Continental J. Education Research 7 (1): 24 - 29, 2014 ISSN: 2141 - 4181
© Wilolud Journals, 2014 http://www.wiloludjournal.com
Printed in Nigeria doi:10.5707/cjeducres.2014.7.1.24.29

ANALYSIS OF SCIENCE STUDENTS' SSCE RESULT USING PRINCIPAL COMPONENT ANALYSIS

[1]Ran Vijay Kumar Singh, [2]Abdulmuahimin Abiola Sanusi and [2]Ahmed Audu
[1] Kebbi State University of Science and Technology, Aliero and [2]Usmanu Danfodiyo University Sokoto

ABSTRACT

This research work intends to detect the subjects that contribute to the inconsistency in the relationship among the subjects of studied by using the method of Principal Component Analysis that was used to investigate the variability among the subjects. The data used for the analysis is students' WAEC results for the year 2011, Federal Science College, Sokoto. Out of fourteen subjects areas which the students sat for, nine subjects were used for the study. Single sets were formed that consists Economics, Geography, English Language, Hausa-Language, Mathematics, Agricultural Science, Biology, Chemistry, and Physics. Nine Components were formed, while only six Components were considered showing the subjects that contribute significantly to the variation and inconsistency among the subjects. However, there should be concentration on the subjects that contribute a high variation to check the students' performance.

KEYWORDS: Principal Component, Maximal Variance, Orthogonal, Eigenvectors.

Received for Publication: 13/01/14 Accepted for Publication: 28/04/14
Corresponding Author: Singhrvk13@gmail.com

INTRODUCTION

Principal component analysis is one of the methods that can be used to analyse multivariate dataset. It can reduce the dimensionality of large data set which consists of a number of interrelated variables to smaller components (Hussain *et. al*. 2011).

Several authors including Anderson (1958), Marada, et. al.(1979), Jolliffe (1986), Hair et. al. (1998), Anderson, *et. al*. (1998) established that in principal component Analysis, we seek to maximize the variance of a linear combination of the variables. Essentially, Principal component is a one-sample technique applied to data with no groupings among the observations and no partitioning of the variables into subset y and x. Principal components are concerned only with the core structure of a single sample of observations on p variables. None of the variables is designated as dependent, and no grouping of observations is assumed (Rencher, 2002). The first principal component is the linear combination with maximal variance; we are essentially searching for a dimension along which the observations are maximally separated or spread out. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on.

Principal components are used to reduce the number of dimensions. Another useful dimension reduction device is to evaluate the first two principal components for each observation vector and construct a scatter plot to check for multivariate normality, outliers, and so on. Everitt and Dunn (1991) and Antonio, *et. al*.(2010) pointed out that, in general, the first few principal components are sensitive to outliers that inflate variances or distort co variances, and the last few are sensitive to outliers that introduce artificial dimensions or mask singularities.

For example, we might want to rank students based on their scores on achievement test in English, Mathematics, Reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights, we can spread the students out further on the scale and obtain a better ranking. In the present paper

our aim is to locate the subjects or group of subjects that contribute inconsistency in the academic performance of the students by using principal component analysis.

Data used for the Analysis

The data used for the study were collected from Federal Science College, Sokoto.
The data consist of scores of 100 students for 2011 WEST AFRICA EXAMINATION COUNCIL (WAEC) Exams, Sokoto, Nigeria. Out of fourteen subjects areas which the students sat for, nine subjects are used. The selection was based on the number of students that sat for the interested subjects of study. The nine subjects were put in a single Set. The Set consists of Economics, Geography, English Language, Hausa-Language, Mathematics, Agricultural Science, Biology, Chemistry, and Physics.

Principal Component Analysis

The steps involved in the analysis of principal component analysis include the following methods as below. Algebraically, principal components are particular linear combinations of the p random variables.

Geometrically, these linear combinations represent the selection of new coordinate system obtained by rotating the original system with their development and does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant ellipsoids.

Step 1: get the data

Consider the linear combinations:

$$\begin{aligned} Y_1 &= a_1^{'}X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ Y_2 &= a_2^{'}X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &\cdot \\ &\cdot \\ &\cdot \\ Y_p &= a_p^{'}X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p \end{aligned} \qquad (1)$$

Step 2: standardise the data

Sometimes it makes sense to compute principal component for raw data. This is appropriate when all the variables are in the same units. Standardizing the data is often preferable when the variables are in different units or when the variance of different columns is substantial. This can be done by subtracting the means of each column and dividing by its standard deviation namely:

$$Z = \frac{(X - \mu_i)}{\sqrt{\sigma_{1i}}} \quad , \qquad i = 1, 2, ..., p \qquad (2)$$

In matrix notation, it is given by:

$$Z = (V^{1/2})^{-1}(X - \mu) \qquad (3)$$

where $V^{1/2}$ is the diagonal standard deviation matrix. From this, we obtain mean of Z equal zero, E(Z) = 0.

Step 3: calculate the covariance matrix.
Further, the covariance matrix of Z is calculated using the formula below

$$COV(Z) = (V^{1/2})^{-1} \sum (V^{1/2})^{-1} = \rho \quad (4)$$

where ρ also known as correlation.

Step 4: calculate the eigenvectors and Eigen values of the covariance matrix
The principal components of Z may be obtained from eigenvectors of the correlation matrix ρ of X, refer to equation (3).

The i^{th} principal component of the standardised variables $Z^{'} = [Z_{1,} Z_{2,} ... Z_P]$ with Cov(Z) = ρ is given by

$$Y_i = \rho_i^{'} Z = \rho_i^{'} (V^{1/2})^{-1} (X - \mu), \qquad i = 1, 2, ..., p \quad (5)$$

The eigenvectors of correlation matrix are also known as principal components coefficients or principal component loadings

Moreover,

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i}^{p} Var(Z_i) = q \quad (6)$$

and

$$\rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i}, \qquad i, k = 1, 2, ..., p \quad (7)$$

In this case, $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$ are the Eigen values-eigenvectors pairs with $\lambda_1 \geq \lambda_2 \geq ... \lambda_p \geq 0$.
As seen from equation (5), the total (standard variables) population variance is simply q, the sum of the diagonal elements of the matrix ρ ,then the proportion of the total variance explained by the K^{th} principal component of Z is :

$$\frac{\lambda_k}{q}, \qquad k = 1, 2, ..., p \quad (8)$$

where the $\lambda_{k's}$ are the Eigen values of ρ.

In short, principal component analysis consists of finding linear transformations $Y_1, Y_2, ..., Y_P$ of the original variables $X_1, X_2, ..., X_P$, that have the property of being uncorrelated.

The Y variables are chosen in such a way that Y_1 has maximum variance, Y_2 has maximum variance to being uncorrelated with Y_1, and so on.

Bartlett's Test
Bartlett's Test is used to test the homogeneity of variance in the components. Test hypothesis :

$$H_o: \ \delta_1 = \delta_2 = \ldots = \delta_k$$
$$H_1: \ \delta_i \neq \delta_j \quad for\ atleast\ one\ pair\ (i, j)$$

Test Statistic**:**

$$T = \frac{(N-k) \ln S_p^2 - \sum_{i=1}^{k} (N_i - 1) \ln S_i^2}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i - 1)) - 1/(N-k))} \quad (9)$$

In the above, S_i^2 is the variance of the i^{th} group, N is the total sample size, N_i is the sample size of the i^{th} group, k is the number of groups, and S_p^2 is the pooled variance. The pooled variance is weighted average of the group variance and is defined as: $S_p^2 = \sum_{i=1}^{k} (N_i - 1)S_i^2 / (N - k)$. (10)

Significance Level: $\alpha = 0.05$

Critical Region: The variances are judged to be unequal if, $T \succ \chi^2_{(\alpha, k-1)}$

Where $\chi^2_{(\alpha, k-1)}$ is the upper critical value of the chi-square distribution with k-1 degree of freedom and a significance level of α.

ANALYSIS AND DISCUSSION

Bartlett's test

Test hypothesis:

$H_O : \delta_1 = \delta_2 = \ldots = \delta_K$

$H_1 : \delta_i \neq \delta_j$ For at least one pair (i , j)

Reject H_O *if* $p - value \prec \alpha = 0.05$

From the Bartlett's Test, Approximation chi-square equals 113.04 with degree of freedom of 36, probability level of 0.000, at $\alpha = 0.05$, we therefore reject H_O and conclude that the variances across the variables are not equal.

In this regard, this calls for the use of Principal Component Analysis; to see the variables i.e. the subjects that posse's high variability contribution to the set of components considered.

Therefore, Table 1 shows the Eigen values in column two, which are the proportions of total variance in all the variables, which are accounted for by the components. From the output component, one gives the highest variance explained followed by component two which gives the second highest variance explained and so on. The second component is formed from the variance remaining after those associated with the first component has been extracted, thus this account for the second largest amount of variance. It is worthwhile to note that the principal component coefficient that gives the variance explained for each component gives the value of less than 30% of the variance explained. Therefore, more than one component is needed to describe the variability of the data. In other to obtain a meaningful interpretation of the principal component analysis, we need to reduce to fewer than nine (9) components. In this study, we use the common decision in which we retain only the component with about 80% of variance explained. Therefore, from column 3 i.e. extraction Eigen Values for the retained components, we observed that six components are retained together with their percentage of variance explained by each component. The cumulative variance give as well, shows that the first six components account for about 82.37% of the total variance in the data. Rencher (2002)

A component's Eigen value may be computed as the sum of its squared component loadings for the entire variable. A component's Eigen value divided by the number of variables (which equals the sum of variances because the variance of a standardized variance equals to 1.0) gives the percentage of variance in all the variables, which it explains. The ratio of Eigen values is the ratio of explanatory importance of the component with respect to the variable. And, so if a component has a low Eigen value less than the standardized variance i.e. 1, then it is

contributing little to the explanatory importance of variance in the variable and may be ignored as redundant with more important components.

Table 1: Total Variance Explained by each Component

Components	Initial Eigen Values			Extraction Eigen Values for the retained components		
S/NO	Total	% of variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.4726	27.47	27.47	2.4726	27.47	27.47
2	1.3276	14.75	42.22	1.3276	14.75	42.22
3	1.0777	11.97	54.20	1.0777	11.97	54.20
4	0.8765	9.74	63.94	0.8765	9.74	63.94
5	0.8360	9.29	73.23	0.8360	9.29	73.23
6	0.8231	9.15	82.37	0.8231	9.15	82.37
7	0.6276	6.97	89.35			
8	0.5257	5.84	95.19			
9	0.4330	4.81	100.00			

Table 2 shows the communalities which measures the percent of variance in a given row explained by all the components. That is, the communality is the squared multiple correlation for the variable using the components as predictors. Communality for a variable is the sum of squared components loadings for that variable (row) and is the percent of variance due to the variable explained by all the components. For full orthogonal Principal Component Analysis, the communality will be 1.0 and all of the variance in the variables will be explained by all the components. Which their number equals that of the variables and is written under initial. The extracted communalities, is the percent of variance in a given variable explained by the components are the extracted, which are normally fewer in number than the original variables which led the coefficient to be less than 1.0.

Table 2: Communalities Extracted By each Variable

Variables	Initial	Extraction
Economics	1.0000	0.9769
Geography	1.0000	0.7832
English Language	1.0000	0.8567
Hausa Language	1.0000	0.7856
Mathematics	1.0000	0.8570
Agric. Science	1.0000	0.8222
Biology	1.0000	0.8650
Chemistry	1.0000	0.6963
Physics	1.0000	0.7706

Table 3: Component Loading

Variables	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Economics	-0.3435	-0.0590	0.6317	0.4752	0.1008	-0.4694
Geography	-0.6309	-0.1332	0.0061	-0.1475	-0.5694	-0.1466
Eng. Language	-0.4650	0.0448	0.4910	-0.6077	0.1571	-0.0590
Hau. Language	-0.4235	0.6985	-0.0289	0.3117	-0.0172	0.1415
Mathematics	-0.5520	-0.9970	-0.2723	-0.0619	0.6793	-0.0549
Agric. Science	-0.4050	0.6056	-0.3188	-0.2570	-0.0845	-0.3417
Biology	-0.5623	0.0694	0.3291	0.0222	-0.0251	0.6592
Chemistry	-0.6901	-0.1269	-0.3273	0.3036	-0.0354	0.0508
Physics	-0.5470	-0.6472	-0.2130	-0.0065	-0.0791	-0.0289

Table 3 is going to be used for interpretation. Interpretations for component loadings in principal components are similar to interpretation of coefficients for factor analysis and coefficients in multiple regressions as well as canonical loadings in Canonical correlation Analysis. We want to have some criterion, which helps us determine which of these are large and which of these are considered to be negligible.

1. In Component 1, the percentage amount of variability explained, contributed by the coefficient of each variable. Economics has the highest coefficient with (-0.3435) followed by Agricultural Science (-0.4050) and Hausa Language (-0.4650), and to a lesser extent Chemistry (-0.6901).
2. Component 2, Hausa Language with (0.6985) has the highest contribution to the variability, followed by Agricultural Science (0.6056) and Biology (0.0694) and English Language (0.0448).
3. Component 3 is primarily related to Economics (0.6317), English Language (0.4910), and Biology (0.3291). As Economics increases, the other two subjects increase as well.
4. Component 4 is primarily related to Economics (0.6793), Hausa Language (0.3117) and Chemistry (0.3036). As Economics increases, the two other subjects increase as well, almost all the other subjects decrease.
5. Component 5 is primarily related to Mathematics (0.6592), and English Language (0.1517).
6. Component 6 is primarily related to Biology (0.6592) only. As From the Table 2 of communalities, it can be seen that all the causes are well represented, we can think of the value as multiple R^2 values for regression model predicting the variable of interest. The communality for a given variable can be interpreted as the proportion of variance in that variable explained by the 6 factors. In other word, if multiple regressions is performed on Biology against the 6 factors, therefore R^2 = 0.865 which is about 86.5% of the variable due to variation in Biology is explained by factors model. The results suggest that principal component analysis does the best job of explaining level of variation in Biology.

CONCLUSION

Principal component analysis was applied and showed six groups of closely inter-related subjects based on the fact that six components were used. It is also shown in Table 3 that values that close zero correlating a variable and a component can be dropped which indicates variable reduction. The strongest inter-related subjects are found in the beginning column of table 3 and decrease through the last column. The 27% of the variability captured by the inter-related variables is due to the contribution of all the subjects but Economics, Agricultural Science and Hausa Language contribute significantly. However, there should be more concentration on the subjects that contribute to the inconsistency in the students' performance.

REFERENCES

Anderson, T. W. (1958): *An Introduction to Multivariate statistical Analysis.* First Edition, John Wiley and Sons. New York.

Anderson R. L. Tathan and William C. Bank (1998): *Multivariate Data Analysis*, 5^{th} Edition Prentice Hall, New York.

Antonio D.; Francisco J.; and Eliseo N. (2010): A principal Component Analysis of the Spanish Volatility Term Structure. *International Research Journal of Finance and Economics. **4**:9 Issue 49 ISSN 1450-2887*

Everitt, B.S.; and Dunn G. (1991): *Applied Multivariate Data Analysis*. Edward Arnoid. London. Pp 219-220.

Hair, J.F., Anderson, R.E, Tathaman, R.L, & Black,W.C (1998): *Multivariate data Analysis*, Fifth edition. NJ. Prentice Hall.

Hussain, F.; Zubairi, Y.Z.; and Hussain, A.G, (2011): Some Application of Principal Component Analysis on Malaysian Wind Data. *Scientific Research and Essays*. **15**:3172-3181.

Jolliffe IT (1986). Principal Component Analysis. Springer-Verlag, pp. 1-6.

Marada K.V.; Kent J. T.; and Bibby J.M. (1979): *Multivariate Analysis*, fifth edition Academic Press Inc, London.

Rencher, A.C (2002): *Methods of Multivariate Analysis*. Second edition, John Wiley & Sons. Inc. New York.