COMPARATIVE STUDY OF TWO DATA REDUCTION TECHNIQUES IN PRINCIPAL COMPONENT ANALYSIS

BY

HARUNA SULEIMAN

DEPARTMENT OF MATHEMATICS
AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA

AUGUST, 2015

COMPARATIVE STUDY OF TWO DATA REDUCTION TECHNIQUES IN PRINCIPAL COMPONENT ANALYSIS

BY

Haruna SULEIMAN, B.Sc. STATISTICS (UNIMAID) 2008
MSC/SCI/08691/2010-2011

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, AHMADU BELLO UNIVERSITY, ZARIA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A
MASTERS DEGREE IN STATISTICS.

DEPARTMENT OF MATHEMATICS,
FACULTY OF SCIENCE
AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA

AUGUST, 2015

DECLARATION

I hereby declare that this Dissertation Titled "Comparative Study of Two Data Reduction Techniques in Principal Component Analysis" has been written by me in the Department of Mathematics under the supervision of Prof. S. U. Gulumbe and Dr. H. G. Dikko. All information derived from the literature has been duly acknowledged in the text and a list of reference provided. No part of this Dissertation is previously presented for another degree at any University.

HARUNA SULEIMAN **DATE**

CERTIFICATION

This Dissertation titled “Comparative Study of Two Data Reduction Techniques in Principal Component Analysis” by Haruna Suleiman met the regulations for the award of the degree of Master of Statistic of Ahmadu Bello University (ABU) Zaria, and is approved for its contribution to knowledge and literary presentation.

Chairman, Supervisory Committee Dr. H.G. Dikko	Date
Member, Supervisory Committee Prof. S. U. Gulumbe	Date
Head of Department Prof. B. Sani	Date
Dean, School of Postgraduate Studies Prof. Kabir Bala	Date

DEDICATION

This Dissertation is dedicated to my biological parents; Alhaji Ja'afar Suleiman Dan'asabe Maichitta and Hajiya Aishatu Ja'afar Suleiman, who gave me a tight and sound moral training and also to my late uncles; Alhaji Abdurra'uf Suleiman Dan'asabe Maichitta and Alhaji Salihu Suleiman Dan'asabe Maichitta.

ACKNOWLEDGEMENTS

All praise is due to Allah, the Lord of the universe, the Lord of mankind, the Master of the Day of Judgment, the day of resurrection and compensation. May the peace and blessing of Allah the Almighty, the omnipotent the omnipresent, be upon our Noble Prophet Muhammad (PBUH).

I am most grateful to my Major Supervisor Dr. H. G. Dikko for his encouragement, critical examinations and criticisms throughout this research work. I am also indebted to my second supervisor, Prof. S.U Gulumbe whose efforts to inculcate the skills and techniques I have employed for the achievement of this Dissertation can never be over emphasized.

My worthy appreciation also goes to the Dean of Postgraduate Studies, Prof. Kabir Bala, Dean Faculty of Sciences, Prof. A.K. Adamu, Head of Mathematics Department, Prof. Babangida Sani and the Departmental Postgraduate Coordinator Dr. Abubakar Yahaya to mention but a few and also to the entire Academic Staff of Ahmadu Bello University.

My sincere appreciation goes to my dear friends and course mates, Mallam. Sani Salihu, H.O.D Department of Statistics, Kano University of Science and Technnology, Wudil. Mallam. Jamilu Garba of Mathematics Department A.B.U. Zaria for their supports, encouragements and contributions.

My profound gratitude goes to my biological father Alhaji. Ja'afar Suleiman Dan'asabe Maichitta who sponsored my education right from elementary level to Masters Degree and also my caring mother Hajiya Aishatu Ja'afar Suleiman for her encouragement and discipline. May Almighty Allah reward them abundantly.Ameen.

I will also like to use this medium as a privilege to offer my special thanks and appreciation to my first mentor Prof. I.K Zubairu of the Department of Building A.B.U Zaria and to my Second

mentor Dr. Ibrahim Abdullahi of Department of Statistics, North-West University, Kano for their encouragement and advise.

I will also like to use this medium as a prerogative to acknowledge Mallam Mukhtar Aliyu for putting me through on how to conduct my analysis using Matlab Package and to the entire Academic and non-academic Staff of Faculty of Computing and Mathematical Sciences, Kano University of Science and Technology, Wudil whose names are too numerous to be mentioned, may Allah reward them all.

Finally, my special thanks goes to the entire members of my family and to my lovely wife Mu'azatu, Abdulra'uf Haruna and my little son Abdurra'uf Haruna whose understandings strengthen my effort throughout the conduct of my research work. May Allah reward you all with eternal life in Paradise. Ameen.

ABSTRACT

Principal Components Analysis is a mathematical procedure of data reduction technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables. This transformation is defined in such a way that the first principal components have the largest possible variance. Different methods for selecting the principal component to be retained exist in many literatures. In this dissertation, a comparative study of the two methods was carried out. The two methods considered are "proportion of variance accounted for" and "eigenvalue one-criterion". The method was applied on data of Cholesterol level in Human for PC's selection. Model adequacy checking was used to test the fitted models. It was found that six principal components were retained by using "proportion of variance accounted for" R-squared and R-squared-Adjusted from this research were 6.5% and 2.3%. For the "eigenvalue one-criterion" three principal components were retained, R-squared and R-squared adjusted were 10.9%, and 3.7% respectively. By considering the result obtained it is clear that eigenvalue one criterion is more preferred and desired in dimensionality reduction in Principal Component Analysis, for the fact that the first three PC's could replace twelve variables by sacrificing a negligible information about the total variation in the system. For authentication, simulation was carried out, R-squared and R-squared adjusted of the fitted model were 28.3% and 7.1% by eigenvalue one-criterion and 23.6% and 5.3% by proportion of variance accounted for respectively. So this authentication, shows without any reasonable doubt that eigenvalue one-criterion method is more preferable in the reduction of a set of data for cholesterol level in human body over the proportion of variance accounted for, also for the fact that the first five PC's retained account for most of the variance, and could replace twelve variables by sacrificing a negligible information about the total variation in the system.

TABLE OF CONTENTS

DECLARATION iii
CERTIFICATION iv
DEDICATION v
ACKNOWLEDGEMENTS vi
ABSTRACT viii
TABLE OF CONTENTS ix
LIST OF TABLES xi
LIST OF FIGURES xii
1.0 CHAPTER ONE: INTRODUCTION 1
1.1 Background of the Study 1
1.1.1 Application of Principal Component Analysis 3
1.2 Statement of the Problems 4
1.3 Aim and Objectives 4
1.4 Significance of the Study 5
1.5 Scope and Limitation 5
1.6 Method of Data Collection 5
1.7 Method of Data Analysis 6
1.8 Definition of Terms 6
2.0 CHAPTER TWO: LITERATURE REVIEW 9
2.1 Development of PCA (Historical Perspective of PCA) 9
2.2 Significance of Principal Component Analysis 9
2.3 Advantages of Principal Component Analysis 10
2.4 Restraints of Principal Component Analysis 11
2.5 The Eigen Value One-Criterion 12
2.6 Proportion of Variance Accounted for 14
3.0 CHAPTER THREE: METHODOLOGY 15
3.1 Preamble 15
3.2 Procedures Involved In Carrying Out PCA 15
3.3 Computation of Principal Component using Cumulative Proportion Procedure 18
3.4 Variance-Covariance Matrix 20

3.4.1 Variance 20
3.4.2 Covariance 21
3.4.3 Variance- Covariance 21
3.5 Correlation Matrix Computation Illustration (Standardization) 22
3.6 *R*-(Squared) - Coefficient of Determination 23
4.0 CHAPTER FOUR: DATA ANALYSIS 25
4.1 Data Presentation: Analysis and Discussions 25
4.2 Illustrations 25
4.3 Computed variance-covariance matrix from the original data set. 27
4.4 Results Obtained by the Real Life Data Set 31
4.5 Analysis Using Simulated Data 33
4.6 Computed variance-covariance Matrix of the Simulated data set. 35
4.7 Results by the Data Simulated from the Real Life Data 41
4.8 Contribution to Knowledge 41
5.0 CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS 43
5.1 Summary 43
5.2 Conclusion 43
5.3 Recommendation 44
Publication 45
References 46
Appendices 48

LIST OF TABLES

TABLE 4.1: Data set of the 12 different variables of 40 Observations..... 36

TABLE 4.2: Variance-covariance Matrix from the original data set …….. 38

TABLE 4.3: Correlation matrix from the variance covariance matrix …… 39

TABLE 4.4: The three retained PC's computed by MATLAB package …….. 41

TABLE 4.5: The Simulated data of 12 variables of 40 observations..... 44

TABLE 4.6: Variance-covariance matrix of the simulated data……. 46

TABLE 4.7: Correlation matrix computed from the variance-covariance matrix of the Simulated data …… 46

LIST OF FIGURES

Figure 3.1: Stages in Principal Component Analysis.…….……..…….. 15

1.0 CHAPTER ONE: INTRODUCTION

1.1Background of the Study

Principal Component Analysis (PCA) – is a statistical analytical tool that is used to explore, sort and group data. What PCA does is to take a large number of correlated (interrelated) variables and transform them into a smaller number of uncorrelated variables (Principal Components) while retaining maximal amount of variation, thus, making it easier to operate the data and make predictions using the retained PC as a predictor or Criterion Variables in Subsequent Analysis. According to Smith (2002) "PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences" since patterns can be hard to find in high- dimension, where the luxury of graphical representation is not available, PCA can be a powerful tool for data analysis.

Principal components (PC's) are linear combinations of random or statistical variables having special properties in terms of variance. For instance, the first PC is a normalized linear combination (that is the sum of squares of the coefficients being one) with the maximum variance. In effect, transforming the original vector of variable to the vector of PC's amount to a rotation of coordinates system that has inherent statistical properties. This choice of a coordinate system is to be contrasted with many problems that are to be treated, where the coordinate system is unrelated.

The PC's turn out to be the characteristics vector of the covariance matrix or the correlation matrix. Thus, the study of PC's can be considered as putting into statistical terms the usual developments of characteristics roots and vectors (for positive semi definite matrices).

From the point of view of statistical theory, the set of principal components yields a convenient set of coordinates, and the accompanying variables of the components characterize their statistical properties. In statistical practice, the method of PC's is used to find the linear combinations with large variance, Hotelling (1933).

In PCA 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 variables in the data as possible), and each succeeding component in turn has the highest variance possible under the constraints that it be orthogonal to (i.e. uncorrelated with) the preceding component. PCA is sensitive to the relative scaling of the original variables. Depending on the field of application; it is also named the discrete Karhunen – Loe've transform (KLT), the Hotelling transform or proper orthogonal decomposition (POD).

Principal component analysis is one of the simplest of the multivariate methods of analysis. The objective of the analysis is to take P variables i.e. X_1, X_2,…, X_P and find the combination of these to produce indices i.e. coded or standardized Z_1, Z_2,..., Z_P that are uncorrelated. The lack of correlation is a useful property because it means that the indices are measuring different dimension on the data.

These indices are ordered so that Z_1 displaced largest amount of variation and Z_2 displayed the second largest variation i.e. V (Z_1) V (Z_2), …,V(Z_P), these variation of Z_i are called the principal components.

1.1.1 Application of Principal Component Analysis

i. The most important use of principal components analysis is reduction of data. It provides the effective dimensionality of the data. If first few components account for most of the variation in the original data, then first few components' scores can be utilized in subsequent analysis in place of original variables.

ii. Plotting of data becomes difficult with more than three variables. It is often possible to account for most of the variability in the data by retaining the first two principal components, and it is possible to plot the values of first two components scores for each individual. Thus, principal component analysis enables plotting the data in two dimensions. Particularly detection of outliers or clustering of individuals will be easier through this technique. Often, use of principal component analysis reveals grouping of variables which would not be found by other means.

iii. Reduction in dimensionality can also help in analysis where the number of variables is more than the number of observations, for example, in discriminants analysis and regression analysis. In such cases, principal component analysis is helpful by reducing the dimensionality of data.

iv. Multiple regressions can be dangerous if independent variables are highly correlated. Principal component analysis is the most practical technique to solve the problem. Regression analysis can be carried out using principal components as repressors' in place of original variables. This is known as principal component regressions

1.2 Statement of the Problem

When there are too many data set to be analyzed in which some of the variables are correlated to one another that is, they are redundant, this make the entire conduct of the analysis difficult. In Principal Component Analysis there are different Methods to be used in reducing the dimensionality of a given data set but among this Method, there is no preference to the Methods. Therefore becomes necessary to compare two of the Methods with the view to finding the most suitable one in reducing the dimensionality of a given data set. Based on the above problem under consideration two Methods, "Eigenvalue one criterion and Proportion of Variance accounted for", are selected for the comparison. Number of PC's retained from both the two Methods and Model adequacy checking will be used to determine the most appropriate approach in dealing with any set of data of cholesterol level of Human body in PCA.

.

1.3 Aim and Objectives of the Study

The aim of this research work is to carry out a comparative study of the two methods of PCA—Eigenvalue-one criterion and proportion of variance accounted for.

The specific objectives include:

1. Simulation of data for the comparison and using data collected from ABUTH
2. Apply the two methods on the simulated data Set.
3. Compare the two Methods using model adequacy checking and by the number of PC's retained to see which one is more preferred and desired in the reduction of data for cholesterol level of Human body.

1.4 Significance of the Study

1. Variation in the data set can be adequately described by the few PC's.

2. Some degree of economy is achieved, for the initial dimension of the data set is reduced so therefore cost implication is minimized in the further analysis.

3. Variation in the P original X Variables is accounted for by a smaller number of Z Variable(s).

4. Also Variation in the say P simulated X Variables is accounted for by a number of Q variable(s).

1.5 Scope and Limitation

The work is limited to two Methods of data reduction techniques (Proportion of Variance Accounted for and Eigen value One-Criterion). Secondary data of twelve variables of forty observations for the decomposition of cholesterol level in human body collected from the Department of Chemical Pathology at ABUTH were used to conduct the analysis, and simulation was equally carryout to conduct same analysis to authenticate the findings.

1.6 Method of Data Collection

Data used for this research were collected from the Department of Chemical Pathology in Ahmadu Bello University Teaching Hospital, (ABUTH) Zaria. Twelve (12) variables for determining the cholesterol level of human body were considered, under each variable 40 observations were collected.

1.7 Method of Data Analysis

Matlab package was used for the computation of the variance- covariance matrices, correlation matrices and the corresponding PC's for both the original and the Simulated data set. The outputs of the Matlab package will be shown in chapter 4.

1.8 Definition of Term

i. Eigenvalues and Eigenvectors: are a special set of Scalars associated with a linear system of equations (i.e. a matrix equation) that are sometimes also known as characteristics roots, characteristic Values (Hoffman and Kunze, 1971), proper Values, or latent roots (Marcus and Minc, 1988). It could also be defined mathematically as; A scalar λ is called an *eigenvalue* of the $n \times n$ matrix A if there is a nontrivial solution X of $|A-\lambda|x=0$ such as X is called an

eigenvector corresponding to the eigenvalue λ

ii. Variance-Covariance Matrix: In probability theory and Statistics, a covariance matrix (also known as dispersion matrix or Variance-covariance matrix) is a matrix whose element in the i,j position is the Covariance between the i^{th} & j^{th} elements of a random variables (that is, of a Vector of random variables). Similarly, it can be defined mathematically as;

Let X be a $kx1$ random Vector. The Covariance matrix of X, or Variance-covariance matrix of X, denoted by $Var[X]$ is defined as;

$$Var[X] = E[x-E(x)][x-E(x)^{T}] \quad \text{.....(1)}$$

provided the above expected values exist and are well defined. On the other hand, it's also defined as a multivariate generalization of the definition of variance for a Scalar random variable Y:

$$Var[Y] = E\left[\left[y - E(y)\right]\left[y - E(y)^{T}\right]\right] \quad \text{.....(2)}$$

iii. Correlation Matrix: Correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values.

iv. $\boldsymbol{R^2}$ is a number that indicate how well data fit a statistical model— sometimes simply a line or curve. It's a statistic used in the context of statistical model whose main purpose is either the prediction of future outcomes of the testing of hypothesis, on the basis of other related information, it provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.

v. $\boldsymbol{R^2}$**-adjusted:** is a modified version of *R*-Squared that has been adjusted for the number of predictors in the model. The adjusted *R*-Squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model less than expected by chance. The adjusted *R*- Squared can be negative, but it's always lower than the *r*-Squared.

vi. Model: Is simply a Hypothetical description of a complex entity.

vii. Principal Component – a linear combination of the original variables (1^{st} Principal Component explains most of the variation in the data, 2^{nd} Pc explains most of the rest of the variance and so on)

2.0 CHAPTER TWO: LITERATURE REVIEW

2.1 Development of PCA (Historical Perspective of PCA)

According to Jolliffe (2002) it is generally accepted that PCA was first described by Pearson (1901). In his article "On lines and planes of closest fit to systems of points in space" he also discussed the graphical representation of data and lines that best represents the data. He concluded that "The best-fitting straight line to a system of points coincides in direction with the maximum axis of the correlation ellipsoid". He also stated that the analysis used in his paper can be applied to multiple variables.

However, PCA was not widely used until the development of computers. It is not really feasible to do PCA by hand when number of variables is greater than four, but it is exactly for larger amount of variables that PCA is really useful, so the full potential of PCA could not be used until after the spreads of computers Jolliffe (2002).

According to Jolliffe (2002) significant contributions to the development of PCA were made by Hotelling (1933). In 1960s as the interest in PC's rose, important contributors were Anderson (1963) with a theoretical discussion and a numerous new ideas concerning uses, interpretations and extensions of PCA, Gower (1966) with discussion about links between PCA and other statistical techniques and Jeffers (1967) with a practical application in two case studies.

2.2 Significance of Principal Component Analysis

PCA is a standard tool in modern data analysis – in diverse fields from neuroscience to computer graphics – because it is a simple, non-parametric method for extracting relevant information

from confusing data sets. With minimal effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the almost all of the hidden Information. The importance of PCA is manifested by its use in so many different fields of science and life. PCA is very much used in neuroscience, for example. Another field of use are pattern recognition and image compression, therefore PCA is suited for use in facial recognition software for example, as well as for recognition and storing of other biometric data..According to Jolliffe (2002) PCA is also used in Agricultural research, biology, chemistry, climatology, demography, ecology, food research genetics, meteorology, oceanography, psychology, quality control etc.

PCA has been used in economics and finance to study changes in stock markets, commodity markets, economic growth, exchange rates etc. PCA has mostly been used to compare different stock markets in search for diversification opportunities, especially in earlier studies like the ones by Makridakis (1974) and Phillipatos *et al*. (1983).

2.3 Advantages of Principal Component Analysis

PCA is a special case of Factor Analysis that is highly useful in the analysis of many times series data and the search for patterns of movement common to several series (true factor analysis makes different assumptions about the underlying structure and solve eigenvectors of a slightly different matrix). This approach is superior to many of the Bivariate statistical techniques used earlier in that, it explores the interrelationships among a set of variables caused by common "factor" mostly economic in nature. Philippatos *et al*. (1983).

PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. A primary benefit of PCA arises from quantifying the

importance of each dimension for describing the variability of a data set. PCA can also be used to compress data, by reducing the number of dimensions, without much loss of information.

When using principal component analysis to analyze a data set, it is usually possible to explain a large percentage of the total variance with only a few components. Principal components are selected so that each successive one explains a maximum of the retaining variance; the first component is selected to explain the maximum of the remaining variance etc. Therefore, the principal component solution is a particularly appropriate test for the existence of a strong market factor. Lessard (1973).

PCA is completely nonparametric: any data set can be plugged in and an answer comes out, requiring no parameters to tweak and no regard for how the data was recorded. From one perspective, the fact that PCA is non-parametric (or plug-and-play) can be considered a positive feature because the answer is unique and independent of the user.

2.4 Restraints of Principal Component Analysis

Restraints in PCA occur mainly due to main assumptions and the data at hand. PCA is not a statistical method from the viewpoint that there is no probability distribution specified for the observation. Therefore, it is important to keep in mind that PCA best serves to represent data in simpler, reduced form.

It is often difficult, if not impossible, to discover the true interpretation of PC's since the new variables are linear combinations of the original variables. In addition, for PCA to work, if the

data has multiple unit of measurement, one should use standardized data so that the mean is zero and the unbiased estimate of variance is unity which is carried out as below;

$$Z_i = \frac{x_i - \mu_x}{\sigma_x}$$

where $Z_i = i^{th}$ standardized variable.

This is because it is often the case that the scales of the original variables are not comparable and that (those) variable (variables) with high absolute variance will dominate the first principal component.

This is one major drawback to standardization, however, standardizing means that PCA results will come out with respect to standardized variables. This makes the interpretation and further application of PCA results even more difficult. Malava (2006).

The mission when using PCA is often to get rid of correlation and interdependence of variables. PCA succeeds in getting rid of second order dependences, but it has trouble with higher-order dependencies. This problem might be solved by using kernel PCA or independent component analysis. The fact that PCA is agnostic to the source of the data is also a weakness.

2.5 The Eigen Value One-Criterion

In principal components analysis, one of the most commonly used criteria for solving the number-of-components problem is the eigenvalue-one criterion, also known as the Kaiser criterion). With this approach, you retain and interpret any component with an eigenvalue greater than 1.00.

The rationale for this criterion is straightforward. Each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an eigenvalue greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained.

On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variable. The purpose of Principal Components Analysis is to reduce a number of observed variables into a relatively smaller number of components; this cannot be effectively achieved if you retain components that account for less variance than had been contributed by individual variables. For this reason, components with Eigenvalues less than 1.00 are viewed as trivial, and are not retained.

The eigenvalue-one criterion has a number of positive features that have contributed to it's popularity. Perhaps the most important reason for its widespread use is its simplicity: You do not make any subjective decisions, but merely retain components with Eigenvalues greater than one. On the positive side, it has been shown that this criterion very often results in retaining the correct number of components, particularly when number of variables were analyzed and the variable communalities are high. Stevens (1986), have investigated the accuracy of the eigenvalue-one criterion, and recommends its use when less than 30 variables are being analyzed and communalities are greater than 70, or when the analysis is based on over 250 observations and the mean communality is greater than or equal to 60.

There are a number of problems associated with the eigenvalue-one criterion, however, it can lead to retaining the wrong number of components under circumstances that are often encountered in research (e.g., when many variables are analyzed, when communalities are small). Also, the mindless application of this criterion can lead to retaining a certain number of components that are trivial, Kaiser (1960).

2.6 Proportion of Variance Accounted for

The second approach in solving the number of factors problem involves retaining a component if it accounts for specified proportion (or percentage) of variance in the data set. For example, you may select any component that accounts for at least 5% or 10% of the total variance. This proportion can be calculated with a simple formula.

In principal components analysis, the "total Eigenvalues of the covariance matrix" is equal to the total number of variables being analyzed (because each variable contributes one unit of variance to the analysis).When researchers use the "cumulative percent of variance accounted for" as the criterion for solving the number-of- components problem, they usually retain enough components variance accounted for at least 70 % (and sometimes 90%), (Kim & Mueller,1978). The proportion of variance criterion has a number of positive features. For example, in most cases, you would not want to retain a group of components that, combined and account for only a minority of the variance in the data set (say, 30%). Nonetheless, the critical values discussed earlier (10% for individual components and 70%-80% for the components) are obviously arbitrary. Because of those and related problems, this approach has sometimes been criticized for its subjectivity Muller (1978)

3.0 CHAPTER THREE: METHODOLOGY

3.1 Preamble

Principal components analysis is one of the most suitable methods for reducing data of a high dimension which is a subset of multivariate data reduction techniques in statistical world of today.

In this chapter, we discuss at length the pattern in which artificial variables were retained using Proportion – of – Variance accounted for and eigenvalue –one – criterion to reduce dimensionality and get the most appropriate PC's worth to be considered in order to avoid redundancy/superfluousness or duplication of components.

3.2 Procedures Involved In Carrying Out PCA

The procedure involves finding the Eigen Value of the sample covariance matrix most especially when the variables are standardized. The variances of the principal components are Eigen values of the covariance matrix; there are P of them, some of which may be zero (0).

Assuming that the Eigen values are ordered $X_1 X_2 X_3, \ldots, X_p = 0$ then X_i corresponds to the i^{th} Principal component.

The constants $a_{11}, a_{12}, a_{13} \ldots a_{1p}$ are the elements corresponding to Eigen Vectors scale so that;

$a_{112} + a_{122} + \ldots + a_{1p2} = 1$ must be satisfied during the computation.

According to Shlens (2009), the basic steps involved in carrying principal component analysis are;

1. Start by coding the variables $X_1, \ldots, X_P$ to have a zero mean and a unit variance.

2. Calculate the covariance matrix of the coded variances i.e. the correlation matrix.
3. Find the Eigenvalues $X_1,\ldots, X_P$ and the corresponding Eigen Vectors $a_{11}, a_{12},\ldots,a_{1p}$. The coefficient of the i^{th} PC's are then given as X_{ij} while the X_i is its variance.
4. Discard any component that only accounts for a small proportion of the variation in the data.

The main steps of the PCA algorithms are as shown below;

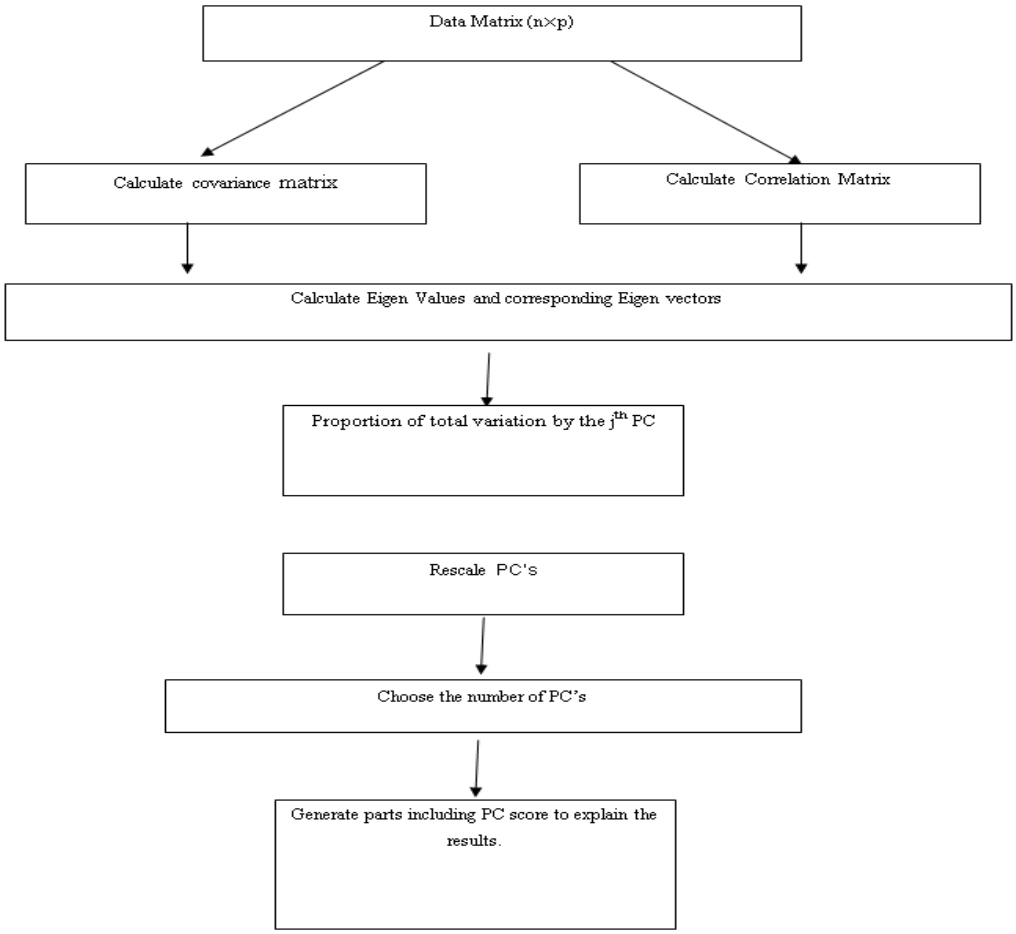


Figure 3.1: Stages in Principal Component Analysis.

The mathematical equations are presented below:

Let $X_1, X_2, X_3,\ldots, X_P$ be the variables under study, then first principal component may be defined as;

$$Z = a_{11}x_1 + a_{11}x_2 + ... + a_{1p}x_p \quad \text{.....(1)}$$

such that variance of Z_1 is as large as possible subject to the condition that;

$$a_{12} + a_{22} + ... + a_1 p_2 = 1 \quad \text{.....(2)}$$

This constant is introduced because if this is not done, then var (Z_1) can be increased simply by multiplying any $a_{ij's}$ by a constant factor. The second principal component is defined as;

$$Z_2 = a_{21}x_1 + a_{22}x_2 + \ \ldots \ + a_{2p}x_p \quad \text{.....} \ (3)$$

such that var(Z_2) is as large as possible next to var(Z_1) subject to the constraint that.

$$a_{12} + a_{22} + \ldots \ + a_2 p_2 = 1, \text{ and } Cov\left(Z_1, Z_2\right) = 0 \text{ and so on}$$

It's quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial P variables in subsequent analysis, thus, reducing the effective dimensionality of the problem. An analysis of principal component often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, PCA is more of a means to an end rather than an end in itself because this frequently serves as an intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It's a mathematical technique which doesn't require specifying the statistical model or assumption about distribution of original variable. It may also be mentioned that principal components are artificial variables and often it's not possible to assign physical meaning to them. Further, since PCA transforms original set of variables to new set of uncorrelated variables, it is worth stressing that if original variables are uncorrelated, then there is no point in carrying out PCA.

3.3 Computation of Principal Component using Cumulative Proportion Procedure

Principal component analysis involved at least three Methods in deciding which component is to be retained for further analysis. In this research, emphasis is given to two of the three well known as presumed, that is, Proportion of Variance Accounted for (Kim & Mueller, 1978) and eigenvalue one criterion (Kaiser criterion) before which computing PCA using the covariance method is most useful.

STEP 1: Organize the data Matrix (nxp)

i. Calculate the variance-covariance Matrix that is

the covariance matrix of the sample $S = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{m1} & \cdots & s_{mn} \end{pmatrix}$

ii. Find the P × P Empirical covariance matrix say Q from the outer product of matrix P, and

then, standardize to remove correlation

iii. Find the Eigen values denoted by Δ and the corresponding Eigen vectors denoted by V of the covariance or the correlation matrix.

iii. Compute the matrix V of the Eigen vectors which the diagonalizes the covariance matrix Q.

$$V^{-1}CV = \Delta \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .. (4)$$

iv. Matrix V, also of dimension P × P, contains P column vectors, each of length P, which

represent the P Eigen vectors of the covariance matrix C.

v. The Eigen values and Eigen vectors are ordered and paired. The Pth Eigen values corresponding to Pth Eigen vectors as;

$$\Delta = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ \lambda_n \end{pmatrix} \text{ correspond to } V = \begin{pmatrix} a_{11} \\ a_{12} \\ . \\ . \\ a_{1n} \end{pmatrix}$$

And rearrange the V's and theΔ's

vii. Sort the columns of the eigenvector matrix V and eigen values followed by the smallest.

$\lambda_1, \lambda_2, ..., \lambda_n$ Meaning any of the eigen values with a high value will be first considered to get the proportion.

viii. Make sure to maintain the correct pairings between the columns in each matrix.

STEP 2: Obtain The PC's Using Proportion Of Variance Accounted For Approaches.

The principal components for the set of data P will be as

$$\begin{aligned} Z_1 &= a_1 x_1 + a_2 x_2 +, ..., + a_n x_n \\ Z_2 &= a_2 x_1 + a_2 x_2 +, ..., + a_n x_n \\ &. \\ &. \\ &. \\ Z_n &= a_n x_1 + a_n x_2 +, ..., a_n x_n \end{aligned} \quad \text{………………………………………………………} (5)$$

The variance of Principal components will be Eigen values i.e.

$$V\ (Z_e)\ =\ V(Z_1) + (Z_2) + \ldots + V\ (Z_{in}) \quad \text{……………………………………………..} (6)$$

I.e. The total variation explained by the original variables.

The total variation explained by the PC's as the values of Δ corresponds to the total variations explained by the original variables.

i.e. $$\sum_{z_1}^{z_n} V(z_i) = \sum_{x_j}^{x_n} \lambda_i \quad \text{………………………………………………………} (7)$$

and vice-versa.

The proportion of total variation accounted for by the principal components, that is from the first to the final on the basis of the proportion we have selected which is 90% is computed by the formula shown below;

$$\text{Proportion} = \frac{Eigenvalue\ for\ the\ component\ of\ interest}{Total\ eigenvalues\ of\ the\ correlation\ matrix}$$

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 +, ..., +\lambda_n} \quad \text{..... (8)}$$

by considering the λ with the highest value, and then we continue to obtain the components accounting for a proportion of the total variance.

Therefore, the principal components Z_1 to Z_i could replace the total variables by sacrificing negligible information about the total variation in the system.

3.4 Variance--Covariance Matrix

The below information explains how to use matrix methods to generate variance-covariance matrix data.

3.4.1 Variance

Variance is a measure of the variability or spread in a set of data. Mathematically, it's the average squared deviation from mean score. We use the following formula to calculate a Sample variance

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) \quad \text{.....} \quad \text{..... (9)}$$

Where;

(n-1) is the number of scores in a set of scores and -1 is meant to clear off errors likely to creeps into the data set.

$\bar{x}$ the mean of the (n-1) scores

X_i is the i^th^ raw score in the set scores.

$S^2{}_i$ is the sample variance of all the scores in the set.

3.4.2 Covariance

Is a measure of the extent for which corresponding elements from two sets of ordered data move in the same direction, we use the following formula to compute covariance.

$$Cov(\text{x, y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / (n-1) \quad \text{.....} \quad (10)$$

Where;

(n-1) is the number of scores in a set of scores with -1 to clear errors.

$\bar{x}$ is the mean of scores of the first sample.

$\bar{y}$ is the mean of scores of the second sample.

x_i is the ith raw in the set score.

y_i is the ith deviation score in the second set of data.

$\text{cov}(x_{i,} y_i)$ is the covariance of correspond score in the two sets of data

3.4.3 Variance- Covariance

The variances appear along the diagonal and covariance appears in the off—diagonal elements, as shown below.

$$V = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 / \mathrm{N} & \sum_{i=1}^{n} x_1 x_2 / N \cdots\cdots & \sum_{i=1}^{n} x_1 x_c / N \\ \sum_{i=1}^{n} x_2 x_1 / \mathrm{N} & \sum_{i=1}^{n} x_2^2 / \mathrm{N}..... & \sum_{i=1}^{n} x_2 x_c / \mathrm{N} \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} x_c x_1 / N & \sum_{i=1}^{n} x_c x_2 / \mathrm{N} & \sum_{i=1}^{n} x_c^2 / \mathrm{N} \end{pmatrix}$$

3.5 Correlation Matrix Computation Illustration (Standardization)

$$D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{s_{11}} & \cdots & 0 \\ \vdots & \sqrt{s_{22}} & \vdots \\ \sqrt{s_{n1}} & 0 & \sqrt{s_{pp}} \end{pmatrix}, \text{ then } D^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{s_{11}}} & 0.... & 0 \\ \vdots & \frac{1}{\sqrt{s_{11}}}..... & 0 \\ \frac{1}{\sqrt{s_{n1}}} & 0..... & \frac{1}{\sqrt{s_{pp}}} \end{pmatrix}$$

Since,

$$\mathrm{S} = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \cdots & s_{pp} \end{pmatrix} \qquad R = \begin{pmatrix} \frac{s_{11}}{\sqrt{s_{11}}\sqrt{s_{11}}} & \cdots & \frac{s_{1n}}{\sqrt{s_{11}}\sqrt{s_{1n}}} \\ \vdots & \ddots & \vdots \\ \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{1p}}} & \cdots & \frac{s_{pp}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \end{pmatrix}$$

$$\therefore R = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

The above formula is for finding *R* i.e. the sample correlation matrix from the given covariance matrix.

3.6 *R*-(Squared) - Coefficient of Determination

It's a statistics used in a content of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypothesis on the basis of other related information.

There are several definitions of R- (Squared) that are only sometimes equivalent. One class of such cases includes that of Simple Linear Regression where Coefficient of Determination (r^2) is used instead of R^2.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad \text{...} \quad (11)$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{, total sum of squares}$$

$$SS_{residual} = \sum_{i=1}^{n} (y_i - f_i)^2 \text{, Residuals sum of squares}$$

Where f_i is the predicted value (or modeled) value sometimes written as $\hat{y}_i$

In some cases the total sum of squares equals the sum of two others sums of squares defined above

$$SS_{residual} + SS_{regression} = SS_{total} \quad \text{..} \quad (12)$$

where SS_{reg} = regression sum of squares also called the residual sum of squares.

R^2 is a statistics that will give some information about the goodness of fit of a model. In regression, the R^2 (coefficient of determination) is a statistical measure of how well the regression line approximates the real data points. And R^2indicates that the regression line perfectly fits the data.

R^2 (Adjusted).

The use of an adjusted R^2 (often written as $\bar{R}^2$ as pronounced (R bar squared) is an attempt to take account of the phenomenon of the R^2 automatically and spuriously increasing when extra exploratory variables are added to the model. The adjusted R^2 is defined as

$$R_{adj}{}^2 = R^2 - (1 - R^2)\frac{P}{n - p - 1} \quad \text{…..…… (13)}$$

where P is the total number of repressors in the linear model (not counting the constant term), and n is the sample size. Adjusted R^2 can also be written as

$$\bar{R}^2 = 1 - \frac{SS_{res} / df}{SS_{tot} / df} \quad \text{…..…….…….. (14)}$$

Where is the degree of freedom df n-1 of the estimate of the population variance of the dependent variable, and is the degree of freedom df n-p-1 of the estimate of the underlying population error variance.

4.0 CHAPTER FOUR: DATA ANALYSIS

4.1 Data Presentation: Analysis and Discussions

The computation of the Principal component vividly explained in Chapter three and the proportion of variance accounted for as well as the eigenvalue one criterion in determining the number of meaningful component to retain are the two approaches we employed for the research, and later model adequacy checking for verification.

4.2 Illustrations

The data set comprises (12) different variables X_1, $X_{2,}\ldots,X_{12}$ of forty (40) female with cholesterol level in mg . X_1 represents Age in years, X_2 =Height in m, X_3 = Weight in kg, X_4 = Waist circumference in cm of a plump female , X_5 = pulse rate in beats per minute, X_6 = Systematic blood pressure in cm Hg, X_7 = Diastolic blood pressure in cm Hg, X_8 =BMI in kgm^{-2}, X_9 =Leg length in cm, X_{10}= Elbow breadth in cm, X_{11}= Wrist breadth in cm, X_{12}=Arm circumference in cm. Under each variable, 40 observations were considered. We need to note here that BMI is measured as Body weight in kilograms by Height in Meters Squared, so the weight unit has to be taken in kg and the height in meters.

Table 4.1: Data Set of the 12 Different Variables of 40 Observations.

S/No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
1.	17	1.633	52.07	67.2	76	10.4	6.1	19.5260	41.6	6	4.6	23.6
2.	32	1.687	67.72	82.5	72	9.9	6.4	23.7950	42.8	6.7	5.5	26.3
3.	25	1.582	48.89	66.7	88	10.2	6.5	19.5340	39	5.7	4.6	26.3
4.	55	1.582	72.89	93	60	11.4	7.6	29.1245	40.2	6.2	5	32.6
5.	27	1.514	57.65	82.6	72	9.4	5.8	25.1505	36.2	5.5	4.8	29.2

Cont'd of Table 4.1												
6.	29	1.615	55.83	75.4	68	10.1	6.6	21.4055	43.2	6	4.9	26.4
7.	25	1.519	50.66	73.6	80	10.8	6.1	22.0126	38.7	5.7	5.1	27.9
8.	12	1.608	49.89	81.4	64	10.4	4.1	19.2945	41	6.8	5.5	33
9.	41	1.725	99.24	99.4	68	12.3	7.2	33.3512	43.8	7.8	5.8	38.6
10.	32	1.559	49.98	67.7	68	9.3	6.1	20.5636	37.3	6.3	5	26.5
11.	31	1.694	85.41	100.7	80	8.9	6.2	29.7637	42.3	6.6	5.2	34.4
12.	19	1.645	47.8	72.9	76	11.2	6.2	17.6644	39.1	5.7	4.8	23.7
13.	19	1.603	61.73	85	68	10.7	4.8	24.0231	40.3	6.6	5.1	28.4
14.	23	1.608	82.73	85.7	72	11.6	6.2	31.9803	48.6	7.2	5.6	34
15.	40	1.694	70.64	126	96	18.1	10.2	24.6398	33.2	7	5.4	35.2
16.	23	1.697	49.35	74.5	72	9.8	6.1	17.1366	43.4	6.2	5.2	24.7
17.	27	1.643	53.97	74.5	68	10	5.3	19.9933	41.5	6.3	5.3	27
18.	45	1.653	73.43	94	72	12.7	7.4	26.8738	40	6.8	5	35
19.	41	1.572	79.97	92.8	64	10.7	6.7	32.3607	38.2	6.8	4.7	33.1
20.	56	1.633	82.19	105.5	80	11.6	7.1	30.8208	38.2	6.9	5.4	39.6
21.	22	1.610	56.38	75.5	64	9.7	6.4	21.7507	38.2	5.9	5	27
22.	57	1.542	48.39	70	76	10.6	5.9	20.3512	38.1	6.1	5	23.6
23.	24	1.610	68.02	98	76	11	7	26.2412	38	7	5.1	34.3
24.	37	1.590	73.98	104.7	76	10.5	6.9	29.2630	36	6.7	5.1	34.4
25.	59	1.539	42.77	67.8	80	11.8	8.2	18.0578	32.1	5.4	4.2	23.3
26.	40	1.613	72.43	99.3	104	13.3	8.3	27.8384	31.1	6.4	5.2	35.6
27.	45	1.488	73.84	91.1	88	11.3	7.5	33.3498	39.4	7.1	5.3	31.8

Cont'd of table 4.1												
28.0	52	1.529	58.96	74.5	60	11.3	6.6	25.2202	40.2	5.9	5.1	27
29.0	31	1.717	81.6	95.5	76	10.7	6.7	27.6797	39.2	6.2	5.0	32.8
30.	32	1.628	67.04	79.5	72	9.5	5.9	25.2942	39	6.3	4.9	31
31.	23	1.592	51.21	69.1	72	10.8	7.2	20.2051	36.6	5.9	4.7	27
32.	23	1.557	88.72	105.5	88	11.4	7.9	36.5976	27	7.5	5.5	41.2
33.	47	1.478	56.33	78.8	80	10.4	7.3	25.7862	38.5	5.6	4.7	25.5
34.	36	1.605	61.23	85.7	60	12.5	7.3	23.7694	39.9	6.4	5.2	30.9
35.	34	1.651	64.13	92.8	72	12.4	8.5	23.5270	37.5	6.1	4.8	27.9
36.	37	1.536	56.2	72.7	88	9.2	4.6	23.8216	39.7	5.8	5	26.5
37.	18	1.569	61.46	75.9	88	11.9	8.1	24.9664	39	6.3	4.9	27.8
38.	29	1.727	45.67	68.6	124	9.3	6.4	15.3126	41.6	6	5.3	23
39.	48	1.701	43	68.7	64	10.6	6.4	14.8614	33.8	5.6	4.6	26.4
40.	16	1.448	52.07	67.2	76	10.4	6.1	24.8342	41.6	6	4.6	23.6

4.3 Computed variance-covariance matrix from the original data set.

The estimated covariance matrix of the data above is given by; .

Table 4.2: Variance-Covariance Matrix from the Original Data Set.

188.6479	-32.8232	52.8015	130.6302	84.6634	-29.8374	4.2403	-2.8534	-16.9561	-1.9011	-1.1921	-5.3931
-32.8232	32.3148	-24.1091	-90.4001	-84.3930	31.1286	-0.0540	8.2932	4.9584	0.9074	0.4891	7.9216
52.8015	-24.1091	210.8905	224.4949	62.3677	-20.2717	3.0982	29.3787	-5.9800	3.7710	1.3485	37.6925
130.6302	-90.4001	224.4949	426.7218	256.2729	-86.7083	3.4864	11.3863	-25.2522	0.8159	-0.0867	20.3816
84.6634	-84.3930	62.3677	256.2729	384.8589	-85.8648	2.1694	-24.1153	-22.7673	-3.1848	-1.2720	-25.0855
-29.8374	31.1286	-20.2717	-86.7083	-85.8648	31.9184	0.6811	10.2901	4.1222	1.1702	0.5657	9.8987
4.2403	-0.0540	3.0982	3.4864	2.1694	0.6811	0.9367	1.7356	-1.4635	0.0616	-0.0430	1.1899
-2.8534	8.2932	29.3787	11.3863	-24.1153	10.2901	1.7356	27.1685	-3.6695	1.4993	0.4877	15.9903
-16.9561	4.9584	-5.9800	-25.2522	-22.7673	4.1222	-1.4635	-3.6695	13.3933	0.3450	0.3765	-2.0108
-1.9011	0.9074	3.7710	0.8159	-3.1848	1.1702	0.0616	1.4993	0.3450	0.3087	0.1376	2.1063
-1.1921	0.4891	1.3485	-0.0867	-1.2720	0.5657	-0.0430	0.4877	0.3765	0.1376	0.1038	0.9392
-5.3931	7.9216	37.6925	20.3816	-25.0855	9.8987	1.1899	15.9903	-2.0108	2.1063	0.9392	22.9285

where S_{ij} is the sample variance or covariance between variables x_i and x_j. The values of variance covariance matrix were obtained by dividing the original values by 1000

Table 4.3: Correlation Matrix from the Variance-Covariance Matrix

1.0000	-0.4204	0.2647	0.4604	0.3142	-0.3845	0.3190	-0.0399	-0.3373	-0.2491	-0.2694	-0.0820
-0.4204	1.0000	-0.2920	-0.7698	-0.7568	0.9693	-0.0098	0.2799	0.2383	0.2873	0.2671	0.2910
0.2647	-0.2920	1.0000	0.7484	0.2189	-0.2471	0.2204	0.3881	-0.1125	0.4674	0.2882	0.5420
0.4604	-0.7698	0.7484	1.0000	0.6324	-0.7430	0.1744	0.1057	-0.3340	0.0711	-0.0130	0.2061
0.3142	-0.7568	0.2189	0.6324	1.0000	-0.7747	0.1143	-0.2358	-0.3171	-0.2922	-0.2013	-0.2670
-0.3845	0.9693	-0.2471	-0.7430	-0.7747	1.0000	0.1246	0.3494	0.1994	0.3728	0.3108	0.3659
0.3190	-0.0098	0.2204	0.1744	0.1143	0.1246	1.0000	0.3440	-0.4132	0.1145	-0.1379	0.2568
-0.0399	0.2799	0.3881	0.1057	-0.2358	0.3494	0.3440	1.0000	-0.1924	0.5177	0.2904	0.6407
-0.3373	0.2383	-0.1125	-0.3340	-0.3171	0.1994	-0.4132	-0.1924	1.0000	0.1697	0.3193	-0.1147
-0.2491	0.2873	0.4674	0.0711	-0.2922	0.3728	0.1145	0.5177	0.1697	1.0000	0.7689	0.7918
-0.2694	0.2671	0.2882	-0.0130	-0.2013	0.3108	-0.1379	0.2904	0.3193	0.7689	1.0000	0.6088
-0.0820	0.2910	0.5420	0.2061	-0.2670	0.3659	0.2568	0.6407	-0.1147	0.7918	0.6088	1.0000

Above is the correlation matrix from variance-covariance matrix.

The variances of the Principal Components will be the same as the Eigen values. Below is for the entire (12) variables under consideration.

And the total variation explained by the original variable is

$V(X_1) + V(X_2) + V(X_3) + V(X_4) + V(X_5) + V(X_6) + V(X_7) + V(X_8) + V(X_9) + V(X_{10}) + (X_{11}) + V(X_{12})$

$=1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000=12.$

The total variation explained by Principal Components is

$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 + \lambda_7 + \lambda_8 + \lambda_9 + \lambda_{10} + \lambda_{11} + \lambda_{12}$

$4.2966+3.2835+1.6816+.7074+0.5650+0.4999+0.3615+0.0145+0.0440+0.227+0.01379+0.1523 = 12.$

It can be seen clearly that the total variation explained by the Principal components i.e. the number of Eigenvalues is the same as the number of variables from the original data sets. (for the corresponding Eigen vectors, see Appendix IV).

The computation of the Proportion of Variance Accounted is as shown below;

$$\frac{\lambda_1}{\lambda+,...,+\lambda_{12}} = 0.3580 \approx 36\%$$

$$\frac{\lambda_1 + \lambda_2}{\lambda_1+,...,+\lambda_2} = 0.6316 \approx 63\%$$

$$\frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 +, ..., +\lambda_{12}} = 0.7718 \approx 79\%$$

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{\lambda_1 +, ..., +\lambda_{12}} = 0.8307 \approx 85\%$$

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}{\lambda_1 +, ..., +\lambda_{12}} = 0.8778 \approx 89\%$$

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6}{\lambda_1 +, ..., +\lambda_{12}} = 0.8991 \approx 90\%$$

From the above illustration, it shows that the first six PC's that is $Z_{1,}Z_2, Z_3$,Z_4,Z_5,andZ_6 could replace 12 variables by sacrificing negligible information about the total variation in the system. The scores of the principal components can be obtained by substituting the values of X's in equation of Z_i's denoted by Z_1, Z_2, and Z_3 as for the PC's retained by eigenvalue one-criterion.

$Z_1 =$ (-0.2687)X_1 +(0.4338)X_2 +(-0.0756)X_3 + (-0.3246)X_4 + (0.3900)X_5 + (0.4434)X_6 + (-0.0313)X_7 + (0.1959)X_8 + (0.1830)X_9 + (0.2832)X_{10} + (0.2639)X_{11} + (0.2391)X_{12}

Z_2= (0.1374)X_1 + (-0.1135)X_2 + (0.4705)X_3 + (0.3741)X_4 + (0.1020)X_5 + (-0.0609)X_6 +(0.2283)X_7 + (0.3481)X_8 +(-0.1580)X_9 + (0.3723)X_{10} + (0.2690)X_{11} + (0.4340) X_{12}

Z_3 = (0.2648)X_1 + (0.1933)X_2 + (-0.1315)X_3 + (-0.1313)X_4 + (-0.1035)X_5 + (0.2412)X_6 + (0.5458)X_7 + (0.2453)X_8 + (-0.4923)X_9+(-0.1994)X_{10}+(-0.3854) X_{11} + (0.0196)X_{12}.....(1)

The whole data set with twelve (12) variables can be represented to a new data set of six principal components denoted by Z_1, Z_2, Z_3, Z_4 , Z_5, and Z_6 for proportion of variance accounted for, and it can also be represented by three principal components Z_1, Z_2 and Z_3 as for the case of Eigenvalue one criterion. Meaning the first six (6) PC's account for most of the variation in the data, as for the proportion of variance accounted for method, while for eigenvalue one criterion the first three PC's explains most of the variation of the data set. Number of PC's retained and the Model Adequacy Checking was also employed to play a pivotal role in the verification exercise. To see which among the two methods is more considerable in dealing with a data of cholesterol level of Human body.

Table 4.4: The three retained PC's computed by matlab package for eigenvalue one criterion method.

PC1	PC2	PC3
-34.8338	74.199	-31.0793
-41.5238	91.995	-30.6841
-41.2970	76.422	-28.1642
-44.4394	105.171	-21.4108
-40.5032	88.665	-27.0311
-36.6310	81.698	-28.8352
-39.3545	80.635	-27.9666
-30.6477	80.747	-33.9194
41.7419	46.257	-6.5639
-50.9966	113.309	-31.6376
46.3881	42.803	-2.5052
-36.5358	76.229	-24.8466

Cont'd of Table 4.4

-50.2718	111.379	-36.2460
-34.5678	79.284	-26.2166
-36.3855	88.973	-32.2965
-38.7874	99.198	-39.6515
-35.9061	80.171	-29.4420
-45.9182	106.030	-25.8114
-42.5164	108.555	-25.3148
-55.1979	120.578	-25.0515
-34.1435	81.763	-28.0277
-46.5004	79.946	-18.6982
-44.0602	102.412	-31.4996
-50.2404	110.819	-28.1123
-49.0281	77.165	-13.1773
-59.8350	112.197	-25.7517
-51.3849	108.046	-25.8098
-38.7735	87.193	-19.5701
-46.3428	108.102	-31.3881
-40.2353	93.327	-27.9400
-35.2834	77.644	-25.8480
-50.4308	124.858	-28.2995
-47.6193	88.736	-21.8677
-36.9779	91.777	-25.1602
-45.1006	95.799	-26.3287
-46.2752	84.998	-27.4861
-40.9829	87.923	-30.4716

Cont'd of Table 4.4

-57.9879	76.837	-33.6106
-39.9601	74.253	-17.5773
-33.6055	75.930	-30.077

4.4 Results Obtained by the Real Life Data Set

The computation using the proportion of variance accounted for shows that, 6 PC's are to be retained for further analysis which yields or accounts for most of the variance of the original data set. A model was fitted using the six retained PC's and coefficient of determination was used to measure the adequacy of the fitted model; it was found that R^2 is 10.9% and R^2-(adjusted) is 3.7%.Similarly, for eigenvalue one criterion it was found that 3PC's are to be retained and R^2 was 6.5% and R^2- (adjusted) is 2.3% ,the adequacy of the model as well as the number of PC's retained conclude comparatively that Eigenvalue one criterion methods of data reduction techniques in principal component analysis is more suitable and preferable in the reduction of data of cholesterol level in Human body.

4.5 Analysis Using Simulated Data

The simulated data of the same dimension with the original data set, was computed from the original data set by the use of SPSS package.

The following data are simulated using SPSS package, the distribution used was normal, and the procedures employed is as follows: One variable was typed and the mean and the standard deviation was obtained, click on transform in the menu bar, select compute variable, a dialogue box appeared from there, and then we type-in target variable name. And in the function group, select random numbers, the function and special variables appeared, we double click R.V.

normal. In numeric expression, we change the two question marks to mean and standard deviation so far obtained from the original variable respectively. We then click on OK, so also is done for the second variable to the third, until the simulation of the final variable is performed.

Table 4.5: The Simulated Data of 12 Variables of 40 Observations.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
19.43	2.65	55.99	69.18	92.56	4.6	5.86	19.70	41.26	6.47	4.86	31.01
31.42	5.11	49.95	83.39	63.31	14.25	7.31	24.25	35.83	6.47	4.72	31.32
28.15	2.99	49.44	52.95	49.52	6.23	5.76	23.01	41.65	6.23	4.85	38.92
26.98	5.15	32.31	68.26	87.16	15.34	6.92	22.56	40.63	6.91	5.21	34.57
6.44	3.56	55.09	87.73	90.97	12.92	6.55	14.77	40.41	6.28	4.61	32.85
7.66	1.03	33.51	74.36	43.6	18.76	6.84	15.25	41.76	7.23	5.19	31.15
27.57	7.53	37	84.5	85.69	11.51	6.02	22.79	38.11	6.54	5.01	30.27
54.37	3.21	67.44	89.9	70.25	17.7	7.73	32.96	37.06	6.02	4.81	31.53
29.37	3.17	72.31	24.87	65.46	14.67	7.39	23.47	40.4	6.59	5.16	19.67
27.58	4.59	53.82	71.41	78.99	8.24	5.97	22.79	34.65	5.69	4.61	29.93
27.26	3.94	67.56	83.8	100.07	16.02	6.91	22.67	36	5.86	4.56	27.14
16	10.76	86.46	81.43	44.74	16.7	6.93	18.40	38.94	6.28	5.14	30.46
51.54	7.35	63.12	96.89	99.83	11.76	6.04	31.88	41	7.28	5.06	21.59
19.96	8.95	72.75	46.63	95.5	12.25	6.6	19.90	39.66	5.62	5.07	34.25
25.79	5.16	42.64	75.55	112.44	10.08	8.14	22.11	38.8	6.65	4.57	25.36
7.41	2.68	46.96	84.81	83.33	20.16	7.49	15.14	35.69	6.56	5.29	31.56
11	0.28	47.01	87.28	62.2	10.02	6.25	16.50	38.9	6.65	4.92	29.5

Cont'd of Table 4.5											
45.85	16.18	58.72	107.79	91.36	14.76	6.93	27.19	42.2	5.9	4.74	31.63
42.03	0.99	65.79	55.02	118.61	9.22	6.8	28.27	37.38	6.16	5.51	41.77
39.95	0.08	43.65	90.49	35.62	10.44	8.38	27.49	38.39	5.48	5.11	26.83
19.84	8.35	51.82	79.72	48.76	10.19	5.07	24.52	44.47	6.25	31.9	31.9
52.76	3.5	43.4	110.3	56.63	20.88	5.18	19.86	40.47	6.52	5.62	34.33
28.27	12.85	68.29	76.04	71.49	11.88	7.13	32.35	42.26	7.21	4.68	29.88
31.73	5.57	65.47	69.22	74.78	8.57	6.82	23.06	39.48	6.71	5.22	37.22
50.66	0.94	47.38	47.94	66.44	11.7	7.63	24.37	40.36	6.22	4.79	20.93
28.99	2.21	77.41	85.82	45.84	24.06	6.23	31.55	30	6.26	4.97	29.27
41.48	8.73	53.88	87.34	96.55	17.32	6.94	22.78	38.05	5.93	4.47	32.97
18.46	9.25	83.09	88.95	73.56	18.59	7.38	19.33	40.47	5.58	4.87	30.28
31.89	12.56	71.94	90.63	114.91	17.2	6.24	24.43	41.21	6.97	5.16	33.25
46.46	0.32	62.32	63.09	63.63	14.73	7.28	29.96	40.51	6.97	5.19	29.46
12.91	5.47	74.26	74.36	49.85	13.08	6.15	17.23	41.36	6.59	5.18	29.72
28.08	2.38	84.12	54.81	56.97	18.34	6.39	22.28	41.63	6.21	5.15	40.05
21.22	5.15	58.57	52.6	91.48	8.93	5.79	20.38	45.1	7.35	4.83	30.6
27.61	7.78	38.27	42.31	102.67	12.79	8.27	22.80	38.67	6.49	5.17	26.16
16.29	0.13	53.56	63.09	73.6	15.32	8.89	18.51	38.07	7.75	5.35	24.69

Cont'd of Table 4.5											
47.69	2.3	73.49	68.4	57.92	21.95	6.31	30.42	39.31	6.76	5.17	35.74
50.54	8.47	54.97	81.88	85.19	17.74	6.3	31.50	39.05	6.13	5.2	33
15.15	7.17	39.59	93.94	63.7	14.23	6.3	18.07	41.4	5.62	4.9	28.68

4.6 Computed variance-covariance Matrix of the Simulated data set.

The estimated variance-covariance matrix of the simulated data set above is given below;

$$\begin{pmatrix}
180.7853 & 3.0738 & 15.6073 & 30.6824 & 31.5425 & 8.6394 & 0.3107 & 68.5982 & -2.7860 & -0.7268 & 0.2244 & 0.4176 \\
3.0738 & 35.8460 & -10.1851 & -1.1435 & 31.0703 & -5.4516 & -0.5411 & 1.1668 & 1.6567 & 0.0256 & -0.4690 & -2.1625 \\
15.6073 & -10.1851 & 220.4470 & -21.8070 & -12.4115 & 19.8996 & -1.1369 & 5.9315 & -3.7579 & -1.2799 & -0.1950 & 6.4327 \\
30.6824 & -1.1435 & -21.8070 & 409.7723 & 4.5532 & 23.7676 & -0.8268 & 11.6556 & -0.5247 & -1.2144 & 0.0825 & 14.1605 \\
31.5425 & 31.0703 & -12.4115 & 4.5532 & 452.0573 & -19.5423 & 2.3194 & 11.9459 & 0.1362 & 0.7404 & -1.1515 & 3.5507 \\
8.6394 & -5.4516 & 19.8996 & 23.7676 & -19.5423 & 19.3489 & 0.3015 & 3.2780 & -4.1372 & -0.0767 & 0.2050 & 0.1025 \\
0.3107 & -0.5411 & -1.1369 & -0.8268 & 2.3194 & 0.3015 & 0.7550 & 0.1180 & -0.4568 & 0.0353 & -0.0061 & -1.2902 \\
68.5982 & 1.1668 & 5.9315 & 11.6556 & 11.9459 & 3.2780 & 0.1180 & 26.0293 & -1.0568 & -0.2755 & 0.0852 & 0.1589 \\
-2.7860 & 1.6567 & -3.7579 & -0.5247 & 0.1362 & -4.1372 & -0.4568 & -1.0568 & 7.7101 & 0.4356 & 0.1666 & 1.8069 \\
-0.7268 & 0.0256 & -1.2799 & -1.2144 & 0.7404 & -0.0767 & 0.0353 & -0.2755 & 0.4356 & 0.2843 & 0.0481 & -0.2728 \\
0.2244 & -0.4690 & -0.1950 & 0.0825 & -1.1515 & 0.2050 & -0.0061 & 0.0852 & 0.1666 & 0.0481 & 0.0812 & 0.4064 \\
0.4176 & -2.1625 & 6.4327 & 14.1605 & 3.5507 & 0.1025 & -1.2902 & 0.1589 & 1.8069 & -0.2728 & 0.4064 & 23.1072
\end{pmatrix}$$

where S_{ij} is the sample variance covariance between variable x_i and x_j from the simulated data.

The values were obtained by dividing each and every element in the matrix notation by 1000.

Table 4.7.Correlation Matrix Computed From The Variance-Covariance Matrix Of The Simulated Data:

1.0000	0.0382	0.0782	0.1127	0.1103	0.1461	0.0266	1.0000	-0.0746	-0.1014	0.0586	0.0065
0.0382	1.0000	-0.1146	-0.0094	0.2441	-0.2070	-0.1040	0.0382	0.0997	0.0080	-0.2749	-0.0751
0.0782	-0.1146	1.0000	-0.0726	-0.0393	0.3047	-0.0881	0.0783	-0.0912	-0.1617	-0.0461	0.0901
0.1127	-0.0094	-0.0726	1.0000	0.0106	0.2669	-0.0470	0.1129	-0.0093	-0.1125	0.0143	0.1455
0.1103	0.2441	-0.0393	0.0106	1.0000	-0.2090	0.1255	0.1101	0.0023	0.0653	-0.1901	0.0347
0.1461	-0.2070	0.3047	0.2669	-0.2090	1.0000	0.0789	0.1461	-0.3387	-0.0327	0.1636	0.0048
0.0266	-0.1040	-0.0881	-0.0470	0.1255	0.0789	1.0000	0.0266	-0.1893	0.0762	-0.0247	-0.3089
1.0000	0.0382	0.0783	0.1129	0.1101	0.1461	0.0266	1.0000	-0.0746	-0.1013	0.0586	0.0065
-0.0746	0.0997	-0.0912	-0.0093	0.0023	-0.3387	-0.1893	-0.0746	1.0000	0.2942	0.2106	0.1354
-0.1014	0.0080	-0.1617	-0.1125	0.0653	-0.0327	0.0762	-0.1013	0.2942	1.0000	0.3166	-0.1064
0.0586	-0.2749	-0.0461	0.0143	-0.1901	0.1636	-0.0247	0.0586	0.2106	0.3166	1.0000	0.2967
0.0065	-0.0751	0.0901	0.1455	0.0347	0.0048	-0.3089	0.0065	0.1354	-0.1064	0.2967	1.0000

Above is the correlation matrix of the simulated data set;

Here equally the variances of the Principal Components will be the same as the Eigen values. Below is for the entire twelve variables of the simulated data under consideration. And the total variation explained by the Original variable is:

$V(X_1) + V(X_2) + V(X_3) + V(X_4) + V(X_5) + V(X_6) + V(X_7) + V(X_8) + V(X_9) + V(X_{10}) + V(X_{11}) + V(X_{12})$

$=1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000+1.0000$

$1.0000+1.0000+1.0000=12.$

The total variation explained by the Principal Components is

$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 + \lambda_7 + \lambda_8 + \lambda_9 + \lambda_{10} + \lambda_{11} + \lambda_{12}$

$2.2094+0.0000+1.7696+1.6626+1.4086+0.3966+0.3564+1.0797+0.5850+0.7016+0.9460+0.8845 = 12$

Here, we are to retain all Eigenvalues that are greater than one which were

$\lambda_1, \lambda_3, \lambda_4, \lambda_5,$ and λ_8

(For the corresponding eigenvectors, see Appendix IX).

The total variation explained by the original variables is the sum of the diagonal of the correlation matrix of the simulated data.

The computations of the proportion accounted for approach is as shown below;

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.1843 \approx 18\%$$

$$\frac{\lambda_1+\lambda_3}{\lambda_1+\lambda_2+\lambda_3+...+\lambda_{12}} = 0.3315 \approx 33\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.4701 \approx 47\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.5875 \approx 59\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5+\lambda_6}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.6774 \approx 68\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5+\lambda_8+\lambda_{11}}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.7160 \approx 72\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5+\lambda_8+\lambda_{11}+\lambda_{12}}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.7898 \approx 79\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5+\lambda_8+\lambda_{11}+\lambda_{12}+\lambda_{10}}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.8480 \approx 85\%$$

$$\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5+\lambda_8+\lambda_{11}+\lambda_{12}+\lambda_{10}+\lambda_9}{\lambda_1+\lambda_2+\lambda_3+,...,+\lambda_{12}} = 0.894 \approx 90\%$$

It also shows from the above illustration that, the first 9PC's represented by $Q_1, Q_2 \ldots Q_9$, could represent 12 variables by sacrificing negligible information about the total variation in the system.

The linear combination for finding the five PC's retained by eigenvalue one criterion are as shown below;

$Q_1=(0.5987)X_1+(-0.0624)X_2+(0.1945)X_3+(0.1945)X_4+(0.0283)X_5+(0.1395)X_6+(0.0485)X_7+$
$(0.5987)X_8+(-0.2347)X_9+(-0.2040)X_{10}+(0.0190)X_{11}+(0.0342)X_{12}$

$Q_2=(0.7071)X_1+(0.0001)X_2+(0.0002)X_3+(0.0002)X_4+(-0.0002)X_5+(0.0002)X_6+$
$(0.0001)X_7+(-0.7071)X_8+(-0.0001)X_9+(0.0001)X_{10}+(0.0000)X_{11}+(0.0000)X_{12}$

$Q_3=(-0.1566)X_1+(-0.4583)X_2+(0.1982)X_3+(0.1310)X_4+(-0.4233)X_5+(0.3720)X_6+$
$(-0.1551)X_7+(-0.1565)X_8+(0.0372)X_9+(0.0779)X_{10}+(0.4840)X_{11}+ (0.3180)X_{12}$

$Q_4=(-0.2669)X_1+(-0.1498)X_2+(0.2094)X_3+(-0.0290)X_4+(-0.1457)X_5+(0.2884)X_6+$
$(0.2566)X_7+(-0.2669)X_8+(-0.5353)X_9+(-0.3262)X_{10}+(-0.3618)X_{11}+(-0.103)X_{12}$

$Q_5=(0.0959)X_1+(-0.2267)X_2+(-0.2320)X_3+(-0.1834)X_4+(-0.0299)X_5+(0.0983)X_6+(0.5647)X_7+$
$(0.0959)X_8+(-0.0204)X_9+(-0.4909)X_{10}+(0.2782)X_{11}+(-.04342)X_{12}$ .....(2)

The above iteration is equally carried out from Q_1 to obtain the entire retained PC's of the simulated data set as shown below:

PC1	PC2	PC3	PC4	PC5
42.1340	-0.1906	-11.3085	-47.3592	-31.2653
57.0808	5.0768	1.2149	-43.5414	-28.8631
44.4111	3.6402	4.5756	-47.3867	-27.8679
46.9416	3.1217	-11.4690	-52.3256	-24.5726

Cont'd of retained Pc's

37.7243	-5.8881	-2.5889	-40.5952	-36.0748
32.0715	-5.3474	14.3241	-36.8496	-24.3855
50.5622	3.3806	-11.8807	-50.6355	-28.2617
81.9548	15.1480	-0.1507	-48.7509	-30.5918
47.1860	4.1721	-4.7662	-35.3151	-17.9792
51.2074	3.3901	-7.7852	-43.8952	-29.5605
59.2389	3.2466	-6.4832	-40.4635	-31.2080
51.5847	-1.6826	27.6612	-25.9578	-34.5031
77.2140	13.9055	-18.6409	-55.4185	-29.6753
44.2792	0.0407	-11.4503	-42.3127	-33.0318
48.6353	2.5967	-24.3887	-50.6716	-25.3191
39.9285	-5.4652	3.6143	-35.6038	-29.6547
38.4509	-3.8821	6.4735	-38.5681	-30.6107
74.3267	11.4123	-13.3647	-56.7479	-37.5171
63.7889	9.7239	-20.3210	-56.2872	-31.4851
62.1035	8.8233	10.2769	-44.0399	-23.8637
66.5017	5.3887	-1.6423	-37.6026	-37.0272
43.2001	-0.0045	7.9493	-43.8821	-32.4600
80.0518	14.4419	9.9730	-53.6832	-28.9806
53.8927	3.6915	-3.4779	-44.6321	-33.6688
55.3645	5.2099	-2.4638	-47.5422	-33.4832
63.87591	3.5128	-10.7654	-49.4817	-13.6962
65.0044	4.0160	19.2778	-27.7002	-33.4926
42.3056	3.3781	-7.6272	-35.8887	-18.1187
68.2419	9.4832	-13.7490	-52.5665	-31.9874

Cont'd of Pc's retained

55.5405	-0.6064	15.0030	-31.9888	-35.9141
63.7510	5.2757	-16.5092	-51.1795	-39.3506
67.1671	11.6725	-0.1494	-47.3593	-23.8207
42.5667	-3.0429	14.0543	-32.7716	-34.9761
55.6613	3.6145	13.7439	-37.9334	-35.2185
41.6308	0.5910	-7.6354	-46.6448	-25.8706
43.3566	3.3917	-25.5522	-49.7888	-18.1880
41.3576	-1.5684	-1.0020	-36.4526	-22.5160
74.1858	12.2197	10.8761	-43.7979	-29.1097
75.4650	13.4648	-2.8402	-52.1987	-25.8532
42.2823	-2.0586	2.4877	-42.6063	-30.9391

4.7 Results by the Data Simulated from the Real Life Data

The computation using the proportion of variance accounted for shows that, 9 PC's are to be retained for further analysis. A model was fitted using the 9 PC's and R^2 was found to be 23.6% and R^2 (adjusted) was 5.3% and for eigenvalue one criterion 5 PCs were retained and R^2 was 28.3% and R^2 (adjusted) is 7.1% This result enhances and authenticates the previous result obtained by the original data, i.e. it confirmed that eigenvalue one criterion of data reduction techniques, is more reasonable for the reduction of a data of cholesterol level of Human body.

4.8 Contribution to Knowledge

The research has been able to detect that Eigenvalue- one criterion of PCA is more suitable in the reduction of a data of cholesterol level of Human body.

The research illustrates the use of Principal component Analysis technique of Multivariate statistics in the reduction of data of a high dimension.

5.0 CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary

The study was able to use Principal Component Analysis technique to determine the reasonable components that are worth to be retained in observing cholesterol composition of human body using data collected from Ahmadu Bello University Teaching Hospital, Shika, Zaria. Moreover, the study has been able to illustrate the use of proportion of variance accounted for and Eigenvalue one criterion to retain the variables that will account for most of the variability in the data set and select the best from the two by the use of Model Adequacy Checking procedure and the number of PC's retained for further analysis. The use of Eigenvalue one Criterion is recommended as the best method to be employed in dealing with a data of high dimension based on the result obtained from the analysis.

5.2 Conclusion

Principal component analysis is primarily the most important aspect of multivariate techniques in the reduction of data of a high dimension. In view of the fact that some of the variables are correlated with one another because they are measuring the same proposition, we believe that it should be possible to reduce the observed variables into a smaller number of Principal Components (artificial variables) that will account for most of the variance in the observed variables. The principal component obtained is then used as predictor or criterion variables in subsequent analysis. Eigenvalue-one criterion method of data reduction in PCA is what was found to be more preferred instead of proportion of variance accounted for approach in dimensionality reduction for the set of data of cholesterol level of human body.

5.3 Recommendation

We recommends the use of eigenvalue- one criterion methods of data reduction technique than the proportion of variance accounted for approach in PCA, which is one of the multivariate statistical Techniques for the reduction of a data of Cholesterol level of Human body in the Health sector globally.

Suggestion for Further Studies

Comparative study of the other various Methods of data reduction techniques in principal component analysis could be done.

The use of various Model Adequacy checking tools to make a comparison of the various approaches of dimensionality reduction in principal component analysis can be investigated.

The use of multivariate data set from different areas to strongly testify which among the various Methods is more reliable in the reduction of any set of data of high-dimension.

.

Publication

Gulumbe, S. U, Dikko, H. G and Haruna, S. (2015) Comparing two Approaches of Data Reduction Techniques in Principal Component Analysis. *International Journal of Innovative Research in Technology & Science (IJIRTS)*, 3(1), 01-04.

References

Alvin, C.R. (2002). *Method of Multivariate Analysis*, Second Edition, John Wiley and Sons, Inc. publications), pp. 112-120.

Anderson, T.W. (1963). Asymptotic theory for principal component analysis, *Annals. Mathematics and Statistics*,34(2), 122-148.

Autar, J.T. (2000). *Adequacy of regression model*. University of south Florida. Retrieved August 24, 2012, from U.S.Akar@eng.usf.edu.

Chatfield, C. and Collin, A.J. (1990). *Introduction to Multivariate Analysis.* Chapman and Hall Publication,. pp. 246-413

Diamantas, K.I and Kung, S.Y. (1996). *Principal Component Neural Networks-Theory and Applications.* John Wiley, New York. p. 272.

Feeney, H. (1967). *Applied Multivariate Analysis Using Bayesian And Frequentist Method.*Springer, p. 155.

Flury, B. (1988) *Common Principal Components and Related Multivariate Models.* John Wiley, New York.

Forzani, L. and Tolmasky, C.F.(2001) . On the Spectral Decomposition of Empirical Correlation Matrices. *Journal of Knot Theory & Ramifications*, 10(8): 1201-1214.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate data analysis. *Biometrika*,(53): 315-328.

Hoffman, K. and Kunze, R.A. (1971). *Linear Algebra,* Second Edition, Springer, pp. 6-25.

Hotelling, H. (1933). Analysis of complex of statistical variables into principal components *Journal of Educational Psychology,* (4):419-441 and(4): 498-520.

Jeffer, J.N.R., (1967). Two case studies in the application of principal component analysis. *Journal of Applied statistics.*(16):225-236.

Jim, F. (2013). Multiple Analysis: Use of Adjusted R^2& Prediction R^2 to include the Correct Number of Variables, *The Minitab log, Wikipedia Encyclopedia.*

Johnson, R.A. and Wichern, D.W (1996). *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice Hall of India, Upper Saddle River, NJ. p.450

Jolliffe, I.T (2002). *Principal Component Analysis*, second edition, New York :Springer-Verlag New York, p. 647.

Jonathan, S. (2009). A tutorial on principal component analysis, derivation, discussion and singular value decomposition. *International Journal Of Computer Application,* vol(54): 233-245

Kaiser, H.F. (1960). The application of electronic compters to factor analysis. *Educational and Psychological Measurement* vol.20: 141-151

Kim, E. and Paul, G. (1987). Principal Component Analysis. *Elsevier Science Publishers B.V Amsterdam,* vol (2): 31-52.

Kim, J.O. and Muller,C.W. (1978). *Factor Analysis*. Statistical methods and practical issues; Newbury Park.

Lessard, D. R. (1973).International portfolio diversification: A multivariate analysis for a group of latin American Countries. *Journal of Finance,* 28(3): 619-633.

Makridakis,S . G. (1974). An analysis of the Interrelationships among the Major World Stock exchanges. *Journal of Business Finance and Accounting,*1(2):195-208

Malava, A .(2006). Principal Component Analysis on Term Structure of Interest Rate.Independent Research Project in Applied Mathematics, Helsinki University of Technology. Department Engineering of Physics and Mathematics. Unpublished.

Marcus, F. and Minc P. L. (1988). *C R C Concise Encyclopedia of Mathematics*, Second Edition, p. 144.

Neil, H. L. (2002*).* Applied Multivariate Statistics. In: Springer text in statistics, *Applied Multivariate Analysis* p. 445-448. Courses and Seminars held at the University of Pittsburgh USA, Published in Library of Compress Cataloguing-in-Publication.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6): 559-572.

Peter, M.C. (2002). What is a Statistical Model, University of Chicago.*The Annals of Statistics*,(30)5: 1225-1310.

Philippatos, G.. C., Christofi, A., & Christofi, P. (1983). The inter-temporal stability of international stock market relationships: Another view. *Financial Management,*12(4): 63-69.

Smith, L. (2002). A tutorial On Principal Components Analysis. Paper Presented At a Seminar Held in the Department of Mathematics, University of Otago In Dunedin, New Zealand. Unpublished Article.

Appendices

Appendix 1: Samples of 40 Patients from Ahmadu Bello University Teaching Hospital for Cholesterol Composition.

X_1	X_2	X_3	X_4		X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
17.0	1.633	52.07	67.2	76.0	10.4	6.1	19.5260	41.6	6.0	4.6	23.6	
32.0	1.687	67.72	82.5	72.0	9.9	6.4	23.7950	42.8	6.7	5.5	26.3	
25.0	1.582	48.89	66.7	88.0	10.2	6.5	19.5340	39.0	5.7	4.6	26.3	
55.0	1.582	72.89	93.0	60.0	11.4	7.6	29.1245	40.2	6.2	5.0	32.6	
27.0	1.514	57.65	82.6	72.0	9.4	5.8	25.1505	36.2	5.5	4.8	29.2	
29.0	1.615	55.83	75.4	68.0	10.1	6.6	21.4055	43.2	6.0	4.9	26.4	
25.0	1.519	50.66	73.6	80.0	10.8	6.1	22.0126	38.7	5.7	5.1	27.9	
12.0	1.608	49.89	81.4	64.0	10.4	4.1	19.2945	41.0	6.8	5.5	33.0	
4.1	19.295	41.00	6.8	5.5	33.0	7.2	33.3512	43.8	7.8	5.8	38.6	
41.0	1.725	99.24	99.4	68.0	12.3	6.1	20.5636	37.3	6.3	5.0	26.5	
7.2	33.351	43.80	7.8	5.8	38.6	6.2	29.7637	42.3	6.6	5.2	34.4	
32.0	1.559	49.98	67.7	68.0	9.3	6.1	17.6644	37.3	6.3	5.0	26.5	
31.0	1.694	85.41	100.7	80.0	8.9	6.2	24.0231	42.3	6.6	5.2	34.4	
19.0	1.645	47.80	72.9	76.0	11.2	6.2	31.9803	39.1	5.7	4.8	23.7	
19.0	1.603	61.73	85.0	68.0	10.7	4.8	24.6398	40.3	6.6	5.1	28.4	
23.0	1.608	82.73	85.7	72.0	11.6	6.2	17.1366	48.6	7.2	5.6	34.0	
27.0	1.643	53.97	74.5	68.0	10.0	5.3	19.9933	41.5	6.3	5.3	27.0	
45.0	1.653	73.43	94.0	72.0	12.7	7.4	26.8738	40.0	6.8	5.0	35.0	
41.0	1.572	79.97	92.8	64.0	10.7	6.7	32.3607	38.2	6.8	4.7	33.1	
56.0	1.633	82.19	105.5	80.0	11.6	7.1	30.8208	38.2	6.9	5.4	39.6	

22.0	1.610	56.38	75.5	64.0	9.7	6.4	21.7507	38.2	5.9	5.0	27.0
57.0	1.542	48.39	70.0	76.0	10.6	5.9	20.3512	38.1	6.1	5.0	23.6
24.0	1.610	68.02	98.0	76.0	11.0	7.0	26.2412	38.0	7.0	5.1	34.3
37.0	1.590	73.98	104.7	76.0	10.5	6.9	29.2630	36.0	6.7	5.1	34.4
59.0	1.539	42.77	67.8	80.0	11.8	8.2	18.0578	32.1	5.4	4.2	23.3
40.0	1.613	72.43	99.3	104.0	13.3	8.3	27.8384	31.1	6.4	5.2	35.6
45.0	1.488	73.84	91.1	88.0	11.3	7.5	33.3498	39.4	7.1	5.3	31.8
52.0	1.529	58.96	74.5	60.0	11.3	6.6	25.2202	40.2	5.9	5.1	27.0
31.0	1.717	81.60	95.5	76.0	10.7	6.7	27.6797	39.2	6.2	5.0	32.8
32.0	1.628	67.04	79.5	72.0	9.5	5.9	25.2943	39.0	6.3	4.9	31.0
23.0	1.592	51.21	69.1	72.0	10.8	7.2	20.2051	36.6	5.9	4.7	27.0
23.0	1.557	88.72	105.5	88.0	11.4	7.9	36.5976	27.0	7.5	5.5	41.2
47.0	1.478	56.33	78.8	80.0	10.4	7.3	25.7862	38.5	5.6	4.7	25.5
36.0	1.605	61.23	85.7	60.0	12.5	7.3	23.7694	39.9	6.4	5.2	30.9
34.0	1.651	64.13	92.8	72.0	12.4	8.5	23.5270	37.5	6.1	4.8	27.9
37.0	1.536	56.20	72.7	88.0	9.2	4.6	23.8216	39.7	5.8	5.0	26.5
18.0	1.569	61.46	75.9	88.0	11.9	8.1	24.9664	39.0	6.3	4.9	27.8
29.0	1.727	45.67	68.6	124.0	9.3	6.4	15.3126	41.6	6.0	5.3	23.0
48.0	1.701	43.00	68.7	64.0	10.6	6.4	14.8614	33.8	5.6	4.6	26.4
16.0	1.448	52.07	67.2	76.0	10.4	6.1	24.8342	41.6	6.0	4.6	23.6

Appendix II: Principal Components Retained by Eigenvalue One Criterion from the Real Life Data.

PC_1	PC_2	PC_3
-34.8338	74.199	-31.0793
-41.5238	91.995	-30.6841
-41.2970	76.422	-28.1642
-44.4394	105.171	-21.4108
-40.5032	88.665	-27.0311
-36.6310	81.698	-28.8352
-39.3545	80.635	-27.9666
-30.6477	80.747	-33.9194
41.7419	46.257	-6.5639
-50.9966	113.309	-31.6376
46.3881	42.803	-2.5052
-36.5358	76.229	-24.8466
-50.2718	111.379	-36.2460
-34.5678	79.284	-26.2166
-36.3855	88.973	-32.2965
-38.7874	99.198	-39.6515
-35.9061	80.171	-29.4420
-45.9182	106.030	-25.8114
-42.5164	108.555	-25.3148
-55.1979	120.578	-25.0515
-34.1435	81.763	-28.0277
-46.5004	79.946	-18.6982

Cont'd of Appendix II

-44.0602	102.412	-31.4996
-50.2404	110.819	-28.1123
-49.0281	77.165	-13.1773
-59.8350	112.197	-25.7517
-51.3849	108.046	-25.8098
-38.7735	87.193	-19.5701
-46.3428	108.102	-31.3881
-40.2353	93.327	-27.9400
-35.2834	77.644	-25.8480
-50.4308	124.858	-28.2995
-47.6193	88.736	-21.8677
-36.9779	91.777	-25.1602
-45.1006	95.799	-26.3287
-46.2752	84.998	-27.4861
-40.9829	87.923	-30.4716
-57.9879	76.837	-33.6106
-39.9601	74.253	-17.5773
-33.6055	75.930	-30.0778

Appendix III: Principal Components Retained by Proportion Of Variance Accounted For, From The Real Life Data.

PC1	PC2	PC3	PC4	PC5	PC6
-34.8338	74.199	-31.0793	-17.3088	-14.3373	-35.2316
-41.5238	91.995	-30.6841	-33.0354	-5.5528	-35.7482
-41.2970	76.422	-28.1642	-17.1600	-18.8924	-29.1530
-44.4394	105.171	-21.4108	-52.9106	5.1317	-28.3858
-40.5032	88.665	-27.0311	-23.3985	-5.0043	-31.7627
-36.6310	81.698	-28.8352	-30.0598	-8.5053	-33.0259
-39.3545	80.635	-27.9666	-19.6173	-13.2235	-30.3071
-30.6477	80.747	-33.9194	-16.2120	-6.6973	-34.6984
41.7419	46.257	-6.5639	-28.1890	5.0359	-34.1800
-50.9966	113.309	-31.6376	-46.1642	5.8962	-37.6733
46.3881	42.803	-2.5052	-32.5526	5.9635	-31.3146
-36.5358	76.229	-24.8466	-28.4374	-10.6988	-24.8944
-50.2718	111.379	-36.2460	-32.9294	-1.1379	40.6240
-34.5678	79.284	-26.2166	-14.7139	-9.3000	-36.7042
-36.3855	88.973	-32.2965	-21.8857	-3.3675	-37.9768
-38.7874	99.198	-39.6515	-33.2794	-5.4469	-42.7950
-35.9061	80.171	-29.4420	-27.3881	-8.6001	-31.1414
-45.9182	106.030	-25.8114	-41.9282	-0.4054	-31.1112
-42.5164	108.555	-25.3148	-41.0638	6.7040	-35.8121
-55.1979	120.578	-25.0515	-47.5738	2.8181	-29.0305

Cont'd of Appendix III

-34.1435	81.763	-28.0277	-23.6797	-5.8145	-33.3707
-46.5004	79.946	-18.6982	-44.2656	-11.9788	-16.0046
-44.0602	102.412	-31.4996	-22.9517	-3.0570	-37.4315
-50.2404	110.819	-28.1123	-32.4389	1.4779	-34.1354
-49.0281	77.165	-13.1773	-40.9250	-15.8396	-10.5935
-59.8350	112.197	-25.7517	-23.1820	-10.9494	-28.7072
-51.3849	108.046	-25.8098	-35.4400	-4.8211	-33.3584
-38.7735	87.193	-19.5701	-48.2827	-1.8980	-23.7834
-46.3428	108.102	-31.3881	-31.4837	0.4064	-39.3613
-40.2353	93.327	-27.9400	-30.3217	-3.8615	-33.0403
-35.2834	77.644	-25.8480	-20.1962	-11.6931	-29.9196
-50.4308	124.858	-28.2995	-15.3524	3.4750	-40.3198
-47.6193	88.736	-21.8677	-36.7153	-9.6024	-25.3636
-36.9779	91.777	-25.1602	-37.1146	-1.3247	-30.6325
-45.1006	95.799	-26.3287	-31.6346	-4.8200	-32.4920
-46.2752	84.998	-27.4861	-27.1037	-13.2737	-27.6742
-40.9829	87.923	-30.4716	-14.1150	-14.3864	-38.0656
-57.9879	76.837	-33.6106	-9.8356	-35.6271	-26.6536
-39.9601	74.253	-17.5773	-38.9958	-10.1331	-13.9137
-33.6055	75.930	-30.0778	-15.6666	-12.4149	-37.5325

Appendix IV: Eigenvectors from Correlation Matrix of a Real Life Data

-0.2687	0.1374	0.2648	-0.7399	0.0442	0.4162	-0.2903	-0.0072	-0.0165	0.0351	0.1690	0.0434
0.4338	-0.1135	0.1933	-0.0669	0.0514	0.0634	0.1006	-0.5985	0.3486	0.5005	0.0772	0.0646
-0.0756	0.4705	-0.1315	-0.2175	0.2150	-0.2648	0.4010	-0.0207	-0.3919	0.4158	-0.0070	-0.3228
-0.3246	0.3741	-0.1313	-0.0179	0.1482	-0.0820	0.1493	0.1215	0.7883	-0.0015	-0.2101	0.0760
-0.3900	0.1020	-0.1035	0.3262	-0.3891	0.0101	-0.2694	0.0735	-0.0415	0.5742	0.3201	0.2386
0.4434	-0.0609	0.2412	-0.0870	-0.0596	0.0451	0.0894	0.7768	0.1451	0.3044	0.0465	-0.0327
-0.0313	0.2283	0.5458	-0.0785	-0.6600	-0.3347	0.1114	-0.1156	0.0039	-0.1573	-0.2049	-0.0514
0.1959	0.3481	0.2453	0.1677	0.3723	-0.3528	-0.6955	-0.0092	-0.0152	-0.0045	-0.0298	-0.0640
0.1830	-0.1580	-0.4923	-0.4962	-0.1995	-0.5372	-0.1838	0.0186	0.0432	0.0078	0.0050	0.2976
0.2832	0.3723	-0.1994	0.0371	-0.2002	0.0774	0.0485	-0.0466	0.1620	-0.3259	0.7090	-0.2329
0.2639	0.2600	-0.3854	0.0098	-0.3363	0.4189	-0.2741	-0.0425	-0.0242	0.0975	-0.5144	-0.2718
0.2391	0.4340	0.0196	0.0676	0.0709	0.1927	0.1857	-0.0045	-0.2268	-0.1116	-0.0800	0.7762

Appendix V: Eigenvalues Of The Correlation Matrix Of A Real Data

4.2966	0	0	0	0	0	0	0	0	0	0	0
0	3.2835	0	0	0	0	0	0	0	0	0	0
0	0	1.6816	0	0	0	0	0	0	0	0	0
0	0	0	0.7074	0	0	0	0	0	0	0	0
0	0	0	0	0.5650	0	0	0	0	0	0	0
0	0	0	0	0	0.4999	0	0	0	0	0	0
0	0	0	0	0	0	0.3615	0	0	0	0	0
0	0	0	0	0	0	0	0.0145	0	0	0	0
0	0	0	0	0	0	0	0	0.0440	0	0	0
0	0	0	0	0	0	0	0	0	0.2557	0	0
0	0	0	0	0	0	0	0	0	0	0.1379	0
0	0	0	0	0	0	0	0	0	0	0	0.1523

Appendix VI: Data Simulated from the Original Data Set.

19.43	2.65	55.99	69.18	92.56	4.6	5.86	19.7	41.26	6.47	4.86	31.01
31.42	5.11	49.95	83.39	63.31	14.25	7.31	24.25	35.83	6.47	4.72	31.32
28.15	2.99	49.44	52.95	49.52	6.23	5.76	23.01	41.65	6.23	4.85	38.92
26.98	5.15	32.31	68.26	87.16	15.34	6.92	22.56	40.63	6.91	5.21	34.57
6.44	3.56	55.09	87.73	90.97	12.92	6.55	14.77	40.41	6.28	4.61	32.85
7.66	1.03	33.51	74.36	43.6	18.76	6.84	15.23	41.76	7.23	5.19	31.15
27.57	7.53	37	84.5	85.69	11.51	6.02	22.79	38.11	6.54	5.01	30.27
54.37	3.21	67.44	89.9	70.25	17.7	7.73	32.96	37.06	6.02	4.81	31.53
29.37	3.17	72.31	24.87	65.46	14.67	7.39	23.47	40.4	6.59	5.16	19.67
27.58	4.59	53.82	71.41	78.99	8.24	5.97	22.79	34.65	5.69	4.61	29.93
27.26	-3.94	67.56	83.8	100.07	16.02	6.91	22.67	36	5.86	4.56	27.14
16	-10.76	86.46	81.43	44.74	16.7	6.93	18.4	38.94	6.28	5.14	30.46
51.54	7.35	63.12	96.89	99.83	11.76	6.04	31.88	41	7.28	5.06	21.59
19.96	8.95	72.75	46.63	95.5	12.25	6.6	19.9	39.66	5.62	5.07	34.25
25.79	5.16	42.64	75.55	112.44	10.08	8.14	22.11	38.8	6.65	4.57	25.36
7.41	-2.68	46.96	84.81	83.33	20.16	7.49	15.14	35.69	6.56	5.29	31.56
11	0.28	47.01	87.28	62.2	10.02	6.25	16.5	38.9	6.65	4.92	29.5
45.85	16.18	58.72	107.79	91.36	14.76	6.93	29.72	42.2	5.9	4.74	31.63
42.03	-0.99	65.79	55.02	118.61	9.22	6.8	28.27	37.38	6.16	5.51	41.77
39.95	0.08	43.65	90.49	35.62	10.44	8.38	27.49	38.39	5.48	5.11	26.83
32.13	8.86	87.3	83.57	84.51	19.6	6.55	24.52	37.78	5.87	4.74	26.35
19.84	8.35	51.82	79.72	48.76	10.19	5.07	19.86	44.47	6.25	5.43	31.9
52.76	-3.5	43.4	110.3	56.63	20.88	5.18	32.35	40.47	6.52	5.62	34.33
28.27	12.85	68.29	76.04	71.49	11.88	7.13	23.06	42.26	7.21	4.68	29.88
31.73	5.57	65.47	69.22	74.78	8.57	6.82	24.37	39.48	6.71	5.22	37.22
50.66	-0.94	47.38	47.94	66.44	11.7	7.63	31.55	40.36	6.22	4.79	20.93
28.99	2.21	77.41	85.82	45.84	24.06	6.23	23.33	30	6.26	4.97	29.27
27.56	9.3	59.38	15.63	60.55	15.05	5.2	22.78	35.29	6.07	4.49	23.84
41.48	8.73	53.88	87.34	96.55	17.32	6.94	28.07	38.05	5.93	4.47	32.97
18.46	-9.25	83.09	88.95	73.56	18.59	7.38	19.33	40.47	5.58	4.87	30.28
31.89	12.56	71.94	90.63	114.91	17.2	6.24	24.43	41.21	6.97	5.16	33.25

46.46	0.32	62.32	63.09	63.63	14.73	7.28	29.96	40.51	6.97	5.19	29.46
12.91	5.47	74.26	74.36	49.85	13.08	6.15	17.23	41.36	6.59	5.18	29.72
28.08	2.38	84.12	54.81	56.97	18.34	6.39	22.98	41.63	6.21	5.15	40.05
21.22	-5.15	58.57	52.6	91.48	8.93	5.79	20.38	45.1	7.35	4.83	30.6
27.61	7.78	38.27	42.31	102.67	12.79	8.27	22.8	38.67	6.49	5.17	26.16
16.29	0.13	53.56	63.09	73.6	15.32	8.89	18.51	38.07	7.75	5.35	24.69
47.69	-2.3	73.49	68.4	57.92	21.95	6.31	30.42	39.31	6.76	5.17	35.74
50.54	-8.47	54.97	81.88	85.19	17.74	6.3	31.5	39.05	6.13	5.2	33
15.15	7.17	39.59	93.94	63.7	14.23	6.3	18.07	41.4	5.62	4.9	28.68

Appendix VII: Principle Component Retained by Eigenvalue One Criterion from the Simulated Data.

PC1	PC2	PC3	PC4	PC5
42.1340	-0.1906	-11.3085	-47.3592	-31.2653
57.0808	5.0768	1.2149	-43.5414	-28.8631
44.4111	3.6402	4.5756	-47.3867	-27.8679
46.9416	3.1217	-11.4690	-52.3256	-24.5726
37.7243	-5.8881	-2.5889	-40.5952	-36.0748
32.0715	-5.3474	14.3241	-36.8496	-24.3855
50.5622	3.3806	-11.8807	-50.6355	-28.2617
81.9548	15.1480	-0.1507	-48.7509	-30.5918
47.1860	4.1721	-4.7662	-35.3151	-17.9792
51.2074	3.3901	-7.7852	-43.8952	-29.5605
59.2389	3.2466	-6.4832	-40.4635	-31.2080
51.5847	-1.6826	27.6612	-25.9578	-34.5031
77.2140	13.9055	-18.6409	-55.4185	-29.6753
44.2792	0.0407	-11.4503	-42.3127	-33.0318
48.6353	2.5967	-24.3887	-50.6716	-25.3191
39.9285	-5.4652	3.6143	-35.6038	-29.6547
38.4509	-3.8821	6.4735	-38.5681	-30.6107
74.3267	11.4123	-13.3647	-56.7479	-37.5171

Cont'd of Appendix VII

63.7889	9.7239	-20.3210	-56.2872	-31.4851
62.1035	8.8233	10.2769	-44.0399	-23.8637
66.5017	5.3887	-1.6423	-37.6026	-37.0272
43.2001	-0.0045	7.9493	-43.8821	-32.4600
80.0518	14.4419	9.9730	-53.6832	-28.9806
53.8927	3.6915	-3.4779	-44.6321	-33.6688
55.3645	5.2099	-2.4638	-47.5422	-33.4832
63.8759	13.5128	-10.7654	-49.4817	-13.6962
65.0044	4.0160	19.2778	-27.7002	-33.4926
42.3056	3.3781	-7.6272	-35.8887	-18.1187
68.2419	9.4832	-13.7490	-52.5665	-31.9874
55.5405	-0.6064	15.0030	-31.9888	-35.9141
63.7510	5.2757	-16.5092	-51.1795	-39.3506
67.1671	11.6725	-0.1494	-47.3593	-23.8207
42.5667	-3.0429	14.0543	-32.7716	-34.9761
55.6613	3.6145	13.7439	-37.9334	-35.2185
41.6308	0.5910	-7.6354	-46.6448	-25.8706
43.3566	3.3917	-25.5522	-49.7888	-18.1880
41.3576	-1.5684	-1.0020	-36.4526	-22.5160
74.1858	12.2197	10.8761	-43.7979	-29.1097
75.4650	13.4648	-2.8402	-52.1987	-25.8532
42.2823	-2.0586	2.4877	-42.6063	-30.9391

Appendix VIII: Principle Component Retained by Proportion of Variance Accounted for from the Simulated Data set.

PC	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	
42.1340	-0.1906	-11.3085	-47.3592	-31.2653	-14.1369	8.5634	37.6311	-7.7419	
57.0808	5.0768	1.2149	-43.5414	-28.8631	-2.9609	12.5629	48.0588	0.3564	
44.4111	3.6402	4.5756	-47.3867	-27.8679	4.8183	5.4200	21.0716	7.5114	
46.9416	3.1217	-11.4690	-52.3256	-24.5726	-10.4719	-5.9058	49.3581	-6.2763	
37.7243	-5.8881	-2.5889	-40.5952	-36.0748	-14.2380	9.7978	56.4059	-6.7458	
32.0715	-5.3474	14.3241	-36.8496	-24.3855	0.5577	0.8433	49.7127	4.7413	
50.5622	3.3806	-11.8807	-50.6355	-28.2617	-11.9971	5.7840	58.7229	-7.1764	
81.9548	15.1480	-0.1507	-48.7509	-30.5918	-3.1849	17.2777	42.2522	-1.2605	
47.1860	4.1721	-4.7662	-35.3151	-17.9792	-10.1953	-1.1732	-9.9975	3.7277	
51.2074	3.3901	-7.7852	-43.8952	-29.5605	-9.3770	11.4723	38.3800	-4.5728	
59.2389	3.2466	-6.4832	-40.4635	-31.2080	-17.6154	11.4012	44.8770	-14.4339	
51.5847	-1.6826	27.6612	-25.9578	-34.5031	1.5787	23.7309	26.3579	3.0907	
77.2140	13.9055	-18.6409	-55.4185	-29.6753	-18.5130	17.1449	52.8137	-11.8892	
44.2792	0.0407	-11.4503	-42.3127	-33.0318	-14.4114	2.4011	14.4229	-0.4697	
48.6353	2.5967	-24.3887	-50.6716	-25.3191	-22.6709	1.2201	52.6281	-15.2014	
39.9285	-5.4652	3.6143	-35.6038	-29.6547	-12.1271	4.3092	57.9175	-9.0140	
38.4509	-3.8821	6.4735	-38.5681	-30.6107	-5.5030	13.9977	53.8616	-2.1726	
74.3267	11.4123	-13.3647	-56.7479	-37.5171	-12.1652	17.7358	65.5227	-2.1087	
63.7889	9.7239	-20.3210	-56.2872	-31.4851	-15.2535	0.6138	23.8945	-13.8854	
62.1035	8.8233	10.2769	-44.0399	-23.8637	4.7051	18.4220	49.5165	5.0521	
66.5017	5.3887	-1.6423	-37.6026	-37.0272	-13.4313	19.3959	33.3224	-0.6021	
43.2001	-0.0045	7.9493	-43.8821	-32.4600	-0.4987	13.3797	42.4240	8.2433	
80.0518	14.4419	9.9730	-53.6832	-28.9806	1.6201	13.6401	67.8883	-4.8152	
53.8927	3.6915	-3.4779	-44.6321	-33.6688	-7.0603	15.0676	34.5675	5.3569	
55.3645	5.2099	-2.4638	-47.5422	-33.4832	-3.9191	13.0299	30.0861	1.5802	
63.8759	13.5128	-10.7654	-49.4817	-13.6962	-7.4057	-1.0272	16.5295	-3.7100	
65.0044	4.0160	19.2778	-27.7002	-33.4926	1.2806	22.5060	35.5560	5.3757	
42.3056	3.3781	-7.6272	-35.8887	-18.1187	-6.9943	-6.4650	-10.1587	6.2793	

Cont'd of Appendix VIII

68.2419	9.4832	-13.7490	-52.5665	-31.9874	-12.9839	7.2313	53.5169	-7.2608
55.5405	-0.6064	15.0030	-31.9888	-35.9141	-8.3269	18.6409	38.4025	-5.1596
63.7510	5.2757	-16.5092	-51.1795	-39.3506	-20.3267	11.3709	51.4970	-8.4067
67.1671	11.6725	-0.1494	-47.3593	-23.8207	-2.6599	7.6259	22.6068	0.4285
42.5667	-3.0429	14.0543	-32.7716	-34.9761	-1.7731	18.9480	28.7640	9.0309
55.6613	3.6145	13.7439	-37.9334	-35.2185	2.3229	9.7872	8.6864	9.8742
41.6308	0.5910	-7.6354	-46.6448	-25.8706	-13.5510	-1.0885	22.4588	-9.7353
43.3566	3.3917	-25.5522	-49.7888	-18.1880	-19.5858	-11.7385	27.7048	-8.6836
41.3576	-1.5684	-1.0020	-36.4526	-22.5160	-10.6717	4.0065	33.4725	-3.4151
74.1858	12.2197	10.8761	-43.7979	-29.1097	2.1168	9.4738	21.6531	2.5503
75.4650	13.4648	-2.8402	-52.1987	-25.8532	-7.9236	5.4250	43.6415	-12.8444
42.2823	-2.0586	2.4877	-42.6063	-30.9391	-7.3403	10.2294	63.5115	0.3620

Appendix IX: Eigenvectors of the Correlation Matrix formed by the Simulated Data

0.5987	0.7071	-0.1566	-0.2669	0.0959	0.0713	0.0050	-0.1160	-0.0289	0.0127	0.1284	0.0225
-0.0624	0.0001	-0.4583	-0.1498	-0.2267	-0.0538	0.0990	0.1008	0.5323	-0.4094	-0.0476	-0.4859
0.1945	0.0002	0.1982	0.2094	-0.2320	0.0064	0.3644	-0.4984	0.1177	0.3314	-0.4791	-0.3015
0.1945	0.0002	0.1310	-0.0290	-0.1834	0.0013	0.3490	0.7867	-0.0669	0.3568	0.0092	-0.1947
0.0283	-0.0002	-0.4233	-0.1457	-0.0299	-0.3350	-0.1600	0.1541	-0.3058	0.0097	-0.7041	0.2167
0.3195	-0.0002	0.3720	0.2884	0.0983	-0.0641	-0.6309	0.1759	0.0254	-0.2094	-0.1958	-0.3883
0.0485	0.0001	-0.1551	0.2566	0.5647	0.2172	0.0092	0.1596	0.5825	0.2401	-0.2201	0.2659
0.5987	-0.7071	-0.1565	-0.2669	0.0959	0.0715	0.0054	-0.1160	-0.0288	0.0128	0.1285	0.0224
-0.2347	-0.0001	0.0372	-0.5353	-0.0204	-0.0761	-0.4381	-0.0852	0.2104	0.6008	0.0494	-0.2095
-0.2040	0.0001	0.0779	-0.3262	0.4969	0.4657	0.1842	0.0115	-0.3291	-0.1633	-0.2486	-0.3888
0.0190	0.0000	0.4840	-0.3618	0.2782	-0.5850	0.2740	0.0083	0.2423	-0.2437	-0.0798	0.1237
0.0342	0.0000	0.3180	-0.3103	-0.4342	0.5084	-0.1044	0.0893	0.2278	-0.2154	-0.2864	0.3940

Appendix X: Eigen Value of The Correlation Matrix formed by the Simulated Data.

2.2094	0	0	0	0	0	0	0	0	0	0	0
0	0.0000	0	0	0	0	0	0	0	0	0	0
0	0	1.7696	0	0	0	0	0	0	0	0	0
0	0	0	1.6626	0	0	0	0	0	0	0	0
0	0	0	0	1.4086	0	0	0	0	0	0	0
0	0	0	0	0	0.3966	0	0	0	0	0	0
0	0	0	0	0	0	0.3564	0	0	0	0	0
0	0	0	0	0	0	0	1.0797	0	0	0	0
0	0	0	0	0	0	0	0	0.5850	0	0	0
0	0	0	0	0	0	0	0	0	0.7016	0	0
0	0	0	0	0	0	0	0	0	0	0.9460	0
0	0	0	0	0	0	0	0	0	0	0	0.8845