Principal Component Analysis in Statistics

Ahmed Sami Abdulghafour Alani

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
January 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 Mahmadov
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.

Assist. Prof. Dr. Yücel Tandoğdu
Supervisor

Examining Committee

1. Prof. Dr. Agamirza Başirov ______________________
2. Assoc. Prof. Dr. Hüseyin Aktuğlu ______________________
3. Asst. Prof. Dr. Yücel Tandoğdu ______________________

ABSTRACT

Researchers and students sometimes need to deal with large volumes of data, causing them to have difficulty in the analysis and interpretation of these data. In the statistical analysis of high dimensional data, it is required to reduce the dimension of data set without losing any important information. One way of achieving this goal is the use the *principal component analysis* (PCA). The PCA objectives are to extract an important part of information from the data set, reducing the size of data with no damage to data and information. This is achieved by finding a new set of independent (uncorrelated) variables called principal components which are obtained as a linear combination of the original variables. The calculation of PCs means the computation of eigenvalues and eigenvectors for a positive-semidefinite symmetric matrix. The first PC has the largest proportion of variance of the data, and the second component has the second largest proportion of variance and is orthogonal to the first principal component. Remaining PCs represents the remainin variance in descending order, and each PC is orthogonal to its prdecesor. After computing the PCs, the first several PCs that represents the large part of variation are selected for use in further analysis. Finally, discussion of correlation between the PCs and original variables and determine which variable has more influence on each PC.

Keywords: Principal Component Analysis (PCA), orthogonal matrix, eigenvalue, eigenvector, singular value decomposition (SVD), covariance, correlation.

ÖZ

Araştırmacılar ve öğrenciler çalışmalarında büyük veri kitleleri ile çalışmak durumunda kalabilirler. Bu durum verilerin analizinde ve yorumunda güçlükler yaratabilir. Büyük boyutlu verilerin istatistiksel analizinde verideki önemli bilgileri kaybetmeden veri boyutu indirgemesi yapılması gereksinimi vardır. Bu amaca ulaşmanın yollarından bir taneside *temel bileşenler analizi* (TBA) dir. TBA'nın amacı verideki önemli bilgi içeriğini çıkarmak, veri boyutunu indirgerken veriye ve içerdiği bilgiye hasar vermemektir. Bu hedefe ulaşırken *temel bileşenler* (TB) denen, mevcut değişkenlerin lineer bir kominasyonu olan, birbirinden bağımsız yeni değişkenler tanımlanır. TB'lerin hesabında prensip olarak pozitif-yarıkesin simetrik bir matrisin özdeğer ve özvektörlerinin hesabı gerekir. Birinci TB verideki salınımın (varyasyonun) en büyük kısmını, ikinci TB birinciye orthogonal olub verideki salınımın ikinci en büyük kısmını temsil eder. Benzer şekilde geriye kalan TB'lerde azalan oranda salınımı temsil eder ve her biri kendinden önce gelene ortogonaldir. TB'lerin saptanmasından sonra, verideki salınımın büyük kısmını temsil eden ilk birkaç TB, daha ileri analiz ve yorumda kullanılmak üzere seçilir. TB'ler ile verideki değişkenler arasındaki ilişki ve hangi değişkenlerin TB üzerinde daha büyük etkisi olduğu incelenir.

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

I am dedicating this thesis to my family

ACKNOWLEDGMENTS

I would like to express my thanks and appreciation to my supervisor Asst. Prof. Dr. Yucel TANDOĞDU for his continuous support in my master research, for his patience and guidance, that helped me to write this thesis.

I want to thank all my professors in Mathematics Department at EMU, who contributed to the development of my mental abilities.

My heartfelt thanks to my lovely wife Eman for her support and her great patience to bear the responsibility of the family during my study period.

I would also like to thank all the staff of Al-Iraqia University in Iraq , and especially the Rector Prof. Dr. Ziad al-Ani, for giving me the opportunity to complete my studies for the Master, as well as, the employees in Scientific Affairs Department, for their assistance and continuous communication throughout the period of my studies, especially Miss Adhwaa.

My greatest appreciation goes to all my friends at EMU, specially Mohammed Khaled, Waleed Ghatee and Ghazwan Ahmed on their great support in my academic life in this this country.

TABLE OF CONTECTS

LIST OF FIGURS

LIST OF TABLES

LIST OF SYMBOLS /ABBREVIATIONS

$\mathbf{A}$	Capital bold letter represents a matrix
X	Capital letter represents a random variable
r.v.	Random variable
$\mathbf{x}$	Small bold letter represented to a vector
λ	Eigenvalue of matrix
SVD	Singular value decomposition
μ	Population Mean
$\bar{x}$	Sample Mean
p.d.f.	Probability distribution function
σ^2	Population Variance
σ	Population Standard deviation
s^2	Sample Variance
s	Standard deviation of a sample
$\mathbf{\Sigma}$	Population Covariance matrix
S	Sample Covariance matrix
PC	Principal Component
PCA	Principal Component Analysis
SLC	Standardized Linear Combination

Chapter 1

INTRODUCTION

At the beginning of a statistical study, the researchers often collect a set of data. When the data set and the variables involved are large, processing, analysis and interpretation becomes very demanding. Hence, the principal component analysis PCA method studied in this thesis provides an alternative by finding a set of linear combinations of the variables representing the data.

Initial foundations for PCA was defined by Karl Pearson (1901) [1], and it is now used in many scientific fields. PCA ingredients used to find the most influential variables of data (a combination form) and that illustrate a greater part of the variance in the data.

PCA is a technique used in statistical analysis to transform a large number of correlated variables to a smaller number of uncorrelated (orthogonal) components which is called *principal components,* while maintaining the important information of the original data, and this makes the data easier to understanding and representation.

In the third chapter some mathematical concepts which are important to understanding the PCA technique are introduced. Fourth chapter begins discussing the reduction of the dimensions geometrically, followed by the Mathematics of PCA and its properties are discussed. Third part of the chapter discusses the interpretation of PCA and the correlation between PCs and the original variables and the methods of how to choosing the number of PCs that provides the best explanation of population data. In the final part of chapter 4, a data set is used to highlight the theoretical concepts of PCA in application, as well as interpretation of the results.

Chapter 2

LITERATURE REVIEW

According to Jolliffe (2002) [2], the first description of the PCA was given by Karl Pearson in (1901). In his article ”On lines and planes of closest fit to systems of points in space,” [1], he also discussed the geometrical representation of the data and the best lines representing data. He concluded that “The best-fitting straight line to a system of points coincides in direction with the maximum axis of the correlation ellipsoid”. Also he pointed to the possibility of the using of analysis of several variables.

Jolliffe (2002), Hotelling (1933; 1936) and Girshick (1939) provided significant contributions to the development of PCA.

Hotelling (1933) started with the ideas of factor analysis, enabling the determination of a smaller set of uncorrelated variables which represent the original variables. He also chose the component which maximizes the total variances of original variables [3]. In a further study, Hotelling gave the accelerated version of power method for finding PCs [4].

Girshick (1939) illuminated the asymptotic variances and covariance of the coefficients of PCs [5].

Anderson (1963) discussed the PCA from the theoretical point of view [6]. However, the use of PCA remained limited until the development of computers. Parallel to the rapid developments in the computer hardware and software in 1960s resulted in a significant contribution to PCA.

Rao (1964) found new ideas for the use, techniques and interpretation of PCA [7]. Gower (1966) disscused the relation between the PCA and other statistical techniques [8]. Jeffers (1967) disscused the practical side of PCA through a practical application in two case studies of PCA [9].

Chapter 3

SOME MATIMATICAL AND STATISTICAL CONCEPTS

In this chapter some basic mathematical and statistical concepts that will be required to understand the Principal Components Analysis (PCA) and related topics in subsequent chapters are introduced.

3.1 Matrix Algebra Concepts

3.1.1 Eigenvalue and Eigenvector

In many statistical applications matrix algebra is widely used. Hence, some basic ideas on matrix algebra are given below to facilitate the understanding of the statistical methods introduced in the following chapters. Let **A** be any square matrix of size $n \times n$. If there exist a non-zero vector **x** and scalar λ such that

$$\mathbf{Ax} = \lambda \mathbf{x} \qquad (3.1)$$

then the vector **x** is called eigenvector of **A** corresponding to the eigenvalue λ [10].

3.1.2 Orthogonal Matrix

An $n \times n$ matrix **A** is called *orthogonal* if $\mathbf{A}^T \mathbf{A} = \mathbf{I}_n$.

3.1.3 Singular Value Decomposition (SVD)

Let **A** be a $m \times n$ matrix of real-values of data and with rank = *r*. The SVD of matrix **A** is the factorizing of **A** into the multiplication of three matrices.

$$\mathbf{A = UDQ^T} \tag{3.2}$$

where **U** is a $m \times m$ matrix with orthogonal columns. The columns of **U** are referred to the *left singular vectors* and ($\mathbf{U^TU = I}$), while **Q** is an $n \times n$ orthogonal matrix, the columns of **Q** (or rows of $\mathbf{Q}^T$) are referred to *the right singular vectors* $(\mathbf{Q^TQ = I})$, and **D** is a $m \times n$ rectangular diagonal matrix defined as

$$d(i, j) = \begin{cases} \delta_i & i = j \\ 0 & i \neq j \end{cases}$$

where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$, the values $d(i, j) = \delta_i$ in the main diagonal of **D** is known as the *singular values of* **A** [11].

3.1.4 Quadratic Form

Let **A** be $n \times n$ matrix. Then, the function $f(x) : \mathbb{R}^n \to \mathbb{R}$ definded by

$$f(x) = \mathbf{x^TAx}$$

is called the *quadratic form* of **A.**

3.2 Statistical Concepts

To understand the statistical concepts, suppose that a random sample is taken from population.

3.2.1 The Population Moment, Mean and Variance

Let *X* be a random variable with *p.d.f.* $f(x)$. The k^{th} moment about the origin of a r.v. *X*, denoted by μ_k, is the expected value of X^k;

$$\mu_k = E(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx \tag{3.3}$$

when *X* is continuous and

$$\mu_k = E(X^k) = \sum_x x^k f(x) \qquad k=0,\ 1,\ 2,\ 3\ldots \qquad (3.4)$$

when *X* is discret. The first moment when (*k*=1) $E(X) = \mu$ is called the population mean.

The k^{th} moment about the mean is called the central k^{th} moment of a random variable *X*, and is defined as the expected value of $(X-\mu)^k$ given by

$$E(X-\mu)^k = \int_{-\infty}^{\infty} (X-\mu)^k f(x)dx \qquad (3.5)$$

When *k=2*, we have the variance σ_X^2 and can also be expressed as

$$\sigma_X^2 = E(X-\mu)^2 = E(X^2) - (E(X))^2 \qquad (3.6)$$

The standard deviation σ is the value that gives information on how the values of the random variable are deviating from the population mean, and is given by the square root of the variance.

3.2.2 The Sample Moment, Mean and Variances

Assume we have a sequence of random samples $X_1, X_2, X_3, \ldots, X_p$, the *rth* sample moment for any *n* of random samples is given by

$$\bar{X}_p^r = \frac{1}{p}\sum_{i=1}^{p} X_i^r \quad p = 1,2,3,... \qquad (3.7)$$

The first sample moment is called the average and is defined by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad n = 1,2,3,... \qquad (3.8)$$

Each of random samples has numerical average value $\bar{x}_n$, which is defined by

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

where x_i is the observation value of X_i.

3.2.2(a) The Properties of Sample Moment

a) The expected value of $\bar{X}_n^r$

$$E[\bar{X}_n^r] = \frac{1}{n}E[\sum_{i=1}^{n} X_i^r] = \frac{1}{n}[\sum_{i=1}^{n} E(X_i^r)] = \frac{1}{n}\sum_{i=1}^{n}\mu_{i,r} \qquad (3.10)$$

If the r.vs $X_i;\ i = 1,\dots,n$ are identically distributed, then

$$E[\bar{X}_n^r] = \mu_r. \qquad (3.11)$$

In the case of *r=1* the expected value of $\bar{X}_n$ is the mean (μ).

b) The $Var(\bar{X}_n^r)$, where we have $X_1, X_2, ..., X_n$ samples

$$Var(\bar{X}_n^r) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i^r) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i^r). \qquad (3.12)$$

When samples are independent,

$$Var(\bar{X}_n^r) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i^r). \qquad (3.13)$$

If the samples are independent and identically distributed (*i.i.d.*), then

$$Var(\bar{X}_n^r) = \frac{1}{n} Var(X_r) \qquad (3.14)$$

when *r=1*

$$Var(\bar{X}_n) = \frac{1}{n} Var(X) = \frac{\sigma^2}{n} \qquad (3.15)$$

3.2.3 The Sample Variance

The sample variance of *n* random samples is denoted by s^2 and given by

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \qquad (3.16)$$

The expected value of sample variance is

$$E(s^2) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} EX_i^{\,2} - nE(\bar{X}^2)\right]$$

Since $E(X_i^2) = \sigma^2 + \mu^2$ then

$$E(s^2) = \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) \right] = \sigma^2 \qquad (3.17)$$

Hence s^2 is an unbiased estimator of σ^2. The purpose of division by $n-1$ in equation (3.16) is to ensure that $S^2$is an unbiased estimator for the varianceσ^2. Division by *n* instead of $n-1$ will introduce a negative bias methodically producing too-small estimator for σ^2

3.2.3(a) Properties of Variance and Covariance

$$Var(\mathbf{a}^T X) = \mathbf{a}^T Var(X)\mathbf{a} = \sum_{i,j} a_i a_j \sigma_{X_i X_j} \qquad (3.18)$$

$$Var(\mathbf{A}X + b) = \mathbf{A}Var(X)\mathbf{A}^T \qquad (3.19)$$

$$Var(X+Y) = Var(X) + Cov(X,Y) + Cov(Y,X) + Var(Y)$$

$$Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$$

$$Cov(\mathbf{A}X, \mathbf{B}Y) = \mathbf{A}Cov(X,Y)\mathbf{B}^T$$

3.2.4 Covariance

The covariance is a measurement tool between two random variables and is defined as

$$\mathrm{cov}(X_i, X_j) = \sigma_{X_i X_j} = E(X_i X_j) - E(X_i)E(X_j) \qquad (3.20)$$

Statistically the sample covariance is

$$\mathrm{cov}(X_i, X_j) = s_{X_i X_j} = \frac{\sum_{i,j=1}^{n} (X_i - \overline{X})(X_j - \overline{X})}{n-1} \qquad (3.21)$$

The covariance between a random variable X_i and itself is the variance $\sigma^2_{X_i}$ of the variable.

3.2.5 Covariance Matrix

If the r.v. X, is *p*-dimensional e.g., $X = \left(X_1 \quad X_2 \quad ... \quad X_p\right)^T$, then theoretical covariance among all elements is given by the covariance matrix $\mathbf{\Sigma}$.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{X_1X_1} & . & . & . & \sigma_{X_1X_p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma_{X_pX_1} & . & . & . & \sigma_{X_pX_p} \end{pmatrix}$$

and the covariance matrix of the sample is denoted by **S**

$$\mathbf{S} = \begin{pmatrix} s_{x_1x_1} & . & . & s_{x_1x_n} \\ . & . & . & . \\ . & . & . & . \\ s_{x_nx_1} & . & . & s_{x_nx_n} \end{pmatrix}$$

S is an unbiased estimator of $\mathbf{\Sigma}$. To show this, assume r.v.s

$X = \left(x_1, x_2, ..., x_n\right)^T$ and $X' = \left(x_1', x_2', ..., x_n'\right)^T$ are given

$$E(\mathbf{S}) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} x_i x_i' - n\overline{x}\overline{x}'\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(x_i x_i') - nE(\overline{x}\overline{x}')\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} (\mathbf{\Sigma} + \mu\mu') - n(V(\overline{x}) + E(\overline{x})E(\overline{x}'))\right)$$

$$= \frac{1}{n-1}\left(n\mathbf{\Sigma} + n\mu\mu' - n\frac{1}{n}\mathbf{\Sigma} - n\mu\mu'\right)$$

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

3.2.6 Correlation Coefficient

Correlation Coefficient ρ is a measure of the linear relationship between two random variables. $(-1 \le \rho \le 1)$

If the correlation between two variables is positive, then an increase (decrease) in the value of one variable corresponds to an increase (decrease) in the value of the other. Similarly a negative correlation would mean an increase (decrease) in the value of one variable will correspond to an decrease (increase) in the value of the other. The case of independence when there is no relation between two variables the correlation is zero. The correlation coefficient denoted by ρ , and is computed by (2.22) [12].

$$\rho_{XY} = corr(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \qquad (3.22)$$

Statistically

$$corr(X,Y) = r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (3.23)$$

To prove this formula, let *X* and *Y* be two random variables with bivariate normal distribution with joint probability density function

$$f(x,y) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \qquad (3.24)$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$ where $\sigma_X > 0$, $\sigma_Y > 0$ and $-1 < \rho < 1$. Consider a set of paired data $\left[(x_i, y_i) : i = 1, 2, ..., n\right]$ where x_i and y_i are values of r.v. from bivariate normal population with the parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and ρ. The estimation of these parameters require the likelihood function given by

$$L = \prod_{i=1}^{n} f(x_i, y_i) \qquad (3.25)$$

Maximization of L starts with differentiation of $\ln L$ with respected to $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and ρ. Equate the result to zero and then solve the system of equations for all parameters. Let us deal with $\frac{\partial \ln L}{\partial \mu_X}$ and $\frac{\partial \ln L}{\partial \mu_Y}$ equated to zero.

$$\frac{\partial \ln L}{\partial \mu_X} = -\frac{1}{2(1-\rho^2)}\left[-\frac{2\sum_{i=1}^{n}(x_i - \mu_X)}{\sigma_X^2} + \frac{2\rho\sum_{i=1}^{n}(y_i - \mu_Y)}{\sigma_X \sigma_Y} \right]$$

$$\frac{\partial \ln L}{\partial \mu_Y} = -\frac{1}{2(1-\rho^2)}\left[-\frac{2\rho\sum_{i=1}^{n}(x_i - \mu_X)}{\sigma_X \sigma_Y} + \frac{2\sum_{i=1}^{n}(y_i - \mu_Y)}{\sigma_Y^2} \right]$$

then

$$\frac{\partial \ln L}{\partial \mu_X} = -\frac{\sum_{i=1}^{n}(x_i - \mu_X)}{\sigma_X^2} + \frac{\rho\sum_{i=1}^{n}(y_i - \mu_Y)}{\sigma_X \sigma_Y} = 0$$

$$\frac{\partial \ln L}{\partial \mu_Y} = -\frac{\rho \sum_{i=1}^{n}(x_i - \mu_X)}{\sigma_X \sigma_Y} + \frac{\sum_{i=1}^{n}(y_i - \mu_Y)}{\sigma_Y^2} = 0$$

By solving this equation system for μ_X and μ_Y, the maximum likelihood estimates for these parameters are obtained as

$$\mu_X = \bar{x} \qquad \mu_Y = \bar{y}$$

Subsequently, by equating $\frac{\partial \ln L}{\partial \sigma_X}$, $\frac{\partial \ln L}{\partial \sigma_Y}$ and $\frac{\partial \ln L}{\partial \rho}$ to zero, substituting $\bar{x}$ and $\bar{y}$ in place of μ_X and μ_Y and Solving the system of equations

$$\hat{\sigma}_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \ , \ \hat{\sigma}_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}} \qquad (3.26)$$

are obtained.

3.2.7 Correlation Matrix

Let $X = (X_1, \ldots, X_n)^T$ be *n*-dimensional random sample, the correlation between r.vs X_i and X_j is denoted by $r_{x_i x_j}$ and given by

$$corr(X_i, X_j) = r_{x_i x_j} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)^2 \sum_{k=1}^{n}(x_{jk} - \overline{x}_j)^2}}$$

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

$$\mathbf{R} = \begin{pmatrix} r_{x_1 x_1} & . & . & r_{x_1 x_n} \\ . & . & . & . \\ . & . & . & . \\ r_{x_n x_1} & . & . & r_{x_n x_n} \end{pmatrix}$$

3.2.8 Relation Between the Correlation Matrix and Covariance Matrix

The correlation matrix **R** formula can be rewrite in algebra matrix

$$r_{X_i X_j} = corr(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)\,\text{var}(X_j)}}$$

$$= \frac{1}{\sqrt{\text{var}(X_i)}} \text{cov}(X_i, X_j) \frac{1}{\sqrt{\text{var}(X_j)}} \qquad (3.27)$$

Let **D** be a diagonal matrix such that the diagonal elements are the same as those of the covariance matrix **S** i.e. ($d_{ii} = s_{ii}$). From (3.27) the relation between the correlation matrix and the covariance matrix is given by (3.28) [13].

$$\mathbf{R = D^{-1/2}SD^{-1/2}} \qquad (3.28)$$

Chapter 4

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a technique used in statistics to facilitate the easy analysis of multivariate data. It works by extracting the important information from the data set and to expressing this information as a set of new orthogonal variables called *principal components* (*PCs*).

4.1 Geometry of Dimension Reduction

Assume that $\mathbf{X}(n \times p)$ is the data matrix composed of *p* variables and *n* observations. Each row $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, $i = 1, 2, 3, ..., n$ is a vector in *p-dimensional* space Figure 4.1.

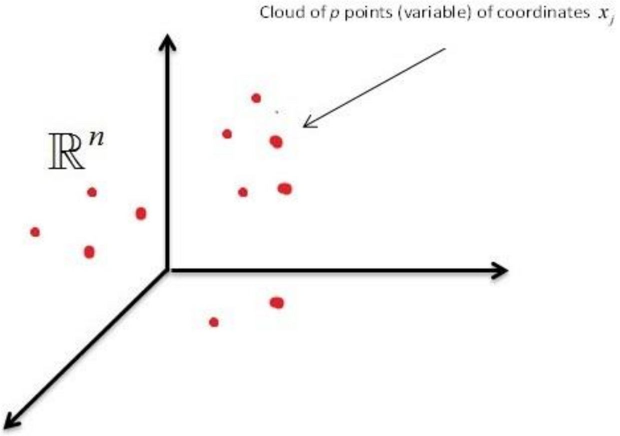


Figure 4-1: Cloud of *n* points (variable) in $\mathbb{R}^n$

Each column $\mathbf{x}_j = (x_{1j}, x_{2j}, ..., x_{nj})$ $j = 1, 2, 3, ..., p$ is a vector in *n-dimensional* space Figur 4.2.

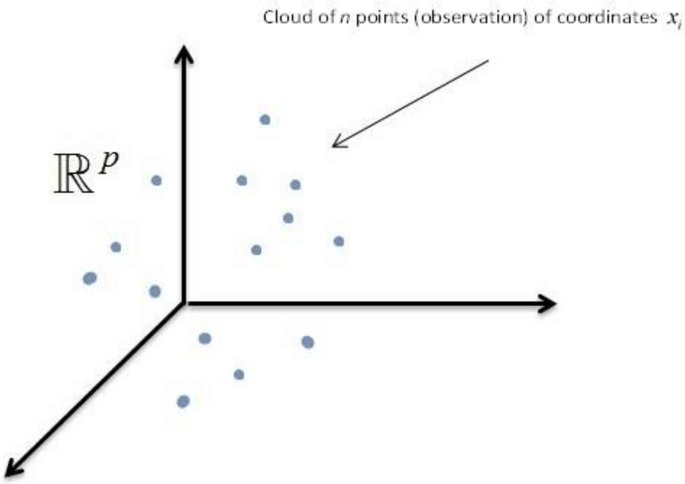


Figure 4-2: Cloud of *n* points (observation) in $\mathbb{R}^p$

4.1.1 Fitting *p-dimensional* Point (observation) Cloud

Let **X** be represented by *n-point (observation)* cloud in *p-dimensional* space. The question is how to reduce the cloud into *r-dimensional* subspace such that $r < p$. The simplest case when *r*=1, the problem is how to project the *n-point* cloud into one-dimensional subspace. Let L be the line of projection, it's direction is given by the unit vector $\mathbf{u} \in \mathbb{R}^p$. For any vector of points $\mathbf{x_i} \in \mathbb{R}^p$, let $\mathbf{x'_i}$ is the projection along the direction **u**. $\boldsymbol{\varepsilon}_\mathbf{i} = \mathbf{x_i} - \mathbf{x'_i}$ is the error vector (figure 4.3). The mean squared error (MSE) is given by [14].

$$\text{MSE }(\mathbf{u}) = \frac{1}{p}\sum_{i=1}^{p}\|\boldsymbol{\varepsilon}_\mathbf{i}\|^2 = \frac{1}{p}\sum_{\mathbf{i=1}}^{\mathbf{n}}\|\mathbf{x_i} - \mathbf{x'_i}\|^2$$

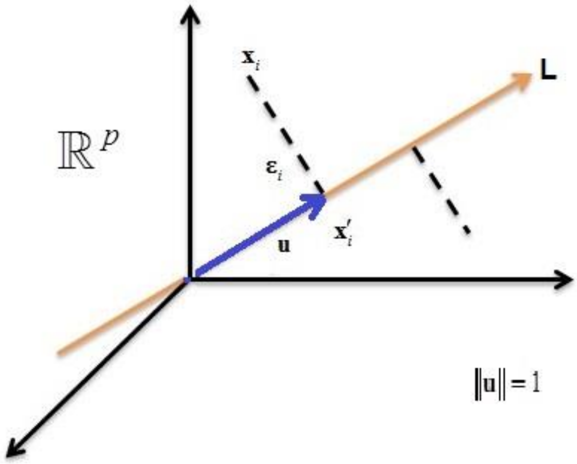


Figure 4-3: The projection of a point on the direction

The MSE (**u**) optimization is

$$\begin{aligned}
\text{MSE }(\mathbf{u}) &= \frac{1}{p}\sum_{i=1}^{p}\|\boldsymbol{\varepsilon}_\mathbf{i}\|^2 = \frac{1}{p}\sum_{\mathbf{i=1}}^{\mathbf{n}}\|\mathbf{x_i} - \mathbf{x'_i}\|^2 \\
&= \frac{1}{p}\sum_{i=1}^{p}(\mathbf{x_i} - \mathbf{x'_i})(\mathbf{x_i} - \mathbf{x'_i})^\mathbf{T} \\
&= \frac{1}{p}\sum_{i=1}^{p}(\|\mathbf{x_i}\|^2 - \mathbf{2x_i x'_i} - (\mathbf{x'_i})^\mathbf{T}\mathbf{x'}) \\
&= \frac{1}{p}\sum_{i=1}^{p}\left(\|\mathbf{x_i}\|^2 - 2(\mathbf{u^T x_i})(\mathbf{x_i^T u}) + \left((\mathbf{u^T x_i})\mathbf{u}\right)^\mathbf{T}(\mathbf{u^T x_i})\mathbf{u}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\left(\|\mathbf{x_i}\|^2 - 2\mathbf{x_i^T}(\mathbf{u^T x_i})\mathbf{u} + \left((\mathbf{u^T x_i})\mathbf{u}\right)^\mathbf{T}(\mathbf{u^T x_i})\mathbf{u}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\left(\|\mathbf{x_i}\|^2 - 2\mathbf{x_i^T}(\mathbf{u^T x_i})\mathbf{u} + (\mathbf{u^T x_i})(\mathbf{x_i^T u})\mathbf{u^T u}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\left(\|\mathbf{x_i}\|^2 - \mathbf{x_i^T}(\mathbf{u^T x_i})\mathbf{u}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\|\mathbf{x_i}\|^2 - \left(\frac{1}{p}\sum_{i=1}^{p}\mathbf{u^T}(\mathbf{x_i^T x_i})\mathbf{u}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\|\mathbf{x_i}\|^2 - \mathbf{u^T}(\frac{1}{p}\sum_{i=1}^{p}\mathbf{x_i^T x_i})\mathbf{u}
\end{aligned}$$

Details of how to reduce MSE (**u**) by finding $\mathbf{u} \in \mathbb{R}^p$ with $\|\mathbf{u}\| = 1$that maximizes $\mathbf{u^T} \mathbf{X}^T \mathbf{Xu}$ are given in Theorems (4.2) and (4.3).

Theorem 4.1: A $p \times p$ symmetric matrix **A** is orthogonally diagonalizable and can be written as

$$\mathbf{A = \Gamma\Lambda\Gamma^T} = \sum_{j=1}^{p} \lambda_j \boldsymbol{\eta}_\mathbf{j} \boldsymbol{\eta}_\mathbf{j}^\mathbf{T} \tag{4.1}$$

where $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_p)$, λ_i being the eigenvalues of **A**, and $\mathbf{\Gamma} = (\boldsymbol{\eta}_\mathbf{1}, \boldsymbol{\eta}_\mathbf{2}, ..., \boldsymbol{\eta}_\mathbf{p})$ is an orthogonal matrix of eigenvectors $\boldsymbol{\eta}_\mathbf{j}$ of **A** [15].

Theorem 4.2 (The Principal Axes Theorem): Let $\mathbf{A = \Gamma\Lambda\Gamma^T}$ be defined as in theorem 4.1, associated with the quadratic form $\mathbf{x^T Ax}$, then the change of variable $\mathbf{x = \Gamma y}$ transforms the quadratic form $\mathbf{x^T Ax}$ into the quadratic form $\mathbf{y^T \Lambda y}$ [16].

$$\begin{aligned} \mathbf{x^T Ax} &= (\mathbf{\Gamma y})^T \mathbf{A\Gamma y} \\ &= \mathbf{y}^T \mathbf{\Gamma}^T \mathbf{A\Gamma y} = \mathbf{y^T \Lambda y} = \lambda_1 y_1 + \lambda_2 y_2 + ... + \lambda_p y_p \end{aligned} \tag{4.2}$$

Theorem 4.3: Let $f(x) = \mathbf{x}^T \mathbf{Ax}$ be the quadratic form of the $p \times p$ symmetric matrix **A** and $\lambda_1 > \lambda_2 > ... > \lambda_p$ be the eigenvalues of **A.** Then the maximum value of $f(x)$ is λ_1. Hence, it occurs when **x** is a unit eigenvector corresponding to λ_1. Generally

$$\max_{\{x:\max \mathbf{xx}^T = 1\}} \mathbf{xAx}^T = \lambda_1 > \lambda_2 > ... > \lambda_p = \min_{\{x:\max \mathbf{xx}^T = 1\}} \mathbf{xAx}^T \ ,$$

The vector which maximizes (minimizes) $\mathbf{xAx}^T$ under the constraint $\mathbf{xx}^T = 1$ is the eigenvector of $\mathbf{A}$ which corresponds to the largest (smallest) eigenvalue of $\mathbf{A}$ [16].

Proof : By Theorem 4.1

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

By the Principal Axes theorem , set $\mathbf{x} = \mathbf{\Gamma y}$ then

$$\mathbf{y}^T\mathbf{y} = (\mathbf{\Gamma y})^T(\mathbf{\Gamma y}) = \mathbf{y}^T\mathbf{\Gamma}^T\mathbf{\Gamma y} = \mathbf{x}^T\mathbf{x} = 1$$

$$\begin{aligned} f(x) = \mathbf{x^T A x} = \mathbf{y^T \Lambda y} &= \lambda_1 y_1 + \lambda_2 y_2 + \ldots + \lambda_p y_p \\ &\leq \lambda_1 y_1 + \lambda_1 y_2 + \ldots + \lambda_1 y_p \\ &\leq \lambda_1 (y_1 + y_2 + \ldots + y_p) \\ &\leq \lambda_1 \mathbf{y}^T\mathbf{y} = \lambda_1 \end{aligned}$$

Thus, $f(x) \leq \lambda_1$ for all $\mathbf{x}$ with $\mathbf{x}^T\mathbf{x} = 1$. Let $\boldsymbol{\eta}_1$ be the eigenvector of $\mathbf{A}$

which corresponds to λ_1, then

$$\mathbf{A}\boldsymbol{\eta}_1 = \lambda_1 \boldsymbol{\eta}_1$$

Thus,

$$f(\boldsymbol{\eta}_\mathbf{1}) = \boldsymbol{\eta}_\mathbf{1}{}^\mathbf{T}\mathbf{A}\boldsymbol{\eta}_\mathbf{1} = \boldsymbol{\eta}_\mathbf{1}{}^\mathbf{T}\lambda_1\boldsymbol{\eta}_\mathbf{1} = \lambda_1\boldsymbol{\eta}_\mathbf{1}{}^\mathbf{T}\boldsymbol{\eta}_\mathbf{1} = \lambda_1$$

Hence, the vector $\mathbf{u}$ which maximizes $\mathbf{u^T X}^T\mathbf{Xu}$ is the eigenvector of $\mathbf{X}^T\mathbf{X}$ that corresponds to the largest eigenvalue.

The point cloud coordinates on a straight line are given by new factorial variable $\mathbf{z}_1$

$$\mathbf{z}_1 = \mathbf{Xu} \qquad (4.3)$$

This factor is a linear combination of the original variables $\left(x_{[1]}, x_{[2]}, \ldots, x_{[p]}\right)$, with coefficients represented by the vector **u**, i.e.

$$\mathbf{z}_1 = u_1 x_{[1]} + u_2 x_{[2]} + \ldots + u_p x_{[p]} \qquad (4.4)$$

In *2-dimensional* subspaces, the projection of a point cloud onto a plane is represented by best linear fitting of $\mathbf{u}_1$ and $\mathbf{u}_2$ ($\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal), i.e.

$$\max_{\mathbf{u}_1, \|\mathbf{u}_1\|=1} \mathbf{u}_1 \mathbf{X}^T \mathbf{X} \mathbf{u}_1 \text{ and } \max_{\substack{\mathbf{u}_2, \|\mathbf{u}_2\|=1 \\ \mathbf{u}_1^T \mathbf{u}_2 = 0}} \mathbf{u}_2 \mathbf{X}^T \mathbf{X} \mathbf{u}_2 \qquad (4.5)$$

Theorem 4.4: The second factorial axis $\mathbf{u}_2$, is the eigenvector of $\mathbf{X}^T\mathbf{X}$ corresponding to the second largest eigenvalue of $\mathbf{X}^T\mathbf{X}$ [17].

The representation of the *n*-point cloud in two-dimensional subspace is given by $\mathbf{z}_1$ and $\mathbf{z}_2$ figure (4.4) such that

$$\mathbf{z}_1 = \mathbf{X}\mathbf{u}_1 \text{ and } \mathbf{z}_2 = \mathbf{X}\mathbf{u}_2$$

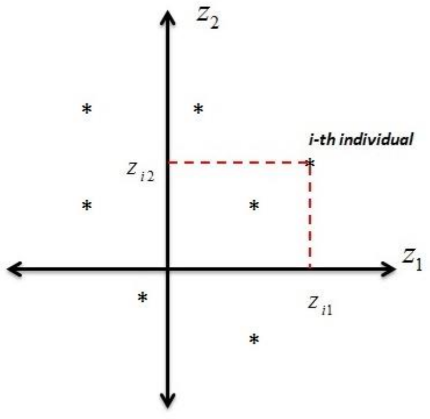


Figure 4-4 Representation of $x_1, x_2, \ldots, x_n$ individuals in 2-dimensional subspace

In *r*-dimensional sub space $2 < r < p$, the factor directions are $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_r$ which denote the eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponds to the largest eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_r$. The coordinates for representing the point cloud of individuals on *r-dimensional* subspace are given by $\mathbf{z}_1 = \mathbf{X}\mathbf{u}_1, \mathbf{z}_2 = \mathbf{X}\mathbf{u}_2, ...$ and $\mathbf{z}_r = \mathbf{X}\mathbf{u}_r$, $\mathbf{z}_r = (z_{1r}, z_{2r}, ..., z_{nr})^T$

$$z_{ir} = \sum_{m=1}^{p} x_{im} u_{mr} \qquad (4.6)$$

4.1.2 Fitting *n*-dimensional Point (variable) Cloud

Let **X** be represented by a *p* point (variable) cloud in *n-dimensional* space. The aim is to reduce the cloud into *q-dimensional* subspace such that $q < n$. Algebraically, this is the same case as *p-dimensional* point cloud (replace $\mathbf{X}$ by $\mathbf{X}^T$).

The representation of *p* variables in *q-dimensional* subspace is done by the same technique of the *n individuals*; the q-subspace is spanned by orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_q$ of $\mathbf{X}\mathbf{X}^T$ corresponding to the eigenvalues $\mu_1 > \mu_2 > ... > \mu_q$ respectively. Representation of the *p* variables on the k^{th} axis are given by the factorial variables

$$\mathbf{w}_k = \mathbf{X}^T \mathbf{v}_k \qquad k = 1, 2, ..., q \qquad (4.7)$$

where $\mathbf{w}_k = (w_{k1}, w_{k2}, ..., w_{kp})$

In 2 dimensional subspace the j^{th} variable is represented as in Figure 4.5.

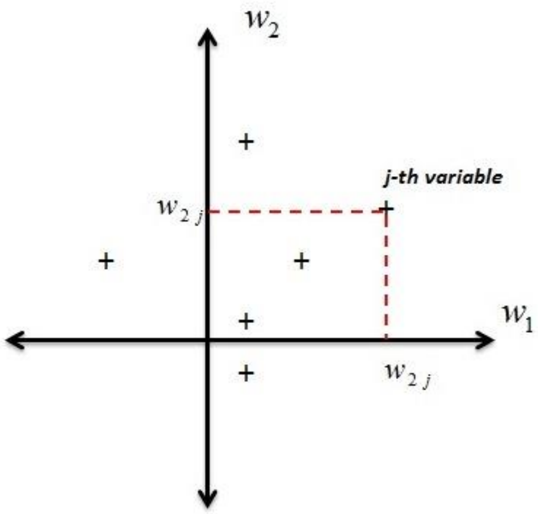


Figure 4-5: Representation of j^{th} variable in tow dimensional subspace

4.1.3 Subspaces Relationships

Illustration of the duality relationship between two models, requires the consideration of the equations of eigenvector in $\mathbb{R}^n$

$$\mathbf{X}\mathbf{X}^T\mathbf{v}_k = \mu_k\mathbf{v}_k \qquad k = 1, 2, \ldots, r \qquad (4.8)$$

where $r = rank(\mathbf{X}\mathbf{X}^T) = rank(\mathbf{X})$. Multiplying (4.8) by $\mathbf{X}^T$ we get

$$\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)v_k = \mathbf{X}^T\mu_k v_k$$

$$\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T v_k\right) = \mu_k\left(\mathbf{X}^T v_k\right) \qquad (4.9)$$

From (4.9), each eigenvector ($\mathbf{X}^T v_k$) of $\left(\mathbf{X}^T\mathbf{X}\right)$ is corresponding to an eigenvector v_k of $\mathbf{X}\mathbf{X}^T$.

Now consider the equations of eigenvectors in $\mathbb{R}^p$

$$\mathbf{X}^T\mathbf{X}\mathbf{u}_k = \lambda_k\mathbf{u}_k \qquad k = 1, 2, \ldots, r \qquad (4.10)$$

Multiplying (4.10) by **X**

$$\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{u}_k = \mathbf{X}\lambda_k\mathbf{u}_k$$

$$\left(\mathbf{X}\mathbf{X}^T\right)\left(\mathbf{X}\mathbf{u}_k\right) = \lambda_k\left(\mathbf{X}\mathbf{u}_k\right) \qquad (4.11)$$

Particularly, assume that $\mathbf{v}_k = \left(\mathbf{X}\mathbf{u}_k\right)$, by rewriting (4.11)

$$\mathbf{X}\mathbf{X}^T\mathbf{v}_k = \lambda_k\mathbf{v}_k \qquad (4.12)$$

This implies that the non-zero eigenvalues of $\mathbf{X}^T\mathbf{X}$ are eigenvalues of $\mathbf{X}\mathbf{X}^T$ as well.

The relation between the eigenvectors $\mathbf{v}_k$ and $\mathbf{u}_k$ is given in Theorem 4.5.

Theorem 4.5 : (Duality Relations) Let *r* be the rank of **X**. For *k* < *r*, the eigenvalues λ_k of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ are the same and the eigenvectors ($\mathbf{u}_k$ and $\mathbf{v}_k$ respectively) are related by

$$\mathbf{v}_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{X}\mathbf{u}_k \qquad (4.13)$$

$$\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{X}^T\mathbf{v}_k \qquad (4.14)$$

4.2 Mathematics of PCA

PCA is a procedure that seeks an *r-dimensional* basis that best captures the variance in the data. The vector that has the largest variance is called the first principal component. The orthogonal vector that captures the second largest variance is called the second principal component, and so on.

4.2.1 Data Pre-treatment

Prior to starting PCA procedure, data are often pre-treated to transform it into suitable form for analysis.

Variables frequently have different numerical units, and different range. For example when there are two variables, the first one being a persons' weight and the second variable is the height, the weight has large range so it has a large variance, but the height has small range, then it has small variance. Since PCA is a method of maximum variance projection, it follows that the variable which has large variance will contribute more than the variable with low-variance [18].

4.2.1(a) Unit Variance (UV) Scaling

In the data matrix each element of a column is divided by the column standard deviation, see figure (3.6) and figure (3.7).

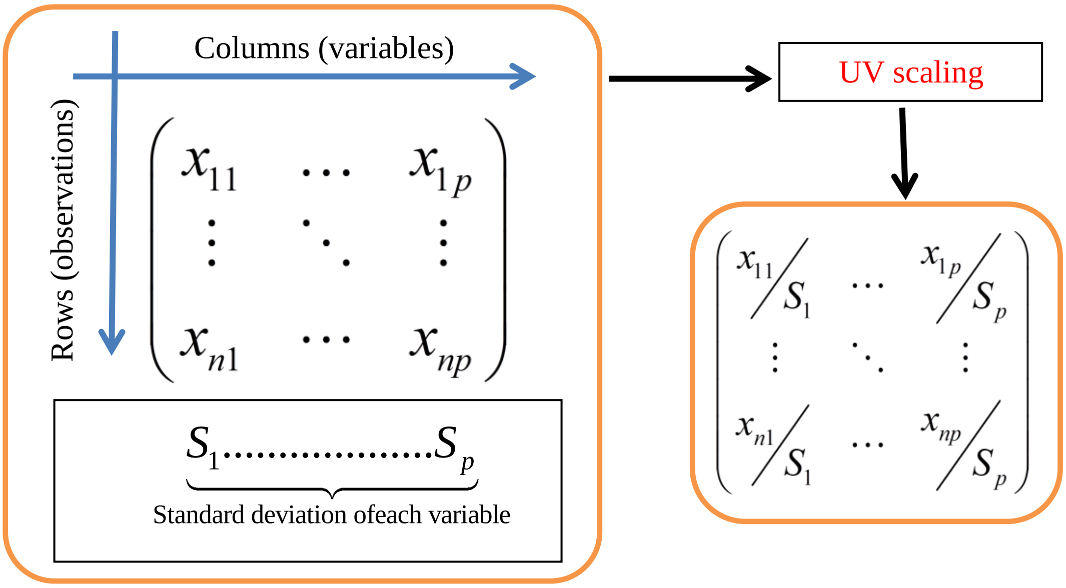


Figure 4-6 Unit Variance (UV) scaling processing

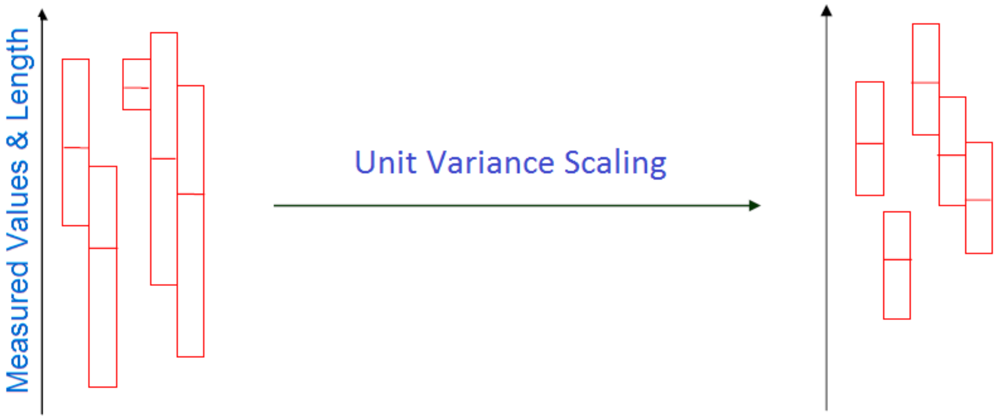


Figure 4-7 Unit Variance (UV) scaling

4.2.1(b) Mean-centering

The second method of pre-treatment of data is mean centering. In this process the mean of each scaled variable are computed and subtract from the UV scaled data.

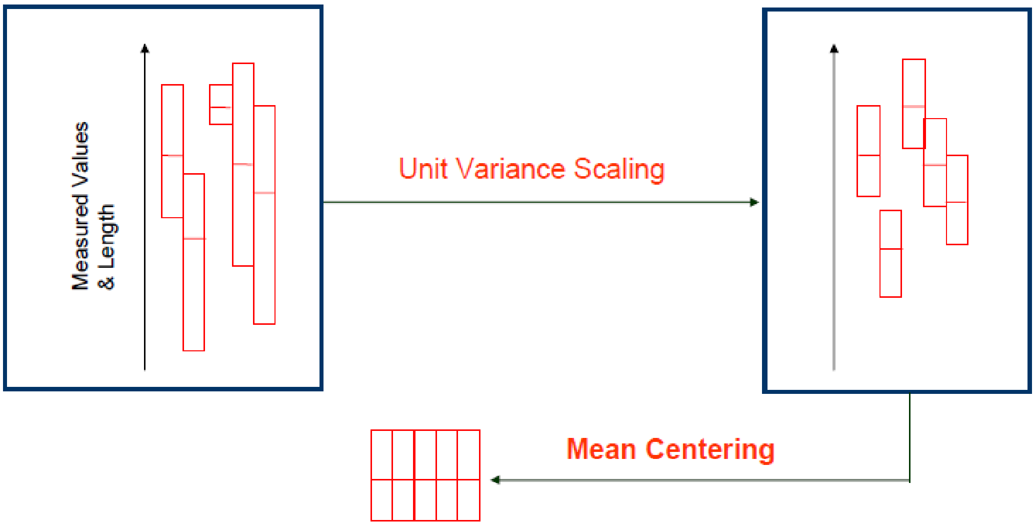


Figure 4-8 UV Scaling and Mean-centering

4.2.2 Centering a Data Matrix Algebraically

Let **X** be $n \times p$ data matrix (p variables and n observations). The “center of gravity" of the columns is a vector $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)$ in $\mathbb{R}^p$ of the means $\bar{x}_j$ of the p variables (columns) which given by:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ . \\ . \\ \bar{x}_p \end{pmatrix} = n^{-1} \mathbf{X}^T \mathbf{1}_n$$

where $\mathbf{1}_n$ is $n \times n$ unit matrix.

The covariance matrix **S** can be written as

$$\begin{aligned} \mathbf{S} = n^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathbf{T}} &= n^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X} - n^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{1}_n\mathbf{1}_n^T\mathbf{X}\right) \\ &= n^{-1}\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{X} \end{aligned}$$

Hence, $\left(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T\right)$ is a centering matrix denoted by **H.** Rewriting the covariance formula

$$\mathbf{S} = n^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{H}\mathbf{X} \qquad (4.15)$$

is obtained.

Note that $\mathbf{H}$ is symmetric and idempotent $(\mathbf{H} = \mathbf{H}^2)$. Then the standardized data matrix is denoted as $\mathbf{X}_*$ and given by

$$\mathbf{X}_* = n^{-1/2}\mathbf{H}\mathbf{X}\mathbf{D}^{-1/2} \qquad (4.16)$$

where $\mathbf{D} = diag(s_{X_iX_i})$

4.2.3 Relationship Between SVD and PCA

Let $\mathbf{X}_c$ be the centered matrix of $\mathbf{X}$ $n \times p$ data matrix. By (2.2) the SVD of $\mathbf{X}_c$ given as

$$\mathbf{X}_c = \mathbf{L}\mathbf{\Delta}\mathbf{Q}^T \qquad (4.17)$$

Now calculate the matrix $\mathbf{X}_c{}^T\mathbf{X}_c$

$$\begin{aligned}\mathbf{X}_c{}^T\mathbf{X}_c &= \left(\mathbf{L}\mathbf{\Delta}\mathbf{Q}^T\right)^T\left(\mathbf{L}\mathbf{\Delta}\mathbf{Q}^T\right)\\ &= \mathbf{Q}\mathbf{\Delta}^T\mathbf{L}^T\mathbf{L}\mathbf{\Delta}\mathbf{Q}^T\\ &= \mathbf{Q}\mathbf{\Delta}^T\mathbf{\Delta}\mathbf{Q}^T\\ &= \mathbf{Q}\mathbf{\Delta}^2_{\mathbf{X}_c}\mathbf{Q}^T \qquad (4.18)\end{aligned}$$

Where $\mathbf{\Delta}^2_{\mathbf{X}_c}$ is $n \times n$ matrix with diagonal entries δ_i^2 for $i = 1, 2, ..., p$ (3.2).

Since $\mathbf{X}_c$ is centered data matrix, the covariance matrix $\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}_c{}^T\mathbf{X}_c$ by theorem (4.2). This can be decomposed as $\mathbf{\Sigma} = \mathbf{U}^{\mathbf{T}}\mathbf{\Lambda}\mathbf{U}$ then

$$\begin{aligned} \mathbf{X}_c{}^T\mathbf{X}_c &= n\mathbf{\Sigma} \\ &= n\mathbf{U}^{\mathbf{T}}\mathbf{\Lambda}\mathbf{U} \\ &= \mathbf{U}^{\mathbf{T}}\left(n\mathbf{\Lambda}\right)\mathbf{U} \end{aligned} \qquad (4.19)$$

By (4.18) and (4.19), $\mathbf{Q}$ (*right singular vectors*) are the same of the eigenvectors of matrix $\mathbf{\Sigma}$, additionally, the singular values of $\mathbf{X}_c$ are related with the eigenvalue of $\mathbf{\Sigma}$.

$$\begin{aligned} n\lambda_i &= \delta_i^2 \\ \lambda_i &= \frac{\delta_i^2}{n} \qquad\qquad i = 1, 2, ..., p \end{aligned}$$

4.2.4 Standardized Linear Combinations (SLC)

A simple way to reducing dimension is to weigh all variables equally. This is undesirable, since all of the elements of vector **x** are measured with equal importance (weight). A more suitable approach is to study a weighted average, namely

Let $\mathbf{x} = (x_1, x_2, ..., x_p)^T$ be a vector, and $\delta = (\delta_1, \delta_2, ..., \delta_p)^T$ weighting vector. Then

$$\delta^T\mathbf{x} = \sum_{j=1}^{p}\delta_j x_j \quad \text{so that} \quad \sum_{i=1}^{p}\delta_i = 1 \qquad (4.20)$$

Equation (4.20) is called a standardized linear combination (SLC). The goal is to maximize the variance of the projection $\boldsymbol{\delta}^T \mathbf{x} = \sum_{j=1}^{p} \delta_j x_j$, i.e., to choose $\boldsymbol{\delta}$ such that

$$\max_{\{\delta:\|\delta\|=1\}} Var(\delta^T X) = \max_{\{\delta:\|\delta\|=1\}} \delta^T Var(X)\delta \qquad (4.21)$$

The weighting vector $\boldsymbol{\delta}$ in (4.21) is found through the spectral decomposition of the covariance matrix, by Theorems (4.2) and (4.3). The direction $\boldsymbol{\delta}$ is given by the eigenvector $\boldsymbol{\eta}_1$ of the covariance matrix $\boldsymbol{\Sigma} = Var(\mathbf{X})$ that corresponds to the largest eigenvalue λ_1.

The SLC with the maximum variance obtained from maximizing (4.21) is the first PC $\mathbf{y_1} = \boldsymbol{\eta}_1{}^{\mathbf{T}}\mathbf{X}$. In orthogonal direction to $\boldsymbol{\eta}_1$ we compute the SLC with the second highest variance $\mathbf{y_2} = \boldsymbol{\eta}_2{}^{\mathbf{T}}\mathbf{X}$, the second PC.

By processing in this way the result for r.v. X with $E(X) = \mu$ and $Var(X) = \boldsymbol{\Sigma} = \boldsymbol{\Gamma\Lambda\Gamma}^T$ the PC transformation can be defined as

$$Y = \boldsymbol{\Gamma}^T (X - \mu) \qquad (4.22)$$

The variable *X* was centered in order to obtain a PC variable *Y* with mean equal to zero.

The next numerical example explains how to calculate the PCs from covariance matrix.

Example (4.1): Let X_1, X_2 and X_3 be the r.vs. and **X** data matrix

$$\mathbf{X} = \begin{pmatrix} 125 & 137 & 121 \\ 144 & 173 & 147 \\ 105 & 119 & 125 \\ 154 & 149 & 128 \\ 137 & 139 & 109 \end{pmatrix}$$

The sample mean $\overline{\mathbf{x}}$ of **X** is $(133 \ 143.4 \ 126 \ )^T$

Covariance matrix $\mathbf{S}$ of $\mathbf{X}$ is

$$\mathbf{S} = \begin{pmatrix} 356.5 & 290 & 68.25 \\ 290 & 390.8 & 191 \\ 68.25 & 191 & 190 \end{pmatrix}$$

The ordered eigenvalues of $\mathbf{S}$ from the highest to the lowest are (729.3961, 183.8405, 24.0634) and the eigenvectors is the columns of next matrix corresponds to the eigenvalues respectively

$$\mathbf{\Gamma} = \begin{pmatrix} 0.6163 & -0.6355 & -0.4651 \\ 0.7146 & 0.2031 & 0.6694 \\ 0.3310 & 0.7449 & -0.5793 \end{pmatrix}.$$

Then the first eigenvector $\boldsymbol{\eta}_1$ which corresponding to the largest eigenvalue is the first column of $\mathbf{\Gamma}$

$$\boldsymbol{\eta}_1 = \begin{pmatrix} 0.6163 \\ 0.7146 \\ 0.3310 \end{pmatrix}$$

The PC transformation is

$$Y = \boldsymbol{\Gamma}^{\mathbf{T}}(X - \bar{\mathbf{X}})$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_1^{\mathbf{T}} \\ \boldsymbol{\eta}_2^{\mathbf{T}} \\ \boldsymbol{\eta}_3^{\mathbf{T}} \end{pmatrix} \begin{pmatrix} x_1 - 133 \\ x_2 - 143.4 \\ x_3 - 126 \end{pmatrix}$$

$$y_1 = 0.6163(x_1 - 133) + 0.7146(x_2 - 143.4) + 0.3310(x_3 - 126)$$

$$y_2 = -0.6355(x_1 - 133) + 0.2031(x_2 - 143.4) + 0.7449(x_3 - 126)$$

$$y_3 = -0.4651(x_1 - 133) + 0.6694(x_2 - 143.4) - 0.5793(x_3 - 126)$$

The first PC is y_1 which corresponds to the largest eigenvalue and the second PC is y_2 is orthogonal to y_1 and corresponds to second largest eigenvalue.

4.2.5 PCs in Practice

The PCs are obtained from the SVD of the covariance matrix. In the principal component transformation, the estimator μ is replaced by $\bar{x}$ and $\boldsymbol{\Sigma}$ is replaced by **S**. Spectral decomposition of the covariance matrix can be written as

$$\mathbf{S = GLG^T} \qquad (4.23)$$

Then the PCs are obtained by

$$Y = (X - \mathbf{1}_n \bar{x}^T)\mathbf{G} \qquad (4.24)$$

where $\mathbf{L} = diag(\ell_1, \ell_2, ..., \ell_p)$ is the diagonal matrix of eigenvalues of **S** and $\mathbf{G = (g_1, g_2, ..., g_p)}$ is a matrix of orthogonal eigenvectors $\mathbf{g_j}$ of **S**.

If all original *p* variables are uncorrelated (orthogonal, independent), then the variables themselves are the PCs. Hence **S** would have the form

$$\mathbf{S} = \begin{pmatrix} s_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{pp} \end{pmatrix}$$

and the eigenvalues ℓ_j of the covariance matrix **S** will be

$$\ell_j = s_{jj} \qquad j = 1, 2, ..., p.$$

Correspondingly the normalized eigenvectors $\mathbf{g}_j$ which have 1 in j^{th} position and zeros else where are

$$\mathbf{g}_j^T = (0, 0, ..., 1, 0, ..., 0) \qquad j = 1, 2, ..., p$$

Thus the j^{th} PC is

$$\mathbf{z}_j = \mathbf{g}_j^T \mathbf{X} = \mathbf{x}_j \qquad j = 1, 2, ..., p$$

As another illustration, in the covariance **S** or correlation matrix **R**, a distinguishing pattern may be identified, from which formulation of the principal components can be deduced. For example, if one of the variables has the highest variance compared with others, this variable will dominate the first component, accounting for the majority of the variance.

Generally, the PCs are computed from **S** rather than **R**, specially if the PCs are used in farther computation. However, in some cases, the PCs will be more interpretable if calculated from **R** [19].

After centering the data matrix $\mathbf{X}_c = \mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^T$, $\mathbf{X}_c^T \mathbf{X}_c$ is the covariance matrix which is used in PCA. When the variables are measured with different unit, the data must be standardized by dividing each variable (each column) by column standard deviation (4.15) (Figure 4.6). In this case $\mathbf{X}_*^T \mathbf{X}_*$ is equal to correlation matrix **R**. Then the analysis referred to correlation PCA [19].

The next simple bivariate example explains how the principal components are changed when computed from original data, centered data and standardized data.

Example 4.2: Dtat given in Table 1 represents the number of engineers in various disciplines with monthly salary, years of experience and working hours

$$X_1 = \text{Experience}(\textit{in years})$$
$$X_2 = \text{Salary}$$
$$X_3 = \text{Work hours}$$

Table 1 Engineering salary

Engineering competence	Experience (years)	Salary (IRD per month)	Work hour (hours/ day)
CAE Analyst	10	900,000	6
Design Engineer	10	900,000	5
Purchase Engineer	11	850,000	8
SCM Enigneer	8	850,000	7
Quality Engineer	11	850,000	5
Production Engineer	9	750,000	9
Maintenance Engineer	12	750,000	6
Mechatronics Engineer	10	800,000	8
OEM Sales Engnineer	9	950,000	7
Engineer	12	800,000	5
Application Engineer	10	800,000	9
Service Engineer	13	600,000	6
Homologation Engineer	10	850,000	9
Management	8	800,000	7
Electronics & Comunication	11	800,000	5
Lead final Assembly Line	11	800,000	8
RAMS Engineers Electrical	10	700,000	6
Structural Design Engineers	9	600,000	7
Configuration Engineers	10	600,000	7
Aerospace Stress Engineer	12	550,000	8

PCs from raw, centered and the standardized data matrices are computed for comparison.

The eigenvalues and eigenvector of $\mathbf{X}^T\mathbf{X}$ are

$$\begin{array}{ll} \lambda_1 = 1.0100 & \mathbf{v}_1^T = (0.0124 \quad 0.9999 \quad 0.0082) \\ \lambda_2 = 0.00007 & \mathbf{v}_2^T = (0.9870 \quad -0.0135 \quad 0.1603) \\ \lambda_3 = 0.000012 & \mathbf{v}_3^T = (0.1604 \quad 0.0062 \quad -0.9870) \end{array}$$

Thus the PCs are

$$\mathbf{y}_{1\mathrm{X}} = 0.0124(X_1 - \overline{x}_1) + 0.9999(X_2 - \overline{x}_2) + 0.0082(X_3 - \overline{x}_3)$$

$$\mathbf{y}_{2\mathrm{X}} = 0.9870(X_1 - \overline{x}_1) - 0.0135(X_2 - \overline{x}_2) + 0.1603(X_3 - \overline{x}_3)$$

$$\mathbf{y}_{3\mathrm{X}} = 0.1604(X_1 - \overline{x}_1) + 0.0062(X_2 - \overline{x}_2) - 0.9870(X_3 - \overline{x}_3)$$

The eigenvalues and eigenvector of the covariance matrix for the centered data $\mathbf{S} = \mathbf{X}_c{}^T\mathbf{X}_c$ are

$$\begin{array}{llrrl} \lambda_1 = 9.3341 & \mathbf{v}_1^T = (-0.0093 & 1.0000 & -0.0005) \\ \lambda_2 = 0.0041 & \mathbf{v}_2^T = (0.5533 & 0.0047 & -0.8329) \\ \lambda_3 = 0.0012 & \mathbf{v}_3^T = (-0.8329 & -0.0080 & -0.5534) \end{array}$$

the PCs are

$$\mathbf{Y}_{1\mathrm{S}} = -0.0093X_1 + 1.0000X_2 - 0.0005\mathrm{X}_3$$

$$\mathbf{Y}_{2\mathrm{S}} = 0.5533X_1 + 0.0047X_2 - 0.8329\mathrm{X}_3$$

$$\mathbf{Y}_{3\mathrm{S}} = -0.8329X_1 - 0.0080X_2 - 0.5534\mathrm{X}_3$$

Eigenvalues and eigenvectors of the correlation matrix after standardizing data ($\mathbf{X}_*$) $\mathbf{R} = \mathbf{X}_*{}^T\mathbf{X}_*$ are

$$\begin{array}{llrrl} \lambda_1 = 1.6673 & \mathbf{v}_1^T = (0.7144 & -0.5457 & -0.4380) \\ \lambda_2 = 1.0282 & \mathbf{v}_2^T = (-0.0084 & -0.6326 & 0.7744) \\ \lambda_3 = 0.3046 & \mathbf{v}_3^T = (-0.6997 & -0.5495 & -0.4565) \end{array}$$

the PCs in third case are

$$\mathbf{Y_{1R}} = 0.7144X_1 - 0.5457X_2 - 0.4380X_3$$

$$\mathbf{Y_{2R}} = -0.0084X_1 - 0.6326X_2 + 0.7744X_3$$

$$\mathbf{Y_{3R}} = -0.6997X_1 - 0.5495X_2 - 0.4565X_3$$

4.2.6 Mean and Variance of PCs

Let $X \sim (\mu, \mathbf{\Sigma})$, $\mathbf{\Sigma} = \mathbf{\Gamma^T \Lambda \Gamma}$ and $Y = \mathbf{\Gamma}^T(X-\mu)$be a linear transformation then the following properties apllies

a) $EY_j = 0$ $\qquad j = 1, 2, ..., p$

$$EY_j = E(\boldsymbol{\eta}_j^T(X-\mu)) = \boldsymbol{\eta}_j^T E(X-\mu) = 0$$

b) $Var(Y_j) = \lambda_j$ $\qquad j = 1, 2, ..., p$

$$Var(Y_j) = Var(\boldsymbol{\eta}_j^T(X-\mu)) \text{ by (3.18) and (3.19)}$$

$$= \boldsymbol{\eta}_j^T Var(X)\eta_j = \lambda_j$$

c) $Cov(Y_i, Y_j) = 0$ $\qquad i \neq j$

$$Cov(Y_i, Y_j) = E(Y_i Y_j) - E(Y_i)E(Y_j) = 0$$

d) Let **S** be the covariance matrix of original variables, and let $\mathbf{Y} = (\mathbf{X} - \mathbf{1}_n \bar{x}^T)\mathbf{\Gamma}$

The covariance matrix of the PCs is

$$\mathbf{S}_Y = \mathbf{\Lambda}$$

where $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ is the eigenvalues of **S,** by (4.1)

$$\begin{aligned}\mathbf{S}_Y &= n^{-1}\mathbf{Y}^T\mathbf{HY} = n^{-1}((\mathbf{X}-\mathbf{1}_n\overline{x}^T)\mathbf{\Gamma})^T\mathbf{H}(\mathbf{X}-\mathbf{1}_n\overline{x}^T)\mathbf{\Gamma} = n^{-1}\mathbf{\Gamma}^T\mathbf{X}^T\mathbf{HX\Gamma}\\ &= \mathbf{\Gamma}^T\mathbf{S\Gamma} = \mathbf{\Lambda}\end{aligned}$$

4.3 Interpreting the Meaning of the PC

PCA produce two items of basic information for interpreting results. First one is the correlation coefficients between the original variables and the PCs which are used in interpreting the meaning of the PCs. The second one is each principal component is associated with an eigenvalue which converts to the proportion of the variation that explained by the PC.

4.3.1 Loading: Correlation Between the r.v. *X* and its PC

The covariance between the original r.v X and the PC Y is given in [2] as

$$\begin{aligned}C\text{ov}(X,Y) &= E(XY^T) - E(X)E(Y^T) = E(XY^T)\\ &= E(XX^T\mathbf{\Gamma}) - \boldsymbol{\mu\mu}^T\mathbf{\Gamma}\\ &= Var(X)\mathbf{\Gamma}\\ &= \mathbf{\Sigma\Gamma}\\ &= \mathbf{\Gamma\Lambda\Gamma}^T\mathbf{\Gamma} = \mathbf{\Gamma\Lambda}\end{aligned} \qquad (4.25)$$

where the covariance matrix $\mathbf{\Sigma} = \mathbf{\Gamma\Lambda\Gamma}^T$ and $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ is the eigenvalues and $\mathbf{\Gamma} = (\boldsymbol{\eta}_\mathbf{1}, \boldsymbol{\eta}_\mathbf{2}, ..., \boldsymbol{\eta}_\mathbf{p})$. This is a matrix of orthogonal eigenvectors $\boldsymbol{\eta}_\mathbf{j}$ of the covariance matrix.

The correlation between each PC and the original variables is denoted by $\rho_{X_iY_j}$ and given by

$$\rho_{X_iY_j} = \frac{\eta_{ij}\lambda_i}{\left(\sigma_{X_iX_i}\lambda_j\right)^{1/2}} = \eta_{ij}\left(\frac{\lambda_i}{\sigma_{X_iX_i}}\right)^{1/2} \qquad \begin{array}{l} i = 1, 2, ..., p \\ j = 1, 2, ..., q \end{array} \tag{4.26}$$

Using actual data, (4.26) translates to

$$r_{X_iY_j} = g_{ij}\left(\frac{\ell_j}{s_{X_iX_i}}\right)^{1/2} \tag{4.27}$$

This correlation coefficient between the r.v *X* and PC is also called "*loading*". Note that sum of squares of loadings is equal to 1.

$$\sum_{j=1}^{p} r_{X_iY_j} = \frac{\sum_{j=1}^{p}\ell_j g_{ij}^2}{s_{X_iX_i}} = \frac{s_{X_iX_i}}{s_{X_iX_i}} = 1 \tag{4.28}$$

4.3.2 Number of PCs to be used

Usually, only the important information is required to be drawn from a data matrix. In this case, the problem is to find how many components are needed to be considered. There are many methods to decide on the number of PCs. Four of them are given below.

4.3.2(a) Scree Plot Test

The Cattell scree test (Cattell, 1966) is based on a graphical representation of the eigenvalues. In this method, the eigenvalues are presented in descending order with corresponding PCs in a scatter plot and drawing the curve. Cattell's scree rule says to drop all PCs after the elbow point. The logic behind this test is that the elbow point divides the major or important PCs (factors) from the trivial or minor PCs (factors). This rule is criticized because of the elbow point selection is subjective and depends on the researcher [20].

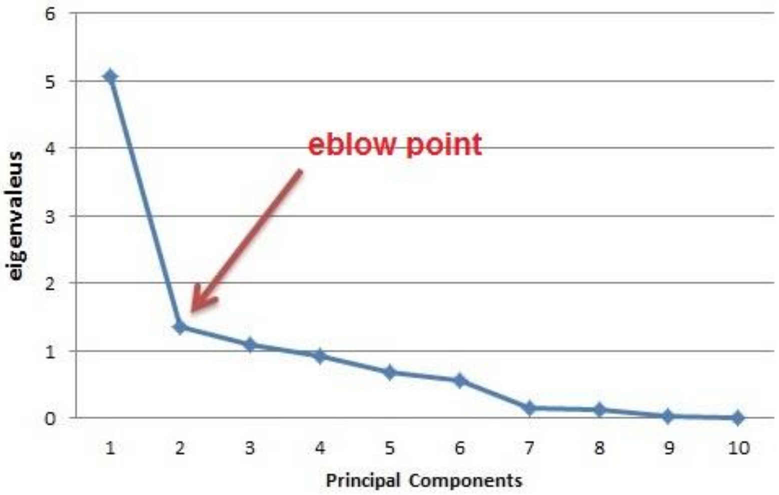


Figure 4-9 Scree plot test

4.3.2(b) Kaiser Criterion

This method is proposed by Kaiser (1960), it’s rule says only the PCs That corresponding to the eigenvalues which are greater than 1 are retained for interpretation [21]. Despite the ease of this method, it carries many weaknesses. One such weakness is in the selection of PCs that do not satisfy the majority of the variance. For instance, it regards a PC with an eigenvalue of 1.01 as ‘major’ and one with an eigenvalue of .99 as ‘trivial’ which is not a very healthy decision.

4.3.2(c) Horn's Parallel Analysis (PA)

This technique based on a simulation method that make a comparison between the observed eigenvalues with those obtained from orthogonal normal variables. A PC is maintained if the corresponding eigenvalue is greater than the 95th of the distribution of eigenvalues derived from the random data [22].

The algorithem of Horn's Parallel Analysis (PA) can be explained as below.

Step 1: Generation of a Random Data

i. Setting up the number of observations and variables in the original data;

ii. Setting up the values taken by original data set (e.g. Likert scale 1-5);

iii. Create a random data set by using SPSS or similar program.

Step 2:Computing Eigenvalues from the Random Data Correlation Matrix

i. Computing the eigenvalues from the random data set, either by a PCA using the SPSS, or any equivalent program;

ii. Note the eigenvalues sequentially in MS Excel or similar software

iii. Repeat Step 1 (iii) and Step 2(i)-(ii) for at least 50 times to create a set of 50 or more parallel eigenvalues.

Step 3: Average Eigenvalues

i. Find the mean, and 95th percentile of all eigenvalues generated by PCA of random data sets;

ii. The result will be a vector of average (and 95th percentile) of eigenvalues. The number of eigenvalues is the same as the number of variables, and in decreasing order.

Step 4: Compare Real Data with Parallel Random Data:

i. Plot eigenvalues from the real and random data sets

ii. Retain only those factors whose eigenvalues are greater than the eigenvalues from the random data.

4.3.2(d) Variance Explained Criteria

The proportion of variance of each PC is calculated by

$$\frac{Var(Y_i)}{\sum_{j=1}^{p} Var(Y_j)} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \qquad (4.29)$$

Let φ_q be the proportion of the sum of first q eigenvalues to $\sum_{j=1}^{p}\lambda_j$

$$\varphi_q = \frac{\sum_{j=1}^{q} Var(Y_j)}{\sum_{j=1}^{p} Var(Y_j)} = \frac{\sum_{j=1}^{q}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \qquad (4.30)$$

Then the number of PCs to be considered are expected to satisfy above 70% of the total variation $\sum_{j=1}^{p}\lambda_j$.

4.3.3 Rotation

Most of the foundations of rotation are developed by Thurstone (1947) and Cattell (1978), who defends the use of rotation to make interpretation of PCs easier and more reliable [23].

After the number of PCs has been selected, an attempt is made to facilitate interpretation and the analysis often based on a rotation of the selected PCs. There are two main kinds of rotation, orthogonal and oblique rotation.

4.3.3(a) Orthogonal Rotation

An orthogonal rotation method is described by a rotation matrix **R**, where the rows represents the original factors and the columns represents the new (rotated) factors. At the intersection of row *i* and column *j* we have the cosine of the angle θ between the original axis and the new axis.

$$\mathbf{R} = \begin{bmatrix} \cos\theta_{1,1} & \cos\theta_{1,2} \\ \cos\theta_{2,1} & \cos\theta_{2,2} \end{bmatrix} = \begin{bmatrix} \cos\theta_{1,1} & -\sin\theta_{1,1} \\ \sin\theta_{1,1} & \cos\theta_{1,1} \end{bmatrix}$$

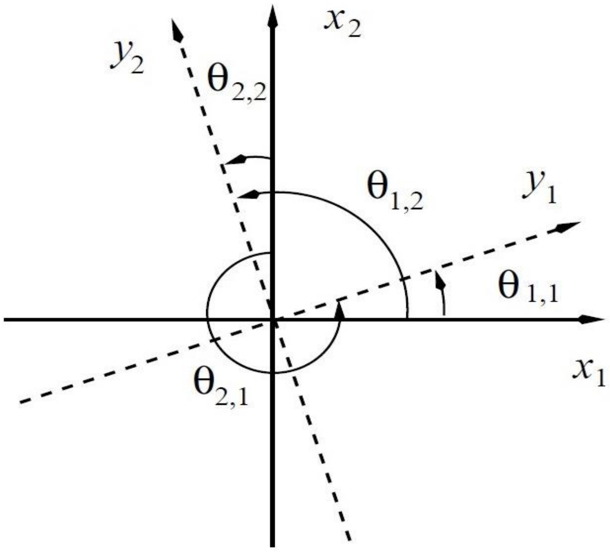


Figure 4-10 Orthogonal rotation in 2-dimensional space

4.3.3(b) VARIMAX

VARIMAX is the most popular orthogonal rotation technique, which was developed by Kaiser (1958) [24]. In statistics, VARIMAX rotation means changing of coordinates used in PCA that maximizes the sum of variances of the squared loadings (squared correlations between variables and PCs).

$$v = \sum (q_{j,\ell}^2 - \overline{q}_\ell^2)^2$$

Where $q_{j,\ell}$ being the loading of j^{th} variable of matrix loadings matrix $\mathbf{Q}$ of PC ℓ and $\overline{q}_\ell^2$ the squared mean of loading. VARIMAX simple solution implies each PC has a small quantity of large loading and a large number of small (or zero) loading.

If the loadings in each column were approximately equal, the variance would be close to 0. As the squared loadings teands 0, the variance will approach a maximum. Thus the VARIMAX technique attempts to make the loadings either large or small to facilitate interpretation [13].

The VARIMAX is available in most of factor (PC) analysis software programs, the output usually includes the rotated loading matrix $\mathbf{Q}^*$, the variance accounted for (sum of squares of each column of $\mathbf{Q}^*$), and the orthogonal rotation matrix **R** that used to obtain $\mathbf{Q}^* = \mathbf{QR}$.

4.3.3(c) Oblique Rotation

The aim of using the Oblique Rotation is to get a simple structure by relocation of factor axes. Oblique rotations strongly recommended by Thurstone [25], since PCs are orthogonal, so they are used more rarely than their orthogonal rotation methods.

4.4 Example

The data in Table A.5 contians library collections, staff and operating expenditures of the 60 largest college and Uni. libraries: Fiscal year 2008 [26]. The following variables are defined on the data set.

X_1 =Number of volumes at end of year (in thousands)

X_2 =Number of e-books at end of year

X_3 =Number of serials at end of year

X_4 =Technician

X_5 =Librarians

X_6 =Other expenses

X_7 =Salaries and wages

X_8 =Public service hours per typical week

X_9 =Gate count per typical week1

X_{10} =Reference transactions per typical week

Since the variables were measured using different units, they are standardized and the correlation matrix is used in PCA. The correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.1926 & 0.4975 & 0.8644 & 0.7801 & 0.8677 & 0.8657 & 0.0631 & 0.2709 & 0.4047 \\ 0.1926 & 1.0000 & 0.2063 & 0.0895 & 0.0231 & 0.1433 & 0.0700 & -0.1365 & 0.0867 & 0.0155 \\ 0.4975 & 0.2063 & 1.0000 & 0.4362 & 0.3050 & 0.4929 & 0.4195 & -0.0193 & 0.0953 & 0.0165 \\ 0.8644 & 0.0895 & 0.4362 & 1.0000 & 0.8906 & 0.9534 & 0.9707 & 0.1516 & 0.3157 & 0.3657 \\ 0.7801 & 0.0231 & 0.3050 & 0.8906 & 1.0000 & 0.8504 & 0.8744 & 0.2453 & 0.2984 & 0.3428 \\ 0.8677 & 0.1433 & 0.4929 & 0.9534 & 0.8504 & 1.0000 & 0.9724 & 0.0787 & 0.1416 & 0.2947 \\ 0.8657 & 0.0700 & 0.4195 & 0.9707 & 0.8744 & 0.9724 & 1.0000 & 0.0902 & 0.2045 & 0.3110 \\ 0.0631 & -0.1365 & -0.0193 & 0.1516 & 0.2453 & 0.0787 & 0.0902 & 1.0000 & 0.2157 & 0.0436 \\ 0.2709 & 0.0867 & 0.0953 & 0.3157 & 0.2984 & 0.1416 & 0.2045 & 0.2157 & 1.0000 & 0.3408 \\ 0.4047 & 0.0155 & 0.0165 & 0.3657 & 0.3428 & 0.2947 & 0.3110 & 0.0436 & 0.3408 & 1.0000 \end{pmatrix}$$

As seen from the correlation matrix, the linear correlation between variables ranges from very strong to very weak.

The ordered eigenvalues of the correlation matrix from highest to lowest are

$$\mathbf{l}^T = \left(5.0787\ \ 1.3459\ \ 1.0911\ \ 0.9236\ \ 0.6705\ \ 0.5527\ \ 0.1624\ \ 0.1318\ \ 0.0259\ \ 0.0175\right)$$

The matrix **G** is made up of eigenvectors $\mathbf{g}_j$ of **R**.

$$\mathbf{G} = \begin{pmatrix} 0.4104 & 0.0839 & -0.0493 & 0.0504 & 0.0041 & -0.0598 & -0.8249 & 0.3647 & -0.0324 & -0.0375 \\ 0.0665 & 0.4226 & -0.6327 & -0.3285 & -0.5435 & -0.0463 & 0.0921 & -0.0039 & -0.0320 & 0.0408 \\ 0.2288 & 0.3902 & -0.0345 & -0.3801 & 0.6724 & -0.3918 & 0.1899 & 0.0618 & -0.0278 & 0.0581 \\ 0.4327 & -0.0146 & 0.0792 & 0.0290 & -0.0443 & 0.1232 & 0.1072 & -0.3177 & -0.7264 & -0.3850 \\ 0.4017 & -0.1287 & 0.1401 & 0.0114 & -0.1591 & 0.1568 & 0.4826 & 0.7193 & 0.0426 & 0.0151 \\ 0.4242 & 0.1341 & 0.1445 & 0.0391 & -0.0982 & 0.0237 & 0.0602 & -0.3393 & 0.6715 & -0.4469 \\ 0.4258 & 0.0558 & 0.1586 & 0.0840 & -0.0871 & 0.1312 & 0.0166 & -0.3374 & 0.0255 & 0.8024 \\ 0.0710 & -0.5486 & 0.2138 & -0.6574 & -0.2511 & -0.3702 & -0.0877 & -0.0864 & 0.0143 & 0.0253 \\ 0.1527 & -0.4465 & -0.5471 & -0.2096 & 0.3805 & 0.5171 & -0.0235 & -0.0755 & 0.1251 & 0.0016 \\ 0.1911 & -0.3521 & -0.4231 & 0.5094 & 0.0114 & -0.6154 & 0.1356 & -0.0514 & 0.0164 & 0.0303 \end{pmatrix}$$

Table 2 lists the eigenvalues of the correlation matrix **R** in the first column, ratio of each eigenvalue to the total in the second column, and the cumulative proportion in the third column. From the third column it is evident that the first 4 eigenvalues which are the variance of the first 4 PCs, represents about 84% of the total variation in the data. Therefore, the use of the first 4 PCs is considered adequate for the representation of the data.

Table 2 : Example 4.4 The proportion of variance of PCs

Eigenvalue	Proportion of variance	Cumulated Proportion
l_i	$l_i \Big/ \sum_{j=1}^{p} l_j$	$\sum_{j=1}^{q} l_j \Big/ \sum_{j=1}^{p} l_j$
5.0787	0.507865	0.51
1.3459	0.134589	0.64
1.0911	0.109109	0.75
0.9236	0.092359	0.84
0.6705	0.067049	0.91
0.5527	0.055269	0.97
0.1624	0.01624	0.98
0.1318	0.01318	0.99566
0.0259	0.00259	0.99825
0.0175	0.00175	1

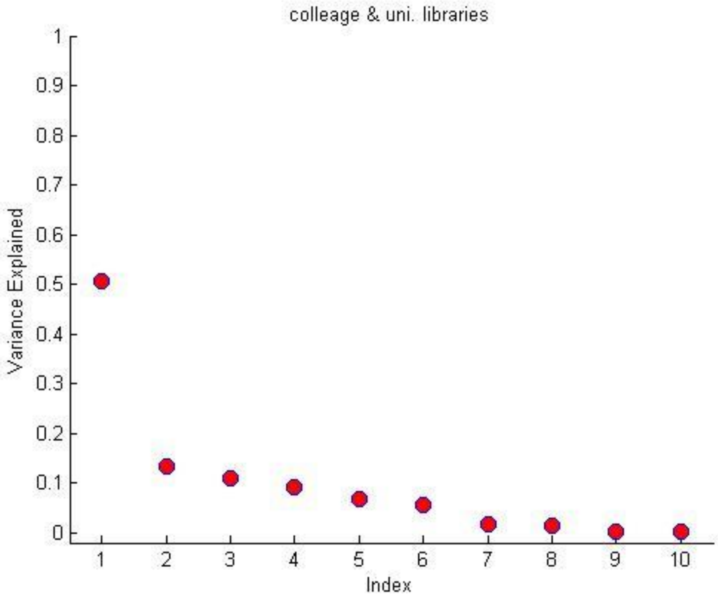


Figure 4-11: Example 4.4, The proportion of variance $l_i \Big/ \sum_{j=1}^{p} l_j$ of PCs

The coefficients used in the computation of the first four PCs that accounts for 84% of total variation are given in Table 3.

Table 3 : Example 4.4 Characteristics coefficients (weights or eigenvectors of the correlation matrix) for first 4 PCs for the PCA of libraries data.

Variables	g_1	g_2	g_3	g_4
X_1	0.4104	0.0839	-0.0493	0.0504
X_2	0.0665	0.4226	-0.6327	-0.3285
X_3	0.2288	0.3902	-0.0345	0.0290
X_4	0.4327	-0.0146	0.0792	-0.3801
X_5	0.4017	-0.1287	0.1401	0.0114
X_6	0.4242	0.1341	0.1445	0.0391
X_7	0.4258	0.0558	0.1586	0.0840
X_8	0.0710	-0.5486	0.2138	-0.6574
X_9	0.1527	-0.4465	-0.5471	-0.2096
X_{10}	0.1911	-0.3521	-0.4231	0.5094
Total	2.3945	2.5772	2.4229	2.2989
$g_{1i} / \sum_{j=1}^{10} g_{ji}$	0.17	0.03	0.02	0.2

The weights of PCs in Table 3 explain which variables are dominant in each PC. The first PC which accounts for 51% of total variation in the data, is highly influenced by the variables X_1, X_4, X_5, X_6 and X_7, and using $\mathbf{y}_j = (X - \mathbf{1}_n \bar{x}^T)\mathbf{g}_j$ can be written as

$$y_1 = 0.4104X_1 + 0.067X_2 + 0.228X_3 + 0.433X_4 + 0.402X_5 + 0.424X_6 + 0.426X_7 + 0.071X_8 + 0.1527X_9 + 0.191X_{10}$$

The second PC accounts for 13.5% of total variation is mainly composed of the difference between X_2, X_3 and X_8, X_9. This is given by

$$y_2 = 0.084X_1 + 0.423X_2 + 0.390X_3 - 0.015X_4 - 0.129X_5 + 0.134X_6 + 0.056X_7 - 0.549X_8 - 0.447X_9 - 0.352X_{10}$$

Similarly other PCs can be interpreted.

Scatter diagrams for PC1 versus PC2 and PC3 versus PC4 are given in Figure 4.12 and Figure 4.13 respectively. To highlight the effect of a variable on the PCs, the points on the scatter diagrams are marked as “o” if the X_1 value involved in the computation of the PC is less than $\bar{X}_1$, and those greater than the $\bar{X}_1$ are marked as “+”. In Figure 4.12 two groups forms reasonably separate scaters mainly due to the high influence X_1 has on PC1 (17% of weights assigned with PC1), compared with its low influence on PC2 (3% of weights assigned to PC2).

In Figure 4.13, two groups of points are mixed as the influence of X_1 on both PC3 and PC4 is about the same, but opposite in sign.

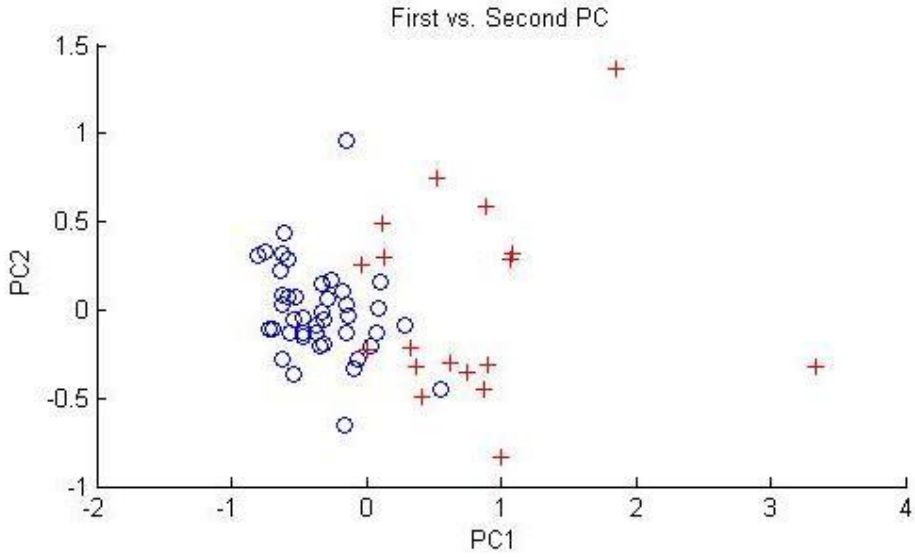


Figure 4-12 Example 4.4: PC1 versus PC2 of the college and Uni. Libraries data.

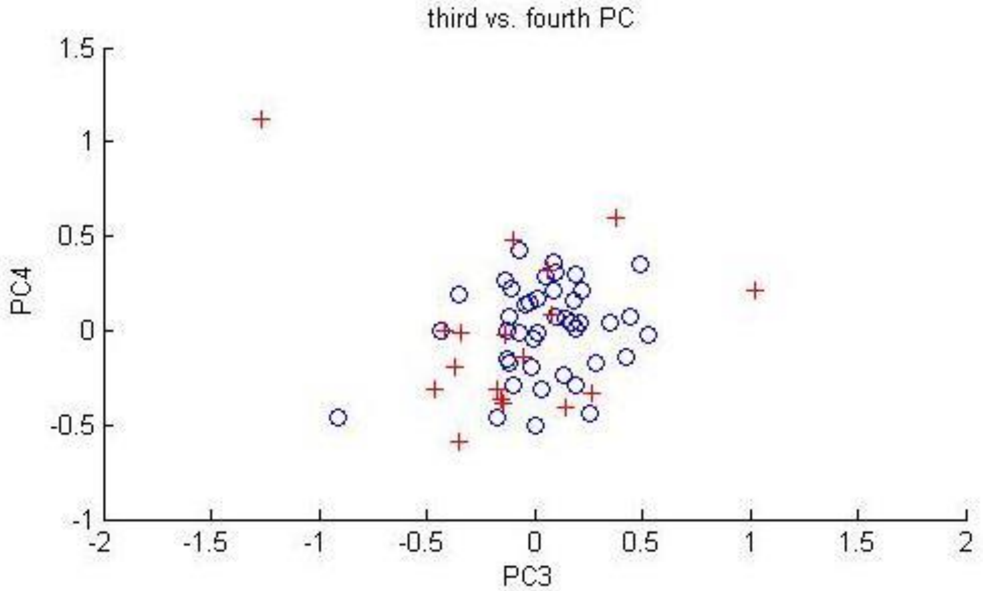


Figure 4-13 Example 4.4: PC3 versus PC4 of the college and Uni. Libraries data.

The correlation between original variable and PCs computed by (4.27) are given in table 4

Table 4: Example 4.4 the correlation between original variable X_i and PCs Y_1, Y_2 and Y_3

variables	$r_{X_iY_1}$	$r_{X_iY_2}$	$r_{X_iY_3}$	$r_{X_iY_4}$	$\sum_{j=1}^{p} r_{X_iY_j}$
X_1	0.9596	0.0973	-0.0514	0.0484	0.935284
X_2	0.1499	0.4903	-0.6609	-0.3157	0.799319
X_3	0.5156	0.4527	-0.0361	-0.3653	0.605528
X_4	0.9752	-0.0169	0.0828	0.0279	0.958935
X_5	0.9249	-0.1493	0.1463	0.0110	0.899255
X_6	0.9053	0.1555	0.1510	0.0376	0.867963
X_7	0.9560	0.0647	0.1657	0.0807	0.952091
X_8	0.1601	-0.6365	0.2233	-0.6318	0.879798
X_9	0.3441	-0.5179	-0.5715	-0.2014	0.753799
X_{10}	0.4307	-0.4085	-0.4419	0.4895	0.787261

From table 4: we can see that the first PC has a positive high correlation with X_1, X_4, X_5, X_6 and X_7. Thus these variables are well explained by first PC. This property is clearly visible in Figure 4.14, as all the correlation values pertaining to these variables lie on the right hand side on the circle. The second PC is well described by the difference between the sum of X_2 and X_3 and the sum of X_8 and X_9. The position of these variables on Figure 4.14 clearly indicates this.

Figure 4.15 shows the same correlation regarding the second PC as in Figure 4.14. X_2,X_9 and X_{10} have negative effect on the third PC as they are below the 0 line on the vertical axis. In Figure 4.16 it is clear to see that the variables X_2,X_9 and X_{10} lie on the left hand side on the circle, this means these variables have negative correlation with 3^{rd} PC. The 4^{th} PC depicts the difference between X_{10} and the sum of X_2,X_3, X_8 and X_9.

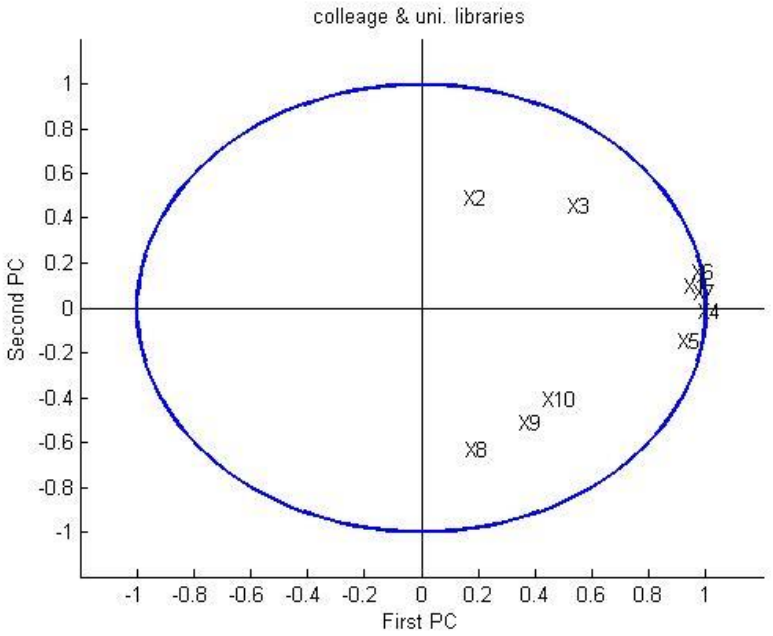


Figure 4-14 Example 4.4 Correlation between original variables X_i and PCs Y_1, Y_2

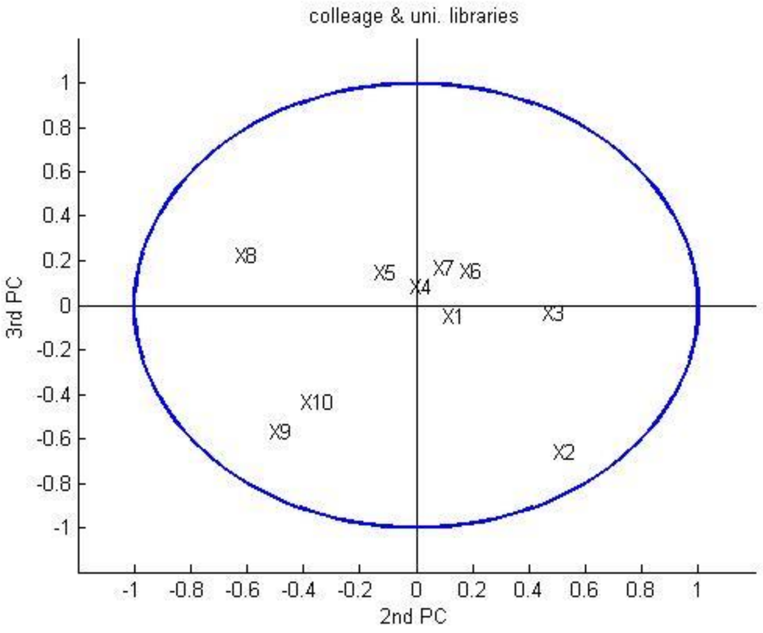


Figure 4-15 Example 4.4 Correlation between original variables X_i and PCs Y_2, Y_3.

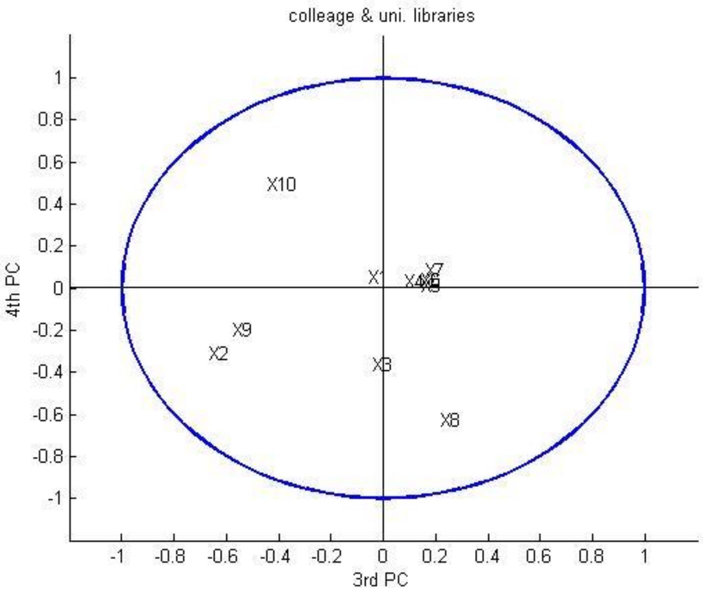
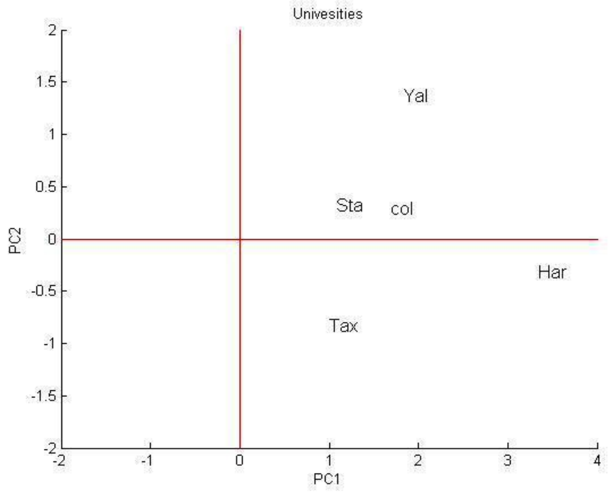


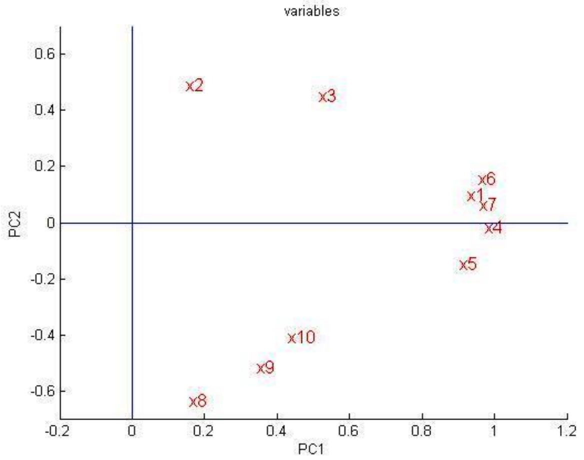
Figure 4-16 Example 4.4 Correlation between original variables X_i and PCs Y_3, Y_4.

The theory given in 4.5 (Duality Relations) is applied to the data in Appendix A shows the relationship between the variables $(X_1, X_2, \ldots\ldots, X_{10})$ and the representation of universites (obsevations) in two dimensions. PCs obtained from $\mathbf{X}^T\mathbf{X}$ (Figure 4.17) and from $\mathbf{X}\mathbf{X}^T$ (Figure 4.18). It indicates that for Harvard Uni. it has the highest full-time equivalent value for Technician and Librarians (X_4 and X_5). Similarly Yale Uni. has the largest number of serials (X_3) at end of year.



Universities	PC1	PC2
Harvard	3.3403	-0.3187
Yale	1.8513	1.3715
Columbia	1.688	0.2845
Texas	1.0027	-0.836
Stanford	1.0783	0.3161

Figure 4-17 Some outliers universities explanation by the first and second PC



variables	PC1	PC2
X_1	0.9249	0.0973
X_2	0.1499	0.4903
X_3	0.5156	0.4527
X_4	0.9752	-0.0169
X_5	0.9053	-0.1493
X_6	0.9560	0.1555
X_7	0.9596	0.0647
X_8	0.1601	-0.6365
X_9	0.3441	-0.5179
X_{10}	0.4307	-0.4085

Figure 4-18 Staff, and operating expenditures of Uni.s (variables) in 2-dimension

Chapter 5

CONCOLUSION

High dimensional data has been reduced by finding an orthogonal transformation. This transform generated a new set of uncorrelated variables called *principal components* that are combination of the original variables without losing the importance of information inherent to the data.

The first component has the largest possible variance, i.e. it represents the largest proportion of the total variance. Second PC has the second largest variance and so on. After the PCs are computed, examine the correlation between the original variables and these components.

PCA is regarded as a data reduction technique. This means, the use of the first few PCs that represents the great majority of variation in the data (preferably over 80%), facilitates the analysis of a large data set with many variables by only analyzing the first few PCs.

An application example with 10 variables with 60 observations for each variable are studied, and it is found that the first 4 PCs represented 84% of the total variation in the data set. This greatly reduces the load of work in the further analysis of the data. Interpretations of the correlation between the variables and the PCs give a good idea about the variables that have high influence on the PCs.

REFERENCE

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical,* vol. 2, no. 6, pp. 559-572., 1901.

[2] I. Jolliffe, Principal Component Analysis, New York: Sipringar, Apr 2002.

[3] H. Hotelling, "analysis of complex of statistical variables in to principal component," *Educational Psychology,* no. 24, pp. 417-441, 498-520, 1933.

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

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

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

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

[8] G. J. .C., "Some distance properties of latent root and vector methods used in,"

Biometrika , no. 53, pp. 325-38, 1966.

[9] J. N. R. Jeffers, "Two Case Studies in the Application of Principal Component Analysis," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 16, no. 3, pp. 225-236, 1967.

[10] C. D. Meyer, Matrix analysis and applied linear algebra, Pheladelphia: Society of indestrial and applied mathematics, Feb. 15, 2001.

[11] K. Baker, "Singular Value Decomposition Tutorial," Ohio State University, Ohio , Jan 2013.

[12] J. R. Movellan, Introduction to Probability Theory and statistics, Javier R. Movellan, August 21, 2008.

[13] A. C. RENCHER, Methods of Multivariate Analysis, Brigham Young University: A JOHN WILEY & SONS, INC. PUBLICATION, 2002.

[14] M. J. Z. &. W. M. Jr., Data Mining and Analysis: Foundations concepts and Algorithms, United kingdom: Cambridge University press, 2013.

[15] D. C. Lay, Linear Algebra and iits applications, New York: Pearson Education Inc, 2012.

[16] D. Poole, Linear Algebra: A Modern Introduction, 3rd edition, Bosten USA: Brooks/Cole, Cengage Learning, 2011.

[17] L. S. Wolfgang Hardle, Applied Multuivariate Statistical Analysis, New York: Sipringer, 29th April 2003.

[18] L. Eniksson, J. Byme, J. Trygg and E. Johansson, Multi- and Megavariate Data Analysis Basic Principles and Applications part1, New York: Umetrics, Inc., 2006.

[19] L. J. Williams and H. Abdi, "Principal Component Analysis," *John Wiley & Sons,* p. 433, july/Aug 2010.

[20] C. RB, "The scree test for the number of factors," *Taylor & Francis Online*, pp. 245-276, Jun 2010.

[21] M. R. Bandalos and D. L. Boehm-Kaufman, "Four common misconceptions in exploratory factor analysis," *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences,* vol. XIX, no. 412, pp. 61-87, 2009.

[22] R. Ledesma and P. Valero-Mora, "Determining the Number of Factors to Retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis," *Practical Assessment Research & Evaluation* , vol. II, no. 12, pp. 1-11, 2007.

[23] A. E. B. T. F. L. Michael S. Lewis-Beck, The SAGE Encyclopedia of Social Science Research Methods, California, USA: SAGE publications Inc., 2004.

[24] K. HF, "The varimax criterion for analytic rotation," *Psychometrika,* no. 23, p. 187–200, 1958.

[25] L. L. Thurstone, "Multiple-factor analysis," *Journal of Clinical Psychology,* vol. 4, no. 2, p. 224, 1948.

[26] U.S. Department of Education, Institute of Education Sciences National, Center for Education Statistics, july 2010. [Online]. Available: http://nces.ed.gov.

[27] R. D. Ledesma, "Determining the Number of Factors to Retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis," *Practical Assessment, Research & Evaluation,* vol. 12, no. 2, pp. 1-11, Feb 2007.

APPENDICES

Appendix A: Table 5 Data of Example 4.4

Collections. staff. and operating expenditures of the 60 largest college and Uni. libraries: Fiscal year 2008 [25].

Institution	Rank order, by number of volumes	Number of volumes at end of year (in thousands)	Number of e-books at end of year	Number of serials at end of year	Full-time-equivalent staff		Operating expenditures (in thousands)		Public service hours per typical week	Gate count per typical week[1]	Reference trans-actions per typical week
					Technician	Librarians	Other expenses	Salaries and wages			
Harvard Uni, (MA)	1	16,250	1,167	110,628	1,229	418	117,884	62,798	168	39,748	5,468
Yale Uni, (CT)	2	12,284	840,000	295,557	735	175	92,248	35,781	111	14,900	1,970
Uni, of California, Berkeley	3	11,020	610,920	87,876	487	92	48,020	24,305	77	27,502	2,100
Uni, of Illinois at Urbana-Champaign	4	10,933	319,533	109,803	473	113	40,571	20,988	144	85,632	6,214
Columbia Uni, in the City of New York	5	9,596	703,121	132,740	616	161	56,089	27,240	108	81,862	3,557
Uni, of Texas at Austin	6	9,447	593,450	56,847	528	130	43,850	20,773	120	87,115	20,693
Uni, of Michigan, Ann Arbor	7	9,175	701,019	69,457	570	169	52,395	25,853	168	73,543	2,884
Stanford Uni, (CA)	8	8,558	419,515	33,903	680	151	78,377	41,382	105	20,100	3,074
Uni, of California, Los Angeles	9	8,467	495,238	175,207	596	125	53,154	28,197	97	64,072	1,843
Uni, of Wisconsin, Madison	10	7,934	766,032	54,164	553	229	43,282	23,459	148	110,368	2,640

Cornell Uni, (NY)	11	7,750	391,897	89,000	549	118	46,798	22,667	146	98,000	1,497
Uni, of Chicago (IL)	12	7,745	851,880	76,607	323	68	34,680	12,638	146	33,881	779
Indiana Uni,, Bloomington	13	7,618	631,617	103,228	445	94	36,282	16,061	168	90,061	2,446
Uni, of Minnesota, Twin Cities	14	6,878	307,082	85,075	394	93	40,734	18,118	100	36,527	2,300
Uni, of Washington, Seattle Campus	15	6,844	387,281	61,847	458	135	36,814	19,345	138	116,000	2,128
Princeton Uni, (NJ)	16	6,779	763,158	51,746	410	97	48,970	18,789	116	13,492	671
Uni, of North Carolina at Chapel Hill	17	6,017	510,110	60,713	452	143	41,124	18,944	146	60,214	2,543
Ohio State Uni,, Main Campus	18	6,016	269,097	78,903	396	62	35,833	16,642	168	39,030	1,476
Duke Uni, (NC)	19	5,829	144,939	61,964	369	117	37,331	16,444	161	9,250	2,638
Uni, of Pennsylvania	20	5,756	340,446	61,676	370	111	37,599	16,991	111	38,589	5,000
Uni, of Pittsburgh, Main Campus (PA)	21	5,657	591,468	59,141	382	120	32,907	12,539	118	84,789	2,587
Pennsylvania State Uni,, Main Campus	22	5,355	42,083	88,668	608	134	47,686	24,437	168	46,247	3,549
Uni, of Arizona	23	5,266	645,463	24,466	239	54	24,676	9,471	142	42,916	531
Uni, of Virginia, Main Campus	24	5,158	374,731	163,032	379	101	35,930	16,921	149	76,424	2,886
Rutgers Uni,, New Brunswick/Piscataway	25	5,081	195,296	74,031	305	66	23,918	13,651	108	53,419	1,216
New York Uni,	26	5,073	545,025	67,960	458	58	44,603	20,703	119	51,500	2,156
Northwestern Uni, (IL)	27	4,843	264,066	82,822	344	97	29,147	12,518	126	28,218	1,427
Michigan State Uni,	28	4,839	66,350	83,460	265	71	23,482	10,714	148	42,367	850
Uni, of Kansas	29	4,799	321,320	60,838	228	54	19,543	9,105	140	42,000	2,350
Uni, of Iowa	30	4,791	486,769	59,442	281	98	27,620	12,335	113	36,273	1,610
Uni, of Oklahoma, Norman Campus	31	4,702	649,929	52,522	158	37	16,253	4,396	117	21,930	523

Uni, of Georgia	32	4,637	128,694	80,748	315	81	24,451	10,106	137	17,700	1,910
Arizona State Uni, at the Tempe Campus	33	4,422	302,266	87,566	332	93	28,571	12,266	149	75,265	2,053
Uni, of Florida	34	4,288	280,238	71,336	402	85	29,731	13,905	111	56,209	1,587
Uni, of Southern California	35	4,084	267,657	70,066	374	75	38,393	17,149	159	53,534	1,173
Louisiana State Uni, and Agricultural & Mechanical College	36	4,067	346,389	101,738	192	52	15,874	—	113	32,228	712
Texas A & M Uni,	37	3,934	461,225	86,737	359	85	34,150	12,329	146	49,683	880
Uni, of Colorado at Boulder	38	3,928	175,377	55,519	216	58	21,454	8,693	104	40,532	1,374
Uni, of South Carolina, Columbia	39	3,885	91,940	21,505	275	71	19,743	7,975	140	31,415	2,969
Johns Hopkins Uni, (MD)	40	3,878	2,003,184	74,701	338	80	32,881	13,282	120	19,373	1,593
Washington Uni, in St, Louis (MO)	41	3,841	382,891	69,400	266	93	32,366	10,219	120	30,000	1,409
Brown Uni, (RI)	42	3,825	284,749	60,499	208	55	19,862	9,162	112	20,064	510
Brigham Young Uni, (UT)	43	3,743	337,546	69,361	383	85	27,167	12,126	105	82,238	3,070
SUNY at Buffalo (NY)	44	3,720	369,721	80,431	242	60	19,972	10,339	168	26,000	562
Uni, of Kentucky	45	3,720	406,014	73,251	287	79	21,414	8,257	135	57,316	1,734
Miami Uni, (OH)	46	3,718	511,114	91,229	146	41	9,488	4,652	168	28,862	1,529
Uni, of Maryland, College Park	47	3,717	88,393	42,393	258	119	32,156	12,600	162	47,982	5,186
Uni, of Rochester (NY)	48	3,701	51,134	28,561	207	89	24,850	8,949	119	4,478	1,004
Uni, of Cincinnati, Main Campus (OH)	49	3,632	459,542	86,363	185	49	21,466	7,729	95	26,700	1,600
Uni, of Hawaii at Manoa	50	3,559	193,133	55,276	237	64	17,860	9,108	135	31,380	1,791
Uni, of Nebraska, Lincoln	51	3,554	321,180	46,865	187	49	12,633	6,465	96	15,004	1,000

Uni, of Missouri, Columbia	52	3,494	25,434	38,364	198	55	17,025	6,386	114	36,426	1,374
Florida State Uni,	53	3,484	477,476	62,093	250	67	18,563	7,090	142	37,531	1,219
North Carolina State Uni, at Raleigh	54	3,477	401,497	67,995	268	98	23,296	10,960	146	37,649	718
Wayne State Uni, (MI)	55	3,454	206,736	20,384	247	53	20,802	9,349	142	38,599	916
Uni, of Notre Dame (IN)	56	3,393	2,295	82,866	260	60	24,077	10,306	126	19,191	497
Uni, of Utah	57	3,373	132,859	48,777	370	69	26,290	12,877	123	39,724	3,680
Uni, of California, San Diego	58	3,373	231,216	34,800	367	63	30,748	16,330	114	51,347	880
Uni, of Connecticut	59	3,368	338,682	71,371	152	61	16,262	9,420	114	51,539	303
Uni, of California, Davis	60	3,354	504,736	50,442	231	53	18,652	9,568	95	33,978	1,129

Appendix B: Matlab Code of Example 4.4

```
close all
clc
clear

x=load('2013.dat');
[n p]=size(x)

y=vertcat(ones(n/2,1),zeros(n/2,1));

h=diag(ones(n,1))-ones(n,n)./n;   % Centering Matrix
y=mean(x);
a=x-repmat(y,n,1);            % Substracts mean
d=diag(1./sqrt(sum(a.*a)'/n));
xs=h*x*d;
xs=xs./sqrt(n);
rr=xs'*xs
[gamma lambda1]=eigs(rr,p,'la')     % Eigenvalues sorted by
size from largest to smallest(Note: Command generates a
Warning(Disregard it))
lambda=(lambda1*ones(p,1))';          % Turns Eigenvalue matrix
into a row vector
w1=gamma.*sqrt(repmat(lambda,p,1))   % coordinates of food
w=w1(:,1:2)                          % Two eigenvectors with
highest eigenvalues

z1=xs*gamma;         % coordinates of families
pc=sqrt(n/p).*z1;    % xs' scaled by square root of p
[f l]=size(pc)
z=pc(:,1:4);
aa=corr(pc);
%pc(:,1:4) =
rotatefactors(pc(:,1:4),'Method','varimax','Coeff',gamma)
s=sum(lambda);
e1=lambda/s;

r=horzcat(pc,a);

r=corr(r);
r1=r(11:20,1:4);

y=vertcat(ones(n/2,1),zeros(n/2,1));
%Plotting relative proportion of variance explained by PCs
nr=1:p;
figure(2)
scatter(nr,e1,75,'MarkerFaceColor','r')
xlabel('Index')
ylabel('Variance Explained')
title('colleage & uni. libraries')
xlim([0.5 6.5])
ylim([-0.02 1])
%plot(nr,e1,'r')
```

```
%Plot the correlation of the original variable with the PCs.
figure
hold on

%Plotting Eigenvalues
subplot(2,2,4,'FontSize',10)
gscatter(pc(:,3),pc(:,4),y,'bb','oo',7,'off')
xlabel('PC3 ')
ylabel('PC4 ')
title('third vs. fourth PC')


%Plot of the first vs. second PC
subplot(2,2,1,'FontSize',10)
gscatter(pc(:,1),pc(:,2),y,'bb','oo',7,'off')
xlabel('PC1 ')
ylabel('PC2 ')
title('First vs. Second PC')

%Plot of the second vs. third PC
subplot(2,2,2,'FontSize',10)
gscatter(pc(:,2),pc(:,3),y,'bb','oo',7,'off')
xlabel('PC2 ')
ylabel('PC3 ')
title('Second vs. Third PC')

%Plot of the first vs. third PC
subplot(2,2,3,'FontSize',10)
gscatter(pc(:,1),pc(:,3),y,'bb','oo',7,'off')
xlabel('PC1 ')
ylabel('PC3 ')
title('First vs. Third PC')
hold off

%Plot the correlation of the original variable with the PCs.
figure
hold on
xlim([-1.2 1.2])
ylim([-1.2 1.2])
line([-1.2 1.2],[0 0],'Color','k')
line([0 0],[1.2 -1.2],'Color','k')
title('colleage & uni. libraries')
xlabel('First PC')
ylabel('Second PC')

circle = rsmak('circle');
fnplt(circle)

text(r1(1,1),r1(1,2),'X1')
text(r1(2,1),r1(2,2),'X2')
text(r1(3,1),r1(3,2),'X3')
text(r1(4,1),r1(4,2),'X4')
text(r1(5,1),r1(5,2),'X5')
text(r1(6,1),r1(6,2),'X6')
text(r1(7,1),r1(7,2),'X7')
text(r1(8,1),r1(8,2),'X8')
text(r1(9,1),r1(9,2),'X9')
text(r1(10,1),r1(10,2),'X10')
hold off

figure
hold on
xlim([-1.2 1.2])
ylim([-1.2 1.2])
line([-1.2 1.2],[0 0],'Color','k')
line([0 0],[1.2 -1.2],'Color','k')
title('colleage & uni. libraries')
xlabel('1st PC')
ylabel('3rd PC')

circle = rsmak('circle');
fnplt(circle)

text(r1(1,1),r1(1,3),'X1')
text(r1(2,1),r1(2,3),'X2')
text(r1(3,1),r1(3,3),'X3')
text(r1(4,1),r1(4,3),'X4')
text(r1(5,1),r1(5,3),'X5')
text(r1(6,1),r1(6,3),'X6')
text(r1(7,1),r1(7,3),'X7')
text(r1(8,1),r1(8,3),'X8')
text(r1(9,1),r1(9,3),'X9')
text(r1(10,1),r1(10,3),'X10')
hold off

figure
hold on
xlim([-1.2 1.2])
ylim([-1.2 1.2])
line([-1.2 1.2],[0 0],'Color','k')
line([0 0],[1.2 -1.2],'Color','k')
title('colleage & uni. libraries')
xlabel('2nd PC')
ylabel('3rd PC')

circle = rsmak('circle');
fnplt(circle)

text(r1(1,2),r1(1,3),'X1')
text(r1(2,2),r1(2,3),'X2')
text(r1(3,2),r1(3,3),'X3')
text(r1(4,2),r1(4,3),'X4')
text(r1(5,2),r1(5,3),'X5')
text(r1(6,2),r1(6,3),'X6')
text(r1(7,2),r1(7,3),'X7')
text(r1(8,2),r1(8,3),'X8')
text(r1(9,2),r1(9,3),'X9')
text(r1(10,2),r1(10,3),'X10')
hold off
figure
[X,Y,Z] = sphere(16);
```

```
xx = pc(:,1);
y = pc(:,2);
z = pc(:,3);
xlabel('First PC')
ylabel('Second PC')
zlabel('Third PC')
scatter3(xx,y,z,'MarkerFaceColor','g')

figure
hold on
xlim([-1.2 1.2])
ylim([-1.2 1.2])
line([-1.2 1.2],[0 0],'Color','k')
line([0 0],[1.2 -1.2],'Color','k')
title('colleage & uni. libraries')
xlabel('3rd PC')
ylabel('4th PC')

circle = rsmak('circle');
fnplt(circle)

text(r1(1,3),r1(1,4),'X1')
text(r1(2,3),r1(2,4),'X2')
text(r1(3,3),r1(3,4),'X3')
text(r1(4,3),r1(4,4),'X4')
text(r1(5,3),r1(5,4),'X5')
text(r1(6,3),r1(6,4),'X6')
text(r1(7,3),r1(7,4),'X7')
text(r1(8,3),r1(8,4),'X8')
text(r1(9,3),r1(9,4),'X9')
text(r1(10,3),r1(10,4),'X10')
hold off

max=load('max.dat')
pc1=max(:,1:2)
pc2=max(:,3:4)
namepc1=['Harvard '
         'Yale    '
         'columbia'
         'Taxas   '
         'Stanford'];

   %Universities
figure
hold on
title('Univesities');
xlabel('PC1');
ylabel('PC2');
xlim([-2 4]);
ylim([-2 2]);

line([-2 4],[0 0],'Color','r');
line([0 0],[-2 2],'Color','r');

for i=1:5
    text(pc1(i,1),pc1(i,2),namepc1(i,1:3),'FontSize',12);
```

```
end;



namew=['x1 '
       'x2 '
       'x3 '
       'x4 '
       'x5 '
       'x6 '
       'x7 '
       'x8 '
       'x9 '
       'x10'];

figure
hold on
title('variables');
xlabel('PC1');
ylabel('PC2');
xlim([-0.2 1.2]);
ylim([-0.7 0.7]);

line([-0.2 1.2],[0 0],'Color','b');
line([0 0],[1 -1],'Color','b');

for i=1:p

text(w(i,1),w(i,2),namew(i,1:3),'Color','r','FontSize',12);
end;
```