IJASRM International Journal of Advanced Scientific Research and Management, Vol. 2 Issue 9, Sept 2017.
www.ijasrm.com
ISSN 2455-6378

Principal Component Analysis (PCA) for Beginners

Mr. Ramkumar Gunasekaran[1] and Mr. Tamilarasan Kasirajan[2]

[1] Research Scholar, University of Auckland,
Auckland, New Zealand

[2] Assistant Professor, Anna University,
Chennai, India

Abstract

Principal component analysis (PCA) is a statistical methodology that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or) significant modes of variation. The goal of PCA is to extract the important information from a large dataset to represent it as a set of new orthogonal variables called principal components (or) significant modes of variation [1]. The main function of PCA is to reduce the dimensionality by extricating the smallest number components that account for most of the variation in the original data.

1. Introduction:

Principal component analysis (PCA) is a statistical modeling technique that uses multivariate measurable procedure. PCA was first invented in the year 1901 by Karl Pearson [14]. PCA is a methodology that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, significant modes of variation). The number of modes are less than or equal to the smaller of the number of original variables or the number of observations. This transformation is defined in such a way that the first principal component has the largest possible variance and all the other succeeding components have the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables. Principal component analysis (PCA), otherwise called Karhunen-Loeve extension, is an established component extraction and information portrayal system broadly utilized as a part of the territories of example acknowledgment and PC vision, for example, face recognition [3]. PCA is used in picture separation and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute [1]. The results of a PCA are typically talked about as far as segment scores, now and then called factor scores (the changed variable esteems comparing to a specific information point), and loadings (the weight by which each standardized unique variable ought to be increased to get the segment score) [2].

2. Goals of PCA:

The goals of PCA are:

1.) To extract the most important information from the data table.
2.) To compress the size of the data set by keeping only the important information;
3.) To simplify the portrayal of the data set, and
4.) To analyze the structure of the perceptions and the factors.

3. Steps in conducting PCA:

The steps below are followed in conducting Principal Component Analysis [12]:

- Initial extraction of the components.
- Determine the number of principal components to retain.
- Rotation to a final solution: After deciding the principal components to retain, a rotated factor pattern is created (this is done for ease of interpretation).
- Interpreting the rotated solution
- Create factor scores and summarize the results.

4. Relationship between PCA and Factor Analysis:

Generally, PCA makes variables that are linear combinations of the original variables. Factor analysis is similar to PCA [10], in that factor analysis also involves linear combinations of variables. Not the same as PCA, factor analysis is a relationship centered approach trying to duplicate the inter-correlations among factors, in which the factors "represent the common variance of variables, excluding unique variance" [11].

5. PCA using R:

There are several ways of performing PCA using R. The most common commands are prcomp () and princomp (). We can also focus on the principal () function in the 'psych' package because it has better options.

6. Applications of PCA:

• PCA can be specifically connected to the hazard administration of loan fee subsidiaries portfolios [7].
• A variation of principal components analysis is utilized as a part of neuroscience to recognize the particular properties of a stimulus that increases a neuron's likelihood of creating an activity potential [8].
• PCA as a dimension reduction technique is especially suited to identify composed exercises of expansive neuronal gatherings. It has been utilized as a part of deciding aggregate factors, i.e. arrange parameters, amid stage changes in the cerebrum [9].

7. Advantages of PCA:

The key advantage of PCA is its low noise sensitivity and increased efficiency; the PCA has quite a lot of applications which are listed below:

- Goal is to reduce the total number of variables, but at the same time, preserve most of information (variation).
- Decreases the redundancy of data [4, 5, and 6].
- Reduce complexity in images and reduces the dimensionality of the data [4, 5, 6].
- Smaller database portrayal since just the student pictures are stored as their projections on a reduced basis [4, 6].
- Reduction of noise as the most extreme variation basis is picked, thus the minor variations out of sight are ignored naturally. [4, 6]

8. Limitations of PCA:

Although PCA is a popular SSM methodology, it has certain limitations to be considered. The major limitation of PCA is that it assumes that the variables can be split up into orthogonal modes, which might not be acceptable for all datasets. PCA is used to find the linear correlations between the variables of a dataset which means that if the variables are linearly correlated in the dataset PCA can be used to find the directions that represent the data, but PCA is not enough for non-linear data. Another limitation is that, if the scale is varied for some of the variables in our dataset the PCA results vary. The final limitation is that the PCA doesn't capture even the smallest invariance unless the dataset provides information in a clear and detailed manner [6].
The key disadvantages of PCA are:

- PCA looks for linear combinations of the original factors. The nonlinear combination may even yield better portrayal. PCA has an augmentation for doing this sort of analysis, Nonlinear PCA.
- The segments are not independent but rather uncorrelated. It would be far superior in the event that we have a portrayal which is independent to each other. It is called Independent Component Analysis.
- The covariance matrix is difficult to be evaluated in an accurate manner [4, 6].
- Even the simplest invariance could not be captured by the PCA unless the training data explicitly provides this information [4, 6].

9. Conclusions:

Principal component analysis is an intense and flexible strategy method for giving a review of complex multivariate information [13].

Reference(s):

[1] H Abdi, & LJ Williams (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.
[2] PJ Shaw (2009). Multivariate statistics for the environmental sciences. Wiley.
[3] C Li, Y Diao, H Ma., & Y Li. (2008, December). A statistical PCA method for face recognition. In Intelligent Information Technology Application, 2008. IITA'08. Second International Symposium on (Vol. 3, pp. 376-380). IEEE.

[4] P J. Phillips, P J. Flynn, T. Scruggs, KW. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, "Overview of the Face Recognition Grand Challenge," in Computer vision and pattern recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 947-954.
[5] D. Srinivasulu Asadi, Ch. DV Subba Rao and V. Saikrishna "A Comparative Study of Face Recognition with Principal Component Analysis and Cross-Correlation Technique," International Journal of Computer Applications Vol. 10, 2010.
[6] S. Karamizadeh., .S.M.MAbdullah, A. A., Zamani, & A. Hooman. (2013). An overview of principal component analysis. Journal of Signal and Information Processing, 4(03), 173.
[7] The Pricing and Hedging of Interest Rate Derivatives: A Practical Guide to Swaps, J H M Darbyshire, 2016, ISBN 978-0995455511.
[8] N Brenner., W Bialek., & S de Ruyter, R.R. (2000).
[9] VK Jirsa., R Friedrich, H Haken, & JS Kelso (1994). A theoretical model of phase transitions in the human brain. Biological cybernetics, 71(1), 27-35.
[10] Ijsmi, Editor (2017-04-26). "Tutorial: Factor analysis revisited – An overview with the help of SPSS, SAS and R packages". International Journal of Statistics and Medical Informatics. 3 (1).
[11] TA Brown. (2014). Confirmatory factor analysis for applied research. Guilford Publications.
[12] IT Jolliffe "Principal Component Analysis and Factor Analysis." Principal component analysis. Springer New York, 1986. 115-128.
[13] R Bro., & AK Smilde. (2014). Principal component analysis. Analytical Methods, 6(9), 2812-2831.
[14] K Pearson. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559-572