APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO FACTORS CONTRIBUTING TO STUDENTS' FAILURE

ABDULMUAHYMIN A. SANUSI

Department of Mathematics and Computer Science, Federal University Kashere, Gombe state.

Abstract

Failure among higher institution students has been the order of the day, despite the effort of the Lecturing Staff in the Universities and Federal government along side with the huge amount of money dedicated to Education at higher level of study. In this regard, this research work tends to detect the factors that contribute consistently to the causes of students' failure in higher institution. The data used in this study includes the values obtained through the questionnaires distributed among the final year students of session 2016/2017, Federal University Kashere, Gombe state. The questionnaire is structured to contain some expressions as factors contributing to Students' failure. Some of the expressions used in the questionnaire are: Lack of Financial support and Sponsorship (LFSAS), Poor Family background on Conventional Education (PFBCE), Distress, poor health and Weather Condition (DPHWC), Lack of Determination, Focus and Time Management (LDATM), Poor Infrastructural and Social Amenities for Learning (PIASA), Inappropriate allocation of Course of Study to Students (IACSS), Previous CGPA (PCGPA) Lack of Orientation and Counselling on courses of study (LOACC), Insufficient Professionals and Experts on Education (IPAEE), Insufficient Materials and aids for learning (IMAAL) Lack of Lecturer to Students ratio (LLTSR). Descriptive statistics was ran, Bartlett's test was carried out to investigate the heterogeneity of variation among the factors and Principal Component Analysis was used to investigate the factors that contribute high variability among the set of factors of students' failure. Out of eleven (11) components, seven (7) Components is considered showing 82.55% variation captured by the factors contributing to students' failure, only the first component captured 27.36% variation out of the 82.55% variation;

among the eleven factors contributing to students' failure in the first component; the factors that contribute significantly to the percentage of variation captured are: Poor Infrastructural and Social Amenities for Learning (PIASA), Insufficient Materials and aids for learning (IMAAL), Insufficient Professionals and Experts on Education (IPAEE), Inappropriate allocation of Course of Study to Students (IACSS), Poor Family background on Conventional Education (PFBCE) and; Lack of Financial support and Sponsorship (LFSAS). In fact proper consideration should be given to these factors to cobweb the causes of students' failure.

KEYWORDS: *Descriptive Statistics, Bartlett's test, Principal Components, Factors contributing to students' failure, percentage of variance.*

Introduction

The reasons for students' failure are almost as complex as are the reasons we are unable to turn around underperforming students in vast numbers. These reasons are multifaceted and interrelated, compounding and exacerbating the problem of students' failure. Statistics show that one third of the students who enter high school will drop out before graduation and many students with secondary school certificate are barely able to read or write. The vast majority of students leaving our education system do not have the skills to earn a living in our increasingly technological society and international marketplace. Research shows that what is needed is not more money spent on education but an understanding of why students are turning off to learning and failing in higher institution. Students don't plan to fail in high school or in life. They unfortunately get derailed along the way by internal struggles, environmental and school factors. Looking at the statistics of the graduating students across the nation, calls us to action in examining how to help drive students to success during their days in the higher institution. It is for this reason, we are looking at the failures so everyone can be equipped to help higher institution students combat the failures and be positioned for success. Some of the major causes of students failing in the higher institution may be factorized into Lack of Financial support and Sponsorship. Poor Family background on Conventional Education, Distress, poor health and

Weather Condition. Lack of Determination, Focus and Time Management, Poor Infrastructural and Social Amenities for Learning, Inappropriate allocation of Course of Study to Students, Previous CGPA, Lack of Orientation and Counselling on courses of study, Insufficient Professionals and Experts on Education, Insufficient Materials and aids for learning and Lack of Lecturer to Students ratio.

By looking at the available data for the multivariate factors contributing to students' failure, the dimensionality of the data is reduced to give a simpler understanding of the data. The data obtained from the questionnaires distributed among the Final year students of Federal University kashere Gombe state, comprise a large dataset where readings recorded

comprise of numerous variables related to the factors that contribute to students' failure after a rigorous edition has been done to avoid biasness.

One of the common statistical approaches in analyzing the multivariate data is to obtain a simplified form of the data with few variables that captures the structure of the whole dataset. One classical approach in statistics is by using the Principal component analysis method Hussain et. al. (2011). These principal components, in turn, may be used in the subsequent statistical analyses. Principal component analysis (PCA), also known as empirical orthogonal function (EOF) analysis, essentially is used to reduce the dimensionality of large dataset which consists of a large number of interrelated variables to smaller components (Jolliffe,1986). In other words, PCA is one of the statistical techniques used on multivariate linear data whereby data transformation is applied in search of the relationships in multivariate data sets. Thus, the aim of PCA is to determine a few linear combinations of the original variables in order to summarize the data.

Essentially, Principal component is a one-sample technique applied to data with no groupings among the observations and any 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)

Examples of the application of PCA include the analysis of science students' results (Ran Vijay Singh et al., 2014), understanding the meteorological characteristics helps in predicting the Weather conditions Hussain et. al. (2011), classification of vegetable oils (Rusak et al., 2003), visualization of trace elemental pattern in vegetable after different cooking procedure (Pradova, 2001), identifying the sources of dimensional variation in the automotive body industry, modelling meteorological data (Mohanan, 2000) and many more.

Methodology

1. Descriptive Statistics

Basic descriptive statistics are calculated to 64 bit decimal precision avoiding any of the pocket calculator formulae that led to unnecessary lack of precision (McCullough and Wilson, 1999).

$$\text{Mean} = \overline{x} = \frac{\sum_{i=1}^{n}(x_i)}{n} \quad (1.1)$$

$$\text{Standard deviation} = S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (1.2)$$

2. Principal Component Analysis

The steps involved in the analysis of principal component analysis include the method of getting the data, standardizing the data, calculating the covariance matrix and visualizing the results. 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 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:

$$Y_1 = a_1^{'}X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Y_2 = a_2^{'}X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$
.
.
.
$$Y_p = a_p^{'}X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p \quad (2.1)$$

Step 2: standardize 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:

$Z = \frac{(X - \mu_i)}{\sqrt{\sigma_{1i}}}$, $i = 1, 2, ..., p$ In matrix notation, it is given by:

$$Z = (V^{1/2})^{-1}(X - \mu) \quad (2.2)$$

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

$$COV(Z) = (V^{1/2})^{-1} \sum (V^{1/2})^{-1} = \rho \quad (2.3)$$

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 2.3.

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

$$Y_i = e_i^{'}Z = e_i^{'}(V^{1/2})^{-1}(X - \mu) \ , i = 1, 2, ..., p \quad (2.4)$$

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

Moreover,

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i}^{p} Var(Z_i) = q \qquad (2.5)$$

and

$$\rho_{Y_i,Z_k} = e_{ik}\sqrt{\lambda_i}\ , \qquad i,k = 1,2,...,p \qquad (2.6)$$

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 3.47, 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:

$$\frac{\lambda_k}{q}, \qquad k = 1,2,...,p \qquad (2.7)$$

Where $\lambda_{k's}$ 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.

3. Tests for Significance

Bartlett's Test

The Bartlett's test of hypothesis is given as:

$H_o: \ \delta_1 = \delta_2 = \ldots = \delta_k$

$H_1: \ \delta_i \neq \delta_j \quad for\ atleast\ one\ pair\ (i, j)$

Test Statistic:

$$T = \frac{(N-k)\ln S_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln S_i^2}{1+(1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i-1)) - 1/(N-k))} \qquad (3.1)$$

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:

$$S_p^2 = \sum_{i=1}^{k}(N_i - 1)S_i^2/(N-k). \qquad (3.2)$$

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.

4. Deciding How Many Component to Retain

In every application, a decision must be made on how many principal components should be retained in order to effectively summarize the data. The following guide lines have been proposed:

(i). Retain sufficient components to account for a specified percentage of the total variance, say 80%.

(ii). Retain the components whose Eigen values are greater than the average of the eigen values, $\sum_{i=1}^{p} = \lambda_i / p$. For a correlation matrix, this average is 1.

(iii). Use the scree graph, a plot of (λ_i = Eigen values) versus (i = no of components), and look for a natural break between the ‘‘Large’’ Eigen values and the ‘‘Small’’ Eigen values (Rencher, 2002).

Data Used for the Analysis

The data used in this study includes the values generated from the respondents through the questionnaires distributed among the final year students of session 2016/2017, Federal University Kashere, Gombe state. The questionnaire is structured to contain some expressions as factors contributing to Students' failure with option provided to be ticked to show the degree of agreement in each expression. Some of the expressions used in the questionnaire are: Lack of Financial support and Sponsorship (LFSAS) Poor Family background on Conventional Education (PFBCE). Distress, poor health and Weather Condition (DPHWC). Lack of Determination, Focus and Time Management (LDATM). Poor Infrastructural and Social Amenities for Learning (PIASA). Inappropriate allocation of Course of Study to Students (IACSS). Previous CGPA (PCGPA). Lack of Orientation and Counselling on courses of study (LOACC). Insufficient Professionals and Experts on Education (IPAEE). Insufficient Materials and aids for learning (IMAAL) and Lack of Lecturer to Students ratio (LLTSR).

In short, the data obtained are of multivariate in nature. The purpose of the study is to design a statistical analysis for the multivariate data on factors contributing to students' failure using principal component analysis which focuses numerical values that can be obtained from the analysis. Using SPSS Version 15 and NCSS 2007, the study aims to describe the variations of the multivariate data by reducing the dimensionality of the principal components. In the study, the statistical model used in the PCA is described. Based on the Eigen values obtained in the analysis, the significant components are identified. The numerical results obtained in the analysis are presented in tabular form in which meaningful interpretations can be made from the numerical outputs.

Analysis and Discussion

Table 1: Descriptive statistics

Variable	Frequency	Mean	Standard deviation
LFSAS	108	2.1204	1.4255
PFBCE	108	2	1.3465
DPHWC	108	1.8519	1.4392
LDATM	108	2.0926	1.3009
PIASA	108	2.5278	1.3771
IACSS	108	2.2130	1.3743
PCGPA	108	1.3981	1.4072
LOACC	108	2.1389	1.4627
IPAEE	108	1.8056	1.4817
IMAAL	108	2.2315	1.4443
LLTSR	108	1.8796	1.5390

Table 1, is the descriptive statistics that shows the frequency, mean and standard deviation of the variables (factors) in used, indicating the factor with the highest mean among other means and the factor with the lowest standard deviation as to assume the expected average value and the standard variation that exist among the set of the factors contributing to students' failure.

Bartlett's test:

The Bartlett's test requires measuring the Homogeneity of variance across variables, i.e. the factors.

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* $p-value \prec \alpha = 0.05$

Table 2: Bartlett's Test

Chi-square value	215.79
Degree of freedom	55
Probability value	0.0000

From table 2 which shows the Bartlett's Test, indicating the Approximation chi-square equals 215.79 with degree of freedom of 55, probability level of 0.0000, at $\alpha = 0.05$, we therefore reject H_O and conclude that the variances across the variables are not equal.

In this regard, Principal Component Analysis is suitable for the set of the data; in order to verify the variables (i.e. factors) that contribute significant variation to the set of the principal components that may be considered.

Table 3 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 eleven (11) 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 seven components are retained together with their percentage of variance explained by each component. The cumulative variance give as well, shows that the first seven components account for about 82.55% of the total variance in the data. Rencher (2002)

Table 3: 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	3.0092	27.36	27.36	3.0092	27.36	27.36
2	1.6385	14.90	42.25	1.6385	14.90	42.25
3	1.1434	10.39	52.65	1.1434	10.39	52.65
4	1.0498	9.54	62.19	1.0498	9.54	62.19
5	0.845560	7.69	69.88	0.845560	7.69	69.88
6.	0.765165	6.96	76.83	0.765165	6.96	76.83
7.	0.628955	5.72	82.55	0.628955	5.72	82.55
8.	0.588445	5.35	87.90			
9.	0.500866	4.55	92.45			
10.	0.440858	4.01	96.46			
11.	0.389278	3.54	100.00			

Table 4 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.

Table 4: Communalities Extracted By each Variable

Variable	Initial	Extraction
LFSAS	1.0000	0.7639
PFBCE	1.0000	0.7584
DPHWC	1.0000	0.6764
LDATM	1.0000	0.9121
PIASA	1.0000	0.8443
IACSS	1.0000	0.9435
PCGPA	1.0000	0.8218
LOACC	1.0000	0.8113
IPAEE	1.0000	0.7808
IMAAL	1.0000	0.7711
LLTSR	1.0000	0.5461

Table 5 is the 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 (i.e. above 0.5) and which of these are considered to be negligible (i.e. below 0.5), irrespective of the negative signs in the first four components considered with Eigen values greater than one (1); we therefore interpret as follows:

1. In Component 1, the percentage amount of variability explained, contributed by the coefficient of each variable. PIASA has the highest coefficient with (0.6364) followed by IMAAL (0.5883), IPAEE (0.5853), IACSS (0.5611), PFBCE (0.5523) and LFSAS (0.5227).
2. In Component 2, LFSAS with (0.6319) has the highest contribution to the variability, followed by PFBCE (0.5734).
3. In Component 3, PCGPA is the only significant variable with coefficient value 0.7764.
4. Component 4 is primarily related to LDATM with 0.5731.

Table 5: Component Loading

Variables	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7
LFSAS	0.5227	0.6319	0.1784	0.1747	0.0425	0.1372	-0.0929
PFBCE	0.5523	0.5734	0.1339	0.1920	-0.0097	0.2642	-0.0089
DPHWC	0.4417	0.4868	-0.2852	-0.3888	-0.0675	0.0674	0.0520
LDATM	0.5208	0.0025	0.2745	0.5731	-0.1750	-0.1424	0.4315
PIASA	0.6364	-0.1317	0.3804	-0.1193	0.0679	-0.1606	-0.4825
IACSS	0.5611	0.0159	0.0327	0.3099	0.3137	-0.6427	0.1399
PCGPA	0.3881	0.1052	0.7764	0.2031	-0.0212	-0.0901	0.0861
LOACC	0.5007	-0.3258	-0.4140	-0.3999	-0.0379	-0.0115	-0.3488
IPAEE	0.5853	-0.3837	-0.0135	0.3966	-0.3010	0.2018	0.0472
IMAAL	0.5883	-0.4960	0.1295	0.1645	-0.3259	0.0984	0.1388
LLTSR	0.3934	-0.3812	-0.0107	-0.0830	0.7117	0.3926	0.1683

CONCLUSION AND RECOMMENDATION

The multivariate set of data collected was analyzed using Descriptive statistics, Bartlett's test for homogeneity of variance among the factors and Principal component analysis, seven groups of closely inter-related factors based on the fact that the first component was used. It is also shown in Table 5 that values that close zero correlating a variable and a component can be dropped which indicates variable reduction. The strongest inter-related factors are found in the beginning column of table 5 and decrease through the last column. The 27.36% of the variability captured by the interrelated variables in the first component is due to the contribution of all the factors but Poor Infrastructural and Social Amenities for Learning (PIASA), Insufficient Materials and aids for learning (IMAAL), Insufficient Professionals and Experts on Education (IPAEE), Inappropriate allocation of Course of Study to Students (IACSS), Poor Family background on Conventional Education (PFBCE) and Lack of Financial support and Sponsorship (LFSAS) contribute significantly. However, the Authority and the Management board of Federal University Kashere Gombe state, is advised to give absolute consideration and provide necessary solutions to these factors that contribute significantly to students' failure in the institution.

Reference:

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*

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.

McCullough BD, Wilson B. On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis* 1999;31:27-37.

Mohanan AH (2000). Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System. J. Clim., 13(4): 821-835.

Parry ML (ed.). 2000. *Assessment of Potential Effects and Adaptations for Climate Change in Europe: The Europe ACACIA Project*. Jackson Environment Institute, University of East Anglia: Norwich.

Pradova V (2001). Three-way Principal Component Analysis for the Visualization of Trace Element Patterns in Vegetables after Different Cooking Procedures. J. Food Compost. Anal., 14(2): 207-225.

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

Rusak DA, Brown L, Martin SD (2003). Classification of Vegetable Oils by Principal Component Analysis of FTIR Spectra. J. Chem. Educ., 80(5): 541-548.

Singh et. el (2014). Analysis of Science students' SSCE results using Principal component Analysis. Continental Journal Education Research. 7(1): 24 - 29