Relationship between Principal Component Analysis and Factor Analysis

Shabir Ahmad

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
September 2017
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

____
Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

____
Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science

____
Asst. Prof. Dr. Yücel Tandoğdu
Supervisor

Examining Committee

1. Prof. Dr. Hüseyin Aktuğlu ________________________________

2. Asst. Prof. Dr. Nidai Şemi ________________________________

3. Asst. Prof. Dr. Yücel Tandoğdu ________________________________

ABSTRACT

In every field of scientific research and application, where the masses of data is available in multivariate form, the use of multivariate statistical analysis techniques can be implemented to achieve proper statistical inferences. The statistical modeling of data is the essential part of the multivariate analysis. The model might be the linear combinations of the original data, which can be created though the relationship between Principal Component Analysis (PCA) and Factor Analysis (FA). Such process of converting the entire data into the set of few clusters or linear models is called dimension reduction. Before applying FA, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy test for FA is used (12). Initial factor loadings and the variamx rotated factor loadings are computed via PCA approach. The estimated factor models generated by ordinary least square method, are further used for statistical control charts. Finally the generation of the uncorrelated statistical models using the relationship between PCA and FA is carried out to enable the estimation of the future outcomes.

Keywords: Correlation matrix, KMO test, Reducible Eigen space, dimension reduction, varimax rotation, uncorrelated statistical models, OLS estimated factor scores, statistical control charts.

ÖZ

Bilimsel araştırma ve uygulamanın her alanında, çok değişkenli verilerin var olduğu durumlarda, en uygun sonuçlar çok değişkenli istatistik analiz yöntemleri ile elde edilebilir. Verilerin istatistikslel modellemesi çok değişkenli analizin temel unsurudur. Bu modelleme Temel Bileşenler Analizi (TBA) ve Faktör Analizi (FA) arasındaki ilişkiden yararlanarak veriler arasında doğrusal kombinasyonların oluşturulması şeklinde olabilir. Verilerin alt gruplara veya doğrusal modellere dönüştürülmesine boyut indirgeme denir. FA yapılmadan önce, verilerin FA'ya uygunluğunun saptanması için Kaiser-Meyer-Olkin (KMO) ölçüm hesabı yapılır. İlk faktör yükleri ve varimax metodu ile dönüşümü yapımış faktör yükleri TBA yaklaşımı ile hesaplanır. Minimum kareler yöntemi ile tahmin edilmiş faktör modeli istatistiksel control grafiklerinin oluşturulmasında kullanıldı. Son olarak TBA ve FA arasındaki ilişki kullanılarak ileriki oluşumların tahmininde kullanılmak üzere bağımsız istatistiksel modeller oluşturulmuştur.

Anahtar kelimeler: Korelasyon matrisi, KMO test, indirgenebilir Eigen uzayı, boyut indirgeme, varimaks döndürümü, enküçük kareler metodu ile tahmin edilmiş faktör skorları, istatistiki Kontrol grafikleri.

DEDICATION

I am dedicating this thesis to my parents

ACKNOWLEDGMENT

First and foremost, I would like to thank Allah Almighty for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

I would like to thank my supervisor Asst. Prof. Dr. Yücel Tandoğdu for his continuous support and guidance for preparing of this study. Without his valuable supervision, all my efforts could have been short-sighted.

Finally, I would like to dedicate this thesis to my parents whose dreams for me have resulted in this achievement and without their loving upbringing and nurturing; I would not have been where I am today and what I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{X}$	Sampling data vector
$\bar{\mathbf{X}}$	Mean data vector
$\bar{x}$	Sample mean
μ	Population mean
$\boldsymbol{\mu}$	Population mean vector
$\mathbf{S}$	Sample covariance matrix
$\boldsymbol{\Sigma}$	Population covariance matrix
ρ	Population correlation coefficient
$\mathbf{R}$	Sample correlation matrix
λ	Eigenvalue
V	Eigenvector
c	Statistical distance
$\rho_{Y_i X_i}$	Correlation between the i^{th} PC and the i^{th} variable
l_{ii}	Factor loading of the i^{th} CFs and the i^{th} variable
$h^2{}_i$	i^{th} Communality
ψ_i	i^{th} Uniqueness
$\hat{F}_i$	i^{th} Estimated common Factor
$\hat{F}_i^*$	i^{th} Estimated rotated Common Factor
PCs	Principal components
CFs	Common Factors
PCA	principal components analysis

FA	Factor analysis
KMO	Kaiser-Meyer-Olkin Sampling Adequacy Test
OLS	Ordinary least square method

Chapter 1

INTRODUCTION

In every field of data analysis, the data are typically collected by researchers through the experimental units. These can be inanimate subjects, human subjects, plants, countries and a wide range of other objects. However, in multivariate analysis, it is sometimes tedious to isolate and study each variable individually. It is essential to study all variables simultaneously, to achieve completely understandable structure and clear configuration of the data. From this point of view, the multivariate statistical techniques will help to make proper statistical conclusions. Initially applications of multivariate methods were only limited to the psychological problems of human intelligence, but currently it is broadly used in quality control, pharmaceutical companies, DNA microarrays, marketing research, industries and telecommunications etc [9].

The aim of this study is the creation of statistical models for multivariate data through the basic relationship between Principal Component Analysis (PCA) and Factor Analysis (FA). PCA constructs the linear transformations of the multivariate data using covariance or correlation matrices. Moreover, these transformations are in fact the statistical models and are largely concerned with exploring and explaining the characteristics of the data. Basically, PCA reduces the number of dimensions and it is heavily utilized in FA to determine the appropriate number of factors and variables for subsequent analysis.

But in some fields they are interchangeably used unconsciously. In FA, the investigators make the assumptions that there exists an underlying model for the data, while PCA is just a mathematical model of the original variables without any assumptions about the variance - covariance matrix. It can be simply employed to condense the data without loss of information. In the case, if the factor model is erroneously applied to a particular data and the assumptions about the covariance matrix are completely unspecified, then FA will lead to improper conclusions or vice versa.

Additionally, the main focus of this thesis is on the generation of the factor model as well as a proper interpretation of the factor model through a case study using a multivariate data set. Generally, the factor model is estimated by ordinary least square method. The estimated factor scores of the factor model are very useful in diagnosing the characteristics of the data.

Chapter 2

LITERATURE REWIEW

In 1904 the first idea of FA was proposed by an English statistician Charles Spearman (1) in the field of modern psychology. He discovered that a single artificial factor called *g* factor could be considered as general intelligence factor. The intellectual performance of the human brain depends on many different variables. Spearman associated all the variables to the *g* factor. Subsequently this idea was developed into a new statistical technique called factor analysis, where the association between the variables was examined. His findings was published in the American Journal of Psychology under the title “General intelligence objectively determined and measured”. According to Spearman Theory, all the test measurements of human intelligence are directly associated, such that it can be modeled by a specific underlying factor of the various mantel abilities [1].

In 1940, Raymond B. Cattell (2) extended the Spearman idea of the g factor theory to the multi-factor theory for human intelligence using the same factor analysis. He started his research with a Personality Factor Questionnaire, in which he included 15 different personality factors including the g factor. The Cattell Factor Analysis was based on the correlation matrix, he found that the sixteen factors themselves are correlated and their scores can be measured on the two uncorrelated factors, which he called extraversion and introversion for the human ability test [2, 3, 4].

In 1901, the first concept of PCA was discovered by Karl Pearson. His main idea was that how to transform or rotate the multi-dimensional data to the low dimensional data. He found the method of transforming original coordinate system to the new coordinate system and also the representation of the best fit lines for the system of points in a multi-dimensional scatter plot [5].

In 1930, Thurston found that PCA and FA are both separate techniques for numerical problems. But due to some insufficient knowledge both are interchangeably used [6].

In 1933 Harold Hoteling used the PCA as data reduction technique in factor analysis. His paper published in the Journal of Educational Psychology named "Analysis of a complex statistical variables into principal components" dealt with the statistical process that transforms the huge volume of data to the low volume data by the set of few uncorrelated variables. However, the method for multivariate statistical data analysis could not be applied to real life problems with large multivariate data due to the volume of computation involved. With the advent of electronic computation starting from 1960s onwards, application of PCA and FA became possible [7]. Three years later in 1936, Hotelling introduced the method of computing PCs by using power method [8].

During the World War II in 1939, Girshick gave another derivation of PCs by using maximum likelihood estimation and he also introduced the sampling theory in the field of PCA [10]. In 1966, Gower J C discussed the geometrical and theoretical interpretation of PCA in the field of FA and other statistical analysis [11].

In 1970, Henry Kaiser proposed the idea of testing the measure of sampling adequacy for factor analysis [12]. Later in 1974, this was improved by Kaiser and Rice [13]**.** This statistic was used to compare the square entries of image correlation matrix and usual correlation matrix. This test is usually called Kaiser-Meyer-Olkin KM sampling adequacy test, abbreviated as KMO [13]. In 1972, Vavra used the PCA as a feature extraction technique before conducting the regression analysis for the solution of economic problems [14].

In 1976, Jackson, J. E. and Lawton, W. H. used another application of PCA in cross impact analysis, dealing with estimating the impact of one outcome given that the likelihood of other outcome is already known [15]. In 1988 Brown used a wide application of PCA in field of chemistry for mass spectroscopic and gas chromatographic problems in which the data measured at the various time intervals [16].

In 1999, Fabrigar claimed that PCA and factor analysis are similar techniques in a few statistical fields. He addressed that principal component and factor analysis can yield the same output. But in low communalities cases both methods will provide different outputs. He also proposed that any data which satisfy the assumptions of factor analysis exists as an underlying model, and the results of this model can be more accurate then PCA results [17].

Chapter 3

MATRIX THEORY AS USED IN MULTIVARIATE STASTISTICS

In this study it is aimed to investigate the relationship between PCA and FA, based on the certain multivariate statistical data analysis concepts. The statistical techniques utilizing some matrix theory will be used to detect the structure and pattern of the huge volume of multivariate data. This will be achieved by first computing the variance-covariance and correlation matrices. Then the relationship between PCA and FA will be explained. However, in this chapter the theory establishing a link between matrix algebra and statistical analysis is explored. In Chapter 4 the summarized theory will be used for dimension reduction of data, modelling, exploring, interpreting and making statistical inferences of available data in a multidimensional environment. Application of such theory necessitates the use of advanced statistical software.

3.1 Matrix Terminologies

In multivariate statistical data analysis when the number of variables are more than two, statistical computations necessitates the use of computer software packages. In this section the use of matrix algebra in statistics is explained.

3.1.1 Matrix Representation of Data

Definition 3.1. In multivariate statistical analysis representation of the data in matrix form is essential. A data with *p* variables and *n* observations can be represented by the matrix X of the size $n\times p$, denoted as

$$\mathbf{X}_{(n\times p)} = \begin{pmatrix} x_{11} & x_{12}\cdots & x_{1p} \\ x_{21} & x_{22}\cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & x_{n2}\cdots & x_{np} \\ & & \end{pmatrix} = \left[X_1, X_2, \cdots, X_p\right] \qquad (3.1.1)$$

where *n* represents the number of observations of the data in each column and *p* represents of the variables in each row [18].

3.1.2 Mean Data Matrix

Definition 3.2 Let $\boldsymbol{X} = \left[X_1, X_2, ..., X_p\right]$ be a random vector containing *p* random variables each with *n* observations. Then the sample mean of the *p* variables can be represented by the following vector.

$$\bar{\mathbf{X}} = \frac{1}{n}\left[\sum_{j=1}^{n} x_{j1}, \sum_{j=1}^{n} x_{j2}, ..., \sum_{j=1}^{n} x_{jp},\right] = \left[\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p\right] \qquad (3.1.2)$$

where each sample mean contained in the sample mean vector measures central tendency of the corresponding random variable [18].

3.1.3 Sample Variance

Definition 3.3 Amount of variability of a single random variable with *n* observations $x_1, x_2, \cdots, x_n$, about its mean $\bar{x}$, can be computed as

$$s_k^2 = s_{kk} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 \qquad k = 1,2,\ldots,p\,. \qquad (3.1.3)$$

Here, *k* represents the number of columns and *j* represents the number of rows of the data matrix **X**. This statistic is commonly used to determine the dispersion among the data points around the sample mean and it is also called measure of spread. It helps to understand the shape of the data [18].

3.1.4 Sample Covariance

Definition 3.4 Let $X_1 = \left[x_{11}, x_{21}, \cdots, x_{n1}\right]$ *and* $X_2 = \left[x_{12}, x_{22}, \cdots, x_{n2}\right]$ be a bivariate random sample of size *n* drawn from two populations, assuming that random variables X_1 and X_2 have a joint probability distribution $f\left(x_1, x_2\right)$. Then the joint variability of X_1 and X_2 is given by

$$Cov\left(X_1, X_2\right) = s_{12} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2) \tag{3.1.5}$$

In general, the measure of linear relationship between the i^{th} and k^{th} variables for $i = 1,2,\cdots,p$ & $k = 1,2,\cdots,p$, can be defined as

$$Cov\left(X_i, X_k\right) = s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \tag{3.1.6}$$

It is useful to estimate the linear associations of any two variables under the same unit [18].

3.1.5 Sample Variance Covariance Matrix

Definition 3.5 In general, the covariance of multivariate data can be expressed by the covariance matrix **S**,

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \right\} \tag{3.1.7}$$

Here the diagonal elements of matrix **S** shows variances of the *p* variables while the off diagonal entries are covariances between the variables X_i and X_j [18].

3.1.6 Sample Correlation and Coefficient of Determination

Definition3.6 Correlation measures the linear dependency between two random variables X_i and X_k having different units of measurement. Mathematically it can be written as

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}} \quad for\ i = 1,2,\ldots p\ and\ k = 1,2,\ldots p \qquad (3.1.8)$$

the square of *r* is called *coefficient of determination* (r^2). It is the ratio of the amount of variation explained by regression equation, to the total variation of a data point from the regression equation [18].

3.1.4 Sample Correlation Matrix

Definition 3.7 In a multivariate random sample, the correlation coefficients between variables can be arrange in the matrix form as follows,

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & r_{pp} \end{bmatrix} \qquad (3.1.9)$$

Correlation coefficient between the two distinct variables is symmetrical. That is $r_{ik} = r_{ki}$ for all *i* and *k*. The correlation coefficient of a variable with itself is always one [18]. Therefore, the diagonal elements of the **R** matrix are 1.

3.2 Statistical Techniques

Statistical methods are commonly used to organize, summarize, analyse data, and make inference about the population from where the data is collected. In this section, the normal probability distribution and statistical approaches will be discussed to help clarify the idea of PCA and FA.

3.2.1 Normal Distribution

Normal distribution is one of the widely used continuous probability distributions in the field of statistical data analysis and the estimation of population parameters based on sample data.

3.2.2 Univariate Normal Distribution

Any statistical experiment associated with a probability distribution consisting of a single random variable of a normal population is called univariate normal probability distribution

Definition 3.8 Considers a univariate random variable *X* of a normal population with mean μ and variance σ^2 that is symbolically denoted as $X \sim N\left(\mu,\sigma^2\right)$. Then the probability density function $f\left(x\right)$ of this random variable *X* is called univariate normal probability distribution and is defined as

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};-\infty\leq x\leq\infty \qquad (3.2.1)$$

Graphically,

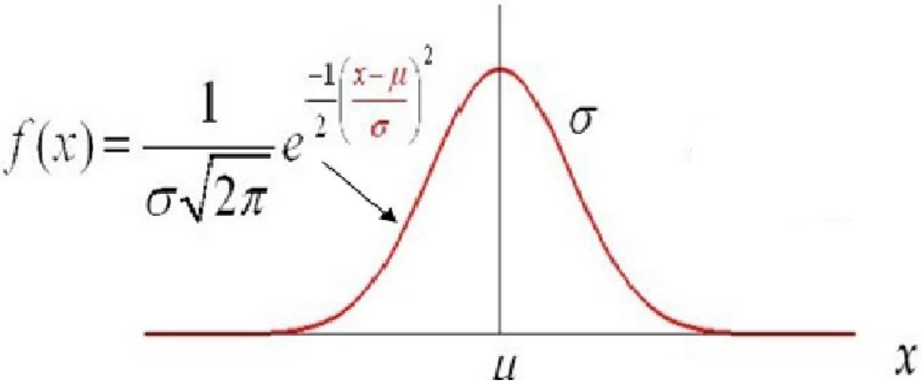


Figure 3.2.1. Graph of univariate normal distribution function.

Graph of the normal distribution is symmetric bell shaped curve. The shape of the curve is determined by two parameters. It is mean μ called centre of the distribution and variance σ^2 called measure of spread [18].

3.2.3 Mean and Variance of the Distribution of Sample Means $\bar{X}$

Definition 3.9 Let $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ with probability density function $f(\bar{x})$, then the population mean μ and variance $\frac{\sigma^2}{n}$ are given by the following.

$$\mu = E(\bar{X}) \tag{3.2.2}$$

Let prove the above quantity, starting from the definition of sample mean, that is

$$E(\bar{X}) = E\left(\frac{X_1 + X_2, \dots, X_n}{n}\right)$$

Using the expectation linear operator property, then

$$E(\bar{X}) = \frac{1}{n}\left[E(X_1) + E(X_2), \dots, E(X_n)\right]$$

As $X_1, X_2, ..., X_n$ are identically distributed this means that all have the identical population mean μ, then simply replacing expectation of the X_i by μ. That is

$$\begin{aligned}E\left(\bar{X}\right) &= \frac{1}{n}[\mu + \mu, ..., \mu] \\ &= \frac{[n\mu]}{n} \\ &= \mu\end{aligned}$$

Hence proved.

And

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \tag{3.2.3}$$

The proof of the equation (3.2.3) is given below

$$\begin{aligned}Var\left(\bar{X}\right) &= Var\left(\frac{X_1 + X_2 +, ..., +X_n}{n}\right) \\ &= Var\left(\frac{X_1}{n} + \frac{X_2}{n} +, ..., + \frac{X_n}{n}\right) \\ &= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) +, ..., + \frac{1}{n^2}Var(X_n) \\ &= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 +, ..., +\sigma^2\right] \\ &= \frac{1}{n^2}\left[n\sigma^2\right] \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Hence proved.

3.2.4 Standard Normal Distribution

Definition 3.10 A special case of the normal distribution with zero mean and unit standard deviation is called standard normal distribution. That is if $X \sim N\left(\mu,\sigma^{2}\right)$, then by definition

$$Z = \frac{x - \mu}{\sigma} \sim N(0,1) \qquad (3.2.4)$$

Therefore the probability density function of the transformed Z random variable is called standard normal probability density function, and is given by

$$f\left(z;0,1\right) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}z^{2}} \;;-\infty \leq z \leq \infty \,. \qquad (3.2.5)$$

This is also called Z distribution and is widely used for testing of hypothesis, and interval estimation in statistical inference [18].

3.2.5 Bivariate Normal Distribution

Definition 3.11 Let us suppose two independent random variables X_1 and X_2 have a bivariate normal distribution. Then the joint probability distribution of X_1 and X_2 is given by the following probability density function

$$f\left(x_{1},x_{2}\right) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\left(1 - \rho_{12}{}^{2}\right)}} exp\left\{-\frac{1}{2(1-\rho_{12}^{2}}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^{2} + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^{2} - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\} \qquad (3.2.6)$$

where σ_{11} and σ_{22} are the population variances of X_1 and X_2 respectively and ρ_{12} is the population correlation coefficient between X_1 and X_2.Graphically the bivariate normal distribution is as shown in Figure 3.2.2.

Geometrically,

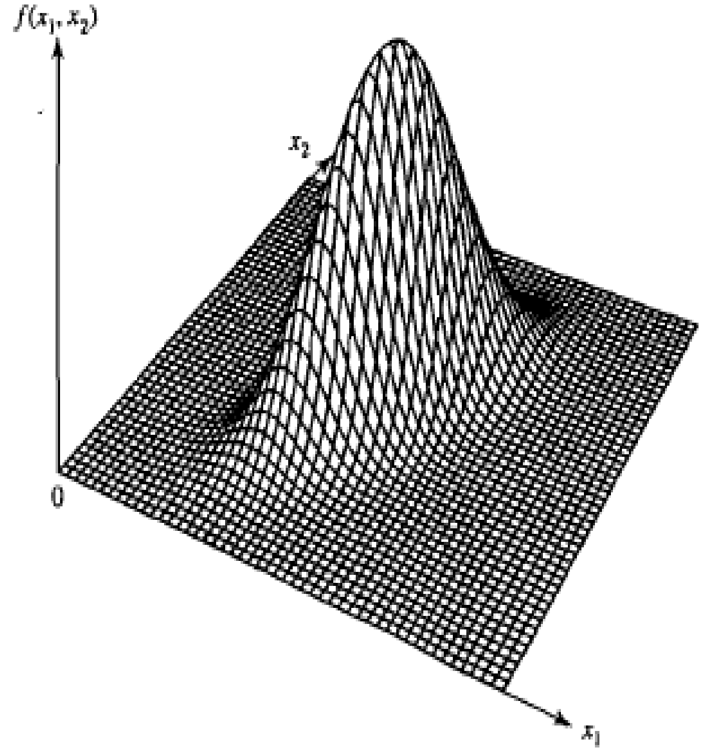


Figure 3.2.2: Graph of a BND figure out as a three dimensional bell shaped object.

The covariance matrix for the bivariate case can be written as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{21} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \tag{3.2.7}$$

Note that due to symmetry of the covariance matrix $\sigma_{12} = \sigma_{21}$.

Let ρ_{12} be population correlation coefficient between X_1 and X_2 given by

$\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$, then matrix $\mathbf{\Sigma}$ can be written as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{22} \end{bmatrix} \tag{3.2.8}$$

Since $\mathbf{\Sigma}$ is invertible matrix so the determinant of the $\mathbf{\Sigma}$ is non-zero and its inverse exists. That is

$$|\mathbf{\Sigma}| = \sigma_{11}\sigma_{22} - \rho_{12}{}^{2}\sigma_{11}\sigma_{22} = \sigma_{11}\sigma_{22}\left(1 - \rho_{12}{}^{2}\right) \tag{3.2.9}$$

and

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22}\left(1 - \rho_{12}{}^{2}\right)}\begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix} \tag{3.2.10}$$

Then the probability density of the bivariate normal probability distribution becomes,

$$f(x) = \frac{1}{(2\pi)|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-(\boldsymbol{X}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})/2} \qquad -\infty < \mathbf{X} < +\infty \tag{3.2.11}$$

with mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2]$ and covariance matrix **Σ**. That is symbolically matrix $\boldsymbol{X} \sim N_{p=2}\left(\mu, \sigma^2\right)$ [18].

3.2.6 Multivariate Normal Distribution

When the number of variables are more than two the joint probability distribution is known as multivariate normal distribution.

Definition 3.12 A data matrix **X** containing the *p* independent random variables $X_1, X_2, ..., X_p$ drawn from a multivariate population with mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_p]$ and covariance matrix **Σ** that is symbolically $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Then joint probability distribution of the *p* variables is given by

$$f(x) = \frac{1}{(2\pi)|\mathbf{\Sigma}|^{\frac{p}{2}}} e^{-(\boldsymbol{X}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})/2} \qquad -\infty < \mathbf{X} < \infty \tag{3.2.12}$$

In the multivariate case the covariance matrix **Σ** is given by

$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{pp} \end{bmatrix}$, when *p*=*1* the univariate normal distribution is obtained [18].

3.3 Relationship between Euclidean Distance and Statistical Distance

Euclidean distance is meaningless when the random fluctuations are involved in a process, since it is deterministic and cannot handle fluctuations in the values attained by the variables. While in statistical distance the fluctuations in variation are due to some random phenomena, and they may be correlated up to a certain degree. Accordingly the proper distance will depend upon the variations of the values taken on by the random variables, and correlation between the variables.

3.3.1 Euclidean Distance

Definition 3.13 Let $\boldsymbol{X} = \left[X_1, X_2\right]$ be a random vector with two random variables X_1 and X_2 with equal standard deviations and both are uncorrelated. Assuming X_1 and X_2 are standard normal, and $P = \left(x_1, x_2\right)$ any arbitrary point from **X**, then according to the Pythagorean Theorem, the Euclidean distance from P to $\mu = \left(0,0\right)$ is given by.

$$\begin{aligned} d\left(\boldsymbol{\mu}, \mathbf{X}\right) &= \sqrt{(x_1 - \mu_1)^2 + (x_1 - \mu_2)^2} \\ &= \sqrt{(x_1 - 0)^2 + (x_1 - 0)^2} \\ &= \sqrt{x_1^{\,2} + x_1^{\,2}} \end{aligned} \tag{3.3.1}$$

By taking the square of equation 3.3.1 the equation of the circle is obtaned. Such that

$$d^2\left(\boldsymbol{\mu}, \mathbf{X}\right) = x_1^{\,2} + x_1^{\,2} = c^2 \tag{3.3.2}$$

According to Euclidean distance, any points that satisfy the equation 3.3.1 will produced a constant distance such as *c*, and all of these points will be equidistance from the origin. This situation can be illustrated graphically as in Figure 3.3.1.

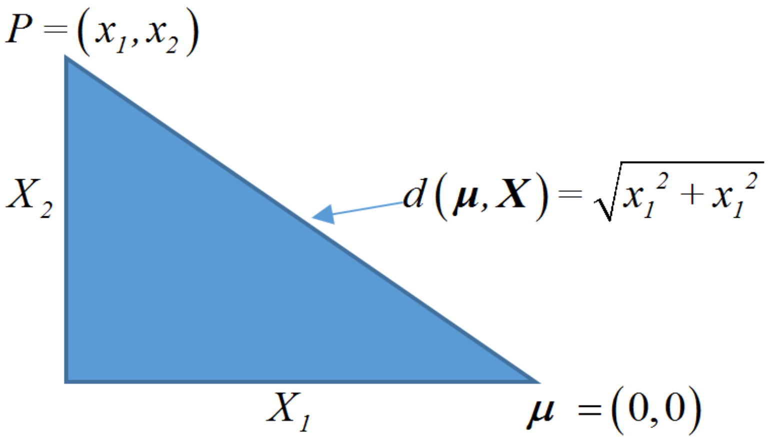


Figure 3.3.1. Representation of Euclidean distance from P to μ.

It is clear from the Figure 3.3.1 that the square Euclidean distance between P and μ. Generates the equation of circle basis on two independent variables having equal magnitudes.

3.3.2 Statistical Distance

Definition 3.14 Let X_1 and X_2 be bivariate random sample with variances s_{11} and s_{22} respectively, and let the $P':\left(\frac{x_1}{\sqrt{s_{11}}},\frac{x_2}{\sqrt{s_{22}}}\right)=\left(x_1^*,x_2^*\right)$ have the standardized coordinates obtained by dividing the coordinates of $P:(x_1,x_2)$ by their respective standard deviations. Then the statistical distance from $P'=\left(x_1^*,x_2^*\right)$ to $\mu\ =(0,0)$ can calculated as follow,

By using equation (3.3.1)

$$\begin{aligned} d(\mu,P') &= \sqrt{\left(x_1^*\right)^2+\left(x_2^*\right)^2} \\ &= \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2+\left(\frac{x_2}{\sqrt{s_{22}}}\right)^2} \qquad (3.3.3) \\ &= \sqrt{\frac{x_1^2}{s_{11}}+\frac{x_2^2}{s_{22}}} \end{aligned}$$

Geometrically,

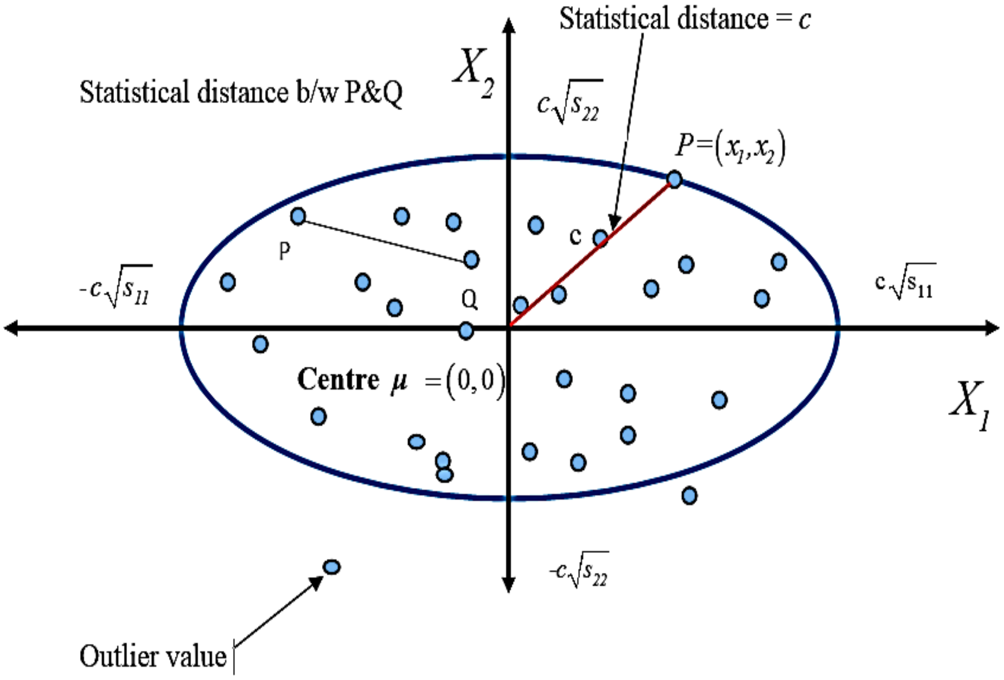


Figure 3.3.2. Graph of statistical distance

By taking the square of equation 3.3.3 the equation of the ellipse is obtained. That is

$$d^2(\mu, P') = \frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2 \tag{3.3.4}$$

It is clear any pair of points of **X** that satisfy the equation 3.3.4 will produce a constant square statistical distance from origin $\mu = (0,0)$ such as c^2.

Remark: An Euclidian distance is the radius of the points to origin, which lies on the circle and is constant. Whereas a statistical distance is the locus of the points from origin lie on the ellipse.

3.3.3 Confidence Ellipsoid

Definition 3.15 Let matrix **X** with *p* variables be normally distributed, that is **X** ~ $N(\boldsymbol{\mu}, \Sigma)$.Then the square statistical distance, produces the hyper ellipsoid that has chi-square distributed with p degrees of freedom. That is if

$$(\mathbf{X}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})=c^2 \qquad (3.3.5)$$

or

$$\mathbf{Z}=\frac{(\mathbf{X}-\boldsymbol{\mu})^2}{\boldsymbol{\Sigma}}\approx\chi^2_{p(\alpha)} \qquad (3.3.6)$$

Then all the **X** values must satisfy the following equation.

$$(\mathbf{X}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\le\chi^2_{p(\alpha)} \qquad (3.3.7)$$

where c^2 is a constant square statistical distance measured from **X** to population mean μ, and generates a hyper ellipsoid that contains $(1-\alpha)\%$ of observations. It can be estimated by the following equations.

$$\mathrm{P}\left[(\mathbf{X}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\le\chi^2_{p(\alpha)}\right]=(1-\alpha)100 \qquad (3.3.8)$$

or

$$\mathrm{P}\left[(\mathbf{X}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\le c^2\right]=(1-\alpha)100 \qquad (3.3.9)$$

Graphically,

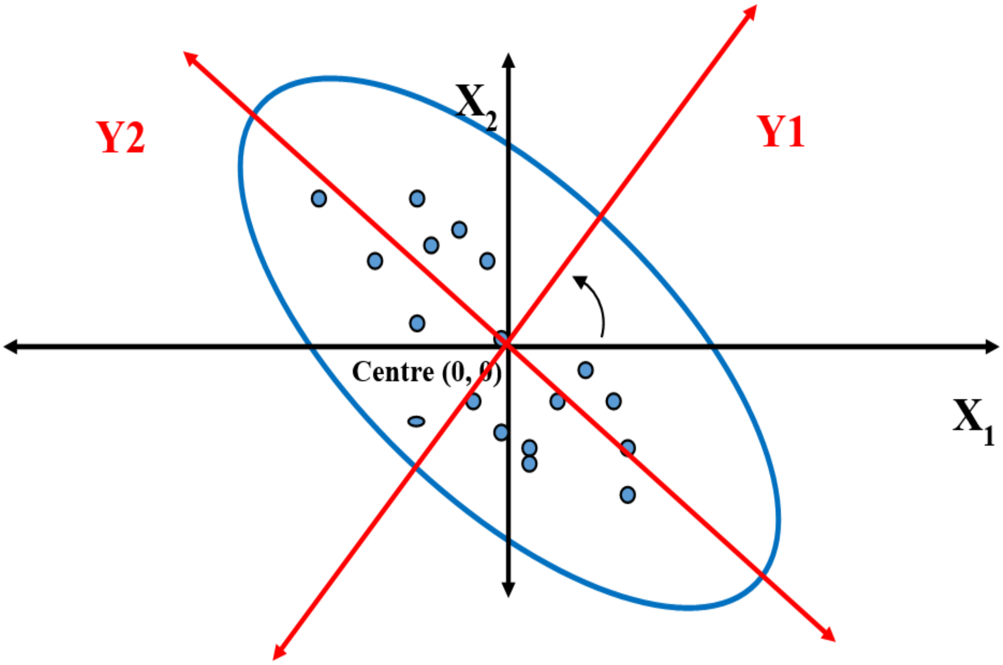


Figure 3.3.3. Representation of confidence ellipsoid for two normal distributions

Remark: The confidence ellipsoid is simply the contour of normal probability density function. It is broadly used for quality control and helps to detect the outliers and clean the data. When a data set is used, equation 3.3.7 becomes as

$$\left(\mathbf{X}-\bar{\boldsymbol{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}-\bar{\boldsymbol{X}}\right)\leq\chi^{2}_{2(.05)} \tag{3.3.10}$$

3.3.4 Example for the Quality Control Ellipse

A clinician wants to test the two different quality of dosage times. A random sample of 12 diverticulosis patients of the age 21- 45 are selected from a case control study and both the dosages are given to them in the two different time periods, the dosage times of each stage are recoded through the patients alimentary canal and it is given in table (3.3.1).

Table 3.3.1. A case control study

	No of Patients	1	2	3	4	5	6	7	8	9	10	11	12
Dosages times (in hours)	Dosage A	63	54	79	68	87	84	92	57	66	53	76	63
	Dosage B	55	62	134	77	83	78	79	94	69	66	72	77

The XLSTAT command from Excel gives the following statistics output of the two different dosage times.

Table 3.3.2. Mean and variance of the dosages time

Sum	842.000	946.000
Mean	70.167	78.833
Variance	174.333	405.242

Here *p* represents total number of dosages and *n* the total number of patients, i.e. *p*=2 and *n*=12

Sample mean vector = $\bar{\boldsymbol{X}} = \left[\bar{x}_A, \bar{x}_B\right] = \left[70.167, 78.833\right]$

Sample covariance matrix = $\boldsymbol{S} = \begin{bmatrix} s_{aa} & s_{ab} \\ s_{ba} & s_{bb} \end{bmatrix} = \begin{bmatrix} 174.333 & 93.757 \\ 93.757 & 405.242 \end{bmatrix}$.

At 95 % confidence quality control ellipse for the dosages data can be obtained via equation 3.3.10 and all the pair of observations must satisfy the condition given in this equation.The critical chi square value at 0.05 significance level is $\chi^2_{2(.05)} = 5.991$.

Then substituting 5.991 into equation 3.3.10 we have $\left(\mathbf{X} - \bar{\boldsymbol{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X} - \bar{\boldsymbol{X}}\right) \leq 5.991$.

Now to check if the dosages time of the patients is under control, all the pairs of observations must fall inside the ellipse. Suppose to see the dosage times $P = \left(63, 55\right)$ of the patient 1 is in the control area or out of control it is necessary to simplify the following equation.

$$\frac{s_{AA}s_{BB}}{s_{AA}s_{BB} - s_{AB}} \left(\frac{\left(x_A - \bar{x}_A\right)^2}{s_{AA}} - 2s_{AB} \frac{\left(x_A - \bar{x}_A\right)\left(x_B - \bar{x}_B\right)}{s_{AA}s_{BB}} + \frac{\left(x_B - \bar{x}_B\right)^2}{s_{BB}} \right) \leq 5.991 \qquad (3.3.12)$$

$$\frac{174.333 \times 405.242}{174.333 \times 405.242 - 93.757} \left(\frac{(63 - 70.167)^2}{174.333} - 2(93.757) \frac{(63 - 70.167)(55 - 78.833)}{174.333 \times 405.242} + \frac{(55 - 78.833)^2}{405.242} \right) \leq 5.991$$

It is straight forward the dosage times of patient No.1 is in the control and still stable and no problem during dosages given to him or her with 5% level of significance. Data pairs are shown in Figure 3.3.4, and data pair No. 1 is well within the limits of the control ellipse.

Graphically,

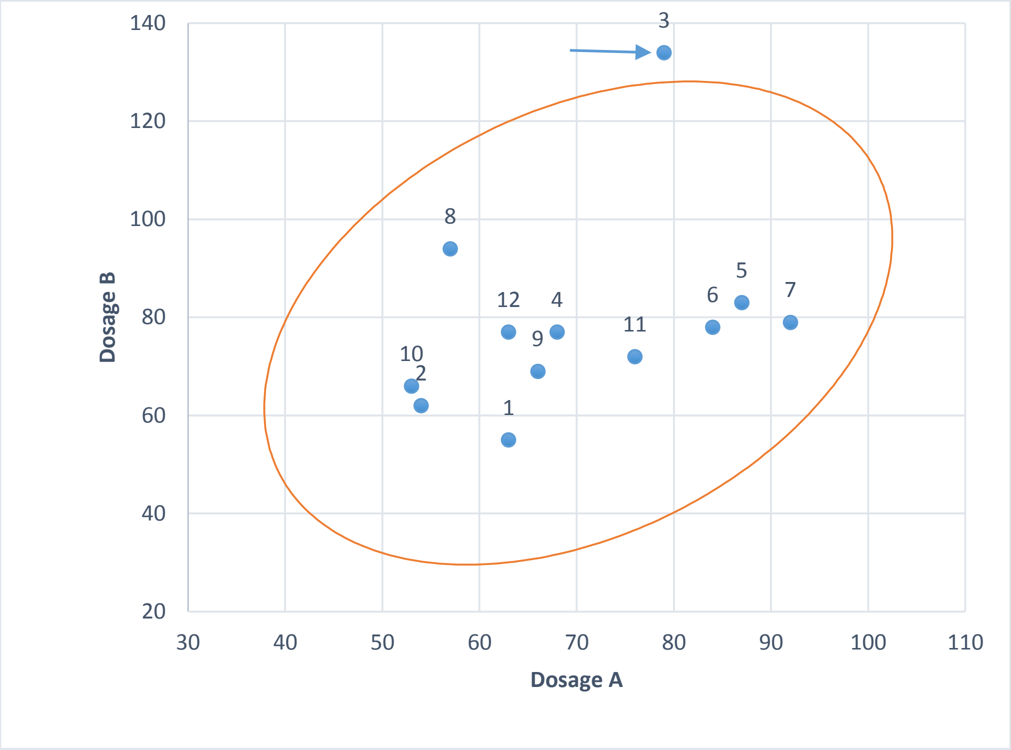


Figure 3.3.4. 95% quality control ellipse for dosages time

The dosage B for patient 3 is statistically out of control with 5% level of significance and it falls outside the control ellipse. That means this point does not satisfy the statistical distance equation from the mean origin. Because it may not contain the actual ingredients given to the patient, or the timing of administration of the dose may not be the same as the other patients. Due to this reason, the effect of dosage B on patient 3 was incorrectly observed in the study. Therefore, the clinician should be aware before investigating or changing the quality of dosages in the future.

Chapter 4

RELATIONSHIP BETWEEN PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS

In this chapter the theoretical concepts will be introduced to understand the fundamental relation between the PCA and FA. In factor analysis, the PCA approach will be used to reduce the dimension of the data. PCA also helps to determine the initial factor loadings and the score coefficients of the FA model. Before discussing the relation it is necessary to understand some basic concepts behind PCA and FA.

Considers the list of the steps involved in the construction of FA model using PCA approach.

1. Compute the covariance $\mathbf{\Sigma}$ or correlation $\boldsymbol{\rho}$ matrices.
2. Calculate eigenvalues and eigenvectors of $\mathbf{\Sigma}$ or $\boldsymbol{\rho}$ matrices.
3. Draw scree plot and determine the number of factors to be used in the model.
4. Calculate the factor loadings matrix using PCA method.
5. Find communalities and specific variances from factor loadings matrix.
6. Rotate the factor loadings matrix for example using varimax rotation technique to interpret the factor loadings easily.
7. Estimate the factor scores using ordinary least square regression.
8. Detect outliers and group the variables by few factors.
9. Interpret the factor scores using statistical control ellipse chart.

4.1 Principal Component Analysis

PCA reduces the high dimensional data into lower dimensional data. In factor analysis, PCA helps to reduce number of factors. Similarly it is also used as a dimension reduction technique in many other multivariate statistical analyses.

4.1.1 Principal Components

Principal components are obtained by linear transformation of the original variables. In the linear transformation process either the covariance or correlation matrices obtained from raw data can be used.

Definition 4.1 Let $X_1, X_2, ..., X_p$ be a set of *p* random variables consisting of *n* observations with covariance matrix **Σ**, then the new set of uncorrelated variables called principal components $Y_1, Y_2, ..., Y_p$ can be expressed as the linear combinations of the original *p* variables [18].

4.1.2 Geometrical Interpretation of PCA

Definition 4.1 Let $\boldsymbol{X} = \left[X_1, X_2, ..., X_p\right]$ be a random vector consisting of *n* observations drawn from a multivariate normal population with a mean vector $\boldsymbol{\mu} = \left[\mu_1, \mu_2, ..., \mu_p\right]$ and covariance matrix **Σ**. It is possible to plot the *n* observations of the multivariate normal data in a $n \times p$ coordinate system. Then the rotated coordinate system of the data, gives a hyper ellipsoid, whose axes are similar to those computed from the Eigen vectors of the covariance matrix **Σ**. Let us consider a constant statistical distance from $\boldsymbol{X} = \left[X_1, X_2, ..., X_p\right]$ to $\boldsymbol{\mu} = \left[0, 0, ..., 0_p\right]$ is defined by

$$\sqrt{\left(\boldsymbol{X} - 0\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X} - 0\right)} = c$$

Then the square statistical distance is

$$\Rightarrow \left(\boldsymbol{X}-0\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}-0\right)=c^2 \Rightarrow \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}=c^2$$

As $\boldsymbol{\Sigma} = \lambda_1\mathbf{e_1}\mathbf{e_1'}+\lambda_2\mathbf{e_2}\mathbf{e_2'}+,\ldots,+\lambda_p\mathbf{e_p}\mathbf{e_p'} \Rightarrow \boldsymbol{\Sigma}^{-1}=\frac{1}{\lambda_1}\boldsymbol{e}_1\boldsymbol{e}_1'+\frac{1}{\lambda_2}\boldsymbol{e}_2\boldsymbol{e}_2'+,\ldots,+\frac{1}{\lambda_p}\boldsymbol{e}_p\boldsymbol{e}_p'$

$$\Rightarrow \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}=\boldsymbol{X}'\frac{1}{\lambda_1}\mathbf{e_1}\mathbf{e_1'}\boldsymbol{X}+\boldsymbol{X}'\frac{1}{\lambda_2}\mathbf{e_2}\mathbf{e_2'}\boldsymbol{X}+\cdots+\boldsymbol{X}'\frac{1}{\lambda_p}\mathbf{e_p}\mathbf{e}_p'\boldsymbol{X}=c^2$$

$$\Rightarrow \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}=\frac{1}{\lambda_1}\left(\mathbf{e_1'X}\right)^2+\frac{1}{\lambda_2}\left(\mathbf{e_2'X}\right)^2+\cdots+\frac{1}{\lambda_p}\left(\mathbf{e_p'X}\right)^3=c^2$$

$$\Rightarrow \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}=\frac{Y_1^2}{\lambda_1}+\frac{Y_2^2}{\lambda_2}+\cdots+\frac{Y_p^2}{\lambda_p}=c^2 \qquad (4.1.1)$$

Thus the square constant statistical distance produces an ellipsoid with axes $Y_1=\mathbf{e_1'X},\ Y_2=\mathbf{e_2'X},\ \ldots,\ Y_p=\mathbf{e_p'X}$, where these axises are actually principal components. Hence semi minor and semi major axes measured by $c\sqrt{\lambda_i}$ in the direction of eigenvector e_i [18].

Geometrically,

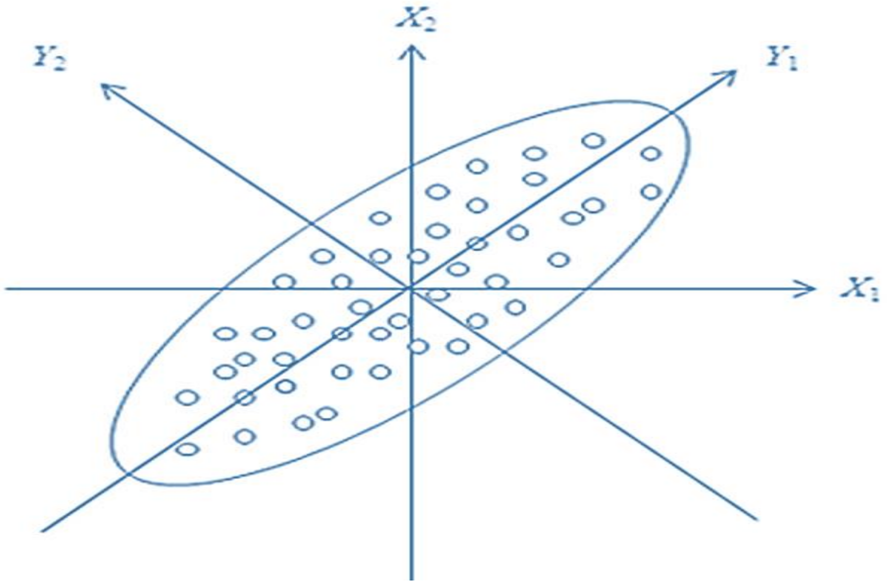


Figure 4.1.1. Graph of PCs Y_1, Y_2 orthogonal to the original coordinate system X_1, X_2

It is clear from the graph that the new Y_1, Y_2 axes passing through the center of the ellipse are obtained by orthogonal rotation of the original coordinate system.

Theorem 4.1.1 Consider the eigenvalue - eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_p, \boldsymbol{e}_p)$ computed from the covariance matrix $\mathbf{\Sigma}$ obtained from the $n \times p$ data matrix, where $\lambda_1 \geq \lambda_2 \geq \geq \lambda_p \geq 0$, and let $Y_1, Y_2, \ldots, Y_p$ be the principal components. Then $Y_1, Y_2, \ldots, Y_p$ are computed as given below.

$$\begin{array}{ccccccccccc} Y_1 & = & \boldsymbol{e}_1'\boldsymbol{X} & = & e_{11}X_1 & + & e_{12}X_2 & + & \cdots & + & e_{1p}X_p \\ Y_2 & = & \boldsymbol{e}_2'\boldsymbol{X} & = & e_{21}X_1 & + & e_{22}X_2 & + & \cdots & + & e_{2p}X_p \\ & \vdots & & & & & & \vdots & & + & \\ Y_3 & = & \boldsymbol{e}_p'\boldsymbol{X} & = & e_{p1}X_1 & + & e_{p2}X_2 & + & \cdots & + & e_{pp}X_p \end{array} \tag{4.1.2}$$

Then $tr(\mathbf{\Sigma}) = \sigma_{11} + \sigma_{22} +, \ldots, + \sigma_{pp}$

where $\sigma_{11} + \sigma_{22} +, \ldots, + \sigma_{pp} = \sum_{i=1}^{p} Cov(X_i, X_i) = \lambda_1 + \lambda_2 +, \ldots, + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$

Proof. By definition the trace of covariance matrix $\mathbf{\Sigma}$ is equal to the sum of it diagonal entries that is

$$tr(\mathbf{\Sigma}) = \sigma_{11} + \sigma_{22} +, \ldots, + \sigma_{pp}. \tag{4.1.3}$$

If $\boldsymbol{P} = [\mathbf{e_1}, \mathbf{e_2} \ldots, \mathbf{e_p}]$ is the matrix containing the eigenvectors of $\mathbf{\Sigma}$ such that $\mathbf{PP' = I}$ and $\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$ is a diagonal eigenvalues matrix, then by definition

$\mathbf{\Sigma} = \mathbf{PDP'}$.

This implies that $tr(\mathbf{\Sigma}) = tr(\mathbf{PDP'}) = tr(\mathbf{D}) = \lambda_1 + \lambda_2 +, \ldots, + \lambda_p$ and

$\sigma_{11} + \sigma_{22} +, \ldots, + \sigma_{pp} = \lambda_1 + \lambda_2 +, \ldots, + \lambda_p$.

Hence proved [18].

4.1.3 PCA for Components Reduction

Each eigenvalue $\lambda_i;\ i = 1,\ldots,p$ represents a certain percentage of total variation in the PCs obtained from the multivariate process under study and is given by $\frac{\hat{\lambda}_\imath}{\hat{\lambda}_1 + \hat{\lambda}_2 +,\ldots,+\hat{\lambda}_p} \times 100$.

It must be pointed out that $Var(Y_i) = \lambda_i$ and $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i)$. Then

$$\tau = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{i=1}^{p} \lambda_i};\ 0 < \tau \leq 1;\ 1 \leq m \leq p \qquad (4.1.4)$$

can be used as a measure to determine the number of PCs to be used. Depending on the nature of the process under study, it is desirable to have τ high to very high. For most applications a value $\tau > 0.8$ is desirabe.

4.1.4 PCA for Variable Reduction

In principal component analysis one of the major issues is to interpret principal components. Sometimes it is difficult to judge high contributed explanatory variables in the component models. The following correlation is used to determine the correlation coefficient between a variable and a principal component.

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\ldots,p \qquad (4.1.5)$$

while the correlation is a measure of the level of relationship between a variable and the PC, the coefficients of the PC measures the contribution of each variable to the PC. Therefore, the two measures should not be compared with each other, but rather be used together for a better interpretation of the individual PC [18].

Theorem 4.1.2 Let $Y_1, Y_2, \ldots, Y_p$ be the set of unobserved random variables (in this case PCs) computed from a population.

Then $\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda}}{\sqrt{\sigma_{kk}}}$ is the correlation coefficient that measures the linear relationship between i^{th} PC and k^{th} variable, where

$$Cov\left(X_k, Y_i\right) = Cov\left(a_k' X, e_i X\right) = a_k' \Sigma e_i ; i,k = 1,2,\ldots,p$$

Proof: Let $a_k' = \left(0,\ldots,0,1,0,\cdots,0\right)$ be the coefficient vector of matrix **X** such that $X_k = a_k' X$ and let $Y_i = e_i X$ be the PCs represented by an equation $a_k' \boldsymbol{\Sigma} e_i = \lambda_i e_i$

By definition

$$Cov\left(X_k Y_i\right) = Cov\left(a_k' X, e_i X\right) = a_k' \boldsymbol{\Sigma} e_i \qquad \text{As } a_k' \boldsymbol{\Sigma} e_i = \lambda_i e_i \tag{4.1.6}$$

$\Rightarrow Cov\left(X_k Y_i\right) = \lambda_i e_i$. Then $V \quad \left(a \quad _k\right) r \quad \sigma$ and $Var\left(Y_i\right) = \lambda_i$ gives

$$Corr\left(X_k Y_i\right) = \frac{Cov\left(X_k Y_i\right)}{\sqrt{Var\left(X_k\right)}\sqrt{Var\left(Y_i\right)}} = \frac{\lambda_i e_{ik}}{\sqrt{\sigma_{kk}}\sqrt{\lambda_i}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} = \rho_{Y_i,X_k} \quad \text{for} \quad i,k = 1,2,\ldots,p$$

Hence proved [18].

4.1.5 Covariance verses Correlation Matrix

When the variables involved in a process have different units or the variations in the data values of some variables are considerably large, it leads to unreliable results in the computation of the principal components and gives ambiguous interpretation of the principal coefficients. To avoid these problems it is necessary to first standardize the data and then compute the principal components using correlation matrix not covariance matrix [18].

4.1.6 Standardized Principal Components

Definition 4.2 Suppose $\boldsymbol{X} = \left[X_1, X_2, \ldots, X_p\right]$ is a random vector consisting of p variables drawn from a multivariate population with mean vector $\boldsymbol{\mu} = \left[\mu_1, \mu_2, \ldots, \mu_p\right]$ and standard deviation matrix is $\mathbf{V}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$. Then the new vector $\mathbf{Z} = \left[Z_1, Z_2, \ldots, Z_p\right]$ with $Z_i = \dfrac{X_i - u_i}{\sqrt{\sigma_{ii}}}$ is called the standard normal vector generated by **X**, and the relation between **Z** and **X** can be expressed as given by

$$\mathbf{Z} = \left(\mathbf{V}^{\frac{1}{2}}\right)^{-1}\left(\mathbf{X} - \boldsymbol{\mu}\right) \tag{4.1.7}$$

The expectation of **Z** is zero. That is

$$E(\mathbf{Z}) = E\left\{\left(\mathbf{V}^{\frac{1}{2}}\right)^{-1}\left(\mathbf{X} - \boldsymbol{\mu}\right)\right\}$$

$$\Rightarrow E\left(\boldsymbol{Z}\right) = E\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1} \times E\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\right]$$

$$\Rightarrow E\left(\boldsymbol{Z}\right) = E\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1} \times \left[E\left(\boldsymbol{X}\right) - E\left(\boldsymbol{\mu}\right)\right]$$

$$\Rightarrow E\left(\boldsymbol{Z}\right) = E\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1} \times \left[\boldsymbol{\mu} - \boldsymbol{\mu}\right] \therefore E\left(\boldsymbol{X}\right) = \boldsymbol{\mu}$$

$$\Rightarrow E\left(\boldsymbol{Z}\right) = E\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1} \times 0 \Rightarrow E\left(\boldsymbol{Z}\right) = \mathbf{0}$$

Hence Proof completed.

Also $$Cov\left(\boldsymbol{Z}\right)=\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1} Cov\left(\boldsymbol{X} \text{ - } \boldsymbol{\mu}\ \right)\left(\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1}\right)' \tag{4.1.8}$$

$$\Rightarrow Cov\left(\boldsymbol{Z}\right)=\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1}\boldsymbol{\Sigma}\left(\left(V^{\frac{1}{2}}\right)^{-1}\right)' = \boldsymbol{\rho}$$

Thus the standardize principal components can also be derived from the correlation matrix $\boldsymbol{\rho}$. See the Theorem 4.1.3 below [18].

Theorem 4.1.3 Let $\boldsymbol{Z}=\left[Z_1, Z_2, ..., Z_p\right]$ be a standard normal vector and $\left(\lambda_1,\mathbf{e}_1\right),\left(\lambda_2,\mathbf{e}_2\right),...,\left(\lambda_p,\mathbf{e}_p\right)$ be pairs of eigenvalues and eigenvectors where $\lambda_1 \geq \lambda_2 \geq....\ \lambda_p \geq 0$ with correlation matrix $Cov\left(\mathbf{Z}\right)=\boldsymbol{\rho}$. Then uncorrelated variables $Y_1, Y_2, ..., Y_p$ can be computed by

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}'\left(\mathbf{V}^{\frac{1}{2}}\right)^{-1}\left(\mathbf{X}\text{ - }\boldsymbol{\mu}\right) \qquad i=1,2,...,p \tag{4.1.9}$$

In this case, each standard normal variable have unit variance and the sum of the variances are equal to the number of variables *p*. That is

$$Var\left(Z_i\right)=\sigma_{ii}=1 \text{. Then } \sum_{i=1}^{p}Var\left(Y_i\right)=\sum_{i=1}^{p}Var\left(Z_i\right)=p \quad \text{for all } i\text{=1, 2, …, } p \tag{4.1.10}$$

Similarly, correlation between k^{th} standard variate Z_k and i^{th} principal component Y_i is defined as

$$Corr\left(Z_kY_i\right)=\frac{Cov\left(Z_kY_i\right)}{\sqrt{Var\left(Z_k\right)}\sqrt{Var\left(Y_i\right)}}=\frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}=e_{ik}\sqrt{\lambda_i}=\rho_{Y_i,Z_k} \quad '' i,k=1,...,p \tag{4.1.11}$$

Consequently, the equation 4.1.11 can be used in determining the number of PCs to be used in representing the process in a lower dimensional space.Since the variance of standardized data is always 1 and forms the diagonal elements of the correlation matrix, then total variance is the same as the number of variables p.

4.2 Factor analysis

Factor analysis is a data classification technique used to group the large number of variables into set of few unobserved variables called factors. The purpose of the factor analysis is to construct a system of equations accommodating the underlying factors in order to capture the maximum information from the data set.

4.2.1 Independent Factor Model

Definition 4.3 Let $\boldsymbol{X} = \left[X_1, X_2, ..., X_p \right]$ be a random vector containing *p* random variables of size *n* that follows a multivariate normal distribution with mean vector $\boldsymbol{\mu} = \left[\mu_1, \mu_2, ..., \mu_p \right]$ and population covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}.$$

Assuming that is $\boldsymbol{X}$ is correlated with $\boldsymbol{F} = \left[F_1, F_2, ..., F_p \right]$ called unobserved factors and $\boldsymbol{\varepsilon} = \left[\varepsilon_1, \varepsilon_2, ..., \varepsilon_p \right]$ called disturbance terms or specific factors, then the *p* deviations model can be expressed as linear combinations of unobserved factors plus error terms and is given as follows,

$$\begin{array}{lllllllllll} X_1 - \mu_1 & = & l_{11}F_1 & + & l_{12}F_2 & + & \cdots & + & l_{1m}F_m & + & \varepsilon_1 \\ X_2 - \mu_2 & = & l_{21}F_1 & + & l_{22}F_2 & + & \cdots & + & l_{1m}F_m & + & \varepsilon_2 \\ \vdots & & & & & & \vdots & & & & \\ X_p - \mu_p & = & l_{p1}F_1 & + & l_{p2}F_2 & + & \cdots & + & l_{1m}F_m & + & \varepsilon_p \end{array} \qquad (4.2.1)$$

In general,

$$X_i - \mu_i = \sum_{j=1}^{m} l_{ij} F_j + \varepsilon_i, \qquad i = 1,2,..,p \qquad (4.2.2)$$

This is called factor analysis model, where l_{ij} is the loading of the i^{th} variable on the j^{th} factor. In other words l_{ij} is the measure of factor loading of the i^{th} variable contribution, on the j^{th} factor [18].

The orthogonal factor model can be expressed in the matrix form as

$$\mathbf{X}_{(p\times1)} - \boldsymbol{\mu}_{(p\times1)} = \mathbf{L}_{(p\times m)}\mathbf{F}_{(m\times1)} + \boldsymbol{\varepsilon}_{(p\times1)} \qquad (4.2.3)$$

where F and $\boldsymbol{\varepsilon}_i$ are unobserved random vectors satisfying the following assumptions.

1. $E(\mathbf{F}) = \mathbf{0}_{m\times1}$, $Var(\boldsymbol{F}) = E(\boldsymbol{FF}') - [E(\boldsymbol{F})]^2 = \boldsymbol{I}_{(\text{m}\times m)}$

Hence $E(\boldsymbol{\varepsilon}) = 0_{p\times1}$ and $Var(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon\varepsilon}') - [E(\boldsymbol{\varepsilon})]^2 = \boldsymbol{\Psi}_{p\times p} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$

2. $Cov(\mathbf{F},\boldsymbol{\varepsilon}) = \mathbf{0}_{m\times p}$, hence F and $\boldsymbol{\varepsilon}$ are independent.

Also $Cov(\boldsymbol{X},\boldsymbol{F}) = \boldsymbol{L}$.

As $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\,\boldsymbol{F} + \boldsymbol{\varepsilon}$

Multiplying of the factor model by $\mathbf{F}'$, then it becomes

$$(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{F}' = (\boldsymbol{L}\,\boldsymbol{F} + \boldsymbol{\varepsilon})\boldsymbol{F}' = \boldsymbol{LFF}' + \boldsymbol{\varepsilon}$$

By taking expectation it is becomes as

$$\begin{aligned} Cov(\boldsymbol{X},\boldsymbol{F}) = E(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{F}' &= E(\boldsymbol{LF}+\varepsilon)\boldsymbol{F}' = E[\boldsymbol{LFF}' + \varepsilon\boldsymbol{F}'] \\ &= \boldsymbol{L}E[\boldsymbol{FF}'] + E[\varepsilon\boldsymbol{F}'] \therefore \boldsymbol{FF}' = \boldsymbol{I} \\ &= (\boldsymbol{L}\times\boldsymbol{I}) + 0 \\ &= \boldsymbol{L} \end{aligned}$$

Hence the proved.

Remark: $\boldsymbol{L}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix}$ is called factor loading matrix, and its elements are the same as the elements of the covariance between i^{th} variable and j^{th} factor i.e. $Cov\left(X_i, F_j\right)=l_{ij}$.

4.2.2 Standardized Orthogonal Factor Model

Let $Z_1, Z_2, \ldots, Z_p$ be the standardized variables and $\boldsymbol{\rho}$ be the population correlation matrix that can be expressed as

$$\boldsymbol{\rho}=\mathbf{LL}'+\boldsymbol{\psi} \tag{4.2.4}$$

where

$$\boldsymbol{\rho}=\begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{np} \end{bmatrix},\ \boldsymbol{L}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix} \text{ and } \boldsymbol{\Psi}_{p\times p}=\begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

Then the *m* common factor model can be written as follows

$$\begin{array}{lllllllllll} Z_1 & = & l_{11}F_1 & + & l_{12}F_2 & + & \cdots & + & l_{1m}F_m & + & \varepsilon_1 \\ Z_2 & = & l_{21}F_1 & + & l_{22}F_2 & + & \cdots & + & l_{1m}F_m & + & \varepsilon_2 \\ \vdots & & & & & & \vdots & & & & \\ Z_p & = & l_{p1}F_1 & + & l_{p2}F_2 & + & \cdots & + & l_{1m}F_m & + & \varepsilon_p \end{array} \tag{4.2.5}$$

System 4.2.5 is called Standardized Orthogonal Factor Model.

Where $l_{ij}=Corr\left(X_i, F_j\right)=\rho_{X_i, F_j}=\sqrt{\lambda_j}e_{ij}$, $Var\left(F_j\right)=1$ and $Corr\left(\varepsilon_i, F_j\right)=0$

4.2.3 Orthogonal Model for Covariance Matrix

Consider the variance-covariance matrix of **X** under the orthogonal factor model [17]. By definition the orthogonal factor model can be written as

$$
\begin{aligned}
&(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = (\boldsymbol{LF} + \boldsymbol{\varepsilon})(\boldsymbol{LF} + \boldsymbol{\varepsilon})' \\
&= (\boldsymbol{LF} + \boldsymbol{\varepsilon})(\boldsymbol{LF} + \boldsymbol{\varepsilon})' \\
&= (\boldsymbol{LF} + \boldsymbol{\varepsilon})((\boldsymbol{LF}') + \boldsymbol{\varepsilon}') \\
&= \boldsymbol{LF}(\boldsymbol{LF})' + \boldsymbol{\varepsilon}(\boldsymbol{LF})' + \boldsymbol{LF\varepsilon}' + \boldsymbol{\varepsilon\varepsilon}'
\end{aligned}
$$

By taking expectation we obtain

$$
\begin{aligned}
&\boldsymbol{\Sigma} = Cov(\boldsymbol{X}) = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' \\
&= \boldsymbol{L}E(\boldsymbol{FF}')\boldsymbol{L} + E(\boldsymbol{\varepsilon F}')\boldsymbol{L}' + \boldsymbol{L}E(\boldsymbol{F\varepsilon}') + E(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}') \qquad (4.2.6) \\
&= \boldsymbol{LL}' + \boldsymbol{\Psi}
\end{aligned}
$$

This gives the covariance structure of **X** for common factors. Diagonal entries of $\boldsymbol{\Sigma}$ can be decomposed as

$$Cov(X_i, X_i) = Var(X_i) = l^2_{i1} + l^2_{i2} +,...,+ l^2_{im} + \psi_i \qquad (4.2.7)$$

$$
\begin{aligned}
&\Rightarrow Var(X_i) = l^2_{i1} + l^2_{i2} +,...,+ l^2_{im} + \psi_i \\
&\Rightarrow Var(X_i) = commuality + uniqueness
\end{aligned}
.
$$

Off diagonal entries of $\boldsymbol{\Sigma}$ can be calculated by

$$Cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \cdots + l_{\mathrm{im}}l_{km} \qquad (4.2.8)$$

4.2.4 Communality and Specific Variance

In case of orthogonal factor model, the $Var(X_i)$ can be split into two parts. First part consists of the sum of square loadings, called *communality* denoted by h^2_i for the i^{th} variable. Communality measure the percentage of the total variation of ***X*** explained by common factors, whereas the last part is symbolized by ψ_i, represents the percentage of variability explained due to some other factors. The variance of error term $Var(\varepsilon_i) = \psi_i$ is called specific variance or uniqueness [18].

4.2.5 Theoretical Relationship between PCA and FA

In sections two types of factor models will be disused. One is called exact factor model and the other is called inexact factor model. The exact model has no error term, for this reason, the exact model is not a suitable model to explore the data. However, the PCA approach will be used to investigate the unknown population parameters of such models.

4.2.6 Exact or Non-Stochastic Factor Model

Let $\left(\lambda_i, \boldsymbol{e}_i\right)$ be the eigenvalue - eigenvector pairs of the covariance matrix $\boldsymbol{\Sigma}$ with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and p=m. Then the covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as

$$\boldsymbol{\Sigma} = \mathbf{PDP}' = \left[\boldsymbol{e}_1 \vdots \boldsymbol{e}_2 \vdots \ldots \vdots \boldsymbol{e}_p\right] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_1' \\ \boldsymbol{e}_2' \\ \vdots \\ \boldsymbol{e}_p' \end{bmatrix}$$

or

$$\boldsymbol{\Sigma} = \lambda_1 \mathbf{e_1 e_1'} + \lambda_2 \mathbf{e_2 e_2'} +, \ldots, + \lambda_p \mathbf{e_p e_p'}$$

$$\boldsymbol{\Sigma} = \left[\sqrt{\lambda_1}\boldsymbol{e_1} \;\vdots\; \sqrt{\lambda_2}\boldsymbol{e}_2 \;\vdots\; \cdots \;\vdots\; \sqrt{\lambda_p}\boldsymbol{e}_p\right] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1' \\ \cdots \\ \sqrt{\lambda_2}\boldsymbol{e}_2' \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_p}\boldsymbol{e}_p' \end{bmatrix}$$

This implies that

$$\boldsymbol{\Sigma}_{p\times p} = \boldsymbol{L}_{p\times p}\boldsymbol{L}'_{p\times p} + 0_{p\times p} = \boldsymbol{L}_{p\times p}\boldsymbol{L}'_{p\times p} \qquad (4.2.9)$$

This provides the covariance structure of **X** in case where the number of common factors are the same as the number of variable $m=p$ and it gives $Var\left(\varepsilon_i\right)=\psi_i=0$ for orthogonal factor model. For this reason it is not a useful method to analyze data with using factor analysis. The value $\sqrt{\lambda_j}\boldsymbol{e}_j$ represents the factor loading of the *jth* column of the loading matrix, without the scale value $\sqrt{\lambda_j}$ factor loading is actually principal component coefficient denoted by $\boldsymbol{e}_j$ [18].

4.2.7 Inexact or Stochastic Factor Model

This approach will be useful when the eigenvalues with not significant contribution to the total variance $\lambda_{m+1},\ldots,\lambda_p$ are eliminated from the following matrix equation

$$\boldsymbol{\Sigma} \ =\lambda_1\boldsymbol{e}_1\boldsymbol{e}_1'+\lambda_2\boldsymbol{e}_2\boldsymbol{e}_2'+,\ldots,+\lambda_m\boldsymbol{e}_m\boldsymbol{e}_m'+\lambda_{m+1}\boldsymbol{e}_{m+1}\boldsymbol{e}_{m+1}'+\lambda_{m+2}\boldsymbol{e}_{m+2}\boldsymbol{e}_{m+2}'+,\ldots,+\lambda_p\boldsymbol{e}_p\boldsymbol{e}_p'.$$

After the exclusion of the terms $\lambda_{m+1}\boldsymbol{e}_{m+1}\boldsymbol{e}_{m+1}'+\lambda_{m+2}\boldsymbol{e}_{m+2}\boldsymbol{e}_{m+2}'+,\ldots,+\lambda_p\boldsymbol{e}_p\boldsymbol{e}_p'$ from the above expression the approximate covariance matrix of **X** can be expressed as

$$\boldsymbol{\Sigma} \ =\lambda_1\boldsymbol{e}_1\boldsymbol{e}_1'+\lambda_2\boldsymbol{e}_2\boldsymbol{e}_2'+,\ldots,+\lambda_m\boldsymbol{e}_m\boldsymbol{e}_m'$$

$$\boldsymbol{\Sigma}=\left[\sqrt{\lambda_1}\boldsymbol{e}_1 \ \vdots \ \sqrt{\lambda_2}\boldsymbol{e}_2 \ \vdots \ \cdots \ \vdots \ \sqrt{\lambda_m}\boldsymbol{e}_m\right]\begin{bmatrix}\sqrt{\lambda_m}\boldsymbol{e}_m' \\ \cdots \\ \sqrt{\lambda_m}\boldsymbol{e}_m' \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_m}\boldsymbol{e}_m'\end{bmatrix}=\boldsymbol{L}_{p\times m}\boldsymbol{L}_{m\times p}'$$

or

$$\boldsymbol{\Sigma}-\boldsymbol{L}_{p\times m}\boldsymbol{L}_{m\times p}' =\boldsymbol{\Psi}$$

$$\mathbf{\Sigma} = \left[\sqrt{\lambda_1}e_1 \;\vdots\; \sqrt{\lambda_2}e_2 \;\vdots\; \cdots \;\vdots\; \sqrt{\lambda_m}e_m\right]\begin{bmatrix} \sqrt{\lambda_m}e'_m \\ \cdots \\ \sqrt{\lambda_m}e'_m \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_m}e'_m \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

and finally,

$$\mathbf{\Sigma} = \boldsymbol{L}_{p\times m}\boldsymbol{L}'_{m\times p} + \mathbf{\Psi} \qquad (4.2.10)$$

where $\mathbf{\Psi}$ is the diagonal matrix whose diagonal entries are specific variances. That is denoted by $Var(\varepsilon_i) = \psi_i$ [18].

This procedure of splitting the covariance matrix of **X** into factor loading matrix plus specific variance matrix is known as principal component approach for factor analysis model.

4.2.8 Factor Analysis Model

Applying the procedure given under section **4.2.7** to a particular data $\boldsymbol{X} = \left[X_1, X_2, ..., X_p\right]$ each variable consisting of the observations $x_1, x_2, ..., x_n$, it is necessary to first transform the data matrix to the deviation matrix. That is,

$$\mathbf{x}_j - \boldsymbol{\mu} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} x_{j1} - \mu_1 \\ x_{j2} - \mu_2 \\ \vdots \\ x_{jp} - \mu_p \end{bmatrix} \text{ for } j=1,2\ ,...,n \qquad (\ 4.2.11)$$

This is sometime called mean corrected data matrix, each observation of this matrix centered by their corresponding population mean. Then the population principal components $Y_1, Y_2, \ldots, Y_p$ can be computed as follows,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \mathbf{e}'\left(\boldsymbol{X} - \boldsymbol{\mu}\right) = \sum_{j=1}^{m} \mathbf{e}_j Y_j + \sum_{j=m+1}^{p} \mathbf{e}_j Y_j = \sum_{j=1}^{m} \sqrt{\lambda_j}\mathbf{e}_j\, Y_j \Big/ \sqrt{\lambda_j} + \sum_{j=m+1}^{p} \mathbf{e}_j Y_j \qquad (4.2.13)$$

$$= \left[\sqrt{\lambda_1}\mathbf{e}_1 \quad \sqrt{\lambda_2}\mathbf{e}_2 \quad \cdots \quad \sqrt{\lambda_m}\mathbf{e}_m\right] \begin{bmatrix} Y_1 \Big/ \sqrt{\lambda_1} \\ \vdots \\ Y_m \Big/ \sqrt{\lambda_m} \end{bmatrix} + \sum_{j=m+1}^{p} \mathbf{e}_j Y_j = \boldsymbol{LF} + \boldsymbol{\varepsilon}$$

Then this yields

$$\boldsymbol{X} \text{ - } \boldsymbol{\mu} = \mathbf{e}Y = \left[\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\right] \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} \qquad (4.2.14)$$

This is called factor analysis model derived by the principal component analysis approach.

$\boldsymbol{L} = \left[\sqrt{\lambda_1}\mathbf{e}_1 \quad \sqrt{\lambda_2}\mathbf{e}_2 \quad \cdots \quad \sqrt{\lambda_m}\mathbf{e}_m\right]$ is called factor loadings matrix.

$\boldsymbol{F} = \begin{bmatrix} Y_1 \Big/ \sqrt{\lambda_1} \\ \vdots \\ Y_m \Big/ \sqrt{\lambda_m} \end{bmatrix}$ represents population common factors generated from the first PCs scaled by the square root of eigenvalues and $\boldsymbol{\varepsilon} = \sum_{j=m+1}^{p} \mathbf{e}_j Y_j$ is called error term generated by the last principal components, whose variances are smaller eigenvalues. This derivation of the model helps to determine the perfect solutions for factor analysis model [18].

Remark 1: $\boldsymbol{\varepsilon} = \sum_{j=m+1}^{p} \mathbf{e}_j Y_j$ is represents all factors having low eigenvalues.

Remark 2: The covariance matrix of **X** computed form original observations or deviation data remains unchanged.

4.2.9 Estimators of Factor Model

Let $\boldsymbol{X} = \left[X_1, X_2,..., X_p\right]$ be the collection of p samples each consisting of the following observations $x_1, x_2, \ldots, x_n$, with sample covariance matrix of the form

$$\mathbf{S} = \hat{\mathbf{P}}\hat{\mathbf{D}}\hat{\mathbf{P}}' = \left[\hat{\mathbf{e}}_1 \vdots \hat{\mathbf{e}}_2 \vdots \ldots \vdots \hat{\mathbf{e}}_p\right] \begin{bmatrix} \hat{\lambda}_1 & 0 & \cdots & 0 \\ 0 & \hat{\lambda}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\lambda}_p \end{bmatrix} \begin{bmatrix} \hat{\mathbf{e}}'_1 \\ \hat{\mathbf{e}}'_2 \\ \vdots \\ \hat{\mathbf{e}}'_p \end{bmatrix}, \text{where } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

Then by the principal component approach when $m < p$ the estimatied facor loading matrix is given by

$$\hat{\mathbf{L}} = \begin{bmatrix} \hat{l}_{11} & \hat{l}_{12} & \cdots & \hat{l}_{1m} \\ \hat{l}_{21} & \hat{l}_{22} & \cdots & \hat{l}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{l}_{p1} & \hat{l}_{p2} & \cdots & \hat{l}_{pm} \end{bmatrix} = \left[\hat{\mathbf{l}}_1, \hat{\mathbf{l}}_2, \cdots, \hat{\mathbf{l}}_m\right] \quad \text{, where } j=1,2,...,m$$

$$= \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 \quad \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 \quad \cdots \quad \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m\right]$$

The matrix $\mathbf{S} - \hat{\mathbf{L}}_{p\times m}\hat{\mathbf{L}}'_{m\times p}$ produce a diagonal matrix whose diagonal entries are specific variances estimated by $\hat{\Psi}$.

That is $\hat{\boldsymbol{\Psi}} = \begin{bmatrix} \hat{\psi}_1 & 0 & \cdots & 0 \\ 0 & \hat{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\psi}_p \end{bmatrix}$

where $\hat{\psi}_i = var\left(X\right) = s_{ii} - \sum_{j=1}^{n} \hat{l}_{ij}^2$

The portion of the total variation of i^{th} variable explained by m factors can be estimated by the following communality value

$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 +,..,+ \hat{l}_{im}^2$ for $i=1,2,...,p$

The number of factors in the factor analysis to be included will be judged based on the following statistic

$$\frac{\hat{\lambda}j}{s_{11}+s_{22}+,...,s_{pp}}=\frac{\hat{\lambda}j}{\hat{\lambda}_1+\hat{\lambda}_2+,...,+\hat{\lambda}_p} \quad ,$$

where $tr\left(\boldsymbol{S}\right)=s_{11}+s_{22}+,...,+s_{pp}$ and

$$\hat{\lambda}_j=\hat{l}_{1j}^2+\hat{l}_{2j}^2+,...,+\hat{l}_{p_j}^2=\left(\sqrt{\hat{\lambda}_j}\boldsymbol{e}'_j\right)'\left(\sqrt{\hat{\lambda}_j}\boldsymbol{e}'_j\right);\ j=1,...m \ ,$$

For standardize variables this can be defined as

$$\frac{\hat{\lambda}j}{\hat{\lambda}_1+\hat{\lambda}_2+\cdots+\hat{\lambda}_p}=\frac{\hat{\lambda}j}{1+1+\cdots+1}=\frac{\hat{\lambda}j}{p}$$

where p is the total number of standardized variables [18].

4.2.10 Factor Rotation

In factor analysis it is difficult to interpret the original factor loadings found by principal component analysis approach. In order to develop a simpler structure of the factor model it is essential to rotate the initial factor loadings. The rotation of factor loadings does not affect the original factor model; only rotate the original factor loadings such that the original factor axes are perpendicular to the new factor axes [18].

Let $\hat{\mathbf{L}}_{p\times m}$ be the estimated factor loadings matrix derived by principal component approach. Then

$$\hat{\boldsymbol{L}}^*_{p\times m}=\hat{\boldsymbol{L}}_{p\times m}\boldsymbol{T}_{m\times m} \tag{4.2.15}$$

where $\boldsymbol{T}$ is transformation matrix. Additionally the rotation of factor loading matrix does not change the covariance and correlation structure. That is

$$\hat{\boldsymbol{L}}_{p\times m}\hat{\boldsymbol{L}}'_{m\times p}+\hat{\boldsymbol{\Psi}}=\mathbf{L}_{p\times m}\boldsymbol{T}_{m\times m}\boldsymbol{T}'_{m\times m}\hat{\boldsymbol{L}}'_{p\times m}=\hat{\boldsymbol{L}}^*_{p\times m}\hat{\boldsymbol{L}}^{*\prime}_{m\times p}+\hat{\boldsymbol{\Psi}}$$

So this suggests that the residual matrix $\boldsymbol{S_n} - \hat{\boldsymbol{L}}_{\boldsymbol{p\times m}}\hat{\boldsymbol{L}}'_{\boldsymbol{m\times p}} + \hat{\boldsymbol{\Psi}} = \boldsymbol{S_n} - \hat{\boldsymbol{L}}^{*}_{\boldsymbol{p\times m}}\hat{\boldsymbol{L}}^{*\prime}_{\boldsymbol{m\times p}} + \hat{\boldsymbol{\Psi}}$ is not affected, and the specific variances and communalities are also unaffected due rotating the original factor loadings. Consequently the original factor model is exactly the same as rotated factor model that is

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\,\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{L}^{*}\,\boldsymbol{F}^{*} + \boldsymbol{\varepsilon} \tag{4.2.16}$$

Remark: Factor rotation helps to determine the appropriate number factor loadings in order to gain more clear idea about the structure of the factor model.

4.2.11 Varimax Rotation

Varimax is an orthogonal factor rotation technique developed by an American statistician Henry F. Kaiser. It helps to achieve the clear configuration of the factor loadings for uncorrelated factors. Let $\tilde{l}^{*}_{ij}$ be the rotated factor loadings whose values lies on the new rotated factor coordinate system defined by the following quantity

$$\tilde{l}^{*}_{ij} = \hat{l}^{*}_{ij} \big/ \hat{h}_i$$

where $\hat{l}^{*}_{ij}$ are the coefficients of the $\hat{\boldsymbol{L}}^{*}_{p\times m}$ rotated factor loadings matrix and $\hat{h}_i$ is the square root of the i^{th} communality.

Then the sum of variance of the $\left(\tilde{l}^{*}_{ij}\right)^{2}$ for j^{th} factor can be expressed as

$$V\left(\left(\tilde{l}^{*}_{ij}\right)^{2}\right) = \sum_{j=1}^{m} \left[\frac{1}{p} \sum_{i=1}^{n} \left(\tilde{l}^{*}_{ij}\right)^{4} - \left\{ \frac{1}{p} \sum_{i=1}^{n} \left(\tilde{l}^{*}_{ij}\right)^{2} \right\}^{2} \right] \tag{4.2.17}$$

In order to achieve the maximum values $\hat{l}^{*}_{ij}$ of the rotated factor loadings it is required to maximize the sum of variance of the square rotated factor loadings $\hat{l}^{*}_{ij}$ on the j^{th} factor [18].

4.2.12 Factor Score

In factor analysis the population parameters of the factor model are usually unknown. These parameters can be estimated by using ordinary least square technique by minimizing the total sum of square residuals of the sample factor model with respect to the estimated value of F_j, such that

$$\frac{\partial \sum_{j=1}^{m} \hat{\varepsilon}_j^2}{\partial \hat{f}_j} = 0$$

This equation yields the following estimator

$$\hat{f}_j = \left(\hat{\boldsymbol{L}}'\hat{\boldsymbol{L}}\right)^{-1} \hat{\boldsymbol{L}}'(x_j - \bar{x}) \qquad (4.2.18)$$

where, $\boldsymbol{L} = \left[\sqrt{\lambda_1}\mathbf{e}_1 \quad \sqrt{\lambda_2}\mathbf{e}_2 \quad \cdots \quad \sqrt{\lambda_m}\mathbf{e}_m\right]$ is the original factor loadings matrix established by PCA approach.

Then $$\hat{F}_{\boldsymbol{j}} = \begin{bmatrix} \frac{\hat{\mathbf{e}}_1'}{\sqrt{\hat{\lambda}_1}} \\ \vdots \\ \frac{\hat{\mathbf{e}}_{\mathbf{m}}'}{\sqrt{\hat{\lambda}_m}} \end{bmatrix} (x_j - \bar{x}) \Rightarrow \hat{F}_j = \frac{\hat{\mathbf{e}}_{\mathbf{j}}'}{\sqrt{\hat{\lambda}_j}} (x_j - \bar{x}) \qquad (4.2.19)$$

$\hat{F}_j$ is actually j^{th} sample principal component $\hat{y}_j = \hat{\mathbf{e}}_{\mathbf{j}}'(x_j - \bar{x})$ scaled by $\frac{1}{\sqrt{\hat{\lambda}_j}}$ That is,

$$\hat{F}_j = \frac{\hat{y}_j}{\sqrt{\hat{\lambda}_j}} = \frac{\hat{\mathbf{e}}_j'(x_j - \bar{x})}{\sqrt{\hat{\lambda}_j}}$$

Moreover, the standardize factor score of the data can be computed by the following estimator is

$$\hat{F}_j = \frac{\hat{e}_j'}{\sqrt{\hat{\lambda}_j}} \left(\frac{x_j - \bar{x}}{s_{ii}}\right) = \frac{\hat{e}_j'}{\sqrt{\hat{\lambda}_j}} z_j \quad \textit{for } j = 1,2,\ldots,m \;\&\; i = 1,2,\ldots,p \qquad (4.2.20)$$

Since the factor scores are useful to judge explanatory variables with high contribution in the factor model, it helps to detect outliers and also obtain a simple structure of the data [18].

Chapter 5

STATISICAL ANALYSIS OF THE WORLD ECOMONIC DATA

The data used in this study represents 12 different economic indicators by country from 185 different countries. Data was originally compiled by the United States, Heritage Foundation of Research and Educational Institute. Each economic indicator is represented by a variable as follows.

X_1 : Gross domestic product (GDP) in billion dollars.

X_2 : GDP/capita.

X_3 : Growth rate.

X_4 : Inflation rate.

X_5 : Interest rate.

X_6 : Income tax rate.

X_7 : Unemployment rate.

X_8 : Corporate tax rate.

X_9 : Tariff rate.

X_{10} : Public debt.

X_{11} : Tax burden .

X_{12} : Government expenditure.

The data was analyzed both numerically and graphically using MATLAB and XLSTAT. The analysis of the data is mainly aimed at generating the new set of economic variables (PCs) and also attempt to establish the relationship between PCA and FA as explained theoretically in Chapter 4. Furthermore a statistical process of how to control the scores of such variables in the future is examined.

5.1 Data Processing

Let $\boldsymbol{X} = \left[X_1, X_2 \ldots, X_{12}\right]$ be a random vector representing the number of variables of the world economic data available in the table given in Table 1 in Appendix I. Initially for the *p*=12 variables and *n*=186 countries the following summary statistics using XLSTAE has been obtained and given in Table 5.1.

Table 5.1. Descriptive Statistics

Variable	Observations	Mean	Variance
GDP	$n_1 = 185$	$\bar{X}_1 = 614.017$	$s_{11} = 114\ 452\ 697.342$
GDP per Capita	$n_2 = 185$	$\bar{X}_2 = 19114.554$	$s_{22} = 445\ 917\ 739.0$
GDP Growth Rate	$n_3 = 185$	$\bar{X}_3 = 2.285$	$s_{33} = 20.187$
Inflation Rate	$n_4 = 185$	$\bar{X}_4 = 4.659$	$s_{44} = 126.092$
Interest Rate	$n_5 = 185$	$\bar{X}_5 = 5.768$	$s_{55} = 31.478$
Income Tax Rate	$n_6 = 185$	$\bar{X}_6 = 27.795$	$s_{66} = 176.763$
Unemployment Rate	$n_7 = 185$	$\bar{X}_7 = 9.612$	$s_{77} = 63.934$
Corporate Tax Rate	$n_8 = 185$	$\bar{X}_8 = 23.703$	$s_{88} = 83.965$
Tariff Rate	$n_9 = 185$	$\bar{X}_9 = 5.434$	$s_{99} = 20.942$
Public Debt	$n_{10} = 185$	$\bar{X}_{10} = 53.181$	$s_{1010} = 1103.030$

Tax Burden	$n_{11} = 185$	$\bar{X}_{11} = 21.625$	$s_{1111} = 126.620$
Gov't Expenditure	$n_{12} = 185$	$\bar{X}_{12} = 33.481$	$s_{1212} = 172.971$

It is clear from the Table 5.1, that all the means differ from each other, and similarly all variances are unequal, exhibiting considerable difference. Under such circumstances it is necessary to standardize the data and use correlation matrix before performing PCA and FA.

The sample correlation matrix for all possible paired observations of twelve variables computed using equation (3.1.8) is,

$$\mathbf{R} = \begin{bmatrix} 1 & 0.145 & 0.052 & -0.027 & -0.089 & 0.172 & -0.125 & 0.082 & -0.127 & 0.153 & 0.058 & 0.034 \\ 0.145 & 1 & -0.143 & -0.154 & -0.405 & -0.059 & -0.240 & -0.311 & -0.434 & 0.020 & 0.290 & 0.255 \\ 0.052 & -0.143 & 1 & -0.305 & -0.129 & 0.106 & -0.035 & -0.006 & 0.012 & -0.034 & -0.015 & -0.081 \\ -0.027 & -0.154 & -0.305 & 1 & 0.395 & -0.089 & 0.008 & 0.066 & 0.107 & -0.057 & -0.142 & -0.111 \\ -0.089 & -0.405 & -0.129 & 0.395 & 1 & -0.120 & 0.102 & 0.196 & 0.332 & -0.057 & -0.181 & -0.152 \\ 0.172 & -0.059 & 0.106 & -0.089 & -0.120 & 1 & 0.002 & 0.591 & 0.102 & 0.274 & 0.287 & 0.109 \\ -0.125 & -0.240 & -0.035 & 0.008 & 0.102 & 0.002 & 1 & -0.006 & 0.175 & 0.112 & 0.121 & 0.300 \\ 0.082 & -0.311 & -0.006 & 0.066 & 0.196 & 0.591 & -0.006 & 1 & 0.326 & 0.096 & -0.078 & -0.126 \\ -0.127 & -0.434 & 0.012 & 0.107 & 0.332 & 0.102 & 0.175 & 0.326 & 1 & -0.008 & -0.392 & -0.187 \\ 0.153 & 0.020 & -0.034 & -0.057 & -0.057 & 0.274 & 0.112 & 0.096 & -0.008 & 1 & 0.236 & 0.191 \\ 0.058 & 0.290 & -0.015 & -0.142 & -0.181 & 0.287 & 0.121 & -0.078 & -0.392 & 0.236 & 1 & 0.559 \\ 0.034 & 0.255 & -0.081 & -0.111 & -0.152 & 0.109 & 0.300 & -0.126 & -0.187 & 0.191 & 0.559 & 1 \end{bmatrix}$$

From a visual inspection of R matrix, it is evident that pairwise correlations are not very high. This suggests that there is no extreme multicollinearity present in the data Moreover highly multicollinearity might affect the univariate contribution of the variable to a factor and may causes problems in conducting factor analysis.

5.2 Detection of Multicollinearity

Multicollinearity occurs when two or more independent variables in a Factor analysis model are highly correlated and one can be expressed as linear combination of the other variables with a certain degree of error. For example $X_3 = 3X_1 + 8X_2 + \varepsilon$. In such cases, the determinant of the correlation matrix will be zero and the factor analysis cannot be performed.

Multicollinearity can be diagnosed by computing the determinant of **R**. In this example $|\mathbf{R}| = 0.0661$ is computed. Since the **R** is invertible matrix and the determinant of **R**>0, it implies that there is no multicollinearity does not exist. In other words, there exists a few set of new uncorrelated variables (PCs) that can be expressed as linear combinations of these variables.

5.3 Kaiser-Meyer-Olkin Sampling Adequacy Test

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy test used to check whether or not the sample data is appropriate for running factor analysis.

The null and alternative hypotheses of KMO sampling adequacy test is give below**.**

H_0 : The sample data is not suitable for factor analysis.

H_1: The sample data is suitable for factor analysis.

If **D** is a diagonal matrix of inverse correlation matrix $\mathbf{R}^{-1}$, i.e. $\mathbf{D} = \mathbf{diag}\left(\mathbf{R}^{-1}\right)$ and $\mathbf{R}^{*} = \mathbf{D}^{-1/2}\mathbf{R}^{-1}\mathbf{D}^{-1/2}$ an anti-image correlation matrix, then KMO test statistic is found by the following formula,

$$KMO = \frac{\sum_{i \neq j}^{p} r_{ij}^{2}}{\sum_{i \neq j}^{p} r_{ij}^{2} + \sum_{i \neq j}^{p} r^{*2}_{ij}} ; 0 < KMO < 1$$

where $\sum_{i \neq j}^{p} r_{ij}^{2}$ is the sum of the square off diagonal entries of the squared correlation matrix $\mathbf{R}^2$ and $\sum_{i \neq j}^{p} r^{*2}_{ij}$ is the sum of the square off diagonal entries of the anti-image correlation matrix $\mathbf{R}^{*2}$.

According to KMO test, if $KMO > 0.5$ the null hypotheses will be rejected, and the sampling will be sufficient.

Since for the data used the KMO is computed as 0.614, it indicates that the data taken as a case study is adequate for running factor analysis [13].

Remark: KMO test is a way of checking whether there is some possible factors that exists leading to dimension reduction of the data. The higher the value of KMO the more powerful the factor analysis will be.

5.4 Dimension Reduction using PCA

The factor extraction and factor retention are obviously judged on the eigenvalues of correlation matrix. As a rule of thumb, the number of common factors is recommended to be the same the number of eigenvalues that are greater than unity. For the case study the eigenvalues of correlation matrix and their cumulative percentages are given in the Table 5.2.

Table 5.2. Eigenvalues and their percentage and cumulative percentage

Factors	Eigen values $Var(F_i) = \lambda_i$	Percentage distribution	Cumulative percentage
F1	2.596	21.629	21.629
F2	1.947	16.228	37.857
F3	1.544	12.870	50.727
F4	1.302	10.850	61.578
F5	0.908	7.565	69.142
F6	0.814	6.781	75.923
F7	0.738	6.153	82.076
F8	0.585	4.878	86.954
F9	0.537	4.471	91.425
F10	0.410	3.420	94.845
F11	0.365	3.042	97.887
F12	0.254	2.113	100.000

Eigenvalues represents the variances of the common factors. It is shown in the Table 5.2 that all eigenvalues are positive and their sum is equivalent to the total number of variables *i.e.* $\sum_{i=1}^{12} \lambda_i = p = 12$. This suggests that the correlation matrix is positive definite and it is possible to obtain the factors from the original data.

It is clear from the Table 5.2 that only first four eigenvalues of $\mathbf{R}$ are greater than unity. Additionally, the percentage of the total standardized population variance due to the first common factor which is the variation explained by the first factor is computed as $\frac{\lambda_1}{p} = \frac{2.596}{12} \times 100 = 21.629\%$.

Similarly, the first two and three factors together accounted for 37.857% and 50.727% of the total standardized sample variance respectively. While the cumulative percentage of the total standardized population variance explained by first four common factors is 61.578%.

Consequently the sample standardized variation is reasonably well summarized by first four common factors. Hence, in place 12 variables, 4 PCs can represent the same data or each of the 12 variables can be represented by 4 common factors.

5.5 Scree Plot

Scree plot is another tool that helps in determining the optimum number of common factors. The percentage of each $\frac{\lambda_j}{\sum \lambda_i}$ gives a visual idea about the distribution of the eigenvalues which are also represent the proportion of sample variance due to the i^{th} factor. Similarly the relative cumulative variance values produces a concave down graph that is helpful in determining the number of factors that can be used in representing each variable. Figure 5.1 clearly shows the decreasing nature of the proportion of sample variance due to the i^{th} factor, via the bar chart. It also shows the concave down relative cumulative variance values, indicating a visible decrease in the slope of the curve at around PC 4 that corresponds to about 60% of the relative cumulative variance. This also means that around 40% of variation is represented by the remaining 8 factors. This relatively high percentage of variation represented by the 8 factors is mainly attributable to the low correlation between variables as can be seen from the correlation matrix.

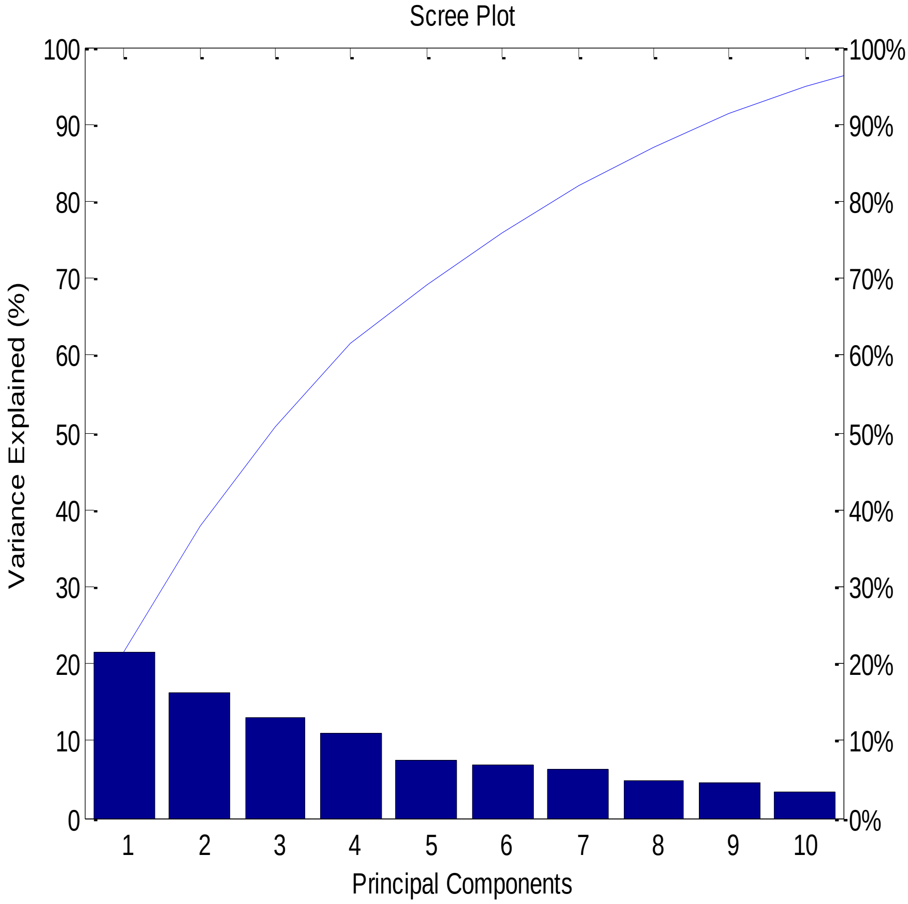


Figure 5.1. Scree plot for dimensions reduction

5.6 Reduced Eigen Space

Reduced Eigen space contains the reduced eigenvalue matrix (**E**) and reduced eigenvector (**V**) matrix of **R** are given below,

$$\mathbf{E}=\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}=\begin{bmatrix} 2.596 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.947 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.544 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.302 \end{bmatrix}$$

and

$$\mathbf{V}=\left[\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots,\boldsymbol{e}_m\right]=\begin{bmatrix} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{21} & e_{22} & e_{23} & e_{24} \\ e_{31} & e_{32} & e_{33} & e_{34} \\ e_{41} & e_{42} & e_{43} & e_{44} \\ e_{51} & e_{52} & e_{53} & e_{54} \\ e_{61} & e_{62} & e_{63} & e_{64} \\ e_{71} & e_{72} & e_{73} & e_{74} \\ e_{81} & e_{82} & e_{83} & e_{84} \\ e_{91} & e_{92} & e_{93} & e_{94} \\ e_{101} & e_{102} & e_{103} & e_{104} \\ e_{111} & e_{112} & e_{113} & e_{114} \\ e_{121} & e_{122} & e_{123} & e_{124} \end{bmatrix}=\begin{bmatrix} -0.133 & 0.158 & 0.249 & 0.365 \\ -0.438 & -0.184 & 0.051 & 0.267 \\ -0.023 & 0.076 & 0.418 & -0.501 \\ 0.255 & -0.069 & -0.343 & 0.480 \\ 0.395 & 0.034 & -0.301 & 0.145 \\ -0.048 & 0.594 & 0.208 & 0.129 \\ 0.041 & 0.203 & -0.481 & -0.433 \\ 0.247 & 0.498 & 0.186 & 0.184 \\ 0.426 & 0.188 & -0.023 & -0.175 \\ -0.138 & 0.373 & -0.100 & 0.091 \\ -0.419 & 0.262 & -0.246 & 0.011 \\ -0.356 & 0.211 & -0.415 & -0.124 \end{bmatrix}$$

Remarks: The eigenvalues measure the variation of the population principal components and the eigenvectors are indicators of the direction of the principal components. Principal components are actually scaled eigenvectors. They span the original coordinate system in the directions of great variability.

5.7 Algorithms for Relationship between PCA and FA.

In multivariate computational statistical analysis, the term algorithms refer to set of rules that can perform calculation or processing the data in order to answer statistical problems, with help of computer software. The procedure of the following algorithms step by step mentioned in Appendix II, the initial factor loadings matrix, variamx rotated factor loadings matrix, commonalities and specific variances obtained by using principal component approach and results are given in Table 5.3.

Table 5.3. Pattern matrices, communalities and specific variances by PCA method.

	Estimated factor loadings $l_{ij} = \sqrt{\lambda_i} e_{ij}$				Rotated estimated factor loadings				Communalities	Uniqueness
Variable	F_1	F_2	F_3	F_4	F_1^*	F_2^*	F_3^*	F_4^*	h_i^2	$1-h_i^2$
GDP	-0.214	0.221	0.309	0.417	-0.363	**0.467**	0.112	0.034	0.364	0.636
GDP per Capita	**-0.706**	-0.257	0.063	0.304	**-0.799**	-0.073	-0.134	0.004	0.661	0.339
GDP Growth Rate	-0.037	0.106	**0.519**	**-0.572**	0.149	0.040	0.132	**-0.754**	0.609	0.391
Inflation Rate	0.411	-0.097	-0.426	**0.548**	0.142	-0.003	0.118	**0.791**	0.66	0.34
Interest Rate	**0.637**	0.047	-0.375	0.165	**0.532**	-0.021	0.105	**0.530**	0.576	0.424
Income Tax Rate	-0.078	**0.828**	0.259	0.147	0.075	**0.845**	-0.190	-0.161	**0.781**	**0.219**
Unemployment Rate	0.065	0.284	**-0.598**	-0.494	**0.507**	-0.187	**-0.628**	-0.019	0.686	0.314
Corporate Tax Rate	0.398	**0.694**	0.231	0.210	0.389	**0.755**	0.117	0.059	**0.738**	**0.262**
Tariff Rate	**0.686**	0.263	-0.028	-0.200	**0.729**	0.134	0.173	0.045	0.581	0.419
Public Debt	-0.223	**0.521**	-0.125	0.104	-0.031	**0.429**	-0.402	0.027	0.348	0.652
Tax Burden	**-0.676**	0.365	-0.306	0.012	-0.358	0.189	**-0.719**	-0.051	0.683	0.317
Gov't Expenditure	**-0.573**	0.295	-0.516	-0.142	-0.180	-0.009	**-0.818**	0.000	**0.702**	**0.298**

5.8 Estimation of Standardized Factor Analysis Model

The estimated factor loadings matrix is computed by $\rho_{X_i,F_j} = \sqrt{\lambda_j} e_{ij} = l_{ij}$. Factor loadings represents correlation between the principal components and the standardized variables, these are computed using (4.2.5)

Like the regression analysis, the standardized score of all variables can be predicted from *m=4* standardized factor model, by the following equations from (4.2.5)

$$\begin{array}{lllllllll}
Z_1 & = & -0.214F_1 & + & 0.221F_2 & + & 0.309F_3 & + & 0.417F_4 + \varepsilon_1 \\
Z_2 & = & -0.706F_1 & + & -0.257F_2 & + & 0.063F_3 & + & 0.304F_4 + \varepsilon_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Z_{12} & = & -0.573F_1 & + & 0.295F_2 & + & -0.516F_3 & + & -0.142F_4 + \varepsilon_{12}
\end{array}$$

where $Z_1,...,Z_{12}$ standard normal variables, the coefficients of $F_1,\cdots,F_4$ are factor loadings and $\varepsilon_1,...\varepsilon_{12}$ are unknown error terms.

It clear from Z_1 model that all the corresponding factor coefficients $l_{11} = -0.214,\ l_{21} = 0.221,\ l_{31} = 0.309,\ l_{41} = 0.417$ of the $F_1,...,F_4$ respectively, are insignificantly contributes to Z_1. This means that Z_1 does not provide the best fit based on scores of $F_1,...,F_4$. Consequently, the highly correlation coefficients between the factors and variables indicates the higher factor loadings on the individual variables.

The communalities are obtained from the sum of square factor loadings, it measures the goodness of fit of the factor model. The communality for the first variable has been obtained from the sum of square factor loadings corresponding to that variable, and the complement of the communality is called uniqueness or specific variance, that is given by

Communality $= h_i^2 = \left(-0.214\right)^2 + \left(0.221\right)^2 + \left(0.309\right)^2 + \left(0.417\right)^2 = 0.364$

and

Uniqueness = $\psi_i = 1 - h_i^2 = 0.636$

This means that 36.4 % of the total standardized variation of GDP (PPP) is captured by factor model and 63.6% are dropped due to some extraneous factors. Similarly, the communalities and uniqueness for each variables are already obtained from equation (4.2.8).

The interpretation of the initial factor loadings found by PCA method is difficult. Rotation of the factor loadings matrix provides a simpler way of interpreting the obtained factor model. During the process of rotation that can be performed using any matrix rotation method, computed communalities and specific variances remains unchanged. In this study the Varimax method of rotation is employed for the rotation of factor loadings matrix by 90^o (orthogonal rotation). The rotated factor loadings are obtained from equations (4.2.14). Following the varimax rotation the new factor model is given.

$$\begin{array}{ccccccccc}
Z_1 & = & -0.363F_1^* & + & \mathbf{0.467}F_2^* & + & 0.112F_3^* & + & 0.034F_4^* + \varepsilon_1^* \\
Z_2 & = & \mathbf{-0.799}F_1^* & + & -0.073F_2^* & + & -0.134F_3^* & + & 0.004F_4^* + \varepsilon_2^* \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Z_{12} & = & 0.149F_1^* & + & 0.040F_2^* & + & 0.132F_3^* & + & \mathbf{-0.754}F_4^* + \varepsilon_{12}^*
\end{array}$$

The rotated factor model is easier to interpret then the un-rotated factor model. Factors that have high effect on the process under study becomes evident. For example for Z_1

the highest influence comes from factor F_2^*. Similar interpretations can be made for all other variables.

It is clear from Table 5.3 that the variables GDP per Capita, Interest Rate and Tariff Rate, have high loadings on factor F_1^* , low or ignorable loadings on other factors. When we look at factor F_2^*, the variables Income Tax Rate, Corporate Tax Rate, and Public Debt are dominant. Thus F_1^* can be called the *economic survival index* factor and the second factor F_2^* the *economic development index* factor. In Figure 5.2 variables with high loadings on F_1^* and F_2^* are clearly visible as they extend along the appropriate axis.

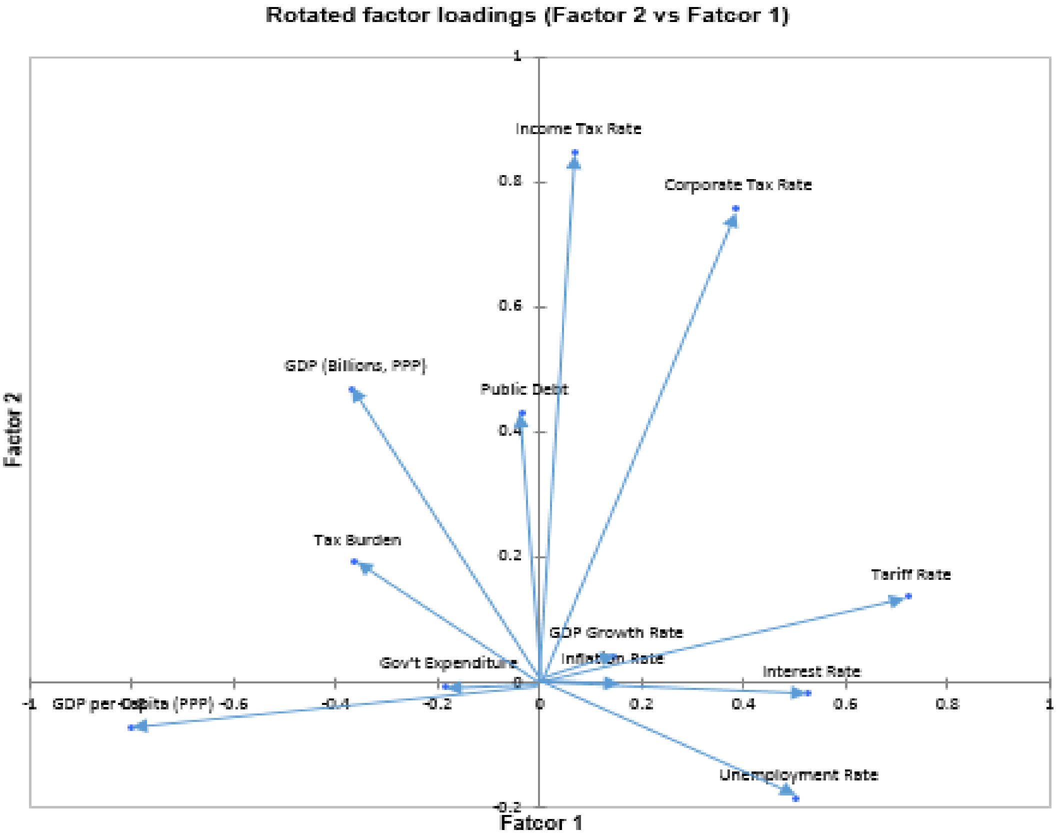


Figure 5.2. Factor loading after varimax rotation F_1^* and F_2^*

From the table 5.3 a plot for rotatated factor loadings of the last two factors F_3^* and F_4^* are used to generated a similar graph to that is given in Figure 5.2.and is given in Figure 5.3. It is demonstrated in the Figure 5.3 that the variables Unemployment Rate, Tax Burden and Gov't Expenditure have the highest loadings on F_3^* .Therefore F_3^* can be called the *economic conservative index* factor. The fourth factor F_4^* receives maximum information from three other variables, GDP Growth Rate, Inflation Rate and Interest Rate. So it can be named as the *economic inconsistency index* factor. The other variables have negligible and considerably low factor loadings on these two factors.

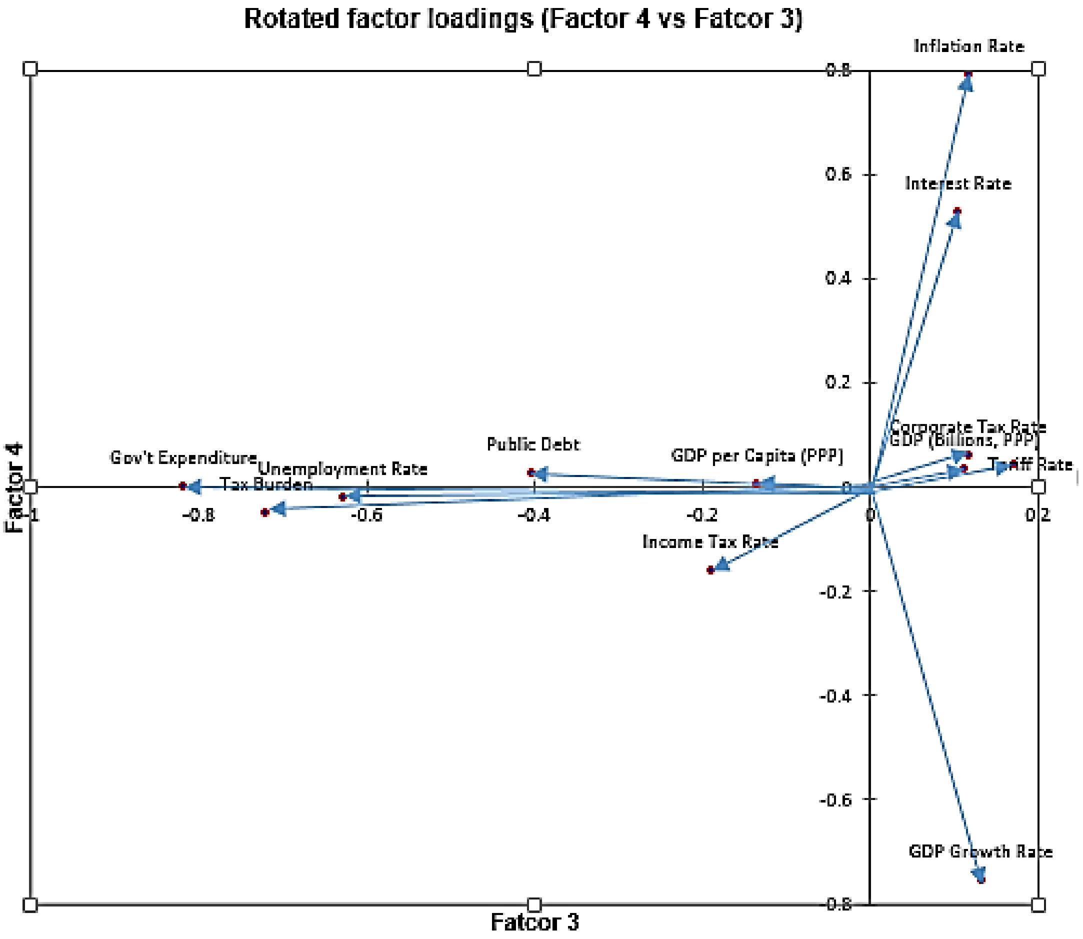


Figure 5.3. Factor loading after varimax rotation for the factors F_3^* and F_4^*.

5.9 Factor Estimation

All the standardized estimated factor score can be predicted by the following equations

$$\hat{F}^*_j = \frac{\hat{e}'^*_j}{\sqrt{\hat{\lambda}^*_j}}\left(\frac{x_j - \overline{x}}{s_{ii}}\right) \qquad for\ j = 1,2,\ldots,m\ \&\ i = 1,2,\ldots,p$$

$$= \frac{\hat{e}'^*_j}{\sqrt{\hat{\lambda}^*_j}} z_j$$

Where $\hat{\lambda}^*_j$ and $\hat{e}'^*_j$ are the j^{th} rotated paired eigenvalues and eigenvectors, for j=1, 2, 3, and 4 the following equations are obtained as

$$\hat{F}^*_1 = \frac{1}{\sqrt{\hat{\lambda}^*_1}}\left[\hat{e}^*_{11}Z_1 + \hat{e}^*_{21}Z_2 + \hat{e}^*_{31}Z_3 + \hat{e}^*_{41}Z_4 +,\ldots,+ \hat{e}^*_{121}Z_{12}\right]$$

$$\hat{F}^*_2 = \frac{1}{\sqrt{\hat{\lambda}^*_2}}\left[\hat{e}^*_{12}Z_1 + \hat{e}^*_{22}Z_2 + \hat{e}^*_{32}Z_3 + \hat{e}^*_{42}Z_4 +,\ldots,+ \hat{e}^*_{122}Z_{12}\right]$$

$$\hat{F}^*_3 = \frac{1}{\sqrt{\hat{\lambda}^*_3}}\left[\hat{e}^*_{13}Z_1 + \hat{e}^*_{23}Z_2 + \hat{e}^*_{33}Z_3 + \hat{e}^*_{43}Z_3 +,\ldots,+ \hat{e}^*_{123}Z_{12}\right]$$

$$\hat{F}^*_4 = \frac{1}{\sqrt{\hat{\lambda}^*_4}}\left[\hat{e}^*_{14}Z_1 + \hat{e}^*_{24}Z_2 + \hat{e}^*_{34}Z_3 + \hat{e}^*_{44}Z_4 +,\ldots,+ \hat{e}^*_{124}Z_{12}\right]$$

Matlab gives the following equations

$$\begin{aligned}\hat{F}^*_1 &= -0.223Z_1 - 0.376Z_2 + 0.132Z_3 - 0.020Z_4 + 0.200Z_5 + 0.021Z_6\\ &\quad +0.330Z_7 + 0.127Z_8 + 0.328Z_9 - 0.002Z_{10} - 0.107Z_{11} - 0.001Z_{12}\end{aligned}$$

$$\begin{aligned}\hat{F}^*_2 &= 0.300Z_1 - 0.001Z_2 - 0.015Z_3 + 0.035Z_4 - 0.012Z_5 + 0.462Z_6\\ &\quad -0.178Z_7 + 0.421Z_8 + 0.045Z_9 + 0.226Z_{10} + 0.084Z_{11} - 0.044Z_{12}\end{aligned}$$

$$\begin{aligned}\hat{F}^*_3 &= 0.129Z_1 + 0.005Z_2 + 0.078Z_3 + 0.032Z_4 - 0.012Z_5 - 0.059Z_6\\ &\quad -0.415Z_7 + 0.069Z_8 + 0.026Z_9 - 0.196Z_{10} - 0.351Z_{11} - 0.438Z_{12}\end{aligned}$$

$$\begin{aligned}\hat{F}^*_4 &= 0.079Z_1 + 0.081Z_2 - 0.534Z_3 + 0.527Z_4 + 0.309Z_5 - 0.073Z_6\\ &\quad -0.057Z_7 + 0.036Z_8 - 0.038Z_9 + 0.052Z_{10} + 0.026Z_{11} + 0.037Z_{12}\end{aligned}$$

After extracting all the small and negligible coefficients and labeling, the following equations created,

$$ESI = -0.376 GDP\ per\ Capita + 0.330 Unemployment\ Rate + 0.328 Tariff\ \ Rate$$

$$EDI = 0.300 GDP + 0.462 Income\ Tax\ Rate + 0.421 Corporate\ Tax\ Rate + 0.226 Public\ Debt$$

$$ECI = 0.415 Unemployment\ Rate + 0.351 Tax\ Burden + 0.438 Gov't\ Expenditure$$

$$EII = -0.534\,Growth\ Rate + 0.527\,Inflation\ Rate + 0.309\,Interest\ Rate$$

5.10 Economic Survival Index (ESI)

The standardized factor scores for all four factors can be computed by using equation (4.2.18) .Out of 185 countries, as an example the standardized factor scores are computed for the top 10 ranking countries. The factor scores for *economic survival index* (ESI) are given in Table 5.4.

Table 5.4. Standardized score of economic survival index

Country	ESI	Ranking
United States	2.977	1
Macau	2.800	2
Qatar	2.731	3
Luxembourg	2.125	4
China	2.096	5
Singapore	2.026	6
Brunei Darussalam	1.983	7
Switzerland	1.848	8
Germany	1.766	9
United Arab Emirates	1.712	10

First ranked United States, has the highest ESI with a score 2.977. United Arab Emirates is ranked 10th, has ESI 1.712. Thus, United States has the highest ESI as compared with other countries.

5.11 Economic Developmental Index (EDI)

The table 5.5 shows the *economic developmental index* (EDI) score for the top 10 ranking countries.

Table 5.5: Economic developmental index score

Country Name	EDI	Ranking
United States	3.816	1
China	3.312	2
Japan	2.533	3
India	1.855	4
France	1.708	5
Belgium	1.630	6
Italy	1.480	7
Greece	1.198	8
Austria	1.176	9
Netherlands	1.143	10

According to EDI given in Table 5.5 United States ranks the first, and Netherlands occupies the tenth place. This index can be used as a measure for the level of development of the concerned country.

5.12 Economic Conservative Index (ECI)

The *economic conservative index* (ECI) score for the top 10 ranking countries are listed in Table 5.6.

Table 5.6. Economic conservative index score

Country Name	ECI	Ranking
Liechtenstein	2.386757	1
Nigeria	1.877187	2
Bangladesh	1.713963	3
Republic of the Congo	1.59834	4
China	1.596723	5
Guatemala	1.535151	6
Madagascar	1.437041	7
Cambodia	1.399659	8
Indonesia	1.38805	9
Hong Kong SAR	1.370173	10

According to Table 5.6 Liechtenstein has the highest ECI, while Hong Kong is in the tenth position. This reveals Liechtenstein has the highest concentration on the change to development of its economy, while those with lower ECI lack this concentration.

5.13 Economic Inconsistent Index (EII)

Economic inconsistent index (EII) score for the highest scoring countries are listed in Table 5.7.

Table 5.7: Economic inconsistent Index score

Country Name	EII	Ranking
Venezuela	7.290538631	1
Yemen	5.182715378	2
Ukraine	4.014900017	3
Sierra Leone	3.223664087	4
Macau	2.870468153	5
Argentina	2.391867354	6
Korea, North	1.874090873	7
Belarus	1.731406576	8
Russia	1.636494569	9
Malawi	1.59328325	10

The EII is an indicator of inconsistencies in the economy of a country mostly due to imbalance in economy and various factors that causes big fluctuations in economy. According to analysis results from the data Venezuela has the highest EII, while Malawi is in the tenth position.

5.14 Statistical Control Ellipse

In the order to see which observations of the factor scores are statistically in or out of control in the future, two control ellipse charts are generated using the four factors.For the purpose of this study the pairwise comparison between the factors (F_1^*, F_2^*) and (F_3^*, F_4^*) are examined.

The 95 % confidence quality control ellipse for all pairs of values of ESI and EDI are obtained the following equation as

$$\frac{\hat{F}_1^{*2}}{\lambda^*_1} + \frac{\hat{F}_2^{*2}}{\lambda^*_2} \le \chi_2^2(.05) \qquad (5.1.2)$$

where

$$\begin{aligned}\lambda^*_1 &= l^{*2}_{11} + l^{*2}_{12} +,...,+ l^{*2}_{112} \\ &= (-0.363)^2 + (\mathbf{-0.799})^2 +,...,+ (-0.516)^2 \\ &= 2.20\end{aligned}$$

$$\begin{aligned}\lambda^*_2 &= l^{*2}_{21} + l^{*2}_{22} +,...,+ l^{*2}_{212} \\ &= (\mathbf{0.467})^2 +,...,+ (-0.009)^2 \\ &= 1.78\end{aligned}$$

Similarly the 95 % control ellipse for ECI and EII factors can be determined. That is

$$\frac{\hat{F}_3^{*2}}{\lambda^*_3} + \frac{\hat{F}_4^{*2}}{\lambda^*_4} \le \chi_2^2(.05) \qquad (5.1.3)$$

where

$$\begin{aligned}\lambda^*_3 &= l^{*2}_{31} + l^{*2}_{32} +,...,+ l^{*2}_{312} \\ &= (0.112)^2 + (-0.134)^2 +,...,+ (\mathbf{-0.818})^2 \text{ and} \\ &= 1.89\end{aligned}$$

$$\begin{aligned}\lambda^*_4 &= l^{*2}_{41} + l^{*2}_{42} +,...,+ l^{*2}_{412} \\ &= (0.034)^2 +,...,+ (0)^2 \\ &= 1.51\end{aligned}$$

The above values of the rotated factor loadings are taken from table 5.3.

This chart was drawn benefiting from equations 3.1.10 and 5.1.2. Using the standardized scores of ESI and EDI

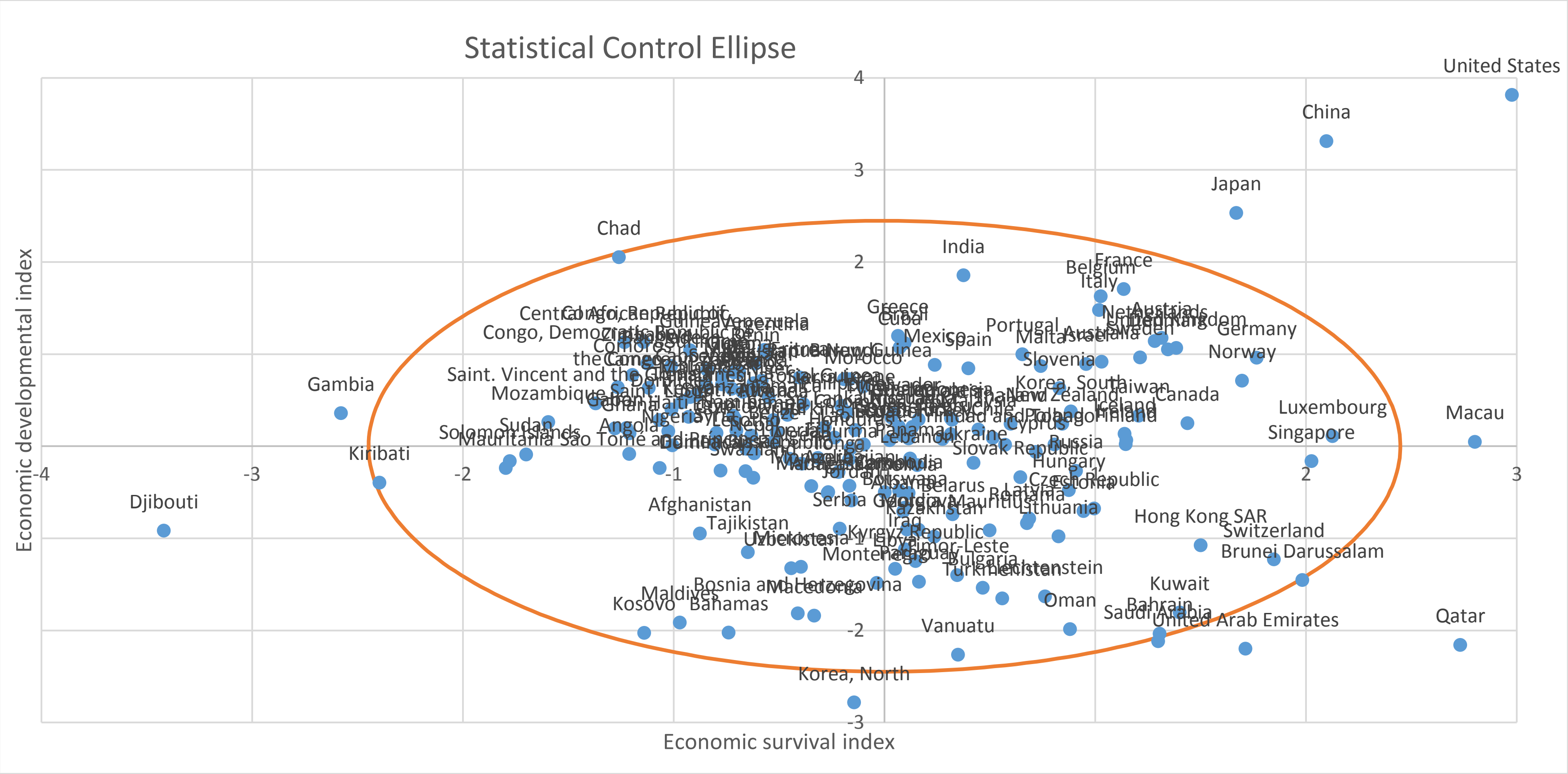


Figure5.4. 95% Statistical Control Ellipse for EDI - ESI pairs

This chart was drawn through the equation 3.1.11 and 5.1.3 and using the standardized scores of ECI and EII

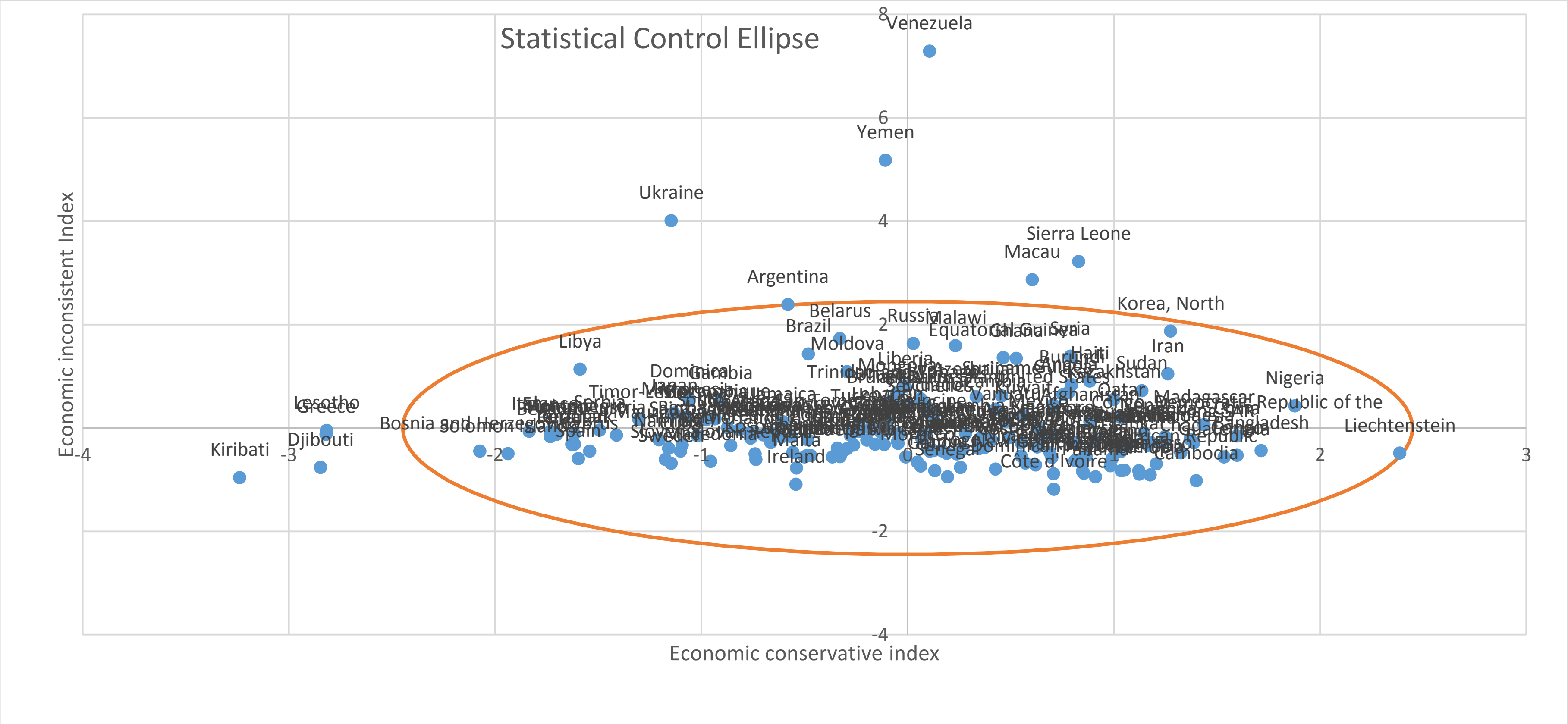


Figure 5.5. 95% Statistical Control Ellipse Chart for ECI and EII pairs

5.14 General Interpretations of Statistical Control Ellipse Charts

The Figure 5.4 shows that 95% control ellipse of the scores for 185 countries. Evidently 9 countries out of the control ellipse. They either have very high or low factor scores on either or on both indexes. Hence, these countries should investigate and change the policy to growth or survival factors influencing the economy in the future.

The control ellipse 5.5 shows that factors score of the 8 countries falls outside the control ellipse. They are statistically out of control with 5% level of significance, which is due to inconsistency and instabilities. Consequently the countries whose scores lie outside the control ellipse, need to take some remedial action to protect their economy against inconsistencies and instability in the future.

Chapter 6

CONCLUSION

In multivariate statistical analysis, the relationship between PCA and FA is investigated. Explored theory is applied to a multidimensional data and obtained results are interpreted. The statistical analysis of such large data without using the PCA – FA relationship would be incomplete. Furthermore, PCA was used as a tool for factors extraction and variable selection in FA.

The link between FA and PCA was investigated through the correlation matrix. Correlation matrix has been chosen due to significant differences between the observations of each variable. More specifically, the correlation matrix would be only applicable when the variation from variable to variable is considerably large, or the units of each variable are not identical.

Before running PCA in FA, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy test was applied to the overall data, where the correlation matrix is utilized to determine the suitability of the data for FA. Obtained test result indicated that the application of FA to the selected sample data is possible. MATLAB is used to compute the eigenvalues and eigenvectors of the correlation matrix. It is worth remembering that the obtained eigenvalues are actually the variances of PCs and PCs are scaled eigenvectors span the original coordinate system to new coordinate system in the directions of the greatest variation of the original data.

In PCA the scree plot helps to determine the number of factors for FA model. These CFs explains the maximum variability in the original data. Moreover, this plot constructed from the eigenvalues of the correlation matrix and provides an evidence to the possible number of highly correlated groups of variables. This enables the selection of the number of factors to be used in establishing the orthogonal factor model for the data.

In PCA, the principal loadings are simply the eigenvectors. In FA the factor loadings are the correlation values between the original variables and underlying common factors. Generally the factor loadings helps to compute the communalities and uniqueness of a particular variable. Additionally, communality can be interpreted for a particular explanatory variable in FA, as it is for a response variable the coefficient of the determination in regression analysis.

In FA sometime is very tedious to interpret the initial factor loadings obtained by PCA approach. In such circumstances the varimax rotation criteria can be used to rotate the initial factor loadings, obtaining a better picture for more elaborate interpretation. Additionally the factor scores estimated by using ordinary least square method based on rotated factor loadings, enables the detection and cleaning outliers from the data. This facility is widely used in the field of statistical quality control and subsequent analysis.

REFRENCEES

[1] Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology*, *15*, 201-293.

[2] Cattell, R. B. (1978). Use of Factor Analysis in Behavioral and Life Sciences. New York: Plenum.

[3] Cattell, R. B. (1973). Personality and Mood by Questionnaire. San Francisco, CA: Jossey-Bass.

[4] Cattell, H. E. P. & Mead, A. D. (2008). The Sixteen Personality Factor Questionnaire (16PF). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds), The Sage Handbook of Personality Theory and Assessment: Vol. 2, Personality Measurement and Testing. Los Angeles, CA: Sage.

[5] Pearson, K. (1901). On lines and planes of closet fit to systems of points in space. Philosophical Magazine , 565.

[6] Robert W. Taft – 2009. Progress in Physical Organic Chemistry, Volume 18,John Wiley & Sons, inc New York ,78.

[7] Hotelling, H. (1933). *Analysis of a Complex of Statistical Variables into Principal Components*. Journal of Educational Psychology. American Psychological Association. *24 (6): 417–441.* doi:10.1037/h0071325.

[8] Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 27, 321-77.

[9] Brian S. Everitt and Torsten Hothorn (2011). An Introduction to applied Multiva - riate Analysis with R. Springer

[10] Girshick, M. A. On the Sampling Theory of Roots of Determinantal Equations. Ann. Math. Statist. 10 (1939), no. 3, 203-224. doi:10.1214/aoms/1177732180.

[11] Gower J. C, Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis. Biometrika 53:325-38, 1966.

[12] Kaiser H. F. (1970) ,A second generation little jiffy. Psychometrika, 35(4):401–415.

[13] Kaiser H. F. and Rice J. (1974) Little jiffy, mark iv. Educational and Psychological Measurement, 34(1): 111–117.

[14] Vavra T.G. (1972), Factor analysis of perceptual change. J. of Marketing Research 9: 193-199.

[15] Jackson, J.E. and Lawton, W. H. (1976) some probability problems associated Cross-impact analysis Analysis, Technology Forecasting and Social 8,263-273.

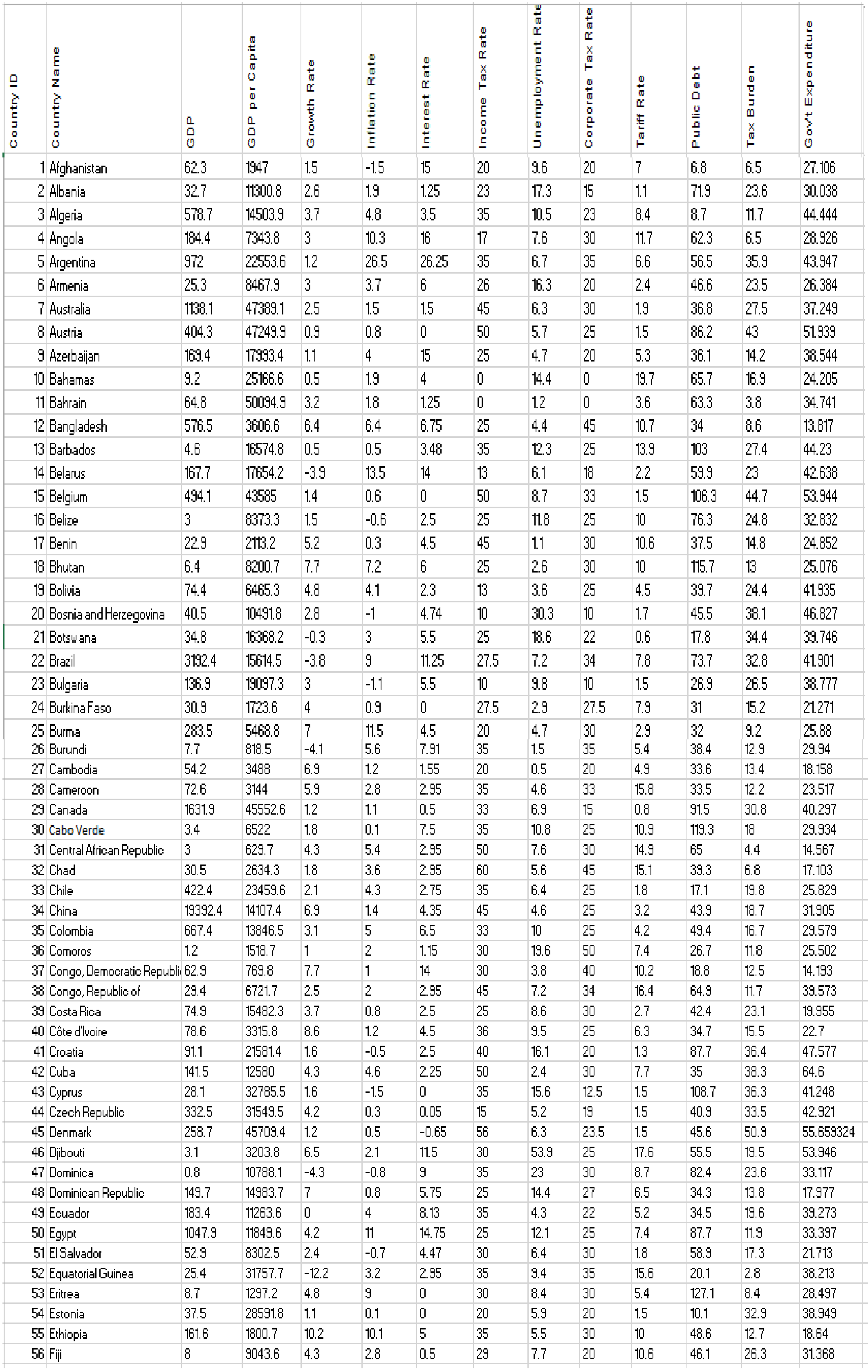
[16] Jackson J. E., 2005, A User's Guide to Principal Components, John Wiley & Sons inch Newyork.

[17] Fabrigar (1999). Evaluating the use of exploratory factor analysis in psychological research. (PDF). Psychological Methods.

[18] Johnson R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* New Jersey: Pearson.

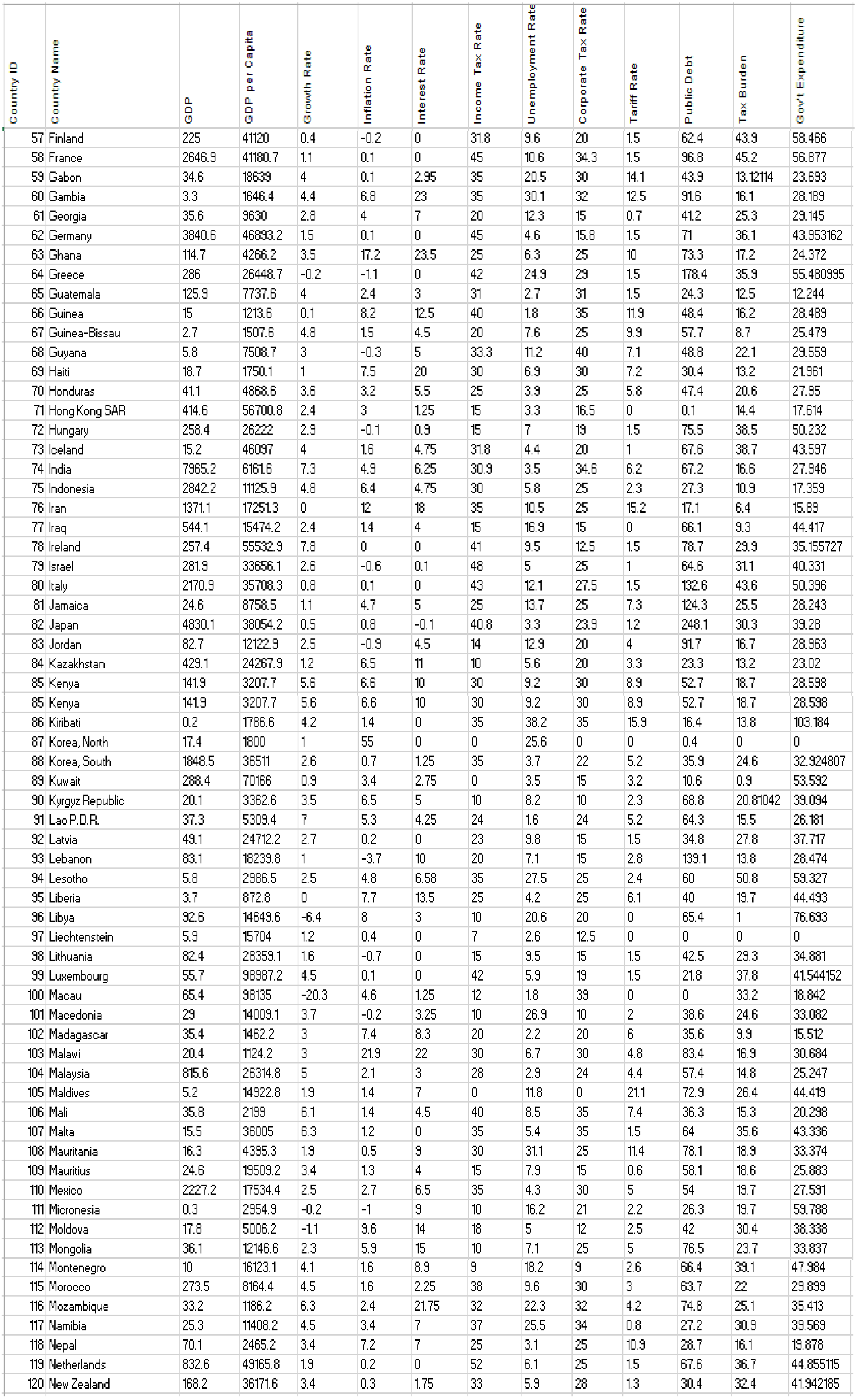
APPENDICES

Appendix A: World Economic Data



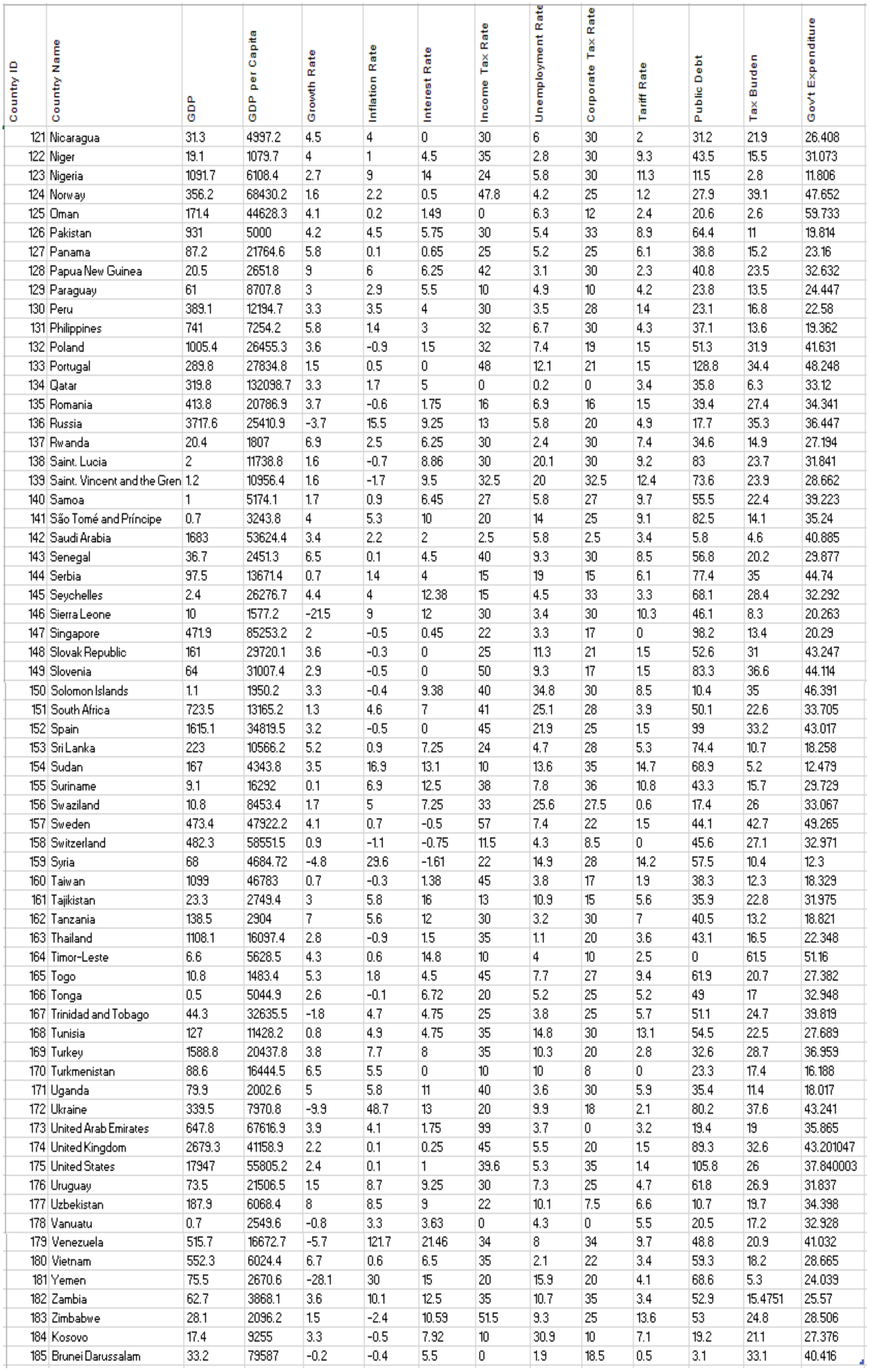
Country ID	Country Name	GDP	GDP per Capita	Growth Rate	Inflation Rate	Interest Rate	Income Tax Rate	Unemployment Rate	Corporate Tax Rate	Tariff Rate	Public Debt	Tax Burden	Gov't Expenditure
1	Afghanistan	62.3	1947	1.5	-1.5	15	20	9.6	20	7	6.8	6.5	27.106
2	Albania	32.7	11300.8	2.6	1.9	1.25	23	17.3	15	1.1	71.9	23.6	30.038
3	Algeria	578.7	14503.9	3.7	4.8	3.5	35	10.5	23	8.4	8.7	11.7	44.444
4	Angola	184.4	7343.8	3	10.3	16	17	7.6	30	11.7	62.3	6.5	28.926
5	Argentina	972	22553.6	1.2	26.5	26.25	35	6.7	35	6.6	56.5	35.9	43.947
6	Armenia	25.3	8467.9	3	3.7	6	26	16.3	20	2.4	46.6	23.5	26.384
7	Australia	1138.1	47389.1	2.5	1.5	1.5	45	6.3	30	1.9	36.8	27.5	37.249
8	Austria	404.3	47249.9	0.9	0.8	0	50	5.7	25	1.5	86.2	43	51.939
9	Azerbaijan	169.4	17993.4	1.1	4	15	25	4.7	20	5.3	36.1	14.2	38.544
10	Bahamas	9.2	25166.6	0.5	1.9	4	0	14.4	0	19.7	65.7	16.9	24.205
11	Bahrain	64.8	50094.9	3.2	1.8	1.25	0	1.2	0	3.6	63.3	3.8	34.741
12	Bangladesh	576.5	3606.6	6.4	6.4	6.75	25	4.4	45	10.7	34	8.6	13.817
13	Barbados	4.6	16574.8	0.5	0.5	3.48	35	12.3	25	13.9	103	27.4	44.23
14	Belarus	167.7	17654.2	-3.9	13.5	14	13	6.1	18	2.2	59.9	23	42.638
15	Belgium	494.1	43585	1.4	0.6	0	50	8.7	33	1.5	106.3	44.7	53.944
16	Belize	3	8373.3	1.5	-0.6	2.5	25	11.8	25	10	76.3	24.8	32.832
17	Benin	22.9	2113.2	5.2	0.3	4.5	45	1.1	30	10.6	37.5	14.8	24.852
18	Bhutan	6.4	8200.7	7.7	7.2	6	25	2.6	30	10	115.7	13	25.076
19	Bolivia	74.4	6465.3	4.8	4.1	2.3	13	3.6	25	4.5	39.7	24.4	41.935
20	Bosnia and Herzegovina	40.5	10491.8	2.8	-1	4.74	10	30.3	10	1.7	45.5	38.1	46.827
21	Botswana	34.8	16368.2	-0.3	3	5.5	25	18.6	22	0.6	17.8	34.4	39.746
22	Brazil	3192.4	15614.5	-3.8	9	11.25	27.5	7.2	34	7.8	73.7	32.8	41.901
23	Bulgaria	136.9	19097.3	3	-1.1	5.5	10	9.8	10	1.5	26.9	26.5	38.777
24	Burkina Faso	30.9	1723.6	4	0.9	0	27.5	2.9	27.5	7.9	31	15.2	21.271
25	Burma	283.5	5468.8	7	11.5	4.5	20	4.7	30	2.9	32	9.2	25.88
26	Burundi	7.7	818.5	-4.1	5.6	7.91	35	1.5	35	5.4	38.4	12.9	29.94
27	Cambodia	54.2	3488	6.9	1.2	1.55	20	0.5	20	4.9	33.6	13.4	18.158
28	Cameroon	72.6	3144	5.9	2.8	2.95	35	4.6	33	15.8	33.5	12.2	23.517
29	Canada	1631.9	45552.6	1.2	1.1	0.5	33	6.9	15	0.8	91.5	30.8	40.297
30	Cabo Verde	3.4	6522	1.8	0.1	7.5	35	10.8	25	10.9	119.3	18	29.934
31	Central African Republic	3	629.7	4.3	5.4	2.95	50	7.6	30	14.9	65	4.4	14.567
32	Chad	30.5	2634.3	1.8	3.6	2.95	60	5.6	45	15.1	39.3	6.8	17.103
33	Chile	422.4	23459.6	2.1	4.3	2.75	35	6.4	25	1.8	17.1	19.8	25.829
34	China	19392.4	14107.4	6.9	1.4	4.35	45	4.6	25	3.2	43.9	18.7	31.905
35	Colombia	667.4	13846.5	3.1	5	6.5	33	10	25	4.2	49.4	16.7	29.579
36	Comoros	1.2	1518.7	1	2	1.15	30	19.6	50	7.4	26.7	11.8	25.502
37	Congo, Democratic Republi	62.9	769.8	7.7	1	14	30	3.8	40	10.2	18.8	12.5	14.193
38	Congo, Republic of	29.4	6721.7	2.5	2	2.95	45	7.2	34	16.4	64.9	11.7	39.573
39	Costa Rica	74.9	15482.3	3.7	0.8	2.5	25	8.6	30	2.7	42.4	23.1	19.955
40	Côte d'Ivoire	78.6	3315.8	8.6	1.2	4.5	36	9.5	25	6.3	34.7	15.5	22.7
41	Croatia	91.1	21581.4	1.6	-0.5	2.5	40	16.1	20	1.3	87.7	36.4	47.577
42	Cuba	141.5	12580	4.3	4.6	2.25	50	2.4	30	7.7	35	38.3	64.6
43	Cyprus	28.1	32785.5	1.6	-1.5	0	35	15.6	12.5	1.5	108.7	36.3	41.248
44	Czech Republic	332.5	31549.5	4.2	0.3	0.05	15	5.2	19	1.5	40.9	33.5	42.921
45	Denmark	258.7	45709.4	1.2	0.5	-0.65	56	6.3	23.5	1.5	45.6	50.9	55.659324
46	Djibouti	3.1	3203.8	6.5	2.1	11.5	30	53.9	25	17.6	55.5	19.5	53.946
47	Dominica	0.8	10788.1	-4.3	-0.8	9	35	23	30	8.7	82.4	23.6	33.117
48	Dominican Republic	149.7	14983.7	7	0.8	5.75	25	14.4	27	6.5	34.3	13.8	17.977
49	Ecuador	183.4	11263.6	0	4	8.13	35	4.3	22	5.2	34.5	19.6	39.273
50	Egypt	1047.9	11849.6	4.2	11	14.75	25	12.1	25	7.4	87.7	11.9	33.397
51	El Salvador	52.9	8302.5	2.4	-0.7	4.47	30	6.4	30	1.8	58.9	17.3	21.713
52	Equatorial Guinea	25.4	31757.7	-12.2	3.2	2.95	35	9.4	35	15.6	20.1	2.8	38.213
53	Eritrea	8.7	1297.2	4.8	9	0	30	8.4	30	5.4	127.1	8.4	28.497
54	Estonia	37.5	28591.8	1.1	0.1	0	20	5.9	20	1.5	10.1	32.9	38.949
55	Ethiopia	161.6	1800.7	10.2	10.1	5	35	5.5	30	10	48.6	12.7	18.64
56	Fiji	8	9043.6	4.3	2.8	0.5	29	7.7	20	10.6	46.1	26.3	31.368

Table World economic data continued



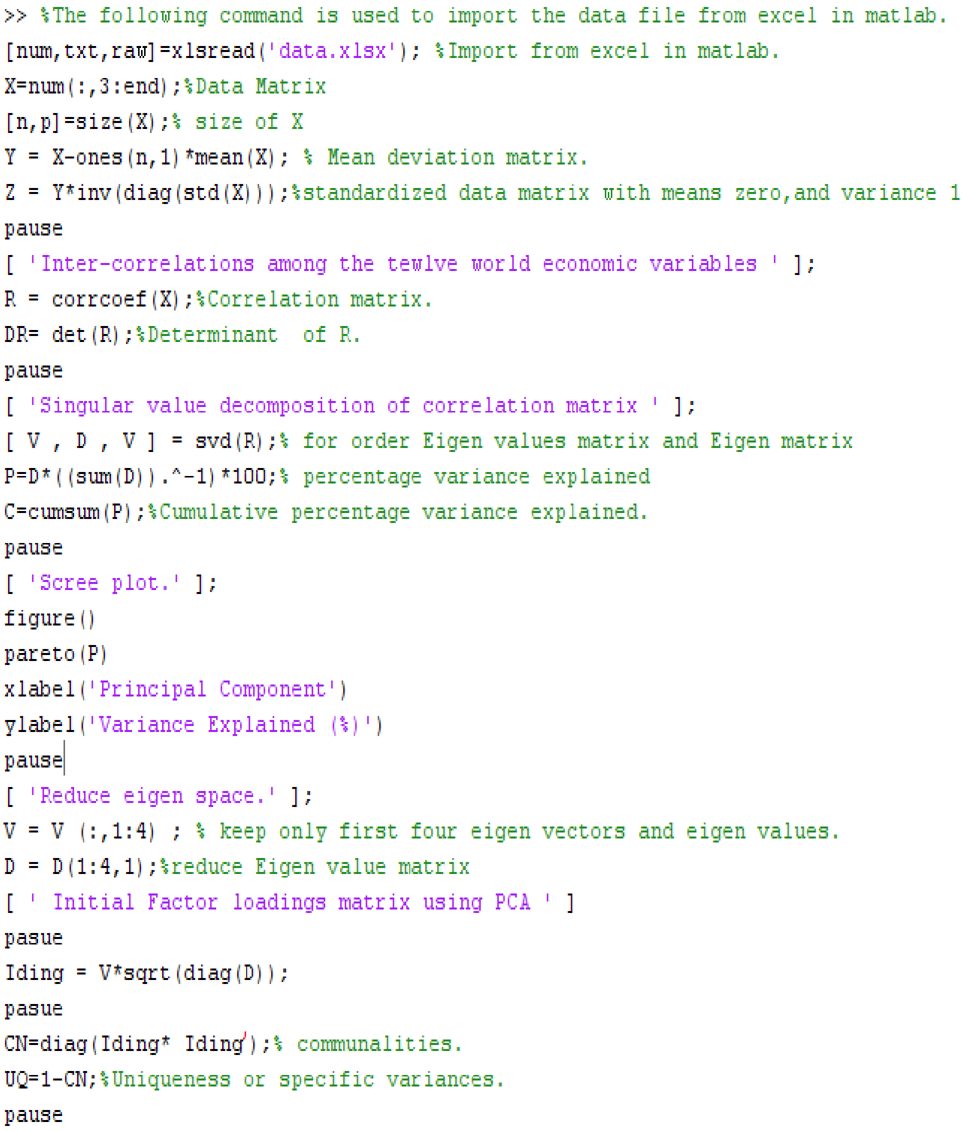
Country ID	Country Name	GDP	GDP per Capita	Growth Rate	Inflation Rate	Interest Rate	Income Tax Rate	Unemployment Rate	Corporate Tax Rate	Tariff Rate	Public Debt	Tax Burden	Gov't Expenditure
57	Finland	225	41120	0.4	-0.2	0	31.8	9.6	20	1.5	62.4	43.9	58.466
58	France	2646.9	41180.7	1.1	0.1	0	45	10.6	34.3	1.5	96.8	45.2	56.877
59	Gabon	34.6	18639	4	0.1	2.95	35	20.5	30	14.1	43.9	13.12114	23.693
60	Gambia	3.3	1646.4	4.4	6.8	23	35	30.1	32	12.5	91.6	16.1	28.189
61	Georgia	35.6	9630	2.8	4	7	20	12.3	15	0.7	41.2	25.3	29.145
62	Germany	3840.6	46893.2	1.5	0.1	0	45	4.6	15.8	1.5	71	36.1	43.953162
63	Ghana	114.7	4266.2	3.5	17.2	23.5	25	6.3	25	10	73.3	17.2	24.372
64	Greece	286	26448.7	-0.2	-1.1	0	42	24.9	29	1.5	178.4	35.9	55.480995
65	Guatemala	125.9	7737.6	4	2.4	3	31	2.7	31	1.5	24.3	12.5	12.244
66	Guinea	15	1213.6	0.1	8.2	12.5	40	1.8	35	11.9	48.4	16.2	28.489
67	Guinea-Bissau	2.7	1507.6	4.8	1.5	4.5	20	7.6	25	9.9	57.7	8.7	25.479
68	Guyana	5.8	7508.7	3	-0.3	5	33.3	11.2	40	7.1	48.8	22.1	29.559
69	Haiti	18.7	1750.1	1	7.5	20	30	6.9	30	7.2	30.4	13.2	21.961
70	Honduras	41.1	4868.6	3.6	3.2	5.5	25	3.9	25	5.8	47.4	20.6	27.95
71	Hong Kong SAR	414.6	56700.8	2.4	3	1.25	15	3.3	16.5	0	0.1	14.4	17.614
72	Hungary	258.4	26222	2.9	-0.1	0.9	15	7	19	1.5	75.5	38.5	50.232
73	Iceland	15.2	46097	4	1.6	4.75	31.8	4.4	20	1	67.6	38.7	43.597
74	India	7965.2	6161.6	7.3	4.9	6.25	30.9	3.5	34.6	6.2	67.2	16.6	27.946
75	Indonesia	2842.2	11125.9	4.8	6.4	4.75	30	5.8	25	2.3	27.3	10.9	17.359
76	Iran	1371.1	17251.3	0	12	18	35	10.5	25	15.2	17.1	6.4	15.89
77	Iraq	544.1	15474.2	2.4	1.4	4	15	16.9	15	0	66.1	9.3	44.417
78	Ireland	257.4	55532.9	7.8	0	0	41	9.5	12.5	1.5	78.7	29.9	35.155727
79	Israel	281.9	33656.1	2.6	-0.6	0.1	48	5	25	1	64.6	31.1	40.331
80	Italy	2170.9	35708.3	0.8	0.1	0	43	12.1	27.5	1.5	132.6	43.6	50.396
81	Jamaica	24.6	8758.5	1.1	4.7	5	25	13.7	25	7.3	124.3	25.5	28.243
82	Japan	4830.1	38054.2	0.5	0.8	-0.1	40.8	3.3	23.9	1.2	248.1	30.3	39.28
83	Jordan	82.7	12122.9	2.5	-0.9	4.5	14	12.9	20	4	91.7	16.7	28.963
84	Kazakhstan	429.1	24267.9	1.2	6.5	11	10	5.6	20	3.3	23.3	13.2	23.02
85	Kenya	141.9	3207.7	5.6	6.6	10	30	9.2	30	8.9	52.7	18.7	28.598
85	Kenya	141.9	3207.7	5.6	6.6	10	30	9.2	30	8.9	52.7	18.7	28.598
86	Kiribati	0.2	1786.6	4.2	1.4	0	35	38.2	35	15.9	16.4	13.8	103.184
87	Korea, North	17.4	1800	1	55	0	0	25.6	0	0	0.4	0	0
88	Korea, South	1848.5	36511	2.6	0.7	1.25	35	3.7	22	5.2	35.9	24.6	32.924807
89	Kuwait	288.4	70166	0.9	3.4	2.75	0	3.5	15	3.2	10.6	0.9	53.592
90	Kyrgyz Republic	20.1	3362.6	3.5	6.5	5	10	8.2	10	2.3	68.8	20.81042	39.094
91	Lao P.D.R.	37.3	5309.4	7	5.3	4.25	24	1.6	24	5.2	64.3	15.5	26.181
92	Latvia	49.1	24712.2	2.7	0.2	0	23	9.8	15	1.5	34.8	27.8	37.717
93	Lebanon	83.1	18239.8	1	-3.7	10	20	7.1	15	2.8	139.1	13.8	28.474
94	Lesotho	5.8	2986.5	2.5	4.8	6.58	35	27.5	25	2.4	60	50.8	59.327
95	Liberia	3.7	872.8	0	7.7	13.5	25	4.2	25	6.1	40	19.7	44.493
96	Libya	92.6	14649.6	-6.4	8	3	10	20.6	20	0	65.4	1	76.693
97	Liechtenstein	5.9	15704	1.2	0.4	0	7	2.6	12.5	0	0	0	0
98	Lithuania	82.4	28359.1	1.6	-0.7	0	15	9.5	15	1.5	42.5	29.3	34.881
99	Luxembourg	55.7	98987.2	4.5	0.1	0	42	5.9	19	1.5	21.8	37.8	41.544152
100	Macau	65.4	98135	-20.3	4.6	1.25	12	1.8	39	0	0	33.2	18.842
101	Macedonia	29	14009.1	3.7	-0.2	3.25	10	26.9	10	2	38.6	24.6	33.082
102	Madagascar	35.4	1462.2	3	7.4	8.3	20	2.2	20	6	35.6	9.9	15.512
103	Malawi	20.4	1124.2	3	21.9	22	30	6.7	30	4.8	83.4	16.9	30.684
104	Malaysia	815.6	26314.8	5	2.1	3	28	2.9	24	4.4	57.4	14.8	25.247
105	Maldives	5.2	14922.8	1.9	1.4	7	0	11.8	0	21.1	72.9	26.4	44.419
106	Mali	35.8	2199	6.1	1.4	4.5	40	8.5	35	7.4	36.3	15.3	20.298
107	Malta	15.5	36005	6.3	1.2	0	35	5.4	35	1.5	64	35.6	43.336
108	Mauritania	16.3	4395.3	1.9	0.5	9	30	31.1	25	11.4	78.1	18.9	33.374
109	Mauritius	24.6	19509.2	3.4	1.3	4	15	7.9	15	0.6	58.1	18.6	25.883
110	Mexico	2227.2	17534.4	2.5	2.7	6.5	35	4.3	30	5	54	19.7	27.591
111	Micronesia	0.3	2954.9	-0.2	-1	9	10	16.2	21	2.2	26.3	19.7	59.788
112	Moldova	17.8	5006.2	-1.1	9.6	14	18	5	12	2.5	42	30.4	38.338
113	Mongolia	36.1	12146.6	2.3	5.9	15	10	7.1	25	5	76.5	23.7	33.837
114	Montenegro	10	16123.1	4.1	1.6	8.9	9	18.2	9	2.6	66.4	39.1	47.984
115	Morocco	273.5	8164.4	4.5	1.6	2.25	38	9.6	30	3	63.7	22	29.899
116	Mozambique	33.2	1186.2	6.3	2.4	21.75	32	22.3	32	4.2	74.8	25.1	35.413
117	Namibia	25.3	11408.2	4.5	3.4	7	37	25.5	34	0.8	27.2	30.9	39.569
118	Nepal	70.1	2465.2	3.4	7.2	7	25	3.1	25	10.9	28.7	16.1	19.878
119	Netherlands	832.6	49165.8	1.9	0.2	0	52	6.1	25	1.5	67.6	36.7	44.855115
120	New Zealand	168.2	36171.6	3.4	0.3	1.75	33	5.9	28	1.3	30.4	32.4	41.942185

Table World economic data continued

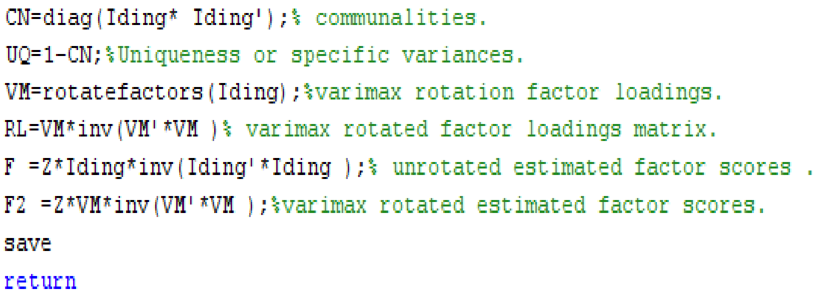


Country ID	Country Name	GDP	GDP per Capita	Growth Rate	Inflation Rate	Interest Rate	Income Tax Rate	Unemployment Rate	Corporate Tax Rate	Tariff Rate	Public Debt	Tax Burden	Gov't Expenditure
121	Nicaragua	31.3	4997.2	4.5	4	0	30	6	30	2	31.2	21.9	26.408
122	Niger	19.1	1079.7	4	1	4.5	35	2.8	30	9.3	43.5	15.5	31.073
123	Nigeria	1091.7	6108.4	2.7	9	14	24	5.8	30	11.3	11.5	2.8	11.806
124	Norway	356.2	68430.2	1.6	2.2	0.5	47.8	4.2	25	1.2	27.9	39.1	47.652
125	Oman	171.4	44628.3	4.1	0.2	1.49	0	6.3	12	2.4	20.6	2.6	59.733
126	Pakistan	931	5000	4.2	4.5	5.75	30	5.4	33	8.9	64.4	11	19.814
127	Panama	87.2	21764.6	5.8	0.1	0.65	25	5.2	25	6.1	38.8	15.2	23.16
128	Papua New Guinea	20.5	2651.8	9	6	6.25	42	3.1	30	2.3	40.8	23.5	32.632
129	Paraguay	61	8707.8	3	2.9	5.5	10	4.9	10	4.2	23.8	13.5	24.447
130	Peru	389.1	12194.7	3.3	3.5	4	30	3.5	28	1.4	23.1	16.8	22.58
131	Philippines	741	7254.2	5.8	1.4	3	32	6.7	30	4.3	37.1	13.6	19.362
132	Poland	1005.4	26455.3	3.6	-0.9	1.5	32	7.4	19	1.5	51.3	31.9	41.631
133	Portugal	289.8	27834.8	1.5	0.5	0	48	12.1	21	1.5	128.8	34.4	48.248
134	Qatar	319.8	132098.7	3.3	1.7	5	0	0.2	0	3.4	35.8	6.3	33.12
135	Romania	413.8	20786.9	3.7	-0.6	1.75	16	6.9	16	1.5	39.4	27.4	34.341
136	Russia	3717.6	25410.9	-3.7	15.5	9.25	13	5.8	20	4.9	17.7	35.3	36.447
137	Rwanda	20.4	1807	6.9	2.5	6.25	30	2.4	30	7.4	34.6	14.9	27.194
138	Saint. Lucia	2	11738.8	1.6	-0.7	8.86	30	20.1	30	9.2	83	23.7	31.841
139	Saint. Vincent and the Gren	1.2	10956.4	1.6	-1.7	9.5	32.5	20	32.5	12.4	73.6	23.9	28.662
140	Samoa	1	5174.1	1.7	0.9	6.45	27	5.8	27	9.7	55.5	22.4	39.223
141	São Tomé and Príncipe	0.7	3243.8	4	5.3	10	20	14	25	9.1	82.5	14.1	35.24
142	Saudi Arabia	1683	53624.4	3.4	2.2	2	2.5	5.8	2.5	3.4	5.8	4.6	40.885
143	Senegal	36.7	2451.3	6.5	0.1	4.5	40	9.3	30	8.5	56.8	20.2	29.877
144	Serbia	97.5	13671.4	0.7	1.4	4	15	19	15	6.1	77.4	35	44.74
145	Seychelles	2.4	26276.7	4.4	4	12.38	15	4.5	33	3.3	68.1	28.4	32.292
146	Sierra Leone	10	1577.2	-21.5	9	12	30	3.4	30	10.3	46.1	8.3	20.263
147	Singapore	471.9	85253.2	2	-0.5	0.45	22	3.3	17	0	98.2	13.4	20.29
148	Slovak Republic	161	29720.1	3.6	-0.3	0	25	11.3	21	1.5	52.6	31	43.247
149	Slovenia	64	31007.4	2.9	-0.5	0	50	9.3	17	1.5	83.3	36.6	44.114
150	Solomon Islands	1.1	1950.2	3.3	-0.4	9.38	40	34.8	30	8.5	10.4	35	46.391
151	South Africa	723.5	13165.2	1.3	4.6	7	41	25.1	28	3.9	50.1	22.6	33.705
152	Spain	1615.1	34819.5	3.2	-0.5	0	45	21.9	25	1.5	99	33.2	43.017
153	Sri Lanka	223	10566.2	5.2	0.9	7.25	24	4.7	28	5.3	74.4	10.7	18.258
154	Sudan	167	4343.8	3.5	16.9	13.1	10	13.6	35	14.7	68.9	5.2	12.479
155	Suriname	9.1	16292	0.1	6.9	12.5	38	7.8	36	10.8	43.3	15.7	29.729
156	Swaziland	10.8	8453.4	1.7	5	7.25	33	25.6	27.5	0.6	17.4	26	33.067
157	Sweden	473.4	47922.2	4.1	0.7	-0.5	57	7.4	22	1.5	44.1	42.7	49.265
158	Switzerland	482.3	58551.5	0.9	-1.1	-0.75	11.5	4.3	8.5	0	45.6	27.1	32.971
159	Syria	68	4684.72	-4.8	29.6	-1.61	22	14.9	28	14.2	57.5	10.4	12.3
160	Taiwan	1099	46783	0.7	-0.3	1.38	45	3.8	17	1.9	38.3	12.3	18.329
161	Tajikistan	23.3	2749.4	3	5.8	16	13	10.9	15	5.6	35.9	22.8	31.975
162	Tanzania	138.5	2904	7	5.6	12	30	3.2	30	7	40.5	13.2	18.821
163	Thailand	1108.1	16097.4	2.8	-0.9	1.5	35	1.1	20	3.6	43.1	16.5	22.348
164	Timor-Leste	6.6	5628.5	4.3	0.6	14.8	10	4	10	2.5	0	61.5	51.16
165	Togo	10.8	1483.4	5.3	1.8	4.5	45	7.7	27	9.4	61.9	20.7	27.382
166	Tonga	0.5	5044.9	2.6	-0.1	6.72	20	5.2	25	5.2	49	17	32.948
167	Trinidad and Tobago	44.3	32635.5	-1.8	4.7	4.75	25	3.8	25	5.7	51.1	24.7	39.819
168	Tunisia	127	11428.2	0.8	4.9	4.75	35	14.8	30	13.1	54.5	22.5	27.689
169	Turkey	1588.8	20437.8	3.8	7.7	8	35	10.3	20	2.8	32.6	28.7	36.959
170	Turkmenistan	88.6	16444.5	6.5	5.5	0	10	10	8	0	23.3	17.4	16.188
171	Uganda	79.9	2002.6	5	5.8	11	40	3.6	30	5.9	35.4	11.4	18.017
172	Ukraine	339.5	7970.8	-9.9	48.7	13	20	9.9	18	2.1	80.2	37.6	43.241
173	United Arab Emirates	647.8	67616.9	3.9	4.1	1.75	99	3.7	0	3.2	19.4	19	35.865
174	United Kingdom	2679.3	41158.9	2.2	0.1	0.25	45	5.5	20	1.5	89.3	32.6	43.201047
175	United States	17947	55805.2	2.4	0.1	1	39.6	5.3	35	1.4	105.8	26	37.840003
176	Uruguay	73.5	21506.5	1.5	8.7	9.25	30	7.3	25	4.7	61.8	26.9	31.837
177	Uzbekistan	187.9	6068.4	8	8.5	9	22	10.1	7.5	6.6	10.7	19.7	34.398
178	Vanuatu	0.7	2549.6	-0.8	3.3	3.63	0	4.3	0	5.5	20.5	17.2	32.928
179	Venezuela	515.7	16672.7	-5.7	121.7	21.46	34	8	34	9.7	48.8	20.9	41.032
180	Vietnam	552.3	6024.4	6.7	0.6	6.5	35	2.1	22	3.4	59.3	18.2	28.665
181	Yemen	75.5	2670.6	-28.1	30	15	20	15.9	20	4.1	68.6	5.3	24.039
182	Zambia	62.7	3868.1	3.6	10.1	12.5	35	10.7	35	3.4	52.9	15.4751	25.57
183	Zimbabwe	28.1	2096.2	1.5	-2.4	10.59	51.5	9.3	25	13.6	53	24.8	28.506
184	Kosovo	17.4	9255	3.3	-0.5	7.92	10	30.9	10	7.1	19.2	21.1	27.376
185	Brunei Darussalam	33.2	79587	-0.2	-0.4	5.5	0	1.9	18.5	0.5	3.1	33.1	40.416

Appendix B: Matlab Code for Relationship between PCA and FA



```
>> %The following command is used to import the data file from excel in matlab.
[num,txt,raw]=xlsread('data.xlsx'); %Import from excel in matlab.
X=num(:,3:end);%Data Matrix
[n,p]=size(X);% size of X
Y = X-ones(n,1)*mean(X); % Mean deviation matrix.
Z = Y*inv(diag(std(X)));%standardized data matrix with means zero,and variance 1
pause
[ 'Inter-correlations among the tewlve world economic variables ' ];
R = corrcoef(X);%Correlation matrix.
DR= det(R);%Determinant  of R.
pause
[ 'Singular value decomposition of correlation matrix ' ];
[ V , D , V ] = svd(R);% for order Eigen values matrix and Eigen matrix
P=D*((sum(D)).^-1)*100;% percentage variance explained
C=cumsum(P);%Cumulative percentage variance explained.
pause
[ 'Scree plot.' ];
figure()
pareto(P)
xlabel('Principal Component')
ylabel('Variance Explained (%)')
pause
[ 'Reduce eigen space.' ];
V = V (:,1:4) ; % keep only first four eigen vectors and eigen values.
D = D(1:4,1);%reduce Eigen value matrix
[ ' Initial Factor loadings matrix using PCA ' ]
pasue
Iding = V*sqrt(diag(D));
pasue
CN=diag(Iding* Iding');% communalities.
UQ=1-CN;%Uniqueness or specific variances.
pause
```



```
CN=diag(Iding* Iding');% communalities.
UQ=1-CN;%Uniqueness or specific variances.
VM=rotatefactors(Iding);%varimax rotation factor loadings.
RL=VM*inv(VM'*VM )% varimax rotated factor loadings matrix.
F =Z*Iding*inv(Iding'*Iding );% unrotated estimated factor scores .
F2 =Z*VM*inv(VM'*VM );%varimax rotated estimated factor scores.
save
return
```