Journal of the Nigerian Association of Mathematical Physics
***Volume 51* (*May, 2019 Issue*), pp315 – 320**
© J. of NAMP

THE USE OF PRINCIPAL COMPONENTS ANALYSIS (PCA) AS A TECHNIQUE FOR DATA REDUCTION

[1]Bello A. H., [2]Folorunsho A.I., [3]Bamigbade T. K. and [4]Obiyemi D. A.

[1]Department of Statistics, School of Sciences, Federal university of Technology, P.M.B. 704, Akure. Nigeria.
[2,3,4]Department of Mathematics/Statistics, Faculty of Science, Osun State Polytechnic, P.M.B. 301, Iree. Osun State, Nigeria.

Abstract

Principal Component Analysis is a multivariate statistical technique that is often useful in dimensionality reduction of a collection of large number of variables to a smaller number of variables without loss in the analytical objectives. This study focuses on the application of Principal Component Analysis (PCA) as a method of analysing inter-correlated variables (subjects) with large data matrix in the academic performance of secondary school students in Nigeria. Shapiro Wilk Test was carried out to check for normality of the data and it was discovered that some of the variables are not normally distributed and transformation was done to normalize the data. The results of the analysis revealed that the first 4 components account for 60.14% of the variability present in the data set that would be retained, since they have an eigenvalue greater than 1. The first component with eigenvalue 3.41648 accounts for 28.47% of the variability in the data set, second PC accounts for 11.76%, the third PC accounts for 11.23% and the fourth accounts for 8.68% of the variability present in the data set. In conclusion the first four components in the PCA are to be retained and then the equation of the principal components is derived after transformation.

***Keywords*:** Principal Component Analysis (PCA), Shapiro Wilk Test, Normal Distribution, Inter-correlation and Transformation.

1.0 Introduction

Principal components analysis is concerned with explaining the variance-covariance structure of a set variables through a few linear combinations of these variables. Its general objectives are data reduction and data interpretation. Although principal components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number K of the principal components. The analysis of principal component often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily have been the result. Analysis of principal components are more of a means to an end and rather than end in themselves, because they frequently serve as intermediate steps in much larger investigations. For instance, principal components maybe input to a multiple regression or cluster analysis.

This study is on the academic performance of students in secondary schools at the junior level and is generally judged by their overall performance in the numerous subjects they offered. The data used was collected from the 2nd term result of a JSS 3 class in Life Spring High School, Akure, Ondo state. Nigeria. However, of the twelve (12) subjects considered for this study, some of them tend to measure the same construct (i.e. multicollinearity is suspected). The aim is to reduce the twelve (12) variables into fewer ones. Hence, the need for dimensionality reduction to reduce the numerous highly correlated variables to a fewer uncorrelated variables (the principal components) in order to eliminate redundancy.

The origin of statistical techniques is often difficult to trace [1]. When a large number of random variables are available for potential study, it may be of interest to inquire initially whether they can be replaced by fewer number of random variables, either a subset of the original or certain functions of them, without loss of much information [2].It can be noted by[3, 4, 5]

Corresponding Author**:** Bello A.H., Email: habello@futa.edu.ng, Tel: +2347033420672

that one can independently derived the singular value decomposition (SVD) in a form that underlies PCA, while [6] used the SVD in the context of two-way analysis of an agricultural trial. However, it is generally accepted that the earliest descriptions of the technique now known as Principal Component Analysis (PCA) were given by [7] and [8]. Hotelling's motivation was that there may be a smaller 'fundamental set of independent variables which determines the values' of the original p variables. He noted that such variables have been called 'factors' in psychological literature, but introduced the alternative term 'components' to avoid confusion with other uses of the word 'factor' in Mathematics. Hotelling's derivation of PCs uses Lagrange multipliers and ended up with an eigenvalue/eigenvector problem. The asymptotic sampling distributions of the coefficients and variances of the sample PCs as discussed [9] and building on the earlier work by [10], has been frequently cited in subsequent theoretical development. The paper by [2] is remarkable for the large number of new ideas concerning uses, interpretations and extensions of PCA; that it introduced. The links between PCA and various statistical techniques as discussed by [11] also provided a number of important geometric insights. The practical side of the subject by [12]also gave an impetus into discussing two case studies in which the uses of PCA go beyond that of a simple dimension reducing tool. Principal Component Analysis may be used under many situations generally as an exploratory instrument to enable us know what is the effective number of dimensions in a dataset or how dominant are certain linear combinations of the variables [13]. The method could also be used when there is high degree of multicollinearity in a data set and the research interest is in determining the fewer sets of variables that could be used for regression analysis. However, PCA can be used for purposes other than in regression analysis. For instance, it is an objective method for the construction of index numbers such as for economic development, production, general price index, etc. Principal components analysis (PCA) summarizes the major variation or information that is contained in many dimensions into a reduced number of uncorrelated dimensions. PCA is an appropriate tool for variable selection and the use of PCA to discard redundant variables has been outlined in [1].

2.0. Methodology

Assuming that there are p variables, we are interested in forming the following p linear combinations:

$$Y_1 = w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p$$

$$Y_2 = w_{21}x_1 + w_{22}x_2 + \ldots + w_{2p}x_p$$

.

.

.

$$Y_p = w_{p1}x_1 + w_{p2}x_2 + \ldots + w_{pp}x_p$$

Where $Y_1, Y_2, \ldots, Y_p$ are the p principal components and w_{ij} is the weight of the j^{th} variable for the i^{th} principal component. Alternatively, Principal Component Analysis can be done by finding the Singular Value Decomposition (SVD) of the data matrix or a spectral decomposition of the covariance matrix.

Eigen Structure of the Covariance Matrix.

Let X be a p-component random vector where p is the number of variables. The covariance matrix, $\sum$, is given by:
$$V(X) = E[X - E(X)][X - E(X)]' \qquad (1)$$

Let $\gamma' = (\gamma_1, \gamma_2 \ldots \gamma_p)$ be a vector of weights to form the linear combination of the original variables, and $Y = \gamma' X$ be the new variable which is a linear combination of the original variable. The variance of the new variable is given by

$$V(Y) = V(\gamma' X)$$
$$= \gamma' V(X)\gamma$$
$$= \gamma' \Sigma \gamma \qquad (2)$$

The problem is reduced to finding the weight vector, γ' such that the variance $\gamma' \Sigma \gamma$ of the new variable is maximum over the class of linear combinations that can be formed subject to the constraint $\gamma' \gamma = 1$.

The first principal component, therefore, is given by the eigenvector, γ_1, corresponding to the largest eigenvalue, λ_1. Let γ_2 be the second p-component vector of weights to form another linear combination.

The next linear combination can be found such that the variance of $\gamma_2' X$ is the maximum, subject to the constraint $\gamma_1' \gamma_2 = 0$ and $\gamma_2' \gamma_2 = 1$.

$\gamma_2^{'}$ is the corresponding eigenvector of λ_2 , the second largest eigenvalue of $\sum$

Similarly, the remaining PC, $\gamma_3^{'}, \gamma_4^{'} ... \gamma_p^{'}$, are the eigenvectors corresponding to the eigenvalues, $\lambda_3, \lambda_4, ..., \lambda_p$, of the covariance matrix, $\sum$

Singular Value Decomposition

Singular Value Decomposition (SVD) expresses any (n×p) matrix, (where n≥p) as a triple product of three matrices, P, D and Q such that

$$X = PDQ^{/} \quad (3)$$

where X is an (n×p) matrix of column rank r,

P is an (n×r) matrix,

D is an (r×r) diagonal matrix,

Q′ is an (r×p) matrix.

The matrices P and Q are orthonormal; that is P′P =I and Q′Q =I.

The P column of Q′ contain the eigenvectors of the X′X matrix and the diagonals of the D matrix contain the square root of the corresponding eigenvalues of the X′X matrix.

Also, the eigenvalues of the matrices X′X and XX′ are the same.

Singular Value Decomposition of the Data Matrix

Let X be an n×p data matrix. Since X is a data matrix, it will be assumed that its rank is p (i.e. r=p) and consequently Q will be a square symmetric matrix.

The columns of Q will give the eigenvectors of the X′X matrix and the diagonal values of the D matrix will give the square root of the corresponding eigenvalues of the X′X matrix.

Let Y be an n×p matrix of the values of the new variables or principal component scores.

Then,

$$Y = XQ$$
$$=(PDQ')Q$$
$$=PDQ'Q$$
$$=PD \quad (4)$$

The covariance matrix, $\sum_Y$, of the new variables is given by:

$$\sum_Y = E(YY^{'}) = E[(PD)'(PD)]$$
$$=E(D'P'PD)$$
$$=E(D^2)$$
$$= \frac{1}{n-1}D^2 \quad (5)$$

Since D is a diagonal matrix. The new variables are uncorrelated among themselves.

As can be seen from the preceding discussion, the SVD of the data matrix also gives the principal components analysis solution.

The weights for forming the new variables are given by the matrix Q, the principal components scores are given by PD, and the new variables are given by $\frac{1}{n-1}D^2$.

Shapiro – Wilk Test for Normality

The Shapiro-Wilk test, proposed in 1965, calculates a W statistic that tests whether a random sample, $x_1, x_2, \cdots\cdots\cdots, x_n$ comes from (specifically) a Normal population.

Hypothesis (Test for Normality)

H_0: Samples came from Normally distributed population

H_1: samples do not come from normally distributed population

Test Statistic: $W = \frac{(\sum_{j=1}^{n} a_j x_{(j)})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ (6)

Decision Rule: Reject H_0 in favour of H_1 at α = 0.05 level of significance if $p - \text{value} < \alpha, otherwise\ do\ not\ reject\ H_0$.

3.0 Results and Discussion

Table 1, shows the descriptive statistics such as the mean and the standard deviation of the original variables. As can be seen, integrated science has the minimum mean of 35.85 with standard deviation of 9.449 while P.H.E has the maximum mean of 68.77 and standard deviation 4.481.

Table 1: Descriptive statistics of the twelve variables under consideration.

Variables	Minimum	Maximum	Sum	Mean	Std. Deviation
English language	41	72	3519	58.65	6.739
Mathematics	51	67	3441	57.35	3.668
Integrated Science	18	57	2151	35.85	9.449
Social Studies	50	77	3962	66.03	6.098
Introductory technology	47	75	3455	57.58	6.400
Business study	40	88	3837	63.95	11.317
Home Economics	48	90	4008	66.80	11.140
Cultural and creative Art	37	63	2864	47.73	4.888
French	48	75	3415	56.92	5.347
Agricultural Science	34	87	3494	58.23	11.278
Computer science	40	73	3056	50.93	6.033
P.H.E	50	77	4126	68.77	4.481

Table 2: Correlation Matrix Associated with the data set

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	1	0.371	0.216	0.146	0.244	0.365	0.401	0.053	0.469	0.392	0.048	0.185
X_2	0.371	1	-0.122	-0.187	-0.018	-0.08	-0.075	-0.31	0.053	-0.122	-0.148	-0.112
X_3	0.216	-0.122	1	0.763	0.485	0.733	0.503	0.671	0.701	0.348	0.421	0.813
X_4	0.146	-0.187	0.763	1	0.391	0.85	0.256	0.852	0.631	0.528	0.616	0.902
X_5	0.244	-0.018	0.485	0.391	1	0.502	0.436	0.282	0.798	0.705	0.535	0.617
X_6	0.365	-0.08	0.733	0.85	0.502	1	0.296	0.826	0.779	0.57	0.549	0.897
X_7	0.401	-0.075	0.503	0.256	0.436	0.296	1	0.154	0.527	0.307	0.168	0.354
X_8	0.053	-0.31	0.671	0.852	0.282	0.826	0.154	1	0.526	0.413	0.454	0.823
X_9	0.469	0.053	0.701	0.631	0.798	0.779	0.527	0.526	1	0.619	0.593	0.808
X_{10}	0.392	-0.122	0.348	0.528	0.705	0.57	0.307	0.413	0.619	1	0.548	0.575
X_{11}	0.048	-0.148	0.421	0.616	0.535	0.549	0.168	0.454	0.593	0.548	1	0.666
X_{12}	0.185	-0.112	0.813	0.902	0.617	0.897	0.354	0.823	0.808	0.575	0.666	1

The variables X_1, X_2,…………………. X_{12} are represented below

English language X_1
Mathematics X_2
Integrated science X_3
Social studies X_4
Introductory technology X_5
Business studies X_6
Home economics X_7
Cultural and arts X_8
French X_9
Agricultural science X_{10}
Computer science X_{11}
P.H.E X_{12}

Table 3: Shapiro-Wilk test for normality of data.

Subject	Z	p-values	Decision	Conclusion
English Language	0.372	0.35488	Accept	Distributed normal
Mathematics	1.366	0.08594	Accept	Distributed normal
Integrated science	1.697	0.04482	Reject	Not Distributed normal
Social Studies	2.955	0.00156	Reject	Not Distributed normal
Introductory tech.	2.237	0.01264	Reject	Not Distributed normal
Business Studies	0.518	0.30215	Accept	Distributed normal
Home Economics	1.773	0.03808	Reject	Not Distributed normal
Cultural and Art	-0.531	0.70233	Accept	Distributed normal
French	1.870	0.03075	Reject	Not Distributed normal
Agricultural Science	-0.562	0.71304	Accept	Distributed normal
Computer	3.847	0.00006	Reject	Not Distributed normal
P.H.E	2.975	0.00147	Reject	Not Distributed normal

From table 3 above, the variables whose p-values < 0.05 are not normally distributed while those whose p-values > 0.05 are distributed normal, using the Shapiro Wilks test for normality. Thus, English, Mathematics, Business Studies, Cultural Art and Agric Science are distributed normal while Integrated Science, Social Studies, Introtech, Home Econs, French, Computer and P.H.E are not distributed normal. Hence the need for data transformation.

Table 4: Results of Principal Component Analysis(Before Transformation)

Component	Eigenvalue	Difference	Proportion	Cumulative
Component 1	3.47772	2.05045	0.2898	0.2898
Component 2	1.42727	.115315	0.1189	0.4087
Component 3	1.31196	.270915	0.1093	0.5181
Component 4	1.04104	.0692355	0.0868	0.6048
Component 5	.971806	.136234	0.0810	0.6858
Component 6	.835571	.0370868	0.0696	0.7554
Component 7	.798485	.128476	0.0665	0.8220
Component 8	.670009	.159849	0.0558	0.8778
Component 9	.51016	.15445	0.0425	0.9203
Component10	.35571	.0426642	0.0296	0.9500
Component11	.313045	.0258207	0.0261	0.9761
Component12	.287225		0.0239	1.0000

Data Transformation

From the analysis it can be observed that: P.H.E., Home Economics, and Social Studies can be transformed by Cubic. Computer Science, Introductory technology and French by inverse of square, while Integrated Science can be transformed by Log.

Results of PCA After Transformation of Data

Table 5: Eigenvalues and proportion of variance explained by the components

Component	Eigenvalue	Difference	Proportion	Cummulative
Component 1	3.41648	2.00547	0.2847	0.2847
Component 2	1.411	0.0631791	0.1176	0.4023
Component 3	1.34782	0.305894	0.1123	0.5146
Component 4	1.04193	0.0486137	0.0868	0.6014
Component 5	0.993315	0.151384	0.0828	0.6842
Component 6	0.841931	0.00919103	0.0702	0.7544
Component 7	0.83274	0.190714	0.0694	0.8238
Component 8	0.642026	0.129453	0.0535	0.8773
Component 9	0.512573	0.149687	0.0427	0.9200
Component 10	0.362885	0.0384821	0.0302	0.9502
Component 11	0.324403	0.0515057	0.0270	0.9773
Component 12	0.272897		0.0227	1.0000

The first 4 components accounting for 60.14% of the variability present in the data set would be retained, since they have an eigenvalue greater than 1. The first component with eigenvalue 3.41648 accounts for 28.47% of the variability in the data set, second PC accounts for 11.76%, the third PC accounts for 11.23% and the fourth accounts for 8.68% of the variability present in the original data set.

From Appendix, the equation of the principal components are;

$$y_1 = 0.3920 english + 0.2653 mathematics + \cdots + 0.1297 PHE \quad (7)$$

$$y_2 = 0.1242 english - 0.4527 mathematics + \cdots + 0.2704 PHE \quad (8)$$

$$y_3 = 0.1479 english - 0.1130 mathematics + \cdots + 0.4715 PHE \quad (9)$$

$$y_4 = 0.1404 english - 0.1944 mathematics + \cdots + 0.1677 PHE \quad (10)$$

Conclusion

Based on the results obtained from the analyses of the data using the PCA, it was observed that after the transformation of seven of the variables which were not normally distributed alongside the five normally distributed variables, using the eigenvalue greater than 1 criterion, four components with respective eigenvalues 3.41648, 1.411, 1.34782, 1.04193 were eventually retained. Though it appears there is no significant difference between the results obtained before and after transformation (comparing tables), we are certain that the normality assumption for PCA holds for the data after transformation. Ideally, the performance of data transformation method should be assessed objectively and quantitatively under different circumstances.

Appendix: Component score coefficient matrix associated with the data set

Subjects	Component 1	Component 2	Component 3	Component 4
English	0.3920	0.1242	0.1479	0.1404
Maths	0.2653	-0.4527	-0.1130	-0.1944
Inter Science	0.2187	0.4888	0.1833	-0.1696
Social Study	0.1496	-0.4524	0.2905	0.5757
Introtech	-0.1611	0.0298	0.5780	-0.0573
Business Study	0.3945	0.0309	0.1859	-0.1146
Home econs	0.4133	-0.2296	0.1497	-0.0581
Cultural and art	0.1534	0.3394	-0.2276	0.4793
French	-0.2967	0.1650	0.1908	-0.1296
Agric Science	0.3526	0.0003	-0.0065	-0.5094
Computer	0.3171	0.2568	-0.3817	0.1685
P.H.E	0.1297	0.2704	0.4715	0.1677

References

[1] Jolliffe I. T., (2002) *Principal Component Analysis* (2nd ed.), New York: Springer-Verlag Inc.

[2] Rao, C. R (1964). The *use and interpretation of Principal Component Analysis in applied research*. Vol. 26, pp 329-358.

[3] Preisendorfer, R. W and Mobley, C. D (1988). *Principal Component Analysis in Meteorology and Oceanography,* Amsterdam. Elsevier.

[4] Beltrami, E. (1873). *Sulle Funzioni Bilineari. Giornale di Mathematiche di Battaglini*, Vol. 11, pp 98-106.

[5] Jordan, M. C. (1874) *Memoire Sur Les Formes Bilineaires*. J. Math. Pures Appl., Vol. 19, pp 35-54.

[6] Fisher, R. A and Mackenzie, W. A (1923). *Studies in Crop Variation II. The manorial response of different potato varieties.* J. Agric. Sci., Vol. 13, pp 311-320.

[7] Pearson, K. (1901) On *lines and planes of closest fit to systems of points in space*. Phil. Mag. (6), Vol. 2, pp 559-572.

[8] Hotelling, H. (1933). *Analysis of a complex of statistical variables into Principal Components.* J. Educ. Psychol., Vol. 24, pp 417-441, 498-520.

[9] Anderson T. W., (2003).*An Introduction to Multivariate Statistical Analysis* (3rd ed.), New Jersey: John Wiley & Sons Inc.

[10] Girshick, M. A (1939). *On the sampling theory of roots of determinant equations.* Ann. Math. Statis., Vol. 10, pp 203-224.

[11] Gower, J. C. (1966). *Some distance properties of latent root and vector methods used in Multivariate Analysis*. Biometrika, Vol. 53, pp 325-338.

[12] Jeffers, J. N. R. (1967). *Two case studies in the applications of Principal Component Analysis*. Appl. Statis., Vol. 16, pp 225-236.

[13] Kendall, M. C and Stuart, A (1976). *The Advanced Theory of Statistics, Vol. I, II, III.* London, Griffin and Co., Ltd..