Journal of the Nigerian Association of Mathematical Physics
Volume **19 (November, 2011), pp 607 – 614**

Principal Component Analysis as an Efficient Performance Measurement Tool

Acha, Chigozie Kelechi

Department Of Mathematics/Statistics
College of Natural and Applied Sciences
Michael Okpara University of Agriculture, Umudike
Abia State. Nigeria

Abstract

This paper uses the principal component analysis (PCA) to examine the possibility of using few explanatory variables (X's) to explain the variation in Y. It applied PCA to assess the performance of students in Abia State Polytechnic, Aba, Nigeria. This was done by estimating the coefficients of eight explanatory variables in a regression analysis. The explanatory variables involved in this analysis show a multiple relationship between a dependent variable and independent variables. A correlation table was obtained from which the characteristic roots were extracted. Also, the orthonormal basis was used to establish the linear independence of the variables. The first principal component accounted for 51.6 percent of the total variation, while the second principal component accounted for 23.3 percent. The descriptive statistics and plots were considered. The principal components yielded good estimates, which leads to the structural co-efficient of the regression model.This led to the conclusion that PCA uses few explanatory variables to explain variations in a dependent variable and is therefore an efficient tool for performance assessment.

Keywords: Orthomormality, Eigenvalue,Diagonalizability, Vector, Standardised.

1.0 Introduction

In the institutions of higher learning, such as Universities, Polytechnics, and Colleges of Education among others, students' academic performances in a semester are judged by their Grade points Average (GPA), the Grade Points Average depends on the grades made by students on the courses offered together with the credit units attached to them. To obtain the Grade Point Average involves taking the summation over the product of the grades made on each course together with the credit units assigned to them and dividing the result by the total credit units assigned to the courses that were offered in that semester, where grades are represented in numbers.

The Grade Point Average could be seen as response variables(X's). There are various statistical techniques used in the estimation of the response variable from the explanatory variable. The major statistical tool for estimation of the coefficients of the explanatory variables is the principal component analysis. The other statistical tools applied are correlation, orthonormality, descriptive statistics and plots or graphs. However, PCA was invented in 1901 by Karl Pearson. PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. The results of a PCA are usually discussed in terms of component scores, that is, the transformed variable values corresponding to a particular case in the data and loadings the weight by which each standardized original variable should be multiplied to get the component score (Shaw [1]).

Corresponding author: E-mail: specialgozie@yahoo.com, Tel. +2348037607057

1. Objectives Of Study

The objectives of this study include the following:

i. To ascertain whether total variation in the dependent variable (Y) could be explained by few explanatory variables(X's).

ii. To ensure the orthonormality of the explanatory variables, such that the principal components are orthogonal.

iii. To solve the problem of multi-colinearity in a multiple regression model which is always present in a model of multiple relationship

2. Literature Review

Principal components analysis is a technique for finding a set of weighted linear composites of original variables such that each composite (a principal component) is uncorrelated with the others. It was originally designed by Pearson [2] though it is more often attributed to Hotelling [3] who proposed it independently. The first principal component is a weighted linear composite of the original variables with weights chosen so that the composite accounts for the maximum variation in the original data. The second component accounts for the maximum variation that is not accounted for in the first. The third component likewise accounts for the maximum given the first two components and so on. These weights are found by a matrix analysis technique called eigen-decomposition which produces eigenvalues. Eigenvalues represent the amount of variation accounted for by the composite and eigenvectors give the weights of the original variables(see http://www.pcp-net.org/encyclopaedia/pca.html[4]).

Principal component analysis as a very useful statistical technique later found application in various fields. PCA is recommended as an explanatory tool to uncover unknown trends in the data. According to Jolliffe[5], Miranda et al[6], "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 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 as high a variance as possible (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 (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–Loèvetransform (KLT), the Hotelling transform or proper orthogonal decomposition (POD).In fact, several data decomposition techniques are available for this purpose: Principal Components Analysis (PCA) is among these techniques that reduces the data into two dimensions. The set of data or elements or numbers arranged in a table (matrix) as rows(row vector) or columns(column vectors) called vectors are being used. Moreover, since the Orthonormal basis is a set of vectors which forms a basis for a vector space and each of these basis vectors are normalized and they are orthogonal to each other.

Axler [7] observed that Orthonormal sets are not especially significant on their own. However, they display certain features that make them fundamental in exploring the notion of diagonalizability of certain operator on vector spaces.Wang et al [8] confirmedthat PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.878, 0.478) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean. According [9] and [10] confirms that orthonormal basis is a set of vectors which forms a basis for a vector space and each of these basis vectors are normalized and they are orthogonal to each other. Principal Components Analysis (PCA) can also be seen as a special case of the more general method of factor analysis whose sole aim is to construct a set of variables of new variables (Pi) called principal components which are linear combination of the X's(see [11]).

2. MATERIALS AND METHODS

The data for this work was collected from the Statistics Department of Abia State Polytechnic, Aba, Nigeria. It shows the Grade Points(GP) and Grade Point Average(GPA) obtained by 2009/2010 National Diploma final year students of the Department.The statistical method include table of correlation co-efficient to check if there is any relationship among the explanatory variables, descriptive statistics describing the features of the data, orthonormality plot to overcome multi-collinearity and show trend or pattern of the explanatory variables and the principal components which accounts for the variation among the variables.

Table 1 shows courses offered and their credit units, the grades obtained in the various courses and the Grade Points Average (GPA). All the analyses were carried out using *Eviews 7* software.

TABLE 1: Students' Grades and Grade Point Average

Courses	GNS 201	COM 211	STA 211	STA 212	STA 213	STA 214	STA 215	STA 216	G.P.A	REMARK
Credit Units	3	3	3	2	2	4	2	2		
Grades	AB	A	AB	B	AB	A	A	AB	3.66	PASS
	AB	AB	A	BC	B	B	A	AB	3.23	PASS
	B	C	AB	BC	C	B	BC	A	2.80	PASS
	AB	AB	B	C	BC	A	B	BC	3.11	PASS
	BC	BC	B	B	AB	BC	C	BC	2.70	PASS
	B	BC	C	B	B	BC	BC	BC	2.61	PASS
	BC	C	B	C	BC	BC	C	B	2.36	PASS

GRADE POINTS
A=4.00.AB=3.50,B=3.00
BC=2.50,C=2.00,F=0.00

5. Analysis And Discussion Of Results

A. Descriptive statistics for the set of data

Table 2: Descriptive statistics for the set of data

	GNS201	COM211	STA211	STA212	STA213	STA214	STA215	STA216
Mean	3.071429	2.857143	3.142857	2.571429	2.857143	3.071429	2.857143	2.928571
Median	3.000000	2.500000	3.000000	2.500000	3.000000	3.000000	2.500000	2.500000
Maximum	3.500000	4.000000	4.000000	3.000000	3.500000	4.000000	4.000000	4.000000
Minimum	2.500000	2.000000	2.000000	2.000000	2.000000	2.500000	2.000000	2.000000
Std. Dev.	0.449868	0.801784	0.626783	0.449868	0.556349	0.672593	0.852168	0.731925
Skewness	-0.272380	0.235217	-0.570697	-0.272380	-0.192012	0.615800	0.476426	0.260009
Kurtosis	1.493080	1.486626	2.871901	1.493080	1.856509	1.716759	1.668234	1.609630
Jarque-Bera	0.748875	0.732553	0.384764	0.748875	0.424388	0.922701	0.782112	0.642702
Probability	0.687676	0.693311	0.824992	0.687676	0.808808	0.630432	0.676342	0.725169
Sum	21.50000	20.00000	22.00000	18.00000	20.00000	21.50000	20.00000	20.50000
Sum Sq. Dev.	1.214286	3.857143	2.357143	1.214286	1.857143	2.714286	4.357143	3.214286
Observations	7	7	7	7	7	7	7	7

Source: Authors computation using *Eviews 7* software.

The descriptive statistics shows the unique features the data that is being used. For instance, inTable 2, the mean value of STA 211(3.142857) is the highest among others but the median of (GNS201, STA211, STA213, STA214) is 3.000000 while the median for the remaining variables is equal to 2.500000.Table 2 also shows that 4.000000 is the maximum and 2.000000 the minimum scores collected.STA215 is having the highest standard deviation while GNS201 and STA211having 0.449868 each are the best in terms of selecting variable with minimum standard deviation. The values of skewness and kurtosis were also computed for the seven observations. In fact, using the probability of the explanatory variables computed in Table 2 at 5% level of significance, we conclude that all the variables used in this work are statistically significant.

B. Pie chart of the explanatory variables

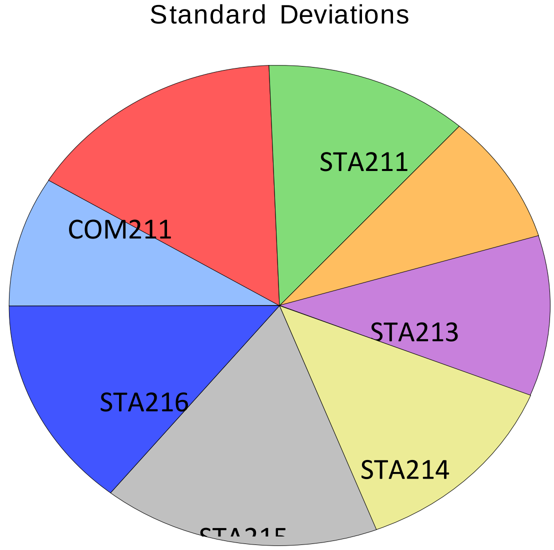

Figure 1: Visual plot of the explanatory variables using standard deviations. Figure 1 confirms the results we have in Table 2 about the standard deviation that STA215 is having the highest standard deviation while GNS201 and STA211having 0.449868 each are the best in terms of selecting variable with minimum standard deviation.

C. Graphs of the explanatory variables

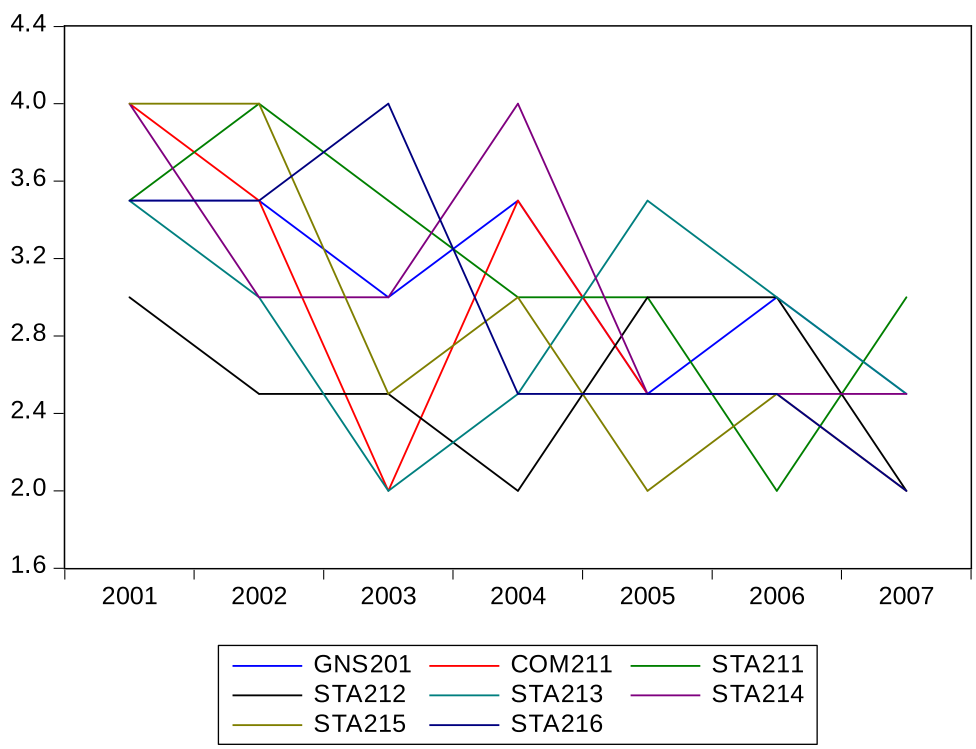

Figure 2: Visual plot of the explanatoryvariables using raw data

It is pertinent to note that the sole aim of this section is to show the patterns or trends of the data. From the visual plot in Figure 2, there are trends in all the explanatory variables, which indicate that the data set is not stationary.It also shows that all theexplanatoryvariables relate to each other.

C. Correlation analysis

Table 3: Correlation Table for the explanatory variables

	COM211	GNS201	STA211	STA212	STA213	STA214	STA215	STA216
COM211	1.000000	0.841625	0.379023	0.148522	0.507072	0.794719	0.879892	0.263718
GNS201	0.841625	1.000000	0.401090	-0.029412	0.047565	0.806562	0.900552	0.524249
STA211	0.379023	0.401090	1.000000	-0.189990	-0.051209	0.367109	0.590642	0.661724
STA212	0.148522	-0.029412	-0.189990	1.000000	0.713477	-0.157378	0.139741	0.271163
STA213	0.507072	0.047565	-0.051209	0.713477	1.000000	0.031814	0.301321	-0.131559
STA214	0.794719	0.806562	0.367109	-0.157378	0.031814	1.000000	0.675036	0.350647
STA215	0.879892	0.900552	0.590642	0.139741	0.301321	0.675036	1.000000	0.582142
STA216	0.263718	0.524249	0.661724	0.271163	-0.131559	0.350647	0.582142	1.000000

Source: Authors computation using *Eviews 7* software.

It is pertinent to note that Table 3 is a table of correlation coefficients between each pair of variables in which principle components can be computed. Table 3 confirms that there exists a relationship between the variables.

D. Orthogonality

Orthogonality occurs when two things can vary independently, they are uncorrelated, or they are perpendicular. The essence of this section is to ensure that the explanatory variables are linearly independent also to check multi-colinearity among the variables.The following procedures can be used to compute the principal component manually. The following results were obtained from the correlation table (Table 3).

$$L_{ij} = \frac{\sum_{i=1}^{K} r.i}{\sqrt{\Sigma}} \quad (1)$$

Where

$$\sum_{i=1}^{K} r.i = 29.246 \, , \quad \sqrt{\sum_{r=i}^{K}} = 5.408, \ i=1,2,3,\ldots, k, \ k = 8.$$

The L_{ij} are the loadings for first the principal component denoted as P_1.

$$P_1 = I_{11}X_1 + I_{12}X_2 + I_{13}X_3 + I_{14}X_4 + I_{15}X_5 + I_{16}X_6 + I_{17}X_7 + I_{18}X_8 \quad (2)$$

The Eigen value of characteristic root for the first principal component was obtained as:

$$\lambda_1 = \sum_{i=v}^{k} I_{11}^2 = I_{11}^2 + I_{12}^2 + I_{13}^2 + I_{14}^2 + I_{15}^2 + I_{16}^2 + I_{17}^2 + I_{18}^2 \quad (3)$$

Percentage of variation accounted for by P_1 (P_1%) is

$$P_1\% = \frac{\lambda_1}{k} x \frac{100}{1} \quad (4)$$

Following the same procedure, the second principal component was obtained using the correlation table as follows:

$$P_2 = I_{21}X_1 + I_{22}X_2 + I_{23}X_3 + I_{24}X_4 + I_{25}X_5 + I_{26}X_6 + I_{27}X_7 + I_{28}X_8 \quad (5)$$

The Eigen value of characteristic root for the second principal component was obtained as:

$$\lambda_2 = \sum_{i=v}^{k} I_{22}^2 = I_{21}^2 + I_{22}^2 + I_{23}^2 + I_{24}^2 + I_{25}^2 + I_{26}^2 + I_{27}^2 + I_{28}^2 \quad (6)$$

$$P_2\% = \frac{\lambda_2}{k} x \frac{100}{1} \quad (7)$$

In this work the computation of the principal component was done using Eviews 7 as shown in Figure 1

Orthonormal loading plot

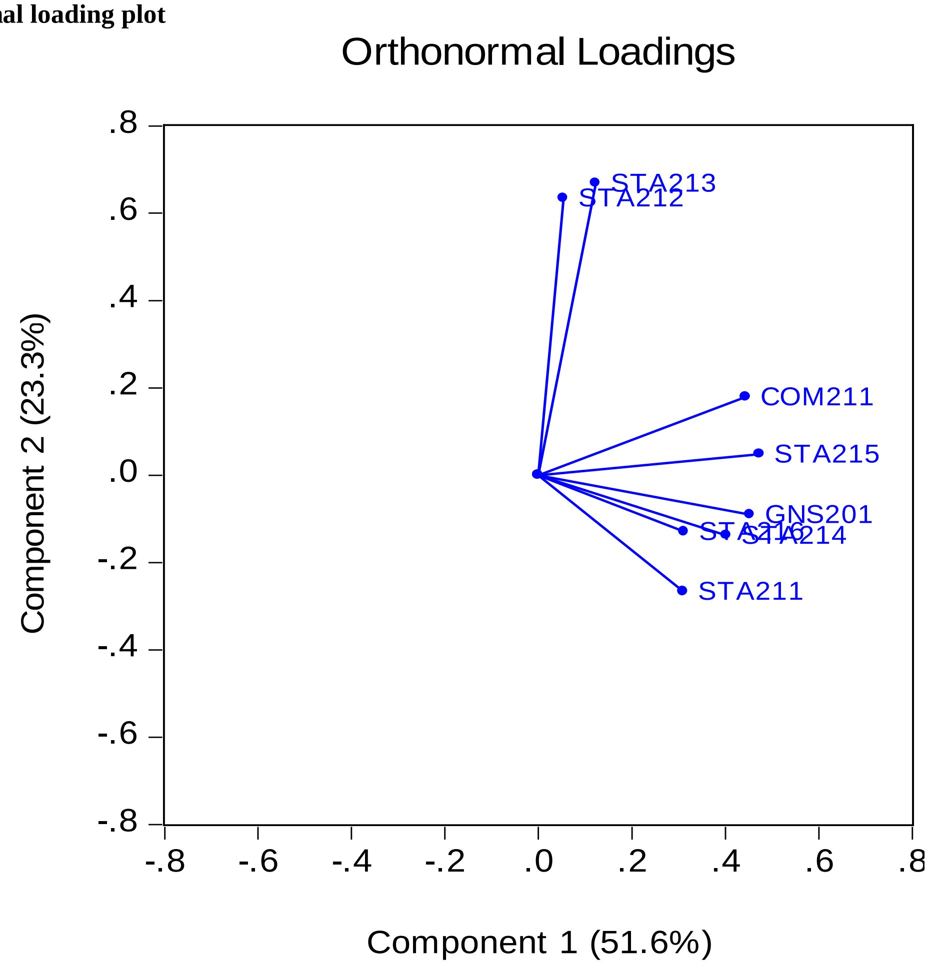

Figure 3: Orthonormal loading plot.
Source: Authors computation using *Eviews 7* software.

In Figure 3, theorthonormal loading of the explanatory variables are plotted and theresults of this plot are as given in Table 4.

Table 4: Result from the orthonormal loadings showing the actual values of the plots.

Explanatory variables	Orthonormal loadings
GNS 201	2(0.44,0.18)
COM 211	5(0.45,-0.09)
STA 211	8(0.31,-0.27)

STA 212	11(0.05,0.63)
STA 213	14(0.12,0.67)
STA 214	17(0.40,-0.14)
STA 215	20(0.47,0.05)
STA 216	23(0.31,-0.13)

Source: Authors computation using *Eviews 7* software.

The Table 4 results are centered at (1, 3) with a standard deviation of 3 in the following directions given in table 4 and of 1 in the orthogonal direction (seeWang et al [8]&http://en.wikipedia.org/wiki/pca[12]). The result is in accordance with the apriority theorem on PCA. It also shows that the explanatory variables are linearly independent.

The component 1 and component 2 of the principal components were plotted on the orthonormal loadings. It was discovered that more than 75 percent approximately of the total variations were explained by the first (two) principal components. The 75 percent accounted for is a very good estimate, which leads to the structural co-efficient of the regression model.

6. Conclusion

This paper examines whether total variation in the dependent variable Y could be explained by few explanatory variables(X's).It starts by analyzing the descriptive statistics and the visual plots of the set of data. The results showed that at 5% level of significance, all the variables used in this work are statistically significant as shown in Table 2.

However, for the orthonormality of the explanatory variables, correlation analysis was carried out which leads to orthogonality of the variables.Orthogonality occurs when two things can vary independently, that is, they are uncorrelated or they are perpendicular.

Furthermore, the result of the orthogonal analysis was shown using orthonormality loading plot. This plot shows the individual plot of the variables. The result of orthonormality shows that there is no multi-colinearity between the variables. The graphs were used to depict or confirm that trend or pattern of the explanatory variables. The paper therefore concludes that having isolated the Principal Components, the first two accounted for more than 75% of the variation set; this gives better estimate for the response variable in the absence of multi- colinearity.With PCA therefore, variations in the response variable can be efficiently explained using few explanatory variables.

Refrences

[1] Shaw P.J.A. (2003) *Multivariate statistics for the Environmental Sciences*, Hodder-Arnold ISBN 0-3408-0763-6.

[2] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space" (PDF). *Philosophical*

*Magazine***2** (6): 559–572. http://stat.smmu.edu.cn/history /pearson1901.pdf.

[3] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417-441.

[4] PCA on-line at: http://www.pcp-net.org/encyclopaedia/pca.html, accessed on 24 August, 2011.

[5] Jolliffe I.T.(2002) Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4.

[6] Miranda, A., Le Borgne, Y. A.&Bontempi, G. (2008) New Routes from Minimal Approximation Error to Principal Component, 27(3), June, Neural Processing Letters, Springer.

[7] Axler, S. (1997), *Linear Algebra Done Right* (2nd ed.), Berlin, New York: Springer-Verlag, ISBN 978-0-387-98258-8.

[8] Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J. (2005) Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365, 671-679.

[9] Kriegel, H. P., Kröger, P., Schubert, E &Zimek, A. (2008). "A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms". *Scientific and Statistical Database Management* **5069**: 418. doi:10.1007/978-3-540-69497-7_27.

[10] Gorban, A. N &Zinovyev,A. Y. (2009)Principal Graphs and Manifolds, In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, Olivas E.S. et al Eds. Information Science Reference, IGI Global: Hershey, PA, USA, 28-59.

[11] Orthonormality on-line at:http://en.wikipedia.org/wiki/orthonormality, accessed on 26 August, 2011.

[12] PCA on-line at: http://www.pcp-net.org/encyclopaedia/pca.html, accessed on 24 August, 2011http://en.wikipedia.org/wiki/pacl, accessed on 22 September, 2011.