Available online at http://www.ssstj.sci.ssru.ac.th

Suan Sunandha Science and Technology Journal

Application of Principal Component Analysis As A Data Reducing Technique

Fayose Taiwo Stephen[1], Egobi Mary Ekundayo[2], Adebara Lanre[3],

[1,3]Department of Mathematics and Statistics, The Federal Polytechnic,
Ado Ekiti, Ekiti State, Nigeria
[2]Department of Statistics, Federal University of Technology,
Akure, Ondo State, Nigeria
Corresponding author e-mail: fayose_ts@fedpolyado.edu.com

Received: *13 Feb 2019* **Accepted:** *3 Apr 2019*

Abstract

This paper is on the application of principal component as a data reducing technique on economic variables for the period of 26 years. The source of data was secondary and was collected from the Central Bank of Nigeria Statistical Bulletin. The aim is to use principal component analysis effectively and profitably to reduce the large and massive economic variables (Data) to a smaller number of PCs while retaining as much as possible of the variation in the original variables. The methodology employed Principal Component which are orthogonal in nature from the original Economic Variables. The criterion for selecting the number of Principal Component to be extracted is the KAISER'S CRITERION which was suggested by GUTTMAN and adopted by KAISER. The result of the analysis revealed that the variables BOP, LR, and INFL have low correlation coefficient with other variables. Furthermore, results showed that the large sample size of economic variables have being reduced and the principal components are extracted in which the first Principal Component have the highest number of variables which are positively highly correlated, the second Principal Component loads positively with Crude Oil production, Lending Rate and Inflation Rate while the third Principal Component load positively with Balance of Payment.

Keywords: Principal Component Analysis, Principal Component, Kaiser's Criterion, Guttman, Kaiser, Karhunen-Loeve Transform, Proper Orthogonal Decomposition

1. Introduction

Principal Component Analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed**,** PCA is sensitive to the relative scaling of the original variables. Depending on the field of application, it is also named the discrete **Karhunen-Loeve Transform (KLT)**, the **Hotelling Transform or Proper Orthogonal Decomposition (POD).**

PCA was invented in 1901 by Karl Pearson. The technique has found application in many diverse fields such as Ecology, Economics, Psychology, Meteorology, Oceanography and Zoology. Now it is mostly used as a tool in exploratory data analysis and for making predictive models that is, to assist in the data analysis, PCA (among other techniques) is generally employed as both a descriptive and data reduction technique.

In the development of PCA, Pearson was interested in constructing a line or a plane that "best fits" a system of points in q- dimension space. Statistically speaking, PCA represents a transformation of a set of q correlated variables into linear combinations of a set of q pair- wise uncorrelated variables called Principal Components. Components are constructed so that the first component explains the largest amount of total variance in the data and

each subsequent component is constructed so as to explain the largest amount of the remaining variance while remaining uncorrelated with (orthogonal to) previously constructed components.

We define the dimension of the data set to be equal to the number of principal component. The set of q principal components is often reduced to a set of size K, where $1 \leq k$ for all q. The objective of dimension reduction is to make analysis and interpretation easier while at the same time retaining most of the information (variation) contained in the data. Clearly, the closer the value of K is to q the better the PCA model will fit the data since more information has been retained, while the closer K is to 1, the simple the model.

Many methods have been proposed to determine the number K, that is, the number of "meaningful" components. Some methods can be easily computed while others are computationally intensive. Methods include (among others): the broken stick model, the Kaiser-Guttman test, Log-Eigen value (LEV) diagram, Velicer's Partial Correlation procedure, Cattel's Scree test, Cross-validation, bootstrapping techniques cumulative percentage of total variance, and Bartlett's test for equality of eigen values.

Data reduction is frequently instrumental in revealing mathematical structure. Karr and Martin note that the percent variance attributed to Principal Components derived from real data may not be substantially greater than that derived from randomly generated data. The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score). PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data.

1.1 Statement of the Problem

In this study, our emphasis is on the application of principal component as a data reducing technique on economic variables for the period of 26 years (1980-2005) by correlating GDP, External Reserve, Exchange Rate, External Debt, Inflation Rate, Lending Rate, Money Supply, Crude Oil Production, Balance of payment, Balance of Trade and Oil Revenue as our selected Economic variables.

We shall seek solutions to the following problems;

- Could the large number of measurement be replaced by a small number of measurement (functions) of them without the loss of information?
- How many components do we extract to avoid loss of information?

1.2 Aim and Objectives of the Study

The aim of this study is to use principal component analysis effectively and profitably to reduce the large and massive economic variables (Data) to a smaller number of PCs while retaining as much as possible of the variation in the original variables. The objectives are:

- To summarize patterns of correlations among observed variables;
- To determine the number of components to be extracted;
- To discover or to reduce the dimensionality of the data set;
- To identify new meaningful underlying variables.

1.3 Materials and Methods

The method of PCA was used to form a new set of variables called the Principal Component which are orthogonal in nature from the original Economic Variables. The criterion for selecting the number of Principal Component to be extracted is the KAISER'S CRITERION which was suggested by GUTTMAN and adopted by KAISER. The criterion states that only Principal Component(s) having Latent root (Eigen value) greater than one are retained in the analysis. That is, we retain P_i iff $\lambda_i > 1$ CHI-SQUARE TEST.

Suan Sunandha Science and Technology Journal

1.4 Method of Data Collection

The method used is transcription from records. The data for this study were collected from the Central Bank of Nigeria (CBN) Statistical Bulletin (Golden Jubilee Edition, 2008). The data were collected for the period of 1980-2005.

Having mentioned the method for collecting the required data and the period, it becomes necessary to state the method for which the data was analyzed. The correlation matrix of the variables was estimated using SPSS version 17.0. An iterative procedure of the Principal Component Analysis was employed in order to obtain the Eigen Values to be extracted (λ_i), the normalized characteristic vectors (V_i), the Principal Components and the values for each Component extracted (P's) using an initial guessed vector (X_o).

1.5 Findings

- The choice of the initial guessed vector of $(1,1,1,1,1,1,1,1,1)^I$ was utilize in all of the iterations without alternating them.
- We discovered that the variables BOP, LR, and INFL have low correlation coefficient with other variables.
- As the extraction continues, higher powers of R are required for the system to converge and at the point in which the principal components were obtained base on KAISER'S CRITERION the power of R begin to reduce for the system to converge.
- Our major findings was that the large sample size of economic variables have being reduced and the principal component are extracted in which the first Principal Component have the highest number of variables which are positively highly correlated, the second Principal Component loads positively with Crude Oil production, Lending Rate and Inflation Rate while the third Principal Component load positively with Balance of Payment.

1.6 Conclusion

Principal Component Analysis is a tool imperative as far as data reduction is concern. It is very helpful in determining empirically how many dimensions or underlying construct accounts for most of the variance. From the analysis, we may infer that with the aid of Principal Component Analysis the researchers have been able to project the large variables (ten predictor variables) into three Principal Components called the factors and the Principal Component were found to be a linear combination of the original variables.

The variables that were linearly combined for each of the Principal Component were obtained based on their correlation with the Principal Components. Knowing fully well that there are several other analysis that can be used as a data reducing technique on economic variables, the method of principal component yielded reliable results.

1.7 Recommendation

The followings are suggested as possible recommendation for further study:

- Future study should employ more or less observations with proper model specification and estimation.
- The Screen Test should also be employed as a criterion for selecting PC since there may be some λ which by approximation will be unity (by Kaiser's criterion).
- We recommend that the size of explanatory variables for this study period should be increase.

1.8 Acknowledgement

We want to acknowledge and thank the Head of Department of Statistics, The Federal University of Technology, Akure (FUTA-LISA) for allowing us to use Statistics laboratory for our analysis and research. We also acknowledge the contribution of other scholars.

1.9 References

Journals article

O'Brien, R.M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality and Quantity. International Journal of Methodology*. Springer. *Vol. 4 pp. 673-690*. doi:10.1007/s11135-006-9018-6.

Kaiser, H.F. (1960). The Application of Electronic Computers Factor Analysis, *Education & Psychological Measurement* 20, 141-151. Doi:10.1177/00131646002000116.

Jeffers J.N.R. (1967); Two Case Studies in the Application of Principal Component Analysis. Applied Statistics. *Journal of the Royal Statistical Society*. 16:225-236. Doi: 10.2307/2985919.

Gower J. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Journal Storage. Biometrika Trust, vol. 53. pp:325-338*.

Psych J. E. (1983). Hotelling H: Analysis of a Complex Statistical Variable into Principal Components., 26:417-441.

Books

Johnson R.A., & Wichern D.W. (1982). *Applied Multivariate Statistical Analysis.* (2nd edition). Upper Saddle River, NJ: Prentice Hall.

Johnson R.A., & Wichern D.W. (1988). *Applied Multivariate Statistical Analysis.* (3rd edition) Englewood Cliffs, NJ: Prentice Hall.

Johnson R.A., & Wichern D.W. (2002). *Applied Multivariate Statistical Analysis.* (4th edition), Upper Saddle River, NJ: Prentice Hall.

Morrison D.F. (2007). *Multivariate Statistical Methods.* McGraw-Hill Book Company, New York. XI11 +X338 S., 15 Abb., 32 Tab.

Morrison D.F. (1967). *Multivariate Statistical Methods*, Mathematical Statistics. 338 Pages. McGraw-Hill Book Company, New York.

Sam K. K. (1991). *Multivariate Statistical Analysis,* A Conceptual Introduction. 303 Pages. Radius Press.

Johnson D.E. (1998). *Applied Multivariate Methods For Data Analysts*. 567 Pages. Duxbury Press.

Dillon W.R., Goldstein M. (1984). *Multivariate Analysis: Methods and Applications*. 587 Pages. Wiley.

Anderson D.R, Sweeney D.J, Williams T.A Freeman J. & Shoesmith, E. (2007). "*Statistics for Business and Economics*". 9th edition. Thomson. ISBN-13- 978-1-84480-313-2.

Morrison D.F. (1983). *Applied Linear Statistical Methods*. Englewood Chiff, New Jersy, Pentice Hall.

Stevens J. (1986). *Applied Multivariate Statistics for Social Sciences*. New Jersey United States of America. Lawrence Elbaum Associates, Inc.

Reports

Central Bank of Nigera (2008) Statistical Bulletin Golden Jubilee Edition, CBN Press Abuja. www.cbn.gov.ng/documents/Statbullentin.asp