Significance of the Covariance Matrix in Principal Component Analysis

Yves Yannick Yameni Noupoue

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

Master of Science
in
Mathematics

Eastern Mediterranean University
August 2015
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Serhan Çiftçioğlu
Acting Director

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

Asst. Prof. Dr. Mustafa Kara
Acting 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 Mathematics.

Asst. Prof. Dr. Yücel Tandoğdu
Supervisor

Examining Committee

1. Prof. Dr Agamirza Bashirov

2. Asst. Prof. Dr. Nidai Şemi

3. Asst. Prof. Dr. Yücel Tandoğdu

ABSTRACT

In all the scientific fields, scientist usually deal with big data. Statistical Data Analysis is therefore used to manage data. Depending on the nature of the experiment, its output can be analyzed using univariate, bivariate or multivariate statistics. In the multivariate case when the number of variables is very large, it sometime wise to reduce the number of variable to optimize the analysis of the data. Dimension reduction is used to reduce the number of variables which is also the size of data. In this work, on method of dimension reduction called Principal Component Analysis (PCA) is discussed. The PCA is a method which is based mainly on two matrices , covariance-variance matrix and correlation coefficient matrix obtained from the data. From the mentioned matrices, using the eigenvalues and corresponding eigenvectors, linear combination of the variables called principal components (PC) are established. It is important to mentioned that for the same set of data, the PCs computed using the covariance-variance matrix are different from those computed using the correlation coefficient matrix. The core topic in this work is to studied the conditions under which it is better to use either covariance matrix or correlation coefficient matrix for the PCs computation.

Keywords: Principal Component Analysis (PCA), Principal Components (PCs), Dimension Reduction, Variance-covariance matrix, Correlation Coefficient Matrix

ÖZ

Bilmin hemen her dalında bilim insanları büyük verilerin analizi ile uğraşmak durumundadır. İstatistiki veri analizi verilerin değerlendirilmesinde kullanılır. Deneyin doasına bağlı olarak, elde edilen veriler, tek veya çok değişkenli istatistik yöntemlerle değerledirilebilir. Değişken sayısının çok fazla olduğu durumlarda, daha hızlı analiz imkanını elde etmek için boyut indirgemesi yapılabilir. Bu amaçla Temel Bileşenler Analizi (TBA) yöntemi kullanılır. TBA metodu verinin kovaryans veya korelasyon matrislerine bağımlı bir sistemdir. Bu matrislerin özdeğer ve özvektörlerinden yararlanarak, Temel Bileşenler (TB) denen değişkenlerin lineer kombinasyonları oluşturulur. Ancak kovaryans ve korelasyon matrisleri kullanılarak oluşturulan TB ler, bir birinden farklıdır. Bu çalışmanın temel amacı, hangi şartlar altında kovaryans veya korelasyon matrislerinin kullanılabileceğinin incelenmesidir.

Anahtar kelimeler: Temel bileşenler analizi (TBA), Temel Bileşenler (TB), özdeğer, özvektör, tekil değer ayrışımı (TDA), kovaryans, korelasyon.

DEDICATION

I am dedicating this thesis to my family

ACKNOWLEDGMENT

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

A special thank to my parents Jean Noupoue and Matilde Noupoue Kanyep.

A special thanks goes to the following family members Noupoue's family Cameroon; Kouembitie's France; Ntchankwe's family Cameroon, without whom I couldn't probably be able to carry my studies up to this stage.

I would like to say thanks to my brother Seve Landry Nguematcha Noupoue and sisters Marie-Therese Ngamakoua Noupoue; Justine Fidéline Tcheumeni Noupoue and my twins sister Nadège Deumeni Noupoue for the love and support they have given to me.

Thanks to Claude Martial Tanguep; Pauline Milaure Ngugnie Diffouo whom have tactically contributed in my progress over the past five years.

Thanks to Alma Krivdic for the studies time we had during our Master program.

I am grateful to Harold Assam Egodji and Abdullah Jangeer who have been very helpful during my studies in Gazimağusa, North Cyprus.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

λ	Eigenvalue
$\mathbf{e}$	Eigenvector
$\bar{x}$	Sample mean
$\bar{\mathbf{x}}$	Sample mean vector
μ	Population mean
$\boldsymbol{\mu}$	Population mean vector
$\boldsymbol{\Sigma}$	population covariance matrix
$\mathbf{S}$	sample covariance matrix
$\boldsymbol{\rho}$	Population correlation coefficient matrix
$\mathbf{R}$	Sample correlation coefficient matrix
$S^{\perp}$	Orthocomplement of a subset S.
$\mathbf{X}'$	Transpose of a vector $\mathbf{X}$.
σ	Standard deviation
Λ	Diagonal matrix of eigenvalues
ρ_{Y_i,X_j}	Correlation between the i^{th} PC Y_i and its j^{th} variable X_j.
$Y_i^{\mathbf{S}}$	ith Principal Component computed using a sample covariance matrix $\mathbf{S}$.
$Y_i^{\mathbf{R}}$	ith Principal Component computed using a sample correlation matrix $\boldsymbol{\rho}$.
PC	Principal Components
PCA	Principal Components Analysis
SVD	Singular Value Decomposition

Chapter 1

INTRODUCTION

In univariate statistic, inference and analysis are based on a single variable data collection. There exits experiment where more than one variable are observed simultaneously. Analysis of such data requires the use of multivariate statistic. Concepts used in univariate statistical analysis can be extended to the multivariate case. The multivariate statistic was historically used for behavioral and biological sciences, but nowadays, its application is found in many other fields of science. Thereby, the multivariate statistics is used in the fields of broadcasting, linguistics, medicine, data mining, mining, psychology and many other areas. In multivariate statistics, all variables are observed simultaneously forming a data matrix with n rows and p columns. Columns represent the variables. The aim of putting those variables together is to enable the processing of such data in a multivariate environment. In this thesis, the method called principal component analysis (PCA) is used to compute linear combinations of variables of interest. These linear combinations are called principal components (PC). The first few (k) PCs represent a high percentage of variation in the original data. Then, data analysis can proceed using the k PCs. Therefore, a dimension reduction is achieved in data analysis. Consider a data matrix of size $n \times p$ representing n observations of p variables. The $p \times p$ covariance or correlation matrix obtained from the data is then used to determine the PCs [7]. Possible number of PC is the same as the number of variable of the data set. Using the PCA, only a few of the PCs $(k \leq p)$ can be used to

adequately explain the total variation in the dataset. It can also help to detect variables of low significance to the process under study. Therefore variables with very low significance can be neglected.

Chapter 2

LITERATURE REVIEW

The first idea of PCA comes from Karl Pearson in 1901. He worked on the geometrical representation of a multivariate data, in a coordinate system. He has established that if the data being processed is univariate, it can be represented in a plane. When the number of variables increases, the data can be represented in a 3-dimensional or even *n*-dimensional space depending on the number of variables. His worked was published in an article named "On Lines and Planes of closet Fit to System of Points in Space. By KARL PEARSON, F.R.S., University College London." The following important result "The line which represented best a system of *n* points in a q-fold space is the line passes through the centroid of the system and which coincides in direction with the least axis of ellipsoid of residuals" which is mainly used in PCA is found in the mentioned article [2].

The PCA method was later developed and named in 1933 by Harold Hotelling [3]. Due to the high dimension of the data processed in PCA, the manual computation is difficult. Therefore, the PCA method hasn't been used widely from the beginning until the appearance of electronics computers and statistical software which can enable the processing of high dimensional data within few second.

From the year 1936 to the year 1946, the American statistician Girshick, Meyer A dedicated his work in the fields of multivariate statistic. His achievement was to

determine the distribution of the squared root as well as characteristic vector which are associated to equations used for testing null hypothesis concerning independence of two sets of variables. The mentioned achievement concerning multivariate statistic and principal components analysis was established in 1939 [4].

In 1963 Anderson T.W has contributed to the development of the fields of principal components analysis. His achievement is the study of the asymptotic properties of the characteristic roots. He established from a covariance matrix that, the characteristics roots are variances and the coefficient of their corresponding characteristic vectors are the principal components coefficients. He also introduced the computation of confidence interval and the hypothesis test of equality of two population roots which are important in the analysis of the principal component significance. He established all the previous results on correlation coefficient matrix as well [3].

In 1964, Rao.C.R contributed in the fields of principal components analysis. He studied the means to introduce more information from the computation of principal components [5].

In 1966, J.C Gower work was based on the study of relation between various statistical techniques and the principal component analysis method [12].

In 1967, Jeffers contribution in the fields of principal components analysis was mainly concerned by the analysis and interpretation of eigenvalues and eigenvectors. He also focused on plotting principal components scores for further analysis, stating eight practical purposes the principal components analysis is used for. Those are the

correlation examination between variables coming from two different sets. The high variability dimension reduction from a set. Discard of variables with lower contribution in a set. Examination of grouping individual in an n –dimensional state. Determination of variable weights. Allocation of individual to a group. Recognition of individual. Regression calculation and orthogonalization [4].

In 1974, Baxter showed that computer graphics facilitates the understanding of principal components scores [12].

In 1982, the regression method was introduced in the fields of principal components analysis by Joliffe with the name of principal components regression [3].

In1997 , Takane and Shibayama developed the concept of Constrained Principal Component Analysis [27].

In 2002, Fotheringham and his coworkers introduced the concept of locally weighted principal components (LWPCA) and the concept of geographically weighted principal components (GWPCA) [28].

Chapter 3

ALGEBRA AND STATISTICS CONCEPTS

The core of our topic is dimension reduction, for a given high dimensional data without losing inherent message carried by the data. Dimension reduction is done by combining various concepts of mathematics. This ranges from basic algebra concepts and statistical interpretation related with data. In this chapter, algebraic topics related with dimension reduction and their implementation to statistics will be discussed.

3.1 Algebraic Concepts

Dimension reduction is done using some basic and advanced algebraic concepts. This section is a review of fundamental algebraic operations and matrices which are useful in data representation and computation [25].

3.1.1 Fields

Definition 3.1: A Field K is a set on which we can define the following two operations. $\oplus$ (addition) and $\cdot$ (multiplication)such that the following conditions hold for any a,b,c given in K:

1 $a \oplus b = b \oplus a$ and $a \cdot b = b \cdot a$ (commutativity)

2. $(a \oplus b) \oplus c = a \oplus (b \oplus c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ (associativity of $\oplus$ and $\cdot$)

3. Existence of identities elements 0 and 1 for addition and multiplication respectively are defined as,if there are two distinct elements in K say 0 and 1 such that satisfies $0 \oplus a = a$ and $1 \cdot a = a$.

4. For each element $a \in K$, and for each element $b \in K$, $b \neq 0$; there exist c and d in K such that $a \oplus c = 0$ and $b \cdot d = 1$ (existence of inverses c and d for addition and multiplication respectively)

5. $a \cdot (b \oplus c) = a \cdot b \oplus a \cdot c$ (distributivity of $\cdot$ over $\oplus$)

For example if $\mathbb{R}$ is the real numbers set, with the usual addition (+) and the usual multiplication ($\times$), then $\mathbb{R}$ is a field.

In what will follow, the most useful field is $\mathbb{R}$. Therefore, $\mathbb{R}$-vector space can be used instead of K-vector space[11].

3.1.2 Vectors

In multivariate data analysis, there is a collection of *n* observations of *p* variables. The observed *p* variables are represented in an arrangement of *p* real values forming a vector called a trajectory. This vector is also called *p*-variate response. Let's denote the i^{th} observation by x_i, where *i= 1,…,p,* then the $p \times 1$ vector is denoted by $\mathbf{x}$ and represented as follow: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ which is a vector of *p* lines and one column.

The transpose of $\mathbf{x}$ is denoted by $\mathbf{x}'$ and is represented by: $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}$.

$\mathbf{x}$ is called a column vector, whereas $\mathbf{x}'$ is called a row vector. The row vector $\mathbf{x}'$ is also called the transpose of the column vector $\mathbf{x}$. The index *p* which represents the number of components in the vector $\mathbf{x}$ is called the order or the dimension of the vector $\mathbf{x}$. Geometrically, $\mathbf{x}$ with its *p* elements is the representation of a point in a *p*-dimensional Euclidean space [18].

Definition 3.2 An ordered set of *p* real numbers, representing a position in a

p-dimensional Euclidean space V_p is called a vector and denoted by $\mathbf{x}_{p\times 1}$.

3.1.3 Vectors Spaces

A real vector space is a collection of $n\times 1$ vectors in a Euclidean space V_n which is closed under the following two vector operations, scalar multiplication and addition.

Definition 3.3 Let K be a given field. A collection of vectors of a set V_n satisfying the following condition is called a n-dimensional vector space over the field $K=\mathbb{R}$ [11].

$$\forall\ \mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\in V_n \quad \mathbf{y}=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\in V_n \ \mathbf{x}+\mathbf{y}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}+\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}=\mathbf{z}=\begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}\in V_n \tag{3.1.1}$$

,

$$\forall\lambda\in\mathbb{R}\ ,\ \forall\ \mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\in V_n\ ,\ \lambda\mathbf{x}=\lambda\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}=\begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{bmatrix}\in V_n \tag{3.1.2}$$

Let's consider for example, $C\left(\left[0,1\right],\mathbb{R}\right)$, the set of continuous functions from $\left[0,1\right]$ into $\mathbb{R}$. If f and g are two functions from $C\left(\left[0,1\right],\mathbb{R}\right)$, and assuming that $\forall x\in\left[0,1\right],\ \ (f+g)(x)=f(x)+g(x)$ and $(\lambda f)(x)=\lambda f(x)$ then $\mathrm{C}\left(\left[0,1\right],\mathbb{R}\right)$ is a $\mathbb{R}$-vector space.

3.1.4 Vectors Subspaces

Definition 3.4 Consider a vector space V_n, a subset S of V_n (i.e. $S\subseteq V_n$) is said to be a vector subspace of V_n if the following hold [9;11]

- $0\in S$
- $\forall\ \mathbf{x},\mathbf{y}\in S,\ \ \mathbf{x}+\mathbf{y}\in S$

- $\forall\, \mathbf{x} \in S$ and $\forall \lambda \in \mathbb{R},\ \ \lambda \mathbf{x} \in S$

Examples:

- $\{0\}$ and V_n are subspaces of V_n
- Let $\mathbb{R}[x]$ be the set of polynomial with their coefficients in $\mathbb{R}$. The set $\mathbb{R}_n[x]$ of polynomial with power less or equals to n is a subspace of $\mathbb{R}[x]$

Definition 3.5 Let U be a $\mathbb{R}$-vector space and Let $V_1, V_2, ..., V_k$ be subspaces of U. The following statement holds.

The summation $V_1 + V_2 + ... + V_k$ is a subspace of U.

3.1.5 Bases

Definition 3.6 Let V be a $\mathbb{R}$-vector space. Let $\mathbf{v}_1, ..., \mathbf{v}_k$ be a set of vectors from V.

The subspace of V spanned by $\mathbf{v}_1, ..., \mathbf{v}_k$ is

$$span(v_1, ..., v_k) = \left\{ \sum_{i=1}^{k} \alpha_i \mathbf{v}_i, \quad \alpha_i \in \mathbb{R}, \ \forall 1 \le i \le k \right\} \tag{3.1.3}$$

Theorem 3.1 $span(\mathbf{v}_1, ..., \mathbf{v}_k)$ is a vector subspace of V.

Theorem 3.2 Let V be a $\mathbb{R}$-vector space, $\forall (\mathbf{v}_1, ..., \mathbf{v}_k) \in V$,

$$span(\mathbf{v}_1, ..., \mathbf{v}_k) = span(\mathbf{v}_1) + \cdots + span(\mathbf{v}_k) \tag{3.1.4}$$

Definition 3.7 Let V be a $\mathbb{R}$-vector space. A set $\mathbf{v}_1, ..., \mathbf{v}_k$ of vectors from V is said to be linearly independent if and only if:

$$\sum_{i=1}^{k} \alpha_i \mathbf{v}_i = 0 \Rightarrow \alpha_i = 0, \ \forall \ 1 \le i \le k \tag{3.1.5}$$

Example 3.1: Consider the following vectors of $\mathbb{R}^3$ and check whether they are linearly independent or linearly dependent. $\mathbf{u}_1 = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}, \mathbf{u}_3 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$.

Solution: To check whether these vectors are linearly dependent or independent, let's solve the equation $\alpha_1 \mathbf{u_1} + \alpha_2 \mathbf{u_2} + \alpha_3 \mathbf{u_3} = 0$ for α_1, α_2 and α_3.

$$\alpha_1 \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \alpha_1 \\ -\alpha_1 \\ 2\alpha_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 2\alpha_2 \\ 1\alpha_2 \end{bmatrix} + \begin{bmatrix} \alpha_3 \\ 0 \\ -\alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The previous is a system of three equation in three unknown α_1, α_2 and α_3

$$\begin{aligned} (1) \quad & \alpha_1 + \quad \alpha_3 = 0 \\ (2) \quad & -\alpha_1 + 2\alpha_2 \quad = 0 \\ (3) \quad & 2\alpha_1 + \alpha_2 - \alpha_3 = 0 \end{aligned}$$

From (1) : $\alpha_1 = -\alpha_3$; (1) in (3) : $\alpha_2 = -3\alpha_1$ and from (2) : $\alpha_2 = 0$ by substitution we have $\alpha_1 = -\alpha_3 = 0$ Such that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. This is the unique solution of the system. So the vectors $\mathbf{u_1}, \mathbf{u_2}$ and $\mathbf{u_3}$ are linearly independents

The previous definitions and theorems lead us to the definition of a base.

Definition 3.8 Let V be a $\mathbb{R}$-vector space. Let $\{\mathbf{v}_1,...,\mathbf{v}_k\}$ be a subset of vectors of the vector space V. $\{\mathbf{v}_1,...,\mathbf{v}_k\}$ is a basis of V if $\{\mathbf{v}_1,...,\mathbf{v}_k\}$ is linearly independent and generates V. That is if $span(\mathbf{v}_1,...,\mathbf{v}_k)=V$ and $\{\mathbf{v}_1,...,\mathbf{v}_k\}$ is linearly independent. Furthermore, the integer k is called the rank or dimension of the vector space V [10].

Theorem 3.3

- If V is a vector space, then V has a basis
- Let V be a vector space, let $\{\mathbf{v}_1,...,\mathbf{v}_k\}$ be a basis of V. Then

$$\forall \mathbf{u} \in V, \exists!\ (\alpha_1,...,\alpha_k) \text{ such that } \mathbf{u}=\alpha_1\mathbf{v_1}+...+\alpha_k\mathbf{v_k} \tag{3.1.6}$$

- Let V be a vector space, if β and δ are two bases of V, then β and δ have same number of vectors.

3.1.6 Vectors Norms

Multivariate statistics deals with multivariate observation. The knowledge of the length of a vector and the angle between two vectors helps determine the relationship between the observations.

Definition 3.9 Let V be a*n*- dimensional vector space, let $\mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$ and $\mathbf{y}=\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}$ be two vectors of V. The inner product of $\mathbf{x}$ and $\mathbf{y}$ is the scalar computed as follow

$$\mathbf{x}'\mathbf{y}=\begin{bmatrix} x_1 & \cdots & x_k \end{bmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}=\sum_{i=1}^{k} x_i y_i; \quad k \le n; \tag{3.1.7}$$

The vectors $\mathbf{x}$ and $\mathbf{y}$ must have the same size; i.e. same numbers of elements.

In what will follow, the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ will be denoted by $<\mathbf{x},\mathbf{y}>$ such that $<\mathbf{x},\mathbf{y}>=\mathbf{x}'\mathbf{y}$

Theorem 3.4 Let V be a vector space over a field $\mathbb{R}$. Let $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}\in V$ and let $\alpha,\beta\in\mathbb{R}$. The following relationships are satisfied by the inner product

- $\mathbf{x}'\mathbf{y}=\mathbf{y}'\mathbf{x}$
- $\mathbf{x}'\mathbf{x}\geq 0$ and $\mathbf{x}'\mathbf{x}=0$ if and only if $\mathbf{x}=0$
- $(\alpha\mathbf{x})'(\beta\mathbf{y})=\alpha\beta(\mathbf{x}'\mathbf{y})$
- $(\mathbf{x}+\mathbf{y})'\mathbf{z}=\mathbf{x}'\mathbf{z}+\mathbf{y}'\mathbf{z}$
- $(\mathbf{x}+\mathbf{y})'(\mathbf{w}+\mathbf{z})=\mathbf{x}'(\mathbf{w}+\mathbf{z})+\mathbf{y}'(\mathbf{w}+\mathbf{z})$

Definition 3.10 From the computation formula of inner product given by the formula (3.1.7) if $\mathbf{x}=\mathbf{y}$ then we have

$$\mathbf{x}'\mathbf{x}=\begin{bmatrix}x_1 & \cdots & x_k\end{bmatrix}\begin{bmatrix}x_1\\ \vdots\\ x_k\end{bmatrix}=\sum_{i=1}^{k}x_i^2 \tag{3.1.8}$$

The scalar $(\mathbf{x}'\mathbf{x})^{1/2}=\sqrt{\sum_{i=1}^{k}x_i^2}$ is called the length of the vector $\mathbf{x}$ or the Euclidean vector Norm of $\mathbf{x}$, and denoted by $||\mathbf{x}||$. It follows that $||\mathbf{x}||^2$ is the norm square of $\mathbf{x}$.

The Euclidean distance or the length between two vectors $\mathbf{x}$ and $\mathbf{y}$ from the vector space V is given by $||\mathbf{x}-\mathbf{y}||=\left[(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})\right]^{1/2}$ (3.1.9)

Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors of a vector space V and let θ be the angle between $\mathbf{x}$ and $\mathbf{y}$. The inner product of $\mathbf{x}$ and $\mathbf{y}$ is also defined by $\mathbf{x}'\mathbf{y}=||\mathbf{x}||\,||\mathbf{y}||\cos\theta$.

Thus $\cos\theta=\dfrac{\mathbf{x}'\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||}$. (3.1.10)

Here the angle θ is such that $0° \le \theta \le 180°$

Theorem 3.5 Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors of a vector space V. $\mathbf{x}$ and $\mathbf{y}$ are said to be orthogonal if their inner product is zero $< \mathbf{x}, \mathbf{y} >= 0$.

Proof: if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal then $\theta = 90°$ and $\cos\theta = \cos 90° = 0$ it follows from the formula $\mathbf{x}'\mathbf{y} = ||\mathbf{x}||\,||\mathbf{y}||\cos\theta$ that $\mathbf{x}'\mathbf{y} = 0$

Definition 3.11 A vector with length 1 is called a unit vector or a normalized vector.

Theorem 3.6 In a vector space, any nonzero vector $\mathbf{x}$ can be normalized by

$$\mathbf{x}_{unit} = \frac{\mathbf{x}}{||\mathbf{x}||}, \tag{3.1.11}$$

where $\mathbf{x}_{unit}$ stands for unit vector or normalized vector obtain from $\mathbf{x}$.

To prove theorem 3, the following lemma should be considered.

Lemma Let V be a vector space over the field $K = \mathbb{R}$. Let $\mathbf{u} \in V$ and $\alpha \in \mathbb{R}$. The following relation holds $||\alpha\mathbf{u}|| = |\alpha| . ||\mathbf{u}||$ (3.1.12)

Proof of Lemma $||\alpha\mathbf{u}||^2 = (\alpha\mathbf{u}')(\alpha\mathbf{u}) = \alpha\alpha(\mathbf{u}'\mathbf{u}) = |\alpha|^2 ||\mathbf{u}||^2$ (3.1.13)

Considering the square root of the formula (3.1.13), we find $||\alpha\mathbf{u}|| = |\alpha| \cdot ||\mathbf{u}||$

Proof: Let $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$ be a vector of the vector space V . The Euclidean distance or the length or the norm of $\mathbf{x}$ is $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{k} x_i^2}$. Let's denoted by $\mathbf{x}_{unit}$ the normalized

vector computed from $\mathbf{x}$ and prove that $\mathbf{x}_{unit}$ has the length 1.

$$\mathbf{x}_{unit} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{1}{\sqrt{\sum_{i=1}^{k} x_i^2}} \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \text{ it follows that } \|\mathbf{x}_{unit}\| = \frac{1}{\sqrt{\sum_{i=1}^{k} x_i^2}} \|\mathbf{x}\| = \frac{\sqrt{\sum_{i=1}^{k} x_i^2}}{\sqrt{\sum_{i=1}^{k} x_i^2}} = 1$$

Example 3.2 Let's consider the following two vectors $\mathbf{u} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$ of a

3-dimensional vector space over the field $\mathbb{R}$. Then let's compute the following.

The length of the vectors $\mathbf{u}$ and $\mathbf{v}$

$$\|u\| = \sqrt{1^2 + (-1)^2 + 2^2} = \sqrt{6} \text{ and } \|v\| = \sqrt{3^2 + 0^2 + 1^2} = \sqrt{10}$$

The distance between $\mathbf{u}$ and $\mathbf{v}$

$$\|u - v\| = \left[(u-v)' (u-v) \right]^{1/2} = \sqrt{(1-3)^2 + (-1-0)^2 + (2-1)^2} = \sqrt{6}$$

The inner product of $\mathbf{u}$ and $\mathbf{v}$

$$u'v = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} = 1 \cdot 3 - 1 \cdot 0 + 2 \cdot 1 = 5$$

The angle between $\mathbf{u}$ and $\mathbf{v}$

Let θ be that angle. $\cos\theta = \dfrac{\mathbf{u}'\mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} = \dfrac{5}{\sqrt{6}\sqrt{10}} \approx 0.645$ thus $\theta = \cos^{-1}\theta \approx 50°$

3.1.7 Orthogonal Basis

Let's consider the usual inner product defined on the canonical basis of $\mathbb{R}^2$, $(\mathbf{e}_1, \mathbf{e}_2)$ or even defined on the canonical usual basis of $\mathbb{R}^n$, $(\mathbf{e}_1, ..., \mathbf{e}_n)$. The following relations hold from those bases:

- $\mathbf{e}_1'\mathbf{e}_1 = 1$, $\mathbf{e}_1'\mathbf{e}_2 = 0$ and $\mathbf{e}_2'\mathbf{e}_1 = 0$ hold in $\mathbb{R}^2$ (3.1.14)

- $\mathbf{e}_i'\mathbf{e}_j = \begin{Bmatrix} 1 \text{ if } i = j \\ 0 \text{ else} \end{Bmatrix}$ hold in $\mathbb{R}^n$ (3.1.15)

$(\mathbf{e}_1, \mathbf{e}_2)$ of $\mathbb{R}^2$ and $(\mathbf{e}_1, ..., \mathbf{e}_n)$ of $\mathbb{R}^n$ are called orthogonal bases in this case. Furthermore, since each vector in the bases $(\mathbf{e}_1, \mathbf{e}_2)$ or $(\mathbf{e}_1, ..., \mathbf{e}_n)$ has the norm 1, there are called orthonormal bases [10].

The idea behind orthogonal basis is to be able for a given *n*- dimensional vector space V and any abstract inner product defined on V, to build a basis of V with vectors of V which has same properties with the foregoing basis $(\mathbf{e}_1, ..., \mathbf{e}_n)$ of $\mathbb{R}^n$ [13].

Definition 3.12 Let V be a*n*- dimensional vector space. Let $\beta = (\mathbf{v}_1, ..., \mathbf{v}_n)$ be a basis of V. β is an orthogonal basis if $\begin{Bmatrix} \mathbf{v}_i'\mathbf{v}_j = 0 \text{ if } i \neq j \\ \mathbf{v}_i'\mathbf{v}_j \neq 0 \text{ if } i=j \end{Bmatrix}$, (3.1.16)

furthermore, if $\begin{Bmatrix} \mathbf{v}_i'\mathbf{v}_j = 0 \text{ if } i \neq j \\ \mathbf{v}_i'\mathbf{v}_j = 1 \text{ if } i = j \end{Bmatrix}$ (3.1.17)

then β is said to be an Orthonormal basis of V.

Theorem 3.7 (Pythagorean Theorem) Let V be a vector space over a field K on which an inner product is defined. $\forall \mathbf{u}, \mathbf{v} \in V$ Such that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, the following formula holds $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \|\mathbf{u} + \mathbf{v}\|^2$. (3.1.18)

Proof :

$$\|\mathbf{u} + \mathbf{v}\|^2 = <\mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v}> = <\mathbf{u}, \mathbf{u}> + <\mathbf{u}, \mathbf{v}> + <\mathbf{v}, \mathbf{u}> + <\mathbf{v}, \mathbf{v}>$$
$$= \|\mathbf{u}\|^2 + 0 + 0 + \|\mathbf{v}\|^2$$

Where $<\mathbf{u},\mathbf{v}>=<\mathbf{v},\mathbf{u}>=0$ because $\mathbf{u}$ and $\mathbf{v}$ are orthogonal. Thus

$\|\mathbf{u}\|^2+\|\mathbf{v}\|^2=\|\mathbf{u}+\mathbf{v}\|^2$

Reminder A *n*-dimensional vector space V over a field K on which an inner product is defined, is called an Euclidean Vector space if and only if the dimension $n<\infty$ and $K=\mathbb{R}$.

Theorem 3.8 If V is an Euclidean vector space, then V has an orthonormal basis.

Theorem3.8 tells us about the existence of an orthonormal basis for any Euclidean vector space. The next theorem is the procedure to obtain an orthonormal basis from any basis of the Euclidean vector space.

Theorem 3.9 (Gram – Schmidt process): Let V be a*n*-dimensional Euclidean vector space and let $(\mathbf{v}_1,...,\mathbf{v}_n)$ be a basis of$V$ [10]. An orthogonal basis $(\boldsymbol{\varepsilon}_1,...,\boldsymbol{\varepsilon}_n)$ of V is obtained from $(\mathbf{v}_1,...,\mathbf{v}_n)$ by the following process:

$$\boldsymbol{\varepsilon}_1=\mathbf{v}_1$$

$$\boldsymbol{\varepsilon}_2=\mathbf{v}_2-\frac{<\mathbf{v}_2,\boldsymbol{\varepsilon}_1>}{<\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_1>}\boldsymbol{\varepsilon}_1$$

$$\boldsymbol{\varepsilon}_i=\mathbf{v}_i-\frac{<\mathbf{v}_i,\boldsymbol{\varepsilon}_1>}{<\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_1>}\boldsymbol{\varepsilon}_1-\cdots-\frac{<\mathbf{v}_i,\boldsymbol{\varepsilon}_{i-1}>}{<\boldsymbol{\varepsilon}_{i-1},\boldsymbol{\varepsilon}_{i-1}>}\boldsymbol{\varepsilon}_{i-1},\quad (2\le i\le n)$$

Example 3.3 Let $\mathbf{u}_1=\begin{bmatrix}1\\1\\0\end{bmatrix},\mathbf{u}_2=\begin{bmatrix}0\\1\\1\end{bmatrix}$ and $\mathbf{u}_3=\begin{bmatrix}1\\0\\1\end{bmatrix}$. Be vectors in $\mathbb{R}^3$. The question here is to find if $(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ form a basis of $\mathbb{R}^3$. It means finding the orthogonal basis of $\mathbb{R}^3$ from $(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ by the Gram-schmidt process.

Solution:

Let's consider α_1,α_2 and α_3 in $\mathbb{R}$ and solve the system $\alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2+\alpha_3\mathbf{u}_3=0$

$$\alpha_1\mathbf{u_1}+\alpha_2\mathbf{u_2}+\alpha_3\mathbf{u_3}=0 \Leftrightarrow \alpha_1\begin{bmatrix}1\\1\\0\end{bmatrix}+\alpha_2\begin{bmatrix}0\\1\\1\end{bmatrix}+\alpha_3\begin{bmatrix}1\\0\\1\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix}\alpha_1\\\alpha_1\\0\end{bmatrix}+\begin{bmatrix}0\\\alpha_2\\\alpha_2\end{bmatrix}+\begin{bmatrix}\alpha_3\\0\\\alpha_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

The above leads us to

$$\begin{aligned}(1)\quad & \alpha_1+\quad\alpha_3=0\\(2)\quad & \alpha_1+\alpha_2\quad=0\\(3)\quad & \quad\alpha_2+\alpha_3=0\end{aligned}$$

From (1): $\alpha_1=-\alpha_3$, (1) in (2) and from (3) give $\alpha_1=\alpha_2=\alpha_3=0$, which is the unique solution of the previous system. This means $(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ is a linearly independent system of $\mathbb{R}^3$, thus $(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ is a basis of $\mathbb{R}^3$. Let's compute the orthogonal basis $(\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_3)$ of $\mathbb{R}^3$ from $(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ using Gram Schmidt process.

$\boldsymbol{\varepsilon}_1=\mathbf{u}_1=(1,1,0)$

$\boldsymbol{\varepsilon}_2=\mathbf{u}_2-\dfrac{<\mathbf{u}_2,\boldsymbol{\varepsilon}_1>}{<\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_1>}\boldsymbol{\varepsilon}_1$, where $<\mathbf{u}_2,\boldsymbol{\varepsilon}_1>=1,\ <\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_1>=2$

$\boldsymbol{\varepsilon}_2=(0,1,1)-\dfrac{1}{2}(1,1,0)=\left(\dfrac{-1}{2},\dfrac{1}{2},1\right)$

$\boldsymbol{\varepsilon}_3=\mathbf{u}_3-\dfrac{<\mathbf{u}_3,\boldsymbol{\varepsilon}_1>}{<\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_1>}\boldsymbol{\varepsilon}_1-\dfrac{<\mathbf{u}_3,\boldsymbol{\varepsilon}_2>}{<\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_2>}\boldsymbol{\varepsilon}_2$, where $<\mathbf{u}_3,\boldsymbol{\varepsilon}_1>=1,\ <\mathbf{u}_3,\boldsymbol{\varepsilon}_2>=\dfrac{1}{2},\ <\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_2>=\dfrac{3}{2}$

$\boldsymbol{\varepsilon}_3=\left(\dfrac{2}{3},-\dfrac{2}{3},\dfrac{2}{3}\right)$

It can easily be checked out that $(\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_3)$ is an orthogonal basis by computing the inner product of each pair of these vectors

$$<\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2>=(1,1,0)\begin{pmatrix}-1/2\\1/2\\1\end{pmatrix}=0,\ <\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_3>=(1,1,0)\begin{pmatrix}2/3\\-2/3\\2/3\end{pmatrix}=0,\ <\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_3>=\left(\frac{-1}{2},\frac{1}{2},1\right)\begin{pmatrix}2/3\\-2/3\\2/3\end{pmatrix}=0$$

An orthonormal basis $(\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3})$ can be computed from the previous orthogonal basis $(\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \boldsymbol{\varepsilon_3})$, by normalizing the vectors of the basis $(\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \boldsymbol{\varepsilon_3})$. We have

$\| \boldsymbol{\varepsilon_1} \| = \sqrt{2}, \ \| \boldsymbol{\varepsilon_2} \| = \frac{\sqrt{6}}{2}, \ \| \boldsymbol{\varepsilon_3} \| = \frac{2}{\sqrt{3}}$ It follows that

$$\mathbf{v_1} = \frac{\boldsymbol{\varepsilon_1}}{\| \boldsymbol{\varepsilon_1} \|} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0 \right), \ \mathbf{v_2} = \frac{\boldsymbol{\varepsilon_2}}{\| \boldsymbol{\varepsilon_2} \|} = \left(\frac{-1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}} \right), \ \mathbf{v_3} = \frac{\boldsymbol{\varepsilon_3}}{\| \boldsymbol{\varepsilon_3} \|} = \left(\frac{1}{\sqrt{3}}, \frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)$$

The basis $(\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3})$ is an orthonormal basis. It is proved by the verification of formula (3.1.17) as follow

$$\begin{cases} \|\mathbf{v}_1\| = \sqrt{\left(\frac{1}{\sqrt{2}} \right)^2 + \left(\frac{1}{\sqrt{2}} \right)^2 + 0^2} = 1 \\ \|\mathbf{v}_2\| = \sqrt{\left(\frac{-1}{\sqrt{6}} \right)^2 + \left(\frac{1}{\sqrt{6}} \right)^2 + \left(\frac{2}{\sqrt{6}} \right)^2} = 1 \\ \|\mathbf{v}_3\| = \sqrt{\left(\frac{1}{\sqrt{3}} \right)^2 + \left(\frac{-1}{\sqrt{3}} \right)^2 + \left(\frac{1}{\sqrt{3}} \right)^2} = 1 \end{cases}$$

$$\begin{cases} \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0 \right) \left(\frac{-1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}} \right)' = 0 \\ \langle \mathbf{v}_1, \mathbf{v}_3 \rangle = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0 \right) \left(\frac{1}{\sqrt{3}}, \frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)' = 0 \\ \langle \mathbf{v}_2, \mathbf{v}_3 \rangle = \left(\frac{-1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}} \right) \left(\frac{1}{\sqrt{3}}, \frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)' = 0 \end{cases}$$

3.1.8 Orthogonal Space

Some vectors space has particular properties which are important in multivariate statistics. Orthogonal space is a requirement for principal components analysis.

Definition 3.13 Let V be a*n*-dimensional vector space over the field K on which an inner product is defined. Let S be a subspace of V this means $S \subseteq V$. The

orthocomplement subspace of S in V is a subspace of V denoted by $S^{\perp}$ and defined by $S^{\perp} = \{\mathbf{v} \in V / <\mathbf{v},\mathbf{u}>=0, \forall \mathbf{u} \in S\}$. This means any vector of S is orthogonal to any vector of its orthocomplement subspace $S^{\perp}$. Then write $V = S \oplus S^{\perp}$ which means the vector space V is a direct sum of its subspaces S and $S^{\perp}$. The relation $V = S \oplus S^{\perp}$ is equivalent to the following two conditions when there are observed together $\dim(V) = \dim(S) + \dim(S^{\perp})$ and $S \cap S^{\perp} = \{0\}$.

Example 3.4 Consider the Euclidean vector space $\mathbb{R}^3$ with its orthogonal basis $(\mathbf{e_1},\mathbf{e_2},\mathbf{e_3})$, on which the usual inner product is defined. Let's consider $S \subset \mathbb{R}^3$ defined by $S = \{\mathbf{v} \in \mathbb{R}^3 \mid \mathbf{v} = \alpha \mathbf{e_1},\ \alpha \in \mathbb{R}\}$, the question is to compute the orthocomplement subspace of S

Solution Let $F = \{\mathbf{u} \in \mathbb{R}^3 \mid \mathbf{u} = \beta \mathbf{e_2} + \lambda \mathbf{e_3},\ \beta,\lambda \in \mathbb{R}\}$. Let $\mathbf{x}_1 \in S$ and $\mathbf{x}_2 \in F$. By the definition of the sets S and F, $\mathbf{x}_1 = \alpha \mathbf{e_1}$ and $\mathbf{x}_2 = \beta \mathbf{e_2} + \lambda \mathbf{e_3}$ where $\alpha, \beta, \lambda \in \mathbb{R}$.

$$
\begin{aligned}
<\mathbf{x_1},\mathbf{x_2}> &= <\alpha \mathbf{e_1}, \beta \mathbf{e_2} + \lambda \mathbf{e_3}> \\
&= <\alpha \mathbf{e_1}, \beta \mathbf{e_2}> + <\alpha \mathbf{e_1}, \lambda \mathbf{e_3}> \\
&= \alpha\beta <\mathbf{e_1},\mathbf{e_2}> + \alpha\lambda <\mathbf{e_1},\mathbf{e_3}> \\
&= \mathbf{0}
\end{aligned}
$$

From a random $\mathbf{x}_1 \in S$ and a random $\mathbf{x}_2 \in F$, we found that $<\mathbf{x_1},\mathbf{x_2}>=0$ this means $F = \{\mathbf{u} \in \mathbb{R}^3 \mid \mathbf{u} = \beta \mathbf{e_2} + \lambda \mathbf{e_3},\ \beta,\lambda \in \mathbb{R}\}$ is the orthocomplement of $S = \{\mathbf{v} \in \mathbb{R}^3 \mid \mathbf{v} = \alpha \mathbf{e_1},\ \alpha \in \mathbb{R}\}$ ie : $S^{\perp} = \{\mathbf{u} \in \mathbb{R}^3 / \mathbf{u} = \beta \mathbf{e_2} + \lambda \mathbf{e_3},\ \beta,\lambda \in \mathbb{R}\}$.

Furthermore $\dim(\mathbb{R}^3) = 3;\ \dim(S) = 1;\ \dim(S^{\perp}) = 2$ which leads us to $\dim(\mathbb{R}^3) = \dim(S) + \dim(S^{\perp})$ it is also obvious that $S \cap S^{\perp} = \{0\}$; actually let's assume that

$$
\begin{aligned}
S \cap S^{\perp} \neq \{\mathbf{0}\} &\Rightarrow \exists \mathbf{u} \in S \cap S^{\perp} \\
&\Rightarrow \mathbf{u} = \alpha \mathbf{e}_1 \text{ and } \mathbf{u} = \beta \mathbf{e}_2 + \lambda \mathbf{e}_3 \\
&\Rightarrow \alpha \mathbf{e}_1 = \beta \mathbf{e}_2 + \lambda \mathbf{e}_3 \\
&\Rightarrow \alpha = \beta = \lambda = 0 \\
&\Rightarrow \mathbf{u} = \mathbf{0}
\end{aligned}
$$

3.1.9 Orthogonal Projection

Theorem 3.10 (Orthogonal Projection) Let V be an *n*-dimensional vector space over a field K. Let E be a finite dimensional subspace of V. The following holds

$$\forall \mathbf{u} \in V,\ \exists!\ \mathbf{v} \in E \mid \|\mathbf{u} - \mathbf{v}\| = d(\mathbf{u}, E) = \inf_{\mathbf{z} \in E} \|\mathbf{u} - \mathbf{z}\|. \tag{3.1.19}$$

Here vector $\mathbf{v}$ is unique in E such that $\mathbf{u} - \mathbf{v} \in E^{\perp}$.

The vector $\mathbf{v}$ is called the orthogonal projection of the vector $\mathbf{u}$ over E.

3.1.1 Matrix

Multivariate data are usually observed in a form of a rectangular arrangement. The arrangement is of the size $(n \times p)$, where *n* is the number of observation in each of the *p* variables .

Definition 3.14 A matrix of size $(n \times p)$ with coefficients in $\mathbb{R}$ is an arrangement of elements of K in a form of *n* rows and *p* columns. A matrix of size $(n \times p)$ is represented by $\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}$ where $a_{ij} \in \mathbb{R},\ \forall i, j \in \mathbb{N}$

The elementary arithmetic of $\mathbb{R}$ is also applicable on matrices such that we can define equality of two matrices, addition of two matrices, and multiplication of two matrices.

Definition 3.15 A matrix $\mathbf{A}$ is said to be a square matrix if it is of size $(n \times n)$, this means if it has same number of rows and columns.

Let's consider the matrices

$$\mathbf{A}=\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}, \mathbf{B}=\begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{pmatrix}, \mathbf{C}=\begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{np} \end{pmatrix}$$

The equality between $\mathbf{A}$ and $\mathbf{B}$ is defined by $\mathbf{A}=\mathbf{B}\Leftrightarrow\forall i, j\in\mathbb{N},\ a_{ij}=b_{ij}$.

The addition of two matrices $\mathbf{A}$ and $\mathbf{B}$ is possible if and only if there are of same size. It is defined by

$$\mathbf{A}+\mathbf{B}=\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}+\begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{pmatrix}=\begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1p}+b_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1}+b_{n1} & \cdots & a_{np}+b_{np} \end{pmatrix} \quad (3.1.20)$$

The multiplication or inner product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is possible if and only if there are of size $(n\times p)$ and $(p\times m)$ respectively. This means the product $\mathbf{A}\times\mathbf{B}$ where $\mathbf{A}$ and $\mathbf{B}$ are of sizes $(n\times p)$ and $(p\times m)$ respectively is possible if the number *p* of columns of the matrix $\mathbf{A}$ is equals to the number *p*of rows of the matrix $\mathbf{B}$. For given $\mathbf{A}=\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{pmatrix}$ the product $\mathbf{A}\times\mathbf{B}$ is defined by

$$\mathbf{A}\times\mathbf{B}=\mathbf{C}=\begin{pmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nm} \end{pmatrix} \text{ where } c_{ij}=\sum_{k=1}^{p}a_{ik}b_{kj},\ 1\le i\le n,\ 1\le j\le m \quad (3.1.21)$$

Remark Generally $\mathbf{A}\times\mathbf{B}\neq\mathbf{B}\times\mathbf{A}$.

The square of a square matrix $\mathbf{A}$ is defined by $\mathbf{A}^2=\mathbf{A}\times\mathbf{A}$. The matrix $\mathbf{A}$ is idempotent if $\mathbf{A}^2=\mathbf{A}$.

Theorem 3.11 Consider the matrices $\mathbf{A,B,C,D}$ and the scalars α and λ. The following properties hold for matrix multiplication and addition

- $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$
- $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$
- $\alpha(\mathbf{A}+\mathbf{B})=\alpha\mathbf{A}+\alpha\mathbf{B}$
- $(\alpha+\lambda)\mathbf{A}=\alpha\mathbf{A}+\lambda\mathbf{A}$
- $(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC})$
- $\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}$
- $(\mathbf{A}+\mathbf{B})\mathbf{C}=\mathbf{AC}+\mathbf{BC}$
- $\mathbf{A}+(-\mathbf{A})=0$
- $\mathbf{A}+0=\mathbf{A}$
- $(\mathbf{A}+\mathbf{B})(\mathbf{C}+\mathbf{D})=\mathbf{A}(\mathbf{C}+\mathbf{D})+\mathbf{B}(\mathbf{C}+\mathbf{D})=\mathbf{AC}+\mathbf{AD}+\mathbf{BC}+\mathbf{BD}$

Definition 3.16 Consider a matrix $\mathbf{A}=\begin{pmatrix} a_{11} & a_{12}\cdots & a_{1p} \\ a_{21} & a_{22}\cdots & a_{2p} \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{np} \end{pmatrix}$, the transpose of the matrix $\mathbf{A}$ is the matrix obtained by changing rows of $\mathbf{A}$ into its columns or vice versa. It is denoted $\mathbf{A}'$ or $\mathbf{A}^T$. In this case, $\mathbf{A}'=\begin{pmatrix} a_{11} & a_{21}\cdots & a_{n1} \\ a_{12} & a_{22}\cdots & a_{n2} \\ \vdots & \vdots & \vdots \\ a_{1p} & a_{2p} & a_{np} \end{pmatrix}$.

Definition 3.17 A square matrix of size *n* is said to be

- Symmetric if $\mathbf{A}'=\mathbf{A}$.
- Skew-symmetric if $\mathbf{A}'=-\mathbf{A}$

Consider $M_n(\mathbb{R})$ to be the vector space of square matrices over the field $\mathbb{R}$. Let $S_n(\mathbb{R}) = \{\mathbf{A} \in M_n(\mathbb{R}) \mid \mathbf{A}' = \mathbf{A}\}$ be the subset of symmetric matrix of $M_n(\mathbb{R})$ and let $A_n(\mathbb{R}) = \{\mathbf{A} \in M_n(\mathbb{R}) \mid \mathbf{A}' = -\mathbf{A}\}$ be the subset of skew-symmetric matrix of $M_n(\mathbb{R})$

Theorem 3.12 $S_n(\mathbb{R})$ and $A_n(\mathbb{R})$ are subspaces of $M_n(\mathbb{R})$. Furthermore;

$\dim(M_n(\mathbb{R})) = n^2$, $\dim(S_n(\mathbb{R})) = \frac{n^2+n}{2}$ and $\dim(A_n(\mathbb{R})) = \frac{n^2-n}{2}$.

$\dim(M_n(\mathbb{R})) = \dim(S_n(\mathbb{R})) \oplus \dim(A_n(\mathbb{R}))$;

Theorem 3.13 Consider the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and the scalars α and λ. The following hold for transposition

- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
- $(\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{A}')' = \mathbf{A}$
- $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$
- $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$
- $(\alpha\mathbf{A} + \lambda\mathbf{B})' = \alpha\mathbf{A}' + \lambda\mathbf{B}'$

Definition 3.18 Let $\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ be a square matrix of size *n*. The matrix $\mathbf{A}$ is said to be a diagonal matrix if and only if $a_{ij} = 0$ if $i \neq j$, where $1 \leq i, j \leq n$ Furthermore, the set $\{a_{ij}\}_{i=j}$ is called the diagonal of the matrix $\mathbf{A}$.

Definition 3.19 For a given square matrix $\mathbf{A}$, the trace of $\mathbf{A}$ is the scalar obtained by the summation of all its diagonal elements. If the trace of $\mathbf{A}$ is denoted $tr(\mathbf{A})$ and computed by $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

Theorem 3.14 Consider two square matrices $\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ and

$\mathbf{B} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix}$. Let α and β be two scalars. The following properties holds

when there are applied on trace operation.

1 $\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$ if $\mathbf{A}$ and $\mathbf{B}$ are of the same size. Ie: if $n=m$

2 $\mathrm{tr}(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha\mathrm{tr}(\mathbf{A}) + \beta\mathrm{tr}(\mathbf{B})$

3 $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$

4 $\mathrm{tr}(\mathbf{A}') = \mathrm{tr}(\mathbf{A})$

5 $\mathrm{tr}(\mathbf{A}'\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{A}') = \sum_{i,j \le n} a_{ij}^2$ and $\mathrm{tr}(\mathbf{A}'\mathbf{A}) = \mathbf{0}$ if and only if $\mathbf{A} = \mathbf{0}$.

From property (5) which computes the trace of the product of a matrix with its transpose, the Euclidean matrix norm is defined.

Definition 3.20 Let $\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ be a square matrix. The Euclidean squared norm of $\mathbf{A}$ is the scalar obtained from the computation of the trace of $\mathbf{A}'\mathbf{A}$. It is computed and denoted as follow $\| \mathbf{A} \|^2 = \mathrm{tr}(\mathbf{A}'\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{A}') = \sum_{i \le n} \sum_{j \le n} a_{ij}^2$ .Such that the Euclidean norm of the matrix $\mathbf{A}$ is simply $\| \mathbf{A} \| = \sqrt{\| \mathbf{A} \|^2}$.

To evaluate the closeness of two square matrices of same size $\mathbf{A}=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{pmatrix}$ the concept of Euclidean squared norm of matrix difference is introduced and computed by $\|\mathbf{A}-\mathbf{B}\|^2=\mathrm{tr}\left[(\mathbf{A}-\mathbf{B})'(\mathbf{A}-\mathbf{B})\right]=\sum_{i,j\le n}(a_{ij}-b_{ij})^2$; such that the "distance" between matrices $\mathbf{A}$ and $\mathbf{B}$ is $\|\mathbf{A}-\mathbf{B}\|=\sqrt{\|\mathbf{A}-\mathbf{B}\|^2}$.

Theorem 3.15 Consider two square matrices $\mathbf{A}$ and $\mathbf{B}$ of size n, the following properties applicable on Euclidean matrix norm are true

- $\|\mathbf{A}\|\ge 0$ and $\|\mathbf{A}\|=0$ if and only if $\mathbf{A}=\mathbf{0}$.
- $\|\alpha\mathbf{A}\|=|\alpha|\cdot\|\mathbf{A}\|,\ \forall\alpha\in\mathbb{R}$.
- $\|\mathbf{A}+\mathbf{B}\|\le\|\mathbf{A}\|+\|\mathbf{B}\|$ (Triangular inequality)
- $\|\mathbf{AB}\|\le\|\mathbf{A}\|\|\mathbf{B}\|$ (Cauchy-Schwarz inequality)

Example 3.5 Consider the matrices $\mathbf{A}=\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}$, $\mathbf{B}=\begin{pmatrix} 5 & 0 \\ -2 & 4 \end{pmatrix}$ and compute the following operations $2\mathbf{A}$, sum of $\mathbf{A}$ and $\mathbf{B}$, product of $\mathbf{A}$ and $\mathbf{B}$, transpose of $\mathbf{A}$, trace of $\mathbf{A}$, norm of $\mathbf{A}$, distance between $\mathbf{A}$ and $\mathbf{B}$

Solution

$$2\mathbf{A}=2\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}=\begin{pmatrix} 4 & 2 \\ -2 & 6 \end{pmatrix}$$

$$\mathbf{A}+\mathbf{B}=\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}+\begin{pmatrix} 5 & 0 \\ -2 & 4 \end{pmatrix}=\begin{pmatrix} 7 & 1 \\ -3 & 7 \end{pmatrix}$$

$$\mathbf{AB}=\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 5 & 0 \\ -2 & 4 \end{pmatrix}=\begin{pmatrix} 8 & 4 \\ -11 & 12 \end{pmatrix}$$

$$\mathbf{BA} = \begin{pmatrix} 5 & 0 \\ -2 & 4 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 10 & 5 \\ -8 & 10 \end{pmatrix};\ \mathbf{AB} \neq \mathbf{BA}$$

$$\mathbf{A}' = \mathbf{A}^T = \begin{pmatrix} 2 & -1 \\ 1 & 3 \end{pmatrix}$$

$$\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}') = 2 + 3 = 5$$

$$\| \mathbf{A} \|^2 = \mathrm{tr}(\mathbf{A}'\mathbf{A}) = \sum_{i,j \le n} a_{ij}^2 = 2^2 + 1^2 + (-1)^2 + 3^2 = 15 \ \Rightarrow\ \| \mathbf{A} \| = \sqrt{15}$$

$$\| \mathbf{A} - \mathbf{B} \|^2 = \mathrm{tr}\left[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})'\right] = 12 \ \Rightarrow \| \mathbf{A} - \mathbf{B} \| = \sqrt{12}$$

3.1.2 Determinant

Beyond elementary matrix operations discussed in the previous section, there exits a second range of operations which are mainly used in principal components analysis. This concerns matrix inverse, determinant and diagonalization [17].

Definition 3.21 Consider the square $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ matrix of size *2*. The scalar $a_{11}a_{22} - a_{21}a_{12}$ is called the determinant of the matrix $\mathbf{A}$ and denoted **det(A)** or $|\mathbf{A}|$.

The determinant is important in the evaluation of covariance and principal component computation. When a square matrix $\mathbf{A}$ has an order $n \geq 3$, the computation of its determinant becomes more difficult than for the case of a matrix of size *2*. To define the determinant of a higher order matrix, the concept of sub matrix is requires.

Definition 3.22 Consider a matrix $\mathbf{A}$ of size $(n \times m)$, a sub matrix $\mathbf{B}$ of size $(p \times q)_{\substack{p \le n \\ q \le m}}$ of the matrix $\mathbf{A}$ is obtained by taking a block of entries of $\mathbf{A}$ of size $(p \times q)_{\substack{p \le n \\ q \le m}}$.

For example, considering the matrix $\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, the matrice

$\mathbf{B}=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $\mathbf{C}=\begin{pmatrix} a_{22} & a_{23} \end{pmatrix}$, $\mathbf{D}=\begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}$,are sub matrices of $\mathbf{A}$ of sizes

$(2\times 2),(1\times 2)$ and (3×2) respectively.

Let consider now a square matrix $\mathbf{A}$ of size $n\geq 3$, when the row *i* and the column *j* of the matrix $\mathbf{A}$ are virtually deleted together, a sub matrix of size (*n-1*) is obtained and denoted $\tilde{\mathbf{A}}_{ij}$. This sub matrix is used for the determinant computation.

Remark A constant is considered to be a matrix of size (1×1). It can then be represented by a_{11} and its determinant is $\det(a_{11})=a_{11}$.

Definition 3.23 Let $\mathbf{A}$ be square matrix of size n. The determinant of $\mathbf{A}$ is defined recursively as follow

If *n*=1 then $\mathbf{A}=a_{11}$ and $\det(a_{11})=a_{11}$, else $\det(\mathbf{A})=\sum_{j=1}^{n}(-1)^{1+j}a_{1j}\det(\tilde{\mathbf{A}}_{1j})$ (3.1.22)

Where a_{1j} is the entrance at the position $(1,j)$ in the matrix $\mathbf{A}$ and $\tilde{\mathbf{A}}_{1j}$ is the submatrix defined above [19].

$(-1)^{1+j}\det(\tilde{\mathbf{A}}_{1j})$ is called the Cofactor of the entry of the matrix $\mathbf{A}$ in the row 1 and the column *j*.

In (3.1.22), the index 1 indicates that the determinant is computed by the cofactor expansion along the first row.

Theorem 3.16 For a given square matrix $\mathbf{A}$ of size *n* the determinant can be computed by expansion along any row ***i*** such a way that the formula (3.1.22) becomes $\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\tilde{\mathbf{A}}_{ij})$ (3.1.23)

Theorem 3.17 Consider a square matrix of size *n*, the following properties applied on matrix determinant hold.

- $\det(\mathbf{A}) = \det(\mathbf{A}')$
- If $\mathbf{A}$ is an upper or lower triangular matrix then $\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}$
- $\det(\mathbf{I}_n) = 1$, where $\mathbf{I}_n$ stands for the identity matrix of size *n*
- $\forall \alpha \in \mathbb{R},\ \det(\alpha \mathbf{A}) = \alpha^n \det(\mathbf{A})$

Matrix determinant is important because it determines the invertibility of a matrix, which is useful for diagonalization process.

Theorem 3.18 Let $\mathbf{A}$ be a square matrix of size *n*. Then $\mathbf{A}$ is invertible if and only if $\det(\mathbf{A}) \neq \mathbf{0}$, in which case there exists a matrix $\mathbf{B}$ of size *n* called inverse of $\mathbf{A}$ and denoted $\mathbf{A}^{-1}$, such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$

Theorem 3.19 Consider two square matrices $\mathbf{A}$ and $\mathbf{B}$ of size *n* $\det(\mathbf{AB}) = \det(\mathbf{A}) \cdot \det(\mathbf{B})$ (3.1.24)

Corollary: If $\mathbf{A}$ is invertible then $\det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{I}_n) = 1 = \det(\mathbf{A}) \cdot \det(\mathbf{A}^{-1})$. It follows that $\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}$. (3.1.25)

Many properties of a given matrix $\mathbf{A}$ are defined based on the computation of its determinant and inverse.

Definition 3.24 A square matrix $\mathbf{A}$ of size *n* is said to be an orthogonal matrix if $\mathbf{A}\mathbf{A}' = \mathbf{I}_n$. Furthermore, every orthogonal matrix $\mathbf{A}$ is invertible and its inverse equals to its transpose, $\mathbf{A}^{-1} = \mathbf{A}'$.

Theorem 3.20 Consider an orthogonal matrix $\mathbf{A}$ the following properties are correct

- $\det(\mathbf{A})$ is either -1 or +1
- The product of two orthogonal matrices is another orthogonal matrix.
- The inverse of an orthogonal matrix is also an orthogonal matrix.
- An orthogonal matrix with determinant equals to 1 is called special orthogonal matrix. Such an orthogonal matrix is a rotation.

3.1.3 Eigenvalues, Eigenvectors of a matrix

Eigenvalues and eigenvectors are some matrix characteristics which help to determine whether or not a matrix is diagonalizable.

Definition 3.25 Consider $\mathbf{A}$ in $M_n(\mathbb{R})$ a scalar λ is said to be an eigenvalue of $\mathbf{A}$ if the following conditions are satisfied

- $\ker(\mathbf{A} - \lambda\mathbf{I}_n) \neq \{\mathbf{0}\}$
- $\det(\mathbf{A} - \lambda\mathbf{I}_n) = \mathbf{0}$
- $\exists \mathbf{x} \in \mathbb{R}^n,\ \mathbf{x} \neq \mathbf{0}$ which verifies $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$

Here $\mathbf{x}$ is called the eigenvector corresponding to the eigenvalue λ.

The subset of $\mathbb{R}$ made of all the eigenvalues of the matrix $\mathbf{A}$ is called the spectrum of $\mathbf{A}$ and is sometime denoted by $Sp_{\mathbb{R}}(\mathbf{A})$ or simply $Sp(\mathbf{A})$ if it is assumed that the field over which the matrix $\mathbf{A}$ is defined is known.

Theorem 3.21 Consider $\mathbf{A}$ in $M_n(\mathbb{R})$ and let the scalar λ be an eigenvalue of $\mathbf{A}$. The set of all the eigenvectors corresponding to the eigenvalue λ is denoted $E_\lambda = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \lambda\mathbf{x}\}$ and E_λ is a vector subspace of $\mathbb{R}^n$.

Definition 3.26 E_λ is called the eigenspace of the matrix $\mathbf{A}$ corresponding to the eigenvalue λ.

In practice, for a given matrix $\mathbf{A}$, there exists a standard process to compute eigenvalues and eigenvectors which involve a real polynomial called characteristic polynomial.

Definition 3.27 Consider $\mathbf{A} \in M_n(\mathbb{R})$. The characteristic polynomial of $\mathbf{A}$ is the polynomial with coefficients over the field $\mathbb{R}$ computed and denoted as follow $p_{\mathbf{A}} = \det(\mathbf{A} - x\mathbf{I}_n)$.

Theorem 3.22 The scalar λ is an eigenvalue of the matrix $\mathbf{A} \in M_n(\mathbb{R})$ if and only if λ is a root of the characteristic polynomial $p_{\mathbf{A}}$.

Example 3.6 Consider $\mathbf{A} = \begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ and compute its eigenvalues and eigenvectors.

Solution The characteristic polynomial of $\mathbf{A}$ is

$p_{\mathbf{A}} = \det\left[\begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix} - x\mathbf{I}_2\right] = (-1-x)(2-x)$, it follows that $\mathbf{A}$ has two distinct eigenvalues which are $\lambda_1 = -1$ and $\lambda_2 = 2$, the spectrum of $\mathbf{A}$ is $Sp(\mathbf{A}) = \{-1; 2\}$, The corresponding eigenvectors are computed as follow: Let $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$, if

theeigenvector corresponding to the eigenvalue λ_1 is denoted e_{λ_1} then

$$\begin{aligned}(\mathbf{A}-\lambda_1\mathbf{I}_2)\mathbf{X}=0 &\Leftrightarrow \begin{pmatrix}0 & 0\\ 1 & 3\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\mathbf{0}\\ &\Leftrightarrow x_1+3x_2=0 \text{ ie: } x_1=-3x_2\\ &\Leftrightarrow \begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}-3x_2\\ x_2\end{pmatrix}=x_2\begin{pmatrix}-3\\ 1\end{pmatrix}\\ &\Leftrightarrow e_{\lambda_1}=\begin{pmatrix}-3\\ 1\end{pmatrix}\end{aligned}$$

As previously, if the eigenvector corresponding to the eigenvalue λ_2 is denoted e_{λ_2} then

$$\begin{aligned}(\mathbf{A}-\lambda_2\mathbf{I}_2)\mathbf{X}=0 &\Leftrightarrow \begin{pmatrix}-3 & 0\\ 1 & 0\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\mathbf{0}\\ &\Leftrightarrow -3x_1=0 \text{ ie: } x_1=0\\ &\Leftrightarrow \begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ x_2\end{pmatrix}=x_2\begin{pmatrix}0\\ 1\end{pmatrix}\\ &\Leftrightarrow e_{\lambda_2}=\begin{pmatrix}0\\ 1\end{pmatrix}\end{aligned}$$

The eigenspace corresponding to the eigenvalue λ_1 is

$$E_{\lambda_1}=\left\{\mathbf{X}\in\mathbb{R}^2 \mid \mathbf{X}=t\begin{pmatrix}-3\\ 1\end{pmatrix},\ t\in\mathbb{R}\right\},$$

The eigenspace corresponding to the eigenvalue λ_2 is

$$E_{\lambda_2}=\left\{\mathbf{X}\in\mathbb{R}^2 \mid \mathbf{X}=t\begin{pmatrix}0\\ 1\end{pmatrix},\ t\in\mathbb{R}\right\},$$

3.1.4 Matrix Diagonalization

In this section, the aim is to study the procedure to transform a given square matrix $\mathbf{A}$ into a product of two matrices with simple structure and a diagonal matrix.

Definition 3.28 Consider $\mathbf{A}$ and $\mathbf{B}$ two square matrices of size *n*. $\mathbf{A}$ is similar to $\mathbf{B}$ if and only if there exits an invertible matrix $\mathbf{P}$ of size *n* such that $\mathbf{P^{-1}AP} = \mathbf{B}$. The statement $\mathbf{A}$ is similar to $\mathbf{B}$ is usually denoted by $\mathbf{A} \sim \mathbf{B}$ [19].

Remark

Assume that $\mathbf{A}$ is similar to $\mathbf{B}$, it follows that

$$\mathbf{P^{-1}AP} = \mathbf{B} \Leftrightarrow \mathbf{P}(\mathbf{P^{-1}AP})\mathbf{P^{-1}} = \mathbf{PBP^{-1}}$$
$$\Leftrightarrow \mathbf{A} = \mathbf{PBP^{-1}}$$

this means $\mathbf{B}$ is similar to $\mathbf{A}$.

Theorem 3.23 Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices such that $\mathbf{A} \sim \mathbf{B}$, the following properties hold

- $\det(\mathbf{A}) = \det(\mathbf{B})$
- $\mathbf{A}$ is invertible if and only if $\mathbf{B}$ is invertible
- $\mathbf{A}$ and $\mathbf{B}$ have same characteristic polynomial and same eigenvalues.

Definition 3.29 A square matrix $\mathbf{A}$ of size *n* is said to be diagonalizable, if there exists a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} \sim \mathbf{D}$.

Theorem 3.24 (Diagonalization theorem) consider a square matrix $\mathbf{A}$ of size *n* with distinct eigenvalues are $\{\lambda_1, \lambda_2, ..., \lambda_k\}_{k \le n}$, the following statements are equivalent:

- $\mathbf{A}$ is diagonalizable
- $\sum_{i=1}^{k} \dim E_{\lambda_i} = n$ where E_{λ_i} is the eigenspace corresponding to the eigenvalue λ_i,
- $\forall i,\ \lambda_i \in \mathbb{R}$ and its geometric multiplicity equals to its algebraic multiplicity.

The three points of the theorem 3.24 are important. In practice, the second point or the third point helps to determine whether a given matrix $\mathbf{A}$ is diagonalizable. When $\mathbf{A}$ is diagonalizable, the invertible matrix $\mathbf{P}$ in the formula $\mathbf{A}=\mathbf{P}^{-1}\mathbf{D}\mathbf{P}$ is computed using all the eigenvectors of $\mathbf{A}$ and the diagonal matrix $\mathbf{D}$ is computed using the eigenvalues of $\mathbf{A}$.

Example 3.7 Consider the following matrix $\mathbf{A}=\begin{pmatrix}2&0&0\\1&2&1\\-1&0&1\end{pmatrix}$, check whether $\mathbf{A}$ is diagonalizable, if so, find the diagonal matrix $\mathbf{D}$ and the invertible matrix $\mathbf{P}$ such that $\mathbf{A}=\mathbf{P}^{-1}\mathbf{D}\mathbf{P}$,

Solution

Let first compute the eigenvalues of the matrix $\mathbf{A}$

$$p_{\mathbf{A}}(x)=\det\left[\mathbf{A}-x\mathbf{I}_3\right]=\det\begin{pmatrix}2-x&0&0\\1&2-x&1\\-1&0&1-x\end{pmatrix}=(2-x)^2(1-x).$$

It follows that $\mathbf{A}$ has two eigenvalues $\lambda_1=2$ with algebraic multiplicity equals to 2 and $\lambda_2=1$ with algebraic multiplicity equals to 1.

Let's compute the eigenvectors and eigenspace corresponding to the previous eigenvalues. Consider in $\mathbb{R}^3$ a vector $\mathbf{x}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}$.

For $\lambda_1=2$, solving the equation $(\mathbf{A}-\lambda_1\mathbf{I}_3)\mathbf{x}=\mathbf{0}$ gives the following eigenvector $\mathbf{e_1}=\begin{pmatrix}0\\1\\0\end{pmatrix}$ and $\mathbf{e}_2=\begin{pmatrix}-1\\0\\1\end{pmatrix}$, the corresponding eigenspace is $E_{\lambda_1}=\left\{\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}-1\\0\\1\end{pmatrix}\right\}$,

For $\lambda_2 = 1$, solving the equation $(\mathbf{A} - \lambda_2 \mathbf{I}_3)\mathbf{x} = \mathbf{0}$ gives the following eigenvector $\mathbf{e}_3 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$ the corresponding eigenspace is $E_{\lambda_2} = \left\{ \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \right\}$,

$\sum_{i=1}^{2} \dim E_{\lambda_i} = \dim E_{\lambda_1} + \dim E_{\lambda_2} = 2 + 1 = 3$, furthermore, the algebraic

Multiplicity of each eigenvector corresponds to its geometric multiplicity; this means the matrix $\mathbf{A}$ is diagonalizable. The diagonal matrix is $\mathbf{D} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and the invertible matrix $\mathbf{P} = \begin{pmatrix} 0 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$, its inverse is $\mathbf{P}^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{pmatrix}$, such that

$$\mathbf{A} = \mathbf{P}^{-1}\mathbf{D}\mathbf{P} \Leftrightarrow \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

If $\mathbf{A}$ is diagonalizable, then it is easy to compute $\mathbf{A}^n$. Actually

$$\begin{aligned} \mathbf{A} = \mathbf{P}^{-1}\mathbf{D}\mathbf{P} \Rightarrow \mathbf{A}^n &= \left(\mathbf{P}^{-1}\mathbf{D}\mathbf{P}\right)^n \\ &= \left(\mathbf{P}^{-1}\mathbf{D}\mathbf{P}\right) \cdot \ldots \cdot \left(\mathbf{P}^{-1}\mathbf{D}\mathbf{P}\right) \\ &= \mathbf{P}^{-1}\mathbf{D}\mathbf{P}\mathbf{P}^{-1}\mathbf{D}\mathbf{P}\ldots\mathbf{P}^{-1}\mathbf{D}\mathbf{P} \\ &= \mathbf{P}^{-1}\mathbf{D}^n\mathbf{P} \end{aligned}$$

In example 3.7,

$$\mathbf{A}^n = \mathbf{P}^{-1}\mathbf{D}^n\mathbf{P} \Leftrightarrow \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 1 \\ -1 & 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2^n & 0 & 0 \\ 0 & 2^n & 0 \\ 0 & 0 & 1^n \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

3.1.5 Singular Value Decomposition

A matrix $\mathbf{A}$ can be written in some easy way to explain more clearly the data it represents. This involves the decomposition of the data matrix. In this section two methods used in the decomposition of a matrix are introduced. These are the spectral decomposition (SD) and singular value decomposition (SVD).

Theorem 3.25 Let $\mathbf{A}$ be a symmetric matrix of size *n*. Then $\mathbf{A}$ can be expressed in terms of its eigenvalues and eigenvectors. This expression is called the spectral decomposition of $\mathbf{A}$. If we denoted by the pair $(\lambda_i, \mathbf{e_i})$ the i^{th} eigenvalue of $\mathbf{A}$ with its corresponding eigenvector, then the spectral decomposition of $\mathbf{A}$ is computed by the following formula $\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{e_i} \mathbf{e_i}'$.

Example 3.1.2 Let consider the symmetric matrix $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, its eigenvalues and eigenvectors are represented by the pairs $\left(1, \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)$ and $\left(2, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)$, such that

$$\mathbf{A} = 1\begin{pmatrix} 1 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}.$$

The idea of the spectral decomposition can be extended for a rectangular matrix $\mathbf{A}$ of size $(n \times m)$, the eigenvalues and eigenvectors in this case are computed from the square matrices $\mathbf{AA}'$ and $\mathbf{A}'\mathbf{A}$. In this case the procedure is called singular value decomposition.

Theorem 3.26 Consider a rectangular matrix $\mathbf{A}$ of size $(n \times m)$ over the field $\mathbb{R}$, then there exits two square orthogonal matrices say $\mathbf{U}$ and $\mathbf{V}$ of sizes $(m \times m)$ and $(n \times n)$ respectively such that $\mathbf{A} = \mathbf{U\Lambda V}'$, where $\mathbf{\Lambda}$ is a diagonal matrix. The elements of $\mathbf{\Lambda}$

are $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$, where $p = \min(m,n)$. The positive real values σ_i are the singular values of the matrix $\mathbf{A}$.

The vectors $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Lambda}$ are computed as follow

- The singular values $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ are the squawroots of the common eigenvalues of both $\mathbf{AA}'$ and $\mathbf{A}'\mathbf{A}$ matrices.
- The matrix $\mathbf{U}$ is made from the eigenvectors of the matrix $\mathbf{A}'\mathbf{A}$.
- The matrix $\mathbf{V}$ is made from the eigenvectors of the matrix $\mathbf{AA}'$.

Example 3.1.3 Consider the (3×2) matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \\ 2 & 1 \end{bmatrix}$, its singular value decomposition is computed as follow

$\mathbf{A}'\mathbf{A} = \begin{bmatrix} 9 & 8 \\ 8 & 9 \end{bmatrix}$, its eigenvalues and its corresponding eigenvectors are $\left\{17, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\}$ and $\left\{1, \begin{pmatrix} 1 \\ -1 \end{pmatrix}\right\}$. Normalization of the eigenvectors yields the matrix $\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$.

$\mathbf{AA}' = \begin{bmatrix} 5 & 6 & 4 \\ 6 & 8 & 6 \\ 4 & 6 & 5 \end{bmatrix}$, similarly to the case of $\mathbf{A}'\mathbf{A}$, after computation and normalization of eigenvalues and eigenvectors, the matrix $\mathbf{U} = \begin{bmatrix} \frac{3}{\sqrt{34}} & \frac{-1}{\sqrt{2}} & \frac{2}{\sqrt{17}} \\ \frac{4}{\sqrt{34}} & 0 & \frac{-3}{\sqrt{17}} \\ \frac{3}{\sqrt{34}} & \frac{1}{\sqrt{2}} & \frac{2}{\sqrt{17}} \end{bmatrix}$.

The singular values of the matrix $\mathbf{A}$ are $\sigma_1 = \sqrt{17}$ and $\sigma_2 = 1$ and $\mathbf{\Lambda} = \begin{bmatrix} \sqrt{17} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$.

Finally

$$\mathbf{A} = \begin{bmatrix} \frac{3}{\sqrt{34}} & \frac{-1}{\sqrt{2}} & \frac{2}{\sqrt{17}} \\ \frac{4}{\sqrt{34}} & 0 & \frac{-3}{\sqrt{17}} \\ \frac{3}{\sqrt{34}} & \frac{1}{\sqrt{2}} & \frac{2}{\sqrt{17}} \end{bmatrix} \begin{bmatrix} \sqrt{17} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

The reduced singular value decomposition of the matrix $\mathbf{A}$ is also defined and computed as follow $\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\mathbf{V}' = \begin{bmatrix} \frac{3}{\sqrt{34}} & \frac{-1}{\sqrt{2}} \\ \frac{4}{\sqrt{34}} & 0 \\ \frac{3}{\sqrt{34}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{17} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$. In this case the third row of the matrix $\mathbf{\Lambda}$ which is made of null elements is deleted to build the matrix $\hat{\mathbf{\Lambda}}$. To keep possible the matrix product, the third column of the matrix $\mathbf{U}$ is deleted to build the matrix $\hat{\mathbf{U}}$ .One important application of the singular value decomposition is to map the data represented in a coordinate system, let's say the orthogonal coordinate system onto a scaled coordinate system under a matrix $\mathbf{A}$. This mapping can help to compress an original image or data such to retain 20% of information.

3.1 Statistics Concepts

Multivariate data analysis is based on the usual concepts of univariate and bi-variate statistics. The basic concepts studied for univariate and bivariate statistics are extended for the case of multivariate statistics.

3.1.1 Sample Space, Random Variable, Probability Distribution

Statistics is concerned with evaluation of data, and based on obtained results interpretation carried out generally for the purpose of helping in sound decision making. Such a study is possible only when it stems from a random variable defined on a sample space, and adheres to the rules of a certain probability function.

Definition 3.30 When a statistical experiment is conducted, the set of all its possible outcomes is called the sample space and it is usually denoted by S [9].

Remark A sample space with a finite number of possibilities is called a discrete otherwise it is called a continuous sample space.

Definition 3.31 A random variable is a measurable function from the sample space that associate each element of the sample space with a real number or any other property that characterizes element from the sample space. A random variable is usually denoted by X .

Remark A random variable with a countable set of outcomes is called a discrete random variable; otherwise it is called a continuous random variable.

Definition 3.32 The set of all the ordered pairs $\left(x, f\left(x\right)\right)$ where x is a possible outcome of the random variable X and $f\left(x\right)$ is the chance of x to appear is called probability function or probability distribution or probability mass function.

3.1.2 Univariate Normal Distribution

It was previously mentioned that there exits two types of variables, which are discrete and continuous variables. Their corresponding distributions functions are also discrete and continuous distributions functions. In the case of continuous distribution, the most interesting one is the normal distribution. This is due to its wide use in many fields of research and application.

Definition 3.33 In the univariate case considers a random variable X with its mean μ and its variance σ^2. The random variable X is said to be normally distributed or a normal random variable if its graphical representation has a bell-shape as shown on the following figure [10].

The probability density of is X computed and denoted by

$$n(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2};\ -\infty<x<\infty\ . \tag{3.2.1}$$

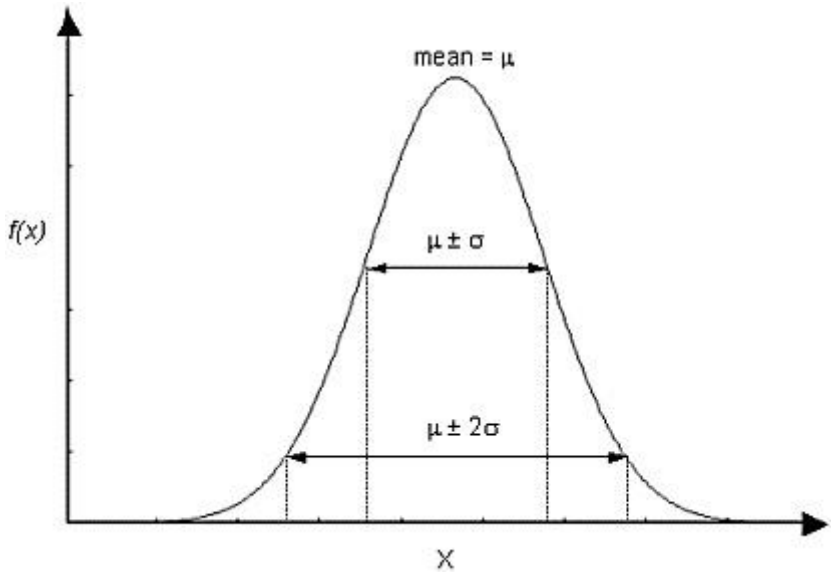

Figure 3.2.1: The normal distribution shape

In practice, a normal distribution is completely defined when the value of μ and σ are specified.

Using the transformation of variable $Z = \frac{X - \mu}{\sigma}$, the normal random variable becomes $n(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$. (3.2.2)

This is a normal distribution with mean 0 and standard deviation 1, also called the standard normal probability density function.

Definition 3.34 A normal random variable, which is distributed with its mean equals to 0 and its variance equals to 1 is said to be a standard normal distribution.

3.1.3 Bivariate Normal Distribution

Consider two normally distributed random variables X and Y. The joint distribution of the variables X and Y is called the bivariate distribution. The variable x and y are dependent. The dependence between the variables X and Y is defined by the correlation coefficient ρ. Probability density of a bivariate normal distribution is given by

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

with $-\infty < x < \infty$; $-\infty < y < \infty$ and ρ is called population correlation coefficient.

(3.2.3)

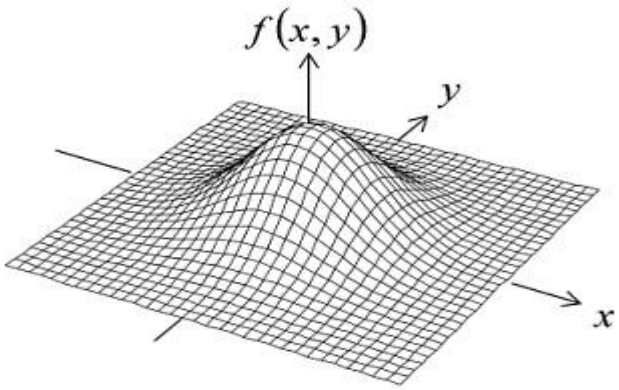

Figure 3.2.2: The bivariate normal distribution shape

3.1.4 Multivariate Normal Distribution

A normal distribution with *p* random variables is generally named as a multivariate normal distribution. Consider the *p=2* (bivariate) case, where the covariance $\mathbf{\Sigma}$ between the random variables X_1 and X_2 can be written in the matrix form

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{X_1}^2 & \rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2} \\ \rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{pmatrix}, \tag{3.2.4}$$

where $\sigma_{X_1}^2$ and $\sigma_{X_2}^2$ are the variances of X_1 and X_2 respectively. $\rho_{X_1X_2}$ is the correlation coefficient between X_1 and X_2 .The determinant of the $\mathbf{\Sigma}$ matrix is

$$|\mathbf{\Sigma}| = \sigma_X^2\sigma_Y^2\left(1-\rho^2\right). \tag{3.2.5}$$

The inverse of the covariance matrix is computed and denoted by

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_X^2} & \frac{-\rho}{\sigma_X\sigma_Y} \\ \frac{-\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2} \end{pmatrix} \tag{3.2.6}$$

Consider the variables X_1, X_2 and their respective mean deviation in form of a matrix $\mathbf{X} = \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}$. Then $\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X} = \chi^2$ which is actually the quadratic form of the vector $\mathbf{X}$, is the chi-square random variable. The bivariate normal distribution function can now be expressed by $f\left(x,y\right) = \left(2\pi\right)^{-1}|\mathbf{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}}$. (3.2.7)

Generally, consider $\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ a random vector of p variables and its vector mean

$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$, the multivariate normal distribution function of $\mathbf{X}$ is

$$f(\mathbf{X}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})} . \quad (3.2.8)$$

3.1.5 Sample Mean - Vector Mean

The central tendency of a dataset is usually measured by the computation of its mean.

Definition 3.35 Let X be a random variable with its distribution function $f(x)$. The expected value of X or the mean of X is defined by

$$\mu = E(X) = \sum_x x f(x) \quad (3.2.9)$$

if X is a discrete random variable and

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (3.2.10)$$

if X is a continuous random variable [9].

Definition 3.36 Consider a random sample of n observations say $(x_1, \cdots, x_n)$, its mean or expected value is the scalar $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. (3.2.11)

Remark Although definition 3.35 and definition 3.36 expressed the mean concept, the computation seems to be different from one definition to another. The difference comes from the fact that in definition 3.35, a distribution function $f(x)$ is given, whereas in definition 3.36, the computation is done from a set of collected data.

The mean concept defined for a sample of n observations can be extended for the case of multivariate data. The concept of mean vector is then defined. Mean vector resulted from the computation of the mean of an extracted vector from a multivariate dataset.

Definition 3.37 Consider the matrix $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & & & & & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$ that

representing a dataset of n observations of p variables. The sample vector mean is

defined by $\bar{\mathbf{x}}' = \frac{1}{n}\mathbf{1}_n' \mathbf{X}$. (3.2.12)

Where $\mathbf{1}_n$ is a column vector of size n, which elements are ones. The sample mean vector $\bar{\mathbf{x}}$ is a column vector of size p, such that it is denoted $\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$. Where each entrance is the mean of its corresponding index column form the data matrix $\mathbf{X}$.

If now the sample consideration is extended for a whole population then the population mean vector is the expectation of the sample mean vector. It is computed and denoted $E(\mathbf{x}) = E\begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_p} \end{pmatrix} = \begin{pmatrix} E(\mathbf{x_1}) \\ E(\mathbf{x_2}) \\ \vdots \\ E(\mathbf{x_p}) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$. (3.2.13)

Let consider a single random variable X_1 with mean μ_1. If X_1 is multiplied by a constant c then its mean value or expectation is expressed by

$E(cX_1) = cE(X_1) = c\mu_1$. (3.2.14)

Furthermore, if a linear combination of two random variables X_1 and X_2 with respective means μ_1 and μ_2 is considered then the mean of its linear combination $c_1X_1 + c_2X_2$ is given by $E(c_1X_1 + c_2X_2) = c_1E(X_1) + c_2E(X_2) = c_1\mu_1 + c_2\mu_2$, (3.2.15)

which can also be expressed in form of a matrix product as follow

$$c_1\mu_1 + c_2\mu_2 = \begin{bmatrix}\lambda_1 & \lambda_2\end{bmatrix}\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} = \mathbf{C}'\boldsymbol{\mu}. \text{ Where } \mathbf{C} = \begin{bmatrix}c_1 \\ c_2\end{bmatrix} \text{ and } \boldsymbol{\mu} = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}. \quad (3.2.16)$$

Theorem 3.27 Consider the following q linear combinations of p random variables

$$\begin{aligned} Z_1 &= c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\ Z_2 &= c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \\ &\vdots \qquad\qquad\qquad \vdots \\ Z_q &= c_{q1}X_1 + c_{q2}X_2 + \cdots + c_{qp}X_p \end{aligned},$$

it can be represented in term of matrices product as follow

$$\mathbf{Z} = \begin{bmatrix}Z_1 \\ Z_2 \\ \vdots \\ Z_q\end{bmatrix} = \begin{bmatrix}c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp}\end{bmatrix}\begin{bmatrix}X_1 \\ X_2 \\ \vdots \\ X_p\end{bmatrix} = \mathbf{CX}. \quad (3.2.17)$$

It follows that $\boldsymbol{\mu}_{\mathbf{Z}} = E(\mathbf{Z}) = E(\mathbf{CX}) = \mathbf{C}\mu_{\mathbf{X}}$. (3.2.18)

3.1.6 Variance and Covariance

It helps to measure the variation from the mean of the random variable. Covariance represents the variation between two variables.

Definition 3.38 Consider a random variable X with mean μ and which has a probability distribution $f(x)$. The variance of X is computed by

$$\sigma^2 = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 f(x) \quad (3.2.19)$$

if X is a discrete random variable, and

$$\sigma^2 = E\left[\left(X-\mu\right)^2\right] = \int_{-\infty}^{+\infty}\left(X-\mu\right)^2 f\left(x\right)dx \tag{3.2.20}$$

if X is a continuous random variable [20;21].

The variance of a random variable is denoted by $Var\left(X\right)$ or σ_X^2 or simply σ^2 [9].

Definition 3.39 Consider a random sample of n observations say $\left(x_1,\cdots,x_n\right)$, with mean $\overline{x}$. Its variance is computed and denoted by $s^2 = \dfrac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n-1}$. (3.2.21)

Theorem 3.28 Consider a linear combination $\mathbf{C}'\mathbf{X} = c_1X_1+\cdots+X_p$ of p random variables $\left(X_1,\cdots,X_p\right)$. Its variance is $Var(\mathbf{C}'\mathbf{X}) = \mathbf{C}'\mathbf{\Sigma}\mathbf{C}$. Here $\mathbf{\Sigma} = \text{cov}(\mathbf{X})$. (3.2.22)

Definition 3.40 If X_1 and X_2 are two random variables with a joint distribution function $f\left(x_1,x_2\right)$, with means μ_{X_1} and μ_{X_2} respectively, then their covariance is

$$\sigma_{X_1X_2} = E\left[\left(X_1-\mu_{X_1}\right)\left(X_2-\mu_{X_2}\right)\right] = \sum_{x_1}\sum_{x_2}\left(x_1-\mu_{X_1}\right)\left(x_2-\mu_{X_2}\right)f\left(x_1,x_2\right) \tag{3.2.23}$$

if the variables X_1 and X_2 are discrete, and

$$\sigma_{X_1X_2} = E\left[\left(X_1-\mu_{X_1}\right)\left(X_2-\mu_{X_2}\right)\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(x_1-\mu_{X_1}\right)\left(x_2-\mu_{X_2}\right)f\left(x_1,x_2\right)dx_1dx_2 \tag{3.2.24}$$

if the variables X_1 and X_2 are continuous.

The covariance of X_1 and X_2 is also denoted $Cov\left(\mathrm{X}_1,\mathrm{X}_2\right)$.

Definition 3.41 Consider a random sample of two variables X and Y with joint distribution of n observations say $\left((x_1,y_1),\cdots,(x_n,y_n)\right)$, its covariance is computed by $\sigma_{XY}^2 = \dfrac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{n-1}$. (3.2.25)

In the case of a random sample with more than 2 variables, it is suitable to use a matrix form to display the computation of the covariance between each pair of variables.

Definition 3.42 Consider p random variables $\left(X_1,\cdots,X_p\right)$. There exist

$\binom{p}{2}=\frac{p!}{2!(p-2)!}=\frac{p(p-1)(p-2)!}{2!(p-2)!}=\frac{p(p-1)}{2}$ possible pairs of elements from

$\left(X_1,\cdots,X_p\right)$. The covariance between elements of each of those pairs can be computed and displayed in form of a symmetric matrix as below

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{X_1X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_p}\\ \sigma_{X_2X_1} & \sigma_{X_2X_2} & \cdots & \sigma_{X_2X_p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{X_pX_1} & \sigma_{X_pX_2} & \cdots & \sigma_{X_pX_p}\end{pmatrix};$$

where the element $\sigma_{X_iX_k}=E\left[\left(X_i-\mu_i\right)\left(X_k-\mu_k\right)\right]$ is the covariance between the variables X_i and X_j . $\boldsymbol{\Sigma}$ is a symmetric matrix so $\sigma_{X_iX_k}=\sigma_{X_kX_i}$. The entrance $\sigma_{X_iX_i}$ seen as the covariance between a variable and itself is simply the variance of X_i.

Theorem 3.29 If X_1 and X_2 are two independent random variables then $\sigma_{X_1X_2}=0$. The vice versa case of theorem 3.29 is not true in general because there are cases where $\sigma_{X_1X_2}=0$ but that X_1 and X_2 are not independent [10].

Theorem 3.30 Consider the following q linear combinations of p random variables

$$\begin{aligned}Z_1&=c_{11}X_1+c_{12}X_2+\cdots+c_{1p}X_p\\ Z_2&=c_{21}X_1+c_{22}X_2+\cdots+c_{2p}X_p\\ \vdots&\qquad\qquad\vdots\\ Z_q&=c_{q1}X_1+c_{q2}X_2+\cdots+c_{qp}X_p\end{aligned},$$

it can be represented in terms of matrix product as follow

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \mathbf{CX}. \tag{3.2.26}$$

It follows that $\boldsymbol{\Sigma}_{\mathbf{Z}} = \text{cov}(\mathbf{Z}) = \text{cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}'$. (3.2.27)

3.1.7 Correlation Coefficient Matrix

The covariance matrix **S** shows the variance values on its diagonal, and the nature of the pairwise linear relationship between the variables in the off diagonal elements. Range of values of the elements of **S**is $-\infty < s_{ij} < \infty$. However, the strength of the linear relationship between the variables is not evident. This handicap is overcome by computing the pairwise linear correlation matrix **R**. Diagonal elements of **R** is equal to 1, since the linear correlation between a variable and itself must be 100%.Off diagonal elements represents the strength of linear correlation between pairs of variables.

Definition 3.43 Consider two random variables X_1 and X_2. The correlation coefficient between them is $\rho_{X_1X_2} = \dfrac{\text{cov}(X_1, X_2)}{\sqrt{var(X_1)var(X_2)}}$ [15]. (3.2.28)

Definition 3.44 Consider p random variables $(X_1, \cdots, X_p)$. There exist $\binom{p}{2} = \dfrac{p!}{2!(p-2)!} = \dfrac{p(p-1)(p-2)!}{2!(p-2)!} = \dfrac{p(p-1)}{2}$ possible pairs of elements from $(X_1, \cdots, X_p)$. The correlation coefficient between elements of each of those pairs can

be computed and displayed in form of a symmetric matrix as follow

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{X_1X_1} & \rho_{X_1X_2} & \cdots & \rho_{X_1X_p} \\ \rho_{X_2X_1} & \rho_{X_2X_2} & \cdots & \rho_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \rho_{X_pX_2} & \cdots & \rho_{X_pX_p} \end{pmatrix}; \qquad (3.2.29)$$

where the element $\rho_{X_iX_k} = \frac{\sigma_{X_iX_k}}{\sqrt{\sigma_{X_i} \cdot \sigma_{X_k}}}$ is the correlation coefficient between the variables X_1, X_2 . $\boldsymbol{\rho}$ is a symmetric matrix so $\rho_{X_iX_k} = \rho_{X_kX_i}$. The entrance $\rho_{X_iX_i}$ seen as the correlation between a variable and itself and it equals to 1.

Chapter 4

COMPUTING PRINCIPAL COMPONENTS USING COVARIANCE AND CORRELATION MATRICES

Multivariate statistics deal with the case where more than one variable is involved. When the number of variables (p) is very large, analysis of data coming from these variables become very demanding in terms of computational time. Therefore, some method has to be developed, that will reduce the number of variables involved in the computation. Principal component analysis (PCA) is one such method that generates a linear combination of all variables starting with the direction of largest variation in the data to the direction where the smallest variation is. Number of principal components (PCs) is the same as the number of variables, but only the first few PCs can account for more than 90% of total variation in the data. Use of the few PCs suffices to produce meaningful and reliable interpretation regarding the data [25;26].

4.1 Population and Sample Principal Components

The principal components of a $n \times p$ dataset representing *n* observations of *p* variables are some particular linear combinations of those *p* random variables. Geometrically, these linear combinations represent a new system of coordinate's axes obtained by rotating the original system $X_1, X_2, \cdots, X_p$ of the coordinate's axes.

In practice, the following steps are used to compute PCs:

- Computation of $\boldsymbol{\Sigma}$ and /or $\boldsymbol{\rho}$ matrices

- Computation of eigenvalues and eigenvectors or either $\boldsymbol{\Sigma}$ and /or $\boldsymbol{\rho}$ matrices.
- The eigenvalues are ordered in descending order.
- The coefficients of the principal components are the eigenvectors of the covariance or correlation matrices $\mathbf{e}_i' = \begin{pmatrix} e_{i1} & . & . & . & e_{ip} \end{pmatrix}$ (4.1.1)

The PCs are computed as follows

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ &\vdots \\ Y_p &= \mathbf{e}_p'\mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned} \quad (4.1.2)$$

Theorem 4.1: Consider a vector of p random variables $\mathbf{X} = \begin{bmatrix} X_1 & \cdots & X_p \end{bmatrix}$, with its associated covariance matrix $\boldsymbol{\Sigma}$, computed from *n* observations. Consider the eigenvalue – eigenvector pairs $(\lambda_1, \mathbf{e}_1), \cdots, (\lambda_p, \mathbf{e_p})$ of the covariance matrix $\boldsymbol{\Sigma}$, such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and the PCs are

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ &\vdots \\ Y_p &= \mathbf{e}_p'\mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

and the following relationship holds

$$\sigma_{X_1X_1} + \sigma_{X_2X_2} + \cdots + \sigma_{X_pX_p} = \sum_{i=1}^{p} \text{var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{var}(Y_i) \text{ [4]} \quad (4.1.3)$$

Proof: The scalar $\sigma_{X_1X_1} + \sigma_{X_2X_2} + \cdots + \sigma_{X_pX_p} = tr(\boldsymbol{\Sigma})$. The matrix $\boldsymbol{\Sigma}$ is diagonalizable and can be represented by $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$. Where $\mathbf{P} = \begin{bmatrix} \mathbf{e_1}, \mathbf{e}_2, \ldots, \mathbf{e}_p \end{bmatrix}$ and

$\mathbf{\Lambda} = diag\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)$. It follows that

$tr\left(\mathbf{\Sigma}\right) = tr\left(\mathbf{P\Lambda P'}\right) = tr\left(\mathbf{\Lambda P'P}\right) = tr\left(\mathbf{\Lambda I}\right) = tr\left(\mathbf{\Lambda}\right) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$. Thus

$$\sum_{i=1}^{p} \text{var}\left(X_i\right) = tr\left(\mathbf{\Sigma}\right) = tr\left(\mathbf{\Lambda}\right) = \sum_{i=1}^{p} \text{var}\left(Y_i\right).$$

The ratio $\dfrac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ represents the proportion of variation of the population due to the k^{th} principal component.

The first few components which represent a high percentage (around 90% or more) of the variation in the population, can be used to replace the initial p variables without losing too much information.

The contribution of the k^{th} variable X_k in the i^{th} PC Y_i can be evaluated from the i^{th} eigenvector $\mathbf{e_i} = \left[e_{i1}, \ldots, e_{ik}, \ldots, e_{ip}\right]$. The importance of the k^{th} variable X_k in the i^{th} PC Y_i is given by the magnitude of e_{ik} .

Theorem 4.2 Consider the PCs

$$Y_1 = \mathbf{e}_1' X = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$
$$\vdots$$
$$Y_p = \mathbf{e}_p' X = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p$$

Computed from the covariance matrix $\mathbf{\Sigma}$, with the associated eigenvalues-eigenvectors couples $\left(\lambda_1, \mathbf{e_1}\right), \cdots, \left(\lambda_p, \mathbf{e_p}\right)$. The correlation coefficient between the i^{th} PC Y_i and the k^{th} variable X_k is computed by

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{X_k X_k}}}, \text{ where } 1 \le i, k \le p \quad [19] \qquad (4.1.4)$$

4.2 Geometric Representation of PCs

Geometrically, PCs represent a new system of coordinate built from an existing one. Furthermore its axes show direction in which there is high variation or where there is accumulation of data for a given dataset. Consider the normally distributed variables $\mathbf{X} = \left(X_1, \ldots, X_p \right)$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\left(N_p \left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \right) \right)$ Let $\lambda_i, \mathbf{e}_i$ be the eigenvalue-eigenvector pairs of the covariance matrix on the $\boldsymbol{\mu}$ centered ellipsoids defined by $\left(\mathbf{x} - \boldsymbol{\mu} \right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu} \right) = c^2$ with axes $(\pm c\sqrt{\lambda_i}\mathbf{e}_i;\ i = 1, 2, \ldots, \mathrm{p})$. Considered that PCs are computed from the covariance matrix $\boldsymbol{\Sigma}$ and represented by $Y_1 = \mathbf{e}_1'\mathbf{x}, \cdots, Y_p = \mathbf{e}_p'\mathbf{x}$ and assuming that $\boldsymbol{\mu} = \mathbf{0}$, it follows that

$$c^2 = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \frac{1}{\lambda_1}\left(\mathbf{e}_1'\mathbf{x} \right)^2 + \frac{1}{\lambda_2}\left(\mathbf{e}_2'\mathbf{x} \right)^2 + \cdots + \frac{1}{\lambda_p}\left(\mathbf{e}_p'\mathbf{x} \right)^2 = \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2. \tag{4.2.1}$$

Since the eigenvalues are positive, this equation is the definition of an ellipsoid in the coordinate system $Y_1, Y_2, \ldots, Y_p$, its axis being the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$. The major axis is given by the eigenvector belonging to the highest eigenvalue λ_1 and the remaining following in sequence of the eigenvalues $\lambda_2, \cdots, \lambda_p$.

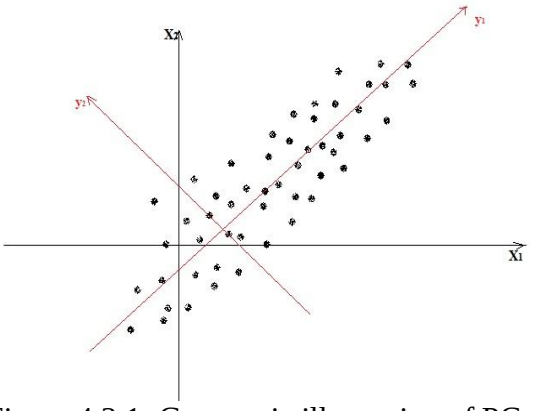

Figure 4.2.1: Geometric illustration of PCs

Figure4.2.1 represents a situation where from the scatter plot of a dataset in the coordinate system (x_1, x_2) . The PCs are computed and the first two are chosen to be enough to represent the data in a new coordinate system $(\mathbf{y_1}, \mathbf{y_2})$ where the new coordinate system fits well to the data.

4.3 Number of PCs Sufficient to Represent the Population Variation

It is mentioned previously that for a dataset of n observation of p random variables, the number of PCs computed is the same as the number of variables. Determination of the number of PCs that can represent the process under study adequately is as follows.

Definition 4.1 Consider the covariance matrix $\mathbf{\Sigma}$ with associated eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e_1}), \cdots, (\lambda_p, \mathbf{e_p})$. The ratio $\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ represents the proportion of variation in the population due to the k^{th} principal component [21]. Furthermore, the variation of the population due to the first q PCs is computed b

$$\Psi = \frac{\lambda_1 + \cdots + \lambda_q}{\sum_{k=1}^{p} \lambda_k} \tag{4.3.1}$$

The first q PCs that represent a high percentage of total variation without a great loss of information is preferred. A value of $\Psi \geq 90\%$ is considered very satisfactory for many practical purposes [23]. Another way to choose the number of eigenvalues sufficient to represent the population variation without loss of information is based on a graphical observation of a scree plot. A scree plot is a two-dimensional graph where the eigenvalues of the covariance matrix are represented on the y-axis and their corresponding index on the x-axis. The eigenvalues are ordered in descending

order, before plotting. The elbow bend of the scree plot show the number of PCs to be considered.

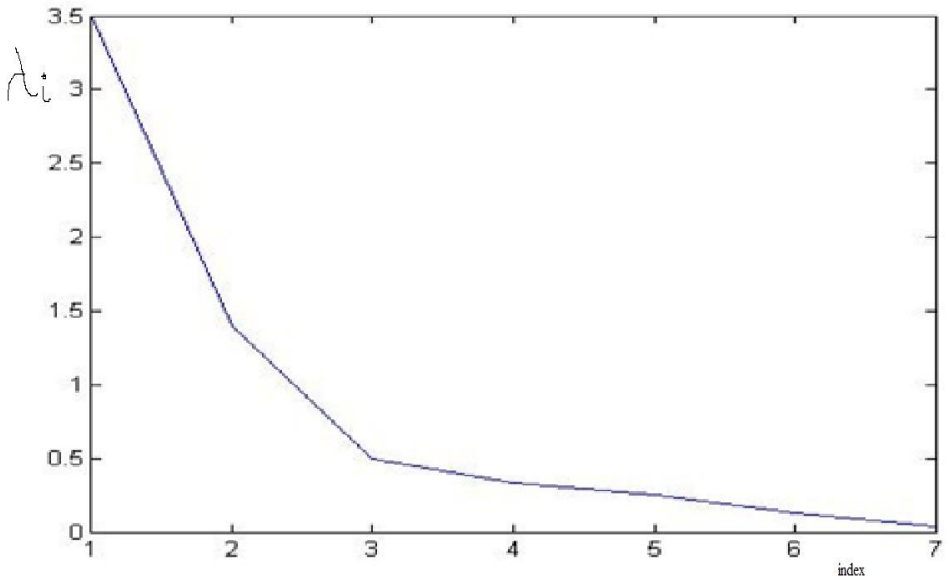

Figure 4.3.1: Illustration of scree plot

Figure 4.3.1 is an illustration of a situation where $p = 7$ random variables are observed. Here, the elbow is observed at the index 3, thus the first 3 PCs are enough to represent the variation of the whole population without loose of information.

4.4 Standardized PCs

The PCs can also be computed from standardized variables. The need of using such variables for PCs computation will be discussed in the upcoming session.

Consider a vector of p random variables $\mathbf{X}' = \left(X_1, \cdots, X_p \right)$, with mean vector $\boldsymbol{\mu}' = \left(\mu_1, \mu_2, \cdots, \mu_p \right)$ and standard deviation matrix

$$\mathbf{V}^{1/2} = diag\left(\sqrt{\sigma_{X_1X_1}}, \sqrt{\sigma_{X_2X_2}}, \cdots, \sqrt{\sigma_{X_pX_p}} \right), \tag{4.4.1}$$

a new vector of p random variables $\mathbf{Z}=\left(Z_1,\cdots,Z_p\right)$ can be defined from

$$\mathbf{X}'=\left(X_1,\cdots,X_p\right),\ \text{where}\ Z_i=\frac{\left(X_i-\mu_i\right)}{\sqrt{\sigma_{X_iX_i}}}. \tag{4.4.2}$$

Definition 4.2 The variables $\mathbf{Z}'=\left(Z_1,\cdots,Z_p\right)$ where $Z_i=\frac{\left(X_i-\mu_i\right)}{\sqrt{\sigma_{X_iX_i}}}$ are called standardized variables [6].

Using the new variables $\mathbf{Z}'=\left(Z_1,\cdots,Z_p\right)$, PCs and all its related concepts can still be computed without loss of information.

The random variable vector $\mathbf{Z}'=\left(Z_1,\cdots,Z_p\right)$ can be expressed in terms of the standard deviation matrix $\mathbf{V}^{1/2}$ and the covariance matrix $\mathbf{\Sigma}$ as follow.

$$\begin{pmatrix} \frac{1}{\sqrt{\sigma_{X_1X_1}}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{\sigma_{X_pX_p}}} \end{pmatrix}\begin{pmatrix} \left(X_1-\mu_1\right) \\ \vdots \\ \left(X_p-\mu_p\right) \end{pmatrix}=\begin{pmatrix} \frac{\left(X_1-\mu_1\right)}{\sqrt{\sigma_{X_1X_1}}} \\ \vdots \\ \frac{\left(X_p-\mu_p\right)}{\sqrt{\sigma_{X_pX_p}}} \end{pmatrix}\Leftrightarrow\left(\mathbf{V}^{1/2}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\right)=\mathbf{Z} \tag{4.4.3}$$

Furthermore, the following relations hold

$$E(\mathbf{Z})=E\left(\left(\mathbf{V}^{1/2}\right)^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)=\left(\mathbf{V}^{1/2}\right)^{-1}E(\mathbf{X}-\boldsymbol{\mu})=\left(\mathbf{V}^{1/2}\right)^{-1}\begin{pmatrix} E\left(X_1-\mu_1\right) \\ \vdots \\ E\left(X_p-\mu_p\right) \end{pmatrix}=\left(\mathbf{V}^{1/2}\right)^{-1}\begin{pmatrix} E\left(X_1\right)-\mu_1 \\ \vdots \\ E\left(X_p\right)-\mu_p \end{pmatrix}=\left(\mathbf{V}^{1/2}\right)^{-1}\begin{pmatrix} \mu_1-\mu_1 \\ \vdots \\ \mu_p-\mu_p \end{pmatrix}=\mathbf{0} \tag{4.4.4}$$

and

$$Cov(\mathbf{Z})=Cov\left(\left(\mathbf{V}^{1/2}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\right)\right)=\left(\mathbf{V}^{1/2}\right)^{-1}Cov\left(\mathbf{X}-\boldsymbol{\mu}\right)\left(\left(\mathbf{V}^{1/2}\right)^{-1}\right)'=\left(\mathbf{V}^{1/2}\right)^{-1}\mathbf{\Sigma}\left(\mathbf{V}^{1/2}\right)^{-1}=\boldsymbol{\rho} \tag{4.4.5}$$

Theorem 4.3 The PCs can be computed from the standardized variables

$\mathbf{Z}' = \left(Z_1, \cdots, Z_p\right)$ where $Z_i = \dfrac{\left(X_i - \mu_i\right)}{\sqrt{\sigma_{X_i X_i}}}$, with the covariance matrix $\text{cov}(\mathbf{Z}) = \boldsymbol{\rho}$ which is actually the correlation matrix. Consider the eigenvalue-eigenvector pairs $\left(\lambda_i, \mathbf{e_i}\right)$ computed from the correlation matrix $\boldsymbol{\rho}$. The PCs are then defined by

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'\left(\mathbf{V}^{1/2}\right)^{-1}\left(X_i - \mu_i\right) = \mathbf{e}_i'\frac{\left(X_i - \mu_i\right)}{\sqrt{\sigma_{X_i X_i}}} \quad . \tag{4.4.6}$$

The variables $\mathbf{Z}' = \left(Z_1, \cdots, Z_p\right)$ being standardized, the variance of each Z_i is unity, this means $\sqrt{\sigma_{Z_i Z_i}} = 1,\ i = 1, \cdots, p$. (4.4.7)

It follows that the correlation coefficient between the i^{th} PC Y_i and the k^{th} variable Z_k is computed by $\rho_{Y_i, Z_k} = \dfrac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{Z_k Z_k}}} = e_{ik}\sqrt{\lambda_i}$. With $1 \le i, k \le p$. (4.4.8)

Moreover, the total variation of the p variables equals to p. Actually,

$$\sum_{i=1}^{p} \text{var}\left(Y_i\right) = \sum_{i=1}^{p} \text{var}\left(Z_i\right) = \sigma_{Z_1 Z_1} + \sigma_{Z_2 Z_2} + \cdots + \sigma_{Z_p Z_p} = 1 + 1 + \cdots + 1 = p \quad . \tag{4.4.9}$$

Similar to the case where computation of PCs is done using the covariance matrix, one can compute the population variation explained by the k^{th} PC by

$$\Psi = \frac{\lambda_k}{p} \tag{4.4.10}$$

or the population variation explained by the first q PCs by

$$\Psi_q = \frac{\lambda_1 + \cdots + \lambda_q}{p} \, . \tag{4.4.11}$$

4.5 Use of Covariance or Correlation Matrices for PC Computation

In the previous sections, it was mentioned that PCs can be computed from both covariance and correlation matrices.

Theorem 4.4 The PCs computed from the covariance matrix are in general not equal to those computed from the correlation matrix [8, 9].

Consider for instance the following set of data where 5 variables are observed and defined as follow. X_1 : Height of an individual in centimeter, X_2 : weight of individual in kg X_3: shoes size of individual in centimeter X_4 : age of individual in years and X_5: monthly expenditure of individual in dollar.

Table 4.5.1: Data of individual parameters

X_1	X_2	X_3	X_4	X_5
170	90	27	28	400
160	60	27	90	1000
180	80	30	21	500
155	70	25	15	150
165	75	28	20	700
130	50	23	11	200
140	30	24	13	100
125	20	19	7	300
190	70	31	30	1500
150	40	26	17	300

The following computations based on covariance matrix and correlation matrix give an illustration of the difference between the PCs computed using one or another of those matrices.

The mean vector of the data is $\bar{\mathbf{x}} = \begin{pmatrix} 156.5 & 58.5 & 26 & 25.2 & 515 \end{pmatrix}$. (4.5.1)

It is also known that $\boldsymbol{\mu} = E(\bar{\mathbf{x}})$.

Computation based on covariance matrix and using table 4.5.1

The covariance matrix is

$$\mathbf{S} = 10^5 \begin{pmatrix} 0.0044 & 0.0039 & 0.0007 & 0.0017 & 0.0648 \\ 0.0039 & 0.0052 & 0.0006 & 0.0014 & 0.0375 \\ 0.0007 & 0.0006 & 0.0001 & 0.0003 & 0.0103 \\ 0.0017 & 0.0014 & 0.0003 & 0.0057 & 0.0599 \\ 0.0648 & 0.0375 & 0.0103 & 0.0599 & 1.9447 \end{pmatrix} \quad (4.5.2)$$

The eigenvalues and eigenvectors of $\mathbf{S}$ are;

$$\boldsymbol{\Lambda}_S = 10^5 \begin{pmatrix} 1.9495 & 0 & 0 & 0 & 0 \\ 0 & 0.0063 & 0 & 0 & 0 \\ 0 & 0 & 0.0039 & 0 & 0 \\ 0 & 0 & 0 & 0.0005 & 0 \\ 0 & 0 & 0 & 0 & 0.0000 \end{pmatrix} \quad (4.5.3)$$

and

$$\mathbf{E}_S = \begin{pmatrix} 0.0333 & -0.5457 & -0.1080 & 0.8126 & 0.1707 \\ 0.0193 & -0.8326 & 0.0568 & -0.5505 & -0.0092 \\ 0.0053 & -0.0869 & -0.0086 & 0.1472 & -0.9852 \\ 0.0308 & -0.0115 & 0.9921 & 0.1207 & 0.0105 \\ 0.9988 & 0.0351 & -0.0280 & -0.0209 & -0.0006 \end{pmatrix} \quad (4.5.4)$$

The PCs computed from this covariance matrix are then

$$\begin{aligned} Y_1^S &= 0.0333X_1 + 0.0193X_2 + 0.0053X_3 + 0.0308X_4 + 0.9988X_5 \\ Y_2^S &= -0.5457X_1 - 0.8326X_2 - 0.0869X_3 - 0.0115X_4 + 0.0351X_5 \\ Y_3^S &= -0.1080X_1 + 0.0568X_2 - 0.0086X_3 + 0.9921X_4 - 0.0280X_5 . \\ Y_4^S &= 0.8126X_1 - 0.5505X_2 + 0.1472X_3 + 0.1207X_4 - 0.0209X_5 \\ Y_5^S &= 0.1707X_1 - 0.0092X_2 - 0.9852X_3 + 0.0105X_4 - 0.0006X_5 \end{aligned} \quad (4.5.5)$$

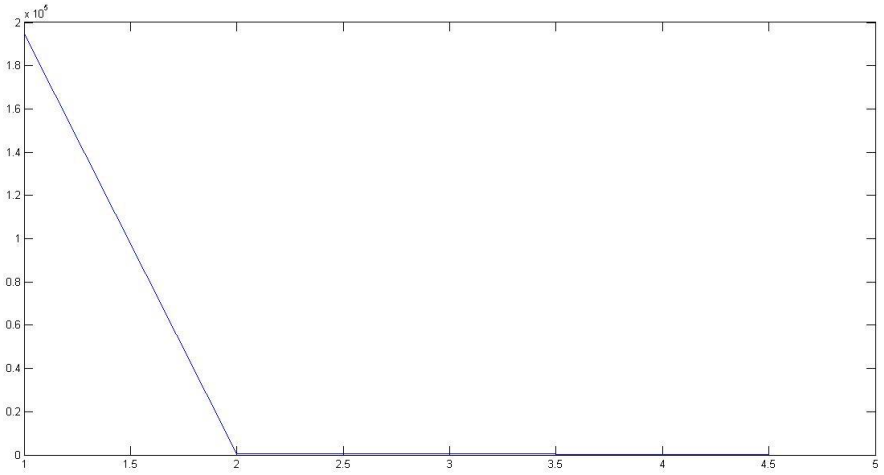

Figure 4.5.1: scree plot of table 4.5.1 from covariance matrix

Figure 4.5.1 represents the scree plot of the eigenvalues computed from the covariance matrix. The elbow bends at $i = 2$ leads to the conclusion that only the first eigenvalue (the first PC) is enough to represent the total variation of the sample without loss of information.

The variation represented by the first PC is $\Psi_1^S = \dfrac{10^5 \cdot 1.9495}{10^5 \cdot 1.9602} = 0.9945$. (4.5.6)

For instance the correlation coefficients between the first PC and the variables are

$\rho_{Y_1X_1} = 0.7009$, $\rho_{Y_1X_2} = 0.3727$, $\rho_{Y_1X_3} = 0.7400$, $\rho_{Y_1X_4} = 0.5696$ and $\rho_{Y_1X_5} = 1.0000$.

According to these results, the variable X_5 has the highest impact in the first PC, Y_1. Because it has a correlation coefficient equals to 1. In other word is 100% correlated to Y_1 . This means if the variable X_5 is omitted within the PC computation, then it will considerably affected the first PC, Y_1 . On the other hand, the variable X_2 with very low correlation to Y_1, would mean an insignificant effect on Y_1. Computation of the first PC, Y_1 with omission of X_2 ,will make a marked difference in Y_1 .

The computation based on correlation coefficient matrix and using data 4.5.1 are the following

The correlation coefficient matrix is

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.8055 & 0.9555 & 0.3336 & 0.7006 \\ 0.8055 & 1.0000 & 0.7647 & 0.2582 & 0.3717 \\ 0.9555 & 0.7647 & 1.0000 & 0.3582 & 0.6666 \\ 0.3336 & 0.2582 & 0.3582 & 1.0000 & 0.5686 \\ 0.7006 & 0.3717 & 0.6666 & 0.5686 & 1.0000 \end{pmatrix} \tag{4.5.7}$$

The eigenvalues and eigenvectors of are;

$$\mathbf{\Lambda}_R = \begin{pmatrix} 3.3960 & 0 & 0 & 0 & 0 \\ 0 & 0.9670 & 0 & 0 & 0 \\ 0 & 0 & 0.4569 & 0 & 0 \\ 0 & 0 & 0 & 0.1466 & 0 \\ 0 & 0 & 0 & 0 & 0.0335 \end{pmatrix} \tag{4.5.8}$$

and

$$\mathbf{E}_R = \begin{pmatrix} -0.5194 & -0.2226 & 0.1481 & -0.1820 & 0.7909 \\ -0.4374 & -0.4186 & -0.5429 & 0.5548 & -0.1757 \\ -0.5119 & -0.1988 & 0.1191 & -0.6130 & -0.5555 \\ -0.2988 & 0.7577 & -0.5476 & -0.1745 & 0.0794 \\ -0.4331 & 0.4020 & 0.6077 & 0.5029 & -0.1694 \end{pmatrix}. \tag{4.5.9}$$

Recall the formula (4.4.2) and the variances $\sqrt{\sigma_{X_1X_1}} = \sqrt{10^5 \cdot 0.0044} = 20.9762$,

$\sqrt{\sigma_{X_2X_2}} = \sqrt{10^5 \cdot 0.0052} = 22.8583$, $\sqrt{\sigma_{X_3X_3}} = \sqrt{10^5 \cdot 0.0001} = 3.1623$,

$\sqrt{\sigma_{X_4X_4}} = \sqrt{10^5 \cdot 0.0057} = 23.8747$ and $\sqrt{\sigma_{X_5X_5}} = \sqrt{10^5 \cdot 1.9447} = 440.9875$ the new

variables are defined by $Z_1 = \dfrac{X_1 - 156.5}{20.9762}$; $Z_2 = \dfrac{X_2 - 58.5}{22.8583}$; $Z_3 = \dfrac{X_3 - 26}{3.1623}$;

$Z_4 = \dfrac{X_4 - 25.2}{23.8747}$; $Z_5 = \dfrac{X_5 - 515}{440.9875}$;

The PCs computed from the correlation coefficient matrix are then defined by

$$
\begin{aligned}
Y_1^R &= -0.5194Z_1 - 0.4374Z_2 - 0.5119Z_3 - 0.2988Z_4 - 0.4331Z_5 \\
Y_2^R &= -0.2226Z_1 - 0.4186Z_2 - 0.1988Z_3 + 0.7577Z_4 + 0.4020Z_5 \\
Y_3^R &= 0.1481Z_1 - 0.5429Z_2 + 0.1191Z_3 - 0.5476Z_4 + 0.6077Z_5 \\
Y_4^R &= -0.1820Z_1 + 0.5548Z_2 - 0.6130Z_3 - 0.1745Z_4 + 0.5029Z_5 \\
Y_5^R &= 0.7909Z_1 - 0.1757Z_2 - 0.5555Z_3 + 0.0794Z_4 - 0.1694Z_5
\end{aligned}
\qquad (4.5.10)
$$

After expanding this expression in term of the variable $Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{X_i X_i}}}$, the equation (4.5.10) becomes

$$
\begin{aligned}
Y_1^R &= -0.0248X_1 - 0.0191X_2 - 0.1619X_3 - 0.0125X_4 - 0.0009X_5 + cte_1 \\
Y_2^R &= -0.0106X_1 - 0.0183X_2 - 0.0629X_3 + 0.0317X_4 + 0.0009X_5 + cte_2 \\
Y_3^R &= 0.0070X_1 - 0.0236X_2 + 0.0377X_3 - 0.0229X_4 + 0.0014X_5 + cte_3 \\
Y_4^R &= -0.0087X_1 + 0.0243X_2 - 0.1938X_3 - 0.0073X_4 + 0.0011X_5 + cte_4 \\
Y_5^R &= 0.0377X_1 - 0.0077X_2 - 0.1757X_3 + 0.0033X_4 - 0.0003X_5 + cte_5
\end{aligned}
\quad , \qquad (4.5.11)
$$

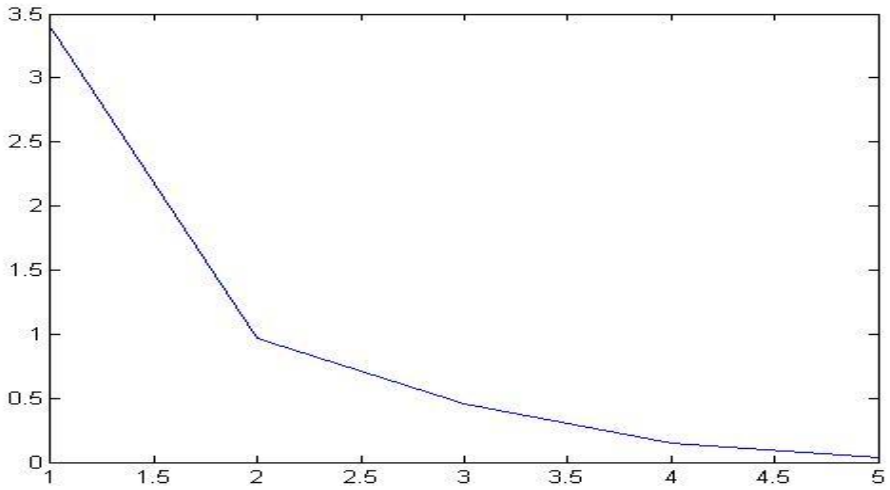

Figure 4.5.2: scree plot of table 4.5.1 from correlation matrix

The variation represented by the first PC is $\Psi_1^R = \frac{3.3960}{5} = 0.6792$, and the variation represented by the first two PCs is $\Psi_2^R = \frac{3.3960 + 0.967}{5} = 0.8726$. So the first two

PCs are enough in this case for the representation of the whole data variation without loss of information. For instance the correlation coefficients between the first PC and the variables are

$\rho_{Y_1Z_1} = -0.9572$, $\rho_{Y_1Z_2} = -0.8060$, $\rho_{Y_1Z_3} = -0.9433$, $\rho_{Y_1Z_4} = -0.5506$ and $\rho_{Y_1Z_5} = -0.7981$

Based on all the computation done so far, the PCs computed from the covariance matrix are different from the PCs computed from the correlation matrix. The relation or the significance of variables in PCs computed using covariance or correlation matrix is different from one case to another. Furthermore, there is neither linear relation nor logarithmic relation between the PCs computed using one or another matrix. This leads to the conclusion that the standardizing of variables has an effect on the PCs' computation. The standardizing of variables is usually required for inhomogeneous variables or for variables which have very high range of variation.

Theorem 4.5 The correlation coefficient matrix is suitable for PCs computation when there are large variations between the data values among the variables and/or when it is an inhomogeneous data [12].

Consider for instance the following set of data representing the measurement of three parameters of an individual, with the variables X_1, X_2 and X_3 being respectively Number of working hours/day, Monthly salary(in Euro) and Years of experience.

Table 4.5.2: Salary

Fields	Number of working hour / day (X_1)	Monthly salary (in Euro) (X_2)	Years of experience (X_3)
Civil Engineer	7	1000	8

Computer Engineer	10	1500	12
Farmer	6	1000	17
Goods deliverer	8	1500	15
Baker	9	1700	11
University lecturer	7	1500	10
High school teacher	6	1500	13
Shop keeper	15	1000	13
Hotel manager	10	1700	19
Waitress	12	1800	10

Table 4.5.2 represents the monthly salaries of ten employees from different sectors, number of working hours per day and number ofyears of experience. Difference between the values of variable X_2 and the other two variables is clearly evident. The PCs will be computed using both covariance and correlation matrix and a conclusion will be drawn.

Mean of three variables are $\bar{\mathbf{x}} = (9;\ 1420;\ 12.8)$. It is also known that $\boldsymbol{\mu} = E(\bar{\mathbf{x}})$.

Computation based on covariance matrix and using data from Table 4.5.2 is

$$\mathbf{S} = \begin{pmatrix} 8.22 & 66.67 & -0.22 \\ 66.67 & 95111 & 26.67 \\ -0.22 & 26.67 & 11.51 \end{pmatrix} \tag{4.5.12}$$

The eigenvalues matrix and its corresponding eigenvectors matrix are

$$\boldsymbol{\Lambda}_S = 10^4 * \begin{pmatrix} 9.5111 & 0 & 0 \\ 0 & 0.0012 & 0 \\ 0 & 0 & 0.0008 \end{pmatrix} \tag{4.5.13}$$

and

$$\mathbf{E}_S = \begin{pmatrix} -0.0007 & -0.0718 & -0.9974 \\ -1.0000 & -0.0002 & 0.0007 \\ -0.0003 & 0.9974 & -0.0718 \end{pmatrix} \tag{4.5.14}$$

The PCs computed from the covariance matrix are

$$\begin{aligned} Y_1^S &= -0.0007X_1 - X_2 - 0.0003X_3 \\ Y_2^S &= -0.0718X_1 - 0.0002X_2 + 0.9974X_3 \\ Y_3^S &= -0.9974X_1 + 0.0007X_2 - 0.0718X_3 \end{aligned} \tag{4.5.15}$$

Computation based on correlation coefficient matrix and using data 4.5.2

The correlation coefficient matrix is

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.0754 & -0.0228 \\ 0.0754 & 1.0000 & 0.0255 \\ -0.0228 & 0.0255 & 1.0000 \end{pmatrix} \tag{4.5.16}$$

its eigenvalues matrix is

$$\mathbf{\Lambda}_R = \begin{pmatrix} 1.0754 & 0 & 0 \\ 0 & 1.0131 & 0 \\ 0 & 0 & 0.9114 \end{pmatrix} \tag{4.5.17}$$

The corresponding eigenvectors matrix is

$$\mathbf{E}_R = \begin{pmatrix} -0.7024 & -0.2744 & 0.6568 \\ -0.7112 & 0.2348 & -0.6626 \\ -0.0276 & 0.9325 & 0.3600 \end{pmatrix} \tag{4.5.18}$$

The PCs computed from the correlation coefficient matrix are then defined by

$$\begin{aligned} Y_1^R &= -0.7024Z_1 - 0.7112Z_2 - 0.0276Z_3 \\ Y_2^R &= -0.2744Z_1 + 0.2348Z_2 + 0.9325Z_3 \text{ , where } Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{X_i X_i}}} . \\ Y_3^R &= 0.6568Z_1 - 0.6626Z_2 + 0.3600Z_3 \end{aligned} \tag{4.5.19}$$

From the foregoing computation it is observed that, the use of the covariance matrix for PCs computation is not efficient enough. Consider for instance the first PC computed from the covariance matrix $Y_1^S = -0.0007X_1 - X_2 - 0.0003X_3$, the variable X_2 is well represented although it is negatively represented, whereas the variables X_1 and X_3 are almost insignificant because there are about 10000 times less representative in the data variation. Consider the first PC computed from the correlation coefficient matrix $Y_1^R = -0.7024Z_1 - 0.7112Z_2 - 0.0276Z_3$, the entire variables are represented the first two have almost the same range of representation whereas the third variable is just about 10 times less representative than other in the data variation. This means for a large scale data, Correlation matrix is more efficient for PCs computation. Furthermore, consider table 4.5.3for a comparative study between the computation of PCs from table 4.5.2 using the covariance on one hand and the correlation matrix on other hand.

Table 4.5.3: ratio between covariance and correlation matrix

	λ_1	λ_2	λ_3
Covariance matrix	$9.5111*10^4$	$0.0012*10^4$	$0.0008*10^4$
Correlation matrix	1.0754	1.0131	0.9114
Ratio $\lambda_i^S / \lambda_i^R (X_i)$	$8.844*10^4$	11.84	8.78

The ratio $\lambda_i^S / \lambda_i^R$ fluctuates widely from$8.844*10^4$ to8.78. This means there is not neither a linear relation nor a logarithmic relation between the PCs computed using covariance and those computed using correlation coefficient matrix.

For more illustration of the concept of large scale data and its effect on PCs computation using covariance or correlation matrix, the following data of same magnitude representing the marks score by 10 students in four subjects X_1, X_2, X_3, X_4. The grading system is out of 20, this means the data record has a normal scale.

Table 4.5.4: Students marks

Student	X_1	X_2	X_3	X_4	student	X_1	X_2	X_3	X_4
1	12	9	11	16	6	14	15	10	11
2	12	13	12	10	7	8	11	13	15
3	18	14	10	9	8	11	15	17	10
4	10	11	17	15	9	18	17	18	7
5	13	19	9	15	10	12	16	17	11

This data is homogenous in nature since the grading system here is out of 20, the difference or the range between marks can not be more than 20.

The mean vector of the data is computed as $\mathbf{x} = (12.8 \;\; 14 \;\; 13.4 \;\; 11.9)$ (4.5.20)

Covariance matrix obtained using data from Table 4.5.3 is

$$\mathbf{S} = \begin{pmatrix} 10.1778 & 4.6667 & -1.4667 & -6.9111 \\ 4.6667 & 9.3333 & 0.3333 & -4.5556 \\ -1.4667 & 0.3333 & 12.2667 & -3.4000 \\ -6.9111 & -4.5556 & -3.4000 & 9.6556 \end{pmatrix} \quad (4.5.21)$$

The eigenvalues matrix is

$$\mathbf{\Lambda_S} = \begin{pmatrix} 20.8197 & 0 & 0 & 0 \\ 0 & 13.2280 & 0 & 0 \\ 0 & 0 & 5.5525 & 0 \\ 0 & 0 & 0 & 1.8331 \end{pmatrix} \tag{4.5.22}$$

Its corresponding eigenvectors matrix is

$$\mathbf{E_S} = \begin{pmatrix} -0.5934 & -0.3064 & -0.3979 & 0.6290 \\ -0.4905 & -0.1194 & 0.8628 & 0.0249 \\ -0.1626 & 0.9334 & 0.0275 & 0.3186 \\ 0.6171 & -0.1435 & 0.3105 & 0.7087 \end{pmatrix} \tag{4.5.23}$$

The PCs computed from the covariance matrix are

$$\begin{aligned} Y_1^S &= -0.5934X_1 - 0.4905X_2 - 0.1626X_3 + 0.6171X_4 \\ Y_2^S &= -0.3064X_1 - 0.1194X_2 + 0.9334X_3 - 0.1435X_4 \\ Y_3^S &= -0.3979X_1 + 0.8628X_2 + 0.0275X_3 + 0.3105X_4 \\ Y_4^S &= 0.6290X_1 + 0.0249X_2 + 0.3186X_3 + 0.7087X_4 \end{aligned} \tag{4.5.24}$$

The total variance expressed by the first PC, the first two PCs and the first three PCs are respectively $\Psi_1^S = \frac{20.8197}{41.4333} = 0.5025$, $\Psi_2^S = \frac{20.8197 + 13.2280}{41.4333} = 0.8217$ and $\Psi_3^S = \frac{20.8197 + 13.2280 + 5.5525}{41.4333} = 0.9558$.

Computation based on correlation coefficient matrix and using data 4.5.3

The correlation coefficient matrix is

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.4788 & -0.1313 & -0.6972 \\ 0.4788 & 1.0000 & 0.0312 & -0.4799 \\ -0.1313 & 0.0312 & 1.0000 & -0.3124 \\ -0.6972 & -0.4799 & -0.3124 & 1.0000 \end{pmatrix} \tag{4.5.25}$$

The eigenvalues matrix computed from the correlation coefficient matrix is

$$\mathbf{\Lambda}_R = \begin{pmatrix} 2.1256 & 0 & 0 & 0 \\ 0 & 1.1093 & 0 & 0 \\ 0 & 0 & 0.5843 & 0 \\ 0 & 0 & 0 & 0.1808 \end{pmatrix} \tag{4.5.26}$$

Its corresponding eigenvectors matrix is

$$\mathbf{E}_R = \begin{pmatrix} -0.5861 & -0.2979 & -0.4124 & 0.6306 \\ -0.5147 & -0.1205 & 0.8487 & 0.0197 \\ -0.1165 & 0.9234 & 0.0520 & 0.3620 \\ 0.6148 & -0.2099 & 0.3271 & 0.6863 \end{pmatrix} \tag{4.5.27}$$

Recall the formula $Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{X_i X_i}}}$ and the variances $\sqrt{\sigma_{X_1 X_1}} = \sqrt{10.1778} = 3.1903$,

$\sqrt{\sigma_{X_2 X_2}} = \sqrt{9.3333} = 3.055$, $\sqrt{\sigma_{X_3 X_3}} = \sqrt{12.2667} = 3.5024$ and

$\sqrt{\sigma_{X_4 X_4}} = \sqrt{9.6556} = 3.1073$ the new variables are defined by $Z_1 = \frac{X_1 - 12.8}{3.1903}$;

$Z_2 = \frac{X_2 - 14}{3.055}$; $Z_3 = \frac{X_3 - 13.4}{3.5024}$; $Z_4 = \frac{X_4 - 11.9}{3.1073}$.

The PCs computed from the correlation matrix are

$$\begin{aligned} Y_1^R &= -0.5861Z_1 - 0.5147Z_2 - 0.1165Z_3 + 0.6148Z_4 \\ Y_2^R &= -0.2979Z_1 - 0.1205Z_2 + 0.9234Z_3 - 0.2099Z_4 \\ Y_3^R &= -0.4124Z_1 + 0.8487Z_2 + 0.0520Z_3 + 0.3271Z_4 \\ Y_4^R &= 0.6306Z_1 + 0.0197Z_2 + 0.3620Z_3 + 0.6863Z_4 \end{aligned} \tag{4.5.28).}$$

After expanding this expression in term of the variable $Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{X_i X_i}}}$, the equation

(4.5.28) becomes

$$\begin{aligned} Y_1^R &= -0.1837\mathrm{X}_1 - 0.2505\mathrm{X}_2 - 0.0333\mathrm{X}_3 + 0.1979\mathrm{X}_4 + cte_1 \\ Y_2^R &= -0.0934\mathrm{X}_1 - 0.0394\mathrm{X}_2 + 0.2636\mathrm{X}_3 - 0.0676\mathrm{X}_4 + cte_2 \\ Y_3^R &= -0.1293\mathrm{X}_1 + 0.2778\mathrm{X}_2 + 0.0148\mathrm{X}_3 + 0.1053\mathrm{X}_4 + cte_3 \\ Y_4^R &= 0.1977\mathrm{X}_1 + 0.0064\mathrm{X}_2 + 0.1034\mathrm{X}_3 + 0.2209\mathrm{X}_4 + cte_4 \end{aligned} \tag{4.5.29}$$

The total variance expressed by the first PC , the first two PCs and the first three PCs are respectively $\Psi_1^R = \frac{2.1256}{4} = 0.5314$, $\Psi_2^R = \frac{2.1256+1.1093}{4} = 0.8087$ and $\Psi_3^R = \frac{2.1256+1.1093+0.5843}{4} = 0.9548$.

Consider the following two tables for a comparative study between the computation of PCs from table 4.5.4 using the covariance on one hand and the correlation matrix on other hand.

Table 4.5.5: ratio of covariance and correlation matrix

	λ_1	λ_2	λ_3	λ_4
Covariance matrix	20.8197	13.2280	5.5525	1.8331
Correlation matrix	2.1256	1.1093	0.5843	0.1808
Ratio $\lambda_i^S / \lambda_i^R$ (X_i)	9.7947	11.9246	9.5028	10.1388

The mean of the Ratio $\lambda_i^S / \lambda_i^R$ is $\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = 10.34$ with a standard deviation of $\sigma_x = \sqrt{\frac{1}{3}\sum_{i=1}^{4}(x_i - \bar{x})^2} = 1.0877$. It can be concluded that there exists a linear correlation between the eigenvalues $\lambda_i^{\mathbf{S}}$ computed from covariance matrix and those λ_i^R computed from correlation coefficient matrix. The eigenvalues computed from the covariance matrix are approximately 10 times those computed from the correlation matrix , when there are considered in the ascending order.

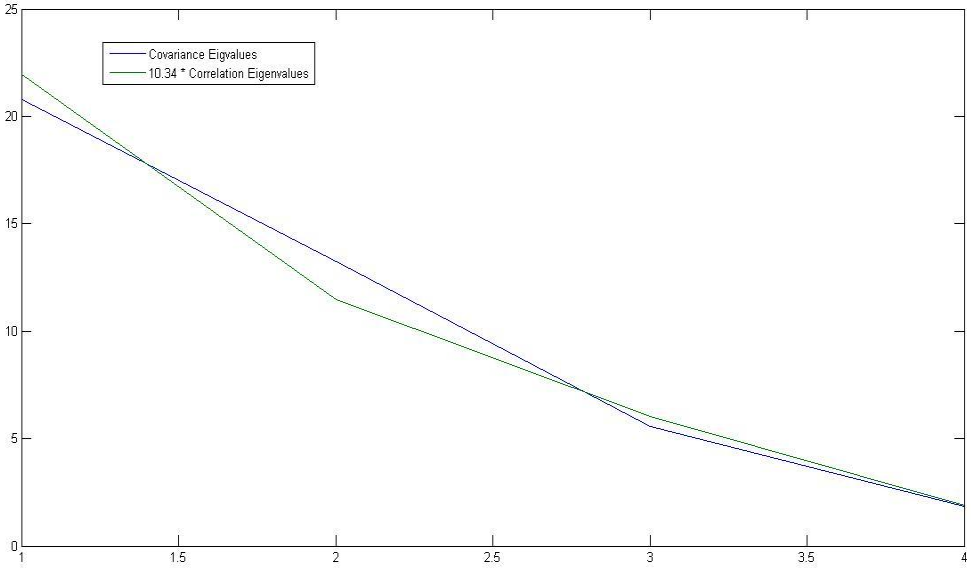

Figure 4.5.3: Scree plot from table 4.5.4 using covariance & correlation matrix

Figure 4.5.3 is a join plotting of the covariance matrix eigenvalues and the correlation matrix eigenvalues which were multiplied by the estimated correlation coefficient $k = 10.34$. It is clear that both of graphs have same variation. They are likely the same.

Table 4.5.6 : Percentage of variation due to cumulative PCs

	Variation attributable to PC1	Variation attributable to PC1 & PC2	Variation attributable to PC1, PC2 & PC3
Covariance matrix	0.5025	0.8217	0.9558
Correlation matrix	0.5314	0.8087	0.9548

From Table 4.5.6 it is evident that when variation between variables is not significant, use of the covariance or correlation matrices to generate the PCs does not make any significant difference. However, there is significant difference between the variation represented by PCs obtained using the covariance and correlation matrices

of the data from Table 4.5.2. This is attributable to the significant difference between the magnitudes of the variables X_2 and those of X_1 and X_3.

Result 4.2 The correlation coefficient matrix is suitable for the PC computation when there is significant difference between the magnitudes of data from different variables, or in the case of inhomogeneous data, i.e. different units for each variable. In case of a standard normal data, $\mathbf{S} = \mathbf{R}$, meaning PCs computed from either covariance or correlation matrix. Before deciding to standardize the data, it is worth computing the PCs from $\mathbf{S}$ and $\mathbf{R}$ matrices. Comparison of the eigenvalues obtained from **S** and **R** matrices will give an initial idea as to the need for standardization or not. A good guide will be the closeness of the eigenvalues obtained from $\mathbf{S}$ and $\mathbf{R}$ matrices. Opinion on the closeness will greatly depend on the nature of the data under study.

4.6 PCA for Outlier Detection and Quality Monitoring

In data collection process and analysis involving a multivariate process where the number of variables may be very large, errors may occur mainly due to human error, or computational errors during data validation. On the other hand some extreme values may not be due to error, but correct values. Such values are called *outliers*. Detection of outliers is crucial, as they may be treated as a separate subset and analyzed separately. Intuitive detection of outliers and suggestion of a solution requires a robust knowledge of the process under study. For example if a value of 93°C is encountered in a variable representing human body temperature, it is an obvious human error in entering data, since the average temperature of a healthy human body is 37°C.Outliers can also be detected through the inspection of the scatter plot of the data.

Consider $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ to be multivariate normally distributed random variables with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The centered data is defined as $\mathbf{X} - \boldsymbol{\mu}$. Assume the first two PCs of the centered data $Y_{j1} = e_1'(\mathbf{x}_j - \bar{\mathbf{x}})$ and $Y_{j2} = e_2'(\mathbf{x}_j - \bar{\mathbf{x}})$ represent above 80% of total variation in the data. Then first two PCs can be used to draw a control ellipse chart. In cases where more than 2 PCs are needed to represent above 80% of total variation in the data, pairwise control ellipses can be generated to check anomalies stemming from different variables. Use of the pairwise PCs leads to2-dimensional representation which is suitable for visualization. The control ellipse chart is defined using the first two PCs by $\frac{Y_1^2}{\lambda_1} + \frac{Y_2^2}{\lambda_2} \leq \chi_2^2(\alpha)$ [17, 29, 31].

Since the considered sample has p PCs, the remaining $p-2$ PCs are used to draw a T^2-chart defined by $T_j^2 = \frac{Y_{j3}^2}{\lambda_3} + \frac{Y_{j4}^2}{\lambda_4} + \cdots + \frac{Y_{jp}^2}{\lambda_p}$, with an upper control limit (UCL) given by $\chi_{p-2}^2(\alpha)$. The T^2-chart is a 2-dimensional graph with index 1 to n on the x-axis (n being the number of observations) and the corresponding computed value $T_j^2 = \frac{Y_{j3}^2}{\lambda_3} + \frac{Y_{j4}^2}{\lambda_4} + \cdots + \frac{Y_{jp}^2}{\lambda_p}$ on the y-axis [21].

Considering the n observations of p random variables $X_1, X_2, \cdots, X_p$, the quality monitoring procedure is done following three steps.

Step 1: Computation of the p PCs and plotting the scatter diagram of the first two PCs Y_1 versus Y_2.

Step 2: Computation and plotting of the control ellipse on the scatter diagram drawn in step 1.

The points from step 1 which fall out of the control ellipse are considered to be outliers. Depending on the process from which data comes from, the cause behind the detected outliers can intuitively be identified. This may be due to human error during data collection, entry, or validation process. Alternately, outliers can be correct values, but due to some conditions of the process they are generated and valid data values. Dealing with outlier data is handled by a branch of statistics named "extreme value analysis".

When the first 2 or 3 PCs do not represent a high percentage of total variation (60% to 80%), then an analysis of the remaining *p-2* or *p-3* PCs through the T^2-chart can be undertaken. In the T^2-chartpoints which are over the UCL are considered to be outliers and the same explanation offered under step 2 is valid.

It is sometime suitable and easy to process the quality monitoring procedure based only on the T^2-chart. In which case, the entire *p* PCs are used to plot the T^2-chart and there is no need to compute and draw the control ellipse chart.

For experiment which results data oscillate between a positive maximum and a negative minimum, there is a need to define also a lower control limit (LCL) as well as the UCL.

Example 4.6.1 Consider the results of 20 students in 4 courses X_1, X_2, X_3 and X_4 where the grading system is out of 20. The marks are given in the following table.

The question here is to use quality monitoring procedure to detect if there are outliers in the data.

Table 4.6.1: Student mark for outliers

Student	X_1	X_2	X_3	X_4	Student	X_1	X_2	X_3	X_4
1	12	9	11	16	11	14	15	10	11
2	12	13	12	10	12	8	11	13	15
3	18	14	10	9	13	11	15	17	10
4	10	11	17	15	14	18	17	18	7
5	13	19	9	15	15	12	16	17	11
6	11	28	12	4	16	12	13	17	10
7	12	13	14	17	17	9	11	10	11
8	40	8	7	10	18	18	10	13	15
9	10	12	12	13	19	14	11	10	7
10	6	3	5	4	20	9	5	1	15

Solution: The value 40 for student 8 seems to be an outlier because the grading system here is out of 20. The following computation is to check at a certain confidence level, whether the value 40 will be detected as an outlier or not. Following the establishment of 4 PCs, the corresponding PC values are computed by substituting X_i values from Table 4.6.1. Resulting $\hat{y}_i$ values are given in Table 4.6.2.

Table 4.6.2: PCs scores for 20 students marks.

Period	$\hat{Y}_{j1}$	$\hat{Y}_{j2}$	$\hat{Y}_{j3}$	$\hat{Y}_{j4}$
1	-1.1500	-4.6242	-3.8316	-1.2371

2	-1.4338	0.4210	1.0265	0.7284
3	4.6635	1.3131	2.6046	-0.5113
4	-4.0458	-0.0103	-5.9955	2.0254
5	-0.9659	2.9340	-0.1979	-7.2067
6	-3.7855	13.8545	8.5094	-3.7460
7	-2.0130	-0.0158	-5.6560	-2.0675
8	27.3196	-2.1274	0.4758	0.0636
9	-3.4343	-1.3182	-1.3323	-0.4585
10	-5.0844	-10.9982	8.6074	5.1788
11	0.5204	1.1327	1.5726	-2.1658
12	-5.5135	-2.3405	-3.4845	-0.4467
13	-3.2661	4.4646	-1.4359	2.7850
14	3.4234	8.2689	0.0591	4.0275
15	-2.4380	5.2256	-2.1388	1.6710
16	-2.0615	2.9669	-1.8825	3.8368
17	-3.9645	-2.8897	1.3227	-0.0109
18	4.4451	-1.6954	-4.5808	0.0099
19	1.1473	-1.3246	4.0169	2.2608
20	-2.3630	-13.2370	2.3410	-4.7366

The axes of ellipsoid are computed using the following formula $\frac{\hat{Y}_1}{\lambda_1}+\frac{\hat{Y}_2}{\lambda_2}\le\chi_2^2(0.05)$ where $\lambda_1=50.0233$; $\lambda_2=35.2862$ and $\chi_2^2(0.05)=5.99$. The major and the minor semi-axes of the ellipsoid are therefore M=$\sqrt{\chi_2^2(0.05)*\lambda_1}=17.31$ and m=$\sqrt{\chi_2^2(0.05)*\lambda_2}=14.53$.

Scatter diagram of the first 2 PCs $\hat{y}_1$, $\hat{y}_2$ and the control ellipse are shown in Figure 4.6.1.

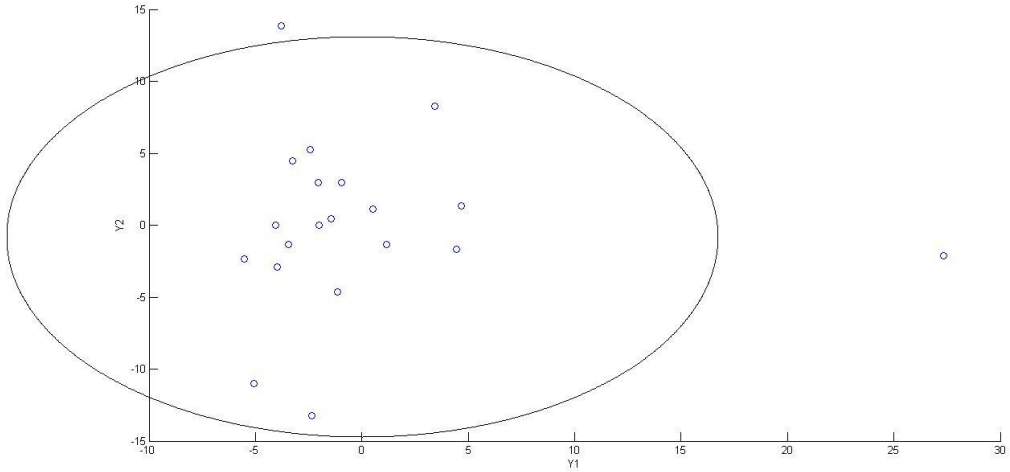

Figure 4.6.1: PC1 versus PC2 from table 4.6.1

From figure 4.6.1, two outliers are detected. The one on the right hand side has a PC1 value well out of the control ellipse with a value around $\hat{y}_1 = 25$. An inspection of $\hat{y}_1$ values identifies this as the 8th value of PC1. This corresponds to the 8th value in X_1 which is the outlier value 40. A similar inspection of the PC2 value around $\hat{y}_1 = 14$ falling out of the control ellipse, points to X_2 variable's 6th value $x_{26} = 28$ which is also an outlier.

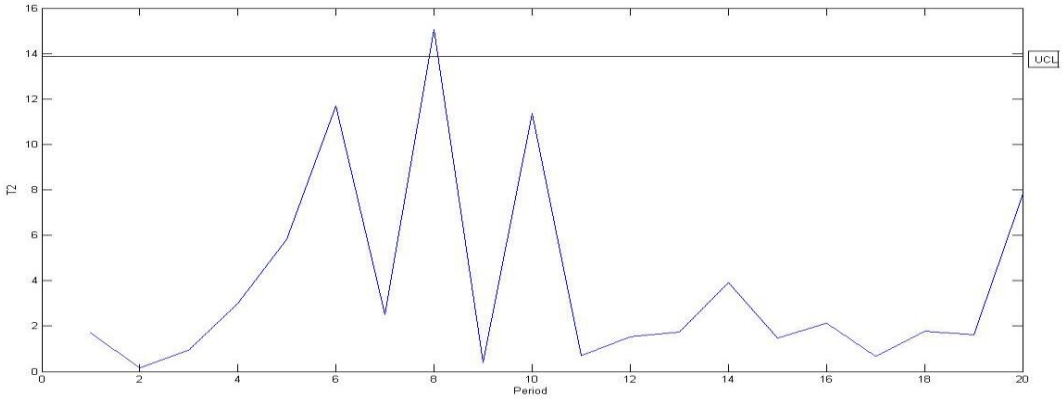

Figure 4.6.2: T2 control chart from table 4.6.1

The UCL is computed using the formula $T^2(p,n) = \frac{np}{n-p+1} F_{p,n-p+1}$ as result UCL=13.98.From figure 4.6.2, there is one value which seems to be an outlier. That value is observed at the period 8. It is obvious that this value is the entrance 40, at the eighth observation.

4.7 Controlling Future Values

In the previous section, from the table 4.6.1, and all the computation done, two outliers where detected. Assuming that a process is stable over the time, one can delete the outliers and use the remaining data to build a new control ellipsoid. This new ellipsoid is use for the prediction of the future PCs.

Consider for instance the following data comes from the deletion of the outliers which were previously observed at the entrance 6 and 8 of data from Table4.6.1. The sixth student had a mark of 28 out of 20 and the eighth student had a mark of 40 out of 20. The following new table consisted of 18 observations of 4 variables is obtained.

Table 4.7.1: marks of students after outliers are deleted

Student	X_1	X_2	X_3	X_4	Student	X_1	X_2	X_3	X_4
1	12	9	11	16	11	14	15	10	11
2	12	13	12	10	12	8	11	13	15
3	18	14	10	9	13	11	15	17	10
4	10	11	17	15	14	18	17	18	7
5	13	19	9	15	15	12	16	17	11
6					16	12	13	17	10
7	12	13	14	17	17	9	11	10	11

8					18	18	10	13	15
9	10	12	12	13	19	14	11	10	7
10	6	3	5	4	20	9	5	1	15

Using the data from table 4.7.1, the PCs and their scores are computed as follow are recapitulated in the following table.

Table 4.7.2: PC scores from marks without outliers.

Period	$\hat{Y}_{j1}$	$\hat{Y}_{j2}$	$\hat{Y}_{j3}$	$\hat{Y}_{j4}$
1	2.5638	4.0800	0.4970	-2.3503
2	-0.4712	-1.6454	-0.1561	0.9027
3	-1.9454	-3.9776	5.3890	-1.0051
4	-2.0509	3.9779	-4.3207	-1.6546
5	-2.5338	2.9007	5.0506	5.2582
6	-1.9865	5.3328	-0.0949	-0.5133
7	0.8427	1.6768	-1.2649	0.9901
8	12.8893	-6.9319	-3.8473	-0.0593
9	-1.0826	-1.1938	3.2495	1.8107
10	1.4809	4.0955	-3.1939	0.8441
11	-4.7881	-1.0484	-3.4546	1.2600
12	-9.2983	-5.3043	0.7959	-1.2853
13	-5.7968	-0.2619	-2.2884	1.3224
14	-3.9409	-1.2819	-3.3119	-0.7912
15	3.2576	-0.2313	-1.3577	1.7107
16	-1.6862	2.0386	3.5174	-5.6132
17	1.4406	-5.1568	1.3807	-0.7262

18	13.1057	2.9311	3.4102	-0.1004

The axes of ellipsoid are computed using the following formula $\frac{\hat{Y}_1}{\lambda_1}+\frac{\hat{Y}_2}{\lambda_2}\le\chi_2^2(0.05)$

where $\lambda_1=31.8334$; $\lambda_2=13.3678$ and $\chi_2^2(0.05)=5.99$. The major and the minor semi-axes of the ellipsoid are therefore $M=\sqrt{\chi_2^2(0.05)*\lambda_1}=13.8088$ and $m=\sqrt{\chi_2^2(0.05)*\lambda_2}=8.9484$ (20)

The control ellipsoid for future values monitoring is the following.

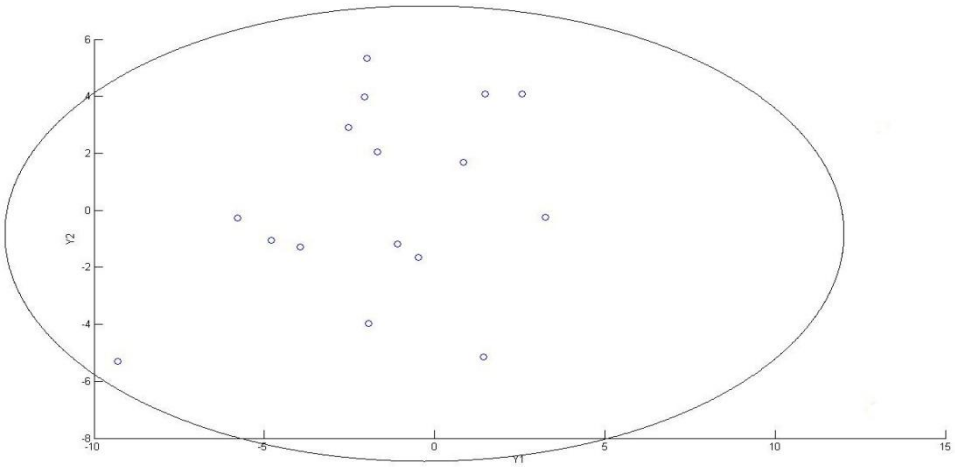

Figure 4.7.1: Control ellipsoid chart for future values monitoring from table 47.1

The T^2 control chart and its UCL are computed and plot based on the following computation.

$T^2(p,n)=\frac{np}{n-p+1}F_{p,n-p+1}$ with $p=4$ and $n=18$. UCL=11.328

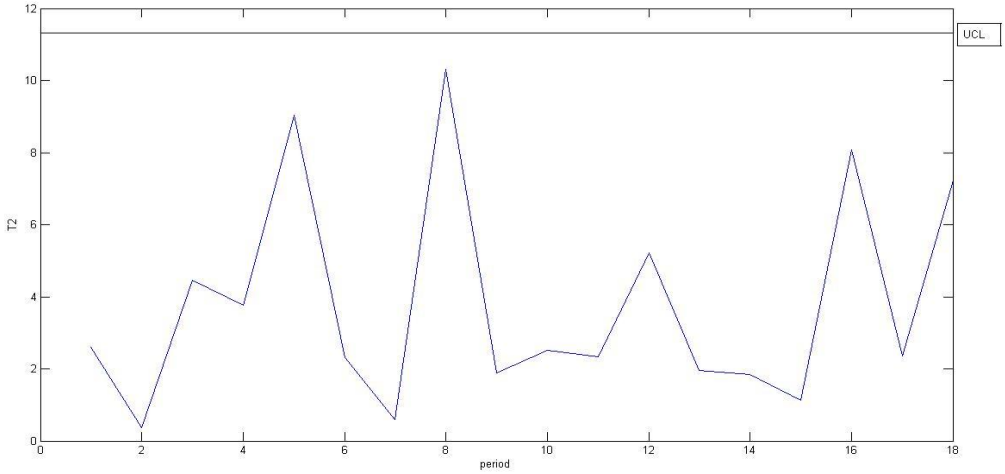

Figure 4.7.2: T^2 chart of data mark for prediction without outliers

From figure 4.7.1 and figure 4.7.2 the process is assumed to be under control because there are not outliers. Assuming that the process is stable over the time, this means for each period the conditions under which the process is done is the same, the bounds of the futures values based are known based on figure 4.7.1 and figure 4.7.2 .

Chapter 5

CASE STUDY : SOLVING PROBLEMS USING PRINCIPAL COMPONENTS ANALYSIS

Having set the framework for PC computation analysis and interpretation in Chapter4, the PCA methodology is implemented in two separate fields of application.

5.1 Case Study 1

The following dataset comes from population census of 1992 in Cameroon. It summarizes the population into 12 subgroups according to age and administrative districts. Age groups are defined as random variables and observations from each district forms the data. The variables are defined as follow where each variable represents a certain age group in years and no overlap between groups [32].

Table 5.1: Population census data

Variable	X_1	X_2	X_3	X_4	X_5	X_6
Years	<1	$1-4$	$5-9$	$10-14$	$15-19$	$20-24$

Variable	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Years	$25-29$	$30-34$	$35-39$	$40-44$	$45-49$	≥ 50

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
14	65	98	70	78	90	95	53	37	27	16	27

17	81	96	111	75	81	101	91	42	23	24	14
13	94	146	141	134	194	213	171	106	93	77	190
7	25	53	22	26	44	45	48	25	18	9	12
76	495	765	574	407	273	228	140	53	34	20	81
7	30	43	61	60	67	77	58	46	21	10	16
7	27	32	26	35	30	47	26	14	5	6	8
2	43	55	54	41	60	43	43	12	16	6	5
4	22	23	24	9	14	22	5	1	2	1	3
23	107	142	127	122	173	193	118	63	30	19	34
7	92	253	201	116	94	77	62	36	12	19	44
49	330	634	440	434	428	413	288	192	177	125	344
42	220	267	201	326	451	472	418	337	335	244	840
19	98	124	157	256	340	271	172	151	73	54	111
8	88	115	88	125	162	158	98	76	50	21	50
12	98	131	131	195	276	262	144	102	47	34	50
6	20	26	20	68	44	47	42	36	24	13	29
9	49	72	72	110	144	135	100	78	47	33	57
14	113	154	180	142	149	187	103	47	28	19	25
0	8	17	8	10	20	24	17	9	6	4	2
5	50	45	64	54	80	84	53	43	33	20	36
31	234	392	333	247	190	188	131	67	31	18	29
40	377	587	572	506	627	679	861	541	628	480	2125
4	46	70	84	116	141	153	150	135	113	74	103

10	72	91	81	121	167	170	127	102	68	50	142
14	149	358	347	524	576	719	643	517	509	336	1075
6	47	66	52	92	128	141	86	61	42	30	42
31	214	368	267	367	517	542	449	361	309	265	460
0	6	8	7	12	16	27	6	6	6	2	4
17	145	242	297	330	393	393	303	213	183	105	258
6	47	65	72	107	131	132	88	50	36	14	47
2	34	47	60	60	80	75	75	41	29	13	37
8	36	62	91	110	149	180	96	49	19	18	16
28	118	130	176	162	235	263	164	114	76	40	66
3	16	24	23	33	61	66	60	44	28	15	31
35	197	282	311	280	347	313	220	150	94	62	156
12	71	91	77	68	97	94	65	26	20	9	12
22	92	133	130	94	104	101	88	45	17	13	11
5	40	40	38	39	30	42	36	19	16	11	6
4	33	45	42	23	34	34	19	6	4	0	1
6	14	16	25	20	33	33	18	9	2	3	4
3	27	30	41	42	75	52	27	8	3	5	2

From the data, the correlation coefficient matrix and its 3D representation are computed and represented as follow

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.9733 & 0.9712 & 0.9335 & 0.8147 & 0.7017 & 0.6551 & 0.5992 & 0.5638 & 0.5486 & 0.5491 & 0.5464 \\ 0.9733 & 1.0000 & 0.9994 & 0.9892 & 0.9141 & 0.8309 & 0.7967 & 0.7605 & 0.7288 & 0.7201 & 0.7209 & 0.7217 \\ 0.9712 & 0.9994 & 1.0000 & 0.9908 & 0.9197 & 0.8377 & 0.8039 & 0.7663 & 0.7358 & 0.7261 & 0.7267 & 0.7259 \\ 0.9335 & 0.9892 & 0.9908 & 1.0000 & 0.9621 & 0.9014 & 0.8748 & 0.8440 & 0.8184 & 0.8095 & 0.8097 & 0.8074 \\ 0.8147 & 0.9141 & 0.9197 & 0.9621 & 1.0000 & 0.9844 & 0.9714 & 0.9451 & 0.9344 & 0.9223 & 0.9210 & 0.9086 \\ 0.7017 & 0.8309 & 0.8377 & 0.9014 & 0.9844 & 1.0000 & 0.9977 & 0.9817 & 0.9792 & 0.9686 & 0.9670 & 0.9523 \\ 0.6551 & 0.7967 & 0.8039 & 0.8748 & 0.9714 & 0.9977 & 1.0000 & 0.9908 & 0.9905 & 0.9820 & 0.9805 & 0.9669 \\ 0.5992 & 0.7605 & 0.7663 & 0.8440 & 0.9451 & 0.9817 & 0.9908 & 1.0000 & 0.9981 & 0.9979 & 0.9976 & 0.9925 \\ 0.5638 & 0.7288 & 0.7358 & 0.8184 & 0.9344 & 0.9792 & 0.9905 & 0.9981 & 1.0000 & 0.9980 & 0.9972 & 0.9885 \\ 0.5486 & 0.7201 & 0.7261 & 0.8095 & 0.9223 & 0.9686 & 0.9820 & 0.9979 & 0.9980 & 1.0000 & 0.9999 & 0.9960 \\ 0.5491 & 0.7209 & 0.7267 & 0.8097 & 0.9210 & 0.9670 & 0.9805 & 0.9976 & 0.9972 & 0.9999 & 1.0000 & 0.9969 \\ 0.5464 & 0.7217 & 0.7259 & 0.8074 & 0.9086 & 0.9523 & 0.9669 & 0.9925 & 0.9885 & 0.9960 & 0.9969 & 1.0000 \end{pmatrix},$$

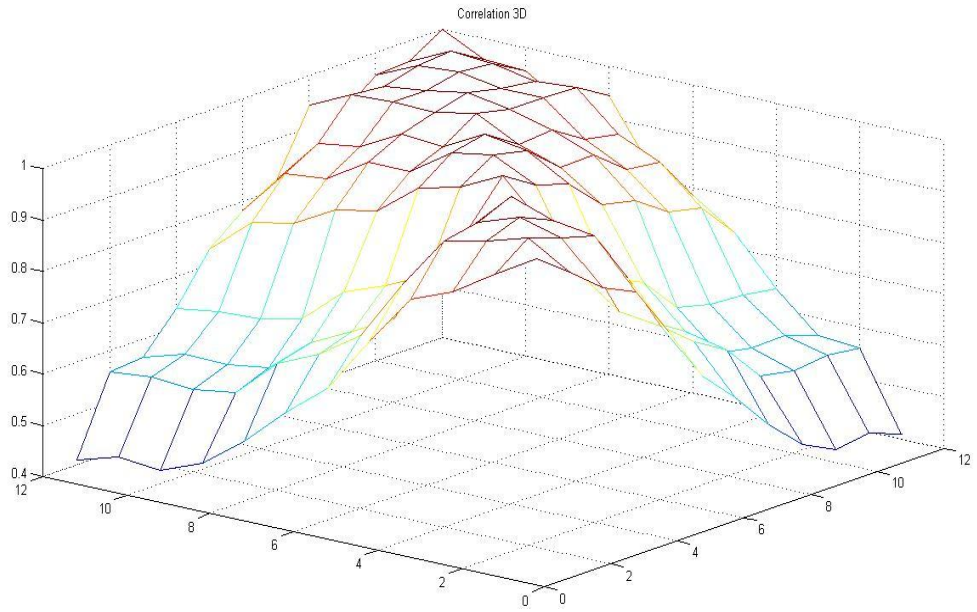

Figure 5.1: Correlation matrix from table 5.1

From the correlation coefficient matrix its 3D graph, one can observed that the correlation between the twelfth variable and the other variables is in an ascending order. It is quite logically because those variables also follow an ascending order. For instance, it is clear that the correlation between the classes in year old $[50,\infty[$ and $[45,50[$ is stronger than the correlation between $[50,\infty[$ and $[5,10[$. From the data the covariance matrix and its 3D representation are computed and represented as follow

$$S = 10^5 \begin{pmatrix} 0.0024 & 0.0154 & 0.0244 & 0.0192 & 0.0166 & 0.0163 & 0.0158 & 0.0138 & 0.0090 & 0.0088 & 0.0066 & 0.0240 \\ 0.0154 & 0.1092 & 0.1788 & 0.1431 & 0.1247 & 0.1234 & 0.1228 & 0.1160 & 0.0755 & 0.0781 & 0.0586 & 0.2212 \\ 0.0244 & 0.1788 & 0.3070 & 0.2432 & 0.2154 & 0.2099 & 0.2123 & 0.1999 & 0.1324 & 0.1368 & 0.1020 & 0.3752 \\ 0.0192 & 0.1431 & 0.2432 & 0.2041 & 0.1831 & 0.1851 & 0.1897 & 0.1821 & 0.1209 & 0.1245 & 0.0917 & 0.3451 \\ 0.0166 & 0.1247 & 0.2154 & 0.1831 & 0.1949 & 0.2141 & 0.2253 & 0.2101 & 0.1515 & 0.1518 & 0.1104 & 0.3859 \\ 0.0163 & 0.1234 & 0.2099 & 0.1851 & 0.2141 & 0.2560 & 0.2711 & 0.2558 & 0.1880 & 0.1877 & 0.1379 & 0.4750 \\ 0.0158 & 0.1228 & 0.2123 & 0.1897 & 0.2253 & 0.2711 & 0.2953 & 0.2826 & 0.2086 & 0.2111 & 0.1545 & 0.5338 \\ 0.0138 & 0.1160 & 0.1999 & 0.1821 & 0.2101 & 0.2558 & 0.2826 & 0.2952 & 0.2132 & 0.2256 & 0.1666 & 0.6133 \\ 0.0090 & 0.0755 & 0.1324 & 0.1209 & 0.1515 & 0.1880 & 0.2086 & 0.2132 & 0.1590 & 0.1654 & 0.1216 & 0.4320 \\ 0.0088 & 0.0781 & 0.1368 & 0.1245 & 0.1518 & 0.1877 & 0.2111 & 0.2256 & 0.1654 & 0.1776 & 0.1308 & 0.4827 \\ 0.0066 & 0.0586 & 0.1020 & 0.0917 & 0.1104 & 0.1379 & 0.1545 & 0.1666 & 0.1216 & 0.1308 & 0.0976 & 0.3592 \\ 0.0240 & 0.2212 & 0.3752 & 0.3451 & 0.3859 & 0.4750 & 0.5338 & 0.6133 & 0.4320 & 0.4827 & 0.3592 & 1.4323 \end{pmatrix},$$

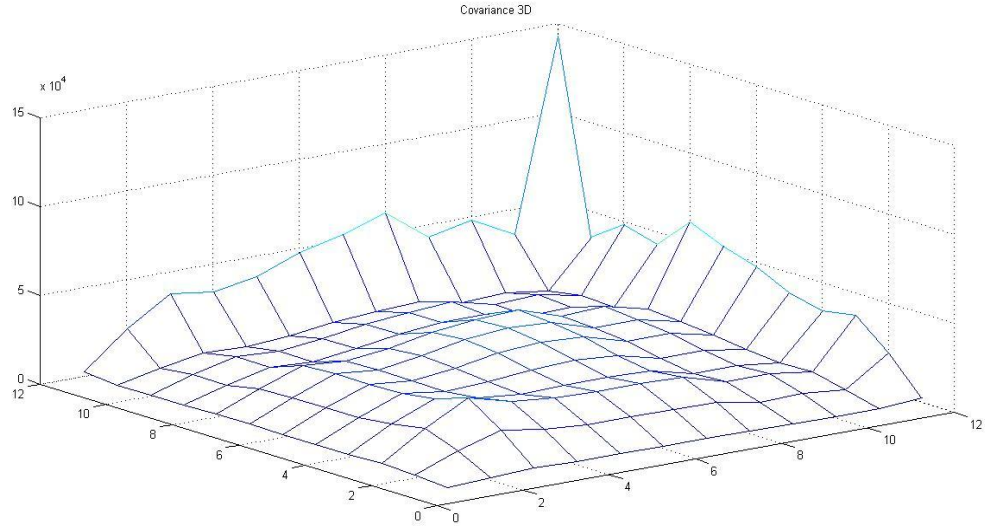

Figure 5.2: covariance matrix from table 5.1

Figure 5.2 is the 3D plot of the covariance matrix **S**. It can be observed that there is very high variation between the twelfth variable and others variables. That high variation is mainly due to the very wide age interval covering all ages above 49. Similarly the variation between the first variable and the others is so low. The low variation is because the age interval is very small compared with the rest of the data, covering only the fewer than one year olds. The eigenvalues computed from the covariance matrix are represented in form of a matrix as follow

$$\mathbf{\Lambda}_S = 10^5 \begin{pmatrix} 0.0001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0003 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0003 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0005 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0019 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0021 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0038 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0064 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0095 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1257 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.3902 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.9900 \end{pmatrix}$$

Its corresponding eigenvectors matrix is

$$\mathbf{E}_S = \begin{pmatrix} -0.9345 & -0.1327 & -0.1986 & 0.0224 & -0.1486 & -0.0972 & 0.0934 & -0.1546 & 0.0346 & -0.0249 & 0.0536 & 0.0168 \\ 0.2416 & 0.0714 & 0.0499 & 0.0472 & -0.3904 & -0.4026 & 0.3666 & -0.5382 & 0.1086 & -0.2334 & 0.3286 & 0.1382 \\ -0.0121 & -0.1849 & -0.0049 & 0.1076 & 0.1999 & 0.3970 & -0.1792 & -0.1246 & -0.4658 & -0.3521 & 0.5585 & 0.2364 \\ -0.0304 & 0.0230 & -0.0723 & -0.3660 & 0.1082 & -0.1340 & 0.2880 & 0.6267 & 0.3162 & -0.2170 & 0.4040 & 0.2090 \\ -0.0580 & 0.3655 & 0.0200 & 0.2603 & -0.2029 & -0.3530 & -0.6545 & 0.1760 & 0.0518 & 0.1835 & 0.2846 & 0.2290 \\ -0.0336 & -0.1252 & 0.1773 & -0.1590 & 0.5557 & -0.0088 & -0.0765 & -0.3992 & 0.4118 & 0.4219 & 0.1789 & 0.2700 \\ 0.0618 & 0.0238 & -0.1372 & -0.1461 & -0.5486 & 0.5568 & 0.1121 & 0.0178 & 0.0752 & 0.4752 & 0.1141 & 0.2952 \\ -0.0500 & 0.0029 & 0.2251 & 0.7257 & 0.1408 & -0.0090 & 0.4574 & 0.2353 & -0.0627 & 0.1804 & -0.0526 & 0.3106 \\ 0.1928 & -0.4311 & -0.6115 & -0.0387 & 0.0755 & -0.3935 & 0.0157 & 0.0536 & -0.3282 & 0.2791 & -0.0581 & 0.2212 \\ -0.1101 & -0.1766 & 0.6524 & -0.3960 & -0.1564 & -0.2367 & 0.0064 & 0.0876 & -0.4447 & 0.1338 & -0.1198 & 0.2367 \\ -0.0831 & 0.7580 & -0.2146 & -0.2283 & 0.2674 & -0.0020 & 0.2294 & -0.1301 & -0.3670 & 0.0738 & -0.0921 & 0.1751 \\ 0.0138 & -0.0336 & -0.0768 & -0.0189 & -0.0383 & 0.0855 & -0.1814 & -0.0685 & 0.2192 & -0.4500 & -0.5089 & 0.6607 \end{pmatrix}$$

From the eigenvectors matrix, the PCs coefficients are computed and represented as follow

$$\begin{pmatrix} Y_{12} \\ Y_{11} \\ Y_{10} \\ Y_9 \\ Y_8 \\ Y_7 \\ Y_6 \\ Y_5 \\ Y_4 \\ Y_3 \\ Y_2 \\ Y_1 \end{pmatrix} = \begin{pmatrix} -0.9345 & 0.2416 & -0.0121 & -0.0304 & -0.0580 & -0.0336 & 0.0618 & -0.0500 & 0.1928 & -0.1101 & -0.0831 & 0.0138 \\ -0.1327 & 0.0714 & -0.1849 & 0.0230 & 0.3655 & -0.1252 & 0.0238 & 0.0029 & -0.4311 & -0.1766 & 0.7580 & -0.0336 \\ -0.1986 & 0.0499 & -0.0049 & -0.0723 & 0.0200 & 0.1773 & -0.1372 & 0.2251 & -0.6115 & 0.6524 & -0.2146 & -0.0768 \\ 0.0224 & 0.0472 & 0.1076 & -0.3660 & 0.2603 & -0.1590 & -0.1461 & 0.7257 & -0.0387 & -0.3960 & -0.2283 & -0.0189 \\ -0.1486 & -0.3904 & 0.1999 & 0.1082 & -0.2029 & 0.5557 & -0.5486 & 0.1408 & 0.0755 & -0.1564 & 0.2674 & -0.0383 \\ -0.0972 & -0.4026 & 0.3970 & -0.1340 & -0.3530 & -0.0088 & 0.5568 & -0.0090 & -0.3935 & -0.2367 & -0.0020 & 0.0855 \\ 0.0934 & 0.3666 & -0.1792 & 0.2880 & -0.6545 & -0.0765 & 0.1121 & 0.4574 & 0.0157 & 0.0064 & 0.2294 & -0.1814 \\ -0.1546 & -0.5382 & -0.1246 & 0.6267 & 0.1760 & -0.3992 & 0.0178 & 0.2353 & 0.0536 & 0.0876 & -0.1301 & -0.0685 \\ 0.0346 & 0.1086 & -0.4658 & 0.3162 & 0.0518 & 0.4118 & 0.0752 & -0.0627 & -0.3282 & -0.4447 & -0.3670 & 0.2192 \\ -0.0249 & -0.2334 & -0.3521 & -0.2170 & 0.1835 & 0.4219 & 0.4752 & 0.1804 & 0.2791 & 0.1338 & 0.0738 & -0.4500 \\ 0.0536 & 0.3286 & 0.5585 & 0.4040 & 0.2846 & 0.1789 & 0.1141 & -0.0526 & -0.0581 & -0.1198 & -0.0921 & -0.5089 \\ 0.0168 & 0.1382 & 0.2364 & 0.2090 & 0.2290 & 0.2700 & 0.2952 & 0.3106 & 0.2212 & 0.2367 & 0.1751 & 0.6607 \end{pmatrix} * \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \end{pmatrix}.$$

Percentage of total variation due to the first eigenvalue is

$\frac{\lambda_1}{\sum_{i=1}^{12}\lambda_i} = \frac{10^5 \cdot 2.9900}{10^5 \cdot 3.5308} = 0.8468$. This means the proportion or percentage of total variance accounted for by the first principal component is: 0.8468 or 84.68%.Similarly the percentage of total variation represented by the first two PCs is

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{12}\lambda_i} = \frac{10^5 \cdot (2.9900 + 0.3902)}{10^5 \cdot 3.5308} = 0.9573 \text{ or } 95.73\%.$$

It follows that the first two principals components are enough to represent the initial 12 variable system with little loss of information.

The first PC is given the following equation

$$Y_1 = 0.0168X_1 + 0.1382 X_2 + 0.2364 X_3 + 0.2090 X_4 + 0.2290 X_5 + 0.2700 X_6 + 0.2952 X_7 + 0.3106 X_8 + 0.2212 X_9 + 0.2367 X_{10} + 0.1751X_{11} + 0.6607X_{12}$$

The second PC is given by the following equation

$$Y_2 = 0.0536X_1 + 0.3286X_2 + 0.5585 X_3 + 0.4040 X_4 + 0.2846 X_5 + 0.1789 X_6 + 0.1141 X_7 - 0.0526 X_8 - 0.0581 X_9 - 0.1198 X_{10} - 0.0921 X_{11} - 0.5089 X_{12}$$

The following table represent the correlation between the PCs Y_1 and Y_2 and the variables $X_1, \ldots, X_{12}$.

Table 5.2: Correlation between PCs and the variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
ρ_{Y_1,X_i}	0.593	0.723	0.738	0.8	0.897	0.923	0.94	0.99	0.96	0.97	0.969	0.955
ρ_{Y_2,X_i}	0.684	0.621	0.63	0.559	0.403	0.221	0.131	-0.06	-0.091	-0.178	-0.184	-0.266

Coefficients of a PC which are the elements of the eigenvectors, are an indicator of the contribution of each variable to the PC. Therefore, there is no direct relation between the contribution of a variable to the PC and the correlation between that PC and the variable. However, a low contribution coefficient coupled with a low correlation can be taken as that variable not being very significant in the computation of that PC. One can decide for instance to omit the variable X_8 in the second PC based on $\rho_{Y_2,X_8} = 0.06$ and the contribution of this variable to PC2 is very low compared with other contributing coefficients of the same PC.

5.2 Case Study 2: PCA Method for Face Recognition

Image processing deals with big data. Consider for instance a two-dimensional square image of size N^2, based on the RGB intensity value, 2^3=8 bits. In this case, a typically 256 by 256 bit image can be viewed as a single vector of dimension 65.536 (256 x 256= 65 536). This image can also be viewed in a 65536-dimensional space as a point. A collection of such images constituted a high dimension space data. Therefore there is a need of dimension reduction to process them. On the other hand, faces have a similar configuration. This mean the distribution of face in a high dimension space is not random. The PCA is therefore used to find out vectors that

best account for the image distribution within the entire high dimensional image/vector space. The subspace built from those vectors is called the "face space or eigenspace ". Each image/vector of the "face space" is a linear combination of original faces images, and it is called "eigenface".

In the following example, the theoretical method and a practical application of face recognition will be discussed. The key words in this example are eigenfaces, eigenspace, orthogonal space, Euclidean distance [1,14,15,16,31].

In the following steps the theoretical procedure and the principles of the face recognition are described.

5.2.1 Theoretical Definitions of the Framework

Step 1: A set of *M* images is built and put into a vector $\mathbf{s}$. Each image being transformed into a vector of size *N*. Then $\mathbf{s} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_M\}$.

Step 2: The set $\mathbf{s}$ is now viewed as a vector so the mean vector of $\mathbf{s}$ which is the mean image of the set is computed by $\boldsymbol{\psi} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\gamma}_i$.

Step 3: For further computation, the difference $\boldsymbol{\varphi}$ between the input image from set $\mathbf{s}$ and the mean image $\boldsymbol{\psi}$ is computed as $\boldsymbol{\varphi}_i = \boldsymbol{\gamma}_i - \boldsymbol{\psi}$.

Step 4: The covariance matrix $\mathbf{C}$ of the system is computed as follow

$\mathbf{C} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T = \mathbf{A}\mathbf{A}^T$, with $\mathbf{A} = \{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \cdots, \boldsymbol{\varphi}_M\}$.

Step 5: A set of orthonormal vectors, $\mathbf{u}_n$ is built. This set should be made of vectors which will describe the distribution in the best way. Therefore, the k^{th} vector $\mathbf{u}_k$ is chosen such that $\lambda_k = \frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{u}_k^T \boldsymbol{\varphi}_i\right)^2$ is maximum. Where λ_k and $\mathbf{u}_k$ are respectively eigenvalues and eigenvectors of the covariance matrix.

Step 6: Computation of the Dirac matrix of the set $\mathbf{A}$ as follow $\mathbf{L}_{ij} = \boldsymbol{\varphi}_i^T \boldsymbol{\varphi}_j$.

Step7:The eigenfaces are then built based on the computed eigenvalues and eigenvectors of the covariance matrix.

The previous seven steps are used to build the experiment system; the following steps are the recognition procedure.

Step 1: Each new face is now transformed into its eigenface components. Then, the input image is compared with the mean image and their difference is multiplied by each eigenvector of the matrix Dirac matrix $\mathbf{L}$. Resulting vector is denoted by $\boldsymbol{\omega}$ .That is $\omega_k = \mathbf{u}_k'\left(\boldsymbol{\gamma} - \boldsymbol{\psi}\right)$ and $\boldsymbol{\omega}' = \left[\omega_1, \omega_2, \cdots, \omega_M\right]$.

Step 2: The class of image which describes the best the input image is determined by computing the Euclidean distance $\varepsilon_k = \left\|\boldsymbol{\omega} - \boldsymbol{\omega}_k\right\|^2$.

Step 3: If ε_k is bellow the defined threshold θ_ε then the input face is considered to belong to a class, and the image is considered to be a known face. A second threshold $\theta_{\varepsilon 2}$ can be defined such a way that if the value ε_k is above the threshold θ_ε but below the threshold $\theta_{\varepsilon 2}$ then the image is considered as an unknown face. If the

value ε_k is instead above the second threshold $\theta_{\varepsilon 2}$ then the image is concluded not to be a face.

5.2.2 Application of the Defined Framework.

As application the "faces" here are a set of 20 images. Those images are logos which have in common the round shape. All the thresholds mentioned above are intuitively fixed. Each image from the training set contributes at the same range in the mean image computation. This means the Euclidean distance between every image of the training set and the mean image is the same. This means, assuming that the training set's images and the input image are recorded in the same conditions, the threshold can be chosen to be the common Euclidean distance between image from training set and their mean. In this case for a given input image, if its Euclidean distance equals to the threshold, then the image is a known face i.e., the image is in the database. One can also assume that the image from the training set and the input image are recorded with some little difference in this case, the threshold can be chosen to be a little bit greater than the common Euclidean distance between training set's images and their mean.

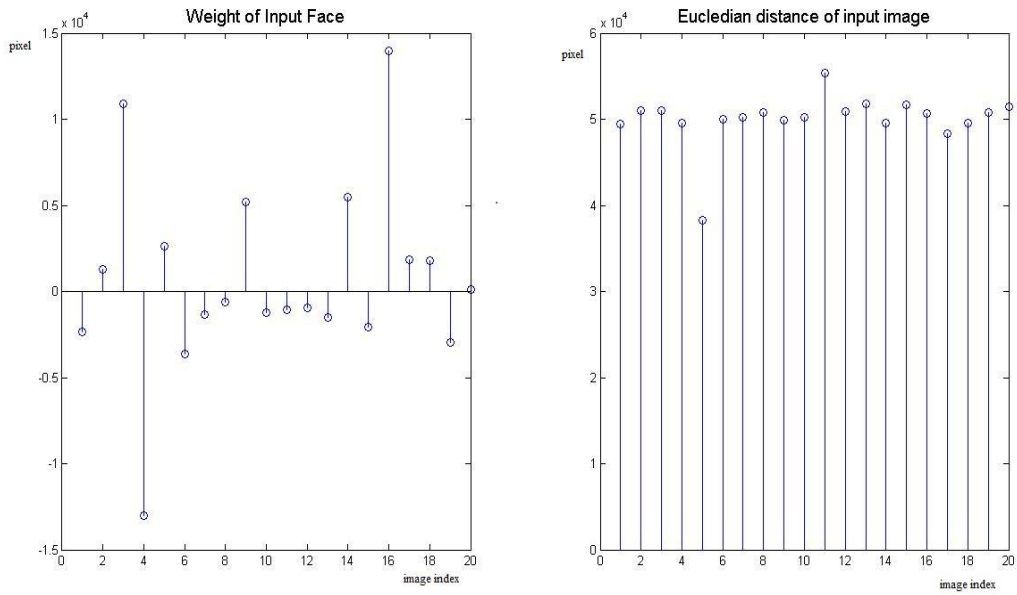

Figure 5.3: Weight and Euclidean distance of a face from the training set

An image from the training set is chose to be an input image. From the figure 5.3, the common Euclidean distance between images from the training set is $\varepsilon_k = 38283$. Furthermore, the minimum distance between images from the training set is $\varepsilon_{\min} = 38283$ and the maximum distance between images from the training set is $\varepsilon_{\max} = 55324$.

The threshold values, $\theta_k = 38283$, $\varepsilon_{\min} = 38283$ and $\varepsilon_{\max} = 55324$ lead to the following conclusion. If an input image has its minimum Euclidian distance equals to $\theta_k = 38283$ then it is said to be a recognized image. If its Euclidean distances are values between $\varepsilon_{\min} = 38283$ and $\varepsilon_{\max} = 55324$ then might be an unknown face. But if its Euclidean distances are rather bigger than $\varepsilon_{\max} = 55324$ then the image is not a face. In this case where an image of the training set is used to setup threshold values, let $I_t = [38283; 55324]$, the interval which lower limit is $\varepsilon_{\min} = 38283$ and upper limit is $\varepsilon_{\max} = 55324$.

Let consider now a logo with round which doesn't belong to the training set. The computation displayed the following results:

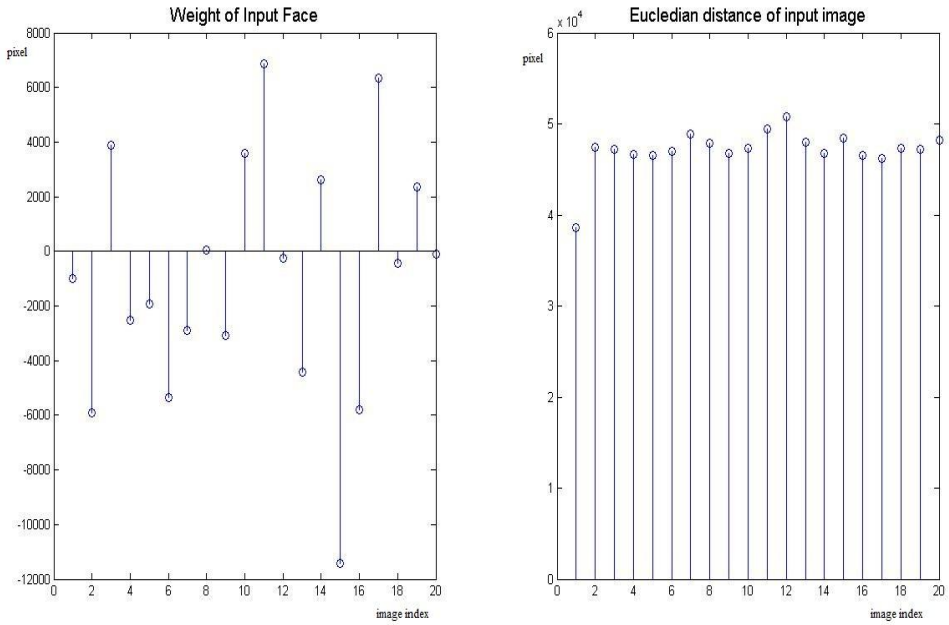

Figure 5.4: weight and Euclidean distance of unknown face.

The maximum Euclidean distance between the input image and the images from the training set is $\varepsilon_{\max} = 50826$ and the minimum is $\varepsilon_{\min} = 38635$. The values $50826; 38635 \in I_t = [38283; 55324]$ this means the input image is an unknown face.

Let consider an image which is not obviously a logo. This means an image without any round shape. The computation displayed the following results:

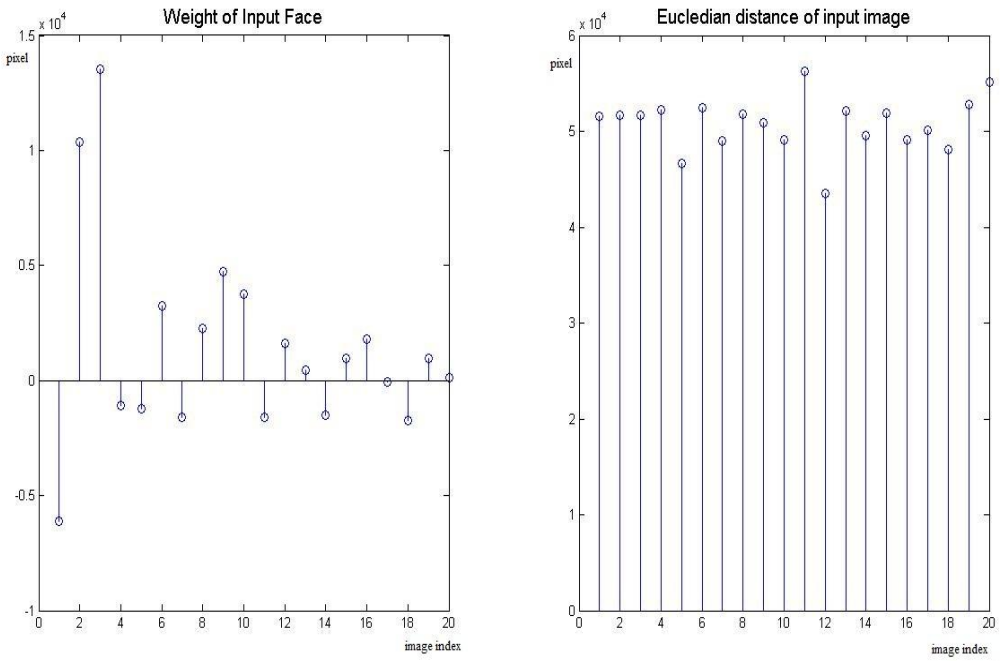

Figure 5.5: weight and Euclidean distance of an image else than a face

The maximum Euclidean distance between the input image and the images from the training set is $\varepsilon_{\max} = 56254$ and the minimum is $\varepsilon_{\min} = 43502$. Here at least one value which is the maximum value of Euclidean distance $56254 \notin I_t = [38283; 55324]$. This leads to the conclusion that the input image was not a face.

Chapter 6

CONCLUSION

Multivariate statistics is of great importance because its application is found in many fields. Whenever one deal with data that comes from a process with multivariate observations, it is wise to reduce the dimension of the dataset to enable the easy processing. Dimension reduction using the PCA has been discussed in this work.

PCA is viewed over two ways. Algebraically, PCs is some special linear combination of variables, built from a multivariate data analysis. Geometrically, the PCA is an orthogonal transformation which changes the initial coordinate system in which data are scatter into a new orthogonal coordinate system. The axes of the new coordinate system are directed by the variation of the scatter data.

PCs can be computed using either covariance matrix or correlation coefficient matrices. For a given data, the PCs computed using covariance matrix and those computed using correlation matrix are not generally the same. In general, the covariance matrix is used for the computation of PCs. However, when the data is inhomogeneous or when the magnitudes of data values for different variables are significantly different, the correlation matrix is preferred.

Consider a dataset of n observations of p variables. Then p PCs can be computed. Only the first few PCs, where the corresponding eigenvalues represent the high

percentage of total variation in the data are enough for further processing and deriving conclusions on the whole dataset.

Coefficients in front of each constituent variable of a PC, indicates the level of contribution of each variable to the PC. The correlation between a PC and its constituent variable show the degree of linear relationship between a variable and the PC.

REFERENCES

[1] Banerjee, A. (2012). *Impact of Principal Component Analysis in the Application of Image Processing.* International Journal of Advanced Research in Computer Science and Software Engineering .

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

[3] Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer.

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

[5] Kruger, U., Zhang, J., & Xie, L. (s.d.). *Developments and Applications of Nonlinear Principal Component Analysis - a Review*.

[6] Catignon, H. (2010). *Multivariate Normal Distribution. Dans Statistical Analysis of Management Data.* Springer.

[7] Duby, C., & Robin, S. (2006). *Analyses en Composantes Principales*. Grignon: Institut National Agronomique Paris.

[8] Gonzalez, P.-L. (s.d.). *Analyses en Composantes Principales.*

[9] Morineau, A. (s.d.). *ACP-Analyses en Composantes Principales*.

[10] Margairaz, F., & Cuneo, N. *Algèbre Lineaire I & II* . Lausanne: Ecole Polytechnique Fédérale de Lausanne.

[11] MALBOS, P. *Analyses Matricielle et Algèbre Linéaire Appliquée* . Lyon: Université Claude Bernard lyon 1.

[12] Neil H, T. (2002). *Applied Multivariate Analysis*. New York: Springer.

[13] Vaillancourt, R. (1995). *Calcul Matriciel*. Ottawa.

[14] Turk, M., & Pentland, A. (1991). *Eigenfaces for Face Detection/Recognition.* Journal of Cognitive Neuroscience , 71-86.

[15] Lalann, C., Georges, S., & Pallier, C. (s.d.). *Statistiques Appliquées à l'expérimentation en Sciences Humaines.*

[16] Kim, K. (s.d.). *Face Recognition using Principle Component Analysis*. Maryland, USA: University of Maryland.

[17] Le-Rademacher, J. G. (2008). *Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data.*

[18] Dimitrov, D. (2008). *Geometric Applications of Principal Component Analysis*. Berlin.

[19] RENCHER, A. C. (2002). *Methods of Multivariate Analysis. John Wiley & Sons.*

[20] Marden, J. I. (2013). *Multivariate Statistics.* Illinois: University of Illinois at Urbana-Champaign.

[21] Tilière, B. d. (2009). *Analyses statistiques multivariées.*

[22] Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2002, 2007, 2012). *Probability & Statistics for Engineers & Scientists.* New Jersey: Prentice Hall.

[23] Richardson, M. (2009). *Principal Component Analysis.*

[24] Baker, K. (2005, 2013). *Singular Value Decomposition Tutorial.*

[25] Gournay, A. (2012, Septembre). *Analyse statistique mutlivariée*. Neuchâtel, Suisse: Université de Neuchâtel.

[26] Chen, H. (s.d.). *Principal Component AnalysisWith Missing Data and Outliers*. Rutgers University.

[27] Härdle, W. K., & Simar, L. (2003, 2007, 2012). *Applied Multivariate Statistical Analysis*. Berlin Heidelberg : Springer.

[28] Zhang, W. (2012). *Regression based principal component analysis for sparse functional data with applications to screening pubertal growth paths*. Columbia : Columbia University.

[29] Anderson, T. W. (1963). *Asymptotic Theory for Principal Component Analysis.*

[30] Balbi, S., & Misuraca, M. (2010). *A Doubly Projected Analysis for Lexical Tables.*

[31] Charlton, M., Brunsdon, C., Demšar, U., Harris, P., & Stewart. (2010). *Principal Components Analysis*: from global to local. Guimarães, Portugal.

[32] Cameroun, R. d. (1992). *Deuxième recensement général de la population et de l'habitat au Cameroun.*

[33] University, D. (s.d.). Eigenface_tutorial. Consulté le juin 1, 2015, sur http://www.pages.drexel.edu/~sis26/Eigenface%20Tutorial.htm

APPENDIX

Appendix A: Matlab Code of Face Recognition [33]

```
% Face recognition by Santiago Serrano
clearall
closeall
clc
% number of images on your training set.
M=20;

%Chosen std and mean.
%It can be any number that it is close to the std and mean of most
of the images.
um=100;
ustd=80;

%read and show images(bmp);
S=[];   %img matrix
figure(1);
for i=1:M
str=strcat(int2str(i),'.bmp');   %concatenates two strings that form
the name of the image
eval('img=imread(str);');
subplot(ceil(sqrt(M)),ceil(sqrt(M)),i)
imshow(img)
if i==3
title('Training set','fontsize',18)
end
drawnow;
    [irow icol]=size(img);    % get the number of rows (N1) and
columns (N2)
temp=reshape(img',irow*icol,1);     %creates a (N1*N2)x1 matrix
    S=[S temp];         %X is a N1*N2xM matrix after finishing the
sequence
%this is our S
end


%Here we change the mean and std of all images. We normalize all
images.
%This is done to reduce the error due to lighting conditions.
for i=1:size(S,2)
temp=double(S(:,i));
    m=mean(temp);
st=std(temp);
S(:,i)=(temp-m)*ustd/st+um;
end

%show normalized images
figure(2);
for i=1:M
str=strcat(int2str(i),'.jpg');
img=reshape(S(:,i),icol,irow);
img=img';
eval('imwrite(img,str)');
subplot(ceil(sqrt(M)),ceil(sqrt(M)),i)
imshow(img)
drawnow;
if i==3
title('Normalized Training Set','fontsize',18)
```

```
end
end


%mean image;
m=mean(S,2);   %obtains the mean of each row instead of each column
tmimg=uint8(m);   %converts to unsigned 8-bit integer. Values range
from 0 to 255
img=reshape(tmimg,icol,irow);    %takes the N1*N2x1 vector and
creates a N2xN1 matrix
img=img';       %creates a N1xN2 matrix by transposing the image.
figure(3);
imshow(img);
title('Mean Image','fontsize',18)

% Change image for manipulation
dbx=[];   % A matrix
for i=1:M
temp=double(S(:,i));
dbx=[dbx temp];
end

%Covariance matrix C=A'A, L=AA'
A=dbx';
L=A*A';
% vv are the eigenvector for L
% dd are the eigenvalue for both L=dbx'*dbx and C=dbx*dbx';
[vv dd]=eig(L);
% Sort and eliminate those whose eigenvalue is zero
v=[];
d=[];
for i=1:size(vv,2)
if(dd(i,i)>1e-4)
        v=[v vv(:,i)];
        d=[d dd(i,i)];
end
end

%sort,  will return an ascending sequence
 [B index]=sort(d);
ind=zeros(size(index));
dtemp=zeros(size(index));
vtemp=zeros(size(v));
len=length(index);
for i=1:len
dtemp(i)=B(len+1-i);
ind(i)=len+1-index(i);
    vtemp(:,ind(i))=v(:,i);
end
 d=dtemp;
 v=vtemp;


%Normalization of eigenvectors
for i=1:size(v,2)       %access each column
kk=v(:,i);
temp=sqrt(sum(kk.^2));
v(:,i)=v(:,i)./temp;
end
```

```
%Eigenvectors of C matrix
u=[];
for i=1:size(v,2)
temp=sqrt(d(i));
    u=[u (dbx*v(:,i))./temp];
end

%Normalization of eigenvectors
for i=1:size(u,2)
kk=u(:,i);
temp=sqrt(sum(kk.^2));
u(:,i)=u(:,i)./temp;
end


% show eigenfaces;
figure(4);
for i=1:size(u,2)
img=reshape(u(:,i),icol,irow);
img=img';
img=histeq(img,255);
subplot(ceil(sqrt(M)),ceil(sqrt(M)),i)
imshow(img)
drawnow;
if i==3
title('Eigenfaces','fontsize',18)
end
end


% Find the weight of each face in the training set.
omega = [];
for h=1:size(dbx,2)
    WW=[];
for i=1:size(u,2)
        t = u(:,i)';
        WeightOfImage = dot(t,dbx(:,h)');
        WW = [WW; WeightOfImage];
end
omega = [omega WW];
end


% Acquire new image
% Note: the input image must have a bmp or jpg extension.
%       It should have the same size as the ones in your training
set.
%       It should be placed on your desktop
InputImage = input('Please enter the name of the image and its
extension \n','s');
InputImage = imread(strcat('C:\Users\YAMENI\Desktop\Image
Test\',InputImage));
figure(5)
subplot(1,2,1)
imshow(InputImage); colormap('gray');title('Input
image','fontsize',18)
InImage=reshape(double(InputImage)',irow*icol,1);
temp=InImage;
me=mean(temp);
st=std(temp);
```

```
temp=(temp-me)*ustd/st+um;
NormImage = temp;
Difference = temp-m;

p = [];
aa=size(u,2);
for i = 1:aa
pare = dot(NormImage,u(:,i));
    p = [p; pare];
end
ReshapedImage = m + u(:,1:aa)*p;    %m is the mean image, u is the
eigenvector
ReshapedImage = reshape(ReshapedImage,icol,irow);
ReshapedImage = ReshapedImage';
%show the reconstructed image.
subplot(1,2,2)
imagesc(ReshapedImage); colormap('gray');
title('Reconstructed image','fontsize',18)

InImWeight = [];
for i=1:size(u,2)
    t = u(:,i)';
    WeightOfInputImage = dot(t,Difference');
    InImWeight = [InImWeight; WeightOfInputImage];
end

ll = 1:M;
figure(68)
subplot(1,2,1)
stem(ll,InImWeight)
title('Weight of Input Face','fontsize',14)

% Find Euclidean distance
e=[];
for i=1:size(omega,2)
    q = omega(:,i);
    DiffWeight = InImWeight-q;
mag = norm(DiffWeight);
    e = [e mag];
end

kk = 1:size(e,2);
subplot(1,2,2)
stem(kk,e)
title('Eucledian distance of input image','fontsize',14)

MaximumValue=max(e)
MinimumValue=min(e)
```