Effect of Centereing Data in Principal Component Analysis

Bilal Sami Mohammad Ghadaireh

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
July 2014
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

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

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

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

Asst. Prof. Dr. Yücel Tandoğdu
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev
2. Prof. Dr. Agamirza Başirov
3. Asst. Prof. Dr. Yücel Tandoğdu

ABSTRACT

In the analysis of multivariate data, the processing and extracting meaningful results becomes very difficult due large number of variables and data. Therefore, statistical techniques to deal with such data, by finding linear combinations of existing variables, such that each variable is assigned a coefficient or score that determines its contribution to that linear combination. These linear combinations are called Principal Components (PC) and the methodology used in the determination of the PCs is called Principal Component Analysis (PCA). In general the number of PCs is expected to be the same as the number of variables. However, the PCs are determined such that the great percentage of variation (usually over 90%) in the data accumulates in the first few PCs. Then, the remaining PCs become redundant, and the information contained in a large number of variables is reduced into a few new variables (PCs) that are linear combinations of original variables. Therefore, a technique used in determining the PCs is very important. In this work, theory of PCA with related mathematical background is explained and using a certain data set, various ways of the application of PCA technique is investigated, obtained results are interpreted.

Keywords: Principle component analysis, data, eigenvalue, eigenvector, covariance, correlation, standardized data, centered data.

ÖZ

Çok değişkenli veri analizinde özellikle değişken sayısının çok fazla olduğu durumlarda işlem yapıp sonuç çıkarma oldukca zordur. Bu şartlar altında veri analizini yapabilmek için geliştirilmiş istatistik teknikler, mevcut değişkenlerin lineer kombinezonlarından oluşan ve biribirinde bağımsız yeni değişkenlerin hesaplanmsını mümkün kılar. Bu değişkenlere Temel Bileşenler ve bu bileşenlerin hesaplanmasınada kullanılan yöntemlerede Temel Bileşenler Analizi denir. Hesaplanan temel bileşen sayısı, değişken sayısı kadardır. Ancak, verideki toplam değişimin çok büyük bir kısmı ilk birkaç temel bileşen tarafından temsil edilir. Sadece bunların analiz ve yorumlamada kullanılması, hesaplamalardaki yoğunluğu ciddi miktarda azaltırken, elde edilen sonuçlar tüm kitleyi 90%'ın üstünde bir temsiliyeti sahiptir. Geriye kalan ve verideki toplam değişimin çok az bir kısmını temsil eden temel bileşenler işleme sokulmaz. Böylece, çok yüksek sayıdaki veri miktari çok aza indirgenmiş olur. Bu nedenle temel bileşenlerin hesabında kullanımlan yöntemler çok önemlidir. Bu çalışmada temel bileşenler analizinin matematiksel temelleri izah edilmiş, belli bir veri seti kullanılarak metodun farklı yaklaşımlarla uygulaması yapılıp, elde edilen sonuçlar yorumlanmıştır.

Anahtar kelimeler: Temel bileşenler analizi, veri, özdeğer, özvektör, kovaryans, korelasyon, standartlaştırılmış veri, merkezileştirilmiş veri.

DEDICATION

I am dedicating this thesis to my family

ACKNOWLEDGMENT

Express my sincere thanks and appreciation to all, who contributed to the completion of this modest effort, led by Asst. Prof. Dr. Yucel TANDOĞDU the supervisor of this thesis who didn't spare days in the counseling, guiding and encouraging me in my studies. I also greatly appreciate the contributions of all members of Mathematics department, and the chairman of department Prof. Dr. Nazim Mahmudov, during my studies for my Master degree.

My special thanks goes to those friends whose help and support made my life and mission easier.

I want to thank all the professors of the Department of Mathematics from whom I learned a lot and enriched my knowledge on mathematics.

I would like to thank my family for their precious support in every aspect during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF SYMBOLS /ABBREVIATIONS

$\mathbf{A}$	Used to represent a matrix
$\mathbf{I}$	Identity matrix
X	Denotes a random variable
$r.v$	Abreviation for random variable
$\mathbf{x}$	Denotes a vector
λ	Denotes an eigenvalue
μ	Population mean
$\bar{x}$	Sample mean
σ^2	Population variance
σ	Population standard deviation
s^2	Sample variance
s	Sample standard deviation
$\sum$	Population covariance matrix
$\mathbf{S}$	Sample covariance matrix

Chapter 1

INTRODUCTION

Processing a data set with large number of variables necessitates special techniques. Principal Component Analysis (PCA) is one of such methodologies, widely used for this purpose. PCA method is based on finding linear combinations of the variables, such that they represent the directions of variation in the multivariate data in ascending order. Number of PCs is the same as the number of variables. However, only few of the PCs are generally enough to represent the process in question, since the large percentage (over 90%) of variation in the process tends to be explained by these few PCs.

Early work by Karl Pearson [1] laid down the foundations on PCA. Interest in PCA and its applicatios in data analysis started increasing in 1970s, leading to the developments vitnesed today. Many valuable contributions made by different researchers. A brief review of this is given in Chaper 2 under literature survey.

Chapter 3 explains the necessary mathematical and statistical background necessary to understand and develop the PCA methodology.

In PCA a multivariate data set is considered as an $n \times p$ dimensional marix, p being the number of variables $X_i;\ i = 1, \ldots, p$ and n number of observations. The data set is manipulated, such that a new set of independent variables $Y_i;\ i = 1, \ldots, p$ consisting of linear combinations of the initial variables, named as Principal Components (PC).

The PCs are determined so as the first is in the direction where the largest variation occurs in the raw data, followed by the remaining PCs representing the direction variations in descendin order. Hence, the first few PCs tend to represent the great majority of variation in the data set. This provides the facility of understanding the process that generated the data by only analizing these few PCs. Theory involved in the computation of PCs, ways of their application such as using the raw data, centered, or standardized data are explained under Chapter 4.

A data set consisting of 5 variables is used as a case study to apply the theory to a real life example concerning the factors that affect the life of a battery.

Chapter 2

LITERATURE REVIEW

Initial work on PCA dates back to 1901 in the work of Karl Pearson [1], who discused the the best way of graphically representing data. Hotelling (1933) [2], [3] and and Girshick (1936; 1939) [4] are the researchers that contributed to the theory of PCA in 1930s.

Among other early researchers worth metioning are the work of Anderson (1963) [5] who elaborates on the asymptotic properties of PCs. Rao (1964) [6] talks about the use and interpretation of PCs, and als published a book on "Generalized Inverse of Matrices and its Applications" in 1971 [9]. Jeffers (1967) [8] offers case studies on the application of PCs.

Mardia et.al. (1979) [23] published what may be considered as one of the first books that combines multivariate analysis together with associated theory and applications to statistics.

In the post 1980s period with the advances of computing power, interest has rapidly grown in PCA, resulting in many theoretical work as well as successful applications in many different fields of endveour. Some of related work that benefited from is listed in the references.

Chapter 3

GENERAL REVIEW OF SOME MATHEMTICS AND STATISTICS RELATED CONCEPTS

In this chapter important mathematical and statistical principals which are necessary to understand the concepts and methods used in principal component analysis are summarized, Centering Matrix and its function in PCA are introduced.

3.1 Matrix Algebra Concepts

The use of matrix algebra in many statistical applications is essential. Certain basic concepts from matrix algebra are introduced in order to enable the comprehension of stastistics to be used in later chapters.

3.1.1 Inverse of a Matrix

Let a square matrix $\mathbf{A}$ of size $n \times n$, which is non-singular $|\mathbf{A}| \neq 0$ be given. Then there exists a matrix $\mathbf{A}^{-1}$ which is called the inverse of $\mathbf{A}$ such that:

$$\mathbf{A}\mathbf{A}^{-1}=\mathbf{A}^{-1}\mathbf{A}=\mathbf{I} \qquad (3.1)$$

where $\mathbf{I}$ is the identity matrix.

The inverse of a sequare $n \times n$ matrix $\mathbf{A}$ can be found by using the following equation

$$\mathbf{A}^{-1} = \frac{\text{adj(A)}}{\det(\text{A})} \qquad (3.2)$$

where the $adj(\mathbf{A})$ denotes the adjoint of matrix $\mathbf{A}$. The adjoint can be calculated by;

Let $\mathbf{B} = b_{ij}$ to be the matrix whose coefficients are found by taking the determinant of the $(n-1)\times(\mathrm{n}-1)$ sub-matrix (minor) obtained deleting the i^{th} row and j^{th} column of **A**, and multiplying b_{ij} by $(-1)^{i+j}$. The obtained matrix **B** is known as the cofactor matrix of **A** [19]. Transpose of gives the adjoint matrix of **A**. Then from equation 3.2 the inverse is obtained.

3.1.2 General Inverse

A general $n \times n$ matrix can be inverted using methods such as the Gauss-Jordan elimination, Gaussian elimination, or LU decomposition. The inverse of a product **AB** of matrices **A** and **B** can be expressed in terms of $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$.

Consider the following properties on matrices. Let

$$\mathbf{C} = \mathbf{AB}$$

then

$$\mathbf{B}=\mathbf{A}^{-1}\mathbf{AB}=\mathbf{A}^{-1}\mathbf{C}$$

and

$$\mathbf{A}=\mathbf{ABB}^{-1}=\mathbf{CB}^{-1}.$$

Therefore,

$$\mathbf{C}=\mathbf{AB}=(\mathbf{CB}^{-1})(\mathbf{A}^{-1}\mathbf{C})=\mathbf{CB}^{-1}\mathbf{A}^{-1}\mathbf{C},$$

so

$$\mathbf{CB}^{-1}\mathbf{A}^{-1}=\mathbf{I},$$

Where, **I** is the identity matrix, and

$$\mathbf{B}^{-1}\mathbf{A}^{-1}=\mathbf{C}^{-1}=(\mathbf{AB})^{-1}.$$

Definition If **A** is an $m \times n$ matrix, then **G** is a generalized inverse of **A** if **G** is an $m \times n$ matrix with

$$\mathbf{AGA} = \mathbf{A} \qquad (3.3)$$

If $\mathbf{A}$ has an inverse in the usual sense, that is if $\mathbf{A}$ is $n \times n$ and has a two-sided inverse $\mathbf{A}^{-1}$, then

$$\mathbf{A}^{-1}(\mathbf{AGA})\mathbf{A}^{-1} = (\mathbf{A}^{-1}\mathbf{A})\mathbf{G}(\mathbf{A}\mathbf{A}^{-1}) = \mathbf{G}$$

while by (3.3)

$$\mathbf{A}^{-1}(\mathbf{A})\mathbf{A}^{-1} = (\mathbf{A}^{-1}\mathbf{A})\mathbf{A}^{-1} = \mathbf{A}^{-1}$$

Thus, if $\mathbf{A}^{-1}$ exists in the usual sense, then $\mathbf{G} = \mathbf{A}^{-1}$. This justifies the term generalized inverse. Any $m \times n$ matrix $\mathbf{A}$ has at least one generalized inverse $\mathbf{G}$ [9].

3.1.3 Eigensolutions (Eigenvalue and Eigenvector)

Matrix $\mathbf{A}$ is a square matrix having size $n \times n$. Also the non-zero vector $\mathbf{x}$ and scalar λ are given. Then, in

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{3.4}$$

The vector $\mathbf{x}$ is called eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ [12]. Excpression of the determinant ($|\mathbf{A}|$) and trace ($\mathrm{tr}(\mathbf{A})$) of matrix $\mathbf{A}$ is given as below.

$$|\mathbf{A}| = \prod_{j=1}^{p} \lambda_j \tag{3.5}$$

$$tr(\mathbf{A}) = \sum_{j=1}^{p} \lambda_j$$

3.1.4 Orthogonal matrix

Matrix $\mathbf{A}$ of size $n \times n$ is orthogonal if

$$\mathbf{A}\mathbf{A}^T = \mathbf{I},$$

where $\mathbf{A}^T$ is the transpose of $\mathbf{A}$ and $\mathbf{I}$ is the identity matrix. In particular, an orthogonal matrix is always invertible, and in component form, $\mathbf{A}^{-1} = \mathbf{A}^T$

$$a_{ij}^{-1} = a_{ij}^T$$

where a^{-1}_{ij} and a^{T}_{ij} are the i, j elements of matrix $\mathbf{A}^{-1}$ and $\mathbf{A}^{T}$ respectively. These relation make orthogonal matrices particularly easy to compute with, since the transpose operation is much simpler than computing an inverse [10].

3.1.5 Orthonormal Matrix

The conditions of achieving orthonormality for two vectors in an inner product space are orthogonal and unit vectors. A set of vectors form an orthonormal set, if all vectors in the set are mutually orthogonal and all are unit length. An orthonormal set which forms a basis is called an orthonormal basis.

Definition Let $\mathbf{V}$ be an inner-product space. A set of n-vectors $\{\mathbf{u}_1,...,\mathbf{u}_n\} \in \mathbf{V}$ is called orthonormal if and only if

$$\forall i, j :< \mathbf{u}_i, \mathbf{u}_j >= \delta_{ij}, \begin{cases} 1 & if < u_i, u_j >= 1 \\ 0 & otherwise \end{cases}$$

where δ_{ij} is the Kronecker delta and $< .,. >$ is the inner product defined over $\mathbf{V}$. Let $\mathbf{A}$ be $n \times n$ matrix as follows: $\mathbf{A} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n)$, $\mathbf{v}_i, i = 1, \ldots, n$ is row vector. This matrix is called orthnormal if it is orthogonal and $\|v_i\|_2 = 1$.

3.2 Decomposition of a Matrix

A matrix $\mathbf{A}$ can be decompose or factored by writing the matrix as the product of two matrices. There are different methods used in matrix decomposition, such as LU decomposition, spectral decomposition (SP), singular value decomposition SVD. Each method finds use among a particular class of problems. In principle component analysis SP and SVD are widely used. Hence, some detail on these methods is given in sections 3.2.1 and 3.2.2.

3.2.1 Spectral Decompositions

This is a method that eatblishes the relationship between a square matrix ant its eigenvalues and eigenvectors. It is also called Jordan decomposition. Theorem 3.1 gives basic idea of the spectral decomposition [21].

Theorem 3.1 Jordan Decomposition. Let $\mathbf{A}(p\times p)$ be a symmetric matrix. Then

$$\mathbf{A}=\mathbf{\Gamma\Lambda\Gamma}^{T}=\sum_{j=1}^{p}\lambda_{j}\boldsymbol{\gamma}_{j}\boldsymbol{\gamma}_{j}^{T}$$

where

$$\mathbf{\Lambda}=diag(\lambda_{1},...,\lambda_{p})$$

$$\mathbf{\Gamma}=(\boldsymbol{\gamma}_{1},...,\boldsymbol{\gamma}_{p})$$

and $\lambda_{1},...,\lambda_{p}$ are the eigenvalues of $\mathbf{A}$. $\mathbf{\Gamma}$ Is an orthogonal matrix consisting of the eigenvectors $\boldsymbol{\gamma}_{1},...,\boldsymbol{\gamma}_{p}$ of $\mathbf{A}$.

Using spectral decomposition powers of a matrix $\mathbf{A}(p\times p)$ can be defined. Suppose $\mathbf{A}$ is a symmetric matrix. Then by Theorem 3.1 and for some $\alpha\in\Re$

$$\mathbf{A}^{\alpha}=\mathbf{\Gamma\Lambda}^{\alpha}\mathbf{\Gamma}^{T} \qquad (3.6)$$

where $\mathbf{\Lambda}^{\alpha}=diag(\lambda_{1}^{\alpha},...,\lambda_{p}^{\alpha})$. From equation 3.6 the inverse of the matrix $\mathbf{A}$ can be obtained by setting $\alpha=-1$,

$$\mathbf{A}^{-1}=\mathbf{\Gamma\Lambda}^{-1}\mathbf{\Gamma}^{T}.$$

3.2.2 Singular Value Decomposition (SVD)

The singular value decomposition (SVD) is a factorization of a real or a complex matrix, with many useful applications in signal processing and statistics. Formally, the singular value decomposition of an $n\times p$ real or complex matrix $\mathbf{A}$ is a factorization of the form:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

where $\mathbf{U}$ is an $n \times p$ and column orthogonal (its columns are eigenvectors of $\mathbf{A}\mathbf{A}^T$), $\mathbf{V}$ is an $p \times p$ and orthogonal (its columns are eigenvectors of $\mathbf{A}^T\mathbf{A}$), and $\mathbf{\Lambda}$ is an $p \times p$ diagonal matrix of the form

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ \vdots & \lambda_P & \vdots \\ \vdots & 0 & \vdots \\ & & \ddots \\ 0 & \cdots & 0 \end{pmatrix}$$

with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ and r = rank($\mathbf{A}$) . $\lambda_1,...,\lambda_p$ are called the singular values of $\mathbf{A}$ [22] .

Example 3.1

$$\mathbf{A} = \begin{bmatrix} 2 & 6 & 8 \\ 3 & 1 & 5 \end{bmatrix}$$

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 104 & 52 \\ 52 & 35 \end{bmatrix} \text{ and } \mathbf{A}^T\mathbf{A} = \begin{bmatrix} 13 & 15 & 31 \\ 15 & 37 & 53 \\ 31 & 53 & 89 \end{bmatrix}$$

The eigenvalues and eigenvectors of $\mathbf{A}\mathbf{A}^T$ are:

$$\boldsymbol{\lambda}_{AA^T} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 7.096 \\ 131.9 \end{bmatrix}, \quad \mathbf{G}_{AA^T} = \begin{bmatrix} 0.4728 & -0.8811 \\ -0.8811 & -0.4728 \end{bmatrix}$$

The eigenvalues and eigenvectors of $\mathbf{A}^T\mathbf{A}$ are:

$$\boldsymbol{\lambda}_{AA^T} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 7.096 \\ 131.9 \\ 0 \end{bmatrix}, \quad \mathbf{G}_{AA^T} = \begin{bmatrix} 0.72 & 0.64 & 0.28 \\ 0.46 & -0.73 & 0.50 \\ -0.52 & 0.23 & 0.82 \end{bmatrix}$$

$$\mathbf{A} = \mathbf{G}_{AA^T}\mathbf{\Lambda}\mathbf{G}^T_{A^TA} = \begin{bmatrix} 2 & 6 & 8 \\ 3 & 1 & 5 \end{bmatrix}.$$

3.2.3 Quadratic Forms

To write a quadratic form $\mathbf{Q}(x)$ a symmetric matrix $\mathbf{A}$ of size $n \times n$ and a vector $\mathbf{x} \in R^n$ are needed.

Then

$$\mathbf{Q}(x) = \mathbf{x}^T \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \qquad (3.7)$$

Let $\mathbf{A}$ denote an $n \times n$ symmetric matrix with real entries and let $\mathbf{x}$ denote an $n \times 1$ column vector [20].

$\mathbf{Q} = \mathbf{x}^T \mathbf{A}\mathbf{x}$ is said to be a quadratic form. Note that

$$\mathbf{Q} = \mathbf{x}^T \mathbf{A}\mathbf{x} = (x_1 ... x_n) \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} (x_1, ..., x_n)^T$$

$$\begin{aligned} &= a_{11}x_1^2 + a_{12}x_1x_2 + ... + a_{1n}x_1x_n \\ &+ a_{22}x_2x_1 + a_{22}x_2^2 + ... + a_{2n}x_n^2 \\ &+ ... \\ &+ ... \\ &+ ... \\ &+ a_{n1}x_nx_1 + a_{n2}x_nx_2 + ... + a_{nn}x_n^2 \\ &= \sum_{i \le j} a_{ij} x_i x_j \end{aligned}$$

For example, consider the matrix

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

The general quadratic form $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{x}^T\mathbf{A}\mathbf{x} = [x_1 \quad x_2 \quad x_3]\begin{bmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

$$= [x_1 \quad 2x_2 \quad 4x_3]\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

$$= [x_1^2 + 2x_2^2 + x_2^3]$$

$\mathbf{Q} = \mathbf{x}^T\mathbf{A}\mathbf{x}$**:** $\mathbf{A}$ quadratic form is said to be:

a**:** negative definite: $\mathbf{Q} < 0$ when $\mathbf{x} \neq 0$

b**:** negative semidefinite: $\mathbf{Q} \leq 0$ for all $\mathbf{x}$ and $\mathbf{Q} = 0$ for some $\mathbf{x} \neq 0$

c**:** positive definite: $\mathbf{Q} > 0$ when $\mathbf{x} \neq 0$

d**:** positive semidefinite: $\mathbf{Q} > 0$ for all $\mathbf{x}$ and $\mathbf{Q} = 0$ for some $\mathbf{x} \neq 0$

e**:** indefinite: $\mathbf{Q} > 0$ for some $\mathbf{x}$ and $\mathbf{Q} < 0$ for some other $\mathbf{x}$.

Theorem 3.2: Let matrices $\mathbf{A}$ and $\mathbf{B}$ be symmetric and $\mathbf{B} > 0$. Then the quadratic form $\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}\mathbf{B}^T\mathbf{x}}$ has a maximum which is the largest eigenvalue of $\mathbf{B}^{-1}\mathbf{A}$.This can be written as

$$\max_{\mathbf{x}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}\mathbf{B}^T\mathbf{x}} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = \min_{\mathbf{x}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{B}\mathbf{x}}.$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$. The vector which maximises (minimizes) $\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}\mathbf{B}^T\mathbf{x}}$ is the eigenvector of $\mathbf{B}^{-1}\mathbf{A}$ which corresponds to the largest (smallest) eigenvalue of $\mathbf{B}^{-1}\mathbf{A}$. If $\mathbf{x}^T\mathbf{B}\mathbf{x} = 1$we get

$$\max_{\mathbf{x}} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = \min_{\mathbf{x}} \mathbf{x}^T\mathbf{A}\mathbf{x}.$$

Proof: $\mathbf{B}^{1/2} = \mathbf{\Gamma}_{\mathbf{B}}\mathbf{\Lambda}^{1/2}{}_{\mathbf{B}}\mathbf{\Gamma}^{T}{}_{\mathbf{B}}$ is symmetric. Then $\mathbf{x}\mathbf{B}^{T}\mathbf{x} = \left\|\mathbf{x}^{T}\mathbf{B}^{1/2}\right\|^{2} = \left\|\mathbf{B}^{1/2}\mathbf{x}\right\|^{2}$.

Set $\mathbf{y} = \dfrac{\mathbf{B}^{1/2}\mathbf{x}}{\left\|\mathbf{B}^{1/2}\mathbf{x}\right\|}$. yields

Then

$$\max_{\mathbf{x}} \frac{\mathbf{x}^{T}\mathbf{A}\mathbf{x}}{\mathbf{x}^{T}\mathbf{B}\mathbf{x}} = \max_{\{\mathbf{y}:\mathbf{y}^{T}\mathbf{y}=1\}} \mathbf{y}^{T}\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{y}. \qquad (3.8)$$

Let $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = \mathbf{\Gamma}\Lambda\mathbf{\Gamma}^{T}$ be the spectral decomposition and vector $\mathbf{z}$ defined as

$\mathbf{z} = \mathbf{\Gamma}^{T}\mathbf{y}$,

Then

$$\mathbf{z}^{T}\mathbf{z} = \mathbf{y}^{T}\mathbf{\Gamma}\mathbf{\Gamma}^{T}\mathbf{y} = \mathbf{y}^{T}\mathbf{y}.$$

Thus (3.8**)** is equivalent to

$$\max_{\{\mathbf{z}:\mathbf{z}^{T}\mathbf{z}=1\}} \mathbf{z}^{T}\mathbf{\Lambda}\mathbf{z} = \max_{\{\mathbf{z}:\mathbf{z}^{T}\mathbf{z}=1\}} \sum_{i=1}^{P} \lambda_i z_i^2$$

but

$$\max_{\mathbf{z}} \sum \lambda_i z_i^2 \le \lambda_1 \max_{\mathbf{z}} \sum z_i^2 = \lambda_1$$

When $\mathbf{z} = (1,0,0,...,0)^{T}$ maximum is obtained. For $\mathbf{y} = \boldsymbol{\gamma}_1$, $\mathbf{x} = \mathbf{B}^{-1/2}\boldsymbol{\gamma}_1$. is

obtained.

The eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$ and $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ are equal. This completes the proof.

Lagrange method can also be used to prove the same theorem. That is maximize $\mathbf{x}^{T}\mathbf{A}\mathbf{x}$ Subject to $\mathbf{x}^{T}\mathbf{B}\mathbf{x}$. Then the Lagrange function is $\mathbf{L} = \mathbf{x}^{T}\mathbf{A}\mathbf{x} - \lambda[\mathbf{x}^{T}\mathbf{B}\mathbf{x} - 1]$. Hence λ is the lag-range constant.

$$\max_{\mathbf{x}} \mathbf{x}^{T}\mathbf{A}\mathbf{x} = \max_{\mathbf{x}}[\mathbf{x}^{T}\mathbf{A}\mathbf{x} - \lambda(\mathbf{x}^{T}\mathbf{B}\mathbf{x} - 1)].$$

Setting the first derivative with respect to $\mathbf{x}$ is equal to $\mathbf{0}$:

$$\frac{\partial L}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} - 2\lambda\mathbf{B}\mathbf{x} = 0$$

So

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

By the definition of eigenvector and eigenvalue, our maximiser $\mathbf{x}^*$ is $\mathbf{B}^{-1}\mathbf{A}$ eigenvector corresponding to eigenvalue λ.

Hence

$$\max_{\{\mathbf{x}:\mathbf{x}^T\mathbf{B}\mathbf{x}=1\}} \mathbf{x}^T\mathbf{A}\mathbf{x} = \max_{\{\mathbf{x}:\mathbf{x}^T\mathbf{B}\mathbf{x}=1\}} \mathbf{x}^T\mathbf{B}\mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \max_{\{\mathbf{x}:\mathbf{x}^T\mathbf{B}\mathbf{x}=1\}} \mathbf{x}^T\mathbf{B}\lambda\mathbf{x} = \max\lambda$$

gives the maximum eigenvalue of $\mathbf{B}^{-1}\mathbf{A}\mathbf{x}$. Corresponding eigenvector is the maximiser $\mathbf{x}^*$.

3.2.4 Derivative

In this section matrix notation for the derivatives will be introduced. Let $f : R^p \to R$ with p variables represented by a $(p \times 1)$ vector $\mathbf{x}$. Let also $\frac{\partial f(x)}{\partial x}$ be the column vector of partial derivatives $\frac{\partial f(x)}{\partial x_j}$, $j = 1,..., p$ and $\frac{\partial f(x)}{\partial x^T}$ be the row vector of the same derivative. $\frac{\partial f(x)}{\partial x}$ is called the gradient of f. Second order partial derivatives are expressed as $\frac{\partial^2 f(x)}{\partial x \partial x^T}$ is the $p \times p$ matrix of elements $\frac{\partial^2 f(x)}{\partial x_j \partial x_j^T}$, $i = 1,..., p$ and $j = 1,..., p$. $\frac{\partial^2 f(x)}{\partial x \partial x^T}$ is called the Hessian of f. When $\mathbf{a}$ is a $(p \times 1)$ vector and $\mathbf{A} = \mathbf{A}^T$ is a $(p \times p)$ matrix

$$\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a},$$
$$\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

can be written. The Hessian of the quadratic form $Q = \mathbf{x}^T \mathbf{A}\mathbf{x}$ is expressed as

$$\frac{\partial^2 f(x)}{\partial x \partial x^T} = 2\mathbf{A}.$$

3.3 Statistical Parameters in Multivariate Case

3.3.1 General on multivariate statistical

Multivariate statistics is the branch of statistics which deals with the analysis of data belonging to many variables. The analysis of simultaneous measurements necessitates the use of multivariate techniques. In this section a brief review of some descriptive statistics concepts pertaining to the multivariate case will be highlighted.

3.3.2 Multivariate sample mean

Let $x_1,...,x_n$ be a particular realization (a random sample of size *n*) of the random variables $X_1,...,X_n$. Then the arithmetic mean $\bar{x}$ of the random sample gives the center of gravity or the average distance from the origin on the real line $\Re$. It is computed as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3.9)$$

3.3.3 Multivariate sample variance

If the random variable *X* represents a particular characteristic of a population, then the variance of the population is defined as $\text{var}(X) = \sigma^2 = E(X-\mu)^2 = E(X^2) - \mu^2$. Variance measures the average squared deviation from the mean. The larger the variance, the more data values are spread around the mean. The sample variance s^2 is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{3.10}$$

or

$$s^2 = \frac{\sum_{i=1}^{n} x^2{}_i - n\bar{x}^2}{n-1} \tag{3.11}$$

Since a sample is a subset of the population, its variance s^2 is can not be expected to be the same as the population variance σ^2. However, s^2 is an unbiased estimator for σ^2, which means $E(S^2) = \sigma^2$.

The following properties on variance – covariance holds.

1. $\sigma^2_{\mathbf{a}^T X} = \mathbf{a}^T \sigma^2_X \mathbf{a} = \sum_{i,j} a_i a_j \sigma_{x_i x_j}$
2. $\sigma^2_{\mathbf{A}X+b} = \mathbf{A}\sigma^2_X \mathbf{A}^T$
3. $\sigma^2_{X+Y} = \sigma^2_X + \sigma_{X,Y} + \sigma_{Y,X} + \sigma^2_Y$
4. $\sigma_{(X+Y,Z)} = \sigma_{(X,Z)} + \sigma_{(Y,Z)}$
5. $\sigma_{(\mathbf{A}X,\mathbf{B}Y)} = \mathbf{A}\sigma_{(X,Y)}\mathbf{B}^T$

3.3.4 Multivariate sample covariance

Let random variables *X* and *Y* with joint probability density function *f*(*x*, *y*) be given. Covariance between these random variables is defined as

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].$$

Here $\mu_X = E(X)$ and $\mu_Y = E(Y)$. If $\sigma_{XY} > 0$, it means r.v. s are simultaneously increasing or decreasing, bu not necessarily at the same rate. $\sigma_{XY} < 0$ would mean an increase in one variable corresponds to a decrease in the other. In r.v.s are independent, then $\sigma_{XY} = 0$. Fr the bivariate case it can be shown that

$$\sigma_{XY} = E(XY) - E(X)E(Y) \ .$$

Addition or multiplication of the random variables X and Y results in a new random variable. Then, if $Z = X + Y$

$$E(Z) = E(X + Y) \; = \; E(X) \; + \; E(Y) \tag{3.12}$$

and if $Z = XY$

$$E(Z) = E(XY) \; = \; E(X)E(Y) \text{, when } X \text{ and } Y \text{ are independent.} \tag{3.13}$$

If f_x and f_y are the marginal densities of r.v.s X and Y respectively, then the r.v.s X and Y are independent iff $f(x, y) = g(x)h(y)$. Equation (3.12) holds regardless the r.v.s being independent or not. In the case of independence the covariance $\sigma_{XY} = 0$, but the vice versa case is not always true.

$$\sigma_{XY} \; = \; E(XY) - E(X)E(Y)$$

Based on the definition of the population covariance, the sample covariance can be written as

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3.14}$$

and it can also be shown that

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{n-1} \tag{3.15}$$

In application it is impossible to have $s_{xy} = \sigma_{XY}$ This means, the chance is almost zero that $P(s_{xy} = \sigma_{XY}) = 0$. s_{xy} is an unbiased estimator for σ_{XY}, i.e. $E(s_{xy}) = \sigma_{XY}$. When $\sigma_{XY} = 0$, it does not mean that any random sample from the same population will have zero covariance. One way of ensuring that a sample from a continuous bivariate distribution will have zero covariance is for the experimenter to choose the values of *x* and *y* so that $s_{xy} = 0$. However, this causes deviation from the concept of

random sampling. One way to see that s_{xy} measures only linear relationships is by seeing that the computation of the slope of simple linear regression line includes s_{xy} as its numerator.

$$\widehat{B} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

Thus s_{xy} is proportional to the slope, which shows only the linear relationship between Y and X. Variables with zero sample covariance can be said to be orthogonal. By definition, if the dot product of the vectors $\mathbf{a}^T = [a_1, a_2, ..., a_n]$ and $\mathbf{b}^T = [b_1, b_2, ..., b_n]$ is $\mathbf{a} \cdot \mathbf{b} = 0$.

3.3.5 Multivariate sample correlation

Since the covariance depends on the scale of measurement of X and Y, it is difficult to compare covariances between different pairs of variables. For example, if we change a measurement from inches to centimeters, the covariance will change. To find a measure of linear relationship that is invariant to changes of scale, we can standardize the covariance by dividing by the standard deviations of the two Variables. This standardized covariance is called a linear correlation coefficient. The population correlation coefficient of two random variables X and Y is

$$\rho_{XY} = corr(x, y) = \frac{\rho_{XY}}{\rho_X \rho_Y} = \frac{E[(X - \mu_X)(y - \mu_X)}{\sqrt{E(X - \mu_X)^2}\sqrt{E(Y - \mu_Y)^2}} \quad . \qquad (3.16)$$

Given *n* pairs of sample data $(x_i, y_i);\ i = 1, \ldots, n$ with respective sample averages $\bar{x}$ and $\bar{y}$, the sample correlation coefficient is defined as

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad . \qquad (3.17)$$

In both the population and sample cases we have $-1 \le \rho_{xy} \le 1$ and $-1 \le r_{xy} \le 1$ respectively.

3.3.6 Variance and covariance matrix

Variance and covariance are often display jointly in a variance-covariance matrix. The variances appear along the diagonal and covariances appear in the off-diagonal elements. If the random variable X is n-dimensional, then the vector $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ represents the random variables, each with finite variance. Then the covariance matrix **Σ,** is the matrix whose (*i*, *j*) entry is the covariance

$$\Sigma_{ij} = \mathrm{cov}(X_i, \mathrm{Y}_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

where

$$\mu_i = E(X_i)$$

is the expected value of the i^{th} random variable in the vector $\mathbf{X}$ .In other words, we have

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{x_1 x_1} & \cdots & \sigma_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ \sigma_{x_n x_1} & \cdots & \sigma_{x_n x_n} \end{pmatrix} \qquad (3.18)$$

The inverse of this matrix $\mathbf{\Sigma}^{-1}$ is the inverse covariance matrix, also known as the concentration matrix or precision matrix [11]. The sample covariance matrix **S** can be written as

$$\mathbf{S} = \begin{pmatrix} s_{x_1x_n} & \cdots & s_{x_1x_n} \\ \vdots & \ddots & \vdots \\ s_{x_nx_1} & \cdots & s_{x_nx_n} \end{pmatrix} \tag{3.19}$$

Expected value of the covariance matrix $\mathbf{S}$ is

$$E(\mathbf{S}) = \frac{n-1}{n}\mathbf{\Sigma} = \mathbf{\Sigma} - \frac{1}{n}\mathbf{\Sigma} \rightarrow E\left(\frac{n}{n-1}\mathbf{S}\right) = \mathbf{\Sigma}$$

It is understood that $[n/(n-1)]\mathbf{S}$ is an unbiased estimator $\mathbf{\Sigma}$, but $\mathbf{S}$ is a biased estimator and the $bias = E(\mathbf{S}) - \mathbf{\Sigma} = -(1/\mathrm{n})\mathbf{\Sigma}$. It can be shown that $E\left(\frac{n}{n-1}\mathbf{S}\right) = \mathbf{\Sigma}$ as below.

$\bar{\mathbf{X}} = (\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_n)/n$.

$$\begin{aligned} E(\bar{\mathbf{X}}) &= \mathrm{E}(\frac{1}{n}\mathbf{X}_1 + \frac{1}{n}\mathbf{X}_2 + \cdots + \frac{1}{n}\mathbf{X}_n) \\ &= \mathrm{E}(\frac{1}{n}\mathbf{X}_1) + \mathrm{E}(\frac{1}{n}\mathbf{X}_2) + \cdots + \mathrm{E}(\frac{1}{n}\mathbf{X}_n) \\ &= \frac{1}{n}\mathrm{E}(\mathbf{X}_1) + \frac{1}{n}\mathrm{E}(\mathbf{X}_2) + \cdots + \frac{1}{n}\mathrm{E}(\mathbf{X}_n) = \frac{1}{n}\boldsymbol{\mu} + \frac{1}{n}\boldsymbol{\mu} + \cdots + \frac{1}{n}\boldsymbol{\mu} = \boldsymbol{\mu} \end{aligned}$$

next,

$$\begin{aligned} (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T &= \left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j - \boldsymbol{\mu}\right)\left(\frac{1}{n}\sum_{l=1}^{n}\mathbf{X}_l - \boldsymbol{\mu}\right)^T \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_l - \boldsymbol{\mu})^T \end{aligned}$$

$$\mathrm{cov}(\bar{\mathbf{X}}) = \mathrm{E}(\bar{\mathbf{X}} - \boldsymbol{\mu})\,\mathrm{E}(\bar{\mathbf{X}} - \boldsymbol{\mu})^T = \frac{1}{n}\left(\sum_{j=1}^{n}\sum_{l=1}^{n}\mathrm{E}(\bar{\mathbf{X}}_j - \boldsymbol{\mu})\,\mathrm{E}(\bar{\mathbf{X}}_l - \boldsymbol{\mu})^T\right)$$

For $j \neq l$ each entry in $E(\bar{\mathbf{X}}_j - \boldsymbol{\mu})\,\mathrm{E}(\bar{\mathbf{X}}_l - \boldsymbol{\mu})^T$ is zero as each entry in the covariance between the independent components of $\mathbf{X}_j$ and of $\mathbf{X}_l$.

Therefore,

$$\mathrm{cov}(\bar{\mathbf{X}}) = \frac{1}{n^2}\left(\sum_{j=1}^{n}\mathrm{E}(\bar{\mathbf{X}} - \boldsymbol{\mu})\,\mathrm{E}(\bar{\mathbf{X}} - \boldsymbol{\mu})^T\right)$$

Since population covariance matrix $\boldsymbol{\Sigma} = E(\bar{\mathbf{X}}_j - \boldsymbol{\mu})\mathrm{E}(\bar{\mathbf{X}}_j - \boldsymbol{\mu})^T$ we can write

$$\mathrm{cov}(\bar{\mathbf{X}}) = \frac{1}{n^2}\left(\sum_{j=1}^{n} E(\bar{\mathbf{X}}_j - \boldsymbol{\mu})E(\bar{\mathbf{X}}_j - \boldsymbol{\mu})^T\right) = \frac{1}{n^2}(\boldsymbol{\Sigma} + \boldsymbol{\Sigma} + \cdots + \boldsymbol{\Sigma})$$

$$= \frac{1}{n^2}(n\boldsymbol{\Sigma}) = \frac{1}{n}\boldsymbol{\Sigma}$$

The $(i,k)^{th}$ element of $(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^T$ is $(X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)$. Sums of products and cross product s are written in matrix form as

$$\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^T = \sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})\mathbf{X}_j^T + \left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})\right)(-\bar{X})^T$$

$$= \sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

Note that $\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}}) = \mathbf{0}$ and $n\bar{\mathbf{X}}^T = \sum_{j=1}^{n}\bar{\mathbf{X}}^T$. Then

$$E\left(\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T\right) = \sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j^T) - nE(\bar{\mathbf{X}}\bar{\mathbf{X}}^T)$$

Remembering the fact given a random vector $\mathbf{V}$ having $E(\mathbf{V}) = \boldsymbol{\mu}_\mathbf{v}$ and $\mathrm{cov}(\mathbf{V}) = \boldsymbol{\Sigma}_\mathbf{v}$,

$E(\mathbf{V}\mathbf{V}^T) = \boldsymbol{\Sigma}_\mathbf{v} + \boldsymbol{\mu}_\mathbf{v}\boldsymbol{\mu}_\mathbf{v}$ leading to

$$E(\mathbf{X}_j\mathbf{X}_j^T) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T \text{ and } E(\bar{\mathbf{X}}\bar{\mathbf{X}}^T) = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T .$$

Based on these results

$$\sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j^T) - nE(\bar{\mathbf{X}}\bar{\mathbf{X}}^T) = n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}^T - n\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T\right) = (n-1)\boldsymbol{\Sigma}$$

can be written and since $\mathbf{S} = \left(\frac{1}{n}\right)\left(\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T\right)$, the desired result

$$E(\mathbf{S}) = \frac{(n-1)}{n}\boldsymbol{\Sigma} \text{ is obtained.}$$

3.3.7 Correlation matrix

The correlation matrix can be seen as the covariance matrix of the standardized random variables.Let $X = (X_1, ..., X_n)$ be n-dimensional random sample, the correlation value among X_i and X_j is denoted by $r_{x_i x_j}$ and give by

$$r_{x_i x_j} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i) \sum_{k=1}^{n} (x_{jk} - \bar{x}_j)}}$$

Obtained $r_{x_i x_j}$ values can be represented in $(n \times n)$ matrix from

$$\mathbf{R} = \begin{pmatrix} r_{x_1 x_n} & \cdots & r_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ r_{x_n x_1} & \cdots & r_{x_n x_n} \end{pmatrix} \qquad (3.20)$$

3.3.8 Relationship between covariance and correlation Matrices

In equations (3.10) and (3.17) computation of sample variance and correlation coefficient are given. In multivariate case the covariance matrix $\mathbf{\Sigma}$ is give in equation (3.18) correlation matrix in (3.20). Relationship between **S** and **R** are explained below. Let $\mathbf{D}^{1/2}$ be defined as the $(p \times p)$ sample standard deviation matrix. Its sinverseis, $(\mathbf{D}^{1/2})^{-} = \mathbf{D}^{-1/2}$. Writing $\mathbf{D}^{1/2}$ and $\mathbf{D}^{-1/2}$ in matrix form we have

$$\mathbf{D}^{1/2}{}_{(p \times p)} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sqrt{s_{pp}} \end{bmatrix}$$

and

$$\mathbf{D}^{-1/2}{}_{(p\times p)} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$$

Since

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix}$$

then

$$\mathbf{R} = \begin{bmatrix} \frac{s_{11}}{\sqrt{s_{11}}\sqrt{s_{11}}} & \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{12}}} & \cdots & \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{1p}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{s_{2p}}{\sqrt{s_{2p}}\sqrt{s_{pp}}} & \cdots & \frac{s_{pp}}{\sqrt{s_{pp}}\sqrt{s_{pp}}} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

Let **D** be a diagonal matrix obtained from **S.** The relation between covariance and correlation matrices is defined as

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

$$\mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$$

Clearly the knowledge of S enables the easy computation of R **and** vice versa [16].

Chapter 4

PRINCIPAL COMPONENT ANALYSIS VIA DIFFERENT APPROACHES TO THE DATA MATRIX

Principal component analysis (PCA) is a dimension reduction technique in a given data matrix of size $n \times p$, when the number of columns representing the variables are very large. This reduction using principal components (PC) becomes essential in order to alleviate the difficulty of interpreting the variation in a large number of variables. Reducing the dimension by means of finding linear combinations of the variables associated with the variation in each variable. Through this approach only the first few PCs tends to account for over 90% of variation in the data. Then, instead of using a large number of variables to figure out the true variation in the data, using only a few (2 or 3) of the PCs will be a much faster way of identifying and explaining the variation within a given data set. Dimension reduction can be applied directly to the raw data, to the centered data, or to the normalized data. Each case has its advantages and disadvantages depending on the nature of the data. In this chapter, the PCA technique will be explained and its application to different data cases will be given in detail. In this chapter we will talk about center the data, raw data and principle component analysis. We will test the original data (raw data) without calculating the center the data and also tested by centering the data and then compare both cases and which one better to use. Now we will talk about principle component analysis.

4.1 Theory of Principle Component Analysis

PCA can be regarded as transforming a given set of *p* random variables to another set of variables (PCs) $\mathbf{Y}^T = \left[Y_1,...,Y_p \right]$. Geometrically, PCs represent the selection of a new coordinate system obtained by rotating the original system $X_1,...,X_p$. The new coordinate system obtained represents the directions with maximum variability. Given a random vector $\mathbf{X}^T = \left[X_1,...,X_p \right]$ representing *p* random variables with covariance matrix **Σ** and an arbitrary $p \times p$ coefficient matrix **A,** the following linear combinations can be written.

$$\begin{aligned} Y_1 &= \mathbf{a}_1^T \mathbf{X} = a_{11}X_1 + \cdots + a_{1p}X_p \\ Y_2 &= \mathbf{a}_2^T \mathbf{X} = a_{21}X_1 + \cdots + a_{2p}X_p \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_p &= \mathbf{a}_p^T \mathbf{X} = a_{p1}X_1 + \cdots + a_{pp}X_p \end{aligned} \tag{4.1}$$

These linear combinations are the uncorrelated PCs. The first principal component has the highest variance. From equation (4.1) variance and covariance are given as

$$Var(Y_i) = \mathbf{a}^T \mathbf{\Sigma} \mathbf{a}, \qquad i = 1,\ldots, p \tag{4.2}$$

$$Cov(Y_i, Y_k) = \mathbf{a}_i^T \mathbf{\Sigma} \mathbf{a}_k, \qquad i,k = 1,\ldots, p \tag{4.3}$$

In place of any arbitrary coefficient vector **a**, vectors with unit length **u** is adopted without loss of generality. Then the first PC $\mathbf{u}_1^T \mathbf{X}$ will be such that $Var(\mathbf{u}_1^T \mathbf{X})$ is maximum subject to $\mathbf{u}_1^T \mathbf{u}_1 = 1$. The i^{th} PC $\mathbf{u}_i^T \mathbf{X}$ will be such that $Var(\mathbf{u}_i^T \mathbf{X})$ is maximum subject to $\mathbf{u}_i^T \mathbf{u}_i = 1$ and $Cov(\mathbf{u}_i^T \mathbf{X}, \mathbf{u}_k^T \mathbf{X}) = 0$ for $k < i$.

Theorem 4.1: Let **B** be a positive definite matrix with eigenva1ues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and associated normalized eigenvectors $\mathbf{e}_1,\ldots,\mathbf{e}_p$. Then

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{B}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_1 \text{ when } \mathbf{x} = \mathbf{e}_1$$
$$\min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{B}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_p \text{ when } \mathbf{x} = \mathbf{e}_p \qquad (4.4)$$

Further

$$\max_{\mathbf{x}\perp\mathbf{e}_1,\ldots,\mathbf{e}_k} \frac{\mathbf{x}^T\mathbf{B}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_{k+1} \text{ when } \mathbf{x} = \mathbf{e}_{1+1},\ k = 1,\ldots,p-1 \qquad (4.5)$$

For proof see [18]

Let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues and $\mathbf{e}_1,\ldots,\mathbf{e}_p$ be the corresponding eigenvectors of the covariance matrix **Σ**. The i^{th} principal component can be written as

$$Y_i = \mathbf{e}_j^T\mathbf{X} = e_{i1}X_1 + \cdots + e_{ip}X_p,\ \ i = 1,\ldots,p$$

with

$$Var(Y_i) = \mathbf{e}_i^T\mathbf{\Sigma}\mathbf{e}_i = \lambda_i,\ \ i = 1,\ldots,p \text{ and } Cov(Y_i,Y_k) = \mathbf{e}_i^T\mathbf{\Sigma}\mathbf{e}_k,\ \ i \neq k \qquad (4.6)$$

Equation (4.6) can be proved based on Theorem 4.1, equation (4.4)

Let matrix **B= Σ** in theorem 4.1. If $\mathbf{a} = \mathbf{e}_1$ and $\mathbf{e}_1$ being a normalized vector ($\mathbf{e}_1^T\mathbf{e}_1 = 1$), then

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \lambda_1 = \frac{\mathbf{e}^T\mathbf{\Sigma}\mathbf{e}}{\mathbf{e}^T\mathbf{e}} = \mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e} = \text{var}(Y_1)$$

Similarly using (4.5) from Theorem 4.1

$$\max_{\mathbf{a}\perp\mathbf{e}_1,\ldots,\mathbf{e}_k} \frac{\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \lambda_{k+1}\ ,\ k = 1,\ldots,p-1$$

When

$$\mathbf{a} = \mathbf{e}_{k+1} \text{ and } \mathbf{e}_{k+1}^T\mathbf{e}_i = 0 \text{ for } i = 1,\ldots k,\ k = 1,\ldots,p-1$$

$$\frac{\mathbf{e}_{k+1}^T\mathbf{\Sigma}\mathbf{e}_{k+1}}{\mathbf{e}_{k+1}^T\mathbf{e}_{k+1}} = \mathbf{e}_{k+1}^T\mathbf{\Sigma}\mathbf{e}_{k+1} = Var(Y_{k+1}) = \lambda_{k+1}$$

To show that when $\mathbf{e}_i^T\mathbf{e}_k = 0,\ i \neq k$ results in $Cov(Y_i,Y_k) = 0$, remember that the eigenvectors of **Σ** are orthogonal if all $\lambda_1,\ldots\lambda_p$ are not equal. Hence, any two eigenvectors will satisfy $\mathbf{e}_i^T\mathbf{e}_k = 0,\ i \neq k$. Multiplying both sides of $\mathbf{\Sigma}\mathbf{e}_k = \lambda_k\mathbf{e}_k$ by $\mathbf{e}_i^T$ gives

$$Cov(Y_i, Y_k) = \mathbf{e}_i^T \boldsymbol{\Sigma}\, \mathbf{e}_k = \mathbf{e}_i^T \lambda_k \mathbf{e}_k = \lambda_k \mathbf{e}_i^T \mathbf{e}_k = 0 \text{ for } i \neq k$$

It is understood that the principal components are uncorrelated and their variances are the eigenvalues of the covariance matrix **Σ**.

Remembering that the diagonal elements of **Σ** are the variances of X_j, $j = 1,\ldots, p$, and then the following relationship becomes evident.

$$\sigma_{11} + \cdots \sigma_{pp} = \sum_{j=1}^{p} Var(X_j) = \lambda_1 + \cdots + \lambda_p = \sum_{j=1}^{p} Var(Y_j)$$

Then total population variance becomes $\sigma^2 = \sigma_{11} + \cdots \sigma_{pp} = \lambda_1 + \cdots + \lambda_p$ .It is also worth mentioning that magnitude of each element of the vector $\mathbf{e}_i^T = (e_{i1}, \ldots, e_{ik}, \ldots, e_{ip})$ indicates the importance of corresponding variable in the PC. The vector element e_{ik} is also proportional to the correlation between Y_i and X_k. This correlation can be computed from

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}},\ i, k = 1, \ldots, p \qquad (4.7)$$

Obviously ρ_{Y_i, X_k} measures the linear correlation between the k^{th} random variable and the concerned i^{th} PC. Tendency is that random variables assigned PC scores $|e_{ik}|$ that are large will have high ρ_{Y_i, X_k} values.

Example 4.1: The following data consisting of 5 variables represent various characteristics of silver zinc battery affecting there life time [15]. Magnitude of data values for each variable is quite different. Therefore, the PC analysis will be applied to the raw data, centered data, and standardized data. Results and interpretations will be explained and compared.

Table 4.1: Battery failure data represented by the data matrix **X**.

$\mathbf{x}_1$ Charge rate(amps)	$\mathbf{x}_2$ Discharge rate(amps)	$\mathbf{x}_3$ Depthof discharge (%ofratedof amperehours)	$\mathbf{x}_4$ Temperature ($^{\circ}C$)	$\mathbf{x}_5$ End of Charge Voltage (volts)
0.375	3.13	60.0	40	2.00
1.000	3.13	76.8	30	1.99
1.000	3.13	60.0	20	2.00
1.000	3.13	60.0	20	1.98
1.625	3.13	43.2	10	2.01
1.625	3.13	60.0	20	2.00
1.625	3.13	60.0	20	2.02
0.375	5.00	76.8	10	2.01
1.000	5.00	43.2	10	1.99
1.000	5.00	43.2	30	2.01
1.000	5.00	100.0	20	2.00
1.625	5.00	76.8	10	1.99
0.375	1.25	76.8	10	2.01
1.000	1.25	43.2	10	1.99
1.000	1.25	76.8	30	2.00
1.000	1.25	60.0	0	2.00
1.625	1.25	43.2	30	1.99
1.625	1.25	60.0	20	2.00
0.375	3.13	76.8	30	1.99
0.375	3.13	60.0	20	2.00

$\mathbf{S}$ Computed from the raw data matrix **X** is:

$$\mathbf{S} = \begin{pmatrix} 0.2251 & -0.0587 & -2.3039 & -0.6414 & 0.0000 \\ -0.0587 & 2.0266 & 4.2253 & -0.0403 & 0.0009 \\ -2.3039 & 4.2253 & 239.1225 & 10.3368 & 0.0030 \\ -0.6414 & -0.0403 & 10.3368 & 99.7368 & -0.0111 \\ 0.0000 & 0.0009 & 0.0030 & -0.0111 & 0.0001 \end{pmatrix}$$

The eigenvalues of $\mathbf{S}$ are $\lambda_1 = 239.98,\ \lambda_2 = 98.98,\ \lambda_3 = 1.95,\ \lambda_4 = 0.20,\ \lambda_5 = 0.0001$ and the corresponds eigenvectors forms the columns of the Γ matrix. The elements of each column of the Γ matrix are the coefficients of the principal components $Y_i,\ i = 1, \ldots, p$.

$$\boldsymbol{\Gamma} = \begin{pmatrix} -0.0098 & 0.0048 & 0.0108 & 0.9999 & -0.0000 \\ 0.0177 & 0.0036 & -0.9998 & 0.0109 & -0.0004 \\ 0.9971 & 0.0735 & 0.0180 & 0.0092 & -0.0000 \\ 0.0735 & -0.9973 & -0.0022 & 0.0055 & 0.0001 \\ 0.0000 & 0.0001 & -0.0004 & 0.0000 & 1.0000 \end{pmatrix}$$

Total variation in the data $\sum_{i=1}^{k} \lambda_k = 341.11$.First 2 eigenvalues represe

$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{k} \lambda_k} = \frac{338.96}{341.11} = 0.994$ or 99.4% of total variation in the data. Using the $\boldsymbol{\Gamma}$ matrix, principal's components are written

$Y_1 = -0.0098X_1 + 0.0177X_2 + 0.9971X_3 + 0.0735X_4$

$Y_2 = 0.0048X_1 + 0.0036X_2 + 0.0735X_3 - 0.9973X_4 + 0.0001X_5$

$Y_3 = 0.0108X_1 - 0.9998X_2 + 0.0108X_3 - 0.0022X_4 - 0.0004X_5$

$Y_4 = 0.9999X_1 + 0.0109X_2 + 0.0092X_3 + 0.0055X_4$

$Y_5 = -0.0004X_2 - 0.0001X_4 + X_5$.

Evidently each PC is dominated by one variable only, while remaining variables have almost negligible influence. Since 99.4% of variation in the data is represented by the first two PCs Y_1 and Y_2, close inspection is necessary.

First PC Y_1 is a linear combination of the variables X_1, X_2, X_3, and X_4. However, the coefficient of X_3 (depth rate of discharge measured as %rated amps/hr) is the largest in absolute terms, dominating Y_1. The temperature (^0C) X_4 also has a notable influence on Y_1. Second PC Y_2 is a linear combination of all 5 variables. Here, X_4 (Temperature (^0C)) is the dominating variable while the depth rate of discharge measured as %rated amps/hr X_3 also has a some influence on Y_2. Since remaining PCs have negligible contribution to the total variation in data, they will not be considered.

The relationship $Var(Y_i) = \lambda_i$ can be checked for this example. For the first PC

$$Var(Y_1) = Var(-0.0098X_1 + 0.0177X_2 + 0.9971X_3 + 0.0735X_4) =$$
$$(-0.01)(0.23) + (0.018)(2.03) + (0.997)(239.12) + (0.074)(99.74) = 245.81 \cong \lambda_1$$

Since PCs are independent as an example Y_1 and Y_2 are checked for independence,

$$\text{cov}(Y_1, Y_2) = C\text{ov}((-0.0098X_1 + 0.0177X_2 - 0.9971X_3 + 0.0735X_4 + 0X_5),$$
$$(0.0048X_1 + 0.0036X_2 + 0.0735X_3 - 0.9973X_4 + 0.0001X_5))$$

Hint: Given random variables $X_1, \ldots, X_n$ and their linear combinations

$Y_1 = \sum_{i=1}^{n} a_i X_i$ and $Y_2 = \sum_{i=1}^{n} b_i X_i$,

$$Cov(Y_1, Y_2) = \sum_{i=1}^{n} a_i b_i Var(X_i) + \sum_{i<j} (a_i b_j + a_j b_i) Cov(X_i, X_j) \qquad (4.8)$$

Using equation 4.8 the covariance between the PCs Y_1 and Y_2 , are given in appendix A. The linear correlation between each PC Y_i and the variables X_i is also worth considering. They are computed as

$$\rho_{Y_i X_j} = \frac{e_{ij}\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}},\ i, j = 1, \ldots, p \qquad (4.9)$$

Equation 4.9 becomes $r_{Y_i X_j} = \frac{e_{ij}\sqrt{\lambda_i}}{s_{jj}};\ i, j = 1, \ldots, p$ for the sample data. Then the linear correlation coefficient between the variables and the first 2 PCs that accounts for 99.4% of total variation in the raw data and the variables are given below.

Table 4.2: Principle component scores and correlation between Y_1 and X_i for raw data.

	X_1	X_2	X_3	X_4
e_{ii}	-0.0098	0.0177	0.9971	0.0735
$r_{Y_i X_i}$	-0.3205	0.193	0.999	0.114

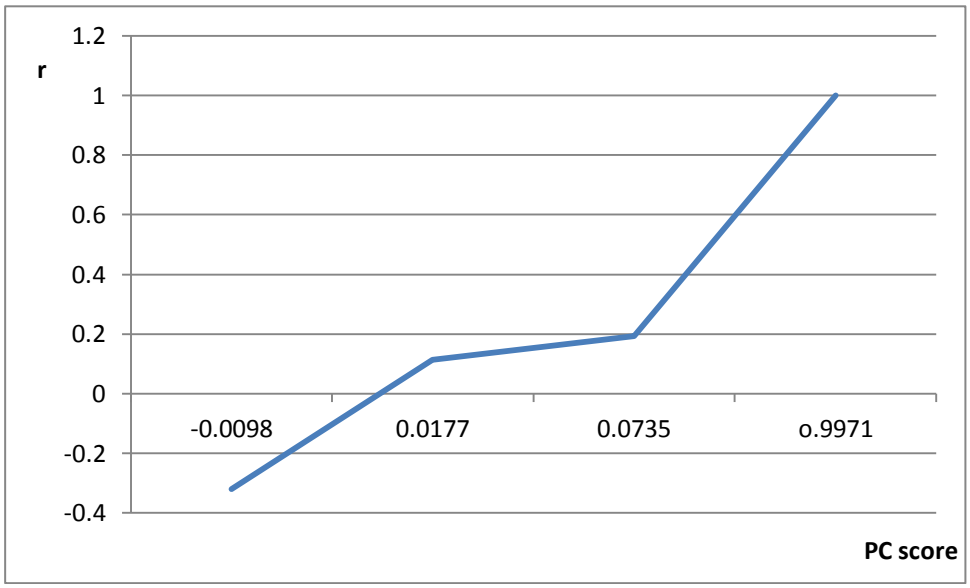

Figure 4.1: Relationship between principle component scores and correlation $r_{Y_1X_i}$ for the raw data.

Table4.3: Principle component scores and correlation between Y_2 and X_i for raw data

	X_1	X_2	X_3	X_4	X_5
e_{ii}	0.0048	0.0036	0.0735	-0.9973	0.0001
$r_{Y_iX_i}$	0.101	0.025	0.0471	-0.011	0.1

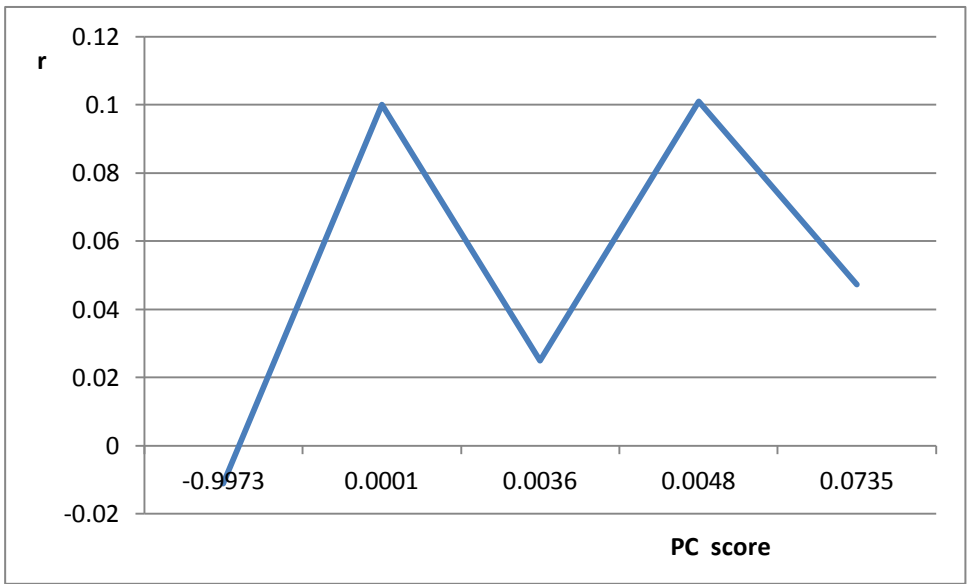

Figure 4.2: Relationship between principle component scores and correlation ($r_{Y_1X_i}$) for the raw data.

From Figures 4.1 and 4.2 it can be observed that in general the higher the contribution of a variable to the PC, leads to a higher linear correlation between that variable and the PC.

4.1.1 Principal components of centered data

In an $n \times p$ data matrix **X,** if the magnitude of the data values belonging to different variables is substantially different than each other, then the variables with bigger values will dominate the total variance. This will reflect on the coefficients of the PCs, leading to misinterpretations. The problem can be alleviated to a certain extent by centering the data matrix, before the computation of the PCs. Here centering means subtracting the mean of each variable $\bar{x}_j;\ j = 1,\ldots, p$ from the values of that variable. That is the expression of the elements of each variable as deviations from its mean $x_{ij} - \bar{x}_j;\ i = 1,\ldots, n;\ j = 1,\ldots, p$. To express this process in matrix form, let $\mathbf{H}_n$ be the centering matrix defined as $\mathbf{H} = \mathbf{H}_n = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$. Here $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{1}$ is the $n \times 1$ vector of 1s.

Then the centering matrix has the following properties [18].

i. It is symmetric and idempotent. $\mathbf{H} = \mathbf{H}^T = \mathbf{H}^{-1},\ \mathbf{H}^2 = \mathbf{H}$.

ii. $\mathbf{H1} = \mathbf{0},\ \mathbf{H1}\mathbf{1}^T = \mathbf{1}\mathbf{1}^T\mathbf{H} = \mathbf{0}$

iii. $\mathbf{Hx} = \mathbf{x} - \bar{x}\mathbf{1},$ where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

iv. Here, premultiplying a column vector by $\mathbf{H}$ results in the deviation values from the mean. If the data matrix $\mathbf{X}$ is premultiplied by the centering matrix, it yields the deviation of each element from its corresponding column mean.

v. $\mathbf{x}^T\mathbf{Hx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

Centering of the sample data matrix $\mathbf{X}$ is given by

$$\mathbf{X}_* = \frac{1}{\sqrt{n}}\mathbf{HXD}^{-1/2}$$

For clarity, such data matrix $\mathbf{X}_*$ will shortly be called centered data.

Table 4.4: Centered data obtained from raw data given in Table

x_1^*	x_2^*	x_3^*	x_4^*	x_5^*
-0.3171	0.0156	-.0416	-0.0246	0.0237
-0.0151	0.0156	0.2082	-0.0373	-0.2133
-0.0151	0.0156	-0.0416	-0.0499	0.0237
-0.0151	0.0156	-0.0416	-0.0499	-0.4503
0.2871	0.0156	-0.2915	-0.0625	0.2607
0.2871	0.0156	-0.0416	-0.0499	0.0237
0.2871	0.0156	-0.0416	-0.0499	0.4977
-0.3171	0.3169	0.2082	-0.0625	0.2607
-0.0151	0.3169	-0.2915	-0.0625	-0.2133
-0.0151	0.3169	-0.2915	-0.0373	0.2607
-0.0151	0.3169	0.5532	-0.0499	0.0237
0.2871	0.3169	0.2082	-0.0625	-0.2133
-0.3171	-0.2874	0.2082	-0.0625	0.2607
-0.0151	-0.2874	-0.2915	-0.0625	-0.2133
-0.0151	-0.2874	-0.2082	-0.0373	0.0237
-0.0151	-0.2874	-0.0416	-0.0752	0.0237
0.2871	-0.2874	-0.2915	-0.0373	-0.2133
0.2871	-0.2874	-0.0416	-0.0499	0.0237
-0.3171	0.0156	0.1963	0.9733	-0.2133
-0.3171	0.0156	-0.0416	-0.0499	0.0237

Covariance computed from centered data is:

$$\mathbf{S} = \begin{pmatrix} 0.0526 & -0.0046 & -0.0164 & -0.0173 & 0.0004 \\ -0.0046 & 0.0526 & 0.0101 & 0.0008 & 0.0034 \\ -0.0164 & 0.0101 & 0.0526 & 0.0106 & 0.0012 \\ -0.0173 & 0.0008 & 0.0106 & 0.0526 & -0.0117 \\ 0.0004 & 0.0034 & 0.0012 & -0.0117 & 0.0526 \end{pmatrix}$$

The eigenvalues of $\mathbf{S}$ are $\lambda_1 = 0.0855,\ \lambda_2 = 0.0623,\ \lambda_3 = 0.0471,\ \lambda_4 = 0.0366, \lambda_5 = 0.0317$ and the corresponding eigenvectors forms the columns of the Γ matrix. The elements of

each column of the Γ matrix are the coefficients or scores of the principal components $Y_i,\ i=1,\ldots,p$.

$$\mathbf{\Gamma} = \begin{pmatrix} 0.5841 & -0.0420 & -0.3709 & 0.2862 & -0.6614 \\ -0.2428 & 0.5382 & -0.7380 & -0.3257 & 0.0234 \\ -0.5339 & 0.2960 & 0.0461 & 0.7692 & -0.1834 \\ 0.5395 & 0.3845 & 0.0446 & -0.3823 & -0.6425 \\ 0.1538 & 0.6878 & 0.5600 & -0.2725 & -0.3398 \end{pmatrix}$$

Total variation in the data $\sum_{i=1}^{k}\lambda_k = 0.2632$. However, due to centering of the data there has been a considerable smoothing, leading to a more uniform distribution of the variation around the mean of each variable. This is visible from the closeness of the variances to each other. Never the less, the first 3 eigenvalues represents $\frac{\lambda_1+\lambda_2+\lambda_3}{\sum_{i=1}^{k}\lambda_k} = \frac{0.1949}{0.2632} \cong 0.74$ or 74% of the total variation of the centered data. But in general all PCs will have significant contribution in representing the centered data. PCs are given below.

$$Y_1 = 0.5841X_1 - 0.2428X_2 - 0.5339X_3 + 0.5395X_4 + 0.1538X_5$$

$$Y_2 = -0.0420X_1 + 0.5382X_2 + 0.2960X_3 + 0.3845X_4 + 0.6878X_5$$

$$Y_3 = -0.3709X_1 - 0.7380X_2 + 0.0461X_3 + 0.0446X_4 + 0.5600X_5$$

$$Y_4 = 0.2862X_1 - 0.3257X_2 + 0.7692X_3 - 0.3823X_4 - 0.2725X_5$$

$$Y_5 = -0.6614X_1 + 0.0234X_2 - 0.1834X_3 + -0.6425X_4 + -0.3398X_5 .$$

Inspection of the first PC Y_1 that represents 33% of total variation in the centered data, reveals that the variables X_1 and X_4 (charge rate and temperature) have the

highest positive influence on Y_1, while X_3 (Depth of discharge) has high negative influence. Similar interpretations can be made for the other PCs by close inspection of their principal component scores. Computed linear correlation coefficients between the first and second PCs, and constituent variables are presented in Tables 4.5, 4.6, and Figures 4.3 and 4.4.

Table 4.5: Principle component scores and correlation between Y_1 and X_i for centered data

	X_1	X_2	X_3	X_4	X_5
e_i	0.5841	-0.2428	-0.5339	0.5395	0.1538
$r_{Y_i X_i}$	0.745	-0.098	-0.6805	0.6877	0.1960

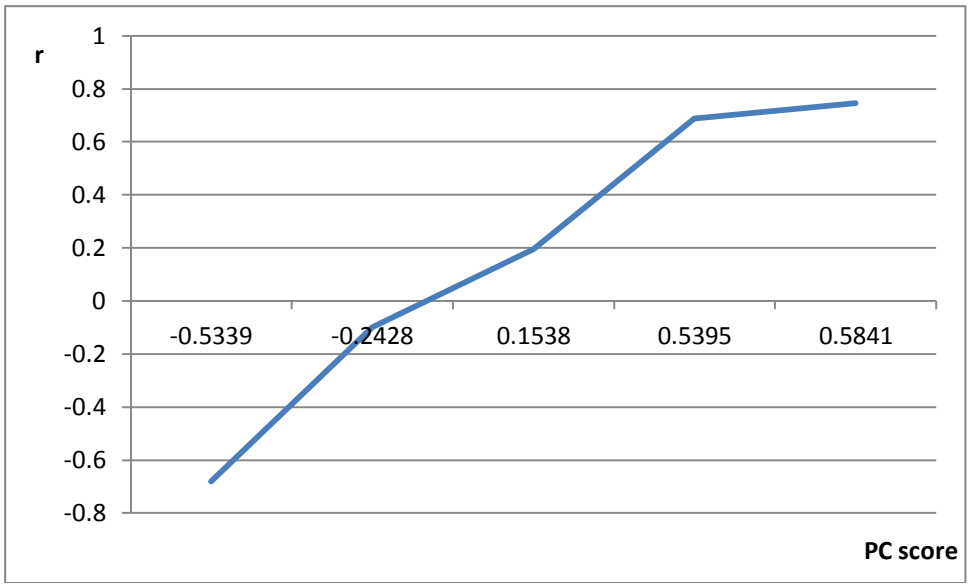

Figure 4.3: Relationship between principle component scores and correlation ($r_{Y_i X_i}$) for the centered data.

Table 4.6: Principle component scores and correlation between Y_2 and X_i for centered data

	X_1	X_2	X_3	X_4	X_5
e_i	-0.0420	0.5382	0.2960	0.3845	0.6878
$r_{Y_i X_i}$	-0.0497	0.5891	0.3240	0.4208	0.7521

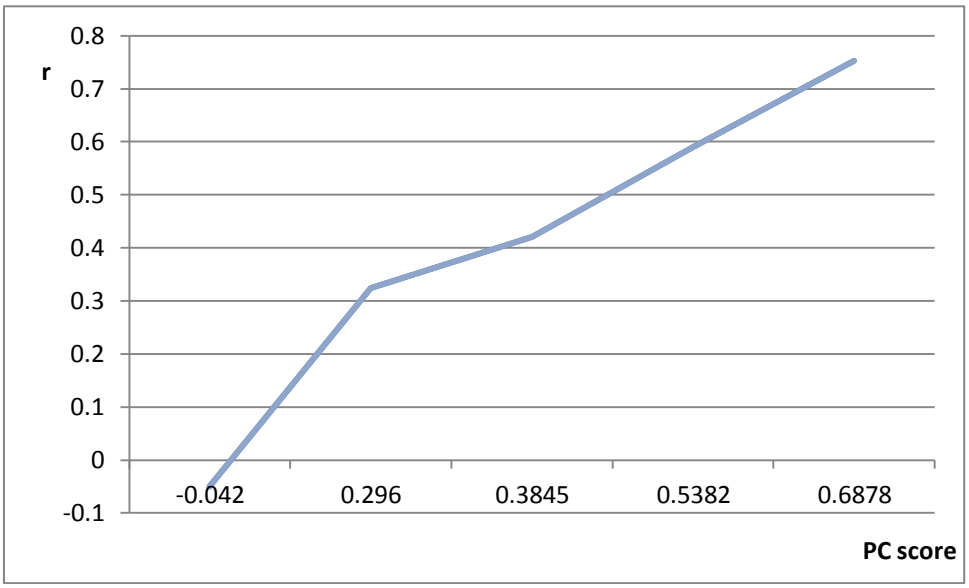

Figure 4.4: Relationship between principle component scores and correlation ($r_{Y_1X_i}$) for the centered data.

Here also the linear correlation between a PC and its constituent variables is compatible with the magnitude of the scores associated with that variable.

4.1.2 Principal components in the multivariate normal case

In the multivariate normal case the random vector **X** has parameters mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. From multivariate normal theory, it is known that the density of **X** is constant. $\boldsymbol{\mu}$ centered ellipsoid is given by

$$(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

with axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, $i = 1,\ldots,p$. Here λ_i and $\mathbf{e}_i$ are the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. Any point on the i^{th} axis has coordinates that are proportional to the vector $\mathbf{e}^T = (e_{i1},\ldots,e_{ip})$ in the coordinate system with origin $\boldsymbol{\mu}$ the i^{th} axis where the point is situated is parallel to the original axis $x_1,\ldots,x_p$.

Remember the facts that the distance from the point $\mathbf{x}^T = \left[x_1,\ldots,x_p\right]$ to the origin is given by the quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$. The square of the distance between $\boldsymbol{\mu}^T = [\mu_1,\ldots\mu_p]$ and any point $\mathbf{x}$ is $(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) = c^2$.

Without loss of generality $\boldsymbol{\mu} = \mathbf{0}$ can be assumed. If $\boldsymbol{\Sigma}^{-1}$ is substituted in place of $\mathbf{A}$ and from spectral decomposition concept

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2 \rightarrow \frac{1}{\lambda_1}(\mathbf{e}_1^T\mathbf{x})^2 + \cdots + \frac{1}{\lambda_p}(\mathbf{e}_p^T\mathbf{x})^2$$

can be written. Here $\mathbf{e}_1^T\mathbf{x},\ldots,\mathbf{e}_p^T\mathbf{x}$ are the PCs y_i, $i = 1,\ldots,p$. Then

$$c^2 = \frac{1}{\lambda_1}y_1^2 + \cdots + \frac{1}{\lambda_p}y_p^2 \qquad (4.9)$$

Since $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$, equation 4.9 represents the ellipsoid with axis $y_1,\ldots,y_p$ in the directions $\mathbf{e}_1,\ldots,\mathbf{e}_p$. The direction of the axes of a constant density ellipsoid is where the PCs lie in. Hence the $\mathbf{x}$ coordinates of any point on the i^{th} ellipsoid are proportional to $\mathbf{e}_i^T = [e_{i1},\ldots,e_{ip}]$. Principal component coordinates will be of the form $y_i = [0,\ldots,0,y_i,0,\ldots,0]$. If $\boldsymbol{\mu} \neq \mathbf{0}$, then the centered PC $y_i = \mathbf{e}_i^T(\mathbf{x}-\boldsymbol{\mu})$ will have $\mu_{y_i} = 0$ and lie in the direction $\mathbf{e}_i$.

Figure 4.5 shows the constant density ellipsoid of a bivariate normal distribution $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ with $\boldsymbol{\mu} = \mathbf{0} = \begin{bmatrix}0\\0\end{bmatrix}$ and ρ=0.75. PCs y_1, y_2 are can also be obtained by rotating the original coordinate axes by an amount equal to θ.

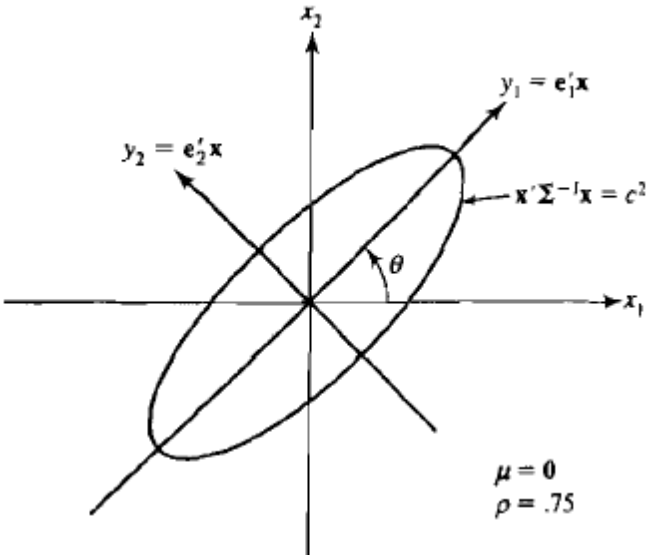

Figure 4.5: Constant density ellipsoid of a bivariate normal distribution

An attempt is made to apply the normal theory for the computation of PCs for the battery data assuming the variables are normally distributed. The overall mean for the whole data is used for the computation of the standardized values. These are given in Table 3.

Table 4.7: Data standardized using the global (overall) mean of the battery data.

z_1	z_2	z_3	z_4	z_5
-0.69104	-0.58103	1.689867	0.891239	-0.62615
-0.66609	-0.58103	2.360715	0.491925	-0.62655
-0.66609	-0.58103	1.689867	0.092611	-0.62615
-0.66609	-0.58103	1.689867	0.092611	-0.62695
-0.64113	-0.58103	1.01902	-0.3067	-0.62576
-0.64113	-0.58103	1.689867	0.092611	-0.62615
-0.64113	-0.58103	1.689867	0.092611	-0.62536
-0.69104	-0.50636	2.360715	-0.3067	-0.62576
-0.66609	-0.50636	1.01902	-0.3067	-0.62655
-0.66609	-0.50636	1.01902	0.491925	-0.62576
-0.66609	-0.50636	3.287123	0.092611	-0.62615
-0.64113	-0.50636	2.360715	-0.3067	-0.62655
-0.69104	-0.6561	2.360715	-0.3067	-0.62576
-0.66609	-0.6561	1.01902	-0.3067	-0.62655
-0.66609	-0.6561	2.360715	0.491925	-0.62615
-0.66609	-0.6561	1.689867	-0.70602	-0.62615
-0.64113	-0.6561	1.01902	0.491925	-0.62655
-0.64113	-0.6561	1.689867	0.092611	-0.62615
-0.69104	-0.58103	2.360715	0.491925	-0.62655
-0.69104	-0.58103	1.689867	0.092611	-0.62615

Covariance of the standardized data using the global mean is computed as

$$\mathbf{S} = \begin{pmatrix} 0.0004 & -0.0001 & -0.0037 & -0.0010 & -0.0000 \\ -0.0001 & 0.0032 & 0.0067 & -0.0001 & 0.0000 \\ -0.0037 & 0.0067 & 0.3813 & 0.0165 & 0.0000 \\ -0.0010 & -0.0001 & 0.0156 & 0.1590 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & -0.0000 & 0.0000 \end{pmatrix}$$

The eigenvalues of $\mathbf{S}$ are $\lambda_1 = 0.3827,\ \lambda_2 = 0.1578,\ \lambda_3 = 0.0031,\ \lambda_4 = 0.0003,\ \lambda_5 \cong 0.0001$.

Principal component scores matrix $\boldsymbol{\Gamma}$ made up of the corresponding eigenvectors is

$$\boldsymbol{\Gamma} = \begin{pmatrix} -0.0098 & 0.0048 & 0.0108 & 0.9999 & -0.0000 \\ 0.0177 & 0.0036 & -0.9998 & 0.0109 & -0.0004 \\ 0.9971 & 0.0735 & 0.0180 & 0.0092 & -0.0000 \\ 0.0735 & -0.9973 & -0.0022 & 0.0055 & 0.0001 \\ 0.0000 & 0.0001 & -0.0004 & 0.0000 & 1.0000 \end{pmatrix}$$

In this case total variation in the data $\sum_{i=1}^{k}\lambda_k = 0.544$. First 2 eigenvalues represents

$\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{k}\lambda_k} = \frac{0.5405}{0.544} \cong 0.9935$ or 99.35% of total variation in the data.

The first two PCs are

$Y_1 = -0.0098Z_1 + 0.0177Z_2 + 0.9971Z_3 + 0.0735Z_4 + 0.0001Z_5$

$Y_2 = 0.0048Z_1 + 0.0036Z_2 + 0.0735Z_3 - 0.9973Z_4 + 0.0001Z_5$.

It is seen that the use of global mean has resulted a significant reduction in the total variation as compared with the raw data which hase total variance $\sum_{i=1}^{k}\lambda_k = 341.11$. On the other hand, PCs for the raw data and standardized using the global mean are the same. However, total variance in the centered data $\sum_{i=1}^{k}\lambda_k = 0.2632$, is about half of the total variance of the standardized data. This is mainly due to the fact that, centering a data matix is based on column averages and standard deviations, which effectively results in greater smoothing of the data values.

Chapter 5

CONCLUSION

Principal component analysis is basically a method designed to transform high dimensional data using an orthogonal transformation. In the process a linear combination of the original variables is computed that forms a new set of independent variables. However, application of the method to any process may not result in a set of PCs that may not reflect the true picture of the original data. The following cases are examined.

When the the variables of the data set have similar scale. Application of the PCA under these conditions will help obtain PCs that are capable of explaining the overall variation without large deviations from the real variation in the data.

When the variables of the data set have different units, or data values of different variables have significant difference in terms of magnitute. In such cases either centering of the data, or standardizing based on global mean can be used.

Centering the data is carried out on the mean of individual variables. This in effect shifts the center of each variable to zero and standardize each variable accordingly. This process reduces the wide variation among variables. PCA is then performed and obtained PCs tends to explain the significance of each variable better than by direct application of PCA to raw data.

Standardizing the data using the global mean smooths the fluctuations in the variances of individual variables. Subsequent application of PCA to standardized data yields the same PCs as those obtained from raw data, indicating that each variable retained its initial significance.

A data set consisting of 5 variables that affects the failure of a battery was used to test the theory given in this thesis. All obtained results are consistent with the theory of PCA.

REFERENCES

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space," Philsophical,vol.2,no.6,1901.

[2] H. Hotelling, "analysis of complex of ststisical variables in to principle component, *Eductional Psychology*,no.24,pp.417-441,498-520,1933.

[3] H. Hotelling, "Simplified calculation of principle component," *Psychometrika* vol.1,pp.27-35,1936.

[4] M . A. Girshick, "On the sampling theory of roots of determinantal equations," *Annals of mathematical statistics*,vol.10,no.3,pp.203-224,1939.

[5] T. Anderson, "Asymptotic theory for principle component analysis," The *annals of Mathematical Statistics*,vol.34,no.1,pp.122-148,1963.

[6] C. Roa, "The use and iterpretation of principal component analysis in applied research," *Sankhia*,vol.A,no.26,pp.329-358,1964.

[7] J. C. Gower, "Some distance properties of latent root and vector methods used in," *Biometrika*,no.53,pp.352-38,1966.

[8] J. N . R. Jeffers, "Two case studies in the Application of Principle Component Analysis," *Journal of the Royal Statistical Society*.Series C (Applied

Statistics),vol.16,no.3,pp.225-236,1967.

[9] C. R. R. Rao & S.k.Mitra, "Generaliaed inverse of matrices and its application ," New York:John Wiley& Sons.p.240,1971.

[10] G. W. Stewart, "The economical storage of plane rotations," Numerische Mathematik,1976.

[11] M. L. Eaton, "Multivariate statistics, a vector space approach,".John Wily and Sons.pp.116-117,1983.

[12] G. Werner, "Linear algebra," 4th ed, Springer-verlag,New York,1975.

[13] R. A.Johnson& D. W. Wichern, "Applied multivariate statistical analysis," Chap.7.1,Oxford University press,New York,1998.

[14] D. C. Montgomery & G.C. Runger, "Applied statistics and propability for engineers," page 201.John Wily & Sons New York,1994.

[15] R. D. W . Wichern, "Applied Multivariate Statistical Analysis," 6th ed,pp.244,2002.

[16] A. C. Rencher, "Methods of multivariate analysis,"Brigham Young University, John Wily & Sons,2002.

[17] I. Jlliffe, "principal component analysis,"New York: Sipringar,Apr.2002.

[18] W. Hardel & L. S, "Applied multivariate statistical analysis,"pp.238,29thApril 2003.

[19] B. Dennis, "Matrix mathematics," Princeton University Press.p.44,2005.

[20] R. A. Johnson & D. W. Wichern, "Applied multivariate statistical analysis," Perntice Hall,6th edition,2007.

[21] B. C .Hall, "Quantum Theory for mathematicians ,"speinger,p.147,2013.

[22] R. M. Smith & T. G . Hurdley, 'The singular value decompostion is related to PCA," Journal of Chromatography Library.vol.57,1984.

[23] K. V. Mardia & J. T. Kent,et.al "Multivariate Analysis," Academic press,1979.

APPENDIX

Appendix: Computation to show independence of PCs y_1 and y_2 obtained from raw data.

$$\begin{aligned}
&\operatorname{cov}(Y_1, Y_2) = (-0.0098 \times 0.0048)\operatorname{var}(X_{11}) + (0.0177 \times 0.0735)\operatorname{var}(X_{22}) \\
&-(0.9971 \times 0.0735)\operatorname{var}(X_{33}) - (0.0753 \times 0.9973)\operatorname{var}(X_{44}) \\
&+0.00005\operatorname{cov}(X_1, X_2) - 0.0060\operatorname{cov}(X_1, X_3) \\
&+0.0101\operatorname{cov}(X_1, X_4) - 0.0001\operatorname{cov}(X_1, X_5) \\
&-0.0023\operatorname{cov}(X_2, X_3) - 0.0174\operatorname{cov}(X_2, X_4) \\
&+0.0001\operatorname{cov}(X_2, X_5) + 0.9998\operatorname{cov}(X_3, X_4) \\
&-0.0001\operatorname{cov}(X_3, X_5) + 0.0001\operatorname{cov}(X_4, X_5) \\
&= (-0.0098 \times 0.0048)(-0.0098) + (0.0177 \times 0.0735)(0.0036) \\
&-(0.9971 \times 0.0735)(0.0180) - (0.0753 \times 0.9973)(0.0055) \\
&+0.00005\operatorname{cov}(X_1, X_2) - 0.0060\operatorname{cov}(X_1, X_3) \\
&+0.0101\operatorname{cov}(X_1, X_4) - 0.0001\operatorname{cov}(X_1, X_5) \\
&-0.0023\operatorname{cov}(X_2, X_3) - 0.0174\operatorname{cov}(X_2, X_4) \\
&+0.0001\operatorname{cov}(X_2, X_5) + 0.9998\operatorname{cov}(X_3, X_4) \\
&-0.0001\operatorname{cov}(X_3, X_5) + 0.0001\operatorname{cov}(X_4, X_5) \\
&= (-0.0098 \times 0.0048)(-0.0098) + (0.0177 \times 0.0735)(0.0036) \\
&-(0.9971 \times 0.0735)(0.0180) - (0.0753 \times 0.9973)(0.0055) \\
&+0.00005(0.0048) - 0.0060(0.0108) \\
&+0.0101(0.9999) - 0.0001(-0.0000) \\
&-0.0023(-0.9998) - 0.0174(0.0109) \\
&+0.0001(-0.0004) + 0.9998(0.0092) \\
&-0.0001(-0.0000) + 0.0001(0.0001) \\
&\cong 0.0000005 + 0.0000005 - 0.001320 \\
&-0.000403 + 0.0000002 - 0.0000650 \\
&+0.0100990 + 0.00000001 + 0.0023 \\
&-0.0002 - 0.00000004 + 0.00919616 \\
&+0.00000001 + 0.00000001 \\
&\cong 0.00121603 \cong 0
\end{aligned}$$