A Review of Modified PCA Techniques

[1]Shivani, [2]Anuradha Bharti, [3]Nishant Behar
[1]student, [2]student ,[3]Assistant Professor
[1]Computer Science Engineering,
[1]Institute Of Technology, Guru Ghasidas University, Bilaspur, India

Abstract: The need to effectively utilize the information in an era of increasing data volumes will easily overwhelm the interpreter. However, recent developments in the field of pattern recognition provide a means to analyze multiple attributes in a single volume. There has been extensive research in pattern recognition techniques. Principal Component Analysis (PCA) is one of the most widely used unsupervised statistical algorithm. PCA has applications in all areas of statistics and machine learning including noise filtering, gene data analysis, stock market predictions, clustering, dimensionality reduction, face recognition, image compression, and visualization. In this paper, all the research work of the recent two decades to modify PCA is reviewed, and all modified techniques are supposed to remove the constraints of standard PCA.

Index Terms **– PCA, Pattern-Recognition.**

I. INTRODUCTION

In the real world, we have multidimensional data. As the dimensions of data increases, it is very difficult to visualize and perform computations on data. So, we reduce the dimensions of data by removing redundant dimensions and taking into consideration the most important dimensions. PCA is an important technique to understand the field of statistics and data science. While using machine learning techniques on multidimensional data, it is recommended to use any dimensionality reduction techniques like PCA because the high dimension of data can make learning algorithm too slow then to speed it up using PCA is the most common application of PCA. Reduction in the dimension of dataset reduces the problem of overfitting and by simplifying the dataset gives us a descriptive and better visualization of data.

II. LITERATURE SURVEY

In 1901, PCA was given by Karl Pearson. Afterward, there have been several efforts in the implementation of PCA in different areas as a dimensionality reduction technique.
IAN T. JOLLIFFE [1] has focused on the misconception about principal components having small eigenvalues will be of no use in regression, and generally, most of the practitioner uses to ignore components of small variances. He demonstrated that these components could be as useful as those with large eigenvalues. To make the computations easier and stable some of the authors Kendall (1957), Hotelling (1957) suggested to use principal components in regression and thus orthogonalizing the regression problem. He contradicted some of the authors like Mansfield et al. (1977, p-38), Gunst and Manson (1980), Mosteller and Turkey (1977, p. 397-398), Hocking (1976, P.31) concerning the rule for deciding principal components with high variance should be kept in regression. In support of low variance, he used the example provided by Kung and Sharif(1980) for the prediction of monsoon onset date from 10 meteorological variables in which tenth principal component is the third most important variables in the regression equation which was just 1 percent of the variability in original data.

Libin Yang [14] in his thesis performed risk analysis on the Australian stock market using the ASX200 index and its components from April 2000 to February 2014 using principal component analysis. He constructed a portfolio based on the principal components. He concluded that the variance explained by the first principal component could serve as a leading indicator of the financial crisis. Partovi and Caputo (2004) first proposed the idea of constructing principal portfolios and motivated researchers along with market practitioners in portfolio management. He found out that a good combination of stock is required to represent the whole data set. Before him, other researchers assumed all combination of stocks are, but they are not. It provides a way for investors to select stocks that will give a proper variation.

Recently Jordan Skaro [16], has used PCA as a post hoc method to reduce crosstalk errors which is a widely observed phenomenon while conducting motion analysis using 3-D motion capture technology. In his study, he concluded that PCA could correct for crosstalk in elliptical training, Gait and Cycling exercises. He recommended adopting the use of a PCA corrected axes set determined from Gait to produce PCA corrected angles.

III. PCA (PRINCIPAL COMPONENT ANALYSIS)

PCA is a mathematical procedure of extracting significant variables (in the form of components) from a large set of variables available in a dataset. It extracts a low dimensional set of features from a high dimensional dataset by keeping as much information as possible. To reduce the dimensions of a d-dimensional dataset, PCA will project it onto a k-dimensional subspace ($where,\ k < d$). With fewer dimensions, visualization, and computation become much more meaningful. PCA is much more meaningful. PCA is more useful when dealing with higher dimensional data.

PCA algorithm steps:

1. Standardize the data.
2. Compute the eigenvectors and eigenvalues from the covariance matrix or correlation matrix, or perform singular vector decomposition.
3. Sort eigenvectors in descending order and choose the x eigenvectors that correspond to the x largest eigenvalues where x shows the number of dimensions of the new feature subspace ($x \leq d$).
4. Compute the projection matrix W from the selected k eigenvectors.
5. Transform the original dataset X via W to obtain a k-dimensional feature subspace Y.

PCA seek for a linear combination of variables such that the maximum variance is extracted from the variables [20][21][22]. And most importantly PCA is always applied on a square symmetric matrix [20][21][22].
Thus, PCA as an analytical approach combines:
•How each variable is associated with one another (Covariance- Matrix).
•The directions in which our data are dispersed (Eigen Vectors).
•The relative importance of these different directions (Eigen Values).

IV. MODIFIED-PCA

1. Malika Heeenaye Mamaode Khan, Naushad Mamode Khan and Raja K. Subramanium [11], while working on their project of feature extraction of dorsal vein pattern, they used the PCA algorithm based on Cholesky matrix decomposition and Lancoz technique to reduce the processing time of pattern recognition. This modified technique reduces the number of computation and hence the processing time decreases. To test the result, they used a database of 200 images along with a threshold value of 0.9 to obtain the false acceptance rate and false rejection rate. This fast and modified PCA algorithm is recommendable when developing biometric security system since it significantly decreases the matching time.

TABLE I [11]

Number of Images	PCA (sec)	MPCACL (sec)	Diff
100	1400	650	-750
80	1130	310	-820
60	843	350	-493
40	560	265	-295
20	278	118	-160
10	135	52	-83

From the table, we can conclude that the average time taken for processing one image by PCA is around 13.5 sec whereas modified PCA took 5.2 sec for one image.

2. XU and Yuille [2] in the year 1995 worked together to make PCA algorithms more robust when outliers are present as in real world, most data are contaminated by some outliers, but conventional PCA algorithm is designed to work on data free from outliers. They observed that due to outliers, the performance of the PCA algorithm is reduced significantly. To make the PCA algorithm robust in the presence of outliers in data, they have a statistical physics approach. Robust rules proposed by them works well in the presence of outliers along with other PCA like tasks such as finding a first principal component vector, first k principal component vector without solving separately. According to comparative experiments of PCA and robust PCA, the results of improved performance in the presence of outliers are shown.

Later in the year 2001, Fernando De La Torse and Michael J. Black [4] worked on the same concept of robustness of PCA and tried to overcome the three major issue occurred in the approach of XU and Yullie as follows: First, due to a single " bad pixel " value an image can lie far enough from a subspace, and the entire sample is treated as an outlier . Second, for computation of distance to the subspace, XU and Yuille used least squares projection of data d[i]; the coefficient used for reconstruction of data d[i] are c[i] = B'd[i]. These reconstruction coefficients can be biased for an outlier. Third, They used binary outlier which can either completely includes or discards a sample. Fernando and Michael introduced a more general outlier process that has computational advantages. They used Robust M. Estimation algorithm for learning high dimensional data such as images. Their approach extended previous work by modeling outliers at a pixel level. They tested on natural and synthetic images to show improved outlier tolerance.

3. Juyang Weng, Yilu Zhang and Wey-Shinan Hwang [5] worked toward fast computation of principal components of high dimensional image vectors. They used an incremental principal component analysis algorithm [IPCA] also known as candid free IPCA [CCIPCA], for computation of principal components of samples incrementally without calculating the covariance matrix. In their result, use of CCIPCA algorithm ensures fast convergence for high dimensional image vectors. In their thesis, they sort out the major issue of computing eigenvectors and eigenvalues of high variance without computing corresponding covariance matrix and without knowing data in advance. Their proposed CCIPCA algorithm used the concept of the efficient estimate for fast convergence of high dimensional data. For improvement of the convergence rate, they used the average amnesic technique.

4. Michael E. Tipping and Christopher M. Bishop [3] worked toward the limitations of PCA that is lack of the probabilistic model for observed data, high computation time in case of large datasets, and could not able to deal properly with missing data. They proposed a probabilistic formulation of PCA with the help of expectation maximization algorithm and latent variable formulation make PCA iterative and computationally efficient. Their work was an extension of the earlier work of Lawley [1953] and Rubin and Anderson [1956]. Probabilistic PCA assumes a Gaussian noise modal and formulates the solution for the Eigenvectors as a

maximal likelihood parameter estimation problem. An EM (Expectation Maximization) algorithm is used for finding the principal axes which will iteratively maximize the likelihood function. According to their conclusion, it is very easy to plugin PPCA as part of more complex problems that are for doing nonlinear dimensionality reduction or subspace clustering.

TABLE II [3]
2DPCA AND BIDIRECTIONAL 2DPCA

	Recognition rates		Running time	
NOPC	2DPCA	$(2D)^2PCA$	2DPCA	$(2D)^2PCA$
5	80	70	1.08	0.6
6	78.3	73.3	1.18	1.1
7	83.3	76.7	1.3	1.18
8	83.3	80	1.43	1.29
9	81.7	81.7	1.51	1.42
10	81.7	81.7	1.63	1.5
20	81.7	81.7	2.63	2.4
30	81.7	81.7	3.62	3.25

5. Due to the extensive use of PCA in different fields including spatial pattern identification created a problem for the scientist to detect the pattern from linear structures as standard PCA is associated with the linear structure. To overcome the problem Ruixin Yang, John Tan and Menas Kafatos [6] proposed a Kernel-based PCA to explore and identify non-linear patterns. In KPCA, this non-linear mapping is never found. Instead, since the data points always appear in a dot product form, the kernel trick is utilized, in which each point is represented using the distances to all other points to form a kernel matrix. Consequently, Eigen Value Decomposition has applied to this new representation. Since the actual high dimensional embedding is not explicitly computed, the kernel-formulation of PCA is restricted in that it does not compute the principal components themselves, but the projections of the data onto those components. Kernel PCA has been demonstrated to be useful in several applications such as novelty detection and image de-noising.

6. Modified PCA for 3D face recognition: To enhance the face recognition performance Omid Gervei, Ahmad Ayatollahi and Navid Gervei [10] in the year 2010 proposed a modified PCA method namely 2DPCA and bidirectional 2DPCA which will utilize 3D face information for extracting principal components. Using these methods they have tested on a public database with a random facial expression which gave 83.3 percent of correct prediction.

TABLE III
COMPARISON CHART: STANDARD PCA AND MODIFIED PCA

Year	Author	Techniques	Improvements	implementation
1995	Xu and Yullie	Statistical physics approach	Robustness in the presence of outliers	Real data spoiled by outlier
2001	Fernando De La Torse and Michael J. Black	Robust M. Estimation algorithm	Improved outlier tolerance	High dimensional data like images contaminated by outlier
1998	Michael E. Tipping and Christopher M. Bishop	Gaussian noise modal and Expectation maximization algorithm	Reduction in computation time for large datasets and can handle missing data	Image Compression and Visualization
2003	Juyang Weng, Yilu Zhang, and Wey-Shinan Hwang	The candid free incremental principal component analysis algorithm	Fast computation of principal components of high dimensional image vectors, fast convergence	Appearance-based image analysis
2006	Ruixin Yang, John Tan and Menas Kafatos	Kernel principal component analysis algorithm	Identification of non-linear patterns	Novelty detection and image de-noising

2010	Malika Heeenaye Mamaode Khan, Naushad Mamode Khan and Raja K. Subramanium	Cholesky matrix decomposition, Lancoz techniques	Improvement in number of computation and processing time	Biometric Security System
2010	Omid Gervei, Ahmad Ayatollahi, and Navid Gervei	2D, bidirectional 2D principal component analysis algorithm	3D face recognition with random facial expression and reduction in processing time.	3D face recognition

IV. RESULTS AND DISCUSSION

In this paper, we reviewed the principal component analysis algorithm and several issues in the standard PCA algorithm. By examining the last two decades of existing literature, we know the use of PCA in different applications along with its strengths and weaknesses. This paper helps to identify first, several major issues in PCA algorithm such as processing time, memory consumption, slow convergence in case of high dimensional image vectors, poor outlier tolerance, spatial pattern identification and second, a summary of the prominent approaches toward modification of PCA algorithm to short out the problem of standard PCA. With the help of this review paper, a practitioner can select the most appropriate modified version PCA algorithm for their application.

REFERENCES

[1] Jolliffe, Ian T, "A Note on the Use of Principal Components in Regression," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 31, no. 3, pp. 300–303, 1982.

[2] Lei Xu and A. L. Yuille, "Robust principal component analysis by self-organizing rules based on statistical physics approach," in *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 131-143, Jan. 1995.

[3] Tipping, M. E., & Bishop, C. M. "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. *61*, no. 3, pp. 611–622, 1999.

[4] Torre, Fernando De la and Michael J. Black, "Robust Principal Component Analysis for Computer Vision," *ICCV* (2001).

[5] Juyang Weng, Yilu Zhang and Wey-Shiuan Hwang, "Candid covariance-free incremental principal component analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 1034-1040, Aug. 2003.

[6] Yang R., Tan J., Kafatos M. A Pattern Selection Algorithm in Kernel PCA Applications, in Filipe J., Shishkov B., Helfert M. (eds) Software and Data Technologies. ICSOFT 2006. Communications in Computer and Information Science, vol. 10, 2006.

[7] Hui Zou, Trevor Hastie & Robert Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265-286, 2006.

[8] Olawale, Fatoki & Garwe, David, "Obstacles to the growth of new SMEs in South Africa: A principal component analysis approach," *Afr. J. Bus. Manag.*, vol. 4, 2009.

[9] Gervei, Omid & Ayatollahi, A & Gervei, N, "3D face recognition using modified PCA methods," World Academy of Science, Engineering and Technology, vol. 63, pp. 264-267, 2010.

[10] Heenaye- Mamode Khan, Maleika & Mamode Khan, Naushad & K. Subramanian, Raja, "Feature Extraction of Dorsal Hand Vein Pattern using a fast modified PCA algorithm based on Cholesky decomposition and Lanczos technique," vol. 61, 2010.

[11] Mukerjee, Amitabha &, Guha. (2019). A Realtime Face Recognition system using PCA and various Distance Classifiers A Realtime Face Recognition system using PCA and various Distance Classifiers,2011.

[12] Shlens, Jonathon, "A Tutorial on Principal Component Analysis," *CoRR* abs/1404.1100 (2014), 2014.

[13] Rea, William & Rea, Alethea & Yang, Libin. (2015). Stock Selection with Principal Component Analysis. 10.13140/2.1.2220.8805, Feb 2015.

[14] Bhargava, Neeraj & Kumar, Abhishek & Kumar, Devesh & Assistant, Meenakshi, "A modified concept of PCA to reduce the classification error using kernel SVM classifier," *IJSER*, vol. 6, 2015.

[15] Skaro, Jordan & Goel, Harsh & Hazelwood, Scott & M Klisch, Stephen, "Principal component analysis of gait and cycling experiments: Crosstalk error reduction and corrected knee axes," *in Proceedings SKARO 2017*, 2017.

[16] Piyush Rai Lecture Notes: Probabilistic Machine Learning. [Access on: 20th October 2018]. [Online]. Available: https://www.cse.iitk.ac.in/users/piyush/courses/pml_fall17/pml_fall17.html

[17] Yu, Chao-Hua & Gao, Fei & Lin, Song & Wang, Jb, "Quantum data compression by principal component analysis," 2018.

[18] Jolliffe, Ian & Cadima, Jorge. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 374. 20150202. 10.1098/rsta.2015.0202.

[19] DOCPLAYER Tutorial: Introduction to Principal Component Analysis and Factor Analysis. [Access on: 15th November 2018]. [Online]. Available: http://docplayer.net/17208775-Introduction-to-principal-components-and-factoranalysis.html

[20] Rochna Ramanayaka PPT: MV Factor. [Access on: 20th November 2018]. [Online]. Available: https://www.scribd.com/presentation/193611351/5-MV-Factor

[21] Yumi Blog: Review on PCA. [Access on: 22nd November2018[Online]Available: http://fairyonice.github.io/Review-on-PCA.html.

[22] Minh Luong PPT: Probabilistic Principal Component Analysis and the E-M Algorithm. [Access on: 10th November 2018]. [Online]. Available: https://people.cs.pitt.edu/~milos/courses/cs3750-Fall2007/lectures/class17.pdf.