCA with optimal regret, *Journal of Machine Learning Research* 17(173) (2016) 1-49.

[19] D Garber, E Hazan, T Ma, Online learning of eigenvectors. *In International Conference on Machine Learning* (2015) 560-56.

[20] A. Allahyar, HS. Yazdi, Online discriminative component analysis feature extraction from stream data with domain knowledge, *Intelligent Data Analysis* 18(5) (2014) 927-951.

[21] M. Herbster, S. Pasteris, M. Pontil, Predicting a switching sequence of graph labelings, *Journal of Machine Learning Research* 16 (2015) 2003-2022.

[22] G. H. Golub, C. F. Van Loan, Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. *Johns Hopkins University Press*, Baltimore, MD,2003, fourth, edition.

[23] J. B. Tenenbaum, V. De Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500)((2000) 2319-2323.

[24] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. *science*, 313(5786) (2006) 504-507.

[25] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001) 245-250.

[26] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500) (2000) 2323-2326.

[27] O. Fontenla-Romero, B. Guijarro-Berdinas, D. Martinez-Rego, B. Perez-Sanchez, D. Peteiro-Barral, Online machine learning. *Efficiency and Scalability Methods for Computational Intellect*, 27 (2013).

[28] G. H. Golub, Some modified matrix eigenvalue problems. *SIAM Review*, 15 (1973) 318334.

[29] M. Gu, S. Eisenstat, A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem, *SIAM J. Matrix Anal. Appl* 15 (1994)12661276.

[30] W. Li, H. Yue, S. Valle-Cervantes, S. Qin, Recursive PCA for adaptive process monitoring, *J. Process Control*, 10 (2000) 471486

[31] A. Hegde, J. Principe, D. Erdogmus, U. Ozertem, Y. Rao, H. Peddaneni, Perturbation-based eigenvector updates for on-line principal components analysis and canonical correlation analysis, *J. VLSI Sign. Process.*, 45 (2006) 8595.

[32] J. Tang, W. Yu, T. Chai, L. Zhao, On-line principal component analysis with application to process modeling. *Neurocomputing*, 82 (2012) 167178.

[33] T. D. Sanger, Optimal unsupervised learning in single-layer linear feedforward neural network, *Neural Netw.* 2(1989) 459-473.

[34] E. Oja, J. Karhunen, On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1) (1985) 69-84.

Open Science Index, Computer and Information Engineering Vol:11, No:8, 2017 waset.org/Publication/10007847